MATERIAL VISIONS: ADVANCING CRYSTAL STRUCTURE PREDICTION WITH POWERFUL TEXT GENERATION MODELS

Reihaneh Maarefdoust[1], Xin Zhang[2], Behrooz Mansouri[3], Yuqi Song[4]

Department of Computer Science, University of Southern Maine, Portland, USA

ABSTRACT

The discovery of new materials has been a protracted and labor-intensive endeavor, relying on iterative trial-and-error methodologies. Recently, materials informatics has been transforming this process by employing advanced data science and computational tools to expedite the discovery of novel materials, such as generative design material formulas, and predict material properties. However, predicting crystal three-dimensional structures remains a challenging task rooted in both the fundamental nature of materials and the limitations of current computational methods. Inspired by the power and success of artificial intelligence (AI) models, especially the deep learning techniques and natural language processing (NLP) algorithms, we consider capturing complex atom descriptions and relationships as text information and explore whether we can use the ability of language models to predict atomic coordinates. In this work, we explore multiple text generation models and employ the Longformer-Encoder-Decoder (LED) model to construct preliminary crystal structures based on detailed atom descriptions. Subsequently, these structures are further refined by a random forest regressor, which generates the final crystal configurations. Our experiments show this method excels in capturing the intricate atom relationships and effectively translating these associations into the specified crystal formats. We also focus on optimizing data representation for both atom descriptions and crystal structures and use clear metrics to evaluate accuracy and stability. Our results indicate that this method has promising potential and could improve the prediction of material crystal structures. Our source code can be accessed freely at https://github.com/RMaarefdoust/Crystal-Structure-Prediction.

KEYWORDS

Material Informatics, Crystal Structure Prediction, Text Generation, Longformer-encoder-decoder

1. INTRODUCTION

TMaterials play a crucial role in driving technological progress and innovation. By developing new materials with specific properties, scientists can create new products or improve the performance of existing ones to address a wide range of challenges. However, traditional methods of designing and developing novel materials are a time-consuming, labor-intensive process that relies heavily on iterative trial-and-error experimentation, or even intuitive insights, such as Density Functional Theory (DFT) calculations [1], which is a quantum mechanical modelling method to investigate the electronic structure of many-body systems, especially in

David C. Wyld et al. (Eds): NLMLT, SEAU– 2024
pp. 01-11, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142001

atoms. The emerging material informatics is set to revolutionize the traditional paradigm, leveraging the powerful synergy between data science and computational tools [2]. This innovative subject integrates machine learning and deep learning techniques with extensive datasets. Material informatics not only accelerates the discovery and selection of materials but also enables a more profound analysis of a vast array of materials data. Consequently, researchers gain deeper insights into the behavior of materials, facilitating swifter and more informed decision-making throughout the development cycle.

Materials informatics has significantly advanced in generating new chemical formulas and predicting material properties [3]. Despite these advances, accurately predicting crystal three-dimensional structures remains challenging. This difficulty primarily arises from the need to model the orderly arrangement of atoms within a lattice. The complexity inherent in this research is further exacerbated by the diverse range of interatomic forces involved, including ionic, covalent, metallic, and van der Waals forces, among others [4]. The intricate interplay of these forces substantially complicates the prediction of how atoms will organize themselves into a stable and coherent structure. As artificial intelligence models become increasingly powerful, researchers are increasingly utilizing deep learning models, machine learning algorithms and evolutionary algorithms to improve the accuracy of crystal structure prediction [5]. For example, Chen et al. proposed the GN-OA [6] model which utilizes graph networks to optimize crystal structures with atomistic potentials. M3GNet [7] was trained on the massive database of structural relaxations performed by the Materials Project [8] over the past ten years and has broad applications in structural relaxation, dynamic simulations and property prediction of materials across diverse chemical spaces. These studies have advanced the progress in crystal structure prediction, but there are extensive opportunities for further exploration and discovery.

