Video Classification-Based Action Recognition with Enhanced Convolutional Neural Networks

Bo Mei

College of Science and Engineering, Texas Christian University,
Fort Worth, TX 76129, USA

Abstract. The classification of videos has become increasingly important in the field of data science research, as it has numerous practical applications in modern society. Compared to image classification, video classification poses a significantly greater challenge. One of the most obvious difficulties is that video classification tasks require more powerful computers due to the large number of features that need to be computed. Additionally, conventional 2D Convolutional Neural Networks (2D CNNs) are not effective in handling such tasks. This paper proposes a novel 2-layer Convolutional Neural Network (CNN) architecture for action recognition that addresses these challenges. The proposed architecture achieved a high test accuracy of 79.66% for classifying large video clips. The results indicate the effectiveness of the proposed approach for video classification tasks.

Keywords: neural networks, video classification, action recognition

1 Introduction

Based on public records from YouTube, a popular video broadcasting website, the total length of video clips stored on its server has surpassed 720,000 hours.[1] Consequently, video clips available on the Internet contain a significant amount of valuable information. Efficiently utilizing this information can bring benefits in various fields. For instance, implementing video classification can enhance security cameras by enabling them to recognize human activities and replace security guards.[2] [3] Moreover, video classification has numerous other applications, such as creating a knowledge base automatically, and enabling smart resource allocation.[4] [5] [6] [7]

Deep neural networks have been frequently used to classify video clips efficiently.[8] [9] [10] Although 2D CNN performs well in image classification, it may not be suitable for video classification because video clips have an additional dimension, which is the frames. In other words, images can be described as a special case of a video clip with only one frame. To address this challenge, a 2-layer CNN model has been proposed in this paper. The first layer of CNN learns the intra-frame information, and the second layer of CNN extracts changes between several consecutive frames. The architecture of the proposed model is shown in Figure 1, and the final test accuracy achieved is 79.66%. This paper primarily focuses on activity recognition for Activities of Daily Living (ADL), and the dataset used in this study covers most ADL activities.

2 Related Work

2.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) resemble traditional Artificial Neural Networks (ANNs) in that they consist of self-optimizing neurons that learn through training. Similar to ANNs, each neuron receives input and performs a scalar operation, followed by a non-linear function. The entire network can be expressed as a perceptive score function, with

David C. Wyld et al. (Eds): NSEC, InWeS, CSIP, BMLI, CACIT, EDUIT -2023
pp. 65-73, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131506

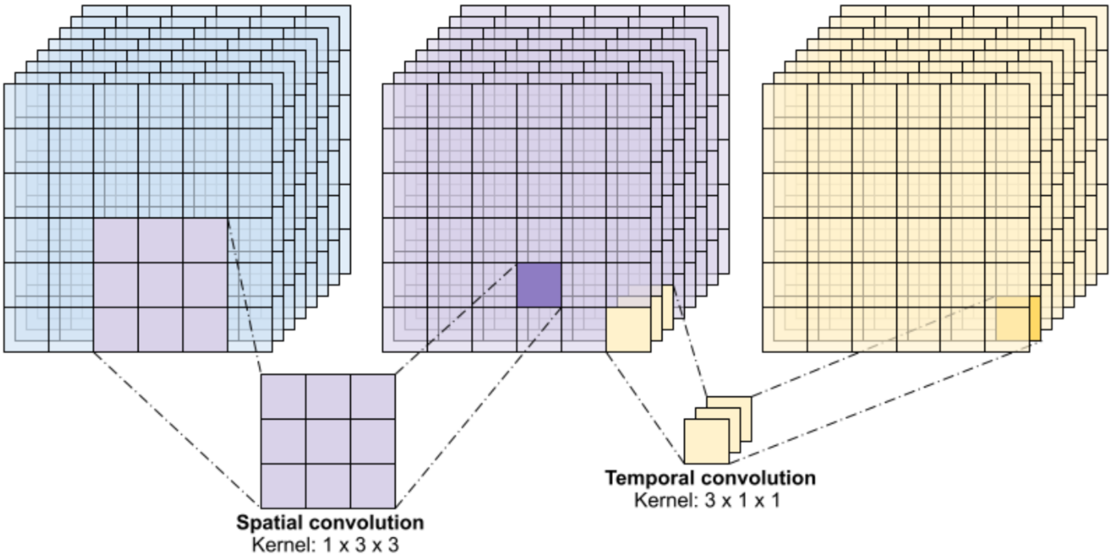

Fig. 1. Overview of proposed model

the final layer containing loss functions associated with different classes. Many of the same techniques used in traditional ANNs can also be applied to CNNs.[11]

