Computer Science & Technology

David C. Wyld,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

- 9th International Conference on Foundations of Computer Science & Technology (CST 2022)
- International Conference on NLP and Machine Learning Trends (NLMLT 2022)
- 13th International conference on Database Management Systems (DMS 2022)
- 3rd International Conference on Cloud and Big Data (CLBD 2022)
- 11th International Conference on Information Technology Convergence and Services (ITCS 2022)
- 3rd International Conference on VLSI & Embedded Systems (VLSIE 2022)

Published By

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Dhinaharan Nagamalai (Eds),
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-925953-79-4
DOI: 10.5121/csit.2022.121901 - 10.5121/csit.2022.121911

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

9th International Conference on Foundations of Computer Science & Technology (CST 2022), November 12 ~ 13, 2022, Chennai, India, International Conference on NLP and Machine Learning Trends (NLMLT 2022), 13th International conference on Database Management Systems (DMS 2022), 3rd International Conference on Cloud and Big Data (CLBD 2022), 11th International Conference on Information Technology Convergence and Services (ITCS 2022), 3rd International Conference on VLSI & Embedded Systems (VLSIE 2022) was collocated with 9th International Conference on Foundations of Computer Science & Technology (CST 2022). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CST 2022, NLMLT 2022, DMS 2022, CLBD 2022, ITCS 2022 and VLSIE 2022. Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically.

In closing, CST 2022, NLMLT 2022, DMS 2022, CLBD 2022, ITCS 2022 and VLSIE 2022 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CST 2022, NLMLT 2022, DMS 2022, CLBD 2022, ITCS 2022 and VLSIE 2022

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld,
Dhinaharan Nagamalai (Eds)

General Chair

Organization

David C. Wyld, Southeastern Louisiana University, USA
Dhinaharan Nagamalai (Eds) Wireilla Net Solutions, Australia

Program Committee Members

Abd Abrahim Mosslah, University of Anbar, Iraq
Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Abdelkader khobzaoui, Djillali LIABES University, Algeria
Abdellatif I. Moustafa, Umm AL-Qura University, Saudi Arabia
Abdelouafi ikidid, Cadi Ayyad University, Morocco
Abderrahim Siam, University of Khenchela, Algeria
Abhishek Shukla, R D Engineering College, India
Addisson Salazar, Universitat Politècnica de València, Spain
Adnan Aldemir, Van Yüzüncü Yil University, Turkey
Adrian Olaru, University Politehnica of Bucharest, Romania
Ahmad A. Saifan, Yarmouk University, Jordan
Ahmed Farouk AbdelGawad, Zagazig University, Egypt
Ajay Anil Gurjar, Sipna College of Engineering and Technology, India
Ajit Singh, Patna University, India
Alexander Gelbukh, Instituto Politécnico Nacional, Mexico
Alireza Valipour Baboli, University Technical and Vocational, Iran
Amal Azeroual, Mohammed V University, Morocco
Amando P. Singun Jr, University of Technology and Applied Sciences, Oman
Amari Houda, Networking & Telecom Engineering, Tunisia
Amizah Malip, University of Malaya, Malaysia
Anand Nayyar, Duy Tan University, Viet Nam
Anas Alsobeh, Yarmouk University, Jordan
Anjan Krishnamurthy, BMS Institute of Technology and Management, India
António Abreu, Polytechnic Institute of Lisbon, Portugal
Arnaud Soulet, University of Tours, France
Ayad Ghany Ismaeel, Rector of Al-Kitab University, Iraq
B.K. Tripathy, Vellore Institute of Technology, India
Benyamin Ahmadnia, California State University, USA
Bo Li, Harbin Institute of Technology, Weihai, China
Borne Pierre, University of Lille, France
Bouchra Marzak, Hassan II University, Morocco
Boukari nassim, Skikda University, Algeria
Brahami Menaouer, National Polytechnic School of Oran, Algeria
Brahim Lejdel, University of El-Oued, Algeria
Cheng Siong Chin, Newcastle University, Singapore
Christian Mancas, Ovidius University, Romania
Corneliu Doroftei, Alexandru Ioan Cuza University of Iasi, Romania
Daniel Hunyad, Lucian Blaga University of Sibiu, Romania
Dário Ferreira, University of Beira Interior, Portugal
Debjani Chakraborty, Indian Institute of Technology, India
Dibya Mukhopadhyay, University of Alabama, USA
Dimitris Kanellopoulos, University of Patras, Greece
Divya C D, Vidyavardhaka College of Engineering, India
Elżbieta Macioszek, Silesian University of Technology, Poland

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

9th International Conference on Foundations of Computer Science & Technology (CST 2022)

International Conference on NLP and Machine Learning Trends (NLMLT 2022)

13th International conference on Database Management Systems (DMS 2022)

3rd International Conference on Cloud and Big Data (CLBD 2022)

11th International Conference on Information Technology Convergence and Services (ITCS 2022)

3rd International Conference on VLSI & Embedded Systems (VLSIE 2022)

GENERIC AND ACCESSIBLE GESTURE CONTROLLED AUGMENTED REALITY PLATFORM

Arya Rajiv Chaloli, K Anjali Kamath,
Divya T Puranam and Prof. Preet Kanwal

Department of Computer Science, PES University Bangalore, India

ABSTRACT

Augmented Reality (AR) is one of the most popular trends in, technology today. Its accessibility has heightened as new smartphones and other devices equipped with depth-sensing cameras and other AR- related technologies are being introduced into the market. AR helps the user view the real-world environment with virtual objects augmented in it. With advances in AR technology, Augmented Reality is being used in a variety of applications, from medicine to games like Pokémon Go, and to retail and shopping applications that allow us to try on clothes and accessories from the comfort of our homes. In the current times, being hands-off is utterly important due to the widespread attack of the COVID19 pandemic. Thus, an application where the necessity to touch a screen is removed would be highly relevant. In such a scenario, AR comes into play. Essentially, it helps to convert the contents on a physical screen to a virtual object that is seamlessly augmented into reality and the user interacts only with the virtual object, thus avoiding any requirement to touch the actual physical screen and making the whole system touch-free.

Hence, the paper aims to propose an Augmented Reality application system that can be integrated into our day-to-day life. With the advent of technology and the digitization of almost all interfaces around us, the opportunity of augmenting these digitized resources has escalated. To demonstrate a generic application that can be used along with any digitized resource, an AR system will be implemented in the project, that can control a personal computer (PC) remotely through hand gestures made by the user on the augmented interface. The methodology proposed targets to build a system that is both generic and accessible. The application is made generic by making the AR system be able to provide an AR interface to any application that can run on a regular PC. The accessibility of the AR system is improved by making it compatible to work on any normal smartphone with a regular camera. There is no necessity for a depth- sensing camera, which is a requirement of popular AR toolkits like ARCore and ARKit.

KEYWORDS

Augmented reality, Gesture Recognition, Generic, Application, Technology, Accessibility.

1. INTRODUCTION

Augmented Reality (AR) provides an interactive experience of virtual objects augmented into the real-world environment by enhancing both real and virtual objects using computer-generated perceptual information, which may span across multiple sensory modalities. It can also be defined as a system consisting of 3 parts namely, the combination of real and virtual worlds, real-time interaction, and the accurate recognition and registration of real and virtual objects. Using this, objects appear as though they are part of our environment but in reality, it is just a simulation. In

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 01-10, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121901

this way, AR can be used to alter one's ongoing perception of the real world and create new experiences [1, 2].

Modified reality systems are of three types. Augmented Reality adds virtual objects to a live view of the real world by making use of the camera on a smartphone or a headset. Virtual Reality, on the other hand, is a completely virtual experience that contains no aspects of the real world. This is done using VR devices like HTC Vive, Oculus Rift, or Google Cardboard. A Mixed Reality (MR) experience combines the elements of both AR and VR, i.e., it is the interaction between real-world and virtual objects.

Augmented Reality is a newly emerging field and has been used in fields such as medical training, retail, design and modelling, education [3], and entertainment, in the past few years. However, due to the field still being in its formative stages, it still lacks the genericness that it should have and is rather very specific to the applications that it is being used for. Some of the current limitations [4] in the field of Augmented Reality include:

- Hardware:
  - Equipment: Very few mobile devices available in the market are equipped with room mapping or depth sensing technology.
  - Specifications: Limited processing power, a small amount of memory, and limited storage available during the deployment phase (phone apps).
- Blending with Reality: Rendering digital data into meaningful and easily understandable graphics and scaling it to fit the perspective of the visual field is challenging.
- AR education: Most consumers are not familiar with AR technology and do not see its applications in their daily lives.
- Possibility of physical harm: There have been many cases of people hurting themselves while playing the popular AR game Pokémon Go.
- Lack of proven business models: There are not many industries that have found an AR-related business model that will work long term, except the gaming industry.
- Privacy and Security: Augmented Reality identification and security policies are not quite developed.

The proposed methodology aims to develop a generic application with high accessibility that would reduce contact with physical surfaces. The user would operate through an AR screen instead of using the surface directly. However, for demonstration and feasibility testing, the scope is restricted to the deployment of an AR interface for a browser application (that runs on a remote PC) on android smartphones With the COVID19 pandemic on the rise, there is a heightened need to maintain social distancing. Washing and sanitizing are of the utmost importance when anything in a public place is touched. As it goes, prevention is better than cure, and hence such situations must be avoided. Through the project, the user can just look at the screen through his/her phone and perform operations using gestures [5, 6] in the air and his/her interaction with the software is completely handsfree. Hence one can avoid coming in contact with public surfaces. Additionally, the screen is completely restricted to the user and hence can provide enhanced security and privacy.

2. RELEVANT WORK

Broadly, the relevant work in this domain falls under three categories: gesture recognition, augmented reality, and gesture recognition in augmented reality.

2.1. Gesture Recognition

There are a variety of methods that are used for extracting the necessary information that is required for gesture recognition systems which include hardware devices like data gloves and colour markers and using the appearance of the hand and skin colour to segment the hand.

Authors Khan, Rafiqul Zaman, and Noor Adnan Ibraheem in [7] explains how gesture recognition systems are mainly classified into three steps: extraction method, features extraction, and classification.

- The extraction method/segmentation is the first step. It consists of dividing the input image into regions separated by boundaries. This is done differently based on whether the gesture is static or dynamic. Some of the methods used to model the hand are the Gaussian Model, Gaussian Mixture Model, and histogram-based techniques.
- In the second step, the features of the segmented image are extracted. Some extraction methods use the contour and silhouette of the hand while others use fingertip positions, palm centres, etc.
- The last step is gesture classification. Various techniques and models like neural networks, FSMs, PCA, Fuzzy CMeans clustering, and genetic algorithms are used for this purpose.

The authors of [7] compare the recognition methods used in hand gesture recognition systems. A summary of extraction methods, feature representation, and recognition of hand gesture recognition systems is listed. They also describe the various methods used for gesture recognition like Neural Networks, HMMs, fuzzy, C-means clustering, etc (which are alternatives for orientation histogram for feature representation). Reference [7] provides more clarity and an insight into the various applications, internal working, limitations, and other specifications of different gesture recognition systems.

2.2. Augmented Reality

Reference [8] deals with the fundamentals of Augmented Reality. As described in [8], the four fundamental parts of Augmented Reality are AR components, AR devices, the applications of Augmented Reality, and visualization issues that could be encountered.

The scene generator, tracking system, and display are the components of an Augmented Reality system. The scene generator is used to set the scene of the Augmented Reality environment. This is a challenge since, unlike a Virtual Reality system, an Augmented Reality system needs to account for the real environment that the virtual object is being placed in. The second component is the tracking system used to track both the virtual and real objects in the Augmented Reality environment. The last component is the display features of the Augmented Reality system.

From [8], the complexities of an AR System are understood, and the importance of a good interface is highlighted multiple times. However, little or no information regarding gesture recognition in Augmented Reality systems is dealt with in the paper. And all the extra information like, types of devices that could be used for AR systems, etc, though helpful for understanding the domain, does not directly align with the project.

2.3. Gesture Recognition in Augmented Reality

Different gesture recognition techniques are proposed by [9] for three basic interaction categories (translation, rotation, and scaling) in a Leap Motion Controller, Augmented Reality framework. In [9], the proposed model is implemented using the Leap Motion Controller (LMC) along with the Unity3D platform. The Leap Motion Controller is a camera sensor developed by Leap Motion that senses natural hand movements and the position of our fingers and allows us to interact with the system by using gestures like swiping, grabbing, pinching, and rotating. In the authors' prototype, they design two different gestures which are of the high and low level of naturalism for three interaction categories, which are translation, scaling, and rotation.

Reading [9] gave us an overview of the types of gestures used in hand gesture recognition systems like translation, scaling, and rotation and helped decide which gestures need to be supported in the project. The user's expectations from an AR application and what could be done to make the user's performance and experience better were also observed from [9].

In [10], the authors aim to develop a natural interaction technique that allows the manipulation of virtual objects on handheld augmented reality devices in 3D space. This method relies on the identification of the position and the movements of the user's fingers. The recognized gestures are then mapped to the corresponding manipulations of the virtual objects in the AR scene. The method implemented in [10] provides 6 degrees of freedom manipulations using natural finger-based gestures for rotating, translating, or scaling a virtual object in a handheld AR system. Markerless 3D gesture-based interaction design has two major components to it: Object Selection and Canonical Manipulations. Reference [10] provided an insight into how the user's hand could be segmented from the surroundings. It also highlighted the different ways in which virtual objects in the scene could be selected and manipulated using hand gestures.

Paper [11] gave an overview of the different basic gestures that users are most likely to use while manipulating 3D objects (As per user studies performed, it was observed that most users physically performed rudimentary object manipulation like zooming, scaling, stretching, and compressing the object which was supported effectively by this system developed). The Client-Server system setup was also demonstrated in [11]. The technicalities of depth and image recognition and tracking in 3D object manipulation and deploying AR applications on desktops by using webcams were observed. Additionally, the fact that gestures need to be as intuitive as possible to reduce mental strain for the end-users was also highlighted in [11].

Reference [12] explains some gesture recognition techniques from the perspective of developing an Augmented Reality media player application. It primarily focuses on an approach that is not computationally intensive, which is more suitable for an AR application. In the final approach, proposed by the authors, the focus is on recognizing the dynamic gestures of the user via a Webcam which could be a built-in or externally attached Camera. Various image processing algorithms like RGB to HSV conversion, blurring, thresholding, blob detection, etc have been integrated to analyse the gestures.

The approach used in [12] had no restrictions on the background of the image. It could handle dynamic gesture recognition via a live video feed, as the runtime processing time is negligible. Also, the colour model that is used is very strong. However, the Augmented Reality part was not dealt with in the paper and the complete integration was not defined. Additionally, [12] provided an overview of a gesture recognition system that is not computationally intensive and an AR Application-oriented approach to gesture recognition.

3. PROPOSED METHODOLOGY

To achieve the target of a generic and accessible AR application, the paper defines the scope of a demonstrable component as an android application that can reflect the activities running on a remote PC host. The android application, which can run on any simple smartphone (without the need for additional depth-sensing cameras) provides an Augmented Reality interface of the PC screen. The user can manipulate the AR interface (which helps avoid physical contact with any device) to operate on the PC remotely. The genericness is reflected in the fact that the applications running on the PC need not be AR compatible. Hence, absolutely any application that can run on the PC, would have an AR interface.

As seen in Figure 1, the approach taken to resolve this issue has two fundamental components. They are the interaction devices and the modes of interaction. There are three interaction components: the AR application, the remote host PC, and the gesture recognition server. The mode of interaction adopted in the chosen approach is a cloud-based interaction system.

3.1. Augmented Reality Application

The Augmented Reality (AR) Application acts as the controller of the entire system. In this part of the project, improvement in the accessibility of AR applications and their use is targeted. The application is deployable on a smartphone rather than a head-mounted device, due to its larger availability across the world. Also, it would only require a regular RGB camera that is built into a standard smartphone, without the necessity of a camera powered with pre-existing depth sensing technology. Since most AR applications use ARCore and ARKit which require smartphones with a depth-sensing camera, the project followed by this paper challenges this popular standard.

The design choice of choosing Unity [3] as the development engine over the Unreal Engine, was a conscious move, to make the AR application deployable to a broader spectrum of devices. However, this choice had to make a compromise on the larger feature set and the stronger and more robust engine that Unreal provides.

3.2. Remote Host

The remote host is the backend PC that is being manipulated by the AR application. The genericness of the system would be demonstrated in this part of the project, as any application visible on a PC screen could be replicated as a holographic AR Screen. The PC can hence be controlled via gestures (touchscreen gestures) in an AR application at any remote destination with a stable internet connection. For the demonstration, a browser application running on the PC is manipulated by the AR application.

3.3. Gesture Recognition Backend

The backend server runs the gesture recognition system. A design choice of using static gesture mapping is made so that the recognition can be quick. However, the challenge that this poses is the fact that not many gestures can be supported, however, the speed benefit outweighs the reduced number of gestures that are supported. Alternatively, an option writing the gesture recognizer in the AR application itself, limited the scope of gesture recognition, as it would have to run in the smartphone, which would be able to provide lower processing power [13]. Hence, an external server (which could also be the host PC itself), runs the gesture recognition system.

For the proof of concept of the methodology, the gesture recognition part involves 2 major parts - Hand segmentation and recognising the gesture in itself [14, 15]. For segmenting the hand from the surroundings, the major steps were background subtraction, thresholding and contour extraction. For recognizing the gestures, the steps involved were: Finding the convex hull and computing the extreme points, finding the centre of the palm and constructing a circle around the palm and fingers, performing bitwise AND between the hand region and the circle and determining the count or number of fingers. This was one of the various alternatives in which static gesture recognition could be implemented and can be easily modified according to the developer's requirements. Lots of research has gone into the field of gesture recognition, and some that are relevant in the current scenario include [16, 17, 18, 19].

3.4. Storage and Communication

The storage is handled using cloud storage systems. Other storage and communication alternatives included the use of a client-server system. But the client-server architecture restricted the AR application to run on the same network as the host PC. This was a clear disadvantage. Hence the choice of using a cloud storage system was made clear. However, this choice had the drawback of being a little slower than its counterpart.

The implementation used Google's Firebase Cloud System [20] for all the storage requirements. The Firebase Real-Time Database (RTDB) was used to store the recognized gesture [21]. The Firebase RTDB is a cloud-based NoSQL database that syncs data in real-time [22, 23], hence it was apt in storing the gesture identified by the gesture recognizer backend, to be used immediately by the remote host to update its state. The Firebase Cloud Storage is used to store the host PC's current state (for updating the AR environment) and the AR environment (for gesture recognition). This Cloud Storage unit of Firebase is a scalable cloud infrastructure [24] capable of storing large binary objects as blobs. The states that are saved in the form of screenshot images are hence stored and retrieved quickly, efficiently, and securely, in the Firebase Cloud Storage.

3.5. Overall System Design

The overall flow of the proposed methodology proceeds as illustrated by the sequence diagram in Figure 1. The process begins with the user enabling the host PC to be accessed by the Augmented Reality application. This is automatically followed by the storage of the initial state (screenshot of the entire PC screen) in the Firebase Cloud Storage. Once the user opens the AR application via the user's smartphone remotely, the state of the host PC is fetched from the Firebase Cloud Storage and rendered as an AR interface in the mobile application. As the user manoeuvres through the AR interface, the state (screenshot) of the AR environment is also stored in the Firebase Cloud Storage. This picture/state of the AR environment is constantly monitored by the gesture recognizer to identify the user's gestures. Once the gesture is identified, it is updated in the Firebase Real-Time database. The host PC reads the gesture from the Firebase Real-Time Database and updates the state of the remote PC. This new state is updated in the Firebase Cloud Storage, which is reflected in the AR application. And the process keeps repeating.

4. RESULTS AND CONCLUSION

The proposed methodology was applied on a browser (Google Chrome, Microsoft Edge as seen in Figure 2) that runs on a laptop with an internet connection. The browser could be accessed remotely via the Augmented Reality interface in an Android application that was connected to the internet, on a separate network. The syncing had minor lags depending on the speed of the internet connection on both the devices (smartphone and remote host PC). However, this

proposed methodology gives a simple, generic, and highly accessible solution for Augmented Reality interfacing for any application.

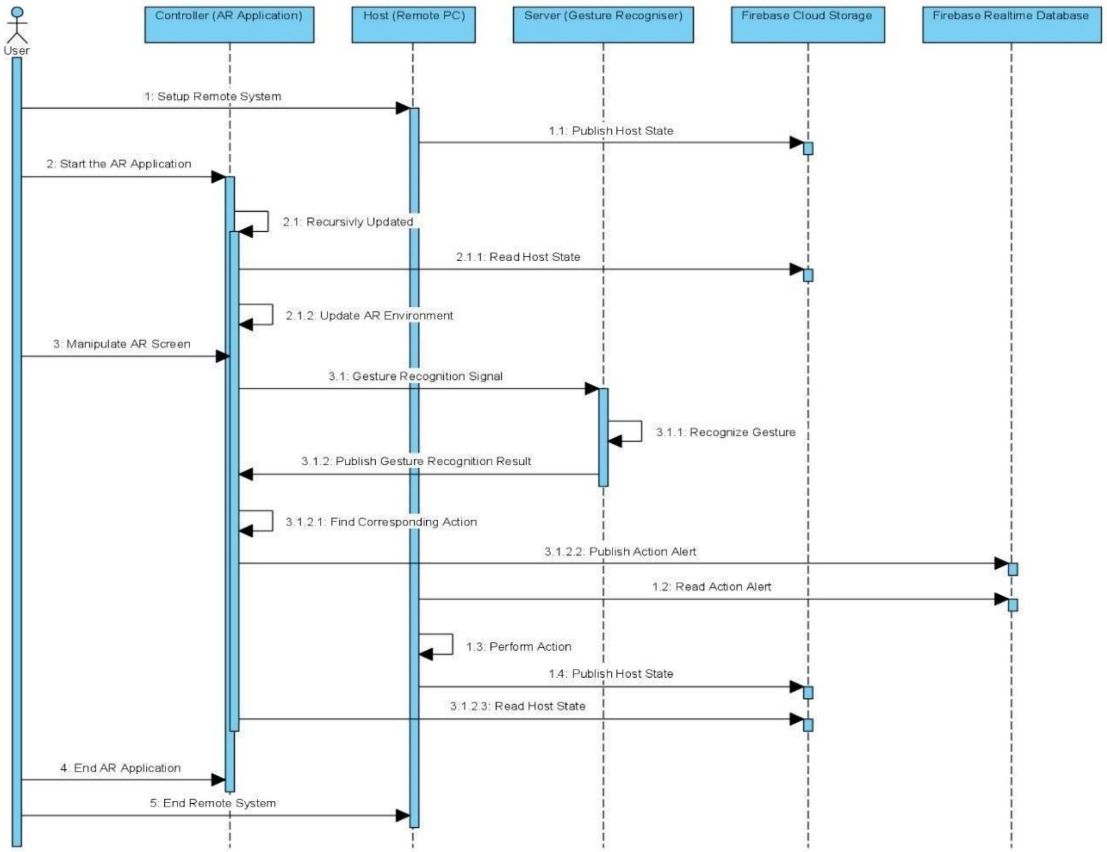

Figure 1. Sequence diagram of the proposed approach

5. FUTURE SCOPE

The methodology proposed in this paper can be further extended to thorough implementation. If funds permit, the Augmented Reality application can be extended to greater graphical quality (with holographic effects, and 3D rendering). Also, a wider range of gestures can be supported by a developer who wishes to extend its functionality. Moreover, there is new scope for innovation in the methodology itself, by taking the platform to an offline access system. This would not only improve the accessibility of the system but would also reduce the lags caused due to internet connectivity issues.

Figure. 2. A view of the AR Application (controlling Microsoft Edge remotely) on a smartphone in a basic demonstration. Only the quadrilateral in the middle is the AR rendering. The remaining is the user's hand and the real environment/surroundings behind.

REFERENCES

[1] Julie Carmigniani et al. "Augmented reality technologies, systems and applications". In: *Multimedia tools and applications* 51.1 (2011), pp. 341–377.

[2] Kevin Bonsor amp; Nathan Chandler. *How Augmented Reality Works*. Feb. 2001. URL: https : / / computer . howstuffworks.com/augmented-reality.html.

[3] Vinh T Nguyen and Tommy Dang. "Setting up virtual reality and augmented reality learning environment in unity". In: *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE. 2017, pp. 315–320.

[4] Daria Dubrova. Sept. 2018. URL: https : / / theappsolutions . com / blog / development / augmented reality-challenges/.

[5] S Siji Rani, KJ Dhrisya, and M Ahalyadas. "Hand gesture control of virtual object in augmented reality". In: *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. 2017, pp. 1500–1505.

[6] *Gesture Interaction for Consumer Devices*. URL: https: //www.crunchfish.com/how- does- the-world- look through-smart-glasses/augmented-reality/.

[7] Rafiqul Zaman Khan and Noor Adnan Ibraheem. “Hand gesture recognition: a literature review”. In: *International journal of artificial Intelligence & Applications* 3.4 (2012), p. 161.
[8] Rodrigo Silva, Jauvane C Oliveira, and Gilson A Giraldi. “Introduction to augmented reality”. In: *National laboratory for scientific computation* 11 (2003).
[9] John Aliprantis et al. “Natural Interaction in Augmented Reality Context.” In: *VIPERC@ IRCDL*. 2019, pp. 50–61.
[10] Huidong Bai et al. “Markerless 3D gesture-based interaction for handheld augmented reality interfaces”. In: *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE. 2013, pp. 1–6.
[11] Mark Billinghurst, Tham Piumsomboon, and Huidong Bai. “Hands in space: Gesture interaction with augmented-reality interfaces”. In: *IEEE computer graphics and applications* 34.1 (2014), pp. 77–80.
[12] Sandeep Vasave and Amol Plave. “Study of Gesture Recognition methods and augmented reality”. In: *arXiv preprint arXiv:1411.5137* (2014).
[13] Arnaud Lemoine et al. “Hand gesture recognition system and method”. In: *Patent US* 6128003 (2000).
[14] Gogul Ilango. *Hand Gesture Recognition using Python and OpenCV - Part 1*. Apr. 2017. URL: https://gogul. dev/software/hand-gesture-recognition-p1.
[15] Gogul Ilango. *Hand Gesture Recognition using Python and OpenCV - Part 2*. Apr. 2017. URL: https://gogul. dev/software/hand-gesture-recognition-p2.
[16] Gur Raunaq Singh. *Introduction to Using OpenCV With Unity*. URL: https://www.raywenderlich.com/5475introduction-to-using-opencv-with-unity.
[17] M Naveenkumar and A Vadivel. “OpenCV for Computer Vision Applications”. In: *Proceedings of National Conference on Big Data and Cloud Computing (NCBDC’15)*. 2015.
[18] Prasham Parikh. *Google Open Sources Real-Time Hand Gesture Recognition Algorithm For Developers*. Aug. 2019. URL: https://in.mashable.com/tech/6130/googleopen- sources- real- time- hand- gesture- recognitionalgorithm-for-developers.
[19] Siddharth S Rautaray. “Real time hand gesture recognition system for dynamic applications”. In: *International Journal of UbiComp (IJU)* 3.1 (2012).
[20] Laurence Moroney, Moroney, and Anglin. *Definitive Guide to Firebase*. Springer, 2017.
[21] Wu-Jeng Li et al. “JustIoT Internet of Things based on the Firebase real-time database”. In: *2018 IEEE International Conference on Smart Manufacturing, Industrial & Logistics Engineering (SMILE)*. IEEE. 2018, pp. 43–47.
[22] Laurence Moroney. “The firebase realtime database”. In: *The Definitive Guide to Firebase*. Springer, 2017, pp. 51–71.
[23] Baudoux Nicolas and Bauwin Lucie. “Real-Time database: Firebase INFO-H-415: Advanced database”. In: *Universite Libre de Bruxelles. Academic year´* 2018 (2017).
[24] Laurence Moroney. “Cloud storage for firebase”. In: *The Definitive Guide to Firebase*. Springer, 2017, pp. 73–92.

AUTHORS

Arya Rajiv Chaloli:
Arya completed her BTech in Computer Science and Engineering (CSE) at PES University, Bangalore. Currently works at Microsoft India, Hyderabad.

K Anjali Kamath:
Anjali completed her Bachelor of Technology in Computer Science and Engineering (CSE) from PES University, Bangalore. Working as a Software Engineer in Citrix Systems, Bangalore.

Divya T Puranam:
Divya completed her BTech in Computer Science and Engineering from PES University, Bangalore. Works as a software engineer at Cisco Systems, Bangalore.

CLASSIFYING CELESTE AS NP COMPLETE

Zeeshan Ahmed, Alapan Chaudhuri, Kunwar Grover,
Ashwin Rao, Kushagra Garg and Pulak Malhotra

International Institute of Information Technology, Hyderabad, India

ABSTRACT

We analyze the computational complexity of the video game "CELESTE" and prove that solving a generalized level in it is NP-Complete. Further, we also show how, upon introducing a small change in the game mechanics (adding a new game entity), we can make it PSPACE-complete.

KEYWORDS

Complexity analysis, NP completeness, algorithmic analysis, game analysis.

1. ABOUT CELESTE

CELESTE [1] is a single-player 2D platformer developed by Maddy Thorson and Noel Berry. The game is about you being Madeline who is climbing to the peak of the mountain "Celeste". The goal of the game is to overcome obstacles on the way and make it to the end of the level. CELESTE consists of 7 levels which consist of sub-parts. In each subpart, you have to reach an exit point without taking damage. Taking damage resets you back to the starting point.

(a) Madeline (b) Directions

Figure 1. CELESTE player

1.1. About the Player

Madeline is the main character whose actions are governed by our controls. She is restricted in 8 directions of motion and primarily has 3 special moves other than basic left and right movement. Those are:

Directions of movement: Madeline can move in 8 directions as described in the left diagram. These and the other moves have fixed button/button combinations assigned to them.

Jump: She has the ability to jump to a certain height, which can be then done again only when she hits the ground.

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 11-28, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121902

Dash: When Madeline has the "charge" she can dash in any direction, this gives her extra momentum and the ability to dash into "space blocks" (described later). The charge gets used up when she dashes and is restored when she hits the ground or passes through the space block. There are other objects which also recharge her, but those are not used in the proof.

There are other objects which also recharge her, but those are not used in the proof.

Grab: Since she is a mountain climber, she has the ability to climb walls and wedges. But she has stamina, which limits the amount of time she can grab onto a wall before sliding down. This stamina is restored when she makes contact with the ground.

2. LEVEL IMPLEMENTATION

2.1. Frames

Figure 2. Bottom left corner is the entry point and the top right corner is the exit point

Frames are areas of games which has a start and an endpoint. You have access to one frame at a time. Using the entries and the exits, you move from 1 frame to another. So Frames serve as the checkpoints in the game. Frames do not have a limit of size since your screen can scroll.

The game has been split into such frames, which are puzzles on their own, which require planning and reflexes to reach the endpoint. There are multiple ways to solve these puzzles due to the game's versatility and the mechanisms in it.

A graph of such frames connected together makes up a level of the game. To construct our proof, we will make frames which we will join together to make our level.

2.2. Objects

Revolving around the basic 3 operations the game has a lot of mechanisms that add to the fun and the difficulty of the game. These are added to the game as the player makes progress in the levels.

For our proof, we will mainly use 3 of the objects. They are:

- Unstable Platform
- Button door
- Space block

We explain how these objects work below. The purpose of these objects will be explained later.

2.2.1. Unstable Platform

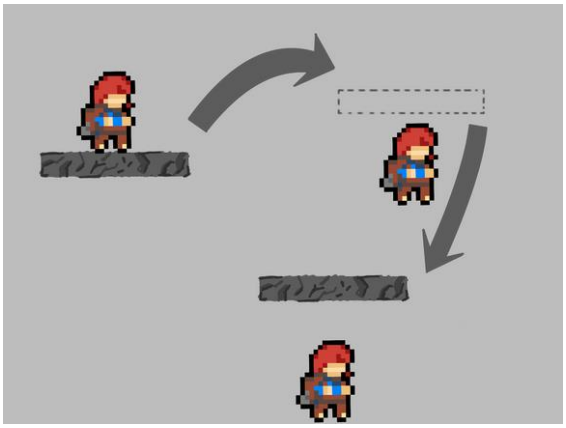

Figure 3. Unstable platform

A stone platform that can float, this platform breaks when Madeline stands on this platform for more than a second. After the platform breaks, Madeline if not jumped will fall down. This platform then forms back in the same place. But it can only be broken when stood upon and can not be broken from below, making it like a trap door if placed correctly.

2.2.2. Button Door

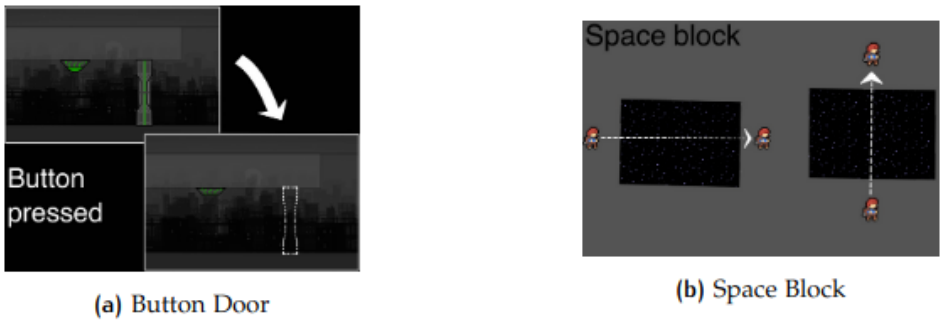

Figure 4. Button door and space block

A door that can only be unlocked using its specific button. This button must be placed in the same frame as the door, but it has the freedom to be placed anywhere in the frame. Once the door is opened it cannot be closed again.

2.2.3. Space Block

A block of celestial material which lets you float into and out in a straight line when you dash into it. Once Madeline dashes into the block, you cannot stop her from reaching the other opening of the block in a straight line. If the other side of the line is blocked with a wall, Madeline dies and respawns at the start of the frame.

2.3. Levels

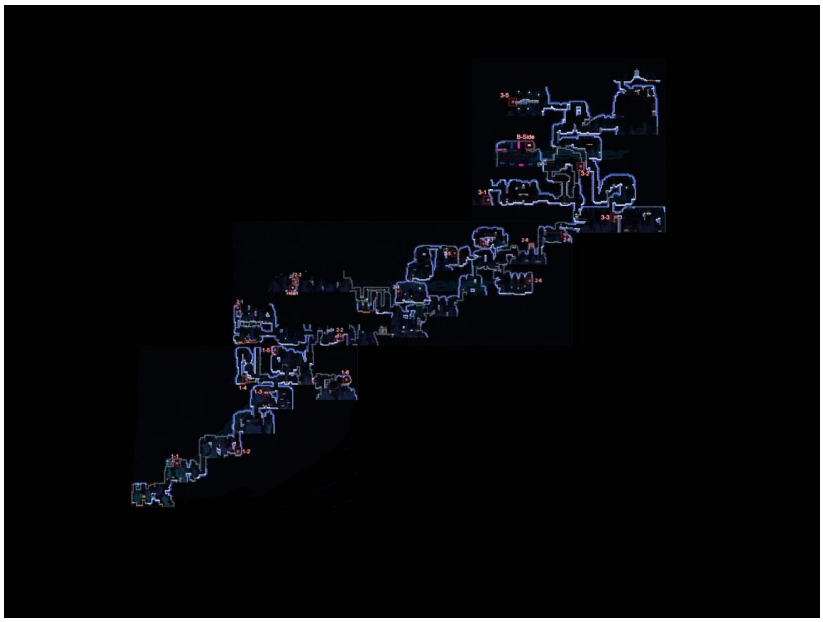

Figure 5. Bottom left being the start and the top right being the end

A level consists of many frames, but there is only 1 flag at the top of the level and there is only 1 initial start point of the level. For most of the levels, there is only 1 linear path from the start to end with a sequence of frames, but some other levels are more complex, which include sub-tasks and other detours. Here is an example, this is the 1st level in the game, the frames have been arranged according to the order.

3. COMPLEXITY CLASSIFICATION

Now that the game has been well defined, we work on classification of the game. Whenever we say "CELESTE belongs to *X* complexity class", we mean to say that the decision problem of deciding whether finish point is reachable from start point.

3.1. Basic Observation

Given a level and a path, that is the moves required to reach the endpoint, you can verify if the path is correct just by applying those moves. The moves are polynomial, why? We repeat the sections only after visiting other sections. Since the number of sections themselves are polynomial, we can only have polynomial moves before we complete the level.

This clearly implies that the game is NP [2]. For example, for the 1st level, we can map the path from start to end as seen below.

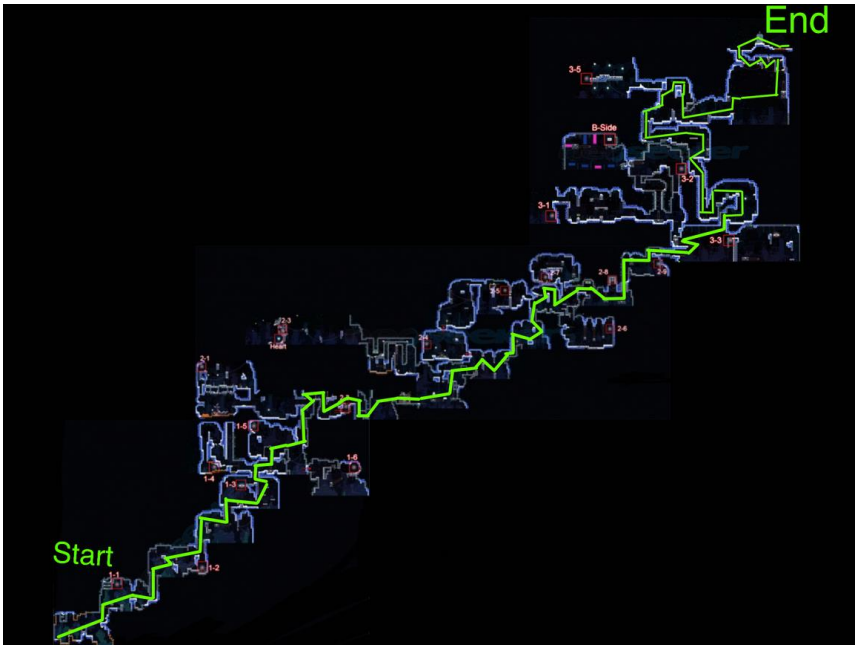

Figure 6. In this pattern we can always map from start to end

Now since we have proven that CELESTE is NP. We can try to prove that it is NP-Hard [2], essentially proving that it is NP-Complete [2].

3.1.1. Intuition behind being NP-Hard

Due to many paths of the game which do not lead to the end and certain mechanisms that lock us out which we will see in the future, it is not possible to tell whether there exists a polynomial-time algorithm to solve the levels. So, we will rather try proving that this game is NP Hard.

