Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/42/20230794

The evolution of transformer models from unidirectional to bidirectional in Natural Language Processing

Yihang Sun

P.C. Rossin College of Engineering Applied Science, Lehigh University, 27 Memorial Drive West, Bethlehem, 18015, The United States

yis722@lehigh.edu

Abstract. Transformer models have revolutionized Natural Language Processing (NLP), transitioning from traditional sequential models to innovative architectures based on attention mechanisms. The shift from unidirectional to bidirectional models has been a remarkable development in NLP. This paper mainly focuses on the evolution of NLP caused by Transformer models, with the transition from unidirectional to bidirectional modeling. This paper explores how the transformer model has revolutionized NLP, and the evolution from traditional sequential models to innovative attention-driven architectures. In this paper, it mainly discusses the limitations of traditional NLP models like RNNs, LSTMs and CNN when handling lengthy text sequences and complex dependencies, highlighting how transformer models, employing self-attention mechanisms and bidirectional modeling (e.g., BERT and GPT), have significantly improved NLP tasks. It provides a thorough review of the shift from unidirectional to bidirectional transformer models, offering insights into their utilization and development. Finally, this paper concludes with a summary and outlook for the entire study.

Keywords: Unidirectional Model, Bidirectional Model, Natural Language Processing.

1. Introduction

Natural Language Processing has experienced a significant revolution, the transformer model emerging as a pivotal factor in this change. It is a significant transformation in NLP from traditional sequential models to innovative architectures based on the attention. Among this transformation, the transition from unidirectional model to bidirectional model has been a remarkable development [1-4].

Unidirectional models, deeply ingrained in the history of NLP, have been foundational in various text processing tasks over the years. These models operate on the principle of sequential language processing, adhering to the order in which words or tokens appear in a text. Their versatility has led to their application in tasks spanning from text classification to machine translation [5-7]. The straightforward nature of unidirectional models, along with their effectiveness, made them the default choice for many NLP applications. At the core of unidirectional models are recurrent neural networks (RNNs) and their variants, such as Long Short-Term Memory networks. These models process input data one token at a time while maintaining a hidden state that carries information from previous tokens to influence the processing of the current token [8-9].

Despite their historical success, unidirectional models face inherent limitations, particularly when dealing with lengthy texts and intricate linguistic dependencies. One of the primary challenges

DOI: 10.54254/2755-2721/42/20230794

unidirectional models encounters is their struggle to effectively capture long-range dependencies within language. Long texts often involve complex relationships between distant words or phrases, posing difficulties for traditional unidirectional models. The recurrent nature of RNNs and LSTMs hinders their ability to efficiently handle such dependencies. In contrast, the transformer model has brought a revolutionary solution to these limitations. It introduces self-attention mechanisms and multi-head attention, enabling it to perform global modeling of text sequences. By employing self-attention, transformers can simultaneously consider all positions within an input sequence, making them highly adept at capturing contextual information and long-range dependencies.

This paper mainly focuses on the evolution of transformer models in the NLP, and its emphasis on the transition from unidirectional to bidirectional modeling. We will talk about the various aspects of transformer models which include architecture, pre-training tasks and performance in the NLP tasks. Through this research, we aim to get a comprehensive understanding of the transformer model's evolution and the implications for the field of Natural Language Processing.

2. Unidirectional To Bidirectional

2.1. Traditional Unidirectional model

The traditional unidirectional model is a foundational approach for processing sequences, and it is marked by its linear handling of input data. Its fundamental characteristic can lie in the sequentiality of input elements in the order of their occurrence. Through this processing, the model can proceed step by step without revisiting or advancing beyond the current input element.

The Convolutional Neural Network (CNN) is indeed a single-directional model, widely employed for tasks such as image recognition, and its application has extended to other domains like Natural Language Processing.

