Researches Advanced in the Development and Application of Transformers

Bo Cheng[1, *, †], Nan Guo[2, †], Yuanqing Zhu[3, †]

[1]Department of Applied Computing, Hang Seng University, Hong Kong, China

[2]College of Computer Science and Technology, China University of Petroleum, Qingdao, China

[3]School of Finance and Business, Shanghai Normal University, Shanghai, China

*Corresponding author: s207011@hsu.edu.hk

†These authors contributed equally

Abstract. The basic task of feature learning is to use algorithms to allow machines to automatically learn useful data and its features during the model building process. The quality of the learned features will greatly affect the results of downstream tasks. Early feature learning methods relied on handcrafted features. Thanks to the development of deep learning, feature learning methods based on convolutional neural networks have greatly improved the quality of features. However, with the increasing scale of training data and the increasing complexity of modeling tasks, deep neural network Transformer based on self-attention mechanism and parallel data processing has gradually become a new research hotspot. Transformer can adaptively and selectively select contextual information and key semantic information in a scene by covering attention networks and fully connected layers and has become an important research area for computer vision and natural language processing tasks. This paper reviews the basic principles and development of Transformer, focuses on its application in CV and NLP, and analyzes effective Transformer-based models. Finally, the challenges faced, and future trends of transformer models are summarized.

Keywords: Transformer, computer vision, natural language processing, self-attention mechanism, neural network.

1. Introduction

Representation learning has always been a popular research topic in the field of artificial intelligence and pattern recognition. The basic task of representation learning is to use algorithms to allow machines to automatically learn useful data and its features in the process of model building. Since the quality of learned features will greatly affect the results of downstream tasks (such as image recognition, object detection, text classification, relation extraction, etc.), designing high-quality feature extraction networks has always attracted the interest of many researchers.

Early feature representations mainly relied on handcrafted features to compute and select low-level features such as textures, corners, and colors as input to the model. Although these features have good physical explanations, their ability to describe and express complex scenes is insufficient and cannot meet the practical application requirements of most tasks. With the continuous development, convolutional neural networks and their various optimizations of deep learning technology have gradually become the main feature learning methods at present. The classic convolutional neural network structure generally includes convolutional layers, pooling layers, activation layers and fully connected layers. By combining low-level features to form more abstract high-level representation attribute categories or features, it can effectively learn semantic information in different scenarios. Convolutional neural networks greatly improve the quality of feature representation and performance on basic tasks. However, as the goals of computer data processing increased, researchers began to focus on processing higher-level data. Transformer is able to adaptively and selectively select contextual information and key semantic information in the scene by covering attention network and fully connected layer, which has further refreshed the results of various tasks in recent years.

Transformer models were originally designed to solve natural language processing (NLP) tasks. The mainstream model in the NLP field, Recurrent Neural Networks (RNNs), cannot be trained in parallel. Also, word coding in one position can only work on word coding in the latter several positions, which makes it is difficult to affect the overall situation especially for the RNN-based text translation task. To solve the above problem, Transformer utilizes the self-attention mechanism Fast parallelism is achieved, and can be increased to very deep depths, significantly improving model accuracy. Since then, the Transformer model has been gradually extended to various tasks such as relation extraction, causal inference, and text classification, and has become the dominant feature learning method in the field of NLP. In the field of computer vision, a wide variety of tasks such as medical image analysis, face recognition, and motion capture require efficient and accurate feature learning methods. Convolutional Neural Networks (CNN) have been the mainstream model for image processing since 2012. The CNN model generally obtains feature maps through convolution operations. The input layer reads the neurons of each layer layer by layer, and obtains different types of features from the cells of different feature maps. Its advantage is that the features, scale and destruction are invariant, which is more suitable for object detection. However, the layer-by-layer correlation of neurons makes CNN ignore the overall characteristics of the data to a certain extent. Recently, with the explosive increase of the amount of modeling data at any time, Transformer has entered the field of CV and has become a new research hotspot due to its excellent performance in processing these massive data.

Focusing on the application of transformer in computer vision and natural language processing, the paper first introduces the essential principles of the Transformer models and comb through those main features of Transformer models such as the self-attentive mechanism and the encoder-decoder structure. After that, the research progress of Transformer in the fields of NLP and CV are discussed in detail, respectively, including the representative Transformer models and their optimizations. In Section 4, the quantitative performance comparisons of different Transformer models are provided. Finally, this paper also summarizes the existing issues in the transformer and gives a look out for its future development.

2. Principle of transformer

The Transformer architecture introduces a self-attention mechanism that avoids recursion in neural networks and relies entirely on the self-attention mechanism to draw global dependencies between inputs and outputs. Through the self-attention mechanism, attention between each word and all words is calculated, so that each word has global semantic information, and long-distance dependency can be captured.