4. FRAMEWORK FOR NP-HARDNESS

To prove the NP-hardness of CELESTE, we here describe a framework for reducing 3-SAT [2] to a 2-D platform game. This framework is based on (link source here). Using this framework in hand, we can prove the hardness of games by just constructing the necessary gadgets [3-5].

The framework is a reduction of the 3-SAT problem. We start from the Start gadget, and end at the Finish Gadget. At each Variable gadget, we make a choice and turn on the Clause gadgets according to the choice, which in essence is making a literal true or false.

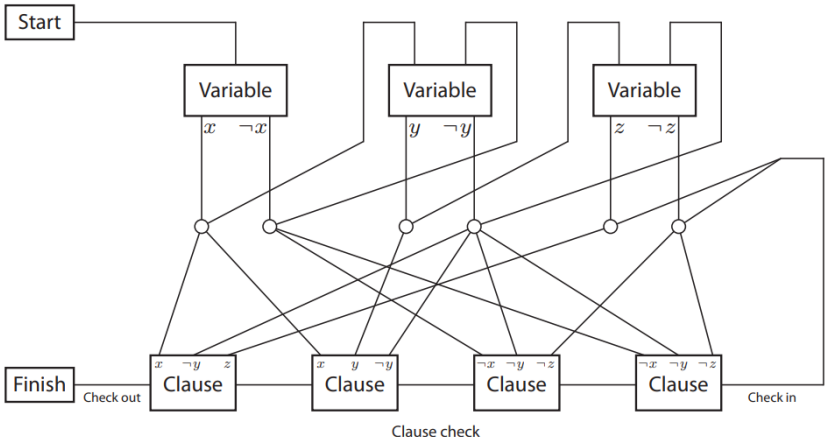

Figure 7. General framework for NP-Hardness

At the end, we make a pass through all of the clauses sequentially and we are able to pass through them if all of them are satisfied. Since the game is 2D, we also need a Crossover gadget. The Crossover gadget makes sure that if there are two overlapping connections, we travel the connections one by one. Below given are more details for the gadgets.

- **Start and Finish:** The start and end gadgets contain the spawn point and the end goal respectively.
- **Variable:** Each variable gadget must force the player to make a binary choice (select *x* or ~*x*). Once a choice is taken the other choice should not be accessible. Each variable gadget should be accessible from and only from the previous variable gadget is such a way that it is independent of the choice of the previous gadget and going back is not allowed.
- **Clause:** Each literal in the clause must be connected to the corresponding variable. Also, when the player visits the clause, there should be a way to unlock the corresponding clause.
- **Check:** After all the variables are passed through, all the clauses are run through sequentially. If the clause is unlocked, then the player moves on to the next clause else loses.
- **Crossover:** The crossover gadget allows passage via two pathways that cross each other. The passage must be such that there is no leakage among them.

If we can build these gadgets using a game, we can reduce 3-SAT to that game using this framework and show that the game is NP-Hard.

5. CELESTE IS NP-HARD

To prove that CELESTE is NP-hard, we will try to use the above framework to reduce 3-SAT to CELESTE.

5.1. Variable Gadget

A Boolean Variable can take 2 values, True or False, and it might have multiple occurrences throughout the formula.

The verification of 3-SAT is done by giving the satisfiable values to the variables, hence the values cannot be changed in the middle of the substitution. For now, we need to take care of the binary and the irreversible nature of boolean variables. We do this with the help of an Unstable Platform.

Figure 8. Exits are covered by Unstable platforms making them one way traps

These exits have an Unstable platform covering them, these have to be broken before the exit can be used. But, why the unstable platform?

Madeline falls from the top on the platform, that is the only entry to the Gadget. The Gadget has 2 exits on the sides of the floor, each leading to a tunnel. The unstable platform makes Madeline seal her choice. Once the path is taken, there is no way to access this frame again other than restarting since the platform will reform blocking the entry.

5.2. Clause Gadget

Each Clause has 3 variables, out of which even if 1 were true the Clause would be true. To implement that in our frame, we use The Button Door.

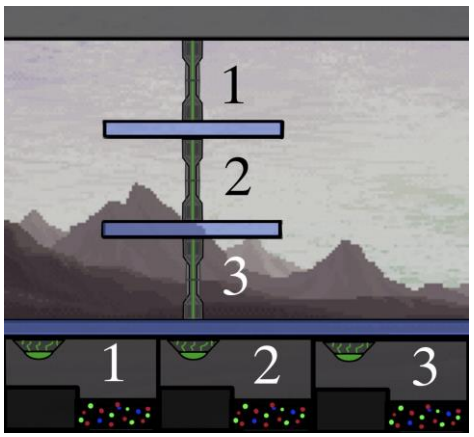

Figure 9. The colourful dots represent space block and the hits are reachable by Madeline

A Clause gadget consists of 3 Parallel Button Doors, the buttons are accessed through the variable tunnels. For now, do not worry about how the tunnels are connected. The main idea is that even if 1 door opens it is sufficient for Madeline to pass through the region.

Madeline presses the buttons according to the values she took for the variables, these will unlock the doors, if the variables made a clause true, the clause would have at least 1 door open.

6. SUPPORT GADGETS

Now these above-mentioned gadgets must be connected and for that, we use our support gadgets that will be constructed as per the requirements.

6.1. The Tunnel

To connect the Variable exit to the Buttons of the Clause, we use a Tunnel gadget.

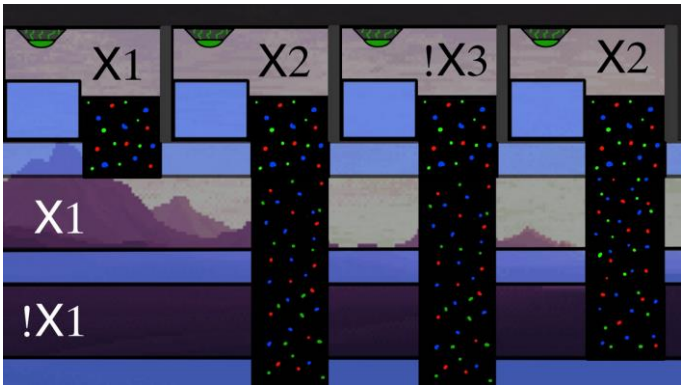

Figure 10. x_1 Tunnel only has access to button which have x_1 as their variable in their clause

Below the buttons, there are Tunnels leading from the exits of the variables. The variables have access to the buttons they can set true according to the Boolean expression.

For example, if x_1 is chosen to be true, then Madeline gets access to the x_1 tunnel and $\sim x_1$ if she had chosen false. x_1 tunnel has access to buttons that open a door to the clause having x_1.

How do we block the variables from accessing the other doors? For that, we have constructed the Crossover Gadget. The space blocks that are displayed in the diagram are used in a specific manner described in the Crossing Frame.

6.2. Crossover Gadget

Since the game is 2D, you cannot avoid paths from crossing each other during the construction of such a level. We can make sure that the intersection of the paths happens only in the form of a cross.

Suppose we want Madeline to go from A_1 to A_2 or B_1 to B_2 or the other direction. But she shouldn't be able to go from an A to a B or vice versa.

The Space block as described before teleports the player from one end to the other in a straight line without any interference. Encountering a wall will kill Madeline and she will respawn at the start of the frame.

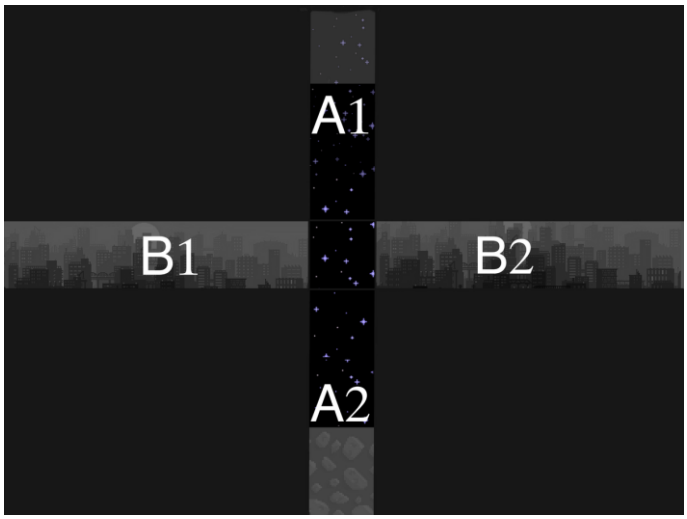

Figure 11. Crossover gadget

So, we put the space blocks in the intersection of the paths in such a way that there are no straight lines connecting the opening side of *A* to *B*. It might seem like you can draw a line from *A* to *B* but remember that Madeline can only move in 8 directions, so the lines can be parallel or 45 degrees inclined with the axis. So, no such line will exist. This means that the only way she can travel through the space block is in a straight line parallel to the axis, hence she cannot access *A* from *B* or vice versa.

7. SEQUENCE OF FRAMES

Now that all the frames have been constructed, we decide the sequence of the frames. Let *n* be the number of variables in the Boolean expression distributed into *k* clauses.

7.1. Variable Order

We will assume the order of the variables to be x_1, x_2, x_3 ... x_n. We select values for these variables in the same order. So the starting position will be in the x_1 gadget since we pick its value first.

7.2. Transition between Variables

From the variable gadget of x_1, we choose its value and go to the respective tunnel. In the tunnel, we press all the accessible buttons, after all the buttons, we reach the end of the tunnel. Now since the values of x_1 have been already picked and substituted, we have to pick a value for the next variable hence we must go to the x_2 variable gadget.

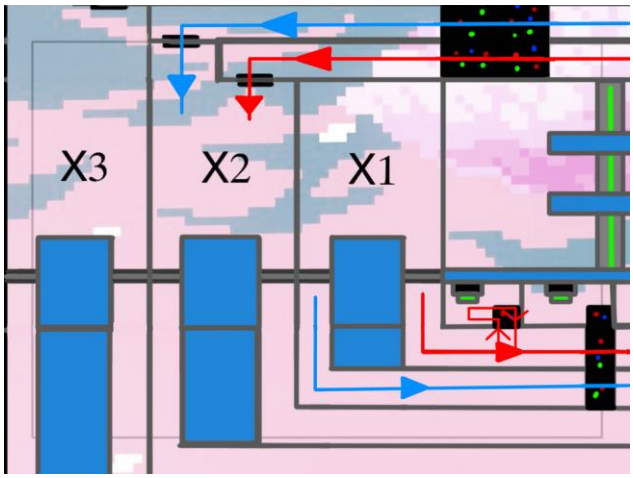

Figure 12. Transition from x_1 to x_2 after completing the path

In such order we pick the value of all the variables, click the buttons which open the clause doors, and continue until we reach the end of the x_n variable. Since we ran out of variables, where do we go now?

7.3. Final Passage

At the end of the x_n passage, we should have an entrance to the Final passage containing the clauses. Madeline after choosing all the variables will now have to reach the flag.

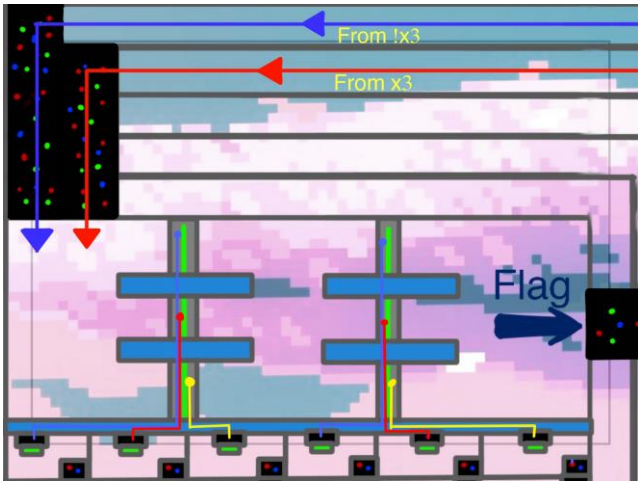

Figure 13. Example case where n = 3

The path to the flag lies on the other side of the passage. If at least one door from each Clause is open only then will she be able to reach to the flag, else she won't be able to complete the level. This was the AND of all the OR clauses, leading to a normal form with 3 variables in each clause.

7.3.1. Why not put the flag at the end?

The flag is always at the top of the level. So we need to redirect the player from the end of the final passage to the flag. Since there are other Paths that come between it, we use crossover frame.

8. FINAL LEVEL

Now that all the separate parts have been explained, we put together our final level. The Boolean expression for which the level has been implemented is:

$$(x_1 \vee x_2 \vee \neg x_3) \wedge (\neg x_1 \vee x_2 \vee x_3)$$

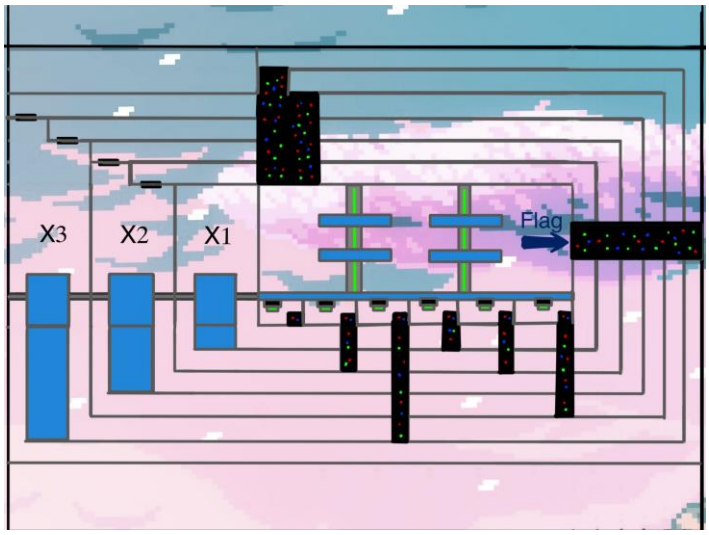

Figure 14. Final level layout

9. EXAMPLE SUBSTITUTION

Let the substitution of the variables be:

- $x_1 = 1$
- $x_2 = 0$
- $x_3 = 1$

We start with the x_1 gadget, take the true tunnel, and activate the door. After which we end up with a tunnel with an exit leading to x_2 gadget.

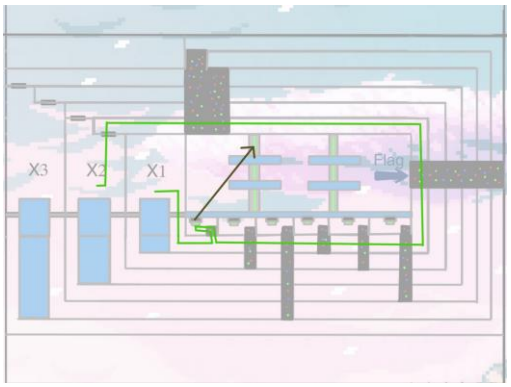

Figure 15. Going from x1 to x2 gadget

Now we are in the x_2, we take the false tunnel, but since there is no x_2 in the expression, there is no door that can be opened from this tunnel. So, we just continue and end up in the x_3 frame. In the x_3 frame, we repeat the same procedure. We open a door in the 2nd OR clause and end up at the end of the tunnel which leads to a space block. This space block when used will take Madeline to the beginning of the 'final passage.'

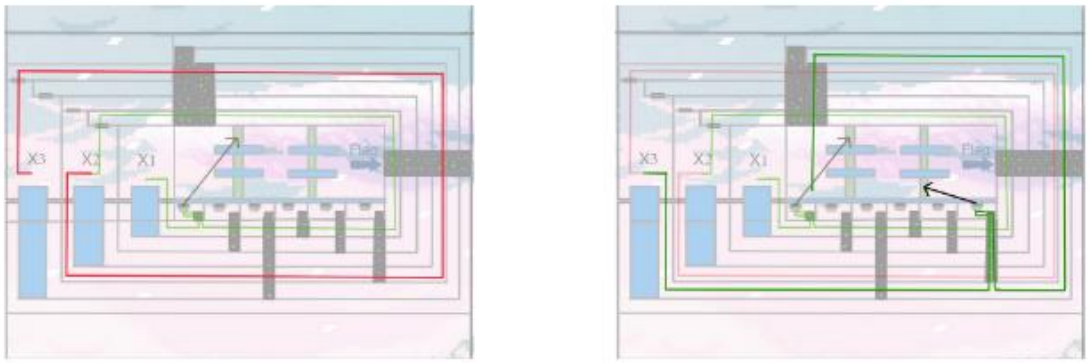

Figure 16. Going to the Final passage

Now one door of each clause has opened, Madeline can pass the final passage and go to the flag.

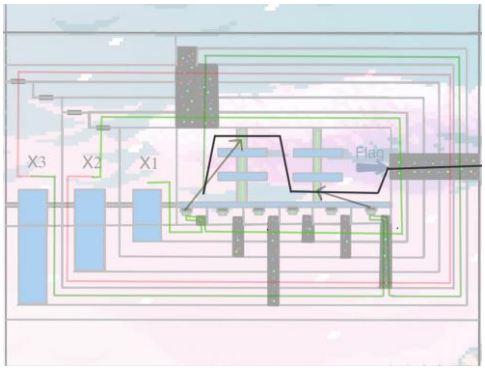

Figure 17. Reaching the flag

Since Madeline was able to reach the flag. The level could be completed, hence the expression as expected is satisfied with the given values.

10. SUMMARY ON NP-COMPLETENESS

This proves that CELESTE is at least as hard as 3-SAT, making it NP-Hard. In conclusion, the game is both NP and NP-Hard, making it an NP-Complete puzzle.

11. MAKING CELESTE PSPACE-COMPLETE

We now describe how making a small change to CELESTE can make it PSPACE-Complete [2]. In the proof the CELESTE is NP-Complete, we used a button door which opened when we pressed the green button and then was obsolete.

We make an addition to the game. The door can now be closed using a red button. When the door is open, the red button is deflated and can be activated and when the door is closed the green button can be activated.

12. CELESTE BELONGS TO PSPACE

To prove that CELESTE is a PSPACE-Complete puzzle, we have to first show that it belongs to PSPACE [2]. Now, it is sufficient to show that CELESTE belongs to NPSPACE [2] since NPSPACE is a subset of PSPACE by Savitch's Theorem [6]. This means that for any given traversal on the level, it has to use polynomial space with respect to the size of the level. Since, the game's element behaviour is a simple (deterministic) function of the player's moves. Therefore, we can solve a level by making moves non-deterministically while maintaining the current game state (which is polynomial).

13. FRAMEWORK FOR PSPACE-HARDNESS

To prove PSPACE-Hardness of CELESTE, we here describe a framework for reducing TQBF problem [2] to a 2-D platform game. This framework is based on (link source here). Using this framework in hand, we can prove hardness of games by just constructing the necessary gadgets. For this framework we need one more gadget: **Pressure Button Door Gadget.**

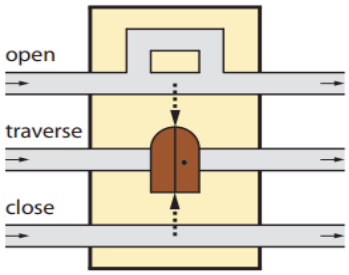

Figure 18. Pressure Button Door Gadget

In CELESTE, to press the button, Madeline has to dash into it. The button in the above frame is thick enough to prevent Madeline to pass through the tunnel without pressing the button. Thus, it forces her to press the button.

- The open path has a button which the player is forced to press.
- The traverse path is the path containing the door which can be traversed if the door is opened.

- The close path button forces the player to close the door.

13.1. A General Framework for PSPACE-Hardness

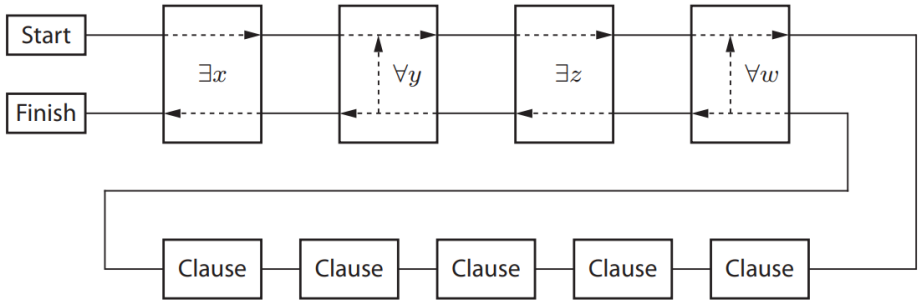

Figure 19. Framework

A given fully quantified Boolean formula

$$\exists x \forall y \exists z \ldots \phi(x, y, z, \ldots),$$

where PHI is in 3-CNF is translated into a row of Quantifier gadgets, followed by a row of Clause gadgets, connected by several paths.

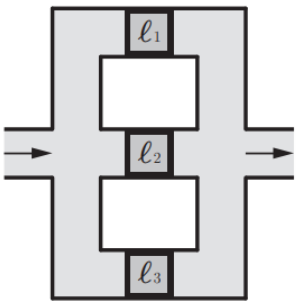

Figure 20. Clause gadget

The clause gadget is built using three gates whose pressure buttons are in quantifier gadgets. And, for building Quantifier gadgets we use a special notation in a tunnel as shown below:

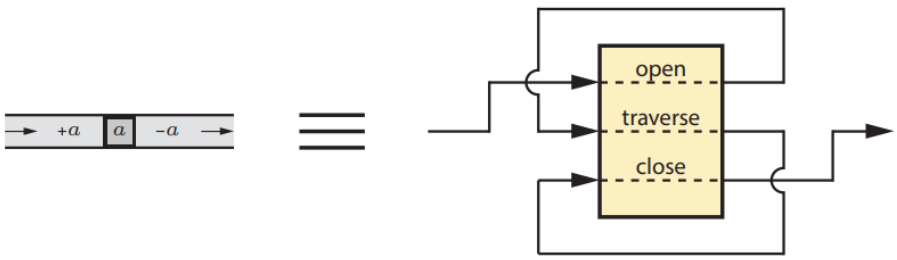

Figure 21. Shorthand notation for tunnels

Here, *+a* opens the gate corresponding to variable and *a* and *-a* closes the gate corresponding to gate *a*. The player is forced to press these buttons as described before.

13.2. Existential Quantifier

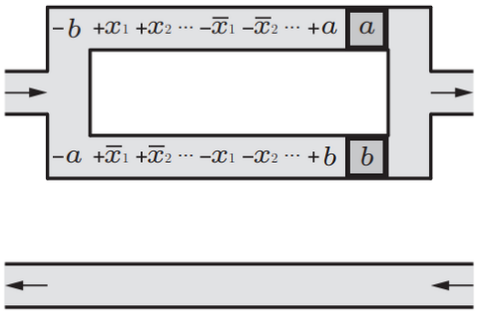

Figure 22. Shorthand notation for tunnels

The player can only select one of the path ways and once it is selected, the player can never change his choice.

13.3. Universal Quantifier

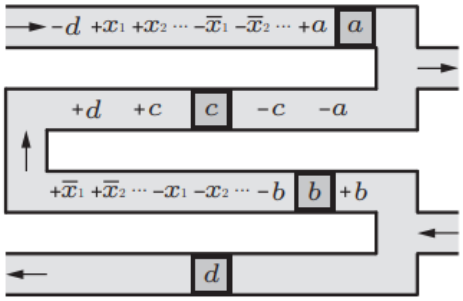

Figure 23. Shorthand notation for tunnels

The player first proceeds to mark the variable as true and while backtracking has to traverse through the clauses again with the variable marked as false. After both possibilities are tried the player is able to move forward.

13.4. Conditions for Traversal

Traversing a quantifier gadget sets the corresponding variable in the clauses. When passing through an existential quantifier gadget, the player can set it according to their choice. For the universal quantifier gadget, the variable is first set to true.

A clause can only be traversed if at least one of the variables is set in it. After traversing all the quantifier gadgets, the player does a clause check and is only able to pass if all the clauses are satisfied. If the player succeeds, they are routed to lower parts of the quantifier gadgets, where they are rerouted to the last universal quantifier in the sequence.

The corresponding variable is then set to False and the clauses are traversed again. This process continues and the player keeps backtracking and trying out all possibilities. We will later show how to build these quantifier gadgets using CELESTE game entities.

14. MODIFIED CELESTE IS PSPACE-HARD

Using the previously described framework, we will build corresponding gadgets and thus show a reduction from the TQBF problem to CELESTE, implying the CELESTE is PSPACE-Hard.

14.1. Door Gadget

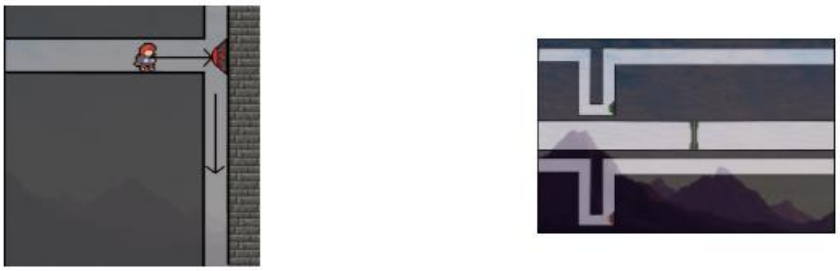

Figure 24. Door gadget and force button

For creating a door gadget we first need to create a way to force the player to dash into a button to activate it. We do this using a narrow tunnel and leaving only enough space to pass if the button is pressed. Now, using this force button, we create a door gadget. The button above will open the door, and the button below will close the door, and since the path is thin, Madeline will not be able to pass through until the button is pressed.

14.2. Multi-Tunnel Gadget

For clubbing multiple open/close symbols together, we use a multi tunnel gadget. This is just a helpful gadget to make Quantifier Gadgets.

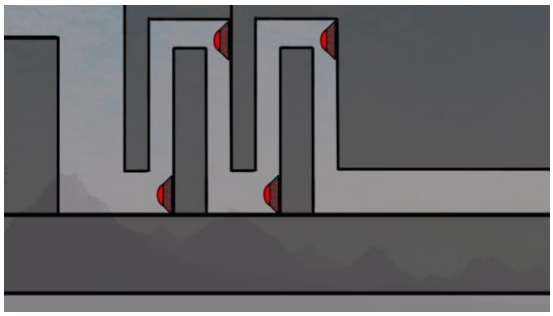

Figure 25. Multi-tunnel gadget

14.3. Putting it all together

Since we were able to make the gadgets described in the framework, we have a reduction from the TQBF problem to Modified CELESTE. Thus, Modified CELESTE is PSPACE-Hard.

15. SUMMARY ON PSPACE-COMPLETENESS

In conclusion, the modified game is both PSPACE and PSPACE-Hard, making it a PSPACE-Complete puzzle.

On adding the close button, we added a requirement to keep track of all the doors that are open. Before once a door was open, it always remained open. Before we had to only keep track of the current state sequentially as all the doors will be opened in a sequence. Knowing that a door is openly implied that all the previous doors were opened to reach the current door.

But after adding the close button, at any point of the game, all the doors are independent and knowledge of the open state of a door gives us no info about the other doors. So, we must keep track of all the other doors. This makes the game harder and makes its PSPACE-Hard instead of NP-Hard.

In short, lack of knowledge about the status of all the doors makes the game PSPACE-Complete.

16. CONCLUSIONS AND FUTURE WORK

In this paper, we have proven that the task of solving levels for the original version of the video game CELESTE is NP-Complete. As described above, this means that this problem is equivalent to solving the Boolean satisfiability problem. Furthermore, addition of complications to the original version in the form of a door-closing button results in the modified game becoming harder [7], PSPACE-Complete to be specific.

This work serves as a stepping stone towards analysing computational complexity of more sophisticated interactive systems and increases the scope of video games that have been studied within this field. Further work involves investigating the hardness of platformers with more involved mechanics, similar to what has been done on analysing the complexity of the popular arcade game Angry Birds [8]. We are hopeful that our work opens up new directions to understand the relationship between the mechanics of interactive systems and computational complexity.

REFERENCES

[1] Wikipedia contributors. Celeste (video game) — Wikipedia, the free encyclopaedia, accessed 2022.
[2] S. Arora and B. Barak, Computational Complexity: A Modern Approach, Cambridge University Press, 2006.
[3] Giovanni Viglietta, Gaming is a hard job, but someone has to do it!, Theory of Computing Systems, *v. 54, p. 595–621*, 2014.
[4] Greg Aloupis *et. al.*, Classic nintendo games are (computationally) hard, Theoretical Computer Science, *v. 586, p. 135-160*, 2015.
[5] Erik D. Demaine *et. al.*, Toward a General Complexity Theory of Motion Planning: Characterizing Which Gadgets Make Games Hard, Innovations in Theoretical Computer Science Conference, 2020.
[6] M. Sipser, Section 8.1: Savitch's theorem, *Introduction to the Theory of Computation*, PWS Publishing, *p. 279-281, ISBN 0-534-94728-X*, 1997.
[7] Joshua Ani *et. al.*, Walking through Doors is Hard, even without Staircases: Proving PSPACE-hardness via Planar Assemblies of Door Gadgets, International Conference on Fun with Algorithms, *v. 157, p. 3:1 – 3:23*, 2021.
[8] Matthew Stephenson *et. al.*, The computational complexity of Angry Birds, Artificial Intelligence *v. 280, p. 103232, issn. 0004-3702*, 2020.

AUTHORS

Zeeshan Ahmed

Senior, Bachelor of Technology and Master of Science Dual Degree in Computer Science at International Institute of Information Technology, Hyderabad.

Alapan Chaudhuri

Senior, Bachelor of Technology and Master of Science Dual Degree in Computer Science at International Institute of Information Technology, Hyderabad.

Kunwar Grover

Senior, Bachelor of Technology in Computer Science and Engineering at International Institute of Information Technology, Hyderabad.

Ashwin Rao

Senior, Bachelor of Technology in Computer Science and Engineering at International Institute of Information Technology, Hyderabad.

Kushagra Garg

Senior, Bachelor of Technology and Master of Science Dual Degree in Computer Science at International Institute of Information Technology, Hyderabad.

Pulak Malhotra

Senior, Bachelor of Technology in Computer Science and Engineering at International Institute of Information Technology, Hyderabad.

PERFORMANCE ANALYSIS OF SUPERVISED LEARNING ALGORITHMS ON DIFFERENT APPLICATIONS

Vijayalakshmi Sarraju[1], Jaya Pal[1] and Supreeti Kamilya[2]

[1]Department of Computer Science and Engineering,
BIT Extension Centre Lalpur, India
[2]Department of Computer Science and Engineering, BIT Mesra, India

ABSTRACT

In the current era of computation, machine learning is the most commonly used technique to find out a pattern of highly complex datasets. The present paper shows some existing applications, such as stock data mining, undergraduate admission, and breast lesion detection, where different supervised machine learning algorithms are used to classify various patterns. A performance analysis, in terms of accuracy, precision, sensitivity, and specificity is given for all three applications. It is observed that a support vector machine (SVM) is the commonly used supervised learning method that shows good performance in terms of performance metrics. A comparative analysis of SVM classifiers on the above-mentioned applications is shown in the paper.

KEYWORDS

The supervised learning algorithm, stock data mining, undergraduate admission scheme, breast lesion detection, and performance analysis.

1. INTRODUCTION

Machine learning is popular today; finding patterns and relationships within highly complex datasets is used in today's era. Several machine learning techniques are making real-world applications possible with recent storage and computational capabilities developments. Machine learning requirements are increasing day by day due to the ever-growing data sets. In 1959, machine learning was first coined by Arthur Samuel, who is an eminent pioneer in the field of artificial intelligence and gaming technology [1]. Machine learning is the branch of artificial intelligence that uses data and algorithms to imitate intelligent human behaviour. In general, machine learning is categorised into three types of learning methods: supervised learning, unsupervised learning and reinforcement learning. Supervised learning needs a trained data set with labelled data [2] with a known output value. Unsupervised learning does not use the training data set. The most common example of an unsupervised learning algorithm is clustering [3]. In Reinforcement learning, the input data from the environment is used as a stimulus to determine how the model should react [2]. In brief, supervised learning requires the training of labelled data with inputs and desired outputs. The labelled input data set of supervised learning is used to train the algorithm. The algorithm improves its estimates by making exact predictions in this training process and re-iterates the algorithm until it achieves the desired level of accuracy. Classification and regression problems are solved through supervised learning [2]. In a regression problem, we try to predict the results within a continuous output, which means that we try to plot input

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 29-35, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121903

variables to some continuous function. In a classification problem, we try to predict the results in discrete output. Supervised learning is used successfully in different fields such as data mining [4], pattern recognition [3], the internet of things [3], health monitoring [3] and market analysis [5]. Three significant and different types of works on supervised learning techniques are based on stock market analysis [5], admission schemes in educational institutes [6] and breast lesion detection [7], where the supervised learning algorithms are used to predict future outcomes based on historical data. Analysing different supervised learning algorithms on applications from different fields is the main motivation of the current work.

This paper is focused on the performance analysis measures of various supervised learning algorithms, such as linear discriminant (LD), logistic regression (LR), naïve Bays (NB), support vector machine (SVM), K-nearest neighbour (KNN), and random forest (RF) methods in the applications of stock market analysis, undergraduate admission data and breast lesion detection. The performance analysis of the said algorithms is first done on the established works of [5]–[7]. Among the various supervised algorithms, SVM is shown to be a good classifier that can be commonly applied to all the aforementioned applications. Therefore, a comparative study of SVM statistics on the said applications is presented in the paper. Section II gives a brief description of machine learning and different supervised learning algorithms. Three applications of supervised learning techniques along with the performance measurements are provided in Section III. Section IV provides a discussion about the performance analysis of the algorithms and a comparative analysis of SVM on three different applications. Finally, the paper is concluded in Section V.

2. SUPERVISED MACHINE LEARNING ALGORITHMS

Machine learning (ML) adds a new dimension to technology where a computer performs any task without human intervention on the basis of constantly learning from past experiences. Therefore, ML has three features: (1) Task, (2) Performance measure, and (3) Experience [8]. Supervised learning (SL) is a type of machine learning in which a function which converts input to output is learned using examples of input-output pairs. Based on labelled training data that comprises training examples, it infers a process. Supervised learning uses labelled datasets to train algorithms to recognise data or properly predict outcomes. The two different tasks in machine learning are classification and regression. In classification, the labelled data is discrete, but in the regression, it is continuous. In the case of unsupervised learning, the training data is not labelled. A classifier is designed by deducing existing clusters in the training data set. The supervised learning algorithms, used in the paper for performance analysis, are discussed as follows. Naive Bayes (NB) is a statistical classification approach used in supervised algorithms. The Bayesian classification's key benefit is that it can handle predicted difficulties. The Bayes theorem is used to underpin this categorization method. The moniker "Naive" stems from the algorithm's strong assumption that all input features are independent of one another and have no correlation in simple probabilistic classifiers like Naive Bayes. Because it is a probabilistic model, Naive Bayes provides a posterior probability of belonging to a class given input features.

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$$

Where A and B are two independent events, P(A) and P(B) are the probabilities of A and B. The conditional probability, P(A/B), is the probability of an occurrence A given an event B. P(B/A) denotes the chance of seeing an event B if A is true.

Logistic regression (LR) is used to determine how much possibility is there for cases where the event is successful, and the identical event is unsuccessful. When the dependent variable is in binary form (having just one of two values), logistic regression is used. That means, it can only have two possible values [9]. Because maximum likelihood calculations are less accurate in small sample sizes than simple least squares and only large sample sizes are needed in logistic regression. The relationship between the dependent and independent variables does not have to be linear.

Because of the ease of interpretation and short calculation time, the K-nearest neighbour (KNN) method is well-known for its simplicity. It saves available cases and categorises new instances based on homogeneity, similar to the distance function [10]. A majority vote of the object's neighbours determines its classification and this process is known as class integration. Following that, the object is assigned to the class with the highest similarity among the K nearest neighbours [10].

A decision tree (DT) acts like a flowchart that categorises instances depending on their characteristics. Each internal node represents a test case; branches show the results of the tests, and leaf nodes show class labels. This strategy will perform better when there are discrete characteristics [11]. In the simplest situation, each test considers a single attribute, and the instance space is partitioned based on the attribute's value. The condition refers to a range in the case of numeric properties.

A support vector machine (SVM) is an ML algorithm that solves problems of classification and regression. The SVM model is supported by the margin calculation idea. This method can be applied to both linear and nonlinear data. Each and every data observation is plotted as a point in n-dimensional space by this algorithm where n represents the number of features. Each feature's value is the matching coordinate's value. It divides the training datasets into classes by finding a line (hyperplane) that divides them. It maximises the margin between the nearest data point (in the classes) and the hyperplane.

Random forest (RF) is a special kind of ensemble learning algorithm used in classification and regression problems. A random forest is a forest of trees, each based on a different bootstrap sample from the training data. When a tree is fitted, some predictor variables are censored at each node. The optimal split is then determined using random forests based on the predictor variables chosen. After the trees have been voted on, they are grown to their full depth and an agreement prediction is made. It creates complex models with high predicted accuracy and highlights the significance of each variable in the categorization model.

A perceptron is a single-layer neural network with weights and biases that may be trained to produce the correct target vector when an acceptable input vector is provided. The training method employed is the perceptron learning rule. Perceptron is mostly used to solve simple pattern classification problems.

Multi-layer perceptron (MLP) networks contain, in general, three layers. Each layer, with the exception of the input layer, operates like a neuron and uses a non-linear activation function. Back-propagation, a supervised learning approach, is used to train the MLP network. The MLP is distinguished from a linear-perceptron by its several layers and non-linear activation function, which allows it to recognise non-linear data. Wide applications of MLPs include speech recognition and speech enhancement.

3. APPLICATIONS

This section discusses three different applications of various supervised learning algorithms. For the applications, we consider the works of [5]–[7] for stock data mining, undergraduate admission scheme and breast lesion detection, respectively.

In [5], the authors have proposed the work on the performance analysis of twelve years of renowned bank stock data. The performance of 4 different SL approaches, that is, SVM, RF, NB, and ANN (artificial neural network), are compared for the particular study. The authors use the values of correctly and incorrectly classified instances, which are used to calculate the performance of predictive classification models. Various statistical metrics, such as accuracy, precision, sensitivity, and specificity of different supervised classifiers are analysed by a renowned Bank's stochastic data set.

In [6], the authors study undergraduate admissions applications. The work focuses on how machine learning techniques can assist admission counsellors in concentrating their efforts on applicants who are more seemingly to join. The admission process is carried out in three stages: stage 0, 1 and 2. In stage 0, the applicants fill up their application forms. In stage 1, admissions counsellors evaluate and select some of the applicants. The applicants accept the requests of counsellors in stage 2. SL techniques are employed to classify stages 1 and 2. Predictive modelling consists of three phases: 1. data processing, 2. classification, 3. feature selection. In data processing, raw data provided by the admission cell is converted into a numerical form and accepted by a classifier. The classifiers are trained to make valuable predictions about future applicants in the classification phase. In the third phase, classifiers are used to determine the features in an application that are dominant in stage 1 and 2 predictions. Classifiers are used to predict stages 1 and Stage 2 relevant to the admission scheme, as discussed in [6]. The first stage indicates whether or not an applicant is accepted, whereas the second indicates whether or not an approved applicant attends the university. In the training process, five classifiers, such as MLP, Perceptron, linear SVM (LSVM), polynomial SVM (SVM POLY) and quadratic SVM (SVM RBF) are trained, and different statistics have been measured to find out the classifier with the highest performance rate. The performance rate is considered in the validation set and the same classifier is used in testing the current data.