Inspired by the notable success of natural language processing algorithms, our study explores the intriguing possibility of transforming complex atomic descriptions and relationships into textual data. This approach aims to leverage the sophisticated predictive capabilities of text models to predict atomic coordinates. To this end, we evaluate various NLP generation models to assess their efficacy and performance in predicting atomic coordinates within specific material formulas. For instance, we investigate the classic Long Short-Term Memory (LSTM) method [9], the emerging and popular BART [10] and Sentence-BERT (SBERT) [11], etc. Our findings highlight the Longformer [12] as particularly adept for crystal structure prediction tasks, especially beneficial in processing extensive documents that require maintaining context across expansive text spans. The Longformer's efficient attention mechanism, coupled with its proficiency in handling lengthy sequences and its robust encoder-decoder architecture, makes it an exemplary tool for this advanced application.

In this paper, we introduce a new approach to predict material structures with advanced computational models with a focus on deep learning techniques tailored for text generation. The cornerstone of our approach is to augment traditional methods by integrating natural language processing algorithms. This integration enables us to accurately model complex atomic relationships and predict crystal structures with enhanced precision and efficiency. Our overarching aim is to propel the field of materials science forward by applying these innovative computational strategies, thereby opening new avenues for research and practical application in material analysis, novel material discovery and material design.

2. RELATED WORKS

2.1. Prediction of Material Crystal Structures

Ryan et al. [13] applied deep neural networks to analyse a large collection of crystallographic data from crystal structure repositories. In their study, multiperspective atomic fingerprints were used to describe coordination topology, which allowed them to distinguish chemical elements and identify structurally similar atomic sites, revealing trends that can be related to the periodic table. Furthermore, this model predicted the likelihood of forming new compounds with a high degree of accuracy, demonstrating its potential to assist in the discovery of new materials, particularly those containing a variety of chemical elements.

The CALYPSO was developed for predicting stable or metastable crystal structures, addressing the challenge of determining structures based solely on chemical compositions and external conditions like pressure. By employing a particle swarm optimization algorithm along with techniques such as symmetry constraints, a bond characterization matrix, and a penalty function, CALYPSO enhances structural diversity and avoids premature convergence. These advancements have demonstrated high efficiency and success in various systems. By leveraging these innovative techniques, CALYPSO can accurately predict crystal structures, making significant advances in the field of material behavior at the atomic level [14].

2.2. Text Generation

Text generation has been revolutionized by deep learning, particularly with the rise of pre-trained language models (PLMs). Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional RNNs (BRNNs), and Convolutional Neural Networks (CNNs) are commonly used in text generation models [15], while BART and SBERT have shown promising results in various text generation tasks.

The T5 (Text-to-Text Transfer Transformer) [16] is a flexible model for tasks like classification and summarization. Pretrained on the C-4 [1] dataset using a method similar to BERT, T5 is fine-tuned in the PoinT-5 [17] methodology to rephrase extracted sentences into brief, informative summaries. The T5 tokenizer prepares input and output sequences, enabling the model to convert narrative sentences into clear and concise summaries, enhancing precision and informativeness in financial narrative summarization.

Unlike traditional Transformer models, which struggle with long sequences due to the quadratic complexity of self-attention, the Longformer employs an attention mechanism that scales linearly with sequence length. Therefore, it can handle documents containing thousands of tokens without truncation or chunking. The model combines local windowed attention with task-motivated global attention, enabling it to capture local context and long-range dependencies. The Longformer has demonstrated state-of-the-art results in various tasks, including reading comprehension, classification, and summarization, outperforming models like RoBERTa on long document tasks. Furthermore, the Longformer-Encoder-Decoder (LED) variant extends the model to support generative sequence-to-sequence tasks. The Longformer's efficient attention mechanism and its ability to process long documents in a single pass make it a powerful tool for a wide range of natural language processing applications [12].