However, CNNs differ from traditional ANNs in that they are primarily used for pattern recognition within images. This enables the encoding of image-specific features into the architecture, making the network more suitable for image-related tasks and reducing the number of parameters required to establish the model.

2.2 Action Recognition

Action recognition is a technology that enables computers to identify and comprehend actions or movements performed by humans or objects in video footage. This technology has a wide range of applications, including surveillance, sports analysis, and human-computer interaction. To perform action recognition, algorithms typically process video footage frame by frame, using techniques like feature extraction and machine learning to identify and classify various actions.[12] [13] [14]

However, accurately identifying and classifying complex actions, such as multiple individuals interacting or objects moving at high speeds, is a significant challenge in action recognition. To overcome this challenge, researchers have developed advanced algorithms that can analyze multiple frames of video simultaneously. Techniques such as Convolutional Neural Networks and Recurrent Neural Networks have been used to enhance performance. Moreover, many action recognition systems incorporate additional sensory information, like audio or depth data, to provide a more comprehensive understanding of the actions being performed.

3 Model

3.1 Data Preprocessing

The original dataset comprised 16,115 video clips with varying lengths. To utilize the data in the CNN, fixed length data is necessary. Consequently, the video clips were clipped to a length of 30 frames, equivalent to one second of footage. Due to the total memory constraint of the NVIDIA RTX 3060Ti GPU, a reduction in resolution was required to make the model functional on non-professional personal computers. A 2D max pooling layer with a stride of two was added before inputting the training data into the GPU, which reduced the total resolution of each frame to 240 × 320 × 3. Finally, the data was reshaped to 3 × 30 × 240 × 480, where 3 represents the number of channels (RGB), 30 represents the number of frames, and 240 and 480 indicate the height and width of each frame, respectively.

3.2 The CNN Layer

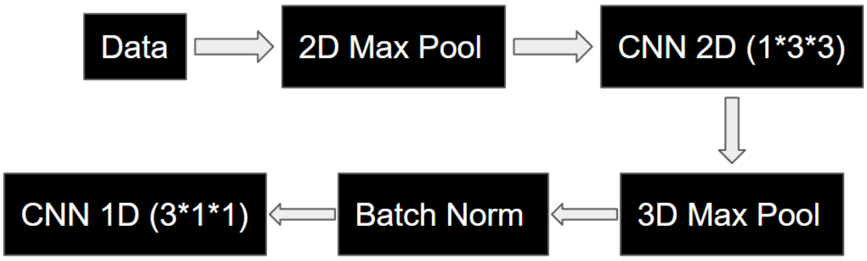

Fig. 2. Detailed information of the proposed CNN model

After data preprocessing, the processed video clips will undergo a 2D convolutional neural network (CNN) layer to extract information within each frame. The initial 2D CNN model is a (3×3) model with three input channels since each frame of the video is a three-dimensional array that contains RGB values represented as floating-point numbers. The primary purpose of the first 2D CNN layer is to extract information from each frame.

However, utilizing CNNs on individual frames of a video clip alone is insufficient for effective classification because a video comprises multiple frames. To capture the changes between frames, an additional 1D CNN model is necessary. Before feeding the data into the second layer, further pooling and normalization are carried out to decrease the computational requirements.

The final layer of the architecture is a fully connected layer that converts the output to probability distributions for each class. The final prediction is made by choosing the class with the highest probability. Although this CNN architecture is simple, it can learn effectively from video clips. The comprehensive experiment results are presented in the Experiment section.