2.1. Attention Mechanism

Attention Mechanism imitates the biological observation process and can improve the observation accuracy of certain regions [1]. Through attention mechanism, important features of sparse data can be extracted quickly. Therefore, it is widely applied in machine translation [2], speech recognition [3], image processing [4] and other fields. Attention mechanism has received extensive attention in the field of neural networks. Its rapid development can be mainly attributed to the following three reasons: Firstly, it is an advanced algorithm that can solve multi-task problems; secondly, it is extensively used to improve the interpretability of neural network; thirdly, it promises to overcome some challenges in RNN, such as performance decline with the increase of input length and low computational efficiency caused by unreasonable input order. The self-attention mechanism [5] is improved based on the attention mechanism. It not only reduces the network's dependence on external information, but also better captures the internal correlation of features.

The attention mechanism is essentially an addressing process, as shown in Figure 1. Given a task-dependent query vector Q. Calculate the Attention distribution of the Key and attach it to the Value to calculate the Attention Value. This process reflects the essence of the attention mechanism, which

alleviates the complexity of the neural network model: It only needs to select part of the task-related information from the input set X to input into the neural network, rather than input all N inputs into the neural network for calculation.

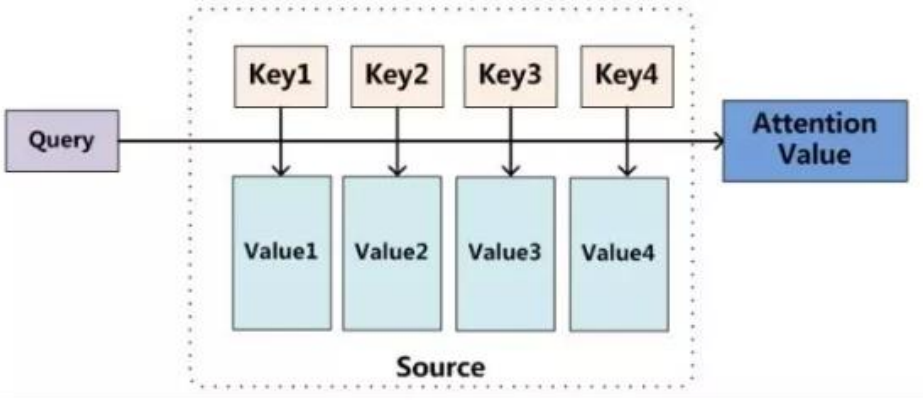

Figure 1. Attention mechanism

The process of the attention mechanism can be divided into the following three steps :

(1) Information input.

$$X = [x1, \ldots xN] \tag{1}$$

Equation (1) represents N input information;

(2) Calculation of attention distribution. Make $Key = Value = X$ and then the attention distribution can be given as equation (2).

$$\alpha_i = softmax\big(s(key_i, q)\big) = softmax\big(s(X_i, q)\big) \tag{2}$$

α_i is called the attention distribution (probability distribution) and it is the scoring mechanism for attention. There are several scoring mechanisms such as additive model, dot product model, scale dot product model and bilinear model, which can be calculated by equation (3)-(6), respectively.

$$s(x_i, q) = v^T \tanh\,(W_{x_i} + Uq) \tag{3}$$

$$s(x_i, q) = x_i^T q \tag{4}$$

$$s(x_i, q) = \frac{x_i^T q}{\sqrt{d}} \tag{5}$$

$$s(x_i, q) = x_i^T W q \tag{6}$$

(3) Weighted average of information. Attention distribution α_i is the attention degree of the ith information in context query Q. Encode input information X with the "soft" information selection mechanism as:

$$attention(q, X) = \sum_{i=1}^{N} \alpha_i X_i \tag{7}$$

2.2. Multiple attention

The essence of the multi-attentional mechanism is to split the query, key and value parameters for multiple times under the condition that the total number of parameters remains unchanged. Each group of split parameters is mapped to different subspaces of high-dimensional space to calculate the attention weight.

