Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/13/20230737

Research on transformer and attention in applied algorithms

Junyan Zhang

Department of Computer Science, Wenzhou-Kean University
Daxue Road 88, Ouhai District, Wenzhou, China, 325060

zhangjun@kean.edu

Abstract. The transformer is an encoder-decoder-based structure and model for deep learning that completely utilizes the self-attention mechanism. It has gained remarkable success in natural language processing and computer vision and is becoming the predominant research direction. This study first analyzes the transformer and attention mechanism, summarizes their advantages, and explores how they help the recommendation algorithm dynamically focus on specific parts of the input that are helpful to perform the current recommendation task. After analyzing the framework of the attention mechanism network and its weight computation for data received. To further enhance the practicality of objects in natural situations and the precision of object recognition, a transformer detection approach based on deformable convolution is presented. And analyzed how the transformer works in the generative pre-trained transformer. These algorithms illustrate the efficacy and robustness of the transformer, indicating that the transformer that incorporates the attention mechanism may satisfy the requirements of the majority of deep learning tasks. However, the unpredictability of demands, the exponential growth of information, and other issues will continue to make it challenging to deal with global interaction mechanisms and a unified framework for multimodal data.

Keywords: deep learning, transformer, attention mechanism, recommendation algorithm, object detection algorithm, generative pre-trained transformer.

1. Introduction

In artificial intelligence (AI), deep learning has emerged in many technologies and networks for processing images and text, such as convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM), and generative adversarial network (GAN), which have numerous applications in computer vision (CV) and natural language processing (NLP). Since deep learning evolved from the study of the human brain, the neural network structure resembles that of the brain itself. The neurons can accept, process input signals, and send output signals. In addition, the relationship between each neuron and other neurons is evaluated by weight, reflecting its importance in the neural network.

An attention mechanism is proposed as a result of the continual development and progress of deep learning, primarily because people live in an era of information explosion. However, the information that humans can receive is limited. With a large amount of information, the human brain can quickly focus on essential areas for critical processing and ignore other information through its powerful information-processing capabilities. When a scene enters the human eye, for instance, humans prefer to concentrate on some essential spots, such as dynamic points or sudden hues, while the remainder of the

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/13/20230737

static picture is momentarily disregarded [1]. This selective perception mechanism dramatically reduces the amount of processed data. This attention mechanism is essential for efficiently allocating information-processing resources and focusing on important information with a higher weight and improving the efficiency and accuracy of information processing.

The transformer has a significant impact on AI, and research on the transformer has many benefits and values. For the whole society, its progress can bring more effective and efficient solutions for various tasks, including improving communication through language translation and strengthening decision-making through improved sentiment analysis, among others.

The future of the transformer is promising, with it expected to continue to evolve and improve. It can be foreseen that future enhancements of the transformer include improving performance and accuracy, integrating with other AI techniques to produce more powerful and general AI systems, increasing efficiency to reduce computational costs for widespread adoption, and designing more specialized models for training in specific domains to improve the efficiency of specific tasks. Therefore, the transformer can offer potential future benefits and developments in almost all areas of AI.

2. Transformer model architecture

Transformers were proposed by a team at Google in 2017 and are increasingly the model of choice for NLP problems [2]. As a new type of network, the transformer network has gradually become a hot topic in deep learning and has achieved good results in computer vision.

2.1. Transformer structure

The transformer network consists of six identical encoder blocks and six identical decoder blocks, and the module structure is shown in Figure 1. The transformer encoder block consists of a multi-head attention layer and a feed-forward neural network. The decoder block consists of a multi-head attention layer, a masked multi-head attention layer, and a feed-forward neural network [3]. Transformer uses a masked multi-head attention mechanism to cover up anonymous information to prevent the model from referring to unknown subsequent information during training. In order to adapt to the task of dealing with variable-length input sequences and improve network stability, the transformer adopts layer normalization operation. It is proposed to use position encoding to solve the problem that the transformer cannot obtain the position information between input elements like RNN and CNN structures. The transformer uses trigonometric functions for position encoding, as shown in formula (1).