4 Dataset

The Toyota Smarthome Trimmed dataset was developed for the purpose of activity classification and comprises 31 activities. The videos were segmented based on the activities, resulting in 16,115 short RGB+D video samples. The activities were performed naturally, resulting in a unique combination of challenges such as high intra-class variation, high-class imbalance, and activities with similar motion and high duration variance. The activities were annotated using coarse and fine-grained labels, which distinguishes Toyota Smarthome Trimmed from other datasets designed for activity classification.

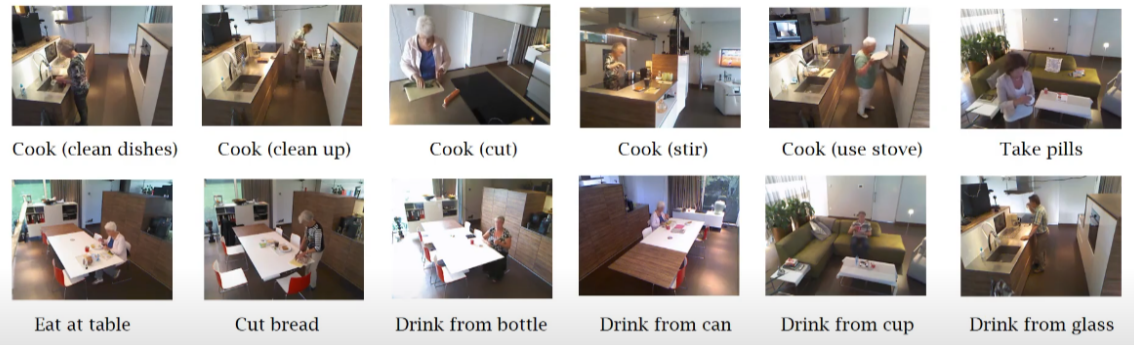

Fig. 3. Classes of Toyota Smarthome Dataset

This dataset was chosen due to its large number of classes and abundant data clips, which will contribute to a more robust test accuracy. Furthermore, the inclusion of similar actions, such as using a tablet and watching TV, poses an additional challenge in training the model.

5 Experiments

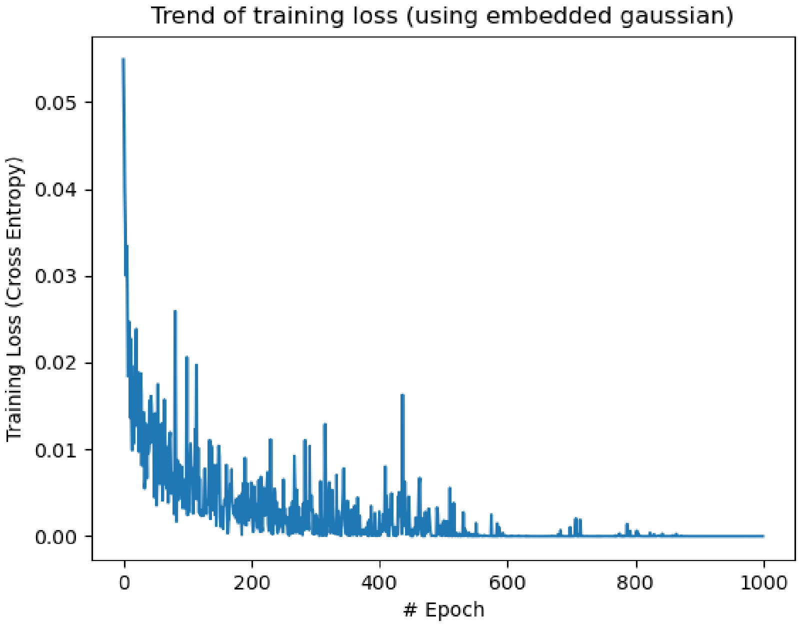

Fig. 4. Trend of training loss using proposed model

5.1 Preparing Dataset

As previously mentioned, the Toyota Smarthome dataset was trimmed to 30 frames in order to reduce preprocessing time. While this may not be sufficient for other video classification tasks, it is adequate for most activities of daily living. Additionally, a function was developed to save the preprocessed data locally and load them from the disk rather than reading the RGB video clips directly, further reducing the time required. This approach enables easy modification and re-training of the model without repeating the preprocessing step. The detailed statistics of the training and testing datasets are presented in Table 1.