To focus on different parts of the input, the attention information in all subspaces is finally merged after several parallel calculations, whose formula is as follows:

$$\begin{cases} MultiHead(Q, K, V) = concat(head_1, head_2, \cdots, head_k)W^0 \\ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \end{cases} \tag{8}$$

Here, W^0, W_i^Q, W_i^K, W_i^V are the parameter matrices of linear transformation. Since attention is distributed differently in different subspaces, multi-headed attention is actually seeking associations between input data from different angles, which can encode multiple relationships and nuances. Transformer adapts an encoder-decoder architecture consisting of six layers of encoders and decoders. Each layer of the encoder contains two sub-layers: multi-head attention layer and feed forward layer. The decoder has three sub-layers, namely, masked multi-head attention, multi-head attention and feed forward. Each sub-layer is followed by a residual connection and a layer normalization, as shown in Figure 2. In one more cover long attention layers in the decoder, due to the training data length is different, in front of the encoder and decoder normally with a maximum length of data as the cell for training, and will only be affected by the data for the current before, do not need follow-up data for reference, so this layer will cover the fall after the current position of the data. Since Transformer's calculations abandon the recursion and convolution of loop structures and cannot simulate the positional information of words in text, artificial additions are required through positional encoding. Number the position of each word in the sentence. Each number corresponds to a vector, and by combining the position vector with the word vector, each word can obtain certain position information. Position coding is introduced through sine and cosine functions, and the formula is as follows:

$$\begin{cases} PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d}\right) \\ PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d}\right) \end{cases} \tag{9}$$

Pos is the position of the word in the sentence, d is the dimension of position encoding, 2i indicates that it is an even dimension, and 2i+1 indicates that it is an odd dimension. Location encoding records the sequential correlation between sequence data. Compared with RNN sequential input, Transformer method can directly input data in parallel and store the location relationship between data, which greatly improves the calculation speed and reduces the storage space[6]. In addition, as the network deepens, the distribution of data changes constantly. To ensure the stability of data feature distribution, layer normalization is introduced to reduce information loss and make training of deep neural network smoother.

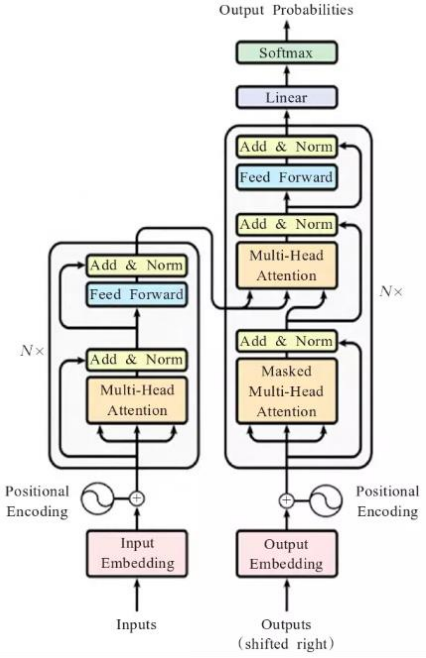

Figure 2. The structure of the transformer

3. Transformer in NLP

3.1. The Far-reaching significance of Pre-trained language models

In the field of Natural Language Processing, the view of Yoshua Bengio is widely accepted [7]that a good representation should capture the implicit linguistic rules and common sense knowledge hiding in text data. Two main types of embedding are Commonly used in neural network-based model structures, consisting of non-contextual and contextual, the former is static and unable to process multiple meanings of words well. To solve this problem, the latter type of dynamic embedding was introduced. People needs to use more data to train the models as well as avoiding overfitting since the parameters of models proliferate as deep learning evolves. In that case, pre-trained language models provide a better model initialization by using data from large text corpora to learn generic text representations in NLP and can help with subsequent tasks.

Before the introduction of the Transformer model, the NLP field used to take pre-trained word embeddings from unlabelled text, such as Word2Vec and Glove, which yielded context-independent embeddings, where the word vector could not change dynamically with the context [8]. Later on, there are LSTMs used as encoders in machine translation tasks to train Seq2Seq models, and ELMo models based on LSTMs trained on large amounts of unlabelled data, which yield linguistic representations that are significantly more effective for subsequent tasks, however, the parameters of these models are mostly fixed and the main model parameters need to be retrained [9].

Transformer introduces more complex structures and more layers of pre-trained language models to help achieve more effective data training, and is widely used in various areas of NLP. Its self-attention mechanism helps the model use fully connected graphs to model the correlation between two words in a sentence and learn the graph structure from it.

3.2. Transformer's development in NLP

(1) GTP Style. The researchers took out the Decoder part of the original Transformer, removed the Multi-Head Attention part, combined with linear transformation and classification to form the model structure of the original GPT, and designed an efficient training strategy. It uses a classical pre-training combined with fine-tuning approach, first training a generative language model based on a primary corpus, and then continuing to train the model using annotated data. Later, the researchers wanted to reduce the use of annotated data by increasing the amount of data which was used to train the language model, thus eliminating the needs for separate annotation of data for individual task training, in which case GPT-2 was developed and adapted to solve the problem of uncertain answers under unsupervised learning. GPT-2 is still inferior to humans in many aspects of language processing, and GPT-3 complements it by being able to do not only the above-mentioned things better and more varied, but also other tasks that have no beginning but only hints, such as writing blogs that humans cannot confirm are written by AI [10] , interacting with people, doing more complex story continuation, generating images, and even writing runnable code based on the user's instructions and even write runnable code. But GPT-3 does not actually use the fine-tuning, as can be seen in the paper on GPT-3: the authors argue that the mainstream of AI models for language processing today is pre-training plus fine-tuning, but that fine-tuning still requires large amounts of supervised data and is still a far cry from the simple examples or instructions that humans need to complete a language task. In contrast, GPT-3's modeling is much closer to that of humans, and can perform a variety of NLP tasks without any fine-tuning at runtime.

