Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/20/20231084

Analysis of transformer-based models for time series data, natural language processing, and computer vision

Huanzhang Chen[1, †], Yunfan Hou[2, †], Tianhao Miao[3, 5, †] and Jiayu Xue[4, †]

[1]SDU-ANU Joint Science College, Shandong University, Weihai, Shandong, 264200, China
[2]School of Telecommunications Engineering, Xidian University, Xi'an, Shaanxi, 710126, China
[3]College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, 201418, China
[4]Faculty of Engineering, University of Sydney, Sydney, NSW 2006, Australia

[5]1000500993@smail.shnu.edu.cn
[†]These authors contributed equally

Abstract. The birth of the Transformer revolutionarily signalled the start of a new epic chapter in the deep learning era. Through an encoder-decoder architecture, including residual connection, multi-head self-attention, etc., it completely reformed the deep models and unified the models used in traditional computer vision (CV) and natural language processing (NLP) problems. In recent years, many papers published have adapted the original Transformer model to better complete tasks in time series analysis, CV, and NLP. In the area of natural language processing, Bidirectional Encoder Representations from Transformers (BERT) employs a two-way transformer structure to learn context-based language representation, whereas Generative Pre-trained Transformer (GPT) employs a one-way transformer but enhances corpus training to enhance the model effect. The Vision Transformer model is the cornerstone of computer vision. It separates the input image into various patches, projects each patch into vectorized features, and then passes the them to Transformer. Based on the idea of the Vision Transformer, Swin Transformer and Biformer further optimized the Transformer and achieved better results. Time series combines the ideas embodied in CV and NLP, and in doing so, improves the specificity and various difficulties of time series problems to lower algorithm complexity and increase prediction accuracy. This article summarizes the uses and improvements of the Transformer in NLP, CV and time series, explores the development history and ideas on algorithm optimization, and predicts the potential developments of Transformer in these three fields.

Keywords: transformer, deep learning, artificial intelligence.

1. Introduction

The birth of the Transformer revolutionarily signalled the start of a new epic chapter in the deep learning era. Researchers proposed a novel neural network for sequential data transduction tasks [1]. The Transformer relies merely on an attention mechanism for achieving global reasoning capabilities.

© 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: 10.54254/2755-2721/20/20231084

The Transformer has been shown to obtain state-of-the-art performances on several sequence transduction tasks, including question answering, machine translation, and text summarization. Transformer architecture is now used in a variety of industries, with its main applications being in NLP, CV, and time series [2].

In the following chapters, representative research, and applications of the aforementioned three fields will be sequentially introduced and analysed.

2. Transformer in time series data

Time series data is a function that takes time as independent variable. It reflects the continuous change of a random variable over time. Traditional Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) and other algorithms have been relatively mature, but still have problems such as insufficient efficiency and insufficient accuracy in handling long series time inputs; while this paper focuses on three points: long-term trend, seasonal variation, and irregular variation, and optimizes the performance of various aspects of time series forecasting by improving the Transformer architecture.

2.1. Informer

For prediction leveraging long time series data, Transformer is optimized from the perspective of efficiency. Figure 1 demonstrates its detailed architecture [3]. Informer implements long sequence input and output, and solves the complexity problem by selecting the highest attention score with ProbSparse. For long-period time series, the input dimension is high, and the time complexity of Transformer increases exponentially with the length of the input sequence; through experiments, the attention score reduces a large amount of computation by only letting the key and vital query form sparse attention in order to improve the efficiency. Therefore, Informer proposes to calculate the Kullback-Leibler (KL) scatter of the attention score scoring distribution and uniform distribution for each query.

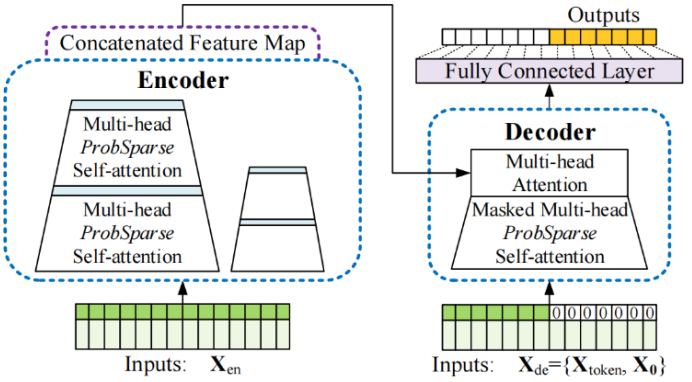

Figure 1. Architecture of Informer [3].