However, T5 was found to be too large for our specific needs, necessitating the use of a smaller pre-trained model. Therefore, we opted for the Longformer Encoder-Decoder due to its efficient attention mechanism and suitability for handling long documents. This choice allowed us to fine-tune and train model effectively for summarizing financial reports.

2.3. Longformer-Encoder-Decoder (Led)

[1] https://www.tensorflow.org/datasets/catalog/c4

The Longformer is highly suitable for text generation tasks, particularly for long documents, due to several key factors. As a result of its efficient attention mechanism, processing of longer texts does not require significant computational resources. By combining local windowed attention with task motivated global attention, the Longformer effectively focuses on relevant parts of the input while capturing the broader context. Regarding sequence-to-sequence tasks, the Longformer-Encoder-Decoder variant is ideal for summarizing and translating content due to its encoder-decoder architecture and efficient attention pattern. The Longformer has achieved state-of-the-art results in long document tasks, such as the arXiv summarization dataset, showcasing its effectiveness in generating coherent and contextually relevant text. In addition, it can be pre-trained on large corpora and fine-tuned for specific tasks, leveraging large datasets to learn general language patterns. Due to its ability to process long documents without truncating or chunking, it is a powerful tool for text generation when maintaining context is important [12].

Recent advancements in natural language processing have significantly benefited from the development of transformer-based models, particularly in handling long document tasks such as question answering and summarization. However, the self-attention mechanism in traditional transformers poses challenges due to its quadratic computational complexity concerning sequence length. Several models have been proposed to address this issue, including Longformer, Hierarchical Transformers, and BigBird, which present a variety of strategies that reduce computation costs while maintaining performance. Among these, the Poolingformer model introduces a novel two-level attention schema that combines sliding-window attention with pooling operations to process long documents efficiently. This approach not only reduces computational cost and memory consumption but also achieves superior performance on tasks such as the Natural Questions (NQ) [18] and TyDi QA [19] benchmarks, demonstrating its effectiveness in long document modelling [20].

2.4. Random Forest

Random Forests is a powerful machine learning model that trains many decision trees on a dataset while also aggregating their predictions. High-dimensional data, missing values, and outliers are handled effectively by this ensemble method, which often only requires minimal hyperparameter tuning. GridSearchCV can be used to enhance performance by optimizing hyperparameters. The Random Forest method involves making bootstrapped samples, selecting random features to decrease overfitting, constructing multiple decision trees, and combining their outputs. Hyperparameters like the number of trees and tree depth can be adjusted during customization. GridSearchCV was used to improve these parameters and select the most appropriate model for each machine learning approach. The model's performance can be improved by choosing the correct values for these hyperparameters. Although there are some drawbacks, such as the high training time for large datasets, Random Forests are versatile and effective for various predictive tasks. The combined use of Random Forests and GridSearchCV for heart disease prediction contributes to the advancement of machine learning applications in healthcare and aims to enhance patient outcomes globally [21]. Additionally, the use of GridSearchCV to fine-tune hyperparameters has been shown to improve performance in predictive tasks, though it may increase training time for large datasets [22].

3. METHODS

We propose a novel approach to material crystal structure generation by employing natural language processing techniques, as shown in Fig. 1. Our method utilizes an LED model to generate crystal structures based on an input string representing the material's atoms. The LED

output is then processed by a random forest regressor to predict the final crystal structure in the desired format. As a result of this. approach, it is possible to capture complex relationships between atoms through the NLP model, while the random forest regressor can be used to map between the LED decoder representation and the structured crystallographic data. However, successful implementation requires careful consideration of data representation for both atom strings and crystal structures, along with the specific role of the regressor in capturing spatial relationships. The evaluation metrics will be crucial for assessing the accuracy and stability of the generated crystal structures.