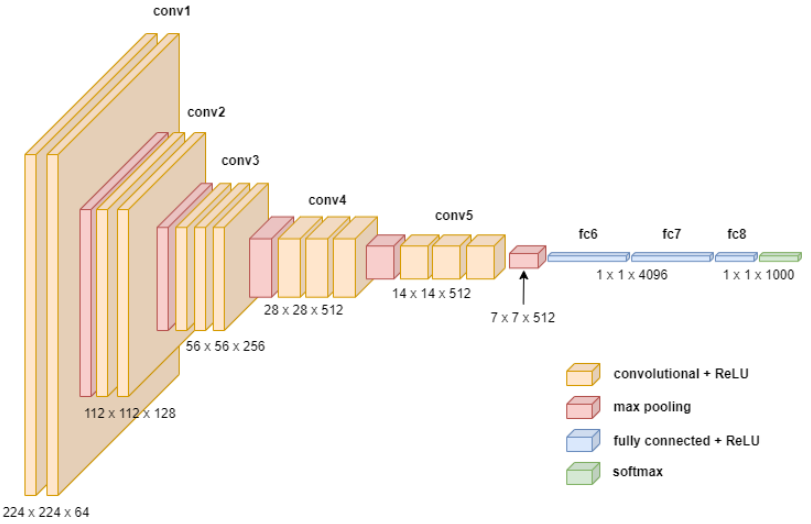

Figure 1. CNN architecture diagram [10].

In the Figure 1, a CNN model designed for image recognition, where the leftmost image represents our input layer, interpreted by the computer as a series of matrices. This CNN architecture features distinctive Convolution Layers with Rectified Linear Unit (ReLU) activation functions. The ReLU activation function is defined as ReLU(x) = max (0, x). Following the Convolution Layers, there are Pooling Layers, another essential element of CNNs. Pooling Layers do not have an activation function. The combination of Convolution Layers and Pooling Layers can be repeated multiple times in the hidden layers, as seen in the diagram, but the specific number of repetitions depends on the model's requirements. Alternatively, one can use combinations like Convolution Layers followed by Convolution Layers or Convolution Layers followed by Convolution Layers and then a Pooling Layer.

DOI: 10.54254/2755-2721/42/20230794

The choice of architecture depends on the specific modeling needs. Following several Convolution and Pooling Layers, there are Fully Connected Layers. These FC layers serve to connect the extracted features from previous layers to the output layer for classification or other tasks. In practice, CNN architectures often consist of several Convolution Layers followed by Pooling Layers, as depicted in the provided CNN structure [10].

Recurrent Neural Networks can be visualized as a cyclic neural network structure that can pass information between different time steps (Figure 2). It enables the network to handle sequential data, and can capture contextual information within the sequences.

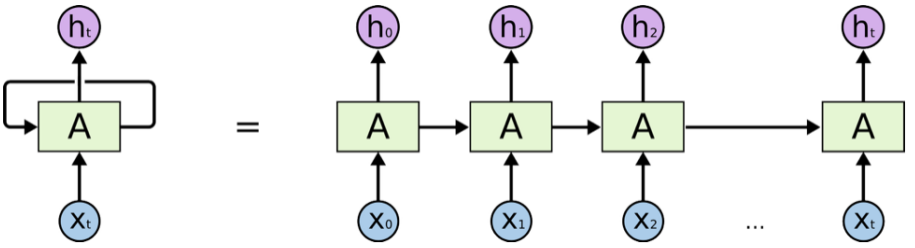

Figure 2. RNN architecture diagram [7].

Recurrent Neural Network consists of artificial neurons and one or more feedback loops. In this structure, x_t represents the input layer, h_t represents the hidden layer with recurrent connections, and y_t represents the output layer. The hidden layer contains a loop, and for better understanding, we can unfold this loop, resulting in the network structure. In the unfolded network structure, the input is a time sequence {..., X_{t-1}, X_t, X_{t+1} , ...}, where xt is associated with the number of input layer neurons. Correspondingly, the hidden layer is {...,h_{t-1}, h_t, h_{t+1} , ...}, where ht is associated with the number of hidden layer neurons [7].

This structure enables the RNN to effectively process sequential data by allowing it to pass information between different time steps and capture contextual information within the sequence. The input and hidden layer state at each time step depend on the previous time steps, making RNN a crucial tool in fields such as time series analysis, natural language processing, and more. Figure 2 illustrates the network structure of an RNN when processing such sequential data, providing us with a clearer visual understanding.