(2) BERT Style. BERT, known as Bidirectional Encoder Representation from Transformer, uses the Encoder of Transformer to pre-train a deep bidirectional representation of a large amount of unlabelled text in by an unsupervised learning manner. In contrast to GPT-2, its training goal is not to predict the next word by clause, but to understand it from context, improving the accuracy of the original GPT for word prediction. In practice, it only requires to fine-tune the pre-trained model with an additional output layer to handle many different NLP tasks without the need to redesign different neural network architectures. On top of BERT, researchers from Facebook have proposed RoBERTa,

which changes the approach to pre-training [11]. On the one hand, it changes from static masking to dynamic masking; while Bert only performs a random mask a fixed number of times for each sample when preparing the training data, RoBERTa dynamically generates a mask each time the model is fed with input, so it is constantly changing. On the other hand, while originally BERT used the NSP to pre-train and thus capture the relationships between sentences, switching to the FULL-SENTENCES method improves its performance on the task of inferring sentence relationships.

Drawing on the phenomenon [12] that a larger Batch size along with a larger learning rate in machine translation can significantly improve model optimisation rates and performance, the researchers used batch_size=8k for training in subsequent experiments. Similar to XLNet, RoBERTa has also been optimised at the data level, using a larger training dataset than Bert.

ELECTRA, also an ingenious improvement on Bert, where the researchers propose a new training method that replaces Masked language model (MLM) pre-training with Replaced token detection (RTD), which uses MLM's Generator to make changes to the input sentences, and discriminator determines whether the current token has been replaced or not. This training method reduces a large amount of computation. According to the researchers' experimental results [13], close performance can be achieved with only 1/4 of the computational effort of RoBERTa, and the ELECTRA model outperforms Bert to the same extent of size.

3.3. Application example

Sentiment classification is one of the subtasks of sentiment analysis and at the same time the most central task of sentiment analysis. Methods for text sentiment classification are currently rule-based, machine-learning-based and deep-learning-based. Deep learning has only been applied in the field of sentiment analysis in recent years, with the most notable result being word vectors. Neural networks (e.g. CNN, RNN, etc.) have also achieved good results in sentiment classification. As attention mechanisms began to be applied in the field of NLP, the combination of neural networks and attention mechanisms became the mainstream approach. The Transformer was introduced to the field of sentiment analysis to solve these problems.

For example, Ke Chen et al. proposed a text sentiment analysis method based on sentiment lexicon and Transformer, which further improves the accuracy of sentiment classification by adding feature information of sentiment words and using Transformer to extract hidden semantic information. Multiple attention mechanisms are used in the Transformer structure, and make it possible for models to learn relevant semantic information in different presentation subspaces. Moreover, the Transformer does not depend on the previous moment of computation, which greatly improves the parallel computing capability of the model. The model achieves good sentiment classification results on different data.

4. Transformer in Computer Vision

4.1. Model comparison of CNN, pure Transformer and hybrid Transformer

There are two main model architectures that employ Transformers for tasks in the computer vision field. One is a pure Transformer structure (ViT), and the other is a hybrid structure combining CNNs and Transformers (DETR). Transformer appeared later than CNN, so it has a simpler and more efficient structure than CNN. However, the relationship between the Transformer of the hybrid structure and the pure Transformer is more like complementation. Scholars believe that the Transformer of the hybrid structure is not enough to deal with the huge amount of data in the future, so there is the emergence of the pure Transformer. Because there is no fine-tuning in the pure Transformer model, although this will make it less accurate in the absence of data, it will be less restricted than the hybrid structure Transformer. Overall, both the pure Transformer and the hybrid structure Transformer are better than the CNN model, but because the CNN is better than the Transformer in local feature capture, the hybrid structure is more efficient than the pure CNN at a smaller model size.