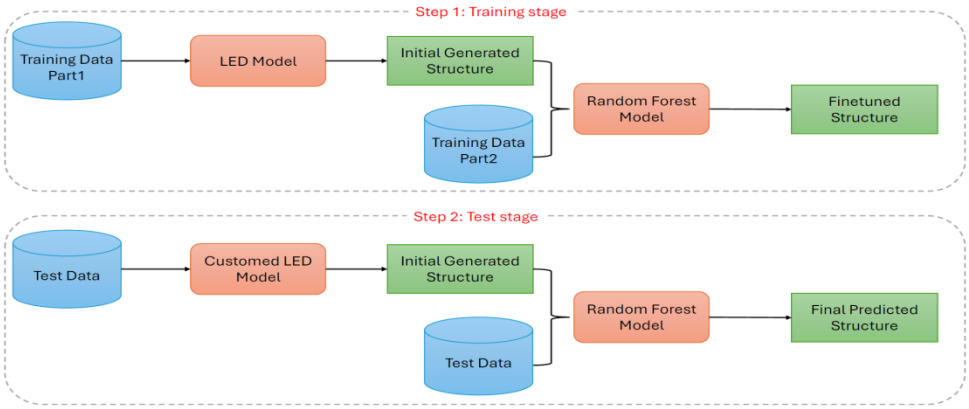

Figure 1. The framework of our method. Initially, we undertake the pre-training of the Longformer-Encoder-Decoder model using a subset of the training dataset. This step allows the model to assimilate the fundamental patterns of crystal atomic coordinates. Subsequently, the pre-trained LED model is employed to generate preliminary atomic coordinates. These initial coordinates are then refined through a random forest regressor, which fine-tunes them to enhance predictive accuracy. In the subsequent stage, we deploy our custom-tailored LED models on a test dataset to assess their performance and to generate the final predicted crystal structures. This two-tier steps not only ensure the precision of the predictions but also establish a robust framework for effectively modeling and predicting complex crystal structures.

3.1. Enhancing Random Forest Regression With Gridsearchcv

A Random Forest Regressor is an ensemble learning method used for regression tasks that consists of constructing numerous decision trees during the training. By averaging predictions for all trees, the algorithm improves accuracy and reduces overfitting. Each tree is built from a random subset of the training data, considering a random subset of features during splits, ensuring robustness and generalization. GridSearchCV, a robust hyperparameter tuning method, optimizes machine learning models by exhaustively searching through a specified parameter grid [22].

3.2. Dataset

The Materials Project is an open resource, providing freely accessible data to all interested parties. To facilitate efficient and large-scale data retrieval, the platform offers bulk download options. Recognizing that manual downloading can be a slow and labor-intensive process, the Materials Project provides an Application Programming Interface (API). This API enables users to download data seamlessly using popular programming languages.

We used the Materials Project API to collect approximately 10,000 CIF files[2] from their website. This approach not only streamlines the data acquisition process but also aligns with the Materials Project's commitment to providing accessible and user-friendly access to their extensive dataset. The data distribution and splitting details for the evaluation runs are summarized in Table 1. The first run utilized 4,900 LED data points with another split of 1053 points for the test and train datasets, as well as a smaller test set consisting of 71 points. As in Run 1, Run 2 involved a larger set of 8,826 LED data, the same allocation of 1,053 points for the test and train datasets, and 71 points for a smaller test set. These consistent data split across both runs ensure a balanced and systematic approach to model training and evaluation, providing a robust foundation for performance comparison and analysis.

Table 1. Data Points Per Splits.

Run	LED	Test	Train	Test
Run 1	4,900	1,053	1,053	71
Run 2	8,826	1,053	1,053	71

3.3. Evaluation Metrics

Our research focused on two important measures: training and validation loss. Our training loss reflects how well our model learns from the data it encounters during training. By measuring validation loss, we can determine whether the model continues to perform well on new data that it has not encountered previously. Keeping an eye on both of these measures enabled us to ensure our model learned well from the training data as well as handled new materials accurately.

4. RESULTS

In our research, we used the Longformer-Encoder-Decoder model, a type of advanced text generation model, to improve the prediction of material structures. Our approach involves inputting data about the elemental composition and sequence of materials into the model, allowing it to learn complex patterns and relationships. After training, we tested the model on a separate dataset to predict material structures. Our method provides a clear and effective framework for predicting material structures, using advanced AI models to advance materials science.