The Long Short-Term Memory algorithm is a special recurrent neural network with a distinctive model structure. Figure 3 shows the architecture of an LSTM. LSTM's point of distinction from traditional RNNs lies in its output mechanism. In addition to the conventional output, represented as 'h,' LSTM introduces an additional pathway that runs vertically through the entire network, known as the cell state. It's important to note that the cell state pathway lacks non-linear activation functions, and consists primarily of simple operations like multiplication and addition. The cell state can be passed to the next unit with relatively minimal changes. The brilliance of LSTM lies in its incorporation of various gating mechanisms, including the input gate, forget gate, and output gate. These gates in LSTM allow it to control the importance of information from the previous step. They use sigmoid functions that output values between 0 and 1 to decide how much old information to keep and how much new information to add to the cell state before passing it to the next step. This controlled flow of information makes LSTM well-suited for tasks involving long-term dependencies and memory retention.

DOI: 10.54254/2755-2721/42/20230794

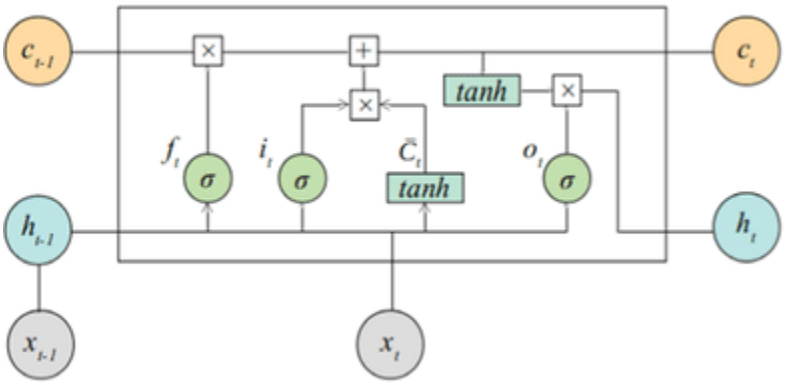

Figure 3. LSTM architecture diagram [9].

2.2. Challenges of Unidirectional model

2.2.1. Challenges in Handling Long-Term Dependencies

In certain situations, recent information is sufficient to address a given task effectively. For instance, If we want to predict the last word in the phrase "the clouds are in the sky," the word should be "sky," and we don't need extensive context. In such cases, the gap between the position to be predicted and the relevant information is relatively close, enabling RNNs to utilize past information. Conversely, there are many scenarios demanding more extensive context information. For example, the text "I grew up in France … I speak fluent French." Recent information suggests that the word to be predicted should pertain to a language, but to determine the specific language, we need background information like "grew up in France" from a more distant context. Theoretically, Unidirectional models have the capability to handle such long-term dependencies, but in practice, they often struggle to address this issue effectively [8]

2.2.2. Challenges in lack of data

Unidirectional models face challenges when adapting to multilingual and multidomain tasks. These challenges include a high demand for annotated data, which is necessary to train the model for specific languages, and it leads to substantial human and time resources. In other words, it means that model will spend high costs. Furthermore, the time and computational resources needed for training unidirectional models for diverse languages, it can pose practical challenges for both researchers and practitioners, especially when quick responses to different language and domain requirements are needed. These models lack adaptability and often require time-consuming retraining or fine-tuning for new languages or domains, which may not always be feasible or efficient. Additionally, in cases like low-resource languages or specialized domains, the scarcity of annotated data can severely hinder the application of natural language processing techniques. These challenges underscore the advantages of more versatile bidirectional models in handling diverse multilingual and multidomain NLP tasks.

3. Bidirectional models

This chapter begins by illustrating the fundamental principles of bidirectional models using the basic transformer model as an example. Subsequently, several improved models based on the Transformer are presented.