In [7], the authors have presented an automatic computer-aided detection and diagnosis system of breast lesions. The developed model is dependent on supervised machine learning algorithms. These techniques are used to classify benign and malignant localised breast lesions. Four different machine learning algorithms are examined for classification: (1) support vector, (2) k nearest neighbours, (3) random forest, and (4) naive Bayes classifiers. The evaluation metrics used to demonstrate the study's success are accuracy, sensitivity, specificity, and precision.

4. DISCUSSION

In the presented work, we discuss the performance analysis of supervised learning techniques in different applications.

From Table I, performance analysis of supervised learning algorithms is observed on various statistics such as accuracy, precision, sensitivity, and specificity. In our comparative analysis, we observe the accuracies of different algorithms. In the case of a stock marketing application, the accuracy for LD is 53.6%, LR is 99.8%, NB is 48.3%, RF is 45.9% and LSVM is 99.1% respectively. Here, we found that the LR technique is the outstanding one out of all the above-mentioned techniques. In the case of the undergraduate admissions application, MLP, perceptron,

LSVM, polynomial support vector machine (SVM POLY), and Quadratic support vector machine (SVM RBF) algorithms are implemented. Overall, we observe that MLP is showing better accuracy than the other algorithms. The third application mentioned in Table I is breast lesion detection. Comparing the accuracies of SVM, ANN, NB and RF algorithms, the RF accuracy is the most accurate that decides tumour aggressiveness. The best precision value classifier in the stock data mining application is LR with 99.8%. Almost all the classifiers show

Table 1. Performance Analysis of Classifier in Three Different Applications

Sl. no.	Reference Applications	Supervised Learning Algorithms	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)
1	Stock Data Mining	Linear Discriminant	53.61	74.58	40.13	14.04
		Logistic Regression	99.82	99.86	53.24	0.18
		Naive Bays	48.33	31.68	16.99	37.84
		Random Forest	45.97	97.93	52.51	1.77
		LSVM	99.12	98.76	52.95	1.03
2	Undergraduate Admission	MLP	94.57	95.62	94.36	94.82
		Perceptron	94.32	95.69	93.71	94.04
		SVM POLY	94.36	95.67	93.87	94.94
		SVM RBF	94.44	95.66	94.01	94.95
		LSVM	94.45	95.70	94.05	94.93
3	Breast Lesion Detection	SVM	82.15	81.0	100	56.66
		KNN	79.36	80.0	96.8	20
		RF	90.36	92.0	96.25	83.33
		NB	87.82	86.0	100	66.67

the same precision value in undergraduate data. However, in the case of breast lesion detection application, the highest precision value classifier is RF with 92%. In comparing maximum and minimum values of sensitivity and specificity in a stock data mining application classifiers, LR is the highly sensitive classifier and Naïve Bayes is less sensitive. Whereas, in the case of specificity, the lowest one is LR and the highest one is NB with 18% and 37.8%, respectively. Let us now compare the values of sensitivity and specificity in the undergraduate application classifiers.

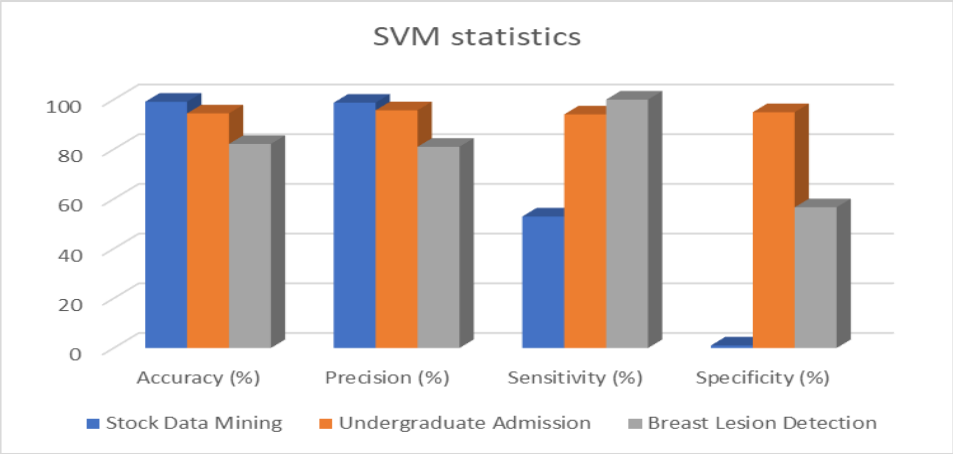

Fig. 1. Comparative analysis of the performance measures of SVM algorithm for the three applications

The highest sensitive rate classifier is MLP with 94.36%, however, specificity is almost identical in all classifiers. In the breast lesion detection application, the SVM classifier has a sensitivity of 100%, and the highest value of the specificity classifier is RF with 83.3%. From the table, we can make one comparative analysis of the three mentioned applications using the SVM classifier (as shown in Fig. 1). The SVM classifier shows better accuracy in the stock data mining application, compared to the other two applications. In terms of other statistics, such as precision, sensitivity and specificity of SVM analysis from the above graph, we found that there is not much variation in the precision value of stock data mining application and undergraduate data application but in comparison to these two applications, moderate interpretation of precision is found in breast lesion detection. Hence, we can say that the precision value is better in the stock data mining application on the overall analysis. Sensitivity is higher in breast lesion detection, lesser in stock data mining and moderately high in undergraduate admission data. Specificity is less in stock data mining applications than in the other two, but quite good in breast lesion detection.

5. CONCLUSION

Performance analysis of different supervised learning algorithms on three existing applications, such as stock data mining, undergraduate admission data and breast lesion detection, have been discussed in the work. It has been observed that, the support vector machine is a commonly used algorithm among all the supervised algorithms and shows good results in terms of different performance metrics. In this paper, we have compared SVM for all three applications. However, some more applications are needed to be explored for detailed analysis of the algorithms. As a future scope of the study, more applications can be used to analyse the best-suited supervised algorithm for classification.

REFERENCES

[1] A. L. Samuel, “Some studies in machine learning using the game of checkers. ii—recent progress,” Computer Games I, pp. 366–400, 1988.

[2] S. B. Kotsiantis, I. Zaharakis, P. Pintelas et al., “Supervised machine learning: A review of classification techniques,” Emerging artificial intelligence applications in computer engineering, vol. 160, no. 1, pp. 3–24, 2007.

[3] U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley, “A brief survey of machine learning methods and their sensor and iot applications,” in 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA). IEEE, 2017, pp. 1–8

[4] N. Saleem and M. I. Khattak, “A review of supervised learning algorithms for single channel speech enhancement,” International Journal of Speech Technology, vol. 22, no. 4, pp. 1051–1075, 2019.

[5] M. Sharma, S. Sharma, and G. Singh, “Performance analysis of statistical and supervised learning techniques in stock data mining,” Data, vol. 3, no. 4, p. 54, 2018.

[6] T. Lux, R. Pittman, M. Shende, and A. Shende, “Applications of supervised learning techniques on undergraduate admissions data,” in Proceedings of the ACM International Conference on Computing Frontiers, 2016, pp. 412–417.

[7] F. Mutlu, G. Cetinel, and S. Gul, “A fully-automated ¨ computer-aided breast lesion detection and classification system,” Biomedical Signal Processing and Control, vol. 62, p. 102157, 2020.

[8] J. Alzubi, A. Nayyar, and A. Kumar, “Machine learning from theory to algorithms: an overview,” in Journal of physics: conference series, vol. 1142, no. 1. IOP Publishing, 2018, p. 012012.

[9] N. Pahwa, N. Khalfay, V. Soni, and D. Vora, “Stock prediction using machine learning a review paper,” International Journal of Computer Applications, vol. 163, no. 5, pp. 36–43, 2017.

[10] R. Choudhary and H. K. Gianey, “Comprehensive review on supervised machine learning algorithms,” in 2017 International Conference on Machine Learning and Data Science (MLDS). IEEE, 2017, pp. 37–43.

[11] D. Dhall, R. Kaur, and M. Juneja, “Machine learning: a review of the algorithms and its applications,” Proceedings of ICRIC 2019, pp. 47–63, 2020.

AUTHORS

Mrs. Vijayalakshmi Sarraju is pursuing her PhD from Birla |Institute of Technology Extension Centre Lalpur. She completed her M. Sc in Mathematics from Nagarjuna University, Andhra Pradesh. She has 22 years of teaching experience. She has expertise in Statistics, Quantitative Techniques, discrete mathematical structures, and Numerical methods. Her research interest is inferential statistics and machine learning.

Dr. Jaya Pal is an assistant professor in the Department of Computer Science and Engineering of Birla |Institute of Technology Extension Centre Lalpur. She completed her M.Sc. and MCA in Mathematics and PhD in Technology. She has 16 years of teaching experience and 10 years of research experience. Her research interests are Fuzzy Logic & its Applications, Soft Computing, Machine Learning, Software Quality Prediction, and Data Mining.

Dr. Supreeti Kamilya is an assistant professor in the Department of Computer Science and Engineering of Birla |Institute of Technology Mesra. She completed her M.Tech and PhD in Computer Science from IIEST, Shibpur. She has 8 months of teaching experience and 7 years of research experience. Her research interests are Theoretical Computer Science, Chaos Theory, Machine Learning and Deep Learning.

AN ANALYSIS OF PHRASE BASED SMT FOR ENGLISH TO MANIPURI LANGUAGE

Maibam Indika Devi[1] and Bipul Syam Purkayastha[2]

[1]Department of Computer Science, IGNTU-RCM, Kangpokpi, Manipur, India
[2]Department of Computer Science, Assam University, Silchar, Assam, India

ABSTRACT

Statistical Machine Translation (SMT) is one ruling approach adopted for developing major translation systems today. Here, we report a phrase-based SMT system from English to Manipuri. The variance in the structure and morphology between English and Manipuri languages and the lack of resources for Manipuri languages pose a significant challenge in developing an MT system for the language pair. In comparison, English has poor morphology and SVO structure and belongs to the Indo-European family. Manipuri language has richer morphology and SOV structure and belongs to the Sino-Tibetan family. Manipuri has two scripts- Bengali script and Meitei script. Here the Bengali script is used for developing the system. Our system uses the Moses toolkit. We train and test the system using the tourism, agriculture and entertainment corpus. Further, we use the BLEU metric to evaluate the systems' performance.

KEYWORDS

Phrase-based SMT, English- Manipuri, Moses, BLEU.

1. INTRODUCTION

Machine Translation (MT) is an important area in Natural Language Processing (NLP) where many systems are being developed worldwide for translation from one language to another. It aids in the translation process, be it a book, movies, official documents from one language to another. English is a simple and easy to learn language. Most documents, books, journals, articles, web pages are available in English. However, with the application of MT, articles or web pages can be viewed in a different language. There are various techniques to MT of which Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) is the most prominent. SMT is the technique where translation is done through statistical models.. In a phrase based SMT model, the translation units are phrases rather than words. To perform translation, phrases in source language will be mapped with target language phrases, by using maximum likelihood estimate; the best translation out of many candidate translations will be selected. We use the open source toolkit Moses[1] to implement the phrase based model of SMT technique.

English is the source language, and Manipuri is the target language. The language pair treated here, English and Manipuri, is a challenging one because of the huge difference in terms of linguistic structure. Manipuri belongs to the Sino-Tibetan family, has SOV structure, tonal, rich morphology, aspect predominance and synthetic category and agglutinating. While English comes under the Indo-European family, has SVO structure, stressed and non-tonal, poor morphology, tense dominant and analytic category. The huge difference in the language structure

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 37-43, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121904

provides a challenge while performing translation. In addition to this, there is a lack of reliable pre-processing tools and resources for the Manipuri language. Though some works are seen for Manipuri language, the existing tools are not satisfactory and reliable.

There are also very limited resources for the Manipuri language. Only a few sentences of bilingual data are available to the public for research. These data are not enough for efficient working and quality output of the system. So, limited corpus and tools, difference in structure and morphology of language under consideration pose challenges in developing the translation system.

The bilingual and monolingual corpora used for training the system consist of varying domains from tourism, agriculture and entertainment. Some of these corpora are downloaded from TDIL[2] website while some are manually developed. For each domain, we train and test separate phrase based SMT systems. We evaluate how the system performs for these domains having different data sizes using automatic evaluation metric.

2. RELATED WORK

[3] has written a survey paper on the various approaches to machine translation and the major translation systems developed for Indian languages. Most of the major Indian languages has well developed machine translation systems. The general purpose Google Translate[4] provides good results. However when dealing with specific domain related translations, tailor made MT systems trained on that domain will better serve its use.

The North-East section of India has a diversity of languages with multiple dialects. [5]has discussed the works carried out in NLP for north-eastern languages covering Hindi, Manipuri, Assamese, Kokborok, Nepali, Mizo, Bodo, Bengali etc. Nevertheless, NLP related advancements are found in the works of Assamese[6], Nepali [7], Bodo[8]. An open-access NLP toolkit[9] dedicated to Bengali is available also. In comparison to them, Manipuri language is lagging far behind. Even though recognized by the Indian Union as one of the scheduled languages, there is little work in NLP applications. The non-availability of resources, language characteristics, and lack of experts poses some of the factors that hinder its development. The survey report[10] covers areas on E-dictionary, Machine Translation, POS tagging, WordNet, Word Sense Disambiguation, Multi-word expressions, Name Entity Recognition, Morphological Analysis. Some of the Manipuri language related MT works are as follows.

[11]developed the Manipuri-English Example-based Machine Translation. The corpora used here is of news domain with POS tagging, NER, morphological analysis and chunking applied. They have measured the output using BLEU and NIST metrics, scoring 0.317 and 3.361, respectively, depending on which claims has been made that the EBMT approach is better than baseline SMT on using the same set of data. [12] developed Manipuri-English Bidirectional SMT systems. Their system used a corpus from the news domain with 10350 sentence pairs for training and 500 sentences for testing. Apart from using the statistics of the corpus, they have incorporated additional morphological information into the system. The English-Manipuri pair has incorporated suffix dependency relations on the source side and case markers on the target side. While for Manipuri-English pair, case markers, POS tags on the source side and suffix and dependency relationships on the target side. Both the translations showed improved results from the baseline system as given by their BLEU score. [13] developed the factored SMT for the English-Manipuri language pair. Suffix and dependency relations are treated as factors on the source side and case markers on the target side. The system is trained using 10350 sentences and tested on 500 sentences. The output shows an improved BLEU score. [14] has carried out the

Phrase-based SMT for Manipuri languages by integrating reduplicated MWE. They have stated that the integration improves the BLEU and NIST scores over baseline SMT. [15] has carried out machine translation from English to Manipuri using SMT and NMT. The output comparison shows NMT having a higher BLEU score as compared to phrasal SMT. In their work "Unsupervised Neural Machine Translation for English and Manipuri" they reported a BLEU score of 3.1 for English-Manipuri translation and 2.7 for Manipuri-English translation. [16] developed the Manipuri-English translation system using the intelligence domain. They used a corpora of 56,678 size from the intelligence domain based on open-source intelligence (OSINT). Evaluation of SMT and NMT is done based on BLEU score, where NMT outperforms SMT. They also incorporated suffix based morphological analysis information which further improves the BLEU score.

3. DEVELOPING THE SYSTEM

3.1. Corpus Preparation

Parallel corpora are a collection of sentences of two different languages which are aligned at the sentence level. The bilingual parallel corpus will be used to train the system. It is the quality and quantity parallel corpus fed into the system that characterizes the result of translations. Therefore bigger the corpora better the system performance. Tourism corpus from TDIL[2] along with newly developed corpus of entertainment, tourism and agriculture are used for training. Table 1 shows the corpus distribution of the different domains used in developing the system.

Table 1. Corpus distribution of different domains used in training the system.

Domain	Translation Model	Language Model
Agriculture	10,000	500
Entertainment	10,000	500
Tourism	25,000	1000

3.2. Preprocessing

The preprocessing steps include tokenizing, true casing and cleaning of parallel data.Tokenizing is the step for identifying tokens such as words, numbers, and punctuations. We use the inbuilt tokenizer for English sentences. Moses inbuilt tokenizers have no support for Manipuri language. So, we use the IndicNLP tokenizer[17]. After this, we perform truecasing and cleaning of the tokenized output. In the cleaning step, we set the length limit to 80.

3.3. English To Manipuri System

In SMT, a source language sequence 'e' (English) is translated into a target language sequence 'm', by computing the most likely translation using the following equation[18] ,

$$p(e,m) = argmax_m p(m|e)(Equation1)$$

Using Bayes Rule [19], Equation 1 is written as-

$$argmax_m p(m|e) = argmax_m p(e|m)p(m)(Equation2)$$

In Equation 2, the component p(m|e) in Equation 1 is decomposed into two components. The component p(m) is the language model, and another component p(e|m) is the translation model,

which is discussed in the latter part of the paper. The English to Manipuri MT system is developed using the phrase-based SMT technique. In phrase-based SMT, the translation units are phrases. A foreign English sentence is segmented into phrases, and each English phrase is mapped into Manipuri phrases. The phrases can also be reordered. As compared to baseline SMT where the translation units are words, the phrase-based SMT provides better results. Various toolkits are available to implement the SMT model of machine translation, Moses being one of them. Moses is open sourced and the most commonly used toolkit for developing SMT systems. The two main components of Moses, the training pipeline and the decoder, form the basis for translation. Apart from this, Moses consists of multiple tools and utilities and also supports various external tools. Developing a translation system from training data requires multiple stages, where the stages are implemented in a pipelined manner, hence the name training pipeline. Moses provides the advantage to add various external tools during the training pipeline. However, the parallel corpora is not used directly for training the system. They are preprocessed first.

3.4. Language Model

In Equation 2, the component p(m) represents the language model. Only the monolingual target side corpus (Manipuri corpus) is required to create the language model. The size of the monolingual corpus used here is separate from that used in training. The language model makes the translation system aware of how the target language should appear and ensures fluent output. For creating language models, the following monolingual corpora are used.

1. Monolingual Manipuri sample general corpus from TDIL.
2. Monolingual Manipuri sample raw corpus from NPLT [20].

Moses supports various language modeling tools such as KenLM, IRSTLM, SRILM and RandLM. Here, the built-in KenLM model is being used. Here, a 3-gram modeling technique is used to compute the probability of Manipuri sentences, denoted by p(m). The component p(m) is calculated based on Markov's Chain Rule [18] as,

$$p(m) = \Pi_{i=0}{}^{m} \quad p(m_i|m_{i-1}, m_{i-2})(Equation3)$$

Where m_i is the current word generated.

3.5. Translation Model

The component p(e|m) in Equation 2 is the translation model. As the component shows, bilingual parallel corpora are involved in creating the translation model. It estimates the lexical correspondence between the languages. The translation model computes the probability of a source sentence for a given target sentence and tries to find the best translation of a given phrase. The probability p(e|m) is computed as the summation of all probabilities with possible alignment 'a' between the phrases of English and Manipuri language,

$$p(e|m) = \sum_a p(e, a|m)(Equation4)$$

Here, the built-in tool GIZA++ aligns the words between the source and target languages. In phrase based SMT, the translation units are phrases; therefore, the translation model is built based on the frequency of occurrences of phrases in the training corpus. This information is stored in a table called a phrase-table which contains the phrases and their frequency over the entire training corpus the higher the frequency of a phrase, the greater the chances of getting a correct

translation. The phrase table forms the translation model for the system. The role of the translation model is to ensure that the source language and target language are good translations of one another.

3.6. Decoder

The decoder takes input sentences in the source language (English) and uses the translation model and language model to translate them into the target language (Manipuri). The decoder is responsible for determining the best translation out of its many candidate translations. It uses the argmax() function to find the maximum translation probability of all candidate translations. There are many tools for decoding in SMT systems. Here, the inbuilt Moses decoder is used.

4. EVALUATING THE SYSTEM

4.1. BLEU

BLEU (Bilingual Evaluation UnderStudy) is the dominant and language-independent metric for measuring translation quality. [21]BLEU score counts the number of matches in a weighted fashion, the consecutive phrases between the machine-translated output and the reference translations made by humans. Out of different BLEU available, we use multibleu.perl to determine the score.

Table 2. BLEU score of the different systems

Domains	Training size	Test data size	BLEU
Agriculture	10,000	1000	7.03
Entertainment	10,000	1000	6.52
Tourism	25,000	1000	14.59

We keep the size of test data the same for all three systems. The training data size is however different. Running a multi-bleu script gives the score in Table 2. As we can see, the size of training data affects the BLEU score. It is common perception that higher the value of BLEU, the better. Moreover, BLEU is directly dependent on the domain size and test data used for training and testing. In our work, however, we do not compare our BLEU scores with those of other MT reports on Manipuri language. This is due to [22] which stated BLEU scores in between papers cannot be compared directly. And that BLEU scores vary with the MT system change, corpus domain, test data, and language pair. Therefore, it is better not to compare the quality of a system, solely based on its BLEU score.

5. CONCLUSION

In our work, we rate the systems of different domains based on BLEU score only to see their result with the little amount of data at hand. The data set used here is insignificant for a well functioning system, however this paper provides an insight on the practicality of the SMT technique on English-Manipuri translations. In our analysis, we get to see that even for such a small amount of training data, the system provides a good output.

SMT is one of the dominating approaches to MT. Nevertheless, there is a recent shift from SMT to NMT in the paradigm of MT. And for both techniques corpus serves as the backbone for functioning. For the language pair English-Manipuri which has got distant linguistic features, it cannot be assumed unless experiment and compare their results which technique will be more

practical. This paper forms a preliminary basis, towards understanding the feasibility and potential of the phrase based SMT system. Our next work will compare NMT and SMT over the same amount of training and testing data.

ACKNOWLEDGEMENTS

We are grateful to Technology Development for Indian Languages (TDIL), National Platform for Language Technology (NPLT) for providing us the corpus and to Professor Chungkham Yashawanta Singh, Fellow at the Indian Institute of Advanced Studies (IIAS) for his support in developing corpus.

REFERENCES

[1] P. Koehn *et al.*, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, Jun. 2007, pp. 177–180. Accessed: Sep. 06, 2021. [Online]. Available: https://aclanthology.org/P07-2045

[2] "English-Manipuri Sentences of Tourism Domain." https://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=456&lang=en (accessed Aug. 17, 2022).

[3] P. Antony, "Machine Translation Approaches and Survey for Indian Languages," *Int J Comput Linguist. Chin Lang Process*, 2013.

[4] "Google Translate." https://translate.google.co.in/ (accessed Aug. 16, 2022).

[5] S. Islam, M. I. Devi, and B. S. Purkayastha, "A Study on Various Applications of NLP Developed for North-East Languages," vol. 9, p. 12, 2017.

[6] R. R. Deka, S. Kalita, M. P. Bhuyan, and S. K. Sarma, "A Study of Various Natural Language Processing Works for Assamese Language," in *Intelligent Techniques and Applications in Science and Technology*, Cham, 2020, pp. 128–136. doi: 10.1007/978-3-030-42363-6_15.

[7] T. B. Shahi and C. Sitaula, "Natural language processing for Nepali text: a review," *Artif. Intell. Rev.*, vol. 55, no. 4, pp. 3401–3429, Apr. 2022, doi: 10.1007/s10462-021-10093-1.

[8] M. Narzary, G. Muchahary, M. Brahma, S. Narzary, P. K. Singh, and A. Senapati, "Bodo Resources for NLP - An Overview of Existing Primary Resources for Bodo," *AIJR Proc.*, Jul. 2021, Accessed: Aug. 17, 2022. [Online]. Available: https://books.aijr.org/index.php/press/catalog/book/115/chapter/1126

[9] S. Sarker, "BNLP: Natural language processing toolkit for Bengali language." arXiv, Dec. 01, 2021. doi: 10.48550/arXiv.2102.00405.

[10] M. I. Devi and B. S. Purkayastha, "Advancements on NLP Applications for Manipuri Language," 2018. doi: 10.5121/ijnlc.2018.7505.

[11] T. D. Singh and S. Bandyopadhyay, "Manipuri-English Example Based Machine Translation System," 2011. /paper/Manipuri-English-Example-Based-Machine-Translation-Singh-Bandyopadhyay/080aa68650d13a770bb7a228c10f17ce31baea21 (accessed Mar. 20, 2021).

[12] T. D. Singh and S. Bandyopadhyay, "Manipuri-English Bidirectional Statistical Machine Translation Systems using Morphology and Dependency Relations," *undefined*, 2010, Accessed: Sep. 06, 2021. [Online]. Available: https://www.semanticscholar.org/paper/Manipuri-English-Bidirectional-Statistical-Machine-Singh-Bandyopadhyay/68d64336fb2ac7d302d3ae45051127484754b174

[13] T. D. Singh and S. Bandyopadhyay, "Statistical Machine Translation of English-Manipuri using Morpho-syntactic and Semantic Information," 2010.

[14] T. D. Singh and S. Bandyopadhyay, "Integration of Reduplicated Multiword Expressions and Named Entities in a Phrase Based Statistical Machine Translation System," 2011.

[15] S. M. Singh and T. D. Singh, "Unsupervised Neural Machine Translation for English and Manipuri," p. 10.

[16] L. Rahul, L. Meetei, and H. Jayanna, "Statistical and Neural Machine Translation for Manipuri-English on Intelligence Domain," 2021, pp. 249–257. doi: 10.1007/978-981-33-6987-0_21.

[17] A. Kunchukuttan, "Indic NLP Resources." Jun. 13, 2022. Accessed: Jun. 25, 2022. [Online]. Available: https://github.com/anoopkunchukuttan/indic_nlp_resources

[18] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation," p. 8.
[19] J. V. Stone, *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*. Sebtel Press, 2013.
[20] "A Gold Standard Manipuri Raw Text Corpus." https://nplt.in/demo/index.php?route=product/product&product_id=1987&search=manipuri (accessed Aug. 18, 2022).
[21] laujan, "Legacy: What is a BLEU score? - Custom Translator - Azure Cognitive Services." https://docs.microsoft.com/en-us/azure/cognitive-services/translator/custom-translator/what-is-bleu-score (accessed Aug. 30, 2022).
[22] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Belgium, Brussels, 2018, pp. 186–191. doi: 10.18653/v1/W18-6319.

ROBERTA GOES FOR IPO: PROSPECTUS ANALYSIS WITH LANGUAGE MODELS FOR INDIAN INITIAL PUBLIC OFFERINGS

Abhishek Mishra[1] and Yogendra Sisodia[2]

[1]Trust Group, India
[2]Conga, India

ABSTRACT

With the advent of large-scale language models in natural language processing (NLP), extracting valuable information from financial documents has gained popularity among researchers, and deep learning has boosted the development of effective text mining models. Prospectus text mining is very important for the investor community to identify major risk factors and evaluate the usage of the amount to be raised during an IPO. In this paper, we investigate how the recently introduced pre-trained language model Roberta can be adapted for this task. We also introduced prospectus-specific sentence transformers for semantic textual similarity along with a dataset to verify the efficacy of our work.

KEYWORDS

IPO, Prospectus, Large Language Models, Semantic Textual Similarity.

1. INTRODUCTION

An Offer Document refers to the prospectus containing information about the public offering or offer for sale. This document contains all the information an investor needs to make an informed investment decision. The prospectus is analysed by merchant bankers, stockbrokers, and the investor community to identify various risk factors and answer questions such as: what are the risk factors involved? What are the related party transactions? Where will the money be deployed after listing, etc. All the financial statements prior to the IPO and any legal disclosures are available in the final offer document. This prospectus contains all the pertinent information an investor needs to make an investment decision. The final offer document must be made public before the company can list in the Indian securities market.

Natural language processing, with the advent of large-scale language models, has been implemented in a variety of fields, including legal and biomedical [1] and [2]. We intend to apply the same methodology to the IPO.

Our contribution is enumerated below:

- We built a large-scale language model for India's IPO. Our solution based on Roberta will help with a wide range of use cases, such as answering questions, recognizing named entities, and classifying sentences and paragraphs.
- We are also presenting sentence transformers based on our large-scale language model. There are a variety of use cases, including semantic similarity and zero-shot learning.

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 45-53, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121905

- We are also making public two datasets:
  - OCR text for 100 prospectuses (PDFs are already in the public domain on the website of the Market Regulator.)
  - One dataset containing pairs of semantically similar sentences was extracted from these prospectuses and annotated by one of the authors, who is a subject matter expert.

2. RELATED WORK

Based on BERT [1], SCIBERT [2] is a pretrained language model for scientific text. SCIBERT was tested on a wide range of scientific domain-specific tasks and datasets. SCIBERT does a lot better than BERT-Base and gets new SOTA results on many of these tasks. LEGAL-BERT [3] is a family of BERT models for the legal domain that achieves state-of-the-art outcomes in many end-tasks. Notably, the performance gains are bigger for the hardest end-tasks (such as multi-label classification in ECHR-CASES and contract header, lease details in CONTRACTS-NER), where domain-specific knowledge is more important. BioBERT [4] is a language representation model that has been pre-trained for biomedical text mining. BioBERT does better than previous models at biomedical text mining tasks like NER, RE, and QA, with only minor changes to the architecture for each task. BioBERT's pre-release version has already been shown to be very good at several biomedical text mining tasks, such as NER for clinical notes. FinBERT is an extension of BERT for the financial domain that was pre-trained on a financial corpus and further fine-tuned for sentiment analysis. The authors achieved state-of-the-art results on both datasets employed by a significant margin. For the classification task, this improved the accuracy of the state-of-the-art by 15%.

The closest things we found that might be related to our domain are FinBert and LegalBert However, on close inspection, we found that FinBert focuses on sentiment analysis and LegalBert focuses more on legal cases and contracts. There isn't.

The most prominent large language model is the BERT model architecture. It is based on a multilayer bidirectional transformer. Instead of the traditional left-to-right language modelling goal, BERT is trained on two tasks: predicting randomly masked tokens and predicting whether two sentences go together or not. There are two primary ways to train a domain-specific language modelling task: a) fine-tuning and b) training from scratch. The key difference is that language model fine tuning begins with a model that has already been trained, whereas training a language model from scratch begins with an untrained, randomly initialised model.

2.1. Fine-tuning the Large Language Model

When refining a language model, a previously pre-trained model (e.g., bert-base, etc.) is retrained on a new unlabeled text corpus (using the original, pre-trained tokenizer). In general, this is advantageous if we intend to use a pre-trained individual for a specific task in which the language employed may be highly technical and/or specialized. The technique was utilised effectively in the SciBERT paper. For computational purposes, we have opted for this method.

2.2. Large Language Model Training from Scratch

A brand-new, randomly initialised model is trained on a massive block of text. This will also improve a tokenizer so that it works best with the data you provide. This comes in especially handy when training a language model for a language that lacks publicly available pre-trained models. However, the computational cost of this method is high.

3. METHODOLOGY

3.1. Dataset Preparation

We downloaded 100 prospectuses from the Market Regulator Website [6]. A prospectus is a very long document with close to 500 pages. The number of pages is 41,697. We used open-source Tesseract OCR [7] to extract text for each document. Text is stored in a JSON file. A prospectus is in Pdf format, so we converted Pdf to images (using Poppler Utilities [8]) and then applied Tesseract. The total number of words is 2,20,60,749.

Figure 1. Corpus Preparation

3.1.1. Prospectus STS

A domain expert author prepared a Semantic Textual Similarity (STS) dataset on 1,598 sentence pairs. The domain expert author was given a random set of sentences (derived from the NLTK sentence tokenizer) and asked to find similar sentences and put them in an STS format. 1,373 pairs are semantically similar. The rest are dissimilar cases.

3.1.2. Prospectus Labels

Also, the domain expert author annotated 18 classes for semantically similar cases. A sentence can have multiple classes, and semantically similar cases have the same classes. These are labels used for identifying various risk factors in a prospectus, e.g., utilisation of funds, operational and currency risk.

Table 1. Semantically Similar Sentences

Sentence 1	**Sentence 2**
This being the first public issue of our corporation, there has been no formal market for the Equity Shares of our Corporation	No assurance can be given regarding an active or sustained trading in the Equity Shares nor regarding the price at which the Equity Shares will be traded after listing.
Investments in equity and equity-related securities involve a degree of risk and investors should not invest any funds in the Offer unless they can afford to take the risk of losing their entire investment.	Investors are advised to read the risk factors carefully before taking an investment decision in the Offer.
Unless the context requires otherwise, the financial information in this Prospectus is derived from the Restated Consolidated Financial Statements of our Corporation comprising	Risk Factors – Significant differences exist between Indian GAAP and other accounting principles, such as U.S. GAAP and IFRS
We have included certain non-GAAP financial measures and certain other selected statistical information related to our business,	non-GAAP financial measures and are significantly different from those of non-insurance companies and may require certain estimates and assumptions in their calculation
Investors may be subject to Indian taxes arising out of the sale of the Equity Shares	unless specifically exempted, capital gains arising from the sale of equity shares held as investments in an Indian company are generally taxable in India

Table 2. Semantically Dissimilar Sentences

Sentence 1	**Sentence 2**
This section of Indian society is characterized by low levels of financial literacy and technology use, lack of financial	judicial precedent may be time consuming as well as costly for us to resolve and may impact the viability of our current business.
Except as disclosed in chapter titled “Financial Statements” beginning on page 1494of this Draft Red Herring .	Stringent quality control is followed during the production process by the quality control department by
Credit risk is the risk of financial loss to our Company if a customer or counterparty to a financial instrument fails to meet	Our total expenses marginally increased by 0.49% to = 14,834.82 million in Fiscal 2020 from % 14,762.64 million
Revenue from contracts with customers is recognised upon transfer of control of promised goods/ services	Except as stated in the chapter titled “Capital Structure” beginning on page 63 of this Red Herring Prospectus
We have a wide variety of 18 different vegetarian and non-vegetarian burgers covering both value and premium	it shall provide reasonable assistance to our Company and the BRLMs in the taking of all steps

Table 3. Class labels

Class	**No of Examples**	**Example Sentence**
Operational Risk	6	human and systems errors when executing complex and high-volume transactions
Intellectual Property	100	We cannot ensure that our intellectual property is protected from copy or use by others, including our competitors, and intellectual property

		infringement actions may be brought against us
Employee Reservation Portion	78	DISCOUNT OF ₹45 PER EQUITY SHARE WAS OFFERED TO THE RETAIL INDIVIDUAL BIDDERS BIDDING IN THE RETAIL PORTION AND THE ELIGIBLE EMPLOYEES BIDDING IN THE EMPLOYEE RESERVATION PORTION
Remuneration	122	Our Directors, Key Managerial Personnel and the Promoter have interests in us other than reimbursement of expenses incurred or normal remuneration or benefits
Currency Risk	14	Fluctuations in the exchange rate between the Rupee and other currencies could have an adverse effect
Utilisation of funds	190	Monitoring Utilization of Funds
Liquidity Risk	4	Our investment portfolio is subject to liquidity risk, which could adversely affect its realizable value
environmental social and governance	12	We continue to undertake various initiatives towards this, including alleviating poverty, pursuing inclusive growth, promoting gender equality, promoting good health, reducing our carbon footprint through consumption rationalisation and using eco-friendly technology
Credit Risk	58	We are subject to the credit risk of the issuers whose debt securities we hold
Risk Disclosure	374	This being the first public issue of our corporation, there has been no formal market for the Equity Shares of our Corporation
Remuneration	120	None of our Directors are entitled to remuneration from our Subsidiaries or Associates
Offer	530	The determination of the Price Band is based on various factors and assumptions.
Market Risk	118	Fluctuations in the exchange rate between the Rupee and other currencies could have an adverse effect
Litigations	384	To material litigation in (iv) above, our Board has considered and adopted the following policy on materiality with regard to outstanding litigation
Investor Taxation	66	Investors may be subject to Indian taxes arising out of the sale of the Equity Shares
Related Party Transactions	192	A summary of related party transactions entered into by our Company with related parties as at and for the nine months ended December
Financial Statements	386	Unless the context requires otherwise, the financial information in this Prospectus is derived from the Restated Consolidated Financial Statements of our Corporation
Audit	54	The Statutory Auditor to the Offer has included certain matters of emphasis in its examination report

As part of next steps Authors want to do more deep work with these classes.

3.2. Methods

3.2.1. Fine-Tuning Roberta

In BERT literature, there are two training objectives: Masked Language Model (MLM) and Next Sentence Forecast (NSP). In MLM random subset of the tokens in the input sequence is selected and replaced with the token special [MASK] Cross-entropy loss in predicting masked tokens is the MLM objective. 15% of the input tokens are chosen uniformly for potential replacement by BERT. 80% of the tokens are replaced with [MASK], 10% are left alone, and 10% are replaced with a randomly selected vocabulary token.

NSP is a binary classification loss for predicting whether two segments in the original text follow one another. The creation of positive examples involves selecting consecutive sentences from the text corpus. Negative examples are created by combining sections from various documents. Positivity and negativity are sampled with equal likelihood.
The purpose of the NSP objective was to improve performance on downstream tasks, such as NLI, that require reasoning about the relationships between pairs of sentences.

When pretraining BERT models, the Roberta [9] Team evaluates several design decisions with great care. They found that performance can be greatly improved by training the model longer, in larger batches, with more data, by removing the goal of predicting the next sentence, by training on longer sequences, and by changing the masking pattern on the training data in real time.

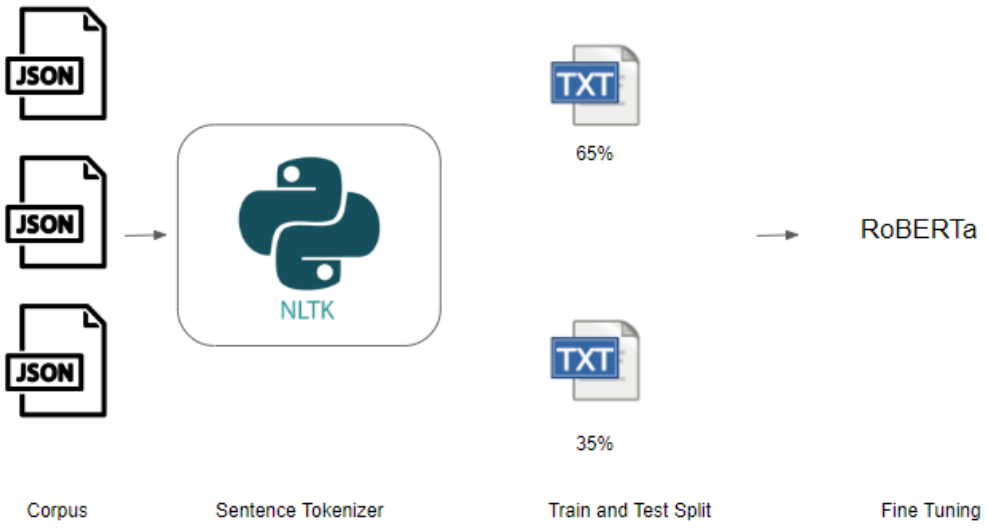

Figure 2. Data Preparation for Roberta Fine-Tuning

We divided the data into 65 and 35% for training and testing. Roberta is fed with sentences derived from NLTK's sentence tokenizer [10]. We fine-tuned our model for 10 epochs. Result is shown in next section.