In addition, Informer introduces self-attention distillation by adding a convolution between every two layers of Transformer to reduce the sequence length by half, which further reduces the training overhead. In the Decoder stage, the method of predicting multiple time-step results at one time is used to alleviate the cumulative error problem compared with the traditional method.

2.2. Autoformer

Architecture of Autoformer is displayed in Figure 2. Encoder inputs historical time series, decoder part of the input consists of two parts: trend term and seasonal term [4]. The trend term is composed of the latter half trend data and data average filling after series decomposition block processing, and the seasonal term is composed of the recent seasonal term and zero filling, both of which indicate that the future data is not known. the encoder part is mainly for modeling complex seasons. decomposition is

DOI: 10.54254/2755-2721/20/20231084

used to extract the seasonal terms from the original series, which is used as information for decoder to predict the future seasonal terms.

The core of Autoformer is the series decomposition block module uses an additive model, which considers that the time series is composed of trend and seasonal terms. The seasonal one is achieved using the sliding average method, where the average value is calculated for each window on the original input time series, and after getting the trend term of each window, the trend term of the whole series is finally obtained. According to the additive model, the trend term is subtracted from the series to obtain the seasonal term.

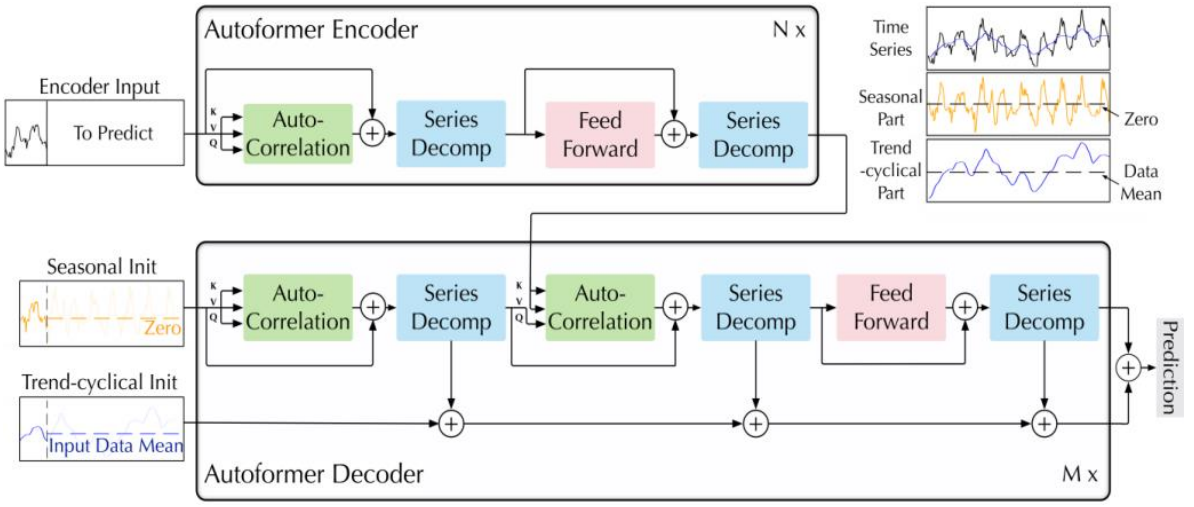

Figure 2. Architecture of Autoformer [4].

Meanwhile, Autoformer upgrades the multiple attention mechanism to Auto-Correlation Mechanism. The auto-correlation is used to find the most correlated segments of the time series. The auto-correlation coefficient of a time series calculates the correlation coefficient and its sliding one step later. Autoformer uses this property to calculate the auto-correlation coefficient of each sliding step and selects the sliding step with top k correlation coefficient. The auto-correlation coefficient is computed by Fast Fourier Transform, which has the advantage of finding similar segments by using the translation of the time series, while the auto-correlation mechanism focuses only on the relationship between points.

3. Transformer in NLP

NLP is a branch of the field of computer science that aims to, interpret, and generate text or speech using computer. NLP utilizes computer algorithms to process language data for downstream tasks such as text classification, generation, speech recognition, etc. The goal of NLP is to enable computers to process and understand natural language as humans do in order to interact and communicate more effectively with humans. GPT and BERT have promoted the development of NLP field, and their development is also an important milestone in the development of NLP.

3.1. GPT

GPT is a series of natural language processing models introduced by OpenAI, which uses unsupervised learning for training. While GPT-1 was excellent at some tasks, it had some problems when generating long text, such as inconsistencies and incoherence [5].