Two activity types in the Toyota Smarthome dataset, namely get up and sit down, were found to be unsuitable for the model since their video clips were too short and did not meet the one-second threshold required by the model. As a result, these two classes were removed, leaving a total of 29 classes for classification by the model.

5.2 Training and Testing

The proposed model was trained and tested using a batch size of 10 over 1000 epochs, and the training and testing process was completed in approximately three hours. The changes in training loss over time are depicted in Figure 4, indicating that the model is learning effectively and improving its performance on the video classification task. The decrease in training loss demonstrates the model's effectiveness and its ability to classify video clips accurately.

Subsequently, all 290 test data clips, consisting of 10 clips for each of the 29 classes, were evaluated using the trained model. The final test accuracy was calculated and recorded after three complete training and testing processes. The model achieved test accuracies of 71.33% (with only 500 epochs), 74.5%, and 79.66% in these processes. These results provide evidence of the model's efficacy in classifying video clips.

Table 1. Statistics of Dataset

Name	Size
Training set	5725
Test set	290
Number of Classes	29

6 Challenges

Although the final testing accuracy indicates that the proposed model is capable of learning from video clips and making accurate predictions, a few issues were encountered during the training period.

6.1 Data Loading

As previously stated, the Toyota Smarthome dataset is a large dataset consisting of over 11 Gigabytes of data. Furthermore, the video clips have varying lengths, making it impossible to load the data directly from the disk without trimming. To address this, the decision was made to trim the data and save it locally as .npy files using the built-in numpy function. However, these .npy files occupy considerably more disk space compared to the original video clips. For example, a preprocessed data clip with 30 frames in the form of .npy requires approximately 2.4 Gigabytes of space. Unfortunately, due to limited disk space on the personal computer, it was necessary to reduce the total number of training files to 5725.

6.2 Training With Other Model

Initially, a self-attention layer was intended to be included between the two designated CNN layers. However, the implementation of this layer resulted in an out of memory error on the machine. Several methods were attempted to resolve this issue, including reducing the batch size, lowering the frame resolution to 96 × 128, and clearing the CUDA memory at the conclusion of each epoch. Unfortunately, none of these approaches proved successful, and the decision was made to remove the layer from the current model.

7 Future Research Plan

While the proposed approach has demonstrated the effectiveness of using a 2-layer CNN in video classification tasks, there are several alternative approaches that have exhibited superior performance. As the adage goes, ”If you want to beat it, you learn it.” Hence, investigating and adopting these models will provide valuable insights and enhance the performance of the proposed model.

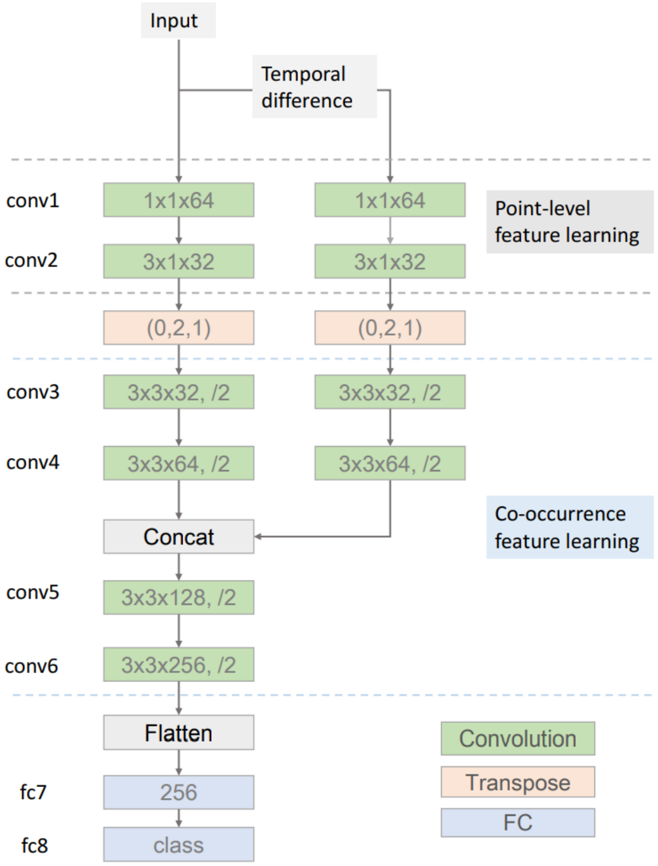

Fig. 5. Example Architecture for Hierarchical Aggregation Model

7.1 Long-Short Term Memory

The current 2-layer CNN model is incapable of handling video clips of varying lengths, which is a common occurrence in daily activities. For instance, certain actions such as get up and sit down were eliminated from the dataset before applying the CNN model.