3.2.2. Sentence Transformer for Semantic Textual Similarity

We used TSDAE-Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning [11] to further get sentence transformer from our pre-trained Roberta TSDAE is a robust method for domain adaptation and pre-training sentence embeddings.

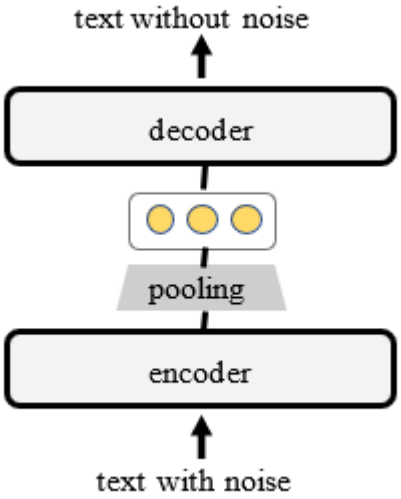

Figure 3. TSDAE Architecture

TSDAE trains sentence embeddings by introducing a specific type of noise (e.g., deleting or exchanging words) into input sentences, encoding the damaged sentences into fixed-size vectors, and then reconstructing the vectors into the original input. The formal training objective is:

$$
\begin{aligned}
(\theta) &= \mathbb{E}_{x\sim D}[\log P_\theta(x|\tilde{x})] \\
&= \mathbb{E}_{x\sim D}[\sum_{t=1}^{l} \log P_\theta(x_t|\tilde{x})] \\
&= \mathbb{E}_{x\sim D}[\sum_{t=1}^{l} \log \frac{\exp(h_t^T e_t)}{\sum_{i=1}^{N} \exp(h_t^T e_i)}]
\end{aligned}
$$

where D is the text corpus, $x = x_1x_2 \cdots x_l$ is the input text training sentence with l no of tokens, $\tilde{x}$ is the equivalent broken sentence, e_t is the word embedding of x_t, N is the vocabulary size and h_t is the hidden state at t decoding step.

All resources for data are available at this link:
https://github.com/scholarly360/ProspectusRoberta

4. RESULTS

4.1. Fine-Tuning Roberta

Perplexity measures, given a model and an input text sequence, the likelihood that the model will generate the input text sequence. It can be used as a metric to evaluate how well the model has learned the distribution of the text it was trained on for the language generation task. Our perplexity on the test dataset was 2.7935, which is a very decent and attainable score.

4.2. Semantic Textual Similarity

The Spearman and Pearson correlation coefficients are normally used for the evaluation of STS datasets. The main difference between the Pearson and Spearman correlation coefficients is that the Pearson value assumes that the two variables are related in a linear way, while the Spearman value also considers monotonic relationships. Table 1 shows our Roberta-based sentence transformer performed better compared to other state-of-the-art sentence transformers [12].

Table 1. Prospectus STS Evaluation Results

Model	**Pearson**	**Spearman**
multi-qa-mpnet-base-dot-v1	0.7228	0.597
all-mpnet-base-v2	0.7369	0.5939
Our Sentence Transformer	0.7859	0.6023

These results show that domain adaption for prospectus is working well. The authors want to further define more problem statements, collect more data, and train larger models.

5. CONCLUSIONS

We have released a Large Language Model based on Roberta for analysing prospectuses. We also put out a Roberta-based sentence transformer that was trained with TSDAE and did a better job on an STS data set with text derived from the Prospectus. This will help investors and the merchant bank community to explore prospectuses in a more automated way, thus saving time.

In the future, we want to explore more with additional use cases such as sentence classification (Zero Shot and Few Shot Classifiers) and token classification. Also, we want to tag our data with the help of multiple annotators so that we can make as accurate a copy as possible of real-world problems.

ACKNOWLEDGEMENTS

The authors would like to thank everyone in their respective organisations for fully supporting them.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee Kristina, Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” 2018. [Online]. Available: arXiv:1810.04805.

[2] Iz Beltagy, Kyle Lo, Arman Cohan, “SciBERT: A Pretrained Language Model for Scientific Text” 2019. [Online]. Available: arXiv:1903.10676.

[3] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, Ion Androutsopoulos, “LEGAL-BERT: The Muppets straight out of Law School” 2020. [Online]. Available: arXiv:2010.02559.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang,” BioBERT: a pre-trained biomedical language representation model for biomedical text mining” 2018. [Online]. Available: arXiv:1901.08746,

[5] Dogu Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.”, 2019. [Online]. Available: arXiv:1908.10063,

[6] Public Issues, SEBI, 2022. [Online]. Available: https://www.sebi.gov.in/sebiweb/home/HomeAction.do?doListing=yes&sid=3&ssid=15&smid=12.
[7] Tesseract Documentation, Tesseract. 2022. [Online]. Available: https://tesseract-ocr.github.io/tessapi/4.0.0/.
[8] Poppler, a PDF rendering library, Poppler. 2022. [Online]. Available: https://github.com/freedesktop/poppler/.
[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov," RoBERTa: A Robustly Optimized BERT Pretraining Approach" 2019. [Online]. Available: arXiv:1907.11692.
[10] nltk.tokenize package, NLTK, 2022. [Online]. Available: https://www.nltk.org/api/nltk.tokenize.html.
[11] Kexin Wang, Nils Reimers, Iryna Gurevych, "TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning" 2021. [Online]. Available: arXiv:2104.06979.
[12] Sentence Transformer Pretrained Models, Sentence Transformer. 2022. [Online]. Available: https://www.sbert.net/docs/pretrained_models.html.

AUTHORS

As Deputy Vice President at Trust Capital, Abhishek Mishra oversees new issues, private placements, and secondary sales in the Indian Fixed Income markets.

Yogendra Sisodia is Director, Machine Learning at Conga. Some current areas of research interest are semi-supervised learning, adversarial machine learning, and deep learning applications in computer vision.

COMPARISON OF SEQUENCE MODELS FOR TEXT NARRATION FROM TABULAR DATA

Mayank Lohani, Rohan Dasari, Praveen Thenraj Gunasekaran, Selvakuberan Karuppasamy and Subhashini Lakshminarayanan

Data and AI, Advance Technology Centers in India,
Accenture, Gurugram, India

ABSTRACT

This paper demonstrates our work on the survey of pre-trained transformer models for text narration from tabular data. Understanding the meaning of data from tables or any other data source requires human effort and time to interpret the content. In this era of internet where data is exponentially growing and massive improvement in technology, we propose an NLP (Natural Language Processing) based approach where we can generate the meaningful text from the table without the human intervention. In this paper we propose transformer-based models with the goal to generate natural human interpretable language text generated from the input tables. We propose transformer based pre-trained model that is trained with structured and context rich tables and their respective summaries. We present comprehensive comparison between different transformer-based models and conclude with mentioning key points and future research roadmap.

KEYWORDS

Survey, NLP (Natural Language Processing), Transformers, Table to Text.

1. INTRODUCTION

In recent years, Natural language processing (NLP) is massively growing field and being used in wide range of applications and features. Table to text generation is a segment, which aims at generating meaningful and descriptive text about defined information in structured data. Few applications in this segment include generating sentences given medical tabular data, descriptions of restaurant menus given meaningful representations, summaries from cricket/football game score tables, generating meaningful text from tables in Wikipedia, converting statistical stock market tables to easily understandable textual form, etc. Existing table to text models has provided a checkpoint for narrating text from tables but those are not completely efficient and reliable as they are exposed to hallucination which means the generated text is meaningful, but it is not related to the source that means it lacks faithfulness.

In this paper we have chosen 3 transformer-based models namely:

- T5
- BART
- mBART

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 55-60, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121906

2. LITERATURE REVIEW

Mike Lewis et al., proposes the research work for BART [1]. It introduces BART as a pre-training model approach that learns to map corrupted documents/text to the original one. This paper contains the working architecture as well as the details about the pretrained model and experimental results along with the comparison with other models.

Yinhan Liu et al., proposed the research work for MBART [2]. In this work, authors presented mBART that is a multilingual sequence to sequence model. mBART is a pre-trained by applying the BART to large-scale monolingual corpus on many languages. It contains the detailed analysis on mBART with different ranges and experimental results.

A study on comparison between Bart, t5 and GPT-2

[3] along with experimental conclusions. Its findings show that BART and T5 perform quite better and gives better results than GPT-2 for the chosen task.

Xinyu Xing et al., published the research work for table to text[4]. It proposes the use of STTP model along with the experimental results and comparison with Bart model.

Yang Yang et al., proposed the research work [5] using transformer. It uses the WIKIBIO dataset and proposed a transformer-based model to study several data to text generation tasks.

Tianyu Liu 1 et al., proposed the text narration from table [8] from an entity-centric view. They have used WIKIPERSON dataset contains around 250000, 30000, and 29000 (table, text) pairs in training, validation, and test sets respectively. They have tried decreasing the hallucinated data and increasing faithfulness by evaluating the faithfulness with two entity-centric metrics, both are proven to have good agreement with human perspective. They have also experimented the comparison of transformer and Bart model.

3. BACKGROUND

3.1. Table to Text

A Table is a widely used type of data source on the web, which has a definite structure and contains useful information. Interpreting the meaning of a table and understanding, explaining its content is an important problem in artificial intelligence, with innovative applications like question answering, in search engines and many more. The task of text narration from tables could be used to support many applications, such as conversational agents and various interpretation browsers. As well as the task can be used to generate meaningful sentences for the well-structured tables on the Internet.

3.2. Transformers

Before transformers were introduced, most SOTA (state-of-the-art) NLP systems used to rely on Recurrent Neural Networks (RNN), such as Long Short-Term Memory (LSTM) and Gated recurrent units (GRUs), with the help of added attention mechanisms. Transformers also tends to make use of attention mechanism but, unlike RNN, they do not have a recurrent structure which means that given enough training data, attention mechanisms can match the performance of RNN with attention.

3.2.1. T5

T5 stands for "Text to Text Transfer Transformer". T5 tries to combine all the downstream tasks into a text to text format. T5 is versioned into various type as per sizes:

- t5-small
- t5-base
- t5-large
- t5-3b
- t5-11b.

3.2.2. BART

BART stands for "Bidirectional Auto-Regressive Transformers." BART is a transformer-based Sequence to Sequence model that makes use of corrupted source text. BART can be seen as combination of BERT and GPT2 which tries generalizing Bert due to the bidirectional encoder and GPT2 with the left to right decoder.

3.2.3. mBART

mBART stands for "Multilingual Bidirectional Auto-Regressive Transformers" MBART is a multilingual encoder-decoder sequence to sequence model which is based on transformers primarily used for translation task but not restricted to that. As the model is multilingual it expects the sequences in a different format.

4. ARCHITECTURE

In this paper we have taken transformer-based models which include T5, BART, mBART for carrying out the comparison task. We have provided our explanation and understanding for the same.

4.1. T5

T5 is an encoder-decoder model which is pre-trained on a multitask which is a mixture of both unsupervised as well as supervised tasks and for which each task is converted into a text to text format. T5 works well on a variety of tasks by prepending a different prefix to the input corresponding to each task, e.g., for translation: translate English to German, summarizing text, predicting similarity score among 2 sentences.

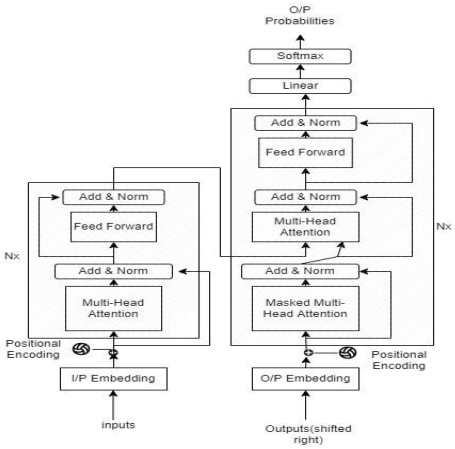

Fig. 1. Architecture of Transformer (T5)

Fig. 1 illustrate the architecture of transformer. The Transformer architecture consists of the Encoder block which is towards the left and the Decoder block which is towards the right.

Encoder: Encoder block consists of a stack of N identical layers. Every layer has a multi-head attention layer.

Decoder: The decoder stack also consists of 6 identical layers. Each decoder layer has 2 multi-head attention layers, followed by a feed forward neural network.

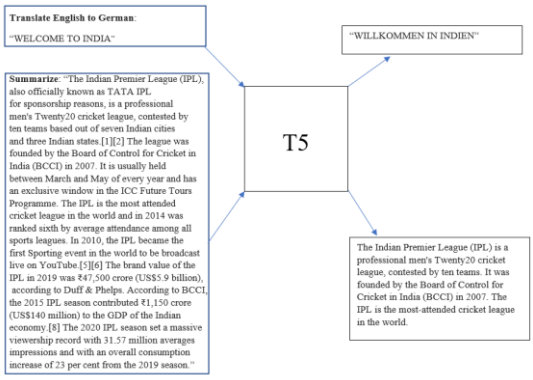

Fig. 2. T5 FLOWCHART

Fig. 2 illustrate the T5 framework. Many tasks can be casted into this framework like language translation, classification task, regression task other sequence to sequence tasks like document summarization for example, summarizing articles from websites, etc.

4.2. BART

BART is a self-supervised Sequence to Sequence auto-encoder model where we send the source data and add some noise to corrupt text, this data we send to the Denoiser which is an encoder decoder model, and then we get the regenerated text at the end. Here Regenerated text gives the feedback back to the Denoiser. i.e., Loss Optimization via Backpropagation. This model works well and gives the best performance when used for Natural Language Generation tasks like translation, summarization but it is also working perfectly for tasks like text classification, Q&A.

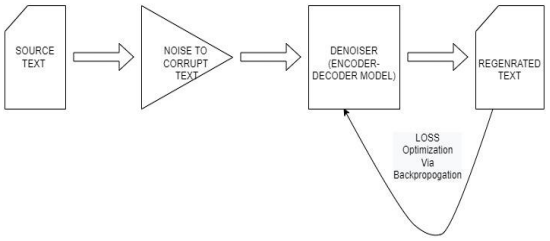

Fig. 3. Architecture of BART

Fig. 3 illustrate the architecture of BART. We have explained this architecture in detail in this paper.

We can say BART is a modified version of Sequence-to-Sequence model made to work as an auto-encoder. Only difference in BART architecture is the use of GELU instead of RELU activation layer. If we compare BART with BERT, BART doesn't make use of a feed-forward network at the top for word prediction while BERT does. BART uses just approximately 10% more parameters compared to equivalent BERT. BART achieves better performance for language generation tasks compared to BERT.

Applications of BART:

We can fine-tune BART achieving better performance for the following tasks:

- Sequence Generation
- Token Classification
- Sequence Classification
- Machine Translation

4.3. mBART

mBART is a sequence to sequence denoising auto-encoder model for pretraining a complete sequence to sequence model by denoising full texts in multiple languages, while earlier approaches have concentrated only on the encoder, decoder, or reconstructing parts of the text. The best part about mBART model is that it learns some structure of the languages during pretraining, and this structure goes beyond linguistic borders and allows intrinsic knowledge transfer between languages. This language-transfer significantly outperform fine-tuning on the target language pair and can be the turning point for some applications.

5. INFERENCE

In this paper we did a comparative survey on three transformer-based models for text generation from tabular data.

- One of the advantages of using T5 model against other models is that it does not provide a label or a span of the inputs as an output to the provided input sentence, but instead it generates the output as a string formatted text.
- BART is extremely flexible and can account for nonlinearities and interactions without overfitting due to the Bayesian priors. In addition, BART's default tuning parameters are effective in many cases.

- In short, we can describe mBART as a multilingual model with encoder-decoder primarily used for translation task but not limited to that. Being multilingual in nature it expects the sequences in a different format.

6. CONCLUSION & FUTURE WORK

In this paper, we have done the survey on three transformer based models for Table to text generation and approaches for the same. As an extension to this research work, we are planning to implement transformer based T5 model as our use case. We will be using the ToTTo dataset which is published by google and train the model on maximum possible epochs and optimize the solution by creating dynamic interface where we can provide our input and predict the meaningful sentence as an output from the model.

REFERENCES

[1] Mike Lewis*, Yinhan Liu*, Naman Goyal*, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer-BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension-BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (aclanthology.org)

[2] Yinhan Liu*, Jiatao Gu*, Naman Goyal*, Xian Li, Sergey Edunov Marjan Ghazvininejad, Mike Lewis, Luke Zettlemoyer-Multilingual Denoising Pre-training for Neural Machine Translation-2001.08210.pdf (arxiv.org)

[3] Auto-regressive Text Generation with Pre-Trained Language Models: An Empirical Study on Question-type Short Text Generation-Anonymous ACL submission-pdf (openreview.net)

[4] Xinyu Xing and Xiaojun Wan Wangxuan Institute of Computer Technology, Peking University The MOE Key Laboratory of Computational Linguistics, Peking University-Structure-Aware Pre-Training for Table-to-Text Generation-Structure-Aware Pre-Training for Table-to-Text Generation (aclanthology.org)

[5] Yang Yang Communication University of China Juan Cao Communication University of China Yujun Wen (wenyujun@cuc.edu.cn) Communication University of China Pengzhou Zhang Communication University of China-Table-to-Text Generation with Accurate Content Copying-https://www.nature.com/articles/s41598-021-00813-6

[6] Google AI Blog: Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer (googleblog.com)

[7] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer-https://openreview.net/pdf?id=5Jcma3gBao

[8] Tianyu Liu 1 *, Xin Zheng 2 3 *, Baobao Chang 1 4, Zhifang Sui 1- Towards Faithfulness in Open Domain Table-to-text Generation from an Entity-centric View- 2102.08585.pdf (arxiv.org)

[9] Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, Dipanjan Das https://arxiv.org/abs/2004.14373

[10] Sha, L., Mou, L., Liu, T., Poupart, P., Li, S., Chang, B., & Sui, Z. Order-planning neural text generation from structured data. *Proc. AAAI.* Vol. 32, No. 1(2018, April)

STUDY OF CONSISTENCY AND PERFORMANCE TRADE-OFF IN CASSANDRA

Kena Vyas and PM Jat

DAIICT, Gandhinagar, Gujarat, India

ABSTRACT

Cassandra is a distributed database with great scalability and performance that can manage massive amounts of data that is not structured. The experiments performed as a part of this paper analyses the Cassandra database by investigating the trade-off between data consistency andperformance. The primary objective is to track the performance for different consistency settings. The setup includes a replicated cluster deployed using VMWare. The paper shows how difference consistency settings affect Cassandra's performance under varying workloads. The results measure values for latency and throughput. Based on the results, regression formula for consistency setting is identified such that delays are minimized, performance is maximized and strong data consistency is guaranteed. One of our primary results is that by coordinating consistency settings for both read and write requests, it is possible to minimize Cassandra delays while still ensuring high data consistency.

KEYWORDS

NoSQL, Cassandra, Consistency, Latency, YCSB, and Performance.

1. INTRODUCTION

Data's relevance has skyrocketed to the point where it is now seen as a precious asset. For any organisation, data is an essential. Every day, massive amounts of data get generated. Data of various formats are seen nowadays in IoT devices such as smartwatches, smart TVs, and home assistants. Every second or minute, data of different kinds gets generated from different devices. As a result, the ability to properly store and retrieve such huge and diverse data is required.

Relational databases have typically been used to store structured data with a high level of consistency. But when it comes to working with unstructured data, they have a number of drawbacks. The rigorous schema constraints of relational databases make it challenging to store massive data, which is typically anticipated to be unstructured or loosely structured. Field lengths are limited in relational databases, which leads to improper handling of unstructured data. Because of the inadequacies of relational databases when it comes to massive data, NoSQL databases have grown in popularity.

NoSQL Databases are non-relational data management systems. It gives a way to save and retrieve data. The data is represented differently than in relational databases, where tabulated relations are used. It does not require a fixed schema. The key advantage of using a NoSQL database is for huge data with dispersed data repositories. Therefore, it's becoming more prevalent in big data and real-time online applications. NoSQL databases have the following features: Flexible schemas, High availability, and Horizontal scaling. NoSQL databases has eventual consistency and hence lacks ACID features.

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 61-77, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121907

1.1. Motivation

The main motivation of this paper is to find optimal setting for Cassandra database such that it provides strong consistency and minimal latency. Understanding this trade-off is crucial for finding a database state that is consistent. The paper examines the trade-offs that NoSQL databases must make between consistency, availability, and latency. It's crucial to understand how different consistency settings affect system latency. There are many NoSQL databases available for use. various industry trends suggest that Apache Cassandra is one of the top three in use today together with MongoDB and HBase [1]. Apache Cassandra is a columnar distributed database that takes database application development forward from the point at which we encounter the limitations of traditional RDBMSs in terms of performance and scalability [2]. Cassandra is a NoSQL distributed database system that is known for managing large amounts of distributed data. It provides high availability without a single point of failure [3].

1.2. Objective

In this paper, the Cassandra database is used to provide a quantitative examination of the fundamental Big Data trade-offs between data consistency and performance. We'd like to provide practical recommendations to developers that use Cassandra as a distributed data storage system, allowing them to forecast Cassandra latency while keeping the required consistency level in mind, and to optimise the consistency settings of operations. A benchmarking approach is developed that optimizes Cassandra's performance that guarantees strong data consistency under the selected workload. A NoSQL database like Cassandra supports database replication in order to maintain availability in the case of event failure or planned maintenance events. Cassandra keeps replicas on several nodes to ensure automatic failover and durability. Depending on the replication mechanism employed, a consistency setting needs be found that maximises performance while minimising latency.

1.3. Outcomes

A benchmarking methodology is created for working with read and write workloads in different proportions. Various workload runs are executed on the deployed cluster and their results are measured. The Cassandra database is monitored for Latency and Throughput values when read and write workloads are executed on it for a varying number of threads. Various combinations of read and write workloads are considered. The outcome of this paper will help the user of the database in identifying a consistency setting that is strong and simultaneously provides sufficient throughput with minimized latency. Two experiments are performed as a part of this work that measured the performance of the Cassandra database for varying read/write workloads, changing threads, and different consistency settings. The first experiment measures the results by separating the read and write workloads. In the second experiment, various proportions of read/write workloads are considered together so that we can get all possible combinations and can measure the results accordingly. From the measured results, regression formulas are generated which can be used for prediction purposes.

1.4. Paper Organization

This paper is organised as follows. In the next section i.e., Section 2 is Cassandra and Consistency where concepts like NoSQL, replication factor and consistency levels are covered, Section 3 talks about Performance Benchmarking with YCSB along with related works, Section 4 is about experimentation, the two experiments performed as a part of this paper are explained in

detail along with their objective, setup, and results. Section 5 concludes the paper with a conclusion.

2. CASSANDRA AND CONSISTENCY

Two Facebook developers, Lakshman and Malik, released Cassandra to the Apache community in 2008. They describe Cassandra as a "distributed storage system for managing very large amounts of structured data spread across many commodity servers while providing highly available service with no single point of failure"[4]. Cassandra is a column-oriented, peer-to-peer NoSQL database that is a distributed and decentralized storage system that is open source. It oversees massive amounts of structured/unstructured/loosely structured data from all around the world. It ensures high availability, which eliminates the possibility of a system failure and provides eventual consistency [5].

Cassandra provides a familiar interface known as Cassandra Query Language (CQL). CQL offers an abstraction layer to the database where implementation specifics are hidden, and native access syntaxes are provided. The data in Cassandra is kept in keyspaces, which are similar to databases in relational database concepts. A column family in the Cassandra database is equivalent to a table in a relational database, and they can be represented as a collection of rows. Rows are formed of columns and their values, which are represented as key-value pairs [6]. The Replication Factor and Strategy can be defined at the time of keyspace creation.

2.1. Cassandra Data Model

The Wikipedia page of Cassandra mentions that the Cassandra data model is “designed for distributed data on a very large scale” [7]. Cassandra runs in main memory and makes asynchronous disc writes on a regular basis. Cassandra comprises ACID properties in order to increase availability and performance. The structure of the Cassandra model is quite different from the relational model.

A Cassandra cluster is a storage unit in the database. It consists of multiple keyspaces. A level of Column families exists beneath the keyspace level. A column family is a logically arranged collection of one or more columns depending on database design. There will be one or more column(s) inside a column family. Within the Cassandra data paradigm, a column is the simplest data structure and is at the lowest level. A column has 3 different attributes namely name, value, and timestamp. The name attribute is used to identify a column. Value attribute stores the actual value related to the name attribute and timestamp is the time when the column is stored, it is mainly used during data replication.

A "row" is similar to a relational database row which is a collection of values linked together. However, there is a difference between the two. The row in the Cassandra model is dynamic and can have a varying number of columns. One of the advantages of Cassandra is the flexibility ofwhat may be stored and the fact that no space is allocated for columns that are not part of the current data set.

2.2. NoSQL

NoSQL is often referred to as "non-SQL" or "non-relational". Eben Hewitt has his own explanation of what NoSQL is all about in his book Cassandra: The Definite Guide [8]. "Comparing NoSQL to relational is basically a shell game," Hewitt argues. Eben Hewitt implies that NoSQL cannot be directly compared to a relational database because it encompasses a wide

range of non-relational database types. Most NoSQL databases provide some level of balance among consistency, availability, partition tolerance, and latency. Although a few databases have made ACID (Atomicity, Consistency, Isolation, Durability) transactions core to their architecture, most NoSQL stores lack these [9].

NoSQL systems can be classified into categories according to their data model. There are four different types of NoSQL databases: Column-oriented, Graph, Document, and Key-value databases. Cassandra, MongoDB, Couchbase, HBase, and Redis are some of the most popular NoSQL databases. Cassandra offers a range of unique features which makes it a good choice for us. Cassandra has no single point of failure because of its peer-to-peer architecture. Scalability is another advantage that Cassandra provides for scaling up or down. It is highly available and fault tolerant because of the data replication it provides. Such benefits provided by Cassandra makes it a great choice.

2.3. Replication and Consistency

Data replication is the process of storing several copies of data in multiple nodes. The replication approach ensures that the same data is available in other nodes if one node fails. Cassandra supports replication in the database to ensure availability in the event of failure or other predefined activity. The process of replicating data from one location to another is known as replication. The replication method for each keyspace determines the nodes where replicas are placed. Cassandra keeps replicas on several nodes to ensure fault tolerance and reliability. The replication factor refers to the total number of replicas in the cluster. A replication factor of one means that each row in the Cassandra cluster has only one copy. At the time of keyspace generation, the Replication Factor can be specified. The replication factor should not be more than the total cluster nodes.

The minimal number of Cassandra nodes that must recognize a read or write operation before it may be declared successful is known as the Cassandra consistency level. Different Edge keyspaces can have different consistency levels allocated to them. When the consistency option is one, it indicates that for a read/write operation to succeed, at least one of the Cassandra nodes in the datacentre must react. Depending on the replication mechanism employed, a consistency setting can be found that maximizes performance while minimizing latency. Cassandra's consistency settings can be set to balance data accuracy and availability. Consistency can be set for a session or for each read or write operation individually.

Hewitt explains three different levels of consistency in his book about Cassandra [8].

Strong Consistency - All data received from the database must be the most current information available. A mechanism for a global timer will be necessary to put a time stamp on the data and actions done to the system. String consistency is essential in areas like financial institutions, e-commerce websites, etc at all times. Strict consistency ensures that the data returned will be consistent and valid. However, one disadvantage is that performance will be degraded because the system will have to verify data with multiple nodes before returning the results.

Most NoSQL systems use the concept of R, W, N where R is the number of nodes from which data is read, W is the number of nodes where data is written and N is the replication factor and when we have R+W>N then, strong consistency can be achieved.

Eventual Consistency - Context here is we have partitioned and replicated data. Any update to such a database needs to be propagated to all replicas. Any read request for a data item following its write should get the last updated value irrespective of a replica from which value is being read.

Eventual consistency is weaker than strong consistency. Whenever eventual consistency is used and a request for data is made, then it may provide data which is one version older than the current one. However, eventual consistency makes sure that the most recent data is available to the user after a certain period of time.

When we make a change to a distributed database, eventual consistency ensures that the change is mirrored across all nodes that store the data, ensuring that we get the same response every time query is made. Eventual consistency offers low latency. Because changes take time to reach replicas throughout a database cluster, early results of eventual consistency data queries may not have the most current updates. The database system guarantees that if no new updates are made to the object, eventually all accesses will return the last updated value [11].

Weak Consistency - Another type of consistency is weak consistency which gives no guarantee that all nodes will have same data at any given time. From time to time, updates are exchanged among nodes such that all nodes have updated data. After a certain period of time, the data in the nodes will reach a consistent state.

2.3.1. Consistency Level (CL) on Write

The number of replica nodes that must acknowledge before the coordinator can report back to the client is determined by the consistency level for write operations. The number of nodes that acknowledge (for a given consistency level) and the number of nodes that store replicas (for a certain replication factor) are almost always different. For e.g., even when only one replica node recognizes a successful write operation with consistency level ONE and RF = 3, Cassandra concurrently replicates the data to two other nodes in the background. Below are write consistency levels that are used in the paper:

Table 1. Consistency levels for Write operation

Level	Description
ONE	It only requires one replica node to recognise it. Because only one copy needs to acknowledge the write operation, it is faster.
QUORUM	It requires 51 percent or a majority of replica nodes across all datacentres to acknowledge it.
ALL	It requires confirmation from all replica nodes. Because all replica nodes must acknowledge the write operation, it is the slowest. Furthermore, if one of the replica nodes fails during the write operation, the write operation will fail, and availability will degrade. As a result, it's advisable not to use this option in production deployment.

2.3.2. Consistency level (CL) on Read

The consistency level for read operations determines how many replica nodes must respond with the most recent consistent data before the coordinator can deliver the data back to the client successfully. Below are read consistency levels that are used in the paper:

Table 2. Consistency levels for Read operation

Level	Description
ONE	Only one replica node returns the data at consistency level ONE. In this scenario, data retrieval is the quickest.
QUORUM	It signifies that 51 percent of replica nodes in all datacentres have responded. The data is then returned to the client via the coordinator.
ALL	It requires confirmation from all replica nodes. The read operation is the slowest in this situation since all replica nodes must acknowledge.

Quorum Calculation - The QUORUM level works with the number of quorum nodes. The following is how a quorum is computed and then rounded down to a whole number:

quorum = floor((sum_of_replication_factors / 2) + 1)

In a cluster of 3 nodes, a quorum is 2 nodes. In a cluster of 6 nodes, a quorum is 4 nodes.

There are mainly 2 ways for setting consistency in a cluster:

1st Way

To set the consistency level for all queries in the current cqlsh session, use CONSISTENCY in cqlsh.

Syntax:

CONSISTENCY [Level]

Example: CONSISTENCY ONE

2nd Way

For setting the consistency level individually for each operation, the consistency can be set in the command line argument (CLI).

-p cassandra.readconsistencylevel=[Level] -p cassandra.writeconsistencylevel=[Level]

Example:-p cassandra.readconsistencylevel=[ONE] -p cassandra.writeconsistencylevel=[ONE]

2.4. CAP Theorem

Being ACID compliance is one of the strengths of relational databases. However, it is hard to achieve serializability in distributed and replicated environment and may leads to delays that are beyond acceptable limits. NoSQL systems have compromised ACID properties in order to achieve better performance when working with large data sets. Because of that, NoSQL systems need to follow some other set of rules that fit the NoSQL criteria. A scientist called Eric Brewer established a theorem called Brewer's CAP theorem. Brewer et.al. [6] realizes this and presents CAP theorem which states that any distributed data store can only provide two of the three (i.e., consistency, availability and partition tolerance) guarantees.

Brewer's CAP theorem categorizes database systems according to their capabilities. The CAP theorem was created to put the different NoSQL solutions together because the bulk of them was obliged to compromise the ACID guarantee in order to focus on more critical aspects for their specific needs. CAP is an acronym that stands for [13]:

- Consistency - At the same moment, all connected nodes see the same data.
- Availability - Even if a request is unsuccessful, it is guaranteed that a response will be received if it is delivered to the database.
- Partition tolerance - There is no single point of failure in the system. If one node fails, the data can still be accessed by another node, and the system will continue to function normally.

Hewitt states in his book about Cassandra that "Brewer's theorem is that in any given system, you can strongly support only two of the three" [8]. The definition says that a database system cannot provide all three properties at the same time. When a system is spread across numerous nodes, it cannot be 100% consistent and available at any given time. When the state of a database is changed (new data added or data updated) due to various reasons it will take a few milliseconds or seconds to propagate the changes to other nodes because of which the system is called eventually consistent.

3. PERFORMANCE BENCHMARKING WITH YCSB

YCSB is an abbreviation for Yahoo cloud serving benchmark. YCSB is a program suite for computing the execution of NoSQL systems. It is used to evaluate/compare the working of different NoSQL systems based on several parameters. YCSB Benchmark is a collection of workloads. It can collect the performance metrics of a system under a specific, pre-defined workload. It makes it easier to compare the performance of the next generation of data serving systems [8]. The YCSB framework is a standard benchmark for evaluating the operation of NoSQL databases such as Redis, MongoDB, HBase, Cassandra, and others. The YCSB framework is made up of a client that generates a workload and a set of basic predefined workloads that cover various aspects of performance. YCSB provides five different workloads. Each workload is a unique combination of read/write queries and data sizes. The operations in the workload are Insert, Update, Read and Scan. The vital feature of the YCSB framework is its extensibility. The workload generating client is extensible which supports the benchmarking of different databases. The workloads are [8]:

Table 3. YCSB default workloads

Workload	Read Weightage	Update Weightage	Insert Weightage	Scan Weightage
A-Update Heavy	50%	50%	0%	0%
B-Read Mostly	95%	5%	0%	0%
C-Read Only	100%	0%	0%	0%
D-Read Latest	95%	0%	5%	0%
E-Short Ranges	0%	0%	5%	95%

3.1. Related Works

Relational databases have been the choice for majority of systems due to their rich set of features. However, they are not suitable for handling huge data. NoSQL databases have gained popularity as they efficiently work with big data [13]. The paper "NoSQL Databases: MongoDB vs Cassandra" talks mainly about NoSQL databases along with their types and also briefs about CAP/ACID theorems. YCSB benchmark is used for the experimentation. The performance parameter which signifies the execution time is taken into consideration for comparing the two databases i.e., MongoDB and Cassandra. In the experiments, six different YCSB workloads are used for testing both the databases. The results indicate that as the data size increased, MongoDB started to reduce performance [13]. However, Cassandra became faster as data size increased.

Yahoo cloud serving benchmark framework is presented in the paper titled "Benchmarking Cloud Serving Systems with YCSB", that facilitates performance comparisons of data serving systems. Four widely used databases like Cassandra, HBase, Yahoo!'s PNUTS, and a simple sharded MySQL implementation are used in the paper for benchmarking. The papers use core workload of YCSB for measuring performance and scalability of the databases. The results show that Cassandra and HBase have higher read latency on a read heavy workload and lower update latency on write heavy workload [14]. Along with that, Cassandra and PNUTS showed better scalability. The paper also explains in details the core workloads provided by YCSB. The paper also talks about the workload generating client that comes with YCSB using which new workloads can be defined.

Our paper focuses on the consistency and latency trade-off aspect mainly. To identify the best setting of threads and read/write workloads such that strong consistency can be obtained. The paper "Consistency Trade-offs in Modern Distributed Database System Design" explains in detail the consistency/latency trade-off. The paper gives a good introduction about CAP theorem. According to CAP, the system must choose between high availability and consistency [10].

The paper "Interplaying Cassandra NoSQL Consistency and Performance: A Benchmarking Approach" puts light on the trade-off between data consistency and performance. The main aim of the paper is to allow the developers to predict the delay in Cassandra by considering the required consistency level. The paper proposes a benchmarking approach for optimising performance of Cassandra such that strong consistency is ensured [12]. In the paper, a Cassandra database is deployed and executed in a real production environment. YCSB benchmark is modified to execute application specific queries. The Cassandra database is benchmarked for various conditions such as different workloads, different consistency settings, etc. After that, regression functions are generated that interpolate the average read/write latency with precision. The paper identifies optimal consistency setting by using regression functions which will help the developers to find out settings such that required consistency level is obtained.

Our presented work shows how different consistency setting affect the Cassandra response time and throughput. Because Cassandra provides the feature of tuneable consistency, it is possible to achieve strong consistency by finding optimal settings. By monitoring various parameters of Cassandra database while different combinations of workload, threads and consistency settings are executed, we try to find certain consistency setting that provides the minimum latency.

4. EXPERIMENTATION

4.1. Experiment Objective

To describe a methodology for benchmarking the performance of Cassandra. To extract experimental results, show how different consistency settings influence the latency and throughput. To understand the relationship between the parameters and generate a regression equation for predicting the parameters. The experiment extracts results based on two scenarios:

1. When the read and write operations are executed individually.
2. When mixed read and write workload are executed.

To narrow down the available options for consistency setting based on results obtained. The objective also includes generating a data set for finding multiple regression equations which can be used to perform predictive analysis and to find an optimal setting such that strong data consistency is guaranteed.

4.2. Cassandra Cluster Setup

A Cassandra cluster of 3 nodes with different IP addresses is deployed on VMware. All the nodes are connected in a cluster by installing Cassandra in all of them and configuring them. A replication factor of 3 is configured for ALL consistency to be applied. The data in the nodes is 3-replicated which means a row in a table has 3 copies in the cluster. The VMware virtual machine uses CentOS operating system that is based on Linux. YCSB benchmark is used in order to evaluate the performance of databases under different workloads. The YCSB Client is a Java program that generates data for database loading and runs the loaded workloads.

Three nodes with IPs: 192.168.29.143, 192.168.29.144, and 192.168.29.145 are deployed in a single cluster such that they are connected and Cassandra is installed on each.

4.3. Results

4.3.1. Experiment 1

In the experiment 1 where the performance of Cassandra is measured by considering the read and write workload individually, the configuration is made as follows. A Cassandra cluster of 3 nodes is deployed on the VMware. In our study, the focus is on examining the dynamic features of Cassandra's performance in various consistency settings. We investigate how the current workload affects database latency and throughput. The following configuration is made for experiment 1.

- A replication factor of 3 is configured.
- Nodes have a Keyspace YCSB and table USERTABLE for experimentation purposes.
- YCSB workload c [read] and workload a [write] parameterized to execute only write operations are used.
- 25,000 records are used for loading and execution
- The results are calculated with
    - a Varying number of threads from 10 to 1000.
    - 3 consistency settings: ONE, QUORUM, and ALL.
- Latency and Throughput for all the combinations are measured for further analyses.
- Regression equations

Read Write Latencies and Throughput Measurements

The tables below show the results of the Cassandra performance benchmarking. The average latency and throughput for read and write requests are shown. For each request, the results are calculated using 25000 records. We may use a mix of average delay and throughput to look at how average read and write delays are affected by the current workload.