$$\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (1)$$

DOI: 10.54254/2755-2721/13/20230737

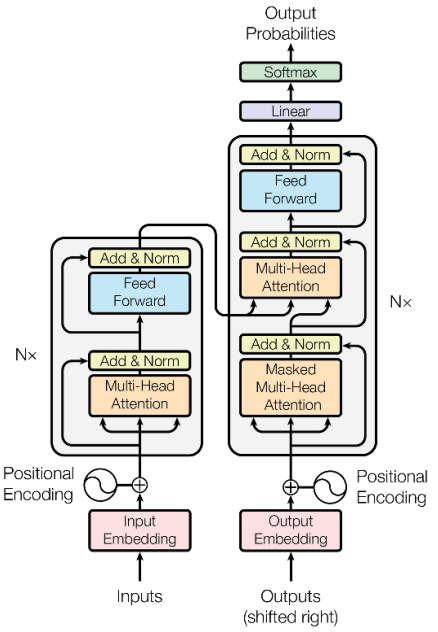

Figure 1. The Transformer - model architecture [2].

2.2. Attention mechanism

An attention function is a mapping between a query and a collection of key-value pairs and an output, where the query, key, value, and output are all vectors. The result is generated as a weighted sum of values, where the compatibility function of the query determines the weight associated with each value and its related key. The multi-head attention mechanism adopted by the transformer structure enables the model to learn different semantic features in different sub-layer spaces and adjust the weights and different projection methods during the projection process to make the model more generalized.

Mnih et al. originally proposed the concept of the attention mechanism [4]. They believed that it highlights the influence of a key input on the output by calculating the weight of the input data, that 5 data (x_1, x_2, x_3, x_4, x_5) provide the input to the attention mechanism network structure. Following end-to-end training, each data will be weighted appropriately (attention score). In the training phase, the goal output determines the weight's value. Figure 2 demonstrates the structure.

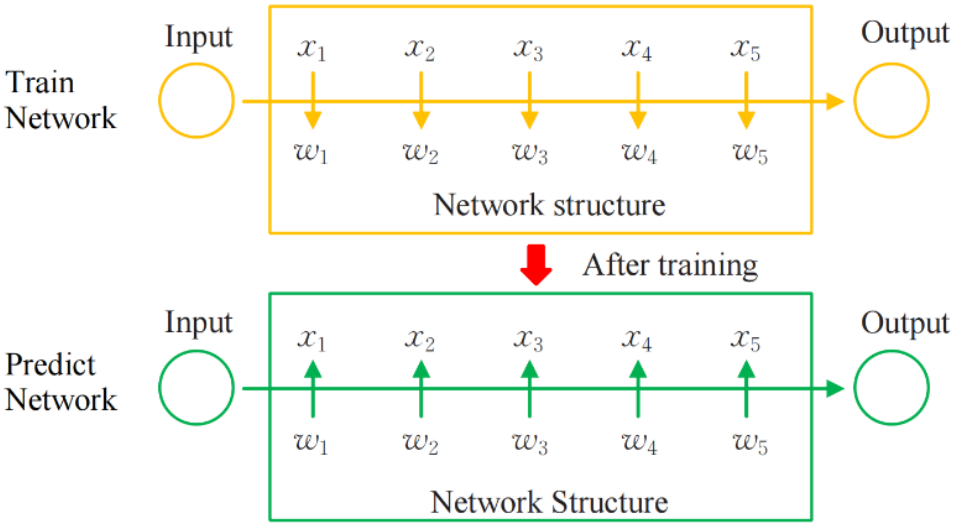

Figure 2. Network of attention mechanism [5].

DOI: 10.54254/2755-2721/13/20230737

The standard attention algorithms are the dot product algorithm and the additive algorithm. Q, K, V vector attention calculation adopts dot product attention algorithm.

To narrow the gap between elements and prevent the softmax from missing discrimination, a scaling factor $\sqrt{d_k}$ is ed in the ordinary dot-product attention mechanism. The attention calculation is shown in formula (2), where Q, K, V are the vector representations obtained by a linear transformation of the original input Token, and T is the transposition operation.

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Compared with the additive attention algorithm, Dot product attention is characterized by easy calculation, more space-saving, and faster practice because it can be implemented using highly optimized matrix multiplication code.