4.2. ViT

Vision Transformer (ViT[14]), also known as Vision Transformer, is the first to replace CNN and use a pure Transformer structure. It is different from the traditional transformer. The input of the Transformer for natural language processing is one-dimensional data, such as words and sentences. The image problem to be solved by ViT is two-dimensional data that is difficult to process directly and needs to be converted into one-dimensional data. sequence of numbers. The solution is shown in Figure 3. The input two-dimensional image is cut into a controllable number of small blocks with a resolution of 16 × 16 and arranged, and then converted into a vector, which can be directly input into the Transformer. Convolutional networks can also accomplish such tasks, but because ViT hardly changes any structure of the original Transformer, ViT performs better than CNN when dealing with the same datasets such as image net [15].

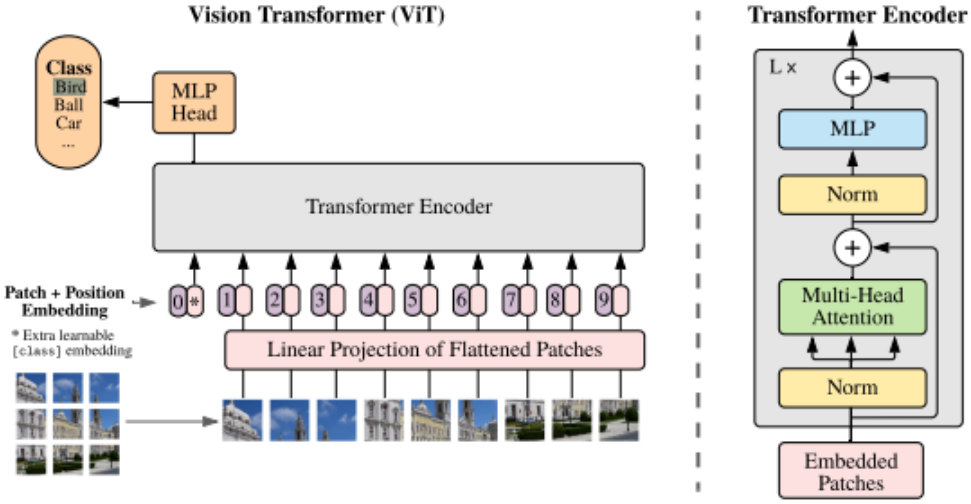

Figure 3. The Components of Vision Transformer [14]

For a two-dimensional image data to be processed, it can be assumed that its resolution is H(Height) x W(Weight) and the number of layers is C(Channel), then an image can be represented by H×W×C. If the cut out image blocks (Patch) are represented by P×P×C and arranged in order, a sequence X p ∈R N×(P×P×C) will be obtained after flattening. Assuming that the dimension used by all layers is fixed to D, use a linear layer to map the image data to the D dimension, the matrix is E, the position encoding is Epos, and finally the matrix input of the Transformer encoder can be obtained as:

$$Z_0 = [X_{class}; X_p^1 E; X_p^2 E; \ldots; X_P^N E] + E_{pos}, E \in R^{(P\times P\times C)\times D}, E_{pos} \in R^{(N+1)\times D} \quad (10)$$

X_{class} in patch embeddings Pre-join is done to complete learnable type tokens in classification tasks [11, 13].

4.3. Application of Transformer in CV field

In the field of computer vision, Transformer's main tasks include image classification, feature extraction and image cutting, etc. Now Transformer only occupies a dominant position in the NLP field, and most other fields are in the trial stage, because other fields need to deal with task objects Most of them are two-dimensional images, which are much more complex than the one-dimensional characters that need to be processed in the NLP field. However, because of the unique advantages of Transformer compared with other models, applications related to Transformer are increasing, such as facial expression recognition (FER), human pose estimation (PoseEstimation) and so on. The former converts the input facial images into visual words and sorts them, and then performs global and local information processing to filter junk information and extract effective information. Facial expression

recognition technology has been widely used in medical, gaming, security and other fields. The latter's human pose estimation is to locate the input human image and extract key information such as human body shape. This kind of technology has been applied to the fields of human-computer interaction and action analysis.

With the development of ViT, more and more ViT algorithms are used in autonomous driving. It is not difficult to find that the unique self - attention mechanism of ViT algorithm will bring great improvements to the algorithms in the field of autonomous driving. As a result, a series of more systematic and closely related autonomous driving algorithms are formed. At present, the data of 3D target detection in the field of unmanned driving mainly come from lidar data and image data. However, these two kinds of data are very different and cannot be combined well and then applied to the use of unmanned driving. For example, when detecting distant objects visible on some images, lidar will cause missed detection due to sparse points. A new CAT - Det framework for object detection is presented in the report of Zhang et al. [17] for the above problems. This framework mainly consists of several parts as shown in Figure 2.2 below. The main function of this framework is to integrate multi-modal data, so that the collection of various data forms a joint perception of multiple organs similar to humans, thereby improving the accuracy and comprehensiveness of the data.