Table 4. Cassandra READ latency statistics

Threads	Average latency, us			Throughput,ops/s		
	ONE	QUORUM	ALL	ONE	QUORUM	ALL
10	22705	23645	44586	400	386	174
50	58040	59129	70685	736	724	605
100	90431	94485	108955	888	872	762
200	150734	154128	167620	1084	1025	938
300	199428	200451	219143	1140	1108	1002
400	251888	253844	272117	1150	1113	1103
500	267875	298233	350792	1203	1166	1049
600	366471	376034	448648	1197	1130	940
700	426929	396847	484428	1186	1235	811
800	481751	500140	571021	1154	1161	946
900	505123	532567	580480	1193	1184	1075
1000	596130	611603	631444	1140	1120	1107

Table 5. Cassandra WRITE latency statistics

Threads	Average latency, us			Throughput,ops/s		
	ONE	QUORUM	ALL	ONE	QUORUM	ALL
10	31100	32908	45034	300	275	196
50	55668	64989	68692	750	666	593
100	78148	75230	91153	1008	1058	884
200	138283	162299	84844	974	983	768
300	172168	192620	218842	1203	1137	988
400	235078	203891	271080	1156	1274	993
500	280063	356982	392521	1208	1192	952
600	308273	312009	359718	1371	1293	1157
700	356290	362819	375992	1372	1322	1218
800	409974	429026	503939	1220	1250	1096
900	449739	466700	504209	1419	1378	1128
1000	531026	544698	620622	1437	1390	1262

Latency graphs

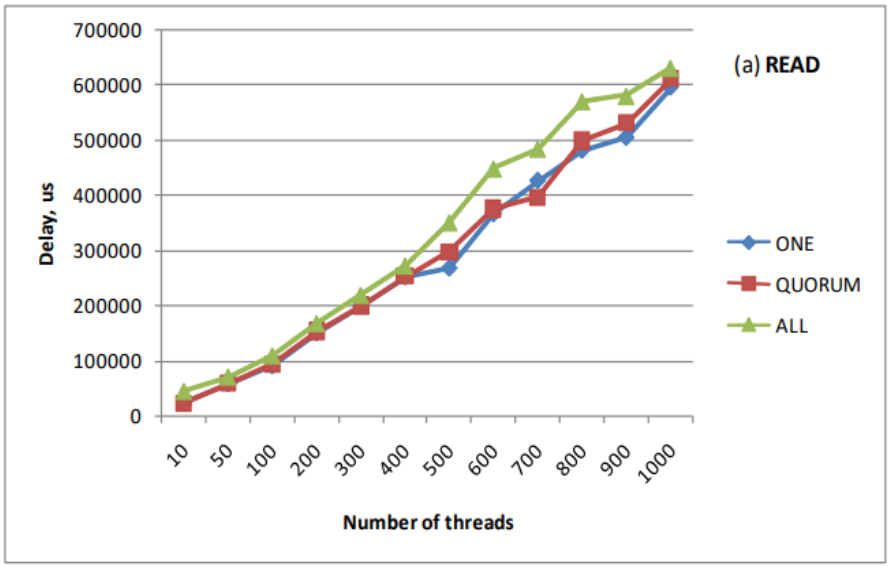

Figure 1. Average Cassandra delay depending on the current workload: reads

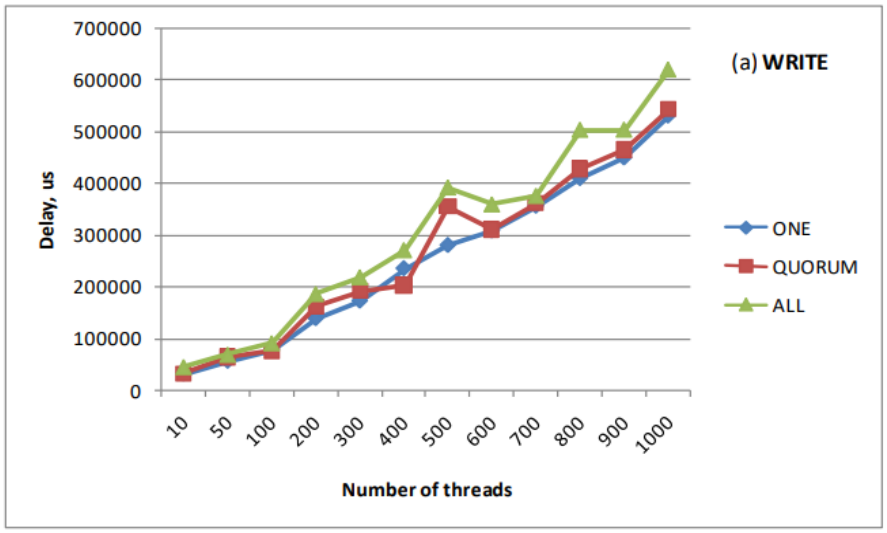

Figure 2. Average Cassandra delay depending on the current workload: writes

Throughput graphs

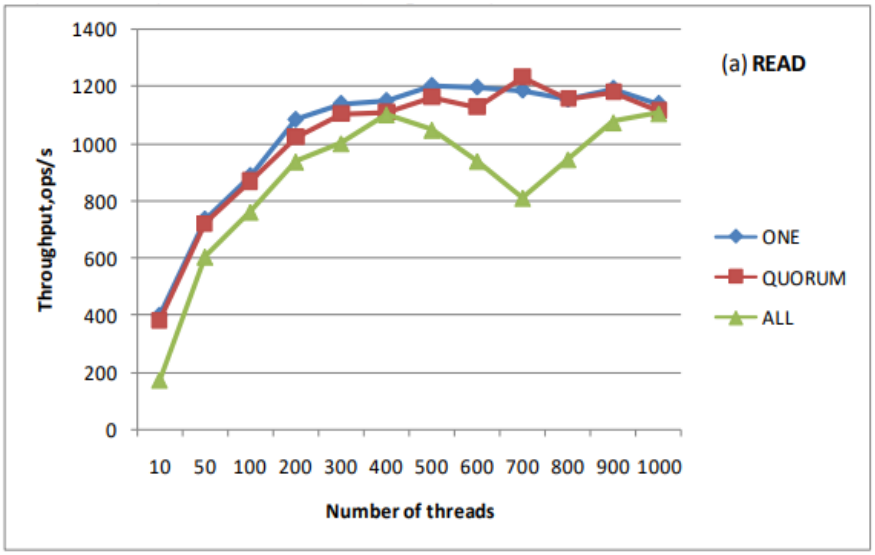

Figure 3. Cassandra Throughput depending on the current workload: reads

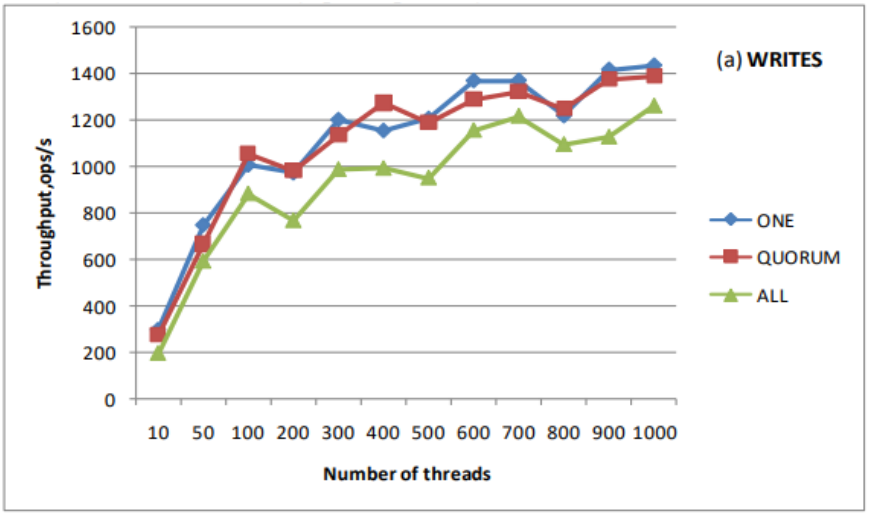

Figure 4. Cassandra Throughput depending on the current workload: writes

Experimental Results

Cassandra reads with the ONE consistency level achieve a maximum throughput of 1203 requests per second, as shown in Table 4. It varies between 1240 and 110 requests per second for the QUORUM and ALL consistency levels. For writes, it is 1437 for ONE consistency level and it fluctuates around 1400 and 1250 for QUORUM and ALL consistency setting respectively.

The graphs in Figure 1 and 2 show the delay experienced for read and write operations individually. The X-axis represents the number of threads running and the Y-axis represents the delay in microseconds. The three lines denote the average latency for ONE, QUORUM, and ALL consistency settings. The average latency for ALL consistency settings is the highest compared

with ONE and QUORUM. However, as shown in Figure 3 and 4, the throughput for ALL consistency settings is the lowest for both read and write operations.

4.3.2. Experiment 2

As already discussed, if the overall number of written and read replicas is more than the factor of replication, the Cassandra database can ensure the maximum data consistency model. This means for a 3-replicated system there are six different read/write consistency settings that can be used to provide high data consistency. They are

- 1R-3W: One read-All write
- 2R-2W: Quorum read-Quorum write
- 3R-1W: All read-One write
- 2R-3W: Quorum read-All write
- 3R-2W: All read-Quorum write
- 3R-3W: All read-All write

Besides, the two settings: 1R-3W and 2R-1W provide the 66.6% of consistency. Finally, the 1R-1W setting can guarantee only the 33.3% of consistency [12]. Whenever a smaller number of replicas are invoked read/write operations in Cassandra executes faster. Hence, in real life experiments, the following consistency should be chosen: 1R-3W, 2R-2W and 3R-1W. All the three combinations follow the rule:

$$(nodes\_written + nodes\_read) > replication\_factor$$

As all the three consistency settings provide strong consistency, a system developer may want to know the performance of those settings for different read/write load proportions and different read/write consistency settings.

Read/Write Latency measurements

For this experiment, 5 different read/write load proportions are taken into consideration: Read/Write-10/90%, Read/Write-30/70%, Read/Write-50/50%, Read/Write70/30% and Read/Write-90/10%. For each of these 5 proportions, read and write latency are measured for 3 consistency settings such as 1) ‘Read ONE – Write ALL’ (1R-3W) 2) ‘Read QUORUM – Write QUORUM’ (2R-2W) 3) ‘Read ALL – Write ONE’ (3R-1W). Table 6 to 10 shows the measured results. The consistency setting that fetches the lowest latencies is highlighted. The tables below show some estimations of Cassandra latency for various configurations, ensuring good consistency in a mixed read/write workload.

Table 6. READ and WRITE latency for ratio: 10/90%

Threads	Read/write	10/90%				
	1R-3W		2R-2W		3R-1W	
	read latency	write latency	read latency	write latency	read latency	write latency
10	30331	31031	191277	50288	42065	31306
50	67304	80219	175507	84587	80445	67491
100	108690	113468	231514	121740	144463	101449
200	149067	153665	308020	198210	187619	154400
300	245266	258348	400140	282352	239309	213881
400	290760	360503	415130	319530	337565	258517
500	372124	481837	455951	366129	431158	323984
600	394041	499268	495030	446155	456912	402001
700	503725	523590	569443	515119	477099	443565
800	612811	669392	760502	593398	637259	611418
900	634603	690514	774205	667434	675742	612999
1000	693981	710275	829634	781465	804184	765775

Table 7. READ and WRITE latency for ratio: 30/70%

Threads	Read/write	30/70%				
	1R-3W		2R-2W		3R-1W	
	read latency	write latency	read latency	write latency	read latency	write latency
10	32741	45848	41353	30100	42690	39421
50	75024	95098	62709	60774	72801	69542
100	111084	138188	148051	121493	246692	224263
200	187287	232319	352954	296963	319321	249553
300	233918	290167	372207	346133	455106	317409
400	256812	301175	404357	370283	517383	363734
500	357096	363032	512833	437759	605245	410471
600	416477	481412	553125	522874	651345	619337
700	533552	597259	638761	603694	677148	618079
800	628928	658895	712611	682823	732824	641235
900	656188	698454	795098	732968	741928	706249
1000	703524	739317	796232	748007	769981	732211

Table 8. READ and WRITE latency for ratio: 50/50%

Threads	Read/write	50/50%				
	1R-3W		2R-2W		3R-1W	
	read latency	write latency	read latency	write latency	read latency	write latency
10	33895	56687	36924	34661	46568	39165
50	73229	95830	89636	87565	99267	89607
100	114747	156710	142052	134773	156400	141365
200	204456	249069	240831	219370	308926	252583
300	253925	297516	328745	299773	353231	301648
400	291299	381609	387702	334432	437584	363444
500	353225	434046	496390	397122	543788	458692
600	427640	501726	534005	461660	550183	472276
700	648993	740161	787573	658604	759187	668920
800	756730	842811	887630	749299	895984	767130
900	783249	879269	892225	792315	907415	857217
1000	853586	964892	970729	885770	973234	871643

Table 9. READ and WRITE latency for ratio: 70/30%

Threads	Read/write	70/30%				
	1R-3W		2R-2W		3R-1W	
	read latency	write latency	read latency	write latency	read latency	write latency
10	24600	31393	30218	29344	39219	32329
50	68325	81604	35040	71423	86173	73635
100	102237	127953	117078	113457	134405	111624
200	180272	235217	199588	184491	230216	184480
300	253253	329891	257724	328285	569271	518260
400	280630	342402	331645	304551	532370	411360
500	406752	529736	408194	360940	763241	486193
600	470798	585201	466305	436402	714981	521906
700	463361	551020	532713	538365	908323	727998
800	499385	589462	669009	629421	819209	669959
900	544131	638540	719122	671458	942113	788877
1000	612822	701998	918065	845165	968210	812331

Table 10. READ and WRITE latency for ratio: 90/10%

Threads	Read/write	90/10%				
	1R-3W		2R-2W		3R-1W	
	read latency	write latency	read latency	write latency	read latency	write latency
10	21729	27576	40283	38839	45563	33537
50	58921	79251	98406	93869	122738	97894
100	93010	138732	151495	142770	167254	129479
200	195954	246562	243233	228524	278509	213629
300	207471	309810	330112	311948	390988	299755
400	286479	366534	408961	300051	381964	538683
500	342750	400440	475445	445930	738210	620529
600	617369	732549	627722	581649	792640	659901
700	644989	726347	675910	632609	811068	690821
800	760282	871117	750107	715626	833562	713990
900	807910	895528	849403	786800	876981	844081
1000	869993	953896	911467	888315	921107	753156

Experimental Results

The 1R-3W configuration delivers the lowest consistency for threads till 200 when the read load proportion is less than 30%. For threads from 200, the 3R-1W setting shows optimal latency among others. When the read load proportion increases, it can be observed that, regardless of the current workload, the 1R-3W option delivers the best latency readings when compared to others. For a read and write proportion of 90/10 %, the 2R-2W setting shows the lowest consistency for a greater number of threads. As the number of requests per second and the fraction of read requests increases, the 2R-2W and specifically the 3R-1W arrangements becomes extremely wasteful. When the percentage of read requests is around 10%, the 3R-1W design still provides the shortest delay in high write-heavy workloads.

4.4. Correlational Analysis

To generalize our results, a multiple regression equation is generated such that it identifies the optimal write consistency factor for the given workload. Syntax of multiple regression equation:

$$Y = \text{Constant } C0 + C1*(X1) + C2*(X2) + C3*(X3) + C4*(X4) \quad (1)$$

The dependent variable Y is the write consistency measure needed to provide strong consistency. There are 4 independent variables: X1-read latency, X2-write latency, X3-threads, and X4-proportion of write workload. To make all of the parameters on the same scale, they are compressed. The following multiple regression formula is created based on the 200 records measured in our experiment:

Table 11. Multiple regression equation static

	Coefficients	*Standard Error*	*P-value*
Intercept	**0.5173**	0.0653	2.57E-13
X1	**-3.876**	0.2967	1.166E-27
X2	**3.5528**	0.4448	1.777E-13
X3	**0.3473**	0.2889	0.231
X4	**0.0739**	0.0853	0.3872

$$Y=0.5173-3.876*X1+3.5528*X2+0.3473*X3+0.0739*X4 \quad (2)$$

Multiple Regression for Read Latency

Table 12. Multiple regression equation for read latency

	Coefficients	*Standard Error*	*P-value*
Intercept	**0.1598**	0.016	6.39E-19
x1	**-0.1257**	0.0142	7.73E-16
x2	**0.8071**	0.0177	2.7E-99
X3	**-0.0708**	0.0204	0.0007

$$Y=0.1598-0.1257*X1+0.8071*X2-0.0708*X3 \quad (3)$$

Here the parameter Y is the read latency measured for various read and write combinations.

Multiple Regression for Write Latency

Table 13. Multiple regression equation for write latency

	Coefficients	*Standard Error*	*P-value*
Intercept	**0.1**	0.0128	5.396E-13
x1	**0.0018**	0.0114	0.8721
X2	**0.7827**	0.0142	7.98E-113
X3	**-0.0981**	0.0164	1.227E-08

$$Y=0.1+0.0018*X1+0.7827*X2-0.0981*X3 \quad (4)$$

Here the parameter Y is the write latency measured for various read and write combinations

5. CONCLUSIONS

To measure Cassandra's latency and performance, we used benchmarking approach. The benchmarking is performed to assess system performance in order to establish how well the system can handle a mixed workload when different consistency settings are employed.

Our research focuses on the relationship between multiple settings for consistency and the performance of the Cassandra column-oriented database. The findings suggest that consistency settings have a considerable impact on Cassandra's response time and throughput, which must be taken into account during system development and monitoring. The Cassandra database gives programmers the ability to fine-tune the consistency setting for each read and write operation request. Software developers can assure strong consistency for their setup by managing the consistency setting by ensuring that the sum of nodes written to and read from is more than the replication factor. In our research, the aim is to choose optimal consistency setting such that strong consistency is provided along with lower latency for our experiment-specific setup.

REFERENCES

[1] Github: Benchmarking Cassandra and other NoSQL databases with YCSB. https://github. com/cloudius-systems/osv/wiki/Benchmarking-Cassandra-and-other-NoSQL-databaseswith-YCSB.
[2] Mishra, V. (2014), Beginning apache Cassandra development. Apress [E-book].
[3] P. Bagade, A. Chandra and A. B. Dhende, "Designing performance monitoring tool for NoSQL Cassandra distributed database," International Conference on Education and e-Learning Innovations, 2012, pp. 1-5, doi: 10.1109/ICEELI.2012.6360579. Eben Hewitt. Cassandra: The Definitive Guide. O'Reilly Media, Inc., 1 edition, 2010.
[4] Datamodel - cassandra wiki. http://wiki.apache.org/cassandra/DataModel.
[5] Daniel Bartholomew. Sql vs. nosql. Linux J., 2010.
[6] Lourenço, J.R., Abramova, V., Vieira, M., Cabral, B., Bernardino, J. (2015). NoSQL Databases: A Software Engineering Perspective. In: Rocha, A., Correia, A., Costanzo, S., Reis, L. (eds) New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing, vol 353. Springer, Cham. https://doi.org/10.1007/978-3-319-16486-1_73.
[7] Abramova, Veronika & Bernardino, Jorge & Furtado, Pedro. (2014). Evaluating Cassandra Scalability with YCSB. 8645. 199-207. 10.1007/978-3-319-10085-2_18.
[8] Eben Hewitt. Cassandra: The Definitive Guide. O'Reilly Media, Inc., 1 edition, 2010.
[9] Pritchett, Dan. (2008). Base an acid alternative. ACM Queue. 6. 48-55. 10.1145/1394127.1394128.
[10] D. Abadi, "Consistency Tradeoffs in Modern Distributed Database System Design: CAP is Only Part of the Story," in Computer, vol. 45, no. 2, pp. 37-42, Feb. 2012, doi: 10.1109/MC.2012.33.
[11] Lakshman, Avinash & Malik, Prashant. (2010). Cassandra — A Decentralized Structured Storage System. Operating Systems Review. 44. 35-40. 10.1145/1773912.1773922.
[12] Gorbenko, A and Romanovsky, A and Tarasyuk, O (2020) Interplaying Cassandra NoSQL Consistency and Performance: A Benchmarking Approach. Dependable Computing - EDCC 2020 Workshops. EDCC 2020. Communications in Computer and Information Science., 1279. pp. 168-184. ISSN 1865-0929 DOI: https://doi.org/10.1007/978-3-030-58462-7_14.
[13] Abramova, Veronika & Bernardino, Jorge. (2013). NoSQL databases: MongoDB vs cassandra. Proceedings of the International C* Conference on Computer Science and Software Engineering. 14-22. 10.1145/2494444.2494447.
[14] Cooper, Brian & Silberstein, Adam & Tam, Erwin & Ramakrishnan, Raghu & Sears, Russell. (2010). Benchmarking cloud serving systems with YCSB. Proceedings of the 1st ACM Symposium on Cloud Computing, SoCC '10. 143-154. 10.1145/1807128.1807152.

APPLICATION OF BAYESIAN OPTIMIZATION AND STACKING INTEGRATION IN PERSONAL CREDIT DELINQUENCY PREDICTION

Jicong Yang and Hua Yin

Guangdong University of Finance and Economics, China

ABSTRACT

The national concept of consumption has changed to excessive consumption, and overdue debts have also increased. The surge of non-performing loans will not only lead to the liquidity difficulties of banks, but also lead to financial risks. Accurate prediction of personal credit overdue is one of the key issues to control financial risks. Traditional machine learning methods build classification models according to the characteristics of credit users, while ensemble learning can ensure high accuracy and prevent model overfitting, which is the mainstream of current application research. The Stacking method can fully combine the advantages of the base model and improve the model performance. The base model and hyperparameter selection have great influence on the prediction accuracy. Therefore, parameter selection according to the studied problem is the core of application. In this paper, the Stacking method is used to integrate multiple single models for credit user overdue prediction, and the parameters of the base model are optimized. The improved Bayesian optimization method is used to select appropriate parameter combinations to improve the model performance.

KEYWORDS

Credit overdue forecast, Stacking integrated learning, Bayesian optimization.

1. INTRODUCTION

With the change in consumer attitudes, the amount of consumer loans to our residents has also grown and the outstanding debt has increased. The proliferation of non-performing loans not only brings the problem of capital turnover difficulties to banks, but also constrains their development and may lead to financial risks, which in turn adversely affects domestic financial development; therefore, accurate prediction of personal credit overdue prediction is a key issue in controlling financial risks.

In this paper, the prediction of personal credit overdue is modeled as a classification problem. Through the personal and loan characteristics of previous credit users and overdue categories, a learning model is established to predict whether personal credit is overdue. The traditional classification algorithm for constructing a single model has the problems of uncertainty and weak generalization. The ensemble learning method integrates diversified weak classifier results to ensure high accuracy while preventing model overfitting [1]. In this paper, XGBoost, random forest and GBDT are used to construct the base learner, and Stacking method is used to integrate. However, the base learner usually needs to set hyperparameters, and the selection and setting of hyperparameters have a great impact on the prediction accuracy. Based on the above problems, this paper improved the Bayesian optimization algorithm and constructed an adaptive balance factor to improve the acquisition function, so that it could dynamically overcome the problem

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 79-92, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121908

that the Bayesian optimization algorithm would fall into the local optimum, optimize the hyperparameters of the base learners of random Forest, GBDT and XGBoost, and construct the optimization Stacking model. The overdue prediction is made based on the real customer data of UnionPay to verify the effect of the model.

Credit risk prediction has been an issue of importance to the financial industry, and in the past studies, researchers have been using various methods to construct credit risk models, and the specific work is as follows.

Wiginton[2] first proposed the use of logistic regression in corporate credit risk management problem and through experimental results it was concluded that logistic regression model has good prediction results in corporate credit risk management problem. Shin et al[3] selected the bankruptcy dataset of Korean listed companies to use SVM to predict the risk of corporate bankruptcy, and the analysis of the results obtained that SVM works better than MDA, Logit and NNs.Chen et al.[4] designed the XGBoost model with improved gradient boosting tree, second order Taylor expansion and also added regularization term to make the performance of the model improved significantly. After the introduction of XGBoost model, a large number of scholars started to apply XGBoost model to the field of risk control. Huang YP et al.[5] used XGBoost model with financial statements of listed companies in Taiwan as the research dataset. The analysis of the results concluded that XGBoost predicted the best results. chang YC et al.[6] used XGBoost models to predict credit risk problems and the results showed that XGBoost models have better results compared with logistic regression and SVM models.

In summary, from the traditional discriminant analysis method to the integrated learning XGBoost method, these models show good results in risk prediction. However, compared with the traditional method and machine learning method, the integrated learning method shows better prediction effect. This paper selects the base learner suitable for the problem studied in this paper to build the model based on Stacking method and referring to the studies of scholars. However, there is a very important factor in the construction of the model: the parameters of the model. Different parameter choices have different applicability to the problem. In view of this problem, this paper does further research.

2. THEORY

2.1. XG Boost

Based on gradient lifting tree algorithm, XGBoost algorithm adds regularization term to the objective function, which can reduce the complexity of the model and avoid overfitting[7]. Its objective function is shown in Equation (1) and Equation (2).

$$\text{obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1} \Omega(f_k) \tag{1}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\omega^2 \tag{2}$$

Where $\hat{y}_i$ is the predicted value, y_i is the true value, $\Omega(f_k)$ is the regular term, f_k is the decision tree, T represents the number of leaf nodes, ω represents the proportion of leaf nodes, γ controls the number of leaf nodes, and λ controls the proportion of leaf nodes.

XGBoost algorithm performs iterative operation and second-order Taylor expansion in the process of solving the objective function, as shown in formula (3) which improves the solving speed and the training speed of the model.

$$obj^{(t)} = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_i) \tag{3}$$

Where g_i and h_i are the first and second derivatives of the loss function, respectively.

$$\begin{aligned} g_i &= \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \\ h_i &= \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \end{aligned} \tag{4}$$

2.2. Random Fores

Random forest is a kind of Bagging method, and decision tree model is used as the base model. The resampling method is used to select multiple sample sets with the same sample size as the given sample each time from the given sample and construct the decision tree based on it . In general, a decision tree divides nodes by selecting a feature from the set of features that can make the model result shift to the best direction. Random forest algorithm adopts the method of random feature selection. Specifically, when building each decision tree, firstly, a subset set containing M ($m \leq m$) features is selected from the feature set to which the node belongs, and the optimal features in this subset are divided. And $n = \log_2 N$ is a random parameter[8].

2.3. GBDT

GBDT is one of Boosting methods. GBDT mainly generates new decision trees, and takes the residuals of the results obtained from the decision trees in this stage as the input of the new decision trees in the next stage, and continues to iterate until the end of the iteration, the cumulative sum of the results of each decision tree is the result of the studied problem. At each iteration, the current decision tree needs to learn the prediction results and residuals of all decision trees in the previous iteration, and build the decision tree with the strategy to reduce the residuals in the subsequent iteration. Its advantage lies in the simple structure of GBDT, has a strong interpretability, the disadvantage is that there is no way to predict the development trend of a problem, that is, only in the scope of the prescribed prediction, can not exceed[9].

2.4. Stacking

The Stacking model fusion method selects multiple basic models and then combines the selected multiple models by specific methods. Because of the differences among models, the purpose of model fusion is to reflect the advantages of different models and make these weak models form strong models by certain methods. However, before adopting the method of model fusion, two criteria of model fusion should be followed. Firstly, the performance of the fused base learners should not be too different, and secondly, there should be discrimination between the learners. Only in this way can model fusion be adopted.

Figure 1 shows the algorithm flow. First, the given data set is divided into five parts, four of which are used for training and the other one is used for testing. Each time, the current training results are taken as the training set of the next layer model. It is also necessary to predict the test

set, take the arithmetic average of the results, and send them to the next layer for prediction. Then, the training results of the first-layer model are taken as the training set of the second-layer model, and the prediction results are taken as the test set of the second-layer model, and all of them are sent to the second layer for training and testing[10] .

According to the Stacking fusion criterion, the base model of the first layer fusion should have good performance, and the performance difference between the models should not be too big. From this perspective, XGBoost model and random forest model were selected as the base model of the first layer, and GBDT model was selected as the Stacking model of the second layer. The structure is shown in Figure 2.

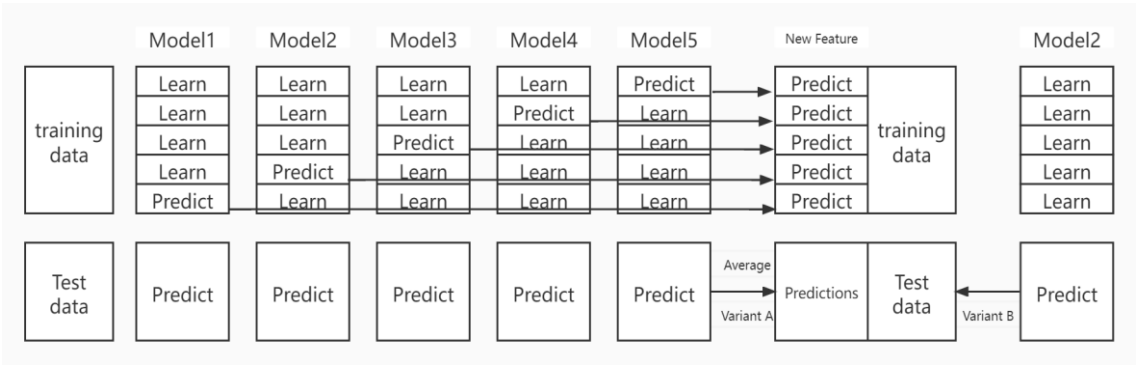

Figure 1. Stacking algorithm process

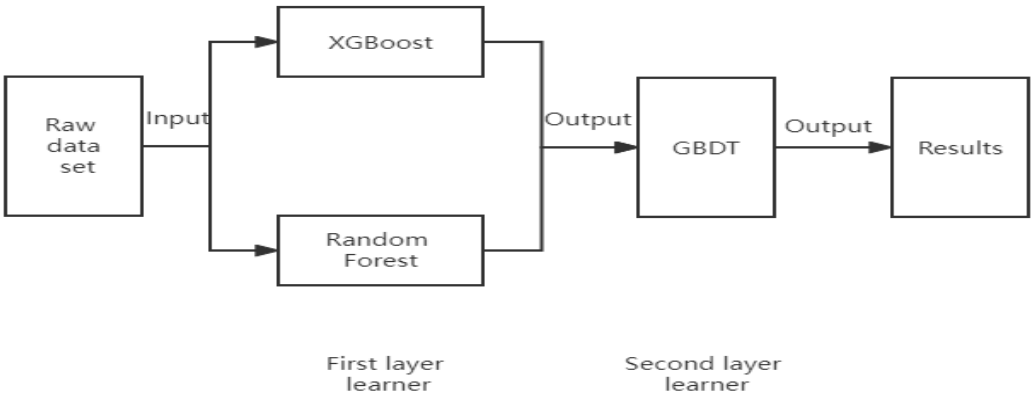

Figure 2. Stacking Model Structure

2.5. Bayesian Optimization Method and its Improvement

(1) Principle of Bayesian optimization algorithm

The idea of Bayesian optimization algorithm is to solve problems in global optimization by approximate approximation. There are two key steps in the execution of the Bayesian optimization algorithm. First, a priori function must be chosen to represent the distribution assumptions of the function being optimized. For this purpose, a Gaussian process is chosen because of its flexibility and ease of handling; second, a collection function must be constructed for determining the next point to be evaluated from the model posterior distribution[11].

In order to carry out Bayesian optimization, it is necessary to consider the establishment of distribution in the objective function, which is usually solved by Gaussian process.

A Gaussian process is an extension of the multidimensional Gaussian distribution to an infinite-dimensional stochastic process. It is defined by the mean value function $\mu(x)$ and the covariance function $k(x,x')$. the Gaussian distribution can be expressed as shown in Equation (5).

$$f(x) \sim GP(\mu(x), k(x,x')) \tag{5}$$

Where $\mu(x) = E(f(x))$, $E(f(x))$ is the mathematical expectation of $f(x)$, and the default value is 0; $f(x)$ denotes the mean absolute error; $k(x,x')$ denotes the covariance function of x.

Assuming that the past information $D_{1:t} = \{x_{1:t}, f_{1:t}\}$ has been obtained, where $f_t = f(x_t)$, then the next value to be searched for is $f_t = f(x_t)$ and the covariance matrix K is noted as：

$$K = \begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_t) \\ \vdots & \ddots & \vdots \\ k(x_t,x_1) & \cdots & k(x_t,x_t) \end{bmatrix} \tag{6}$$

From the Gaussian process, it follows that both f_t and f_{t+1} obey the joint Gaussian distribution. If we set the mean value of both to be 0, then the joint Gaussian distribution can be expressed as shown in Equation (7).

$$\begin{bmatrix} f_{1:t} \\ f_{t+1} \end{bmatrix} \sim N\left(0, \begin{bmatrix} k & k \\ k^T & k(x_{t+1},x_{t+1}) \end{bmatrix}\right) \tag{7}$$

where k can be expressed as：

$$k = [k(x_{t+1},x_1) \cdot k(x_{t+1},x_2) \cdots k(x_{t+1},x_t)] \tag{8}$$

The posterior probability of f_{t+1} is obtained by means of the edge density function is:

$$p(f_{t+1} \mid D_{1:t}, x_{t+1}) = N(\mu_t(x_{t+1}), \sigma_t^2(x_{t+1})) \tag{9}$$

where $\mu_t(x_{t+1})$ and $\sigma_t^2(x_{t+1})$ are calculated as follows:.

$$\mu_t(x_{t+1}) = k^T K^{-1} f_{1:t} \tag{10}$$

$$\sigma_t^2(x_{t+1}) = k(x_{t+1},x_{t+1}) - k^T K^{-1} k \tag{11}$$

From the above calculation it can be estimated that x_{t+1} satisfies a normal distribution at any interval, which in turn enables the sampling function to determine the next most dominant sample point.

By determining the next point to be evaluated through the sampling function, the number of iterations can be reduced and the evaluation cost can be lowered. Usually, the selection of sampling points is considered from two aspects: exploitation and exploration. exploitation is to search around the current optimal solution according to it, so as to find the global optimal solution; exploration is to try to explore the unevaluated sample points to avoid getting into the local optimal solution.

The acquisition function used in this paper is Probability of Improvement, and its acquisition function is shown in Equation (12).

$$PI(x) = \Phi(\frac{\mu(x) - y_{\max} - \delta}{\sigma_t(x)}) \tag{12}$$

where $y_{\max}$ is the current function optimal value, Φ is the standard normal distribution cumulative distribution function, and Φ is the equilibrium parameter that balances the relationship between development and exploration.

(2) Bayesian optimization algorithm improvement

However, influenced by the equilibrium parameter δ, the parameter value is too small will lead to the case of local optimal solution and too large will affect the exploration efficiency. Since the equilibrium parameter δ is a fixed value and cannot be dynamically adjusted according to the optimization condition, it can easily lead to the case of local optimal solution, therefore, this paper constructs the adaptive equilibrium factor to improve the acquisition function so that the acquisition function can avoid falling into the local optimal solution as much as possible. The improved collection function is shown in Equation (13) :

$$PI(x) = \Phi(\frac{\mu(x) - y_{\max} - \varepsilon}{\sigma(x)}) \tag{13}$$

In the formula $\varepsilon = 1 - 1/u$, $u = e^{y_{\max} - y}$, $y_{\max}$ represents the maximum value of the objective function in the current observed data, y represents the objective function value of the collection point in the last iteration, when y is close to $y_{\max}$, ε approaches 0, and the collection function tends to explore the state; when y is far from $y_{\max}$, ε approaches 1, and the collection function tends to develop the state.

2.6. Iv Value and WOE

When building a model, it is usually necessary to judge whether features have predictive ability, while IV refers to the value of information, which can be used to judge whether features can have predictive ability[12].The IV values are calculated as follows:

$$IV_i = (p_{y_i} - p_{n_i}) * WOE = (p_{y_i} - p_{n_i}) * \ln(\frac{p_{y_i}}{p_{n_i}}) = (\frac{y_i}{y_T} - \frac{n_i}{n_T}) * \ln(\frac{y_i / y_T}{n_i / n_T}) \tag{14}$$

Equation (14) is the IV value of a grouping in a variable, which is the sum of the IV values of each grouping , n is the number of variable groupings. In order to reflect the proportion of the sample size of a variable in the current subgroup to the overall, ($p_{yi} - p_{ni}$) is added here before WOE, so as to better reflect the contribution of a variable to the overall, the smaller the proportion, the smaller the contribution, and vice versa.

WOE in Equation (15) means weight of features. It is a way to encode the features. But the features need to be encoded after taking the corresponding grouping. After grouping, The WOE value for group i is calculated as follows:

$$\mathrm{WOE_i} = \ln(\frac{p_{y_i}}{p_{n_i}}) = \ln(\frac{y_i / y_T}{n_i / n_T}) = \ln(\frac{y_i}{y_T}) - \ln(\frac{n_i}{n_T}) \tag{15}$$

In Eq. (20), p_{yi} is the ratio of the number of past due in the group to the overall number of past due, p_{ni} is the ratio of the number of non-past due in the group to the overall number of non-past due, y_i is the number of past due in the group, n_i is the number of non-past due in the group, y_T is the number of all past due in the sample, and n_T is the number of all non-past due in the sample. Therefore, the meaning of WOE is the difference between "the number of past due in the group as a percentage of all past due" and "the number of non-past due in the group as a percentage of overall non-past due".

Usually an IV value less than 0.3 indicates no predictive power.

2.7. Peterson Correlation Coefficient Method

The Pearson correlation coefficient method is a measure of correlation between characteristics [13]. It is calculated as shown in Equation (16).

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 (y - \bar{y})^2}} \tag{16}$$

Where r indicates the correlation between two features. Usually r is less than 0.4 for weak correlation, greater than 0.6 for strong correlation, and greater than 0.8 for very strong correlation.

2.8. Evaluation Indicators

In order to enable comparison of training effects among different models, so the evaluation metrics taken in this paper include confusion matrix, accuracy, precision, recall, F1-score and AUC to measure the performance of a model [14].

The representation of the confusion matrix is shown in Table 1.

Table 1. Confusion Matrix

		Predicted results	
		1	0
True Category	1	TP	FP
	0	FN	TN

The accuracy rate is the proportion of correct samples to the total sample.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

The precision rate is the sample of all positive class samples with correct predictions.

$$Precision = \frac{TP}{TR + FP} \tag{18}$$

Recall is the fraction of all positive class samples that are correctly predicted.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{19}$$

The F1-score is the summed average of the recall and precision rates. It satisfies.

$$\frac{2}{\mathrm{F}_1} = \frac{1}{\mathrm{Precision}} + \frac{1}{\mathrm{Recall}} \tag{20}$$

$$\mathrm{F}_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{TN}} \tag{21}$$

The ROC curve is a visual expression of the model effect. The dynamic relationship between TP and FP in the model is reflected by the drawn curve. To some extent, the differences between different learners can be understood through the ROC curve. The AUC value is the area under the curve, which is used to measure the generalization of the model.