2.3. Transformer advantages

The Transformer is the pioneering model to be built with a pure attention mechanism, enabling not only faster computation, but also superior results in translation tasks, and lays the foundation for subsequent bidirectional encoder representation from transformers models. The transformer model, which replaces LSTM with a complete self-attention structure, yields better translation results. By addressing the slow training shortcoming of RNNs and utilizing a self-attention mechanism, the Transformer enables fast parallelism. Additionally, the Transformer can be scaled to a deeper depth to fully leverage the characteristics of deep neural network models, thereby further improving model accuracy.

Compared with CNN, the number of operations required to calculate the association between two locations does not increase with distance [6]. The correlation between each word can be directly calculated without passing through the hidden layer. Self-attention can produce more interpretable models [7]. The distribution of attention in the model is observable, and each attention head is capable of learning to perform distinct tasks.

For long-range dependencies, the transformer can learn global information more easily. This advantage depends on the usage strategy of attention calculation that all input instances are calculated between two. Transformer models have weaker inductive biases compared to CNNs and RNNs. While theoretically not as powerful as RNNs, given sufficient data and massive scale, transformer models have proven to outperform their "competitors" with strong inductive biases eventually. Transformer models will be particularly suitable for unsupervised pre-training with sufficient computing power and sufficient data scenes.

For the transformer, the remarkable advantages of the attention mechanism promote the strength of the transformer. It can focus on relevant information while ignoring irrelevant information, and directly establishes the dependency between input and output without cycling. It also enhances the degree of parallelism and significantly improves the running speed. In addition, the attention mechanism overcomes some limitations of traditional neural networks, such as the system's performance decreasing with the increase of input length, the computational efficiency of the system being low due to the unreasonable input sequence, and the system lacking feature extraction and enhancement. However, the attention mechanism can well model sequence data with variable lengths, further enhancing its ability to capture remote dependency information and effectively improving accuracy by reducing the depth of the hierarchy.

3. Applied algorithms of transformer and attention

3.1. Recommendation algorithm

At present, the attention mechanism has been applied in classic recommendation scenarios such as click-through rate prediction, multimedia recommendation, score prediction, group, and bundle recommendation, and has been proven to have many advantages in improving recommendation

DOI: 10.54254/2755-2721/13/20230737

performance. First, the attention mechanism could collect global relationships in a single step while also focusing on local data relationships. Secondly, it has a more vital ability to capture long-term dependence; It has higher parallelism and can reduce model training time; Then, it has higher scalability and robustness; Additionally, the structure is relatively simple with few parameters; Moreover, it can enhance the model's interpretability to a limited degree; Finally, it can quickly extract important features of sparse data and their dependencies [8].

It is conceivable for the attention mechanism to assist the recommendation model in rapidly recognizing the most informative features, recommending the most representative items, and enhancing the interpretability of the model to a certain extent. Attention mechanism technology still has some open problems in recommendation diversity, recommendation explainability, and fusion of multiple sources of information, which are worth discussing.

The majority of currently used recommendation models do not have a thorough comprehension of the auxiliary data, which is most evident in the following three areas. First, the supplementary information is more complex and includes heterogeneous properties from multiple sources. Second, no consideration is given to the combination of any various types and any number of qualities. Third, the estimated attributes do not have any higher-order cross features. To generate a semantically richer and more precise representation of the item, how to utilize the attention mechanism to achieve early integration of information from all parts so that all parts complement and inspire each other to uniformly embed auxiliary information into the potential semantic space of the item, ultimately improving the scalability of the model and improving recommendation performance will be the exploration direction and research focus of the academic community in the future.

3.2. DETR detection algorithm based on transformer

With the application of the transformer in computer vision, the transformer has gradually become one of the leading research algorithms in target detection. A detection transformer, as known as DETR, is the most basic transformer-based detection algorithm in target detection. It removes the non-maximum suppression and anchor operations from the current mainstream detectors, making the detection process clearer and more accurate. It is a simple end-to-end target detection processing algorithm.

Encoder and decoder structures in the transformer comprise the core of DETR [9], which fully uses the attention mechanism in both two structures to establish global and regional relationships. Then pays attention to the local area to obtain the target area of interest in the picture and introduces the multi-head attention mechanism, which focuses on the target object in the picture from multiple angles [10]. DETR divides the detection task into feature extraction and target prediction, and the structure mainly includes CNN, transformer, and feed-forward neural network [11]. The CNN serves as the primary network for feature extraction in the DETR architecture. The transformer component of DETR contains both encoders and decoders, while a feed-forward neural network is utilized for predicting the category and location of objects. The DETR architecture employs a residual neural network as its backbone network for generating feature maps which are then transformed into sequences and subjected to a position encoding strategy. These sequences are then fed into the transformer's encoder, and the output of the encoder is combined with a set of object queries. The decoder then processes this input, leading to the prediction of the object’s position and category through a feed-forward neural network.