LED is suitable for tasks that require handling long documents, offering improvements in attention mechanisms. As depicted in Fig 2, the training loss decreases from 0.716 to 0.314 and validation loss from 0.349 to 0.329 over the first six epochs, then both losses stabilize around these values for the remaining epochs.

[2] https://next-gen.materialsproject.org/materials

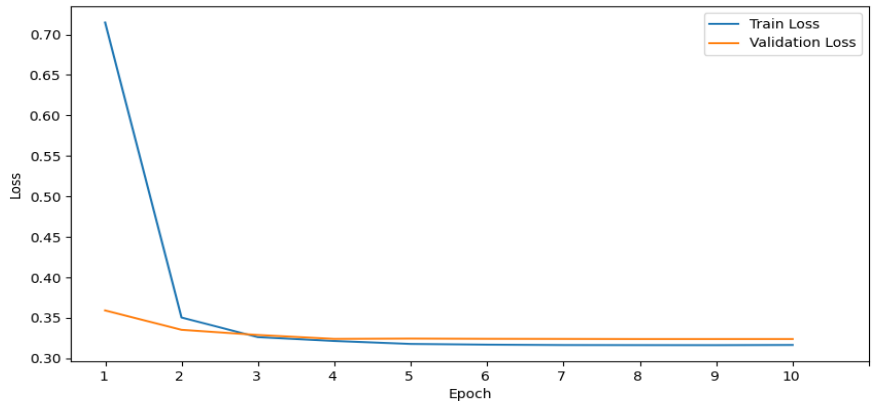

Figure 2. First LED Train and validation loss for 10 epochs for 4900 materials

As shown in Fig 3, both training and validation losses decrease over 10 epochs, with training loss reducing from 0.772 to 0.516 and validation loss from 0.566 to 0.515.

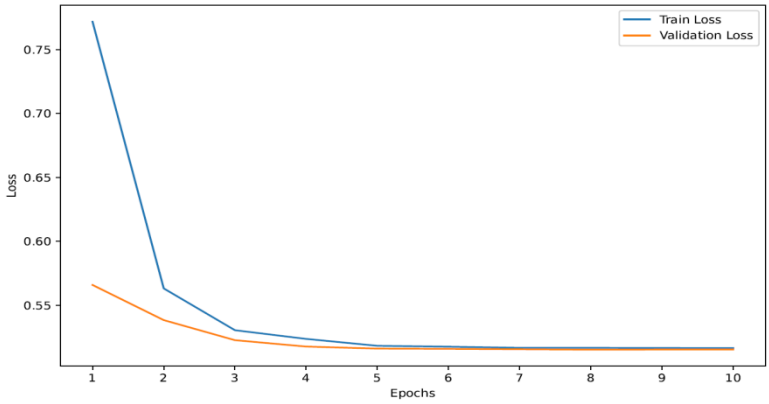

Figure 3. Second LED Train and validation loss for 10 epochs for 8826 materials

The evaluation results presented in Table 2 highlight the performance of two different runs. For Run 1, the Mean Squared Error (MSE) for the training set is 0.088, while the test set MSE is slightly higher at 0.098. This indicates a relatively low error in both the training and testing phases, suggesting good model performance. Similarly, Run 2 shows a training set MSE of 0.087 and a test set MSE of 0.096, demonstrating consistent results across both runs. Both runs involved fitting 405 folds, indicating a thorough cross-validation process to ensure robust and reliable evaluation. These metrics suggest that the models are well-fitted and have good generalization capabilities, as evidenced by the close values of the training and test MSEs.

Table 2. Evaluation Results. MSE indicates the mean square error.

Run	MSE (Train)	MSE (Test)	FittingFolds
Run 1	0.088	0.098	405
Run 2	0.087	0.096	405

Table 3. Selected Parameters Values after Grid Search.