3.1. Transformer model

The Transformer model is a deep learning structure that is employed in natural language processing and various sequence-to-sequence tasks. The core innovation of the Transformer model is the self-attention mechanism. This mechanism allows the model to weigh the importance of each word in the sequence

DOI: 10.54254/2755-2721/42/20230794

relative to all other words, thus capturing rich contextual information. In the seminal work titled "Attention Is All You Need," it is explained that self-attention functions by allowing every word in the input sequence to focus on each other word, and their individual contributions are carefully amalgamated to create an output representation. This innovative method for global modeling represents a significant departure from traditional sequential models like RNNs and LSTMs, which struggled with handling long range dependencies in text [1]

The innovation of the Transformer lies in its extensive incorporation of the attention mechanism into neural network models, and it discarding the conventional LSTM and RNN architectures. Traditional sequential models like LSTM and RNN have trouble fully utilizing GPU parallel computing, making them inefficient in terms of time. The authors' innovative idea is to keep the attention mechanism while getting rid of LSTM and RNN structures, focusing on creating an attention model that can work well with parallel computing. One of the significant contributions of this paper is the attention sub-module's design, with the entire model being a stack of these sub-modules [1].

This model processes the entire input sequence in one go, eliminating the need for step-by-step reasoning as in LSTM and RNN, thus fully exploiting GPU parallel computing capabilities. However, to compensate for the loss of word positional information, position encoding is introduced. The figure 4 shows the detail of model.

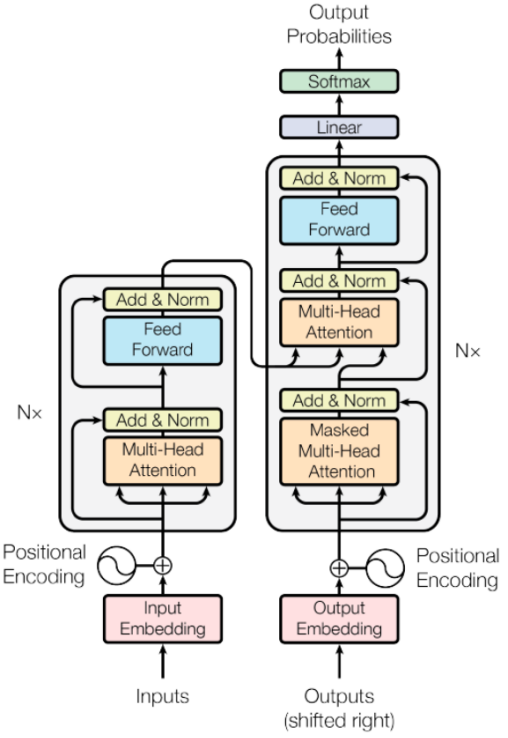

Figure 4. Transformer model [1].

3.2. The Bidirectional Model Based on Transformer Model

Building upon the self-attention mechanism at the core of the Transformer model, Bidirectional Encoder Representations from Transformers (BERT) represents a groundbreaking development in NLP and an embodiment of the power of bidirectional modeling. Traditional unidirectional models can only consider the context preceding a word in text, and it cannot capture information following a word directly. To address this, the BERT model introduced bidirectional self-attention, allowing it to consider both preceding and following context for each word. BERT does not rely on the previous word in the sequence, but it considers both the previous and the next word, and is therefore better at capturing subtle

DOI: 10.54254/2755-2721/42/20230794

relationships and dependencies in the text. This application of bidirectional modeling builds on the foundation laid by the Transformer model's self-attention mechanism.

BERT divides the input text into tokens and then maps them to high-dimensional vectors to create word embeddings that provide semantic information for each token. These embeddings are the basis for generating queries, keys, and values for each token through matrix multiplications. These representations are the basis for computing the self-attention score, which determines the relevance of each token to all other tokens [2]. Self-attention scores, determining token relevance in the context, are calculated by multiplying queries with the transpose of keys and scaling them. BERT then constructs a weighted context representation, connecting each token to all others in the sequence, enabling better context understanding and semantic grasp [3]. To enhance modeling capacity, BERT employs multi-head self-attention and a stacked architecture with multiple encoder layers. Multi-head self-attention captures diverse relationships, while stacking improves token representations and context understanding.