Transformers' detection algorithm has shown positive detection results, but in order to increase object detection performance, DETR only adds transformers after the backbone network. The ResNet backbone still uses CNN convolution to extract features because there is no other workable option. Therefore, this research suggests using deformable convolutions based on transformers to tackle this problem in order to better extract the features of various objects and improve the capability of feature extraction in the backbone network. Deformable convolution can adaptively alter the receptive field in accordance with various properties of various objects to more effectively focus on the target object [12]. Combining the robust modeling of the link between each pixel position with the Transformer and the adaptive adjustment of the feature receptive field of the deformable convolution.

Compared with ordinary convolution, deformable convolution increases the offset, dynamically

DOI: 10.54254/2755-2721/13/20230737

adjusting the sampling lattice. The offset makes the receptive field not limited by the fixed shape during the convolution operation. Therefore, the position information of spatial sampling is further offset and adjusted.

3.3. Generative pre-trained transformer

Transformer architecture based on the attention mechanism is now extensively applied in NLP, and it has been proven to achieve extraordinary performance on many NLP tasks, including text classification, language translation and generation, and question answering. Furthermore, many pre-trained language models based on the transformer have been developed, such as generative pre-trained transformer (GPT) [13] and bidirectional encoder representations from transformers (BERT) [14]. Generative pre-training involves training a deep neural network to predict masked words in a large corpus of text data. Then it produces a new transformer model, GPT, which is further refined through a process of fine-tuning on a downstream task, leading to the achievement of cutting-edge performance [13].

GPT is a series of large-scale NLP models that utilize the transformer as a core component to effectively learn complex relationships between words and produce coherent and human-like responses across a broad spectrum of natural language inputs.

An important advantage of GPT models is their versatility in dealing with various NLP, which can be generally used for any program related to natural language. These models can capture and learn complex language patterns through large-scale pre-training, thereby generating thinking logic close to humans that imitate human language and communication. Additionally, GPT models can rapidly adapt to changes in input data, thus reducing the time and cost associated with developing NLP models.

However, the GPT model has several disadvantages. Developing and training a new GPT model requires significant computing resources, which imposes high software and hardware requirements on the developer and company, so that may not be available to small companies and individual developers. Therefore, the complete deployment of GPT presents significant challenges. Furthermore, while the feedback provided by GPT is trained on a vast quantity of data, the lack of diversity and accuracy in the data can lead to deceitful and biased results, some incorrect data even from users. GPT also does not consider the target user when outputting results, which can lead to offensive, uncomfortable, and unsuitable responses, particularly in sensitive areas like politics, religion, and children.

Although there are disadvantages associated with GPT models, their future of them still shows encouraging potential for further development and application. As more data is collected and new research is carried out, the quality of its generated text can be continuously improved, as well as its ability to handle more complex problems and deal with domain-specific tasks, and bring better user experience.

Research on GPT models contributes to creating new transformer architectures or improving existing ones and exploring new applications. ChatGPT developed by OpenAI is an example of a GPT model that exemplifies the effectiveness of pre-training on a massive corpus of text data and fine-tuning for specific NLP tasks. This pre-trained language model has been proven and applied in various areas, including chatbots, content generation, virtual assistants, and customer service. Its impressive capabilities in language processing and generation make it a useful chatbot.

4. Discussion

Although transformers have demonstrated their capabilities in various tasks, challenges remain. In addition to efficiency and generalization constraints, the transformer may benefit from a subsequent theoretical study, an improved global interaction mechanism, and a unified framework for multimodal data.

The transformer's architecture has been shown to support large-scale datasets with appropriate parameters for training. According to the results of several experiments, transformers have a higher capacity than traditional models like CNNs and RNNs, and thus can handle large amounts of training data and generally perform better when trained on enough data. Nevertheless, the transformer usually has few prior assumptions about interpretability, which requires some theoretical analysis.