3. DATA PRE-PROCESSING

3.1. Data Analysis

The data of this experiment are 11017 real data after desensitization provided by UnionPay. We built an extensive dataset with 199 credit characteristics.

3.2. Missing Value Handling

The source of the individual credit data widely miscellaneous, there may be repeat characteristics and lack of situation, and in the process of personal credit evaluation, the lack of some variable

values will affect the final prediction, if applied to the actual, may result in incalculable losses, so the first step to access to the data set needs to missing features the data set.

Table 2. Missing feature amount and proportion

Features	Missing amount	Missing percentage	Features	Missing amount	Missing percentage
X_121	10963	0.997906	X_110	10913	0.993355
X_120	10963	0.997906	X_063	10913	0.993355
X_119	10963	0.997906	X_071	10896	0.991808
X_118	10952	0.996905	X_072	10896	0.991808
X_102	10952	0.996905	X_073	10896	0.991808
X_103	10952	0.996905	X_107	10877	0.990078
X_104	10952	0.996905	X_115	10870	0.989441
X_111	10914	0.993446	X_116	10870	0.989441
X_064	10914	0.993446	X_117	10868	0.989259
X_062	10913	0.993355	X_108	10846	0.987257
X_109	10913	0.993355	…	…	…

As can be seen from Table 2, the missing ratio of X_062-X_073, X_081-X_087, X_092-X_120, X_128-X_130, X_133, X_135 and X_136 reaches more than 70%. Because the missing ratio is too high, if filling is adopted, It will affect the accuracy of the model. In order to reduce the deviation, the operation of deleting features is adopted in this paper.

3.3. Balanced Processing

As can be seen from the bar chart shown in Figure 3, the studied data set is unbalanced and not overdue: overdue = 4:1. Imbalanced data classification means that the proportion of categories in the data set is unbalanced. If the proportion of one category is large, the algorithm will favor the category with large proportion in classification. In order to eliminate the influence caused by the imbalance problem, this paper uses SMOTE algorithm to adjust the imbalance, so that the ratio of non-overdue class and overdue class in the processed data set reaches 1:1.

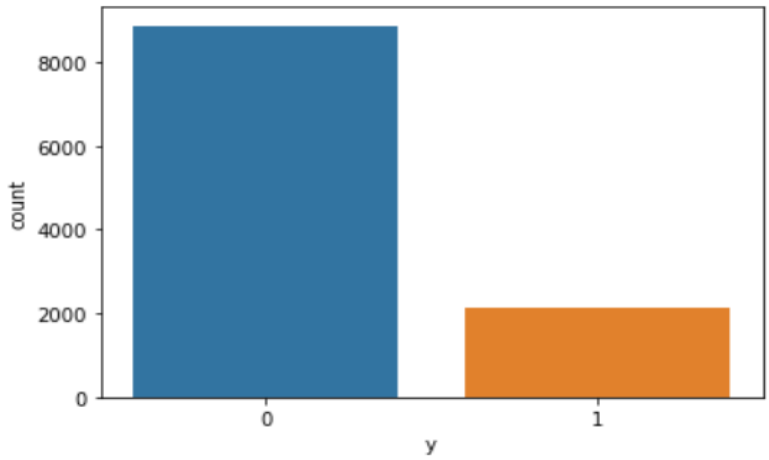

Figure 3. Positive and negative sample proportions

3.4. Feature Dimensionality Reduction

Credit data sets are characterized by high dimensionality and large redundancy among feature sets , so feature dimensionality reduction is needed. There are 146 features left in the dataset after processing for missing values. Feature dimensionality reduction is divided into two steps. Firstly, the IV value is calculated to remove the features with too low IV value. Secondly, the correlation analysis of the dataset was carried out to delete the features with high correlation.
Table 3 shows the statistics of some feature IV values：

Table 3. Value IV of the feature

Characteristics	Iv value	Characteristics	Iv value
X_125	0.640842	X_142	0.575979
X_146	0.631067	X_045	0.567916
X_078	0.622846	X_139	0.567685
X_127	0.615642	X_137	0.556560
X_126	0.606877	X_140	0.555608
X_141	0.601162	X_194	0.548917
X_144	0.593460	X_145	0.546987
X_131	0.590596	X_138	0.544725
X_059	0.582293	X_195	0.543420
X_079	0.576658	X_060	0.542415

After removing the features whose IV value was lower than 0.03 through the first step, 111 feature variables remained.

Correlation analysis was performed on the features to remove the variables with low IV value in the features with high correlation, and Pearson correlation coefficient method was used for processing.

After calculation by Pearson correlation coefficient method, features with correlation higher than 0.7 were removed to obtain the processed data set, which contained 40 feature variables.

4. EMPIRICAL STUDY

In this section, XGBoost, random forest and GBDT were first used to construct XRG-Stacking model and compared with XGBoost, random forest, GBDT, logistic regression and decision tree to verify the performance improvement of the fusion model compared with the single model. Furthermore, the improved Bayesian optimization method is used to optimize the parameters of XGBoost, random forest and GBDT. At the same time, it is compared with the optimization results of Bayesian optimization algorithm, grid search, random search, simulated annealing and genetic algorithm to verify the superiority of the improved Bayesian optimization algorithm. Finally, the IMPBO-XRG-Stacking model and the optimized base model were constructed by optimized XGBoost, random forest and GBDT to prove the improvement of problem accuracy by parameter optimization.

4.1. Experimental Verification

4.1.1. Comparative Analysis of XRG-Stacking Model

The experimental comparison results of XRG-Stacking and random Forest, XGBoost, GBDT, logistic regression and decision tree are shown in Table 4.

Table 4. Analysis of comparative results

	accuracy	precision	Recall	F1-score	AUC
logistic regression	0.6549	0.6683	0.6162	0.6412	0.6767
decision tree	0.7808	0.7630	0.8151	0.7882	0.8613
random Forest	0.7974	0.7938	0.8044	0.7991	0.8869
GBDT	0.8213	0.8457	0.7864	0.8150	0.9130
XGBoost	0.8607	0.9018	0.8267	0.8626	0.9408
XRG-Stacking	0.8779	0.9143	0.8353	0.8830	0.9508

According to the analysis in Table 4, the results of XRG-Stacking method had better effects compared with other basic models, indicating that the XRG-Stacking model that is combined with multiple models has better predictive effects on personal credit overdue problems than a single model. However, the XRG-Stacking model fusion method does not improve the prediction performance of the problem with a single model. By combining with real life, the number of resident loans is hundreds of millions and the number is very large. Therefore, the slight improvement in the performance of personal credit overdue prediction has a huge impact. It also has big implications for the financial industry. When it is difficult to further improve the performance of personal credit overdue problems by using a single model, the Stacking model fusion method can be considered to improve the ability of identifying whether users are overdue, so as to achieve better prediction effect.

4.1.2. Optimization Algorithm Optimization Base Model Comparison Experiment

The parameters to be optimized by the improved Bayesian optimization method for XGBoost, random Forest and GBDT models and the best parameter combination optimized by the improved Bayesian optimization algorithm are shown in Table 5

Table 5. Optimization parameters and optimal values

XGBoost	
parameter	**value**
learning_rate	0.07
n_estimator	177
min_child_weight	4.8
max_depth	10
gamma	0.31
subsample	0.83
colsample_bytree	0.72
Random Forest	
parameter	**value**
n_estimators	48
max_depth	10
min_samples_split	11
min_samples_leaf	14
GBDT	
parameter	**value**
n_estimator	70
learning_rate	0.1
subsample	0.8
max_depth	8
min_samples_split	150
min_samples_leaf	40

The prediction results obtained by feeding the optimization parameters into the model are compared with the optimization results of other optimization methods, as shown in Table 6.

Table 6. Comparative analysis of optimization models

	accuracy	**precision**	**Recall**	**F1-score**	**AUC**
XGBoost	0.8607	0.9018	0.8267	0.8626	0.9408
Random Forest	0.7974	0.7938	0.8044	0.7991	0.8869
GBDT	0.8213	0.8457	0.7864	0.8150	0.9130
Improved Bayesian-XGBoost	0.8781	0.9190	0.8388	0.8734	0.9498
Improved Bayesian-Random Forest	0.8323	0.8403	0.8211	0.8306	0.9181
Improved Bayesian-GBDT	0.8722	0.9107	0.8256	0.8661	0.9442
Bayesian-XGBoost	0.8721	0.9130	0.8301	0.8702	0.9478
Bayesian-Random Forest	0.8291	0.8310	0.8270	0.8290	0.9147
Bayesian-GBDT	0.8675	0.9023	0.8245	0.8616	0.9441

As can be seen from Table 6, compared with other optimization methods, the improved Bayesian optimization method has the best effect in optimizing the three base models.

4.1.3. Comparative Analysis of IMPBO-XRG-Stacking Model

The comparison and analysis results of the IMPBO-XRG-Stacking model constructed with optimized XGBoost, random forest and GBDT and the optimized base model are shown in Table 7.

Table 7. Comparison of evaluation results

	accuracy	precision	Recall	F1-score	AUC
Improved Bayesian-Random Forest	0.8323	0.8403	0.8211	0.8306	0.9181
Improved Bayesian-GBDT	0.8722	0.9107	0.8256	0.8661	0.9442
Improved Bayesian-XGBoost	0.8781	0.9190	0.8388	0.8734	0.9498
IMPBO-XRG-Stacking	0.8879	0.9243	0.8453	0.8830	0.9551

As can be seen from Table 7, the ImpBO-XRG-Stacking model is the highest compared with the single model in accuracy, accuracy, recall, F1-score and AUC.

5. CONCLUSION

The main research content of this paper is to use optimized XGBoost, random forest and GBDT to build Stacking model, select real desensitization data provided by UnionPay as data set, and send it into the model for training after data preprocessing. The comparison experiment with the optimized single model proves that the model fusion and parameter optimization can improve the accuracy of problem prediction.

ACKNOWLEDGEMENT

This work was supported by The Humanity and Social Science Youth foundation of Ministry of Education of China(21YJCZH20);Innovation Team Project of Higher Education of Guangdong "Intelligence Rule of Law Research Team" (2022WCXTD008).

REFERENCES

[1] Zehra W, (2021) Cross corpus multi-lingual speech emotion recognition using ensemble learning, Complex & Intelligent Systems, Vol. 7, No.4, pp1845-1854.
[2] J. C. Wiginton, (1980) A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior,Journal of Financial Quantitative Analysiss, Vol. 15, No.3, pp 757-770.
[3] Shin K S, (2005) An Application of Support Vector Machines in Bankruptcy Prediction Model,Expert Systems with Applications, Vol. 28, No.1, pp 127-135.
[4] T. Chen, (2016) XGBoost: A Scalable Tree Boosting System,KDD'16: Proceedings ofthe 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 785-794.
[5] Huang Y P, (2019)A New Perspective of Performance Comparison among Machine Learning Algorithms for Financial Distress Prediction, SSRN Electronic Journal.
[6] Chang Y C, (2018) Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions, Applied Soft Computing, pp73.
[7] Trizoglou P,(2021) Liu X, Lin Z. Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines, Renewable Energy , Vol.179,pp 945-962
[8] Teles G, Rodrigues J J P C, Rabêlo R A L, et al.(2021) Comparative study of support vector machines and random forests machine learning algorithms on credit operation,Software: Practice and Experience,Vol. 51,No.12,pp 2492-2500
[9] Yang J S, Zhao C Y, Yu H T, et al.(2020) Use GBDT to predict the stock market, Procedia Computer Science,Vol.174,pp 161-171
[10] Khoei T T, Labuhn M C, Caleb T D, et al.(2021)A Stacking-based Ensemble Learning Model with Genetic Algorithm For detecting Early Stages of Alzheimer's Disease,2021 IEEE International Conference on Electro Information Technology,pp 215-222
[11] Xia Y, Liu C, Li Y Y, et al.(2017)A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring,Expert systems with applications,Vol.78,pp 225-241

[12] Wang Z, Zhang P, Sun W, et al.(2021) Application of data dimension reduction method in high-dimensional data based on single-cell 3D genomic contact data,ASP Transactions on Computers , Vol.1,No.2,pp 1-6
[13] Xie A, Yang H, Chen J, et al.(2021) A short-term wind speed forecasting model based on a multi-variable long short-term memory network,Atmosphere, Vol.12,No.5,pp 651
[14] Ahsan M M, Mahmud M A P, Saha P K, et al, (2021)Effect of data scaling methods on machine learning algorithms and model performance[J]. Technologies,Vol. 9,No.3,pp 52

AUTHOR

JICONG YANG is currently pursuing his master's degree at Guangdong University of Finance and Economics, majoring in big data analytics and applications.

A FORMAL COMPOSITION OF MULTI-AGENT ORGANIZATION BASED ON CATEGORY THEORY

Abdelghani Boudjidj[1] and Mohammed El Habib Souidi[2]

[1]Ecole nationale Supérieure d'Informatique (ESI),
BP 68M, 16270, Oued-Smar Algiers, Algeria
ICOSI Lab University,
Abbes Laghrour khenchela BP 1252 El Houria 40004 Khenchela, Algeria
[2]University of Khenchela, Algeria
ICOSI Lab University,
Abbes Laghrour khenchela BP 1252 El Houria 40004 Khenchela, Algeria

ABSTRACT

The application of organizational multi-agent systems (MAS) provides the possibility of solving complex distributed problems such as, task grouping mechanisms, supply chain management, and air traffic control. The composition of MAS organizational models can be considered as an effective solution to group different organizational multi-agent systems into a single organizational multi-agent system. The main objective of this paper is to provide a MAS organizational model based on the composition of two organizational models, Agent Group Role (AGR), and Yet Another Multi Agent Model (YAMAM), with the aim of providing a new MAS model combining the concepts of the composed organizational models. Category theory represents the mathematical formalism for studying and modeling different organizations in a categorical way. This paper is mainly based on the idea of modeling the multi-agent organization AGR and YAMAM in a categorical way in order to obtain formal semantic models describing these organizations of MAS, then compose them using also the theory of categories which represents a very sophisticated mathematical toolbox based on composition.

KEYWORDS

Multi-agent systems, Organizational models, Category theory, composition.

1. INTRODUCTION

Multi-agent systems represent a set of agents that communicate with each other to meet a specific need, a goal or to accomplish a task (or a set of tasks) [1], the evolution of multi-agent systems and the cooperation between agents open the way to the emergence of the notion of organization for this type of complex systems, several models have been presented in the literature that take into consideration the notion of organization [2], such as Agent Group Role (AGR) [3], Yet Another Multi-Agent Model (YAMAM)[4] ,…

MAS organizational models can be defined as MAS mechanisms used in order to coordinate the agents' behaviors to accomplish complex tasks [5]. AGR organizational model [6] can be considered as one the most popular organizational mechanism used in the last years. This model is based on three different concepts, which are, Agent, Group, and Role used simultaneously to

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 93-110, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121909

reflect the system behavior. In relation to complex problems [7], AGR was applied to the Pursuit-Evasion Game (PEG) [8] in order to provide a pursuit coalition formation allowing the grouping of the pursuers to capture the detected evaders.

In relation to AGR model, YAMAM is not based on the concept group, however, it is based on four concepts: Agent, Role, Task and Skills. The main principle of YAMAM can be resumed as follows: in order to play a specific role, an agent must have the appropriate skills in order to be able to improve the specific task. Recently, YAMAM was also applied to the PEG [9] with the aim of providing a pursuit groups access mechanism by the use of the concepts forming model.
To study an organizational multi-agent system, it is necessary to model these elements and the relationships between them, as well as these relationships with the environment in which it is located.

Category theory represents a multidisciplinary mathematical toolbox. It has been used in many fields of computer science. This theory offers a rich body of theory for reasoning about structures (objects and relationships between objects) [10]. In relation to existing formal methods, category theory provides the ability to organize and layer abstractions, as well as to find commonalities between different structures. Nowadays, the use of CT is not only reserved to the pure mathematicians. In other words, it was proved that CT represents a powerful tool in computer science, and industry [11].

Category theory is used to study and formalize organizations and collective phenomena in human societies with the aim of capturing their logics in categorical models. It is based on the idea of using category theory to develop organizational systems multi-agents by taking inspiration from collective phenomena and organizations in human societies in the work of [12]

Category theory is based on composition as in algebraic languages [13], it is at a very high level of abstraction, represented by a set of objects in the form of a category, morphisms between objects that represent the relationships between them, and functors between categories, these notions will be used for the modeling and the composition of organizational multi-agent systems.

The Category theory in its principle as indicated by its definition represents categories of objects and the links between objects as well as links between these categories of objects, (which can be sets or groups), this resemblance with the models of organizational multi-agent systems which are also groups of agents, links between agents and links between groups of the organization, , the agents play roles and perform tasks, these explicit links represent a support for the choice of the category theory for the modeling of organizational MAS

In other words, it is about transforming the Agent-Group-Role (AGR) organizational model in a categorical way in order to obtain a formal semantic model. This formal model allows the analysis, the verification and also the validation of the main concepts of an organization at a high level of abstraction.

The system of systems is a field that arose from the need to deal with specific types of problems where the fact that many capabilities and desired outcomes will be developed through the integration or composition of existing systems[14].

Organizational multi-agent systems are promising systems for managing the emergence of new systems, the main contribution of MAS as a simulation technique is its ability to represent the behavior of human actors, this representation through the organizational SMAs takes into consideration several important concepts in human society [15] such as the role, the group, the

alliances... taking into account a heterogeneous and dynamic representation of the environment on which the system is located.

In this paper, we introduce a formal composition of AGR and YAMAM by the use of category theory principle. The main objective of this work is to provide a new organizational model that take into consideration the benefits of the concepts forming the composed models.

The paper is organized in the following way: in section 2, we detail the principles of category theory as well as the relations between YAMAM and AGR models. In section 3, we introduce the categorical representation of YAMAM and AGR organizational models. Section 4 is devoted to the categorical composition of the two studied organizational models. Finally, we conclude this paper in section 5 by providing a brief summary regarding the usefulness of category theory.

2. PROBLEM DESCRIPTION

This section is devoted to the description of the problem regarding MAS organizational models as well as category theory. The concept of the organization is very important in relation to multi-agent systems. The use of the organization in this field requires a formal framework to mathematically manage it, and validate interesting MAS properties.

Several models have been made to reflect the importance of organization in multi-agent systems, and to lead to effective solutions to complex problems, such as AGR and YAMAM, in the AGR model access to the Role is unconditional for any agent but on the other hand there is a control mechanism for access to the Roles in the YAMAM model, but there is not the notion of the group which exists in the AGR model, this notion of the group plays an important role in the organization,

The relation between AGR and YAMAM models is that they have common concepts Agents and Roles and to develop the relation between these concepts and take the benefits lake the access to roles by agents, the agents have to improve skills to take the role, and use the notion of group to simplify the accomplishment of system tasks, we will integrate them into a new model in a formal way, this new model which will contain the notion principles of the two models (AGR and YAMAM) and use it to resolve several complex problems.

Category theory allows to model AGR and YAMAM in a formal manner, category theory provides many mathematical aspects and concepts at a very abstract level. A categorical representation of the AGR multi-agent system was recently introduced through the use of category theory [16], which will allow us to represent or to transform the concepts related to YAMAM in a formal model, examine them in an abstract way, and formalizing the system as collections of objects (categories) and morphisms with the aim of reasoning about these objects and their relationships or interactions (morphisms).

CT is based on the mathematical composition as a whole of operation between objects, morphisms or functors, or even between categories, Composition: From the arrows :

It includes the following data [17]:

- Objects: A, B, C, etc.
- Morphisms: f, g, h, etc.
- Domain and Codomain: For each arrow f, we give objects: dom (f), cod (f) called domain and codomain of f. We write: f: A → B to indicate that A = dom (f) and B = cod (f).

- Composition: From the arrows f: A → B and g: B → C, that is to say with: cod (f) = dom (g), we have a given arrow: g ◦ f: A → C.
- Identity: For each object A there is a given arrow 1A: A → A, called identity arrow of A.

These components are required to comply with the following laws:

- Associativity: h ◦ (g ◦ f) = (h ◦ g) ◦ f, for all f: A → B, g: B → C, h: C → D.
- Unit: f ◦ 1A = f = 1B ◦ f, for all f: A → B.

f: A → B and g: B → C, that is to say with: cod (f) = dom (g), we have a given arrow: g ◦ f: A → C.

a field called Systems of systems "SoS" is a term hugely used to describe systems composed of systems independent constituent acting to achieve a common goal. System of systems engineering (SOSE) is a field that emerged from the need to address the issues of SoS [18], this field has evolved through the results of the work of integrating existing or legacy systems, integrating one part of the system into another, or integrating the entire system, resulting in a system that provides the capability desired. Authors in [19], and as we said that CT is based on composition, we will compose AGR and YAMAM using CT to have a new model.

The objective of this work is to take advantage of the non-common concepts that make up the two models in order to see the emergence of a new more complete organizational model.

So in this work we are going to treat two essential points, to present YAMAM in a categorical way, the formal model obtained for YAMAM, as well as that of AGR form two categories which will allow us thereafter to compose them that take into consideration the benefits of the concepts forming the composed models in only one system categorically, by using comma category.

3. AGR AND YAMAM MODELING WITH CT

3.1. Categorical representation of the YAMAM model

In what follows, we present the categorical modeling of YAMAM, examine its structure, and represent the main concepts such as: skills, tasks, role and agents, and their relationships via category theory, and thereafter a return to the global system which is the set of these categories, which represent YAMAM using constructions from the Category Theory (CT).

But before we will detail the YAMAM model and its concepts then represent it in a categorical way.

3.1.1. Yet Another Multi-Agent Model (YAMAM)

Based on 4 concepts which are the pillars of this model, Agent, Role, skill and task.

The organization of YAMAM is described by its inherent structure, therefore the relations between the agents are paramount in relation to the agents and their behaviors. The organization is represented by a set of agents who have goals, tasks, skills and Roles, the goals of which the agent works to achieve them, performs a set of actions and plays one or more roles in order to achieve a desired goal. Agents may have new goals after a system update, which causes new actions to perform and roles to play.

Agents communicate with each other and cooperate to achieve the overall goal of the organization.

Agent: is defined as an autonomous entity in an environment equipped with sensors, able to communicate and use skills to complete tasks, these skills can be evolved over time.

Role: is defined as a function or form of agent identification or service. A role can be played or assigned to one or more agents in an environment, a set of tasks to be performed by agents. We consider that an agent can play a role only if he has the skills required to perform the tasks involved for this role.

Skill: is represented by a unit of knowledge necessary for the processing of given tasks. An agent can be aggregated several skills in order to perform the required set of tasks.

Task: it is the operation of a skill or an action that requires one or more skills to complete it.

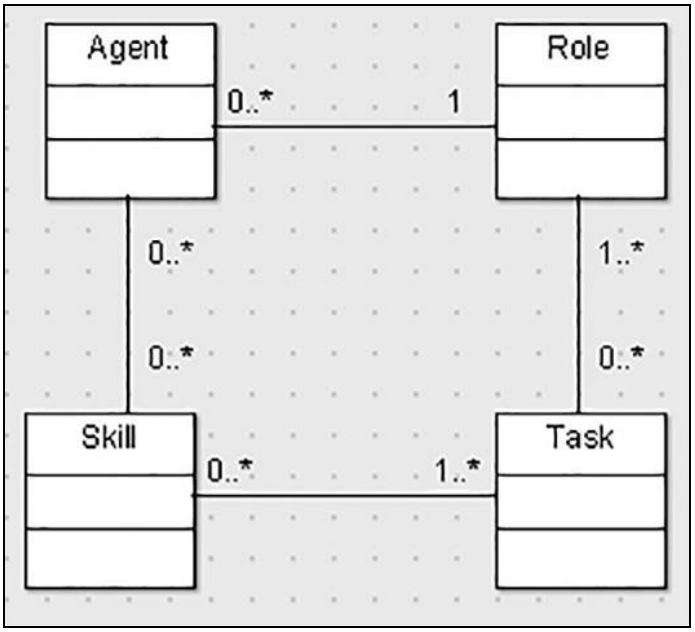

Figure 1. YAMAM meta-model

The following section contains definitions of: Skills, Actions, Tasks, Roles, Agents, and their relationships, which allows us to formalize the YAMAM organization in a categorical way.

3.1.2. Agent Category

The Agent category encompasses all agents of the YAMAM organizational model, represented by these objects, which can be autonomous objects in an environment capable of communicating and using skills to complete tasks. Morphisms are the identity morphisms of each object (agent).

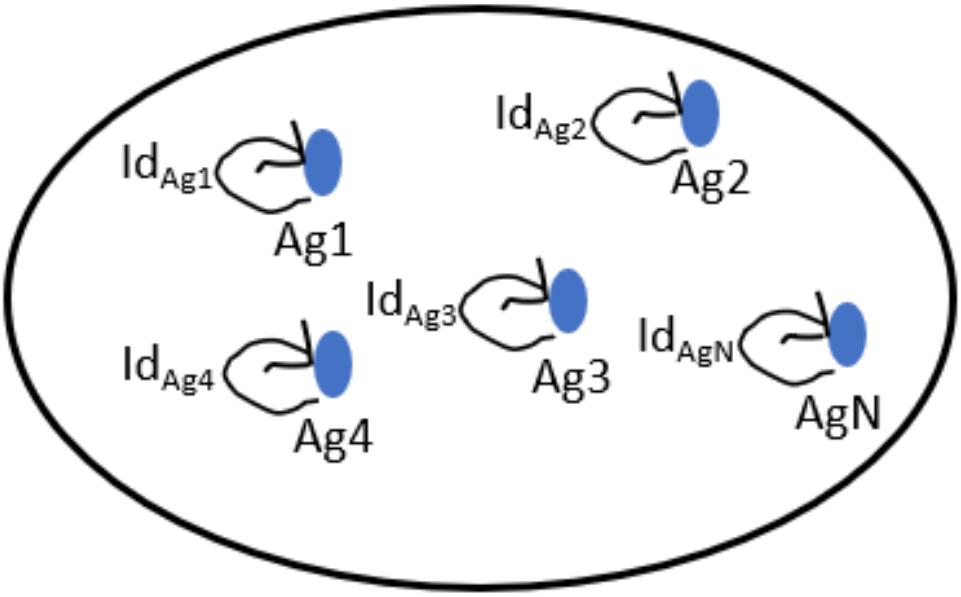

Figure 2. Category Agents

3.1.3. Role Category

In its definition, a role is a function or form of agent identification or a service, a role can be played or assigned to one or more agents in an environment, a set of tasks to be performed by agents. So to represent the role category, you have to define the task category.

3.1.4. Task Category

In the YAMAM model, the Task is the operation of a skill or action that requires one or more skills to complete. So the Task category is linked to the Skill category. An agent is able to complete a given task, if it has the necessary skills (the set of units of knowledge that represents the skill). To define the Task category, you must define the Skill category and then go up to the Role category

3.1.5. Skill Category

The skill category represented by one or more units of knowledge necessary for the processing of given tasks, an agent can be aggregate several skills in order to perform the set of required tasks. Objects and morphisms in this category are represented as follows

- Objects: are a set of knowledge units designated by CM1, CM2, etc.
- Morphisms: identity morphisms

CM objects are categories that encompass one or more knowledge units from the base category Knowledge Units. The knowledge unit is a discrete category, it contains only the identity morphisms and the objects which are the knowledge units.

The following figure presents the category that includes all the units of knowledge for the YAMAM model,

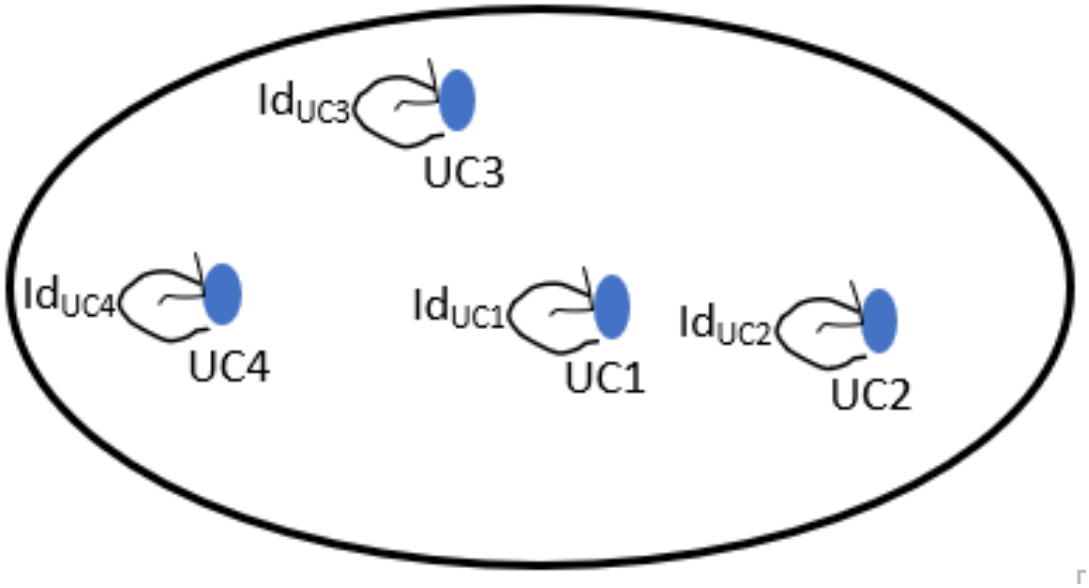

Figure 3. Knowledge units category

From the knowledge unit category, the Skill category is designed, a skill can have one or more knowledge units as shown in the following figure:

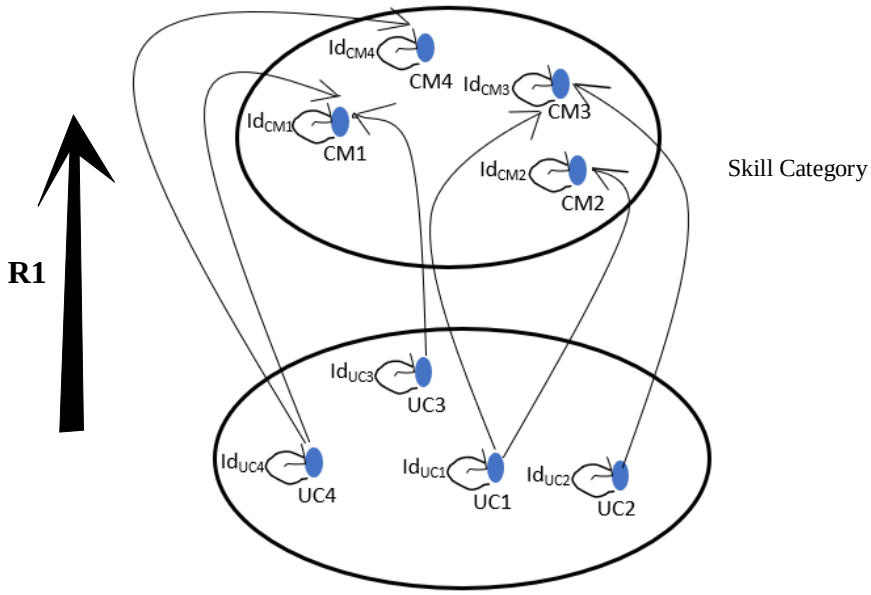

Figure 4. Skill Category creation

Functor R1:
Objects
R1(UC1) = Skill.CM2
R1(UC1) = Skill.CM3
R1(UC2) = Skill.CM3
R1(UC3) = Skill.CM1
R1(UC4) = Skill.CM1

Morphisms
$R1(Id_{UC1}) = Id_{R1(MC2)}$
$R1(Id_{UC1}) = Id_{R1(MC3)}$
$R1(Id_{UC2}) = Id_{R1(MC3)}$
$R1(Id_{UC3}) = Id_{R1(MC1)}$
$R1(Id_{UC4}) = Id_{R1(MC1)}$

Once we have designed the Skill category, we will create the Task category. As its definition indicates, a task is the operation of a skill requiring one or more skills to complete it.

The following figure shows the Task category

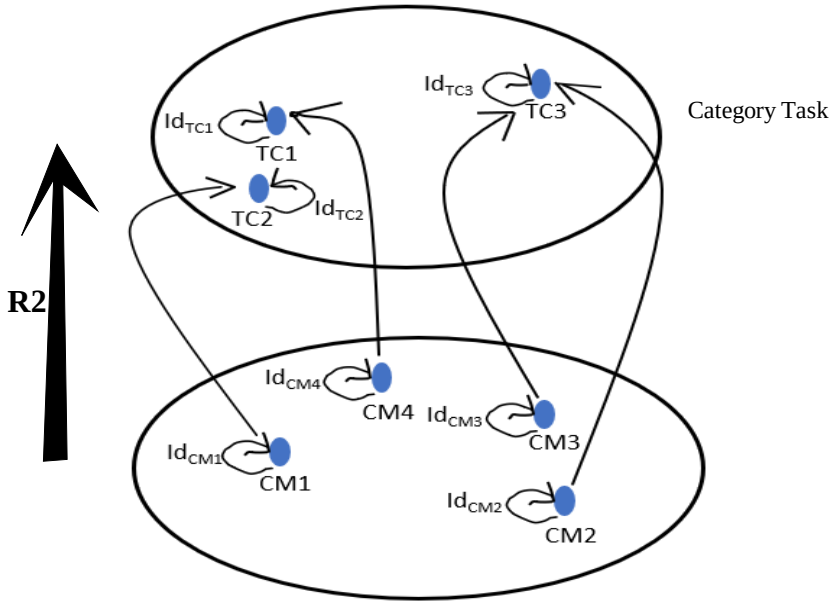

Figure 5. Task category creation

Functor R2:
objets
R2(CM1) = Task.TC2
R2(CM2) = Task.TC3
R2(CM3) = Task.TC3
R2(CM4) = Task.TC1
Morphisms
$R2(Id_{CM1}) = Id_{R2(TC2)}$
$R2(Id_{CM2}) = Id_{R2(TC3)}$
$R2(Id_{CM3}) = Id_{R2(TC3)}$
$R2(Id_{CM4}) = Id_{R2(TC1)}$

The Role category is represented by one or more sequences of tasks to be executed sequentially, in turn it will be designed from the Task category as follows:

Assuming that a RL1 (Role 1) is the sequence of the two tasks TC3 and TC2 represented by the morphism SC1, and RL2 is the execution of the Task TC1 as figure 6 shows.

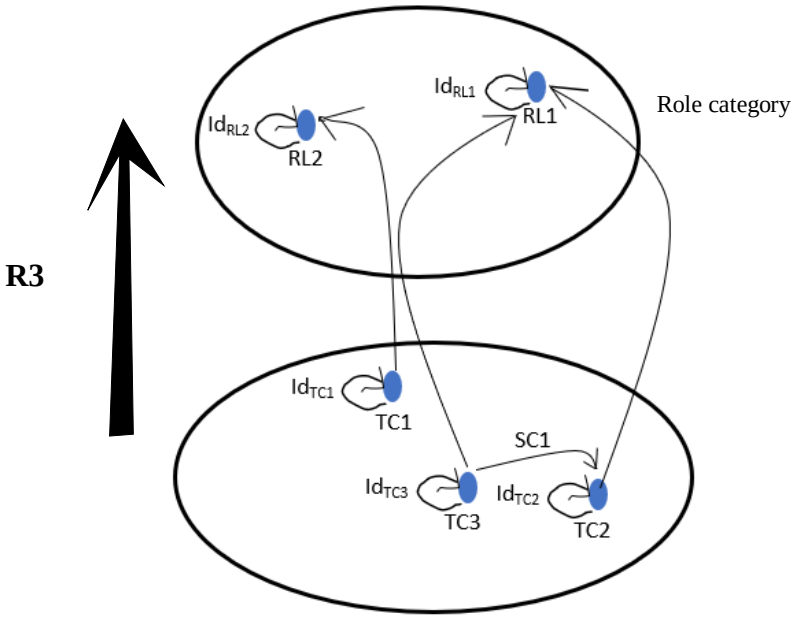

Figure 6. Role category creation

Functor R3 :
Objets
R3(TC1) = Rôle.RL2
R3(TC2) = Rôle.RL1
R3(TC3) = Rôle.RL1
Morphisms
Identity morphisms
R3(Id_{TC1}) = $Id_{R3(RL2)}$
R3(Id_{TC2}) = $Id_{R3(RL1)}$
R3(Id_{TC3}) = $Id_{R3(RL1)}$
Morphism SC1
R3(SC1) = $Id_{R3(RL1)}$

All the categories of the YAMAM organizational model have been presented in categorical form, the following diagram presents the YAMAM categorical model.

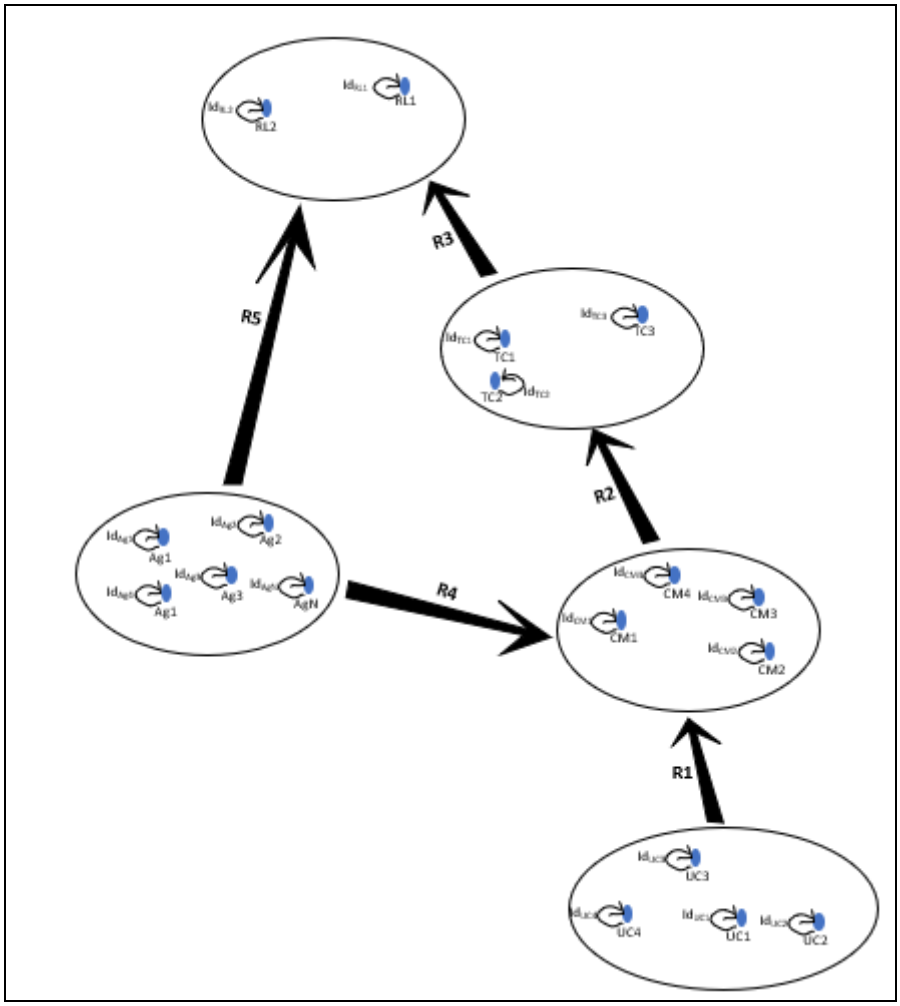

Figure 7. YAMAM Categorical Model

An agent can take the role if he has the skill (or skills) necessary for this role, through the functor R4 the agents point to the objects (skills) which have the abilities to use them to play the roles and complete the tasks. tasks related to them.