GPT-2 is an upgraded version of GPT-1. The most significant improvements are the larger model size and more parameters, up to 150 million, as well as some improvements in training data sets and preprocessing methods. These improvements make GPT-2 even better at not only performing a variety of natural language processing tasks, but also producing realistic natural language text [6].

DOI: 10.54254/2755-2721/20/20231084

GPT-3 and GPT-2 is different in the model size. GPT-3 has 175 billion parameters, 10 times as many as GPT-2. This change in model size improves the performance of many natural language processing tasks, producing excellent text even if only a few examples are given to learn from. As a result, GPT-3 is a significant improvement over GPT-2 in its ability to complete a variety of language tasks [7].

In general, the development process of GPT series models is mainly improved from the model scale, training data set and preprocessing methods, and these improvements have improved the performance.

3.2. BERT

It is another language model proposed in 2018. Unlike the traditional language model, Bert has a two-way structure for extracting context-sensitive language representation, which achieves great results for many NLP tasks at the time. BERT improves on the model structure compared to GPT by adopting a two-way structure for extracting context-sensitive syntax vocabulary. GPT leverages a one-way structure. In the comparative analysis of GPT 1.0 and BERT, it can be seen that the performance improvement of BERT over GPT 1.0 is mainly due to the difference between the bidirectional language model and the unidirectional language model. In BERT model, the [Mask] tag is also introduced, which is used to simulate the data randomness in the pre-training stage, but it leads to the inconsistency between the pre-training and fine-tuning stages.

3.3. Advantages and disadvantages analysis

Transformer model has the following advantages and disadvantages. There are mainly four advantages. Firstly, it could process long sequences. Traditional RNN and Convolutional Neural Network (CNN) models have weak processing capacity for long sequences, while Transformer uses a self-attention mechanism. By weighting information at different positions in the sequence, the representation of each position can take information at other positions into account at the same time, thus better adapting to input sequences of different lengths and effectively processing long sequences [8]. Secondly, it has high parallel computing efficiency. Compared to models such as RNN, Transformer is capable of parallel computing because its self-attention mechanism has the following characteristics: The calculation of each position only needs to use the information of itself and all positions in the sequence, instead of relying on all previous positions in the sequence like RNN, so the calculation can be carried out in parallel [1]. Thirdly, it has excellent machine translation effect. The attention mechanism enables the model to better deal with long-distance dependence problems, and can parallelize the calculation, which speeds up the training speed and reasoning speed of the model. In addition, location coding and residual connection mechanism in Transformer model can also help to improve model stability and prediction ability of data not found in training set. Fourthly, it has strong pre-training capability. Transformer can pre-train on large amounts of unlabelled data, improving its performance on downstream tasks. First, Transformer's self-attention mechanism can model each element in the sequence to capture richer language structure information. Self-attention mechanisms use attention to assign weights between different elements, thus enabling the model to model the entire sequence globally, rather than processing the sequence only step by step based on context, as traditional cyclic neural networks (RNN) do [5].

However, there are still many disadvantages. Firstly, it requires large amount of training data. Transformer model has many parameters and requires big data for training; otherwise, it could overfit [8]. Sensitive to location information: Transformer models are very sensitive to location information in sequences, so Transformer may not perform as well as traditional models such as RNN in some tasks that require consideration of sequential order [8]. Secondly, it requires long training time. Transformer has so many weights, so the training consumption is huge, and strong computing resources are needed to support the training.

DOI: 10.54254/2755-2721/20/20231084

3.4. Representative application

A question-and-answer system is an artificial intelligence technology that automatically answers a user's questions, usually involving text and voice input. To evaluate a question-and-answer system, many aspects are required to be considered, such as accuracy, speed, reliability, scalability, and user experience. Here are some methods and criteria for evaluating question answering systems. Here is an example of a paper assessing question answering systems: An Evaluation of Open-Domain Question Answering Systems [9]. This paper evaluates several open domain questions answering systems, including both traditional and deep learning methods. This paper evaluates the accuracy, speed, reliability, and scalability of these systems, and analyses their shortcomings and limitations. The evaluation results of this paper provide an in-depth understanding of open domain question answering systems and provide valuable guidance for developing better question answering systems.