Run	#Trees	Depth	SS	SL
Run 1	200	10	10	4
Run 2	300	10	10	1

In our work, we employed GridSearchCV from scikit-learn [23] to fine-tune the hyperparameters of the Random Forest Regressor. We performed cross-validation to find the optimal setting for both runs by defining a grid with various values for the following:

- #trees: Number of trees in the forest
- Depth: Maximum Depth of each tree
- SS: Minimum Samples required to Split a node
- SL: Minimum Samples required to be at a Leaf node

The optimal configurations for these parameters were evaluated and determined by two separate runs through systematic cross-validation [21], detailed in Table 3, which displays the values of the parameters after grid search. These finely tuned parameters highly improved the accuracy of our model, achieving an impressive MSE of 0.083 on the test dataset, which demonstrates the effectiveness of GridSearchCV in refining model performance.

5. CONCLUSIONS

Although our method successfully uses the Longformer-Encoder-Decoder (LED) model and a random forest regressor to predict material crystal structures, there are still several areas where we can explore and improve. The model's inability to handle novel or highly complex atomic configurations may be hampered by its reliance on predefined atomic descriptions as input strings. The model's generalizability to a wider range of materials and crystal structures could be enhanced by integrating dynamic input representations that adapt to diverse atomic systems in future work.

Moreover, while a random forest regressor can be effective in refining atomic coordinates, it could be more advantageous to experiment with other predictive models. Deep learning-based regressors, ensemble models, or hybrid approaches can provide better accuracy, scalability, and

computational efficiency. By capturing the intricate relationships between atomic interactions, this could help improve the model's ability to predict accurate crystal structures.

Providing training data that is both available and of high quality is another challenge. Since the LED model's performance is closely tied to the diversity and accuracy of its training data, obtaining larger and more varied datasets will be crucial for refining predictions, especially for materials with complex atomic behaviors. Advancing data augmentation techniques or synthetic datasets should be considered for future work to overcome these limitations.

Finally, while our evaluation metrics have shown promising results, it is important to address the robustness of the model when confronted with noisy, incomplete, or low-quality data, which can occur in real-world scenarios. Future research should look into more complex evaluation metrics that are more compatible with practical applications in materials science. We aim to improve the versatility and accuracy of our approach by tackling these challenges, which could potentially open up the possibility of predicting and discovering new material structures.

REFERENCES

[1] James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton, (2013), " Materials design and discovery with high throughput density functional theory: the open quantum materials database (oqmd)", Jom, 65:1501–1509, 2013.

[2] Jianjun Hu, Stanislav Stefanov, Yuqi Song, Sadman Sadeed Omee, Steph-Yves Louis, Edirisuriya MD Siriwardane, Yong Zhao, and Lai Wei, (2022), " Materials Atlas. org: a materials informatics web app platform for materials discovery and survey of state-of-the-art. npj Computational Materials, 8(1):65, 2022.

[3] Yuqi Song, Edirisuriya M Dilanga Siriwardane, Yong Zhao, and Jianjun Hu, (2021), "Computational discovery of new 2d materials using deep learning generative models", ACS Applied Materials & Interfaces, 13(45):53303– 53313, 2021.

[4] Artem R Oganov, Chris J Pickard, Qiang Zhu, and Richard J Needs, (2019), "Structure prediction drives materials discovery. Nature Reviews Materials", 4(5):331–348, 2019.

[5] Lai Wei, Sadman Sadeed Omee, Rongzhi Dong, Nihang Fu, Yuqi Song, Edirisuriya M. D. Siriwardane, Meiling Xu, Chris Wolverton, and Jianjun Hu." Cspbench: a benchmark and critical evaluation of crystal structure prediction", 2024.

[6] Guanjian Cheng, Xin-Gao Gong, and Wan-Jian Yin. (2022), "Crystal structure prediction by combining graph network and optimization algorithm", Nature communications, 13(1):1492, 2022.