3.1.6. Play the roles by the Agents

In this part we will explain how agents take roles via categories.

Assuming that agent Ag1 has skills **CM1** and **CM2**, and agent **Ag2** has skills **CM2** and **CM3**, and agent **Ag3** has skill **CM4**

So the agent **Ag2** is able to complete **TC3**, Ag1 able to complete **TC2** and **Ag3** able to complete **TC1**

Consider the functor **R4**, **R4**: Agent Category → Skill Category.

Objets
R4 : **Ag1→CM1**, **et R4** : **Ag1→ CM2, so R4** : **Ag1→CM1**, **CM2**
This means that Agent Ag1 has skills CM1 and CM2

Morphismes
R4 : $\mathbf{Id_{Ag1}}$→$\mathbf{Id_{CM1}}$
R4 : $\mathbf{Id_{Ag1}}$→$\mathbf{Id_{CM2}}$
Agent objects Ag2 and Ag3:
R4 : **Ag2→CM2**, **et R4** : **Ag2→ CM3, so R4** : **Ag2→** (**CM2**, **CM3)**
R4: **Ag3 →CM4**
In order, Agent **Ag2** has skills **CM2** and **CM3**, Agent **Ag3** has **CM4** skills,
Now suppose we also have the functor **R2**,
Morphismes Agents Ag2 and Ag3 :
R4 : $\mathbf{Id_{Ag2}}$→$\mathbf{Id_{CM2}}$
R4 : $\mathbf{Id_{Ag2}}$→$\mathbf{Id_{CM3}}$
R4 : $\mathbf{Id_{Ag3}}$→$\mathbf{Id_{CM4}}$
R2: Skill Category →Task Category,
Objets
R2 : **CM1→TC2**,
This means that to accomplish the task TC2, the agent in charge of completing it must have the skill CM1.
Morphisms :
R2 : $\mathbf{Id_{CM1}}$→$\mathbf{Id_{TC2}}$
Objets
R2 : **CM2→TC3**, **R2** : **CM3→TC3**, so **R2** : (**CM2, CM3) →TC3**.
To complete the task **TC3** it is necessary to have skills **CM2** and **CM3**
Morphisms
R2 : $\mathbf{Id_{CM2}}$→$\mathbf{Id_{TC3}}$
R2 : $\mathbf{Id_{CM3}}$→$\mathbf{Id_{TC3}}$
Objets :
R2 : **CM4→TC1**,
The same for the **TC1** task, you must have the **CM4** skill in order to accomplish it,
Morphisms
R2 : $\mathbf{Id_{CM4}}$→$\mathbf{Id_{TC1}}$
Roles are defined as performing task(s), so

R3: Task Category → Role Category,
Objets
R3 : **TC1→RL2**,
Obtaining the Role **RL2** by an agent means that this Agent will execute the task **TC1**, and who has the skill(s) to complete it,
Morphisms
R3 : $\mathbf{Id_{TC1}}$→$\mathbf{Id_{RL2}}$
Objets
R3 : **TC2→RL1**, **R3** : **TC3→RL1**, donc **R3** : (**TC3, TC2) →RL1**,
Obtaining the Role RL1 by an agent means that this Agent will execute the task **TC3** and **TC2**, and who has the skills to complete it,
Morphisms
R3 : $\mathbf{Id_{TC2}}$→$\mathbf{Id_{RL1}}$
R3 : $\mathbf{Id_{TC3}}$→$\mathbf{Id_{RL1}}$
R3 : **SC1**→$\mathbf{Id_{RL1}}$

We are going to apply the composition function between functors as follows:

As an example we take the agent **Ag3** which will be linked to the skill **CM4** which is linked to the task **TC1**, the latter in turn is linked to the role **RL2** as shown in the following figure:

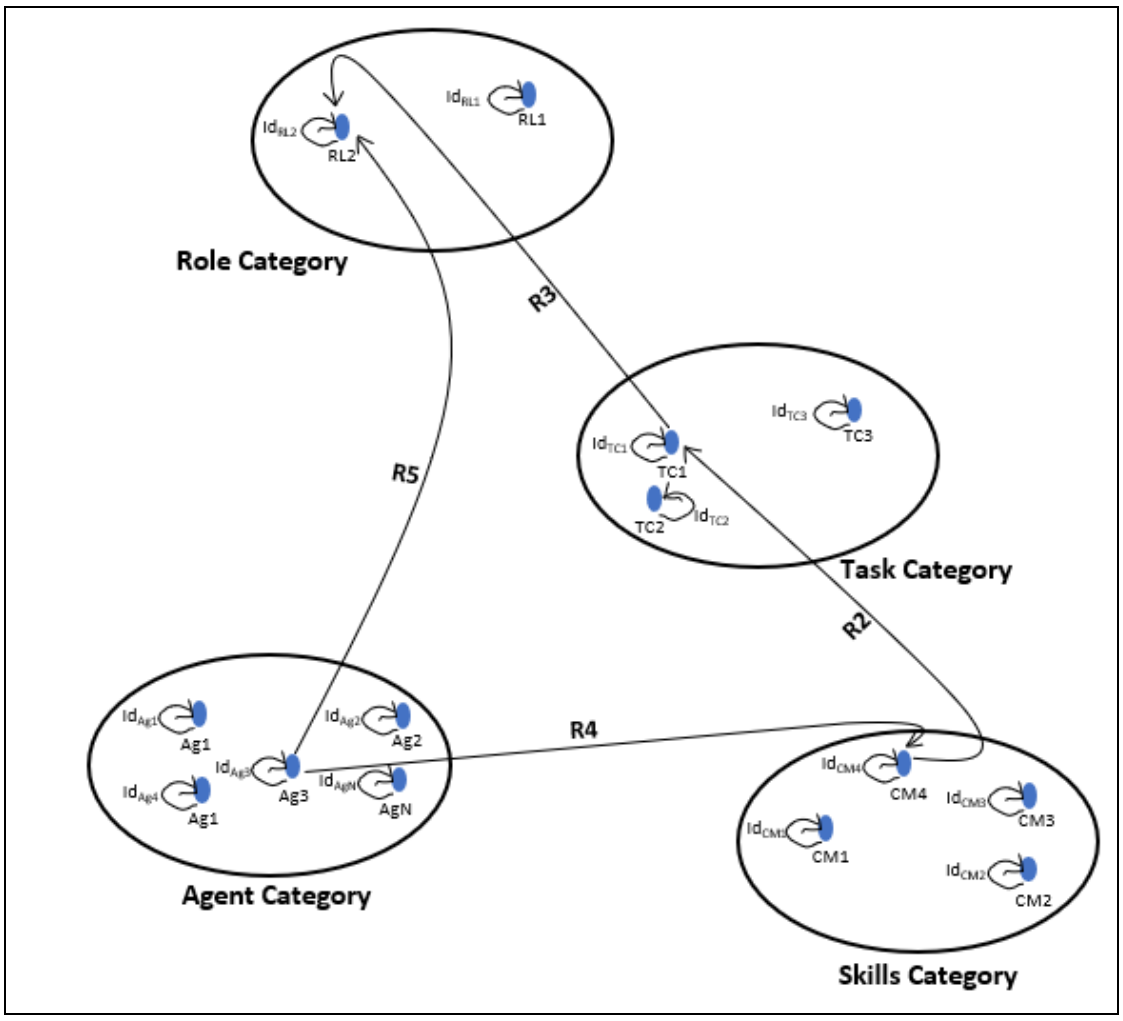

Figure 8. Role playing by an agent in the YAMAM categorical model.

From the previous figure we will obtain the following mathematical equations:

$$\begin{cases} Ag3 \xrightarrow{R4} CM4 \quad (1) \\ Et \\ CM4 \xrightarrow{R2} TC3 \xrightarrow{R3} RL2 \quad (2) \end{cases}$$

From equation 2: **R3 ◦ R2 : CM4 → RL2** called the composite of R2 and R1, so the new equation is: R3 ◦ R2 : CM4 → RL2 **(3)**

Then we will dial 1 and 3: $Ag3 \xrightarrow{R4} CM4 \xrightarrow{R3 \circ R2} RL2$

So: **(R3 ◦ R2) ◦ R4 : Ag3 → RL2**

This last composition is equal to **R5**:

(R3 ◦ R2) ◦ R4 = R5

Through these equations the role RL2 will be played by the Agent Ag3 who has the skill **CM4** that it is necessary for this role, and then Ag3 it will execute the task 3 (**TC3**), and to be able to play a role c' is the same for the other Agents in the Agents category.

4. COMPOSITION OF THE AGR AND YAMAM MODELS

Our organization is made up of agents equipped with environmental sensors to detect obstacles and the position of other agents, to see changes, to communicate with each other, etc. they can also act or react according to their objectives and their roles in the organization.

The environment in our study represented by a set of objects (categories) with certain properties and relationships between these objects.

Category theory is based on Composition, objects, categories and/or morphisms, it allows us to compose two categories to have a new category by ensuring any constraints or rules that may be presented in one of these categories composed, this powerful composition operation, will be used to compose two different organizational systems in a new system. The reformulation by the categories allows us to move on to the composition of the latter.

The categories represent a reformulation of the organizational SMAs, composing the functors between source categories (A, B) towards a target category (C) reflects the composition of a system A (AGR) with a system B (YAMAM) in a new system C , and the result is guaranteed to fit the target categories and satisfy the starting category paths.

In our case, we consider that a member system is represented by an organization of agents more particularly an organizational SMAs (source category A or B). So we are talking about a composition of organizations.

The first organizational system A is represented by the AGR organizational model that we have detailed in our work (Towards a formal multi-agent organizational modeling framework based on category theory).

The second organizational system B is represented by the YAMAM model, Which we modeled in the previous part.

We will use comma category to perform the composition of the two AGR and YAMAM categorical models,

4.1. Composition using comma category

Commas categories [20] are categories where The basic idea is the elevation of morphisms of a category C to objects of other categories.

The complete generality for this category can be obtained by taking a subclass of morphisms - those whose source is in the image of a functor L: A→ C and whose target is in the image of another functor R: B→ C. Comma category (L; R) has as objects, triplets of the form (a; f: L (a) → R (b); b) where a is an object of A and b an object of B. A morphism in (L; R), from (a; f; b) to (a'; f'; b') is a pair of morphisms s: a→ a' and t: b→ b' such that the following square commutes :

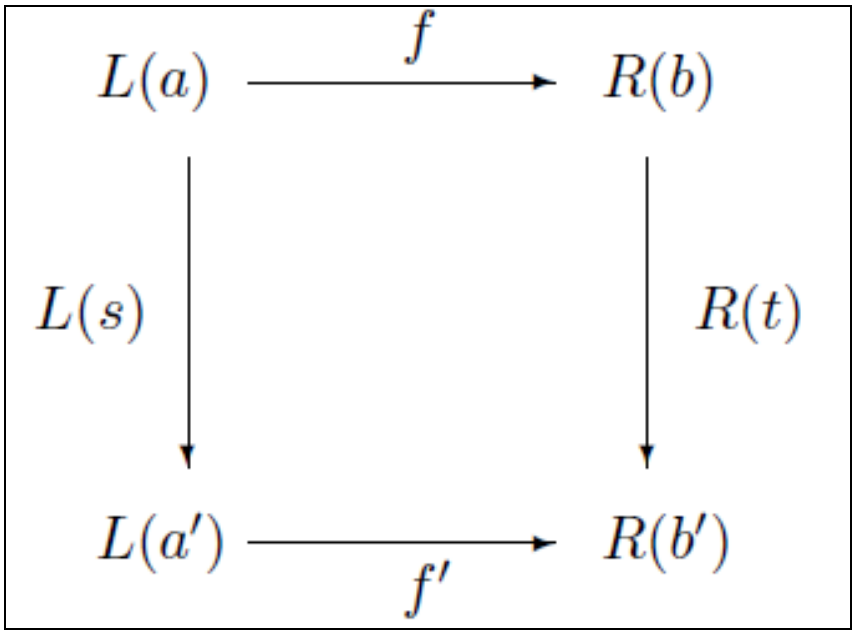

Figure 9. Composition of objects and morphisms in comma category

The composition is defined by component, (s , t) ◦ (s', t') = (s◦s', t◦t') and the identities are pairs of identities. The category commas are associated with two projection functors:

Left: (L, R) → A; right: (L, R) → B defined by: Left (a, f, b) = a
And Left(s, t) = s, And similarly for the right functor.

This category can serve us in the composition of systems, it does not modify the basic starting systems (source), and the new category uses the two source categories in its operation.

After modeling the two systems AGR and YAMAM categorically, in this part we will use the categorical models to compose them via Comma category

The following figure presents the two organizational models on the left AGR and on the right YAMAM,

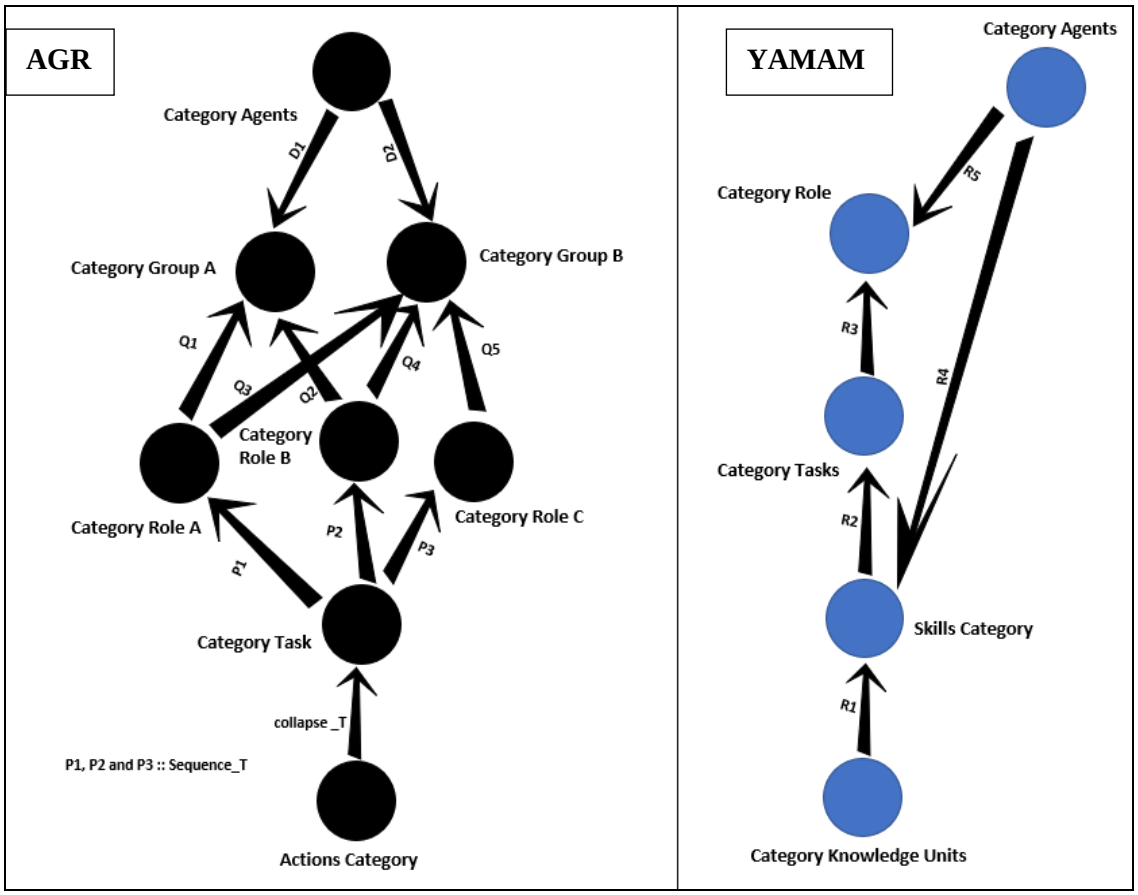

Figure 10. Simplified categorical representation of the two organizational models AGR and YAMAM.

We have presented the categories in the form of objects to simplify the drawing of categories of the two AGR and YAMAM systems, these two models will be composed in a new category, in the middle as shown in the following figure:

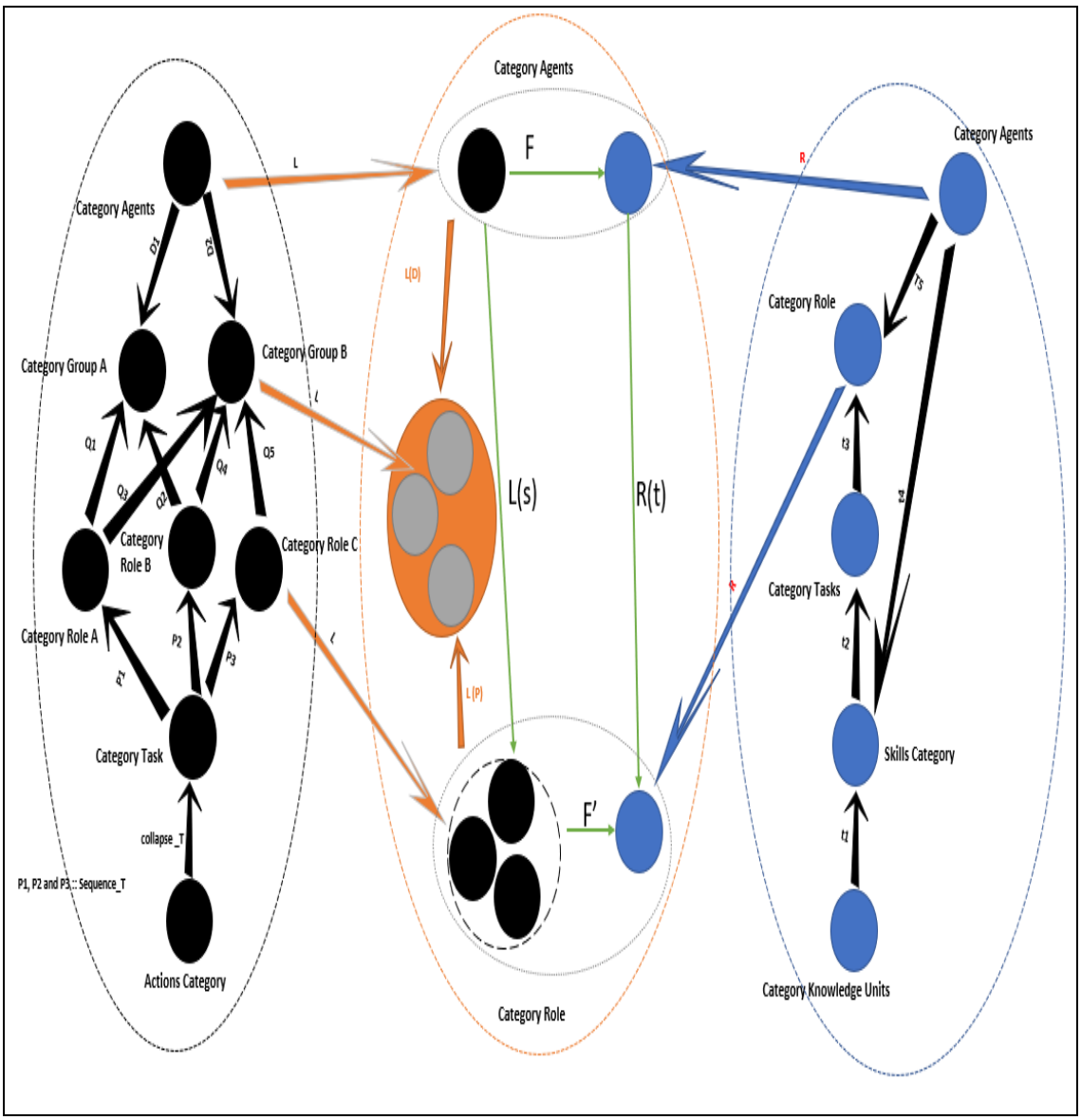

Figure 11. Composition of the two organizational models AGR and YAMAM with comma category

Explanation with comma category

The object (category agent) points to the object (CA) via L, and the object (Category Group B) points to the object (G), and the object (category agent) points to (category group B) which implies a morphism between the object (CA) and the object (G),
Object (category agent) points to object (CA) via R, and object (Category Group B) points to object (G), and object (category agent) points to (category group B) which implies a morphism between the object (CA) and the object (G),
The source is like a functor L: A→ C and whose target is like another functor R: B→ C. Comma category (L; R) has as objects, triplets of the form (a; f: L (a) → R (b); b) where a is an object of A and b an object of B. A morphism in (L; R), from (a; f; b) à (a'; f'; b') is an even pair of morphisms s: a→ a' and t: b→ b' such that the following square commutes:

5. CONCLUSIONS

In our study we are interested in the modeling of Organizational MASs and in these aspects, and compose tow Organizational MASs models (AGR and YAMAM), the complexity of this task requires a formalism that has a great power of expression, analysis and verification, category theory is better placed for this purpose, it is a very sophisticated toolbox for its graphical interpretation.

The advantages offered by category theory, first: visual formalism, mathematical support, modularity, hierarchical specification and easy description of systems, we can consider that it is the most suitable for modeling the dynamics of discrete systems, organizational aspects,

We will then focus on CT as a powerful graphical theory that allows the system to be easily understood and also allows the simulation of its activities,

This formalism is well suited to the formal verification of organizational MASs which contains both states and events represented by the interaction between the different Agents as well as between groups of agents and their roles in the organization.

Second: After the generation of the categorical model, the CT allows us to switch to several known modeling modes such as graphs or sets. This very important link represented by a functor (F) allows us to switch between the mathematical representation, which give us the possibility to reformulate the studied problem via graphs, sets, topos, and this allows us to use the characteristics and properties of each domain to solve the starting problem.

The use of CT in the composition of systems in general, and more specifically multi-agent organizational systems, as a new way of modeling them, opens the way perhaps to solving very complex problems in the future, by mastering the mathematical tools it offers.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

[1] Brazier, F.M., et al., DESIRE: Modelling multi-agent systems in a compositional formal framework. 1997. 6(01): p. 67-94.
[2] Argente, E., V. Julian, and V.J.E.N.i.T.C.S. Botti, Multi-agent system development based on organizations. 2006. 150(3): p. 55-71.
[3] Ferber, J. and A. Gutknecht. Un méta-modèle organisationnel pour l'analyse, la conception et l'execution de systèmes multiagents. in Proceedings of Third International Conference on Multi-Agent Systems ICMAS. 1998.
[4] Savall M., P.J., Chaignaud N., Itmi M., YAMAM, An organisation model for the multiagent systems, Implementation in the Phoenix Platform 3rd Francophone Conference of Modeling and Simulation "Conception, Analyze, Management of Industrial Systems" MOSIM
[5] in Conception, Analyze, Management of Industrial Systems" MOSIM. 2001: Troyes – France.
[6] El Habib Souidi, M., et al., Multi-agent pursuit-evasion game based on organizational architecture. 2019. 27(1): p. 1-11.
[7] Ferber, J., O. Gutknecht, and F. Michel. From agents to organizations: an organizational view of multi-agent systems. in International workshop on agent-oriented software engineering. 2003. Springer.
[8] Dorri, A., S.S. Kanhere, and R.J.I.A. Jurdak, Multi-agent systems: A survey. 2018. 6: p. 28573-28593.

[9] Souidi, M., et al., Coalition formation algorithm based on organization and Markov decision process for multi-player pursuit evasion. 2015. 11(1): p. 1-13.
[10] Souidi, M.E.H., et al., Multi-agent pursuit coalition formation based on a limited overlapping of the dynamic groups. 2019. 36(6): p. 5617-5629.
[11] Wirsing, M., Algebraic specification. Handbook of Theoretical Computer Science (J. van Leeuwen, ed.). 1990.
[12] Fong, B. and D.I.J.a.p.a. Spivak, Seven sketches in compositionality: An invitation to applied category theory. 2018.
[13] Abderrahim, S. and R. Maamri, A Category-theoretic Approach to Organization-based Modeling of Multi Agent Systems on the Basis of Collective Phenomena and Organizations in Human Societies. Informatica, 2018. 42(4).
[14] Awodey, S., Category theory. 2006: Oxford University Press. 2006, USA.
[15] Brunetto, G., Fusion d'entreprises et intégration des systèmes d'information. 2006.
[16] Bardis, P.D.J.S.S., Social interaction and social processes. 1979. 54(3): p. 147-167.
[17] Boudjidj, A., E. Merah, and M.E.H.J.I. Souidi, Towards a formal multi-agent organizational modeling framework based on category theory. 2021. 45(2).
[18] Awodey, S., Category theory. 2010: Oxford University Press.
[19] Nielsen, L.J.S.o.s.e.b.c., model-based techniques, and A.C.S. research directions, Peleska, 2015 Nielsen CB, Larsen PG, Fitzgerald J., Woodcock J., Peleska J. 2015. 48(2): p. 1-18.
[20] Keating, C., et al., System of systems engineering. 2003. 15(3): p. 36-45.
[21] Rydeheard, D.E. and R.M. Burstall, Computational category theory. Vol. 152. 1988: Prentice Hall Englewood Cliffs.

THE CHALLENGES OF INTERNET OF THINGS ADOPTION IN DEVELOPING COUNTRIES: AN OVERVIEW BASED ON THE TECHNICAL CONTEXT

Ayman Altameem

Department of Computer and Engineering Sciences, College of Applied Studies and Community Services, King Saud University, Riyadh, Saudi Arabia

ABSTRACT

The Internet of Things (IoT) has the potential to change the way we engage with our environments. Its prevalence has spread to various areas of industrial and manufacturing systems in addition to other sectors. However, many organizations are finding it increasingly difficult to navigate IoT. To unleash its full potential and create real economic value, it is essential to learn about the obstacles to IoT delivery. There is high potential for IoT implementation and usage in developing countries, and major barriers must be addressed for IoT delivery. This paper explores the challenges that impact the adoption of IoT in developing countries based on the technical context. It also presents a general conclusion in the form of recommendations to capture the maximum benefits of IoT adoption.

KEYWORDS

Internet of Things adoption, Obstacles of IoT in developing countries, IoT Technical Context.

1. INTRODUCTION

IoT offers tremendous potential to transform the globe by linking devices in large interoperable systems managed by analytics and software [1]. IoT is one of the most significant emerging technologies [2], [3]. It allows anyone in any location to connect to anything at any point in time using a device [4]. IoT has been recognized for its impacts on various sectors, including construction and manufacturing, healthcare, logistics, oil and gas, agriculture, and transportation. IoT system has the ability to empower all industries to change from conventional business models to new revenue streams [54]. Figure 1 illustrates the global IoT market share as determined by Fortune Business Insights [6], which divided the market into banking, financial services, and insurance (BFSI), transportation, healthcare, information technology (IT), telecom, manufacturing, government, agriculture, retail, sustainable energy, and other sectors. Accordingly, healthcare and manufacturing were anticipated to have the largest IoT market share in 2021 [6].

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 111-118, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121910

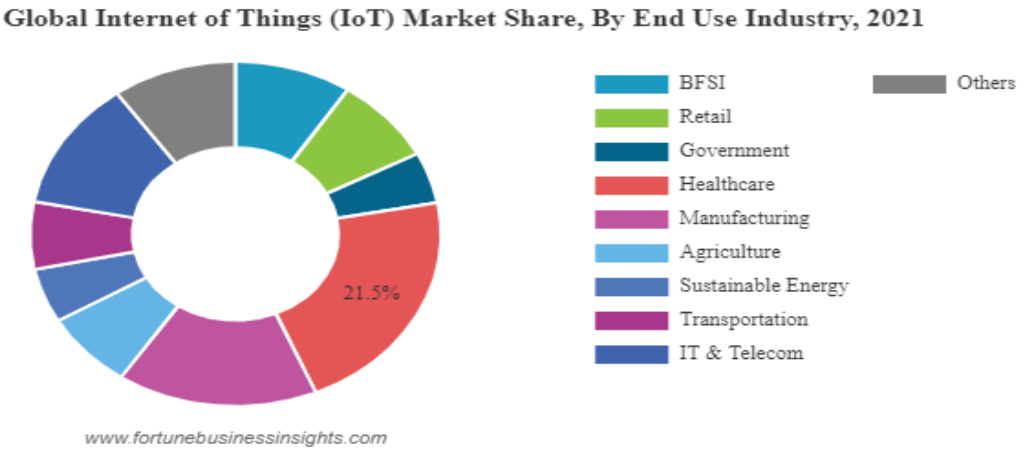

Figure 1. Global IoT market share [6]

The significant influence of IoT on the Internet and global economy is remarkable. It is anticipated that by 2025, there will be close to 100 billion linked IoT devices [7] and a world economic impact in excess of $11 trillion [8], which would equate to around 11% of the global economy based on the World Bank's projection of $99.5 trillion per year in global GDP in 2025 [8]. Figure 2 shows the anticipated rise in the enterprise IoT market [10], which rose by more than 22% between 2020 and 2021 to $157.9 billion. IoT Analytics estimates that the IoT market share will rise at a compound annual growth rate (CAGR) of 22.0% between 2022 and 2027 to eventually total $525 billion.

Customers will capture the utmost of the advantages. [8] showed that the users of IoT, including consumers, businesses, and other organizations, might be able to capitalize on 90% of the value that IoT applications produce. Accordingly, for example, the value of enhancing the health of chronic disease patients over remote monitoring will possibly be approximately $1.1 trillion each year in 2025 [8].

Organizations are utilizing IoT technology to increase efficiency and effectiveness [11], [44] in addition to enhancing decision-making [46], [47] and increasing the value of the business [39], [40], [47], [48]. The active nature and rapid changes in IoT have revealed obstacles and issues that might stand in the way of allowing users to capture its advantages [12]. Thus, the process of understanding the major challenges that influence the adoption of IoT is a very important endeavor.

A further in-depth overview is needed to map the obstacles associated with the implementation and adoption of IoT in developing countries. Experts estimate that by 2025, 40% of the economic value added from IoT will be generated in the developing world [13].

The aim of this paper is to explore the major challenges to implementing IoT in developing countries based on the technical context. This paper provides an inclusive outline to aid in further research in this field. Tackling these challenges will help organization leaders and IT professionals in developing countries take an efficient course of action during IoT delivery.

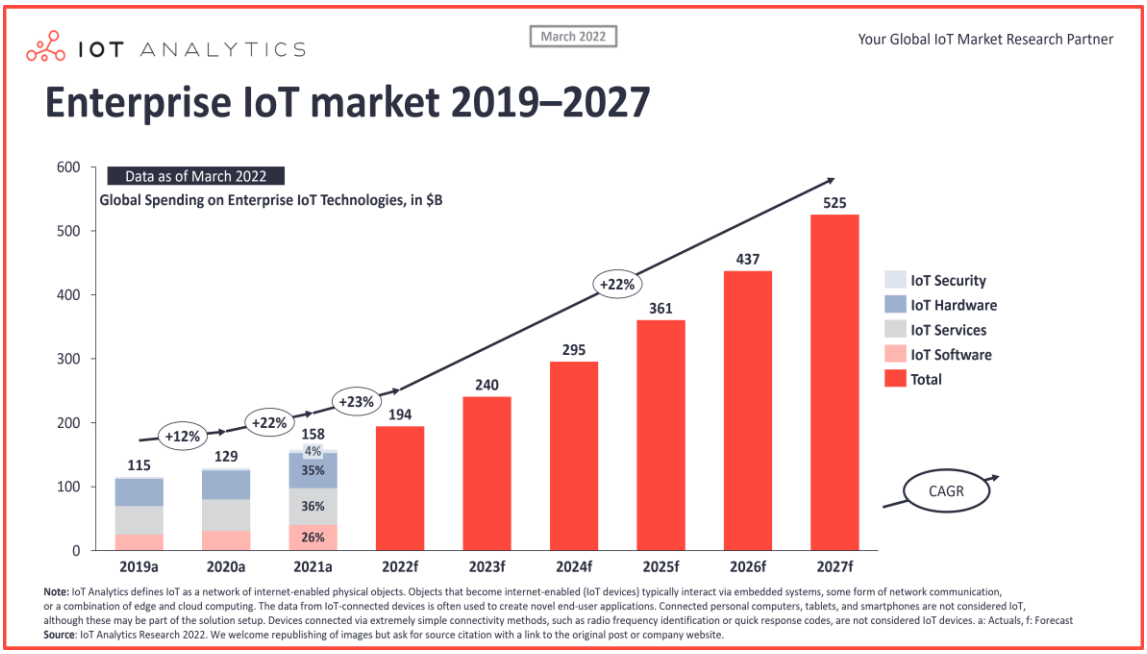

Figure. 2. Growth in the enterprise IoT market [10]

2. INTERNET OF THINGS (IOT)

Researchers have offered several definitions of IoT [12]. [14] defined it as a network of devices linked with electronics, software, sensors, and network connectivity. [15] viewed IoT as "an internetworking of physical objects such as sensors, actuators, personal computers, software, intelligent devices, automobile, and network connectivity that enable them to collect and exchange data without human intervention." According to [16], IoT is a network of connected devices that are uniquely addressable based on standard communication protocols. [17] described IoT as a technology that enables individuals and things to be linked at any place and anytime, linking any service with anyone, and ideally using any path or network. Furthermore, [49] defined IoT as "a network that connects an ordinary physical object with an identifiable address to provide intelligent services." According to the Gartner Group, IoT "is the network of physical objects that contain embedded technology to communicate and sense or interact with their internal states or the external environment" [18].

3. CHALLENGES OF IOT

IoT has been recognized as an innovative technology, and its development has attracted great interest from many sectors worldwide [19]. Despite the technology's many potential benefits, including improving our quality of life [50], organizations in developing countries have encountered several obstacles to IoT adoption. Studies on IoT have addressed the following challenges based on the technical context: access control [20], [21], security [12],[24]–[29], [34],[36],[38], privacy [12], [27], [30]–[34],[36], IoT infrastructure [13], [35], [37], energy requirement [36], [37], [45], IoT expertise [13], [37], compatibility [31], [37], complexity [31], [34], [44], and connectivity [41]–[43]. These IoT challenges are illustrated in Figure 3.

3.1. Security and Privacy

Many researchers have emphasized security and privacy as major challenges to IoT delivery. [9] pointed out that IoT presents opportunities for hackers and has been associated with new security risks that application developers and device manufacturers cannot predict [55]. There are many threats that can impact IoT entities, such as attacks that target various communication channels, identity fabrication, denial of service, and physical threats [56]. End-user privacy can be threatened due to their restricted control and options over the collection, retention, and distribution of their data [55].

3.2. IT Infrastructure

The flexibility of IT infrastructure allows organization leaders to be the most advantaged by an IT system, since it can react to new developments more efficiently [57]. The lack of IT infrastructure could put countries at risk of being left behind in economic terms [58]. IT infrastructure is one of the IoT requirements. Many researchers [59] have noted that organizations could have problems adopting IoT systems due to their lack of IT infrastructure.

3.3. Power Requirement

In addition, reliable power resources are vital to powering IoT in many developing countries [45]. For most objects, energy is crucial. Sometimes, a lack of energy can even limit the lifespan of an object [51]. A stable and reliable power supply is vital to the enabling of these systems, as it is necessary for constant operation over a period of several months to years [52].

3.4. Compatibility

Several researchers have recognized the importance of compatibility in IoT adoption [31], [37]. In the prediction of communication-oriented services, perceived compatibility has been recognized as an important issue in determining a user's adoption of such services [53].

3.5. Complexity

Complexity leads to greater difficulty in the deployment of IoT applications. [60] pointed out that "The complexity of an innovation may be determined by the breadth and depth of knowledge required, and it acts as a barrier to potential adopters of IS innovation." A simplified implementation mechanism and ease of use of the technology are essential to the successful adoption of IoT applications.

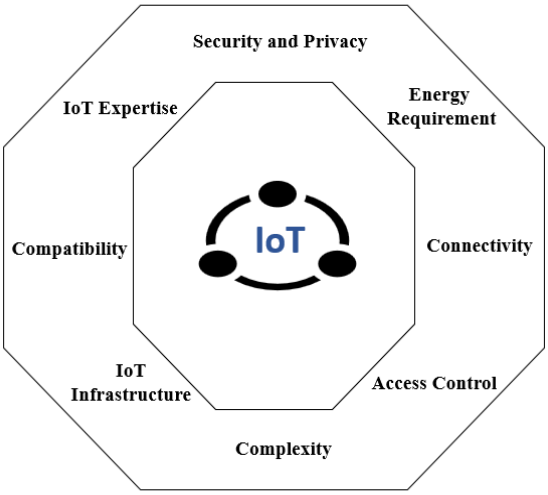

Figure. 3. Challenges of IoT

3.6. IoT Expertise

IT professionals are vital to the continuous development of IoT systems [61]. IoT systems require day-to-day maintenance and updates. Sufficient IT skills are required to increase its likelihood of implementation. [62] noted that the competition for IoT is the competition between technology and professionals.

3.7. Access Control

Access control is another aspect of great significance and sensitivity [63]. [5] basically defined access control systems as "software that is used to control access to files, records, etc." Access is permissible if it fulfils the rules related to the data [64].

3.8. Connectivity

In developing countries, one of the critical technological challenges is providing users with a sufficient Internet speed. IoT demands both scalability and reliable connectivity. Internet connectivity can be either an important barrier to or an enabler in the implementation of IoT.

Organizations must recognize and understand the complex realities inherent in the IoT adoption process. Only with such understanding can they develop the right methods, tools, and solutions to surmount these challenges and derive the maximum benefits from IoT.

4. CONCLUSIONS

As one of the most advanced emerging technologies, IoT is altering many industries and economies. It has a high potential to connect everyday objects to the Internet. Although IoT provides various benefits that improve our quality of life, there are major obstacles to its delivery that should be addressed. This paper examined the major IoT challenges in developing countries based on the technical context. This study presents an inclusive outline to aid further research in this area. This paper found that access control, security, privacy, IoT infrastructure, energy requirements, IoT expertise, compatibility, complexity, and connectivity are key barriers that impact IoT adoption. This paper can aid both practitioners and researchers. For practitioners, this

paper addressed the major obstacles that impact successful IoT adoption in developing countries. Organization leaders and IT professionals should take these obstacles into consideration so that they can take an efficient course of action when facilitating IoT delivery. As for researchers, this paper is a useful reference for further research in this area. Future studies can be conducted to validate these results by developing a tool and taking a survey of organizations.