Convolutional Neural Networks (CNNs) methods have been the mainstream in the field of medical image analysis over the past decade. When Transformer proved to be beneficial for image processing tasks, it became an alternative to CNNs. Scholars Christos [18] and others studied the possibility of ViT completely replacing CNN. In multiple comparative experiments, they found that, except for the limited amount of data, the performance of ViT and CNN on ImageNet pre-training is similar, and in other cases such as transfer learning, self-supervised learning and fine-tuning, the performance of ViT is better. it is good. And in addition to the above comparison, Transformer has some unique advantages. For example, CNN can only provide rough visualization images due to the existence of its pooling layer. The Transformer token can simulate the interaction between each region in the image with the help of self-attention mapping, so as to output an image with more details. Transformer can also generate attention maps of many medical images to ease the pathological analysis work of doctors, such as attention maps of cancer cell distribution maps, attention maps of protein contact maps, and it is difficult for CNN to focus on such granular images as Transformer.

5. Performance analysis

5.1. Common dataset

GLUE: To get the most out of NLU tasks, researchers from New York University, The University of Washington, and others have created a multi-tasking natural Language Understanding benchmark and analysis platform, GLUE (General Language Understanding Evaluation). GLUE contains nine NLU tasks in English. The nine tasks of GLUE involved natural language inference, text implication, emotion analysis, semantic similarity and other tasks.

SST-2: SST-2 (The Stanford Sentiment Treebank), a single-sentence sorting task that contains sentences from movie reviews and human annotations of their emotions. The task involved emotions for a given sentence, divided into two categories: positive and negative. It contains 67, 350 training sets, 873 development sets, and 1, 821 test sets.

MNLI-m: MNLI(The Multi-genre Natural Language Inference Corpus), a task of Natural Language Inference, is a collection of text implication annotation of sentence pairs through crowdsourcing. Given a premise statement and a hypothesis statement, the task is to predict whether the premise statement contains an entailment, a contradiction to the hypothesis, or neither. Premise statements are collected from dozens of different sources, including transcribed phonics, novels, and government reports.It contains training set 392, 702, development set dev-matched 9, 815, development set dev-mismatched 9, 832, test set test-matched 9, 796, test-mismatched9, 847. As MNLI is a text with many styles in different fields, it can be divided into two versions of data set:

matched and mismatched. Matched means that the data sources of training set and test set are consistent, while mismatched means that the data sources of training set and test set are inconsistent.

SQuAD: SQuAD is an extractive QA dataset proposed by Rajpurkar et al. The dataset contains 100,000 triples (questions, texts, answers) from 536 Wikipedia articles. For each article's questions, there are many annotators to mark the answer, and the answer appears in the original text.

BOOKCORPUS and English Wikipedia: The combination of BOOKCORPUS (Zhu et al., 2015) and English Wikipedia is the original data used for BERT's training

CC-NEWS: Facebook researchers collected data from the English section of the CommonCrawl News dataset, which contains 63 million English-language News articles from September 2016 to February 2019.

RACE: RACE is a massive reading comprehension dataset containing approximately 28,000 articles and nearly 100,000 questions. The dataset is from China's English Test, designed for middle and high school students.

5.2. Performance comparison of representative Transformer models in NLP

In comparison to the original Bert, the researchers explored which quantitative metrics had an impact on the pre-trained BERT model while keeping the model architecture unchanged; first maintaining the Bert-base configuration, progressively controlling for variables and comparing the effects in each case, using the datasets SQuAD, GLUE's MNLI-m, SST-2 and RACE to assess the effects. The results show that the RoBERTa model improves its effectiveness by 5%-20% compared to BERT on a wide range of datasets.

Experimental details analysis:

In contrast to the static mask method used in the original Bert, the researchers use an improved version of the static mask and dynamic masking methods respectively. The dynamic masking method generates dynamically each time the input was provided to the model, and the experimental results are as follows:

Table 1. Performance under different masking methods

Masking	SQuAD 2.0	MNLI-m	SST-2
reference	76.3	84.3	92.8
Reimplementation:			
static	78.3	84.3	92.5
dynamic	78.7	84.0	92.9

It can be seen that the modified version of static mask is comparable to the reference version, while dynamic mask is slightly better than static mask.