[7] Chi Chen and Shyue Ping Ong, (2022), " A universal graph deep learning interatomic potential for the periodic table", Nature Computational Science, 2(11):718–728, 2022.

[8] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL materials, 1(1), 2013.

[9] Sepp Hochreiter, J¨urgen Schmidhuber, (1997), "Long short-term memory", Neural computation, 9(8):1735–1780, 1997.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer, (2019), "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension", arXiv preprint arXiv:1910.13461, 2019.

[11] Nils Reimers and Iryna Gurevych, (2019), "Sentence-bert: Sentence embeddings using siamese bert-networks", arXiv preprint arXiv:1908.10084, 2019.

[12] Iz Beltagy, Matthew E Peters, and Arman Cohan, (2020), "Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150, 2020.

[13] Kevin Ryan, Jeff Lengyel, and Michael Shatruk, (2018), "Crystal structure prediction via deep learning", Journal of the American Chemical Society, 140(32):10158–10168, 2018.

[14] Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma, (2012), "Calypso: A method for crystal structure prediction", Computer Physics Communications, 183(10):2063–2070, 2012.

[15] Touseef Iqbal and Shaima Qureshi, (2022),"The survey: Text generation models in deep learning", Journal of King Saud University-Computer and Information Sciences, 34(6):2515–2528, 2022.
[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu, (2020), "Exploring the limits of transfer learning with a unified text-to-text transformer", Journal of machine learning research, 21(140):1–67, 2020.
[17] Abhishek Singh, (2020) "Point-5: Pointer network and t-5 based financial narrative summarization", arXiv preprint arXiv:2010.04191, 2020.
[18] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov, (2019), "Natural questions: A benchmark for question answering research", Transactions of the Association for Computational Linguistics, 7:452–466, 2019.
[19] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi, (2020), "QA: A benchmark for information-seeking question answering in typologically diverse languages", Transactions of the Association for Computational Linguistics, 8:454–470, 2020.
[20] Hang Zhang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan, and Weizhu Chen, (2021), "Poolingformer: Long document modeling with pooling attention", In the International Conference on Machine Learning, pages 12437–12446. PMLR, 2021.
[21] Shagufta Rasheed, G Kiran Kumar, D Malathi Rani, MVV Prasad Kantipudi, and M.Anila. Heart disease prediction using grid-searchcv and random forest. EAI Endorsed Transactions on Pervasive Health and Technology, 10, 2024.
[22] Zaineb M Alhakeem, Yasir Mohammed Jebur, Sadiq N Henedy, Hamza Imran, Lu´ıs FA Bernardo, and Hussein M Hussein. "Prediction of eco-friendly concrete compressive strength using gradient boosting regression tree combined with grid-searchcv hyperparameter-optimization techniques", Materials, 15(21):7432, 2022.
[23] F.Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, (2011), "Scikit-learn: Machine learning in Python. Journal of Machine Learning Research", 12:2825–2830, 2011.

AUTHORS

Reihaneh Maarefdoust

She is a computer science graduate student at University of Southern Maine and Research Assistant at AIIRLab. Her primary research interests center around NLP, Information Retrieval, and Machine Learning methods.

Xin Zhang

He got the Ph.D. degree from the University of South Carolina in 2023. He is an assistant professor in the department of Computer Science at the University of Southern Maine. His research interests include computer vision and image processing.

Behrooz Mansouri

He is an Assistant Professor of Computer Science at the University of Southern Maine. He holds a Ph.D. (2022) in Computer Science from the Rochester Institute of Technology (Rochester, NY). His primary research interests include Information Retrieval and Natural Language Processing along with other topics in data mining and machine learning.

Yuqi Song

She received the Ph.D. degree in computer science from the University of South Carolina in 2023. She is currently an assistant professor in the department of Computer Science at the University of Southern Maine. Her research interests include machine learning, deep learning, data mining and their applications in cross-disciplines, such as materials informatics, recommender systems.