The ALBERT model introduces parameter sharing and cross-layer parameter sharing mechanisms to enhance the efficiency of Transformer models, while reducing the number of parameters. Through parameter sharing, ALBERT shares most of the parameters among different layers, which means that different layers use the same parameters. It can reduce the number of parameters in the model. It also helps to improve the training and inference efficiency. Furthermore, ALBERT introduces cross-layer parameter sharing. It enables the model to better capture semantic relationships between different layers. These innovations significantly improve model performance across various natural language processing tasks [6].

However, the limitations of the Albert model include its reduced performance in certain tasks compared to larger models, primarily due to information loss resulting from parameter sharing. Furthermore, training costs of ALbert model remains relatively high.

Transformer-XL model mainly introduces relative positional encoding and recurrence mechanism. Relative position encoding can make the model better understand the relationship between different positions in the sequence. At the same time, the recursive mechanism enables the model to handle longer text sequences. These two innovations can enhance the model's ability to model context, and it can capture long-distance dependencies. They have a significant impact on a variety of natural language processing tasks, thereby improving the performance and efficiency of the model [4].

Besides, XLNet model introduces self-attention regularization and permutation-based pre training techniques. The self-attention regularization encourages the model to be more selective in attention, and it can make the model focus on the most relevant parts of the input. The permutation-based pre-training techniques consider all possible permutations of the input sequence. This novel approach enables the model to learn from multiple context orders, enhancing its capacity to capture nuanced language dependencies. These enhancements play a crucial role in bolstering the model's robustness and its ability to comprehend intricate language structures [5].

Transformer-XL and XLNet are natural language processing models based on the Transformer architecture, but they have notable differences. Transformer-XL introduces relative positional encodings, allowing it can handle long sequence data and capture long-range dependencies, which is particularly useful for tasks that involve extensive text. In contrast. XLNet employs traditional absolute positional encodings, which exhibit lower performance, when dealing with long sequences. Furthermore, Transformer-XL is an autoregressive model, employing autoregressive mechanisms during both training and generation. XLNet adopts a non-autoregressive training approach, allowing it to use more surrounding context to predict each token. It can offer superior performance in certain tasks. Additionally, XLNet introduces Permutation Language Modeling to learn richer bidirectional dependencies, whereas Transformer-XL relies on conventional unidirectional language modeling. XLNet enhances parameter efficiency, and it performance through a two-stream prediction parameter sharing mechanism. The choice between the two models depends on specific task requirements and performance criteria, and these distinctions endow them with unique advantages in different tasks.

Recently, based on Transformer, the GPT model is developed, The GPT model is a language model that uses a Transformer framework and has strong language understanding and generation abilities

DOI: 10.54254/2755-2721/42/20230794

because it's trained on a lot of text data. The GPT-1 is a pretraining model based on the Transformer architecture, designed to enhance natural language understanding through large-scale unsupervised learning. GPT-1 initially undergoing self-supervised learning on extensive text data and subsequently demonstrating excellent performance across various natural language processing tasks through fine-tuning. This research laid the foundation for subsequent GPT model iterations and has driven significant advancements in the field of natural language processing.

The GPT model has been developed, with significant enhancements. For example, GPT-2 and GPT-3, these equipped with more parameters to better capture language patterns. Furthermore, fine-tuning the pre-trained model on specific downstream tasks has enabled GPT to achieve state-of-the-art performance across various natural language processing tasks. Furthermore, the careful prompt design allows researchers to guide the model perform a wide range of tasks. However, it's important to note that GPT models come with challenges, including substantial computational and resource requirements and concerns about potential bias in content generation.

Each of these models, from BERT to ALBERT, Transformer-XL, XLNet, and GPT not only represent a significant advancement in NLP with unique strengths and applications, but also some limitations and trade-offs. Researchers must choose the suitable model based on requirements and performance criteria of specific task.