In the final RoBERTa model, the researchers eliminate the next sentence prediction method and the training data was obtained consecutively from one document. During the training process, the researcher use four methods to conduct the experiments, which were:

Using sentences or sentence segments plus NSP:

a.SENTENCE-PAIR+NSP

b.SEGMENT-PAIR+NSP

Input from the same document or different documents.

c.FULL-SENTENCES: (no NSP)

d.DOC-SENTENCES (no NSP);

and the overall effect of the experiment is a<b<c<d;

The results of the experiment are shown in the following table and can be summarised as follows:

(1) Segments will be better than sentences;

(2) No NSP task will be better than with NSP;

(3) No cross-document will be better than cross-document.

Table 2. Results of 4 different methods, assessed by SQuAD and MNLI,SST,RACE

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Reimplementation (with NSP loss):				
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
Reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6

The use of large batch training methods in neural machine translation research showed that it can improve final task performance, and researchers have used this approach also for improving Bert. The original Bert used a batch size of 256 sequences and 1 million steps, while in RoBerta, 125K steps for batch size=2K sequences or 31K steps for batch size=8K were trained by gradient accumulation; the experimental results are as follows.

Table 3. Change batch_size and steps, the learning rate,perplexity,and MNLI,SST vary

bsz	steps	lr	ppl	MNLI-m	SST-2
256	1M	1e-4	3.99	84.7	92.7
2K	125K	7e-4	3.68	85.2	92.9
8K	31K	1e-3	3.77	84.6	92.8

Comparing the "perplexity" and final task performance of BERT-base at increasing batch size, it can be observed that large batches training resulted in a significant reduction in perplexity of the masked language modeling target and improved the accuracy of the final task.

5.3. Performance comparison of representative Transformer models in CV

Table 4. Part of representative Transformer models in CV

Category	Model	Dataset	Params/106	Accuracy	AP	mIoU
Image Classification	iGPT	ImageNet	1 362	69	-	-
		CIFAR-10	1 362	96.3	-	-
		CIFAR-100	1 362	82.8	-	-
Object Detection	DETR	COCO	41	-	42	-
Image Segmentation	Segmenter	ADE20K	-	-	-	50.77
		Pascal	-	-	-	80.7
		Cityscapes	-	-	-	55.6
Video Processing	TimeSformer	K400	121.4	78	-	-
		SSv2	121.4	59.5	-	-

There are many Transformer models in the field of computer vision. There are representative Transformer models for each task domain. Because of the huge number of models, only one Transformer model for each type of task is selected for analysis. Table 5.4 is overview of the transformer performance.

(1) IGPT in image classification tasks: IGPT is obtained by Chen [19] and others based on the existing GPT-2. This model uses unsupervised learning, and the author converts the input image into a long bar pixel sequence, and pre-train and fine-tune it. On CIFAR-10, this model can even achieve 99% accuracy with fine-tuning. Even compared with self-supervised standards, its accuracy reaches 69% of Top-1 on ImageNet.

(2) Segmenter in image segmentation tasks: Segmenter is a ViT-based Transformer model for semantic segmentation, which can also be used for image segmentation. Segmenter relies more on the output embedding corresponding to the image patch, it needs to use the encoder to encode the

embedded sequence to obtain the class label, and then perform decoding and pixel classification to obtain the final pixel segmentation map.

(3) DETR in the target detection task: DETR uses a hybrid structure of CNN+Transformer. It first uses CNN to extract the features of the image, and then uses Transformer for preprocessing. Finally, the decoder transmits the output data to the feedforward network. Decode into box coordinates and class labels for object detection tasks. Almost all object detection algorithms after DETR are improved on its basis.

(4) TimeSformer in video processing tasks: In video processing tasks, the best 3D CNN models that have been done before can only use videos of a few seconds in length to avoid excessive computation. TimeSformer is a Transformer structure without convolution. The authors Bertasius[20] et al found divided space-time attention in their research. It can make the image blocks in the video only compared with the image blocks in other spatial positions at the same time, which not only greatly reduces the amount of computation, but also improves the training effect.

6. Discussion

6.1. Existing problems in NLP

(1) Lack of fine-grained semantic representation. The lack of fine-grained representation can have a significant impact on the real task, especially in the face of targeted spoofing, and Bert's ability to encode context does not mask its lack of fine-grained semantic representation, which can have a significant impact on the real task.

(2) Theoretical shortcomings of relative positional coding. Transformer's position coding can be divided into two categories: "absolute position coding" and "relative position coding", with relative position coding performing relatively better in experiments in NLP and CV. In order to determine whether a model has sufficient ability to recognize position, a probe experiment was conceived, which is simply summarized as an experiment in which n zeros are input and position numbers 1 to n are output in an orderly manner, and if the model is capable of completing this experiment, it has the ability to recognize position itself, independent of external input. And since most relative position encodings only modify the Attention matrix before Softmax, the Attention matrix with relative position information is just a probability matrix and the model always outputs the same result for each position.