DOI: 10.54254/2755-2721/13/20230737

Current neural networks may not be able to check the deviations in the system and provide specific explanations for a well-functioning system. A significant benefit of transformers is the ability to represent global dependencies between nodes in the input data through the utilization of attention mechanisms. Many types of research, however, have demonstrated that the majority of nodes do not require complete focus. To a certain degree, computing the attention of all nodes is inefficient. Thus, there is still considerable opportunity for improvement in accurately predicting global interactions.

The self-attention module can be considered a fully connected neural network with dynamic connection weights that aggregates non-local input via dynamic routing. Consequently, it is worthwhile to investigate alternate dynamic routing mechanisms. Global interactions can also be modeled by different types of neural networks, including memory augmentation models.

Integrating multimodal data is essential and beneficial for enhancing task performance. As transformers have produced outstanding outcomes in numerous domains, it is necessary to develop a unified framework that better captures the fundamental relationships between multimodal data. The transformer's existing multimodal attention mechanism should be enhanced.

5. Conclusion

The transformer is the first model built with pure attention, which computes faster and performs better on a variety of tasks, utilizing the technique of self-attention to obtain speedy parallelism. Moreover, the depth of the transformer may be expanded to profound depths. The performance of the recommendation algorithm using the attention mechanism has advantages, such as reducing the training time of the model through parallelism, stronger scalability and robustness, and faster extraction of important features of the data. The core of another algorithm DETR in this paper is to use the transformer architecture to make full use of the attention mechanism to establish the global and regional relationship between the detection target and the image. An efficient idea is to use transformer-based deformable convolutions to improve object detection performance. The last application is GPT, which is a concrete implementation of the transformer designed for NLP, and the author conducts a critical evaluation of its advantages, disadvantages, and bright future.

Although the current network has shortcomings in terms of multi-modality, unified framework, and interpretability, it is believed that after improving and integrating the structure, the network must have high research significance and implementation value in future deep learning. This paper, as research on transformers, lacks an introduction to the history of transformer development. And it only explains and analyzes two applied algorithms related to transformer and attention, and lacks an in-depth discussion on the application of transformer in NLP, CV, etc. In addition, due to time and resource constraints, the author did not conduct experimental verification and cannot prove the actual performance of the transformer. Therefore, future work will focus on the NLP applications of transformers, especially GPT for Conversational AI. Furthermore, future work should also focus on experimental verification. Select a common daily problem as an application scenario to prove the performance of the transformer, analyze the results, then improve model performance, explore more application scenarios, and better utilize the advantages of the transformer.

References

[1] Chaudhari, S., Mithal, V., Polatkan, G., Ramanath, R. An attentive survey of attention models. ACM Transactions on Intelligent Systems and Technology (TIST), 12(5), 1-32 (2021).

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 30 (2017).

[3] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).

[4] Mnih, V., Heess, N., Graves, A. Recurrent models of visual attention. Advances in neural information processing systems, 27 (2014).

DOI: 10.54254/2755-2721/13/20230737

[5] Gao, G. Survey on Attention Mechanisms in Deep Learning Recommendation Models [J]. Computer Engineering and Applications, 2022, 58(9): 9-18 (2022).
[6] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 1580-158 (2020).
[7] Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., Le, Q. V. Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541 (2018).
[8] Gong, Y., Zhang, Q. Hashtag recommendation using attention-based convolutional neural network. In IJCAI, 2782-2788 (2016).
[9] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. End-to-end object detection with transformers. In European conference on computer vision. Springer, Cham, 213-229 (2020).
[10] Ba, J., Mnih, V., Kavukcuoglu, K. Multiple object recognition with visual attention. arXiv preprint arXiv:1412.7755. (2014).
[11] Wang, M., Huang, S., Liu, L. Transformer detection algorithm based on deformable convolution. Information Technology and Informatization (07), 199-201+205. doi:CNKI:SUN:SDDZ.0.2022-07-05. (2022).
[12] Meinhardt, T., Kirillov, A., Leal-Taixe, L. Feichtenhofer, C. Trackformer: Multi-object tracking with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8844-8854 (2022).
[13] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. Improving language understanding by generative pre-training (2018).
[14] Devlin, J., Chang, M. W., Lee, K.,Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).