3.3. Bidirectional Model Application

The bidirectional model's success can be attributed in part to its pre-training tasks, specifically the Masked Language Model and Next Sentence Prediction. Pre-trained on large amounts of textual data, BERT has gained profound linguistic comprehension capabilities, making it an invaluable asset for a variety of NLP tasks. Its applications span the fields of sentiment analysis, named entity recognition and translation.. Furthermore, The capability of bidirectional model proves invaluable for handling complex text with intricate dependencies, making them adept at tasks involving ambiguity, intricate grammar, or context-dependent elements. In the realm of textual analysis, bidirectional models excel. Their comprehensive understanding of contextual relationships allows them to generate text that is more coherent and logically structured. In the domain of question answering, they also provide more accurate answers. Furthermore, in the field of sentiment analysis, where context plays a crucial role, bidirectional models excel in capturing subtle nuances of emotions within text. Named Entity Recognition also can get benefits from bidirectional models as they more effectively capture contextual cues related to entity names, contributing to information extraction and text labeling. Bidirectional language models have advantages in understanding and generating text in context. The applications span a wide range of natural language processing tasks, enhancing accuracy, coherence, and adaptability across various domains and challenges.

4. The trend of natural language processing development

Multimodal fusion, knowledge graph reasoning, cross-domain applications, standardization, and ethical considerations will become the directions for research. The fusion of Graph Neural Network technology with NLP is poised to elevate the capabilities of NLP in handling synthetic data to new heights. It not only enhances the processing of individual textual data but also extends to handling multimodal data, including images, audio, and video, within synthetic information. The innovative potential lies in representing knowledge graphs in graphical form and subsequently integrating them with mainstream NLP models. This integration not only augments the model's capacity to express knowledge but also enhances its ability to fuse information, enabling comprehensive comprehension and analysis of data from diverse sources.

Multimodal knowledge fusion as a significant trend that means integrating not only textual data, but also various data types into knowledge graphs, such as images, audio, and video. This will make NLP systems more comprehensive, capable of handling diverse information. For example, for a particular entity, we can acquire not only its textual description but also related images and audio data, enabling a more comprehensive understanding and analysis. Another crucial trend will be the development of

DOI: 10.54254/2755-2721/42/20230794

knowledge graph reasoning. This implies that knowledge graphs will no longer remain static data repositories but will be used for automatic inference of new knowledge. Through reasoning, systems can automatically discover hidden relationships or attributes among entities, enriching the content of knowledge graphs. This will enhance the dynamic nature and adaptability of knowledge graphs, making them better suited to meet evolving information needs [11]. The combination of these two trends will propel NLP technology to make greater strides in handling multimodal data and knowledge reasoning, providing us with more comprehensive and intelligent information processing capabilities. It will also help NLP systems better understand and respond to the ever-growing and diverse sources of data.

Large language models have emerged as a main trend in the field of Natural Language Processing, and it represent future direction with substantial impact. The large language models are characterized by their vast parameter count and robust computational power, and it have significantly elevated performance across a spectrum of NLP tasks which include language comprehension, generation, and translation. Their remarkable generalization capabilities, acquired through extensive pre-training on large datasets, enable them to adapt seamlessly to diverse tasks and domains, reducing the need for task-specific customization.

For instance, Evaluating the quality of generated text in summarization is a multifaceted challenge. Language assessment encounters a significant gap between established metrics and human judgment, encompassing both objective aspects like grammatical and semantic accuracy and subjective factors such as completeness, brevity, and engagement. Large Language Models provide a comprehensive evaluation framework by comparing generated text with reference text, considering both objective and subjective perspectives. Empirical results from experiments on two real summarization datasets demonstrate our model's competitive performance and strong alignment with human assessors. This underscores the potential of large-scale models in addressing the intricate aspects of text assessment and generation within the realm of NLP [12].