To address this issue, Residual Neural Network (RNN) can be employed as a potential solution. RNN, as a special case, is capable of handling long-term dependencies, and Long-Short Term Memory (LSTM) can extract additional information from extended video clips. Importantly, both RNN and LSTM are capable of handling data with various lengths, making them well-suited for video/audio classification tasks.

7.2 Self Attention

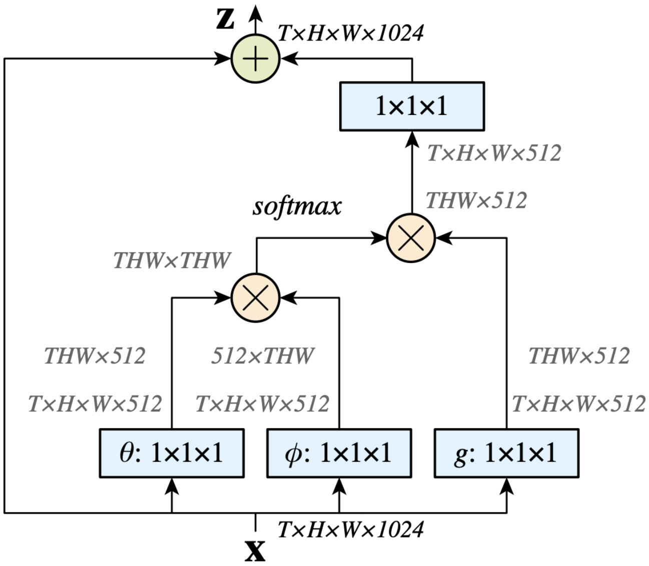

Fig. 6. Example Architecture for Non Local (Self Attention) Model

It has been observed that incorporating self-attention between CNN layers is another effective method to enable the model to extract inter-frame changes within a video clip. This is due to the ability of self-attention to direct the model's attention to specific portions of the input data, aiding in the identification of important features in the data. The detailed architecture of the self-attention model is depicted in Figure 6.

7.3 Hierarchical Aggregation

Video classification necessitates the consideration of both intra-frame and inter-frame changes. While searching for appropriate models, the Hierarchical Aggregation approach was discovered, which could be beneficial for video classification. This approach allows the model to incorporate information from various portions of the video and learn to identify crucial features in the data. The detailed architecture of Hierarchical Aggregation can be found in Figure 7.

As depicted in Figure 7, Hierarchical Aggregation generates a series of additional data that contain temporal differences. These data are then input into the same deep learning model as the original dataset. The original dataset and temporal differences are concatenated to enable the model to identify inter-frame changes, which is beneficial for video classification.

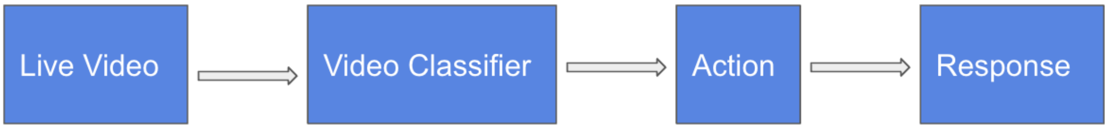

Fig. 7. Simple Smart Home Monitoring System

8 Ablation Study

Precise action recognition and classification by security cameras could have numerous practical applications, such as enhancing home monitoring systems.

Most home monitoring systems currently require human presence as they only display live video on the screen. Individuals monitoring the system must manually classify every action they observe and decide on a corresponding response (most commonly "do nothing"). This solution is suboptimal since it can be replaced by artificial intelligence. Additionally, the expense of hiring an individual to perform this task is unnecessary. Improved video classification accuracy would eliminate the need for continuous human surveillance. After the classifier has made its prediction, simple if-statements can be added. For example, if the classifier predicts "robbery," the system should automatically notify the police. The development of a smart home monitoring system is considerably complex, but it is feasible once the deep learning challenge has been resolved.