4. Transformer in CV

In 1962, David Hubel and Torsten Weisel studied the visual cortex response in cats by using single eyelid suture, and then proposed the concept of receptive field. In 1980, Kunihiko Fukushima proposed the neural network structure Neocognition including convolutional layers and pooling layers. It became the theoretical core of modern CNN development. In 1989, Yann LeCun referred to the BP algorithm and proposed LeNet, becoming the world's first CNN model in the true sense, and further optimized LeNet-5 in 1998. In the following 20 years, the theoretical development of the CV field has advanced by leaps and bounds, and excellent models such as AlexNet, ResNet, and EfficientNet have emerged one after another. Now that the development of CNN has entered a bottleneck again, people urgently need a new direction to help the CV field move forward [10].

4.1. Vision transformer

In 2020, Google proposed Vision Transformer (ViT) in the CV field, which abandoned the traditional CNN method and completely ported the Transformer in the NLP field to the CV field to deal with classification tasks [11]. In this paper, ViT divides the input picture into multiple patches (16x16), and then processes each patch with patch embedding and position embedding. The patches are then fed into the Transformer Encoder, processed by Normalization, Multi-Head Attention, Feedforward Neural Network (FNN), and Multi-Layer Perceptron (MLP), and finally output by MLP Head. Its structure is displayed in Figure 3.

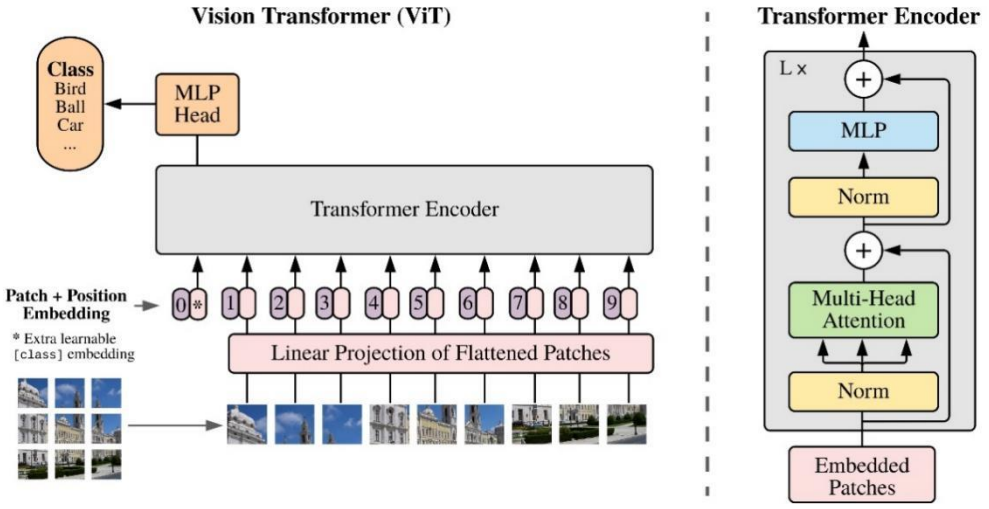

Figure 3. Architecture of ViT [11].

In the classification task, using the ImageNet dataset, the Top-1 Acc of ViT-H/14 model pre-trained by 300M JFT is 88.55%, BiT-L is 87.54, and Noisy Student is 88.5%, and the results are

DOI: 10.54254/2755-2721/20/20231084

similar.but the TPUv3-core-days required are different, the ViT-H/14 model only needs 2.5k, while BiT-L and Noisy Students need 9.9k and 12.3k, respectively. On other datasets such as CIFAR-10 and Oxford-IIIT Pets, ViT-H/14 model exhibits the same time advantage.

However, there are two major problems. Firstly, the variance of visual entities is large, and the performance of visual transformers may not be very good in different scenarios. Secondly, as the image resolution increases, the computational cost also increases. In fact, Transformer shows excellent performance over traditional CNNs on large-scale datasets, but the effect is not satisfactory on smaller datasets.

4.2. Swin transformer

Based on the previous problems, Swin Transformer (ST) introduces the sliding window mechanism and enables the model to process super-resolution images through hierarchical design, to save computation and pay attention to local information [12]. The whole model contains a total of 4 stages. At the beginning of the last three stages, the resolution will be downsampled first. Two similar Swin Transformer Encoders in the block structure are connected in series. The difference is that window multihead self-attention (W-MSA) adopts regular windowing mechanism, and subdivides into 4 windows for calculation based on the patch center point; shifting window multihead self-attention (SW-MSA) adopts shifted windowing mechanism, which splits the original patch into 9 windows according to the center point of 4 windows to realize the shift of the feature map. At the same time, set the mask for attention so that its calculation result is equivalent to the regular windowing mechanism. Overall, ViT uses a global attention mechanism for calculations, which reduces the amount of computation. ST, on the other hand, limits attention computation to each window.