Exploring domain adaptation and transfer learning strategies becomes imperative. It means using NLP methods when you don't have much data or when you're dealing with very different topics. Low-resource languages encompass minority languages, regional dialects, and seldom-used languages. It diminishes the applicability of conventional deep learning approaches. Developing extensive annotated datasets for these languages is an expensive endeavor, necessitating the pursuit of cost-effective annotation methodologies. Cross-lingual transfer methods assume a pivotal role in enhancing the performance of low-resource languages by harnessing insights from other languages. Conversely, cross-domain NLP tasks mandate NLP models to exhibit robust adaptability across diverse domains. These tasks involve reconciling domain-specific disparities in terminology, syntax, and contextual nuances. Given the high expenses associated with annotating data for different domains, exploring domain adaptation and transfer learning strategies becomes imperative. Moreover, crafting models tailored to fulfill the requisites of domain-specific NLP tasks proves pivotal for effective cross-domain NLP. The zero-shot learning which involves text classification, and processing using machine learning models without annotated data. It emphasizes the latest NLP models that possess zero-shot learning capabilities, enabling them to perform text classification without prior labeled data. These models hold great utility for various practical applications, particularly in cases where acquiring large-scale labeled data can be challenging or costly, such as in specific domains or languages [13]. The development of these techniques is expected to further advance research and applications in the field of NLP, enhancing its scalability and adaptability.

5. Conclusion

This paper mainly focuses on the evolution of language processing models in the field of Natural Language Processing, with a special emphasis on the transition from unidirectional to bidirectional modeling. We delved into various aspects of the transformer model, including its architecture, pre-training tasks, and performance in NLP tasks. Through this research, our goal is to gain a comprehensive understanding of the evolution of the transformer model and its implications for the field of NLP.

DOI: 10.54254/2755-2721/42/20230794

The field of Natural Language Processing has undergone a significant evolution from unidirectional models to bidirectional models. At the first, the unidirectional models dominated filed of natural language process. However, it still had limitations in understanding and generating text, since they could only use the context they had observed up to that point. But with the rise of deep learning, bidirectional models like the Transformer that became significant breakthroughs. These models introduced self-attention mechanisms, and enable them to simultaneously consider all words in the context. For the result, it can get substantial improvement in text processing performance. Methods like Transformer, BERT, XL-Net, GPT achieved outstanding pretraining through large-scale self-supervised learning, and delivering remarkable results across various NLP tasks. The NLP field will continue to evolve towards greater diversity and universality, including multimodal fusion, knowledge graphs and reasoning, low-resource languages, and cross-domain applications. Simultaneously, the NLP community needs to address standardization and ethical issues in order to ensure fair and ethical use of the technology. This series of developments will drive the forefront of NLP technology, offering more possibilities for various application scenarios while maintaining sustainability and credibility.

References

[1] Vaswani, Ashish, et al. “Attention is all you need”. Advances in neural information processing systems 30 (2017).

[2] Subakti, A., Murfi, H. & Hariadi, N. “The performance of BERT as data representation of text clustering”. Journal of Big Data 9, 15 (2022).

[3] Lin, Tianyang, et al. "A Survey of Transformers." Artificial Intelligence Open, 34 (2022). https://doi.org/10.1016/j.aiopen.2022.10.001.

[4] Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." arXiv preprint arXiv:1901.02860 (2019).

[5] Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." Advances in neural information processing systems 32 (2019).

[6] Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." arXiv preprint arXiv:1909.11942 (2020).

[7] Salehinejad, H., Sankar, S., Barfett, J., Colak, E., & Valaee, S. Recent Advances in Recurrent Neural Networks. arXiv preprint arXiv:1801.01078 (2017).

[8] Bengio, Y., Simard, P., & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2), 157–166 (1994).

[9] Wadawadagi, Ramesh, and Veerappa Pagi. "Sentiment analysis with deep neural networks: comparative study and performance assessment." Artificial Intelligence Review 53.8: 6155-6195 (2020).

[10] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[11] Schneider, Phillip, et al. "A decade of knowledge graphs in natural language processing: A survey." arXiv preprint arXiv:2210.00105 (2022).

[12] Wu, Ning, et al. "Large language models are diverse role-players for summarization evaluation." arXiv preprint arXiv:2303.15078 (2023).

[13] Davison Joe. 2020a. “New Pipeline for Zero-Shot Text Classification.” Retrieved December 28, (2021).