6.2. Existing problems in ViT

(1) Huge data requirements: The inductive bias capability of Self-Attention is weaker than that of CNN. Self -Attention does not make any assumptions about some unencountered data in advance in order to obtain more flexible hypotheses from a large amount of data. But CNN has the assumption of space invariance, it can process the entire feature map with one weight, so CNN can do better than ViT in the face of small amount of data. The solution is to add a CNN network to help ViT learn, which is why sometimes the Transformer with hybrid structure is better than the pure Transformer structure (ViT).

(2) Huge computing demand: The larger the token value, the more complex the data that vit needs to calculate. Therefore, when the input data is some large feature maps, the amount of matrix operations will be too large. In addition, the number of tokens and hidden size remain unchanged during the original Transformer calculation. The solution to this is to use a pyramid structure, so that the number of tokens in the upper layer is less than that in the lower layer. In the process of generating Q, K, and V, the feature maps or tokens of K and V are pooled, thereby reducing the amount of calculation.

(3) The number of stacking layers is limited: as the number of stacking layers increases, the similarity between different Blocks and Tokens will increase, resulting in the problem of over-smoothing. The solution is mainly to increase hidden size, but this will add a lot of parameters. Other

solutions include increasing the similarity penalty loss term, increasing feature diversity and attention map diversity, etc.

7. Conclusion

Transformer can adaptively and selectively select contextual information and key semantic information in a scene by covering attention networks and fully connected layers, and has become an important research area for computer vision and natural language processing tasks. This paper first reviews the basic principles and development of Transformer; secondly, representative methods of Transformer in computer vision and natural language processing tasks are introduced in detail, including the benefits and drawbacks of these methods. Finally, on the basis of quantitatively comparing the performance of different Transformers, this paper concludes the limitations and future research trends of Transformer models in these two fields.

References

[1] LIU J H. Active and semi-supervised learning based on ELM for multi-class image classification[D].Nanjing: South-east University, 2016.
[2] WANG H, SHI J C, ZHANG Z W. Text semantic relation extraction of LSTM based on attention mechanism[J]. Application Research of Computers, 2018, 35(5): 1417-1420.
[3] TANG H T, XUE J B, HAN J Q. A method of multi-scale forward attention model for speech recognition[J]. Acta Electronica Sinica, 2020, 48(7): 1255-1260.
[4] WANG W, SHEN J, YU Y, et al. Stereoscopic thumbnail creation via efficient stereo saliency detection[J].IEEE Transactions on Visualization and Computer Graphics, 2016, 23(8): 2014-2027.
[5] LIN Z, FENG M, SANTOS C N, et al.A structured self-attentive sentence embedding[C]//Proceedings of the Inter-national Conference on Learning Representations, Toulon, France, 2017.
[6] REN H, WANG X G.Review of attention mechanism[J]. Journal of Computer Applications, 2021, 41(S1): 1-6.
[7] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.
[8] Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: adapt language models to domains and tasks[J]. arXiv preprint arXiv:2004.10964, 2020.
[9] Sarzynska-Wawer J, Wawer A, Pawlak A, et al. Detecting formal thought disorder by deep contextualized word representations[J]. Psychiatry Research, 2021, 304: 114135.
[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[11] Keskar N S, Mudigere D, Nocedal J, et al. On large-batch training for deep learning: Generalization gap and sharp minima[J]. arXiv preprint arXiv:1609.04836, 2016.
[12] Liu Z, Lin W, Shi Y, et al. A Robustly Optimized BERT Pre-training Approach with Post-training[C]//China National Conference on Chinese Computational Linguistics. Springer, Cham, 2021: 471-484.
[13] Clark K, Luong M T, Le Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators[J]. arXiv preprint arXiv:2003.10555, 2020.
[14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
[15] Chen X, Hsieh C J, Gong B. When vision transformers outperform ResNets without pre-training or strong data augmentations[J]. arXiv preprint arXiv:2106.01548, 2021.
[16] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. IEEE transactions on pattern analysis and machine intelligence, 2022.

[17] Zhang Y, Chen J, Huang D. CAT-Det: Contrastively Augmented Transformer for Multi-modal 3D Object Detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 908-917.
[18] Matsoukas C, Haslum J F, Söderberg M, et al. Is it time to replace cnns with transformers for medical images? [J]. arXiv preprint arXiv:2108.09038, 2021.
[19] CHEN M, RADFORD A, CHILD R, et al.Generative pretraining from pixels[C]//International Conference on Machine Learning, 2020: 1691-1703.
[20] BERTASIUS G, WANG H, TORRESANI L.Is space-time attention all you need for video understanding?[J].arXiv: 2102.05095, 2021.