Pre-training using the ImageNet-22K dataset, limited to 384^2 image size. The Swin-L model with 197M parameters made ImageNet Top-1 Acc reach 87.3% under the condition of 103.9G FLOPs, and its single classification processing time was 42.1s, while the original ViT-L/16 model with 307M parameters achieved 85.2% ImageNet Top-1 Acc under the condition of 190.7G FLOPs, and its single classification processing time was 27.3s. In comparison, although the processing time is slightly longer, ST improves the recognition accuracy under the conditions of smaller model parameters and lower computing power, and performs better than the original ViT model.

4.3. BiFormer

In order to overcome the two types of problems mentioned above, introducing sparse properties into the attention mechanism is also a feasible direction. Many studies are also working on optimizing Vanilla Attention mechanisms, such as Swin Transformer's inside local windows and Crossformer's dilated windows [13]. Most of these methods introduce sparseness characteristics through manual partitioning, thereby alleviating the problem of excessive computational complexity. To better optimize the application of the original Transformer in classification tasks, Lei Zhu's team of City University of Hong Kong proposed a dynamic sparse attention mechanism to achieve more flexible computing allocation and content awareness, so that it has dynamic query-aware sparsity. Based on this module, the team built a new general vision network architecture called BiFormer. The architecture of BiFormer is similar to Swin Transformer and consists of 4 stages. Except for the first one, each stage contains a Patch Merging section to downsample the input. In each Block, BiFormer does not directly take positional encoding, but uses 3×3 depth convolution to dark-position encodes relative position information. Bi-level Routing Attention (BRA) is introduced to filters out most of the irrelevant key-value pairs and reduces query sharing in the coarse-grained area by building an affinity map and then performing irrelevant pruning, and then applies fine-grained Tokens-to Token attention in the remaining routing area to achieve fast and low-resource image processing capabilities.

BiFormer-B uses fewer parameters (58M Params) and obtains a higher Top-1 Acc (85.4%) in a lower computing power environment (9.8G FLOPs), while Swin-B uses more parameters (88M Params) in a high computing power environment (15.4G FLOPs) but obtains a lower Top-1 Acc

DOI: 10.54254/2755-2721/20/20231084

(83.5%) than BiFormer-B. It shows that BiFormer's performance has been greatly improved compared to ST.

5. Conclusion

This paper provides an analysis of the use of Transformers in NLP, CV and time series. Overall, the evaluation of Transformer's performance on NLP tasks are covered based on discussion of the distinct architectures of BERT and GPT, the transition from the early development of CV to current applications of Transformer as well as the gradual improvements of Transformer in time series. In the future study of NLP, GPT and BERT models might be employed more frequently in automatic question answering, text classification, semantic analysis, text summarization, and other sectors. With further modification, they are expected to show stronger competence in processing multilingual and cross-lingual contents and may be combined with other technologies to create stronger NLP applications. In the area of CV, combining CNNs and Transformers is a promising approach for improving the performance of computer vision models. CNN + Transformer hybrid architectures combine the strengths of both transformer and CNN architectures, which use CNNs to extract local features from the image, and then use a transformer to learn relationships between these features. They could help to create more efficient and accurate models that could obtain state-of-the-art performances on several tasks such as optimization of computing resources and segmentation. In time series, the primary research focus will continue to concentrate on improving Transformer architecture. To extract the causes of time non-stationarity from text data at the same time, more NLP and CV-related models may be incorporated into the existing model. Then, input will be made sequentially using the method of segmenting time in the CV field, which allows the model to be built with more parameters, thus increasing its accuracy.

References

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

[2] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et, al. (2022). A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence, 45(1), 87-110.

[3] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on artificial intelligence, 35(12), 11106-11115.

[4] Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in Neural Information Processing Systems, 34, 22419-22430.

[5] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training, 1-12.

[6] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et, al. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[8] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et, al. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 38-45.

[9] Noguera, E., Llopis, F., Ferrandez, A., & Escapa, A. (2007). Evaluation of open-domain question answering systems within a time constraint. In 21st International Conference on Advanced Information Networking and Applications Workshops, 1, 260-265.

[10] Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., et, al. (2021). CNN variants for computer vision: history, architecture, application, challenges and future scope. Electronics, 10(20), 2470.

DOI: 10.54254/2755-2721/20/20231084

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et,al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[12] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, 10012-10022.

[13] Zhu, L., Wang, X., Ke, Z., Zhang, W., & Lau, R. (2023). BiFormer: Vision Transformer with Bi-Level Routing Attention. arXiv preprint arXiv:2303.08810.