9 Conclusion

In summary, this approach offers a potential advantage in terms of its simplicity, as the 2-layer architecture allows for efficient computation and training of the model. Furthermore, the use of Convolutional Neural Networks enables the model to learn complex spatial and temporal relationships in the video data, improving its ability to accurately classify actions.

However, several challenges must be considered when implementing this approach in real-world scenarios. For example, the requirement for powerful computers and significant amounts of data may limit the model's practicality in certain situations. Additionally, the complexity of action recognition may necessitate the use of more advanced techniques, such as Recurrent Neural Networks, to achieve high accuracy levels.

In conclusion, the proposed 2-layer Convolutional Neural Network architecture is a promising solution for video classification tasks and has the potential to make significant contributions to the field of action recognition. The exploration of cutting-edge approaches will continue in the hopes of contributing to the field of computer vision.

References

1. Jain, Ritika, Riya Garg, Muskan Verma, and Anuradha Taluja. Classification of YouTube Data based on Opinion Mining. In Proceedings of the International Conference on Innovative Computing & Communication (ICICC). 2021.
2. Vrskova, Roberta, Robert Hudec, Patrik Kamencay, and Peter Sykora. Human activity classification using the 3DCNN architecture. Applied Sciences 12, no. 2 (2022): 931.
3. Sharma, Vijeta, Manjari Gupta, Anil Kumar Pandey, Deepti Mishra, and Ajai Kumar. A review of deep learning-based human activity recognition on benchmark video datasets. Applied Artificial Intelligence 36, no. 1 (2022): 2093705.

4. Gao, Junyu, Tianzhu Zhang, and Changsheng Xu. Learning to model relationships for zero-shot video classification. IEEE transactions on pattern analysis and machine intelligence 43, no. 10 (2020): 3476-3491.
5. Bhardwaj, Shweta, Mukundhan Srinivasan, and Mitesh M. Khapra. Efficient video classification using fewer frames. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 354-363. 2019.
6. Jiang, Xiantao, F. Richard Yu, Tian Song, and Victor CM Leung. Intelligent resource allocation for video analytics in blockchain-enabled internet of autonomous vehicles with edge computing. IEEE Internet of Things Journal 9, no. 16 (2020): 14260-14272.
7. Jiang, Xiantao, F. Richard Yu, Tian Song, and Victor CM Leung. A survey on multi-access edge computing applied to video streaming: Some research issues and challenges. IEEE Communications Surveys & Tutorials 23, no. 2 (2021): 871-903.
8. Rehman, Atiq, and Samir Brahim Belhaouari. Deep learning for video classification: A review. (2021).
9. Islam, Md Shofiqul, Shanjida Sultana, Uttam Kumar Roy, and Jubayer Al Mahmud. A review on video classification with methods, findings, performance, challenges, limitations and future work. J. Ilm. Tek. Elektro Komput. Dan Inform.(JITEKI) 6 (2020): 47-57.
10. Brattoli, Biagio, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4613-4623. 2020.
11. Ghimire, Deepak, Dayoung Kil, and Seong-heum Kim. A survey on efficient convolutional neural networks and hardware acceleration. Electronics 11, no. 6 (2022): 945.
12. Zhang, Hong-Bo, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. "A comprehensive survey of vision-based human action recognition methods." Sensors 19, no. 5 (2019): 1005.
13. Duan, Haodong, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. "Revisiting skeleton-based action recognition." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2969-2978. 2022.
14. Basak, Hritam, Rohit Kundu, Pawan Kumar Singh, Muhammad Fazal Ijaz, Marcin Woźniak, and Ram Sarkar. "A union of deep learning and swarm-based optimization for 3D human action recognition." Scientific Reports 12, no. 1 (2022): 5494.

Authors

Bo Mei received Ph.D in Computer Science from the George Washington University. He did Master of Science in Computer Science from the George Washington University. Currently, he is an Assistant Professor of Computer Science in Texas Christian University. His research interests are Machine Learning and Internet of Things (IoT).