REFERENCES

[1] Tektronix, (2016), A Guide To Building IoT Ready Devices, Available online: https://download.tek.com/document/37W_60226_2_IoT_eBook.pdf (accessed on 1 June 2022)

[2] Kassab, M., DeFranco, J., and Laplante, P, (2020), A systematic literature review on Internet of things in education: Benefits and challenges, Journal of Computer Assisted Learning, 36(2), pp.115-127

[3] Songsom, N., Nilsook, P., and Wannapiroon, P, (2019), The student relationship management system process via the Internet of things, TEM Journal, 8(4), pp. 1426-1432

[4] Zheng, J., Simplot-Ryl, D., Bisdikian, C., and Mouftah, H, (2011), The Internet of Things, IEEE Communications Magazine, 49(11), 30-31.

[5] Herold, R. (2007), Information Security Management Handbook (6th Ed.), Auerbach Publications, Canada, US.

[6] Fortune Business Insights, 2022, Internet of Things (IoT) Market, Available online: https://www.fortunebusinessinsights.com/industry-reports/internet-of-things-iot-market-100307(accessed on 8 Sep 2022).

[7] "Global Connectivity Index." Huawei Technologies Co., Ltd., 2015, Available online: http://www.huawei.com/minisite/gci/en/index.html (accessed on 6 Sep 2015).

[8] Manyika, James, Michael Chui, Peter Bisson, Jonathan Woetzel, Richard Dobbs, Jacques Bughin, and Dan Aharon, "The Internet of Things: Mapping the Value Beyond the Hype." McKinsey Global Institute, June 2015.

[9] Yaqoob, I., Hashem, I. A. T., Ahmed, A., Kazmi, S. A., & Hong, C. S. (2019). Internet of things forensics: Recent advances, taxonomy, requirements, and open challenges. Future Generation Computer Systems, 92, 265-275.

[10] Wegner, P., 2022, Global IoT market size grew 22% in 2021 — these 16 factors affect the growth trajectory to 2027, IoT Analytics' Global IoT Enterprise Spending Dashboard, Available online: https://iot-analytics.com/iot-market-size/ (accessed on 8 Sep 2022).

[11] F. Restuccia, S. D. Oro, and T. Melodia, "Securing the Internet of Things in the Age of Machine Learning and Software-defined Networking," vol. 1, no. 1, pp. 1–14, 2018.

[12] Nord, J.H.; Koohang, A.; Paliszkiewicz, J, (2019), The IoT: Review and theoretical framework, Expert Syst. Appl, 133, 97–108.

[13] T. Scherf, (2016), Internet of Things—Hype or hope for developing countries? KfW Development Research, Available online: https://www.kfw-entwicklungsbank.de/PDF/DownloadCenter/PDF-Dokumente-Development-Research/Internet-of-Things-%E2%80%93-hype-or-hope-for-developing-countries.pdf(accessed on 5 Sep 2022).

[14] D. Mocrii, Y. Chen, P. Musilek, IoT-based smart homes: a review of system architecture, software, communications, privacy and security, Internet Things 1–2 (2018) 81–98, doi:10.1016/j.iot.2018.08.009.

[15] Amiruddin, A., Ratna, A. A. P., and Sari, R. F. (2019), Systematic review of Internet of things security, International Journal of Communication Networks and Information Security, 11(2), 248-255.

[16] L. Atzori, A. Iera, G. Morabito, The internet of things: a survey, Comput. Netw, 54 (2010) 2787–2805, doi:10.1016/j.comnet.2010.05.010

[17] H. Sundmaeker, P. Guillemin, P. Friess, S. Woelfflé, Vision and challenges for realizing the internet of things, Clust. Eur. Res. Proj. Internet Things Eur. Commision, (2010), doi:10.2759/26127

[18] Gartner Group, Internet of Things (IoT), Gartner Glossary, Available online: https://www.gartner.com/en/information-technology/glossary/internet-of-things (accessed on 19 Sep 2022).

[19] A. Al-fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, (2015), "Internet of Things: A Survey on Enabling Technologies, Protocols and Applications Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," vol. 17, no. January, pp. 2347–2376.
[20] Mishra, R., and Yadav, R. (2020), Access Control in IoT Networks: Analysis and Open Challenges, Available at SSRN 3563077.
[21] A. Ouaddah, H. Mousannif, A.A. Elkalam, A.A. Ouahman, (2017), Access control in the internet of things: big challenges and new opportunities, Comput. Netw., 112 (2017), Pages 237-262
[22] Pereira, C., Aguiar, A. (2014), Towards efficient mobile M2M communications: survey and open challenges, Sensors, 14(10), 19582-19608.
[23] Huo, L., and Wang, Z. (2016), Service composition instantiation based on cross-modified artificial Bee Colony algorithm, China Communication, 13(10), 233-244.
[24] Patnaik, R., Padhy, N., and Raju, K.S. 2021, A Systematic Survey on IoT Security Issues, Vulnerability and Open Challenges. In Intelligent System Design, (pp. 723-730). Springer, Singapore.
[25] M. M. Hossain, M. Fotouhi, and R. Hasan, "Towards an Analysis of Security Issues, Challenges, and Open Problems in the Internet of Things," 2015 IEEE World Congr. Serv., pp. 21–28, 2015.
[26] Q. Jing, A. V. Vasilakos, J. Wan, J. Lu, and D. Qiu, "Security of the Internet of Things: perspectives and challenges," Wirel. Networks, vol. 20, no. 8, pp. 2481–2501, 2014.
[27] Shafagh, H., Burkhalter, L., Hithnawi, A., and Duquennoy, S. (2017). "Towards blockchain-based auditable storage and sharing of IoT data," in Proceedings of the 2017 on Cloud Computing Security Workshop (New York), 45–50. doi: 10.1145/3140649.3140656
[28] Khan, A., and Salah, K., (2018), IoT security: Review, blockchain solutions, and open challenges. Future Generation Computer Systems, Vol. (82), pp.395-411.
[29] S. A. Kumar, T. Vealey, and H. Srivastava, "Security in the internet of things: Challenges, solutions, and future directions," Proc. Annu. Hawaii Int. Conf. Syst. Sci., vol. 2016–March, pp. 5772–5781, 2016.
[30] Ng, I., and Wakenshaw, S., (2017), The IoT: Review and research directions, International Journal of Research in Marketing, Vol.34,No.1,pp.3-21.
[31] T. Wansinghe, R. Gosine, L. James, G. Mann, O. de Silva and P. Warian, (2020),The IoT in the Oil and Gas Industry: A Systematic Review, in IEEE IoT Journal, vol.7, No.9, doi: 10.1109/JIOT.2020.2995617.
[32] S. Birkel and E. Hartmann, (2019), Impact of IoT challenges and risks for SCM, Supply-Chain-Management, ISSN: 1359-8546
[33] Weber, H., (2010), Internet of Things – New security and privacy challenges. Computer Law and Security Review, Vol.26,pp.23-30.
[34] L., Da Xu, W. He, and S. Li, (2014),IoT in industries: A survey, IEEE Transactions on Industrial Informatics,Vol.10, No.4,pp.2233-2243.
[35] Laboratoire International de Recherche en Informatique et Mathématiques Appliquées, Available online: Available: https://lirima.inria.fr/fr/focus-on-a-joint-project-team-iot4d/ (accessed on 9 Sep 2022).
[36] Y. Yang, L. Wu, G. Yin, L. Li & H. Zhao, (2017), A Survey on Security and Privacy Issues in IoT, in IEEE Internet of Things Journal, Vol.4, No.5, pp.1250-1258.
[37] Sachin S., Angappa G., Harsh P., Sudhanshu J., Modeling the IoT adoption barriers in food retail supply chains, Journal of Retailing and Consumer Services, ISSN 0969-6989, doi.org/10.1016/j.jretconser.2019.02.020.
[38] M., Ahlmeyer, and M., Chircu, (2016),Securing the IoT:A review. Issues in Information Systems,Vol.17,No.4,pp.21-28.
[39] A., Karkouch, H., Mousannif, H., Al Moatassime, and T., Noel, (2016), Data Quality in IoT: A State of the Art Survey, JNCA, ISSN 1084-8045, doi.org/10.1016/j.jnca.2016.08.002.
[40] N., Côrte, P., Ruivo, & T., Oliveira, (2020), Leveraging IoT and Big Data Analytics Initiatives in European and American Firms, Information and Management, ISSN 0378-7206, doi.org/10.1016/j.im.2019.01.003.
[41] K. Kinder, "The Societal Impact of the Internet of Things," 2013.
[42] J. Ding, M. Nemati, C. Ranaweera and J. Choi, "IoT Connectivity Technologies and Applications: A Survey," in IEEE Access, vol. 8, pp. 67646-67673, 2020, doi: 10.1109/ACCESS.2020.2985932.
[43] S. S. I. Samuel, "A review of connectivity challenges in IoT-smart home," 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), 2016, pp. 1-4, doi: 10.1109/ICBDSC.2016.7460395.

[44] T., Lynn, P., Rosati, & T. Endo, (2018), Toward the Intelligent Internet of Everything: Observations on Multidisciplinary Challenges in Intelligent Systems Research, In: Picazo-Vela, S., Hernández L. R. (eds.) Technology, Science, and Culture: A Global Vision, vol. 116, pp. 52–64.
[45] M. N. S. Miazi, Z. Erasmus, M. A. Razzaque, M. Zennaro, and A. Bagula, ‘‘Enabling the Internet of things in developing countries: Opportunities and challenges,’’ in Proc. 5th Int. Conf. Informat., Electron. Vis. (ICIEV), Dhaka, Bangladesh, May 2016, pp. 564–569.
[46] Rey, A.; Panetti, E.; Maglio, R.; Ferretti, M. Determinants in adopting the Internet of Things in the transport and logistics industry. J. Bus. Res. 2021, 131, 584–590.
[47] P. Rosati, and T. Lynn, (2020), Mapping the Business Value of the Internet of Things. In: Lynn, T., Mooney, J., Lee, B., Endo, P. (eds) The Cloud-to-Thing Continuum. Palgrave Studies in Digital Business and Enabling Technologies. Palgrave Macmillan, Cham. doi.org/10.1007/978-3-030-41110-7_8
[48] P. Rosati, G. Fox, D. Kenny, and T. Lynn., (2017), Quantifying the Financial Value of Cloud Investments: A Systematic Literature Review. 2017 IEEE International Conference on Cloud Computing Technology and Science, pp.194–201.
[49] H.-D. Ma, “Internet of things: Objectives and scientific challenges,” Computer Science and Technology, Springer, vol. 26, no. 6, 2011
[50] J. P. Meltzer, (2019), “Global Views Globalviews Artificial intelligence primer:,” no. 12, 2019.
[51] Dorsemaine, B., Gaulier, J.-P., Wary, J.-P., Kheir, N., & Urien, P. (2015). Internet of Things: A Definition & Taxonomy. 2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies. doi:10.1109/ngmast.2015.71
[52] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, “Internet of things: Vision, applications and research challenges,” Ad Hoc Netw., 10(7), 1497 (2012).
[53] A. H. Crespo, M. M. G. de los Salmones, I. R. del Bosque, ´ et al., “Influence of users perceived compatibility and their prior experience on b2c e-commerce acceptance,” in Electronic Business and Marketing. Springer, 2013, pp. 103–123.
[54] Rybakov, A. (2021), Applying IoT to transform business: top-10 cases, AgileVision, Available online: https://www.agilevision.io/blog/applying-iot-to-transform-business-top-10-cases (accessed on 28 Sep 2022).
[55] Maras, M.-H. (2015). Internet of Things: security and privacy implications. International Data Privacy Law, 5(2), 99–104. doi:10.1093/idpl/ipv004
[56] S. Babar, P. Mahalle, A. Stango, N. Prasad, R. Prasad, Proposed Security Model and Threat Taxonomy for the Internet of Things (IoT), 3rd International Conference on Recent Trends in Network Security and Applications, Chennai, India, 2010, pp.420–429.
[57] Chanopas, A., Krairit, D. and Khang, D. B. (2006). “Managing information technology infrastructure: a new flexibility framework”. Management Research News, Vol. 29, No. 10, pp. 632-651.
[58] Lund S. and Manyika, J., (2016), ``How digital trade is transforming globalisation," The E15 Initiative, Int. Centre Trade Sustain. Develop. (ICTSD) World Econ. Forum, Geneva, Switzerland, Tech. Rep. [Online].Available:http://www.e15initiative.org/andhttp://e15initiative.org/publications/how-digital-trade-is-transforming-globalisation/
[59] I. Ehie, and M. Chlton, (2020), Understanding the influence of IT/OT Convergence on the adoption of IoT in manufacturing organisations, Computers in Industry, Vol.115, ISSN 0166-3615.
[60] Lee, S. and Kim, K. J. (2007) “Factors affecting the implementation success of Internet-based information systems”. Computers in Human Behavior, Vol. 23, No. 4, pp. 1853-1880.
[61] Yan Yu, Jianhua Wang, & Guohui Zhou. (2010). The exploration in the education of professionals in applied Internet of Things Engineering. 2010 4th International Conference on Distance Learning and Education. doi:10.1109/icdle.2010.5606038
[62] Yu Zhongcheng. The Internet of Things, the Next War We Can't Afford to Lose [J].China Computer & Communication,20 10(3)44-46.
[63] Botha, R. A. and Eloff, J. H. P. (2001), “A framework for access control in workflow systems”, Information Management & Computer Security, 9/3, pp. 126-133.
[64] Blobel, B., Nordberg, R., Davis, J. M. and Pharow, P. (2006), “Modelling privilege management and access control”, International Journal of Medical Informatics, 75, pp. 597-623

SCREENING DEEP LEARNING INFERENCE ACCELERATORS AT THE PRODUCTION LINES

Ashish Sharma, Puneesh Khanna, Jaimin Maniyar

AI Group, Intel, Bangalore, India

ABSTRACT

Artificial Intelligence (AI) accelerators can be divided into two main buckets, one for training and another for inference over the trained models. Computation results of AI inference chipsets are expected to be deterministic for a given input. There are different compute engines on the Inference chip which help in acceleration of the Arithmetic operations. The Inference output results are compared with a golden reference output for the accuracy measurements. There can be many errors which can occur during the Inference execution. These errors could be due to the faulty hardware units and these units should be thoroughly screened in the assembly line before they are deployed by the customers in the data centre.

This paper talks about a generic Inference application that has been developed to execute inferences over multiple inputs for various real inference models and stress all the compute engines of the Inference chip. Inference outputs from a specific inference unit are stored and are assumed to be golden and further confirmed as golden statistically. Once the golden reference outputs are established, Inference application is deployed in the pre- and post-production environments to screen out defective units whose actual output do not match the reference. Strategy to compare against itself at mass scale resulted in achieving the Defects Per Million target for the customers

KEYWORDS

Artificial Intelligence, Deep Learning, Inference, Neural Network Processor for Inference (NNP-I), ICE, DELPHI, DSP, SRAM, ICEBO, IFMs, OFMs, DPMO.

1. INTRODUCTION

Artificial Intelligence (AI) is growing in recent years and is expected to be re-shaping industries. Deep Neural networks are in the heart of the AI revolution. Deep Neural Network is composed of layers of simulated neurons with different connectivity schemes. The new computation model is based on massive parallel execution of linear algebra operations. New dedicated architectures that are optimized for Deep Learning execution is required to achieve high efficiency and to meet the market requirements. Deep learning inference accelerators are designed specifically to deliver superior performance, low latency, power efficiency and cost savings for cloud, data centres and other emerging applications.

SpringHill (NNP-I) is an Inference Chip from Intel which is used to accelerate execution of the arithmetic operations. It consists of 12 Inference Compute Engines (ICE) as described in Figure 1. Each ICE contains hardware accelerator IPs DELPHI and DSP. The operations that are supported by DELPHI includes direct CNNs (convolution Neural Network) as well as GEMM (General Matrix Multiplication), nonlinear activation, quantization, and pooling operations. The ICE core is highly programmable and integrates a strong VLIW vector Tensilica DSP as

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 119-126, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121911

described in Figure 2. This allows a variety of operators that are not accelerated by the DL accelerator to be mapped to the DSP and executed in high efficiency. In addition, the ICE includes dedicated memory access blocks: a dedicated Deep Learning DMA (DSE) with dedicated features such as 4D walks, padding and stride; Compression/de-compression engine and MMU (Memory Management Unit). Each ICE has a large local SRAM of size 4MB to store the persistent data. A pair of ICE units are connected via ICEBO which allows the ICEs to share the data with each other. All the ICEs share LLC cache of size 24 MB.

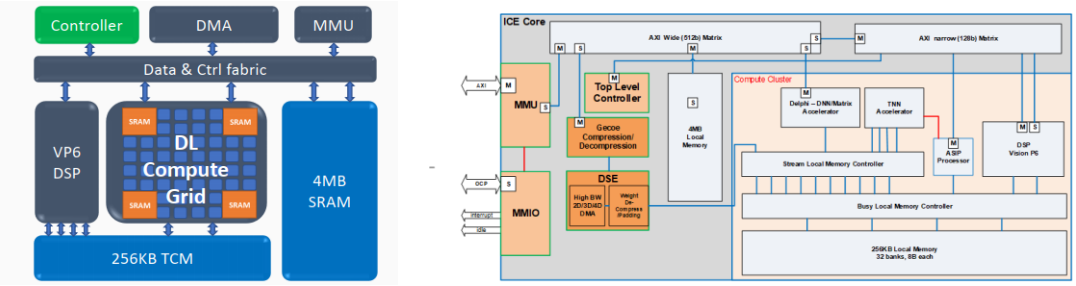

Figure 1. Inference Compute Engine (ICE)

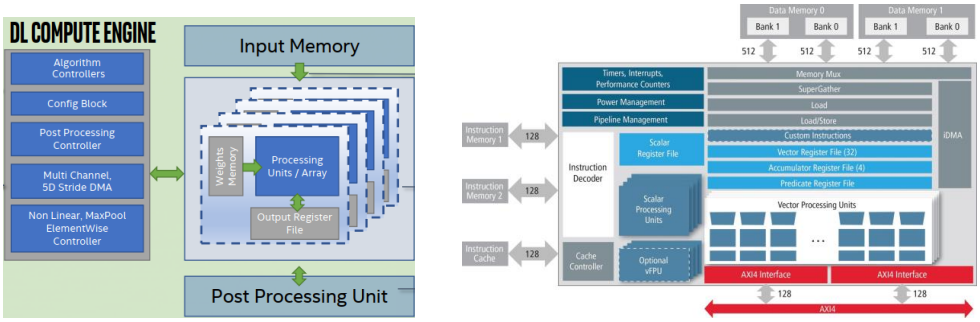

Figure 2. DL Compute Engine and Vector Processing Unit (DSP)

The NNP-I card is connected to the host processor using the PCI. The Inference applications run on the host. The model is loaded onto the card memory along with the inputs and weights. The inference outputs are transferred back to the host and subsequently to the Inference application. The inference accuracy for the same inputs for different inferences must remain the same. The inference application also collects the error statistics from the card.

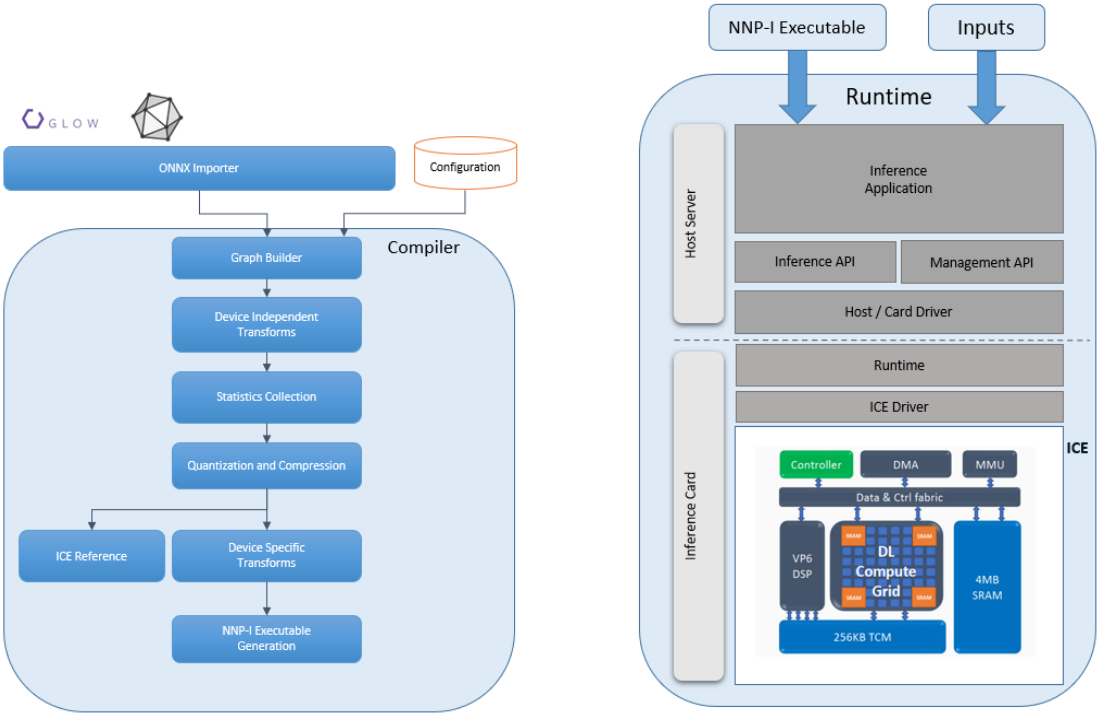

Figure 3. NNPI Software Stack

The NNPI Software stack described in Figure 3 includes the Graph Compiler, Platform/Drivers. The models are compiled using the Graph Compiler which generates the recipe. The recipe is passed to the Inference Application. The Inference Application runs on the host server and interfaces with the card using the platform drivers. The recipe gets executed on the ICEs and the final output is shared back to the Inference Application. The outputs received after the Inference are compared with the expected outputs (software reference) and the Inference efficiency/accuracy is measured.

There are different kinds of errors which can occur during the inference execution due to the hardware issues. These include Byte Mismatch (Computational) errors, Deep SRAM errors, ECC errors, PCIe AER errors, parity errors. The cards should be thoroughly screened at the factory with the help of a dedicated test content so that the faulty cards are not delivered to the customers.

The silicon units are screened at the fabrication assembly line, followed by the screening at the card manufacturer before sending it to the customer ODM. The customer ODM runs extensive regression tests before deploying these cards in the data centre. The customer also keeps on running the periodic sanity tests on the cards to check the health of these cards.

2. SCREENING PROCESS

2.1. Hardware Screener Inference Application

The Hardware Screener Application has been developed based upon the Inference APIs and Management APIs exposed by the NNPI software. The Application creates an Inference context and for each such context a Runtime process instance gets created on the device to run inferences. The Inference library parses the recipe generated by the graph compiler and then allocates

memory buffers on the host/device(s) and pass this information to the Runtime/ICE driver framework. The ICE Driver framework schedules the different networks on the different ICE units. Multiple threads are created in each context to extract the maximum utilization of the ICEs.

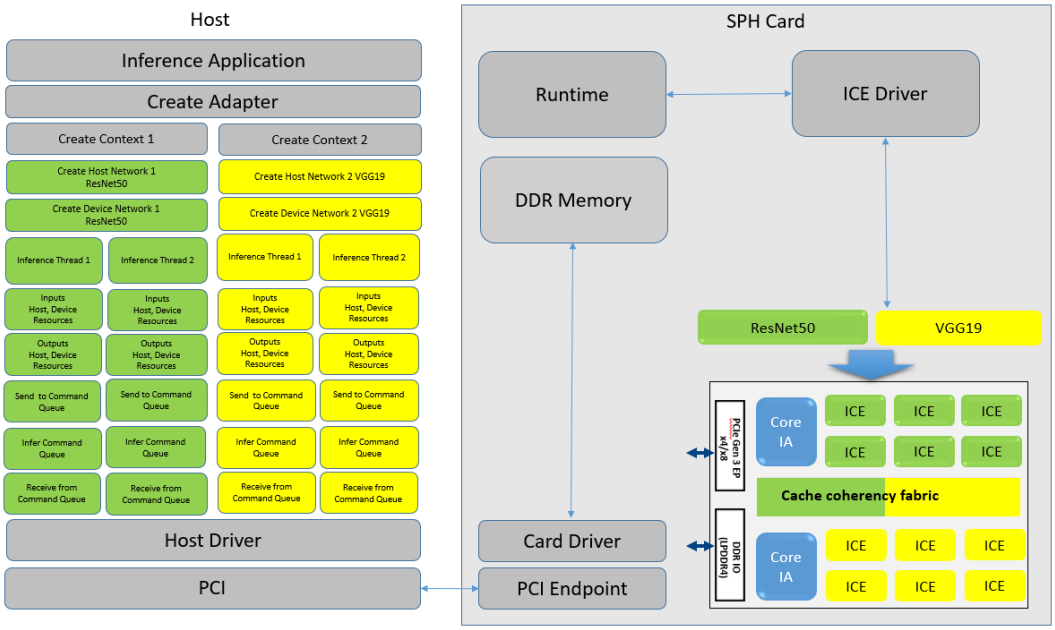

Figure 4. Hardware Screening Application

The Hardware Screening Application described in the Figure 4 is used to infer 50K images from the ImageNet dataset for different models/workloads. This application generates the IFMs (Input Feature Maps) and the OFMs (Output Feature Maps) on one hardware unit. The Inference is executed on multiple hardware units and the outputs are compared with the saved OFMs from the first hardware unit. If the outputs match, then these set of IFMs and OFMs would be used as reference and this package is deployed at the assembly line to screen the units.

The other errors like ECC, DSRAM, MCE, PCIe are identified using the Software counters incremented during the Inference execution. These software counters are being monitored by the Hardware Screener Application using the Management APIs.

The detailed screening flow is described in the Figure 5.

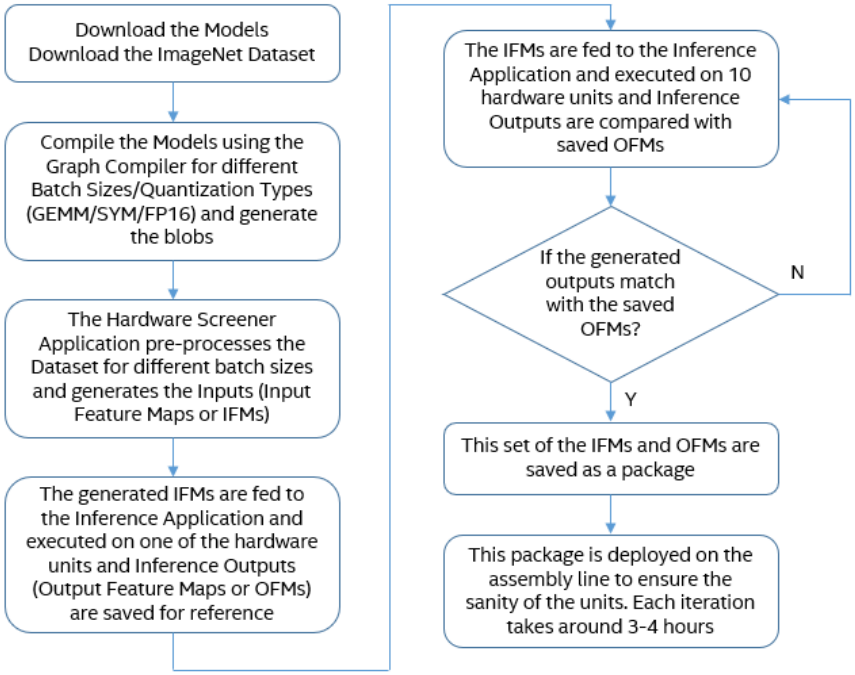

Figure 5. Flowchart of screening process

The AI workloads configurations are described in the Table 1. The workloads are generated for different Quantization Types, Batch Sizes, Number of ICEs. The other parameters include the Persistent Deep SRAM, Shared Read (ICEBO), Class of Service, ICEDC Sync (Barrier/Fence). This test content stresses upon all the different execution flows within the ICEs and helps in determining the errors.

Table 1. AI Workloads and Configurations

Workloads	Quantization Type	Batch Size	Number of ICEs	Persistent Deep SRAM	Shared Read (ICEBO)	Class of Service (CLOS)	ICEDC Sync (Barrier / Fence)	Number of Inputs	Execution Time
ResNet50	GEMMLOWP	12	12	N	Y	Y	Y	50K	30 secs
ResNet50	SYMLOWP	12	12	N	Y	Y	Y	50K	30 secs
ResNet50	GEMMLOWP	2	2x6	N	Y	Y	Y	50K	150 secs
ResNet50	SYMLOWP	2	2x6	N	Y	Y	Y	50K	150 secs
ResNet50	FP16	2	2x6	N	Y	Y	Y	50K	900 secs
ResNet50	GEMMLOWP	1	1x12	Y	N	N	N	50K	350 secs
ResNet50	SYMLOWP	1	1x12	Y	N	N	N	50K	350 secs
ResNet50	FP16	1	1x12	Y	N	Y	Y	50K	2000 secs
ResNext	GEMMLOWP	12	12	N	Y	Y	Y	50K	60 secs
ResNext	SYMLOWP	12	12	N	Y	Y	Y	50K	60 secs
ResNext	GEMMLOWP	2	2x6	N	Y	Y	Y	50K	300 secs
ResNext	SYMLOWP	2	2x6	N	Y	Y	Y	50K	300 secs
ResNext	GEMMLOWP	1	1x12	Y	N	N	N	50K	1000 secs
ResNext	SYMLOWP	1	1x12	Y	N	N	N	50K	1000 secs
ShuffleNet	GEMMLOWP	12	12	N	Y	Y	Y	50K	20 secs
ShuffleNet	SYMLOWP	12	12	N	Y	Y	Y	50K	20 secs
ShuffleNet	GEMMLOWP	2	2x6	N	Y	Y	Y	50K	120 secs
ShuffleNet	SYMLOWP	2	2x6	N	Y	Y	Y	50K	120 secs
ShuffleNet	GEMMLOWP	1	1x12	Y	N	N	N	50K	220 secs
ShuffleNet	SYMLOWP	1	1x12	Y	N	N	N	50K	220 secs
ShuffleNet	FP16	1	1x12	N	Y	Y	Y	50K	1350 secs
ResNet50	GEMMLOWP	1	1x12	Y	N	N	N	50K	350 secs
ResNet50	SYMLOWP	1	1x12	Y	N	N	N	50K	350 secs
ResNet50	SYMLOWP	12	12	Y	N	N	N	50K	30 secs
VGG19	GEMMLOWP	1	1x12	Y	N	N	N	50K	2400 secs
VGG19	SYMLOWP	12	12	Y	N	N	N	50K	180 secs

VGG19	GEMMLOWP	12	12	Y	N	N	N	50K	180 secs
ResNet50	GEMMLOWP	12	12	N	Y	Y	Y	50K	300 secs
ShuffleNet	GEMMLOWP			Y	N	N	N		
ResNet50	GEMMLOWP	2	2x3	N	Y	Y	Y	50K	300 secs
ShuffleNet	GEMMLOWP			N	Y	Y	Y		
ResNet50	GEMMLOWP	2	2x2	N	Y	Y	Y	50K	500 secs
ResNext	GEMMLOWP			N	Y	Y	Y		
ShuffleNet	GEMMLOWP			N	Y	Y	Y		
Total Time									4 hours

2.2. Hardware Screening at Assembly Lines

The hardware units are screened at different stages using the Hardware Screener Application as shown in the Figure 6.

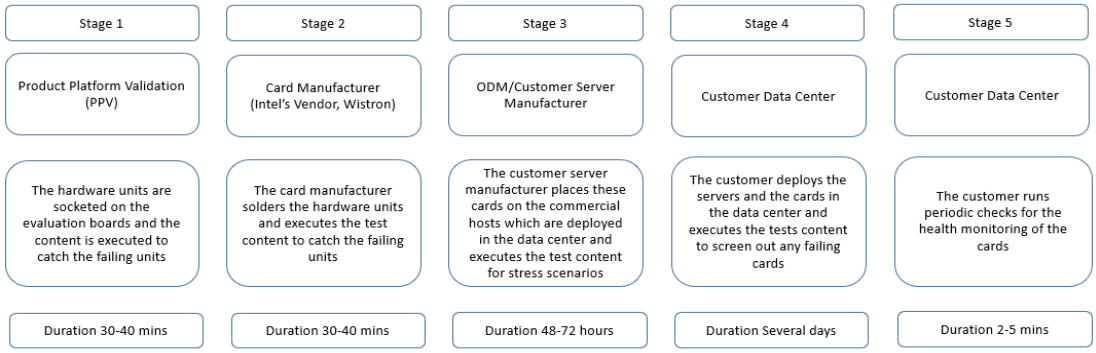

Figure 6. Stages of Deployment

The different error types which lead to failing cards are described in the Table 2.

Table 2. Error Types and their description

Error Type	Description
Byte Mismatch (Computational Errors)	The Inference output should match byte by byte with the golden reference output. If these outputs differ, then it leads to the byte mismatch errors.
ECC Errors	Error correction code memory (ECC memory) uses an error correction code (ECC) to detect and correct n-bit data corruption which occurs in memory. The ECC errors are also divided into the correctable and non-correctable errors.
Deep SRAM Errors	DSRAM arrays are the densest circuitry on a chip. The error correction logic is also added with the SRAMs to detect/correct the DSRAM errors. If the DSRAM errors are not corrected, then it leads to DSRAM single/multi bit errors.
PCI Express Advanced Error Reporting (AER)	PCI Express errors can occur on the PCI Express link itself or on behalf of transactions initiated on the link. The PCI AER errors are divided into the correctable errors and uncorrectable errors. Uncorrectable errors can cause a particular transaction or a particular PCI Express link to be unreliable.
Parity Errors / Memory Check Errors (MCE)	Parity error is an error that results from irregular changes to data, as it is recorded when it is entered in memory. Different types of parity errors can require the retransmission of data or cause serious system errors, such as system crashes.
VMin Errors	VMin errors are seen due to the voltage sensitivity at various operating frequencies of the chip. This also leads to the byte mismatch or the computational errors.

The following table shows the data for the 20 bad units which have been screened using the Hardware Screener application. The errors have been detected in the first iteration using the respective workloads.

Table 3. Units detected using Hardware Screening Application

SNo	Workload	Quantization Type	Batch Size	No of ICEs	No of errors	Failed Iteration	Error Type
1	ResNext	GEMMLOWP	2	2	2	1	Bytes Mismatch
2	ResNext	SYMLOWP	1	1	1	1	DSRAM Error
3	ResNext	SYMLOWP	1	1	35	1	Bytes Mismatch
4	ResNext	GEMMLOWP	2	2	4	1	Bytes Mismatch
5	ResNext	SYMLOWP	1	1	6	1	Bytes Mismatch
6	ResNet50	SYMLOWP	1	1	22	1	Bytes Mismatch
7	ShuffleNet	SYMLOWP	1	1	11	1	Bytes Mismatch
8	ShuffleNet	GEMMLOWP	2	2	3	1	MCE Error
9	ResNet50	GEMMLOWP	2	2	16	1	Bytes Mismatch
10	ResNet50	FP16	1	1	29	1	Bytes Mismatch
11	ResNext	SYMLOWP	1	1	12	1	Bytes Mismatch
12	ResNext	SYMLOWP	2	2	7	1	Bytes Mismatch
13	ShuffleNet	GEMMLOWP	2	12	11	1	Bytes Mismatch
14	ResNext	SYMLOWP	1	1	4	1	Bytes Mismatch
15	ResNext	SYMLOWP	1	1	1	1	ECC Error
16	ShuffleNet	GEMMLOWP	2	2	3	1	Bytes Mismatch
17	ShuffleNet	GEMMLOWP	2	2	14	1	Bytes Mismatch
18	ResNet50	FP16	1	1	36	1	Bytes Mismatch
19	ResNext	SYMLOWP	1	1	22	1	Bytes Mismatch
20	ResNext	SYMLOWP	1	1	3	1	Bytes Mismatch

3. CONCLUSIONS

The Hardware Screener Application to screen out any bad/defective units is currently deployed at pre- and post-production environments within Intel, Intel Card Manufacturers, Customer ODM Manufacturers, and Customer Data Centres.

This work has helped significantly to screen out the cards due to the following errors:

- Byte Mismatch/Computational errors
- Vmin sensitive issues fixing the voltage sensitivity at various operating frequencies
- Power/ Frequency related issues
- Memory Related errors – Deep SRAM, MRC errors
- ECC Correctable/Uncorrectable errors
- Floating point accumulator errors

The execution of this flows has also helped in the achieving the following significant targets:

- Stability of Hardware/Software - Many significant software bugs, hangs were also revealed and fixed in pursuit of executing continuous inferencing over multiple days
- DPMO Target - Defects per million opportunities target has been achieved
- Deviations in expected performance over multiple hours/days of execution

REFERENCES

[1] Quantization Algorithms: https://intellabs.github.io/distiller/algo_quantization.html

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6848-6856).

[5] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
[6] SpringHill (NNPI-1000) Intel's Data Center Inference Chipset, HotChips Conference 2019.
[7] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.
[8] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, "Aggregated Residual Transformations for Deep Neural Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5987-5995, doi: 10.1109/CVPR.2017.634.
[9] Spring Hill - Microarchitectures - Intel - WikiChip. (n.d.). from https://en.wikichip.org/wiki/intel/microarchitectures/spring_hill
[10] Glenn Henry, Parviz Palangpour, Michael Thomson, J Scott Gardner, Bryce Arden, Jim Donahue, Kimble Houck, Jonathan Johnson, Kyle O'Brien, Scott Petersen, Benjamin Seroussi, Tyler Walker, "High-Performance Deep-Learning Coprocessor Integrated into x86 SoC with Server-Class CPUs Industrial Product", 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), pp.15-26, 2020

AUTHORS

Ashish Sharma (ashish3.sharma@intel.com) is an AI Software Engineering Manager at Intel Technologies, Bangalore. He holds a Masters in Computer Systems from BITS Pilani. He has been in R&D for 23 years. He leads the SW Validation and Development activities primarily for Tensorflow, PyTorch, PyLightning frameworks, Security and Linux Kernel drivers on Intel Training/Inference Accelerators. He has been leading multiple work groups within the team focusing on Leveraging Industry Practices, Standardization, Competition in AI.

Puneesh Khanna (puneesh.khanna@intel.com) is a AI Frameworks Architect at Intel Technologies, Bangalore. He holds a Masters in Machine Learning and AI from LJMU University, UK. He has been in R&D for 17 years. He works primarily on Tensorflow and Pytorch frameworks and enabling state of art deep learning models on Intel AI accelerators. He has been also contributing to the opensource code of these AI frameworks. He is leading an AI cohort group of GAR region, solving enterprise specific problems by leveraging AI and ML at Intel.

Jaimin Maniyar (jaimin.maniyar@intel.com) is a AI Solutions Engineer at Intel Technologies, Bangalore. He holds a Masters in Machine Learning and AI from Vellore Institute of Technology. He has been in R&D for 5 years. He works primarily on Tensorflow and Pytorch frameworks and enabling state of art deep learning models on Intel AI accelerators. He has delivered many AI and ML Sessions at Intel collaborating with Dataa Centre Skills Academy.

AUTHOR INDEX