The current status and prospects of transformer in multimodality

Yingjie Zhang

Detroit Green Technology Institute, Hubei University of Technology, Wuhan, Hubei, China

17771443438@163.com

Abstract. At present, the attention mechanism represented by transformer has greatly promoted the development of natural language processing (NLP) and image processing (CV). However, in the multimodal field, the application of attention mechanism still mainly focuses on extracting the features of different types of data, and then fusing these features (such as text and image). With the increasing scale of the model and the instability of the Internet data, feature fusion has been difficult to solve the growing variety of multimodal problems for us, and the multimodal field has always lacked a model that can uniformly handle all types of data. In this paper, we first take the CV and NLP fields as examples to review various derived models of transformer. Then, based on the mechanism of word embedding and image embedding, we discuss how embedding with different granularity is handled uniformly under the attention mechanism in multimodal scenes. Further, we reveal that this mechanism will not only be limited to CV and NLP, but the real unified model will be able to handle tasks across data types through pre-training and fine tuning. Finally, on the specific implementation of the unified model, this paper lists several cases, and analyzes the valuable research directions in related fields.

Keywords: multimodal transformer, multi-headed attention, pre-training.

1. Introduction

Multi-modal refers to the research direction of machine learning with speech, text, vision and other field materials. In the information age, diversified data is expected to become a new direction for the development of machine learning. Current machine learning materials mainly focus on vision (including single or continuous images), speech, and text. At present, the research based on discontinuous images is mainly carried out through Convolutional Neural Network (CNN) network. Based on the sliding invariant and parameter sharing characteristics of convolution kernel, it is able to complete the tasks such as image classification, object detection or tracking. Speech-based machine learning mainly relies on relevant technical means to transform speech into pure text and then conduct multi-modal fusion. And video is essentially a combination of continuous images and voice. Therefore, in the multimodal field, the focus is on finding models that can span CV and NLP.

In 2017, Vaswani et al. revealed the application of multi-head parallel attention learning mechanism Transformer in the AI field [1], and Transformer quickly became one of the most widely discussed models in the field of machine learning. Transformer innovatively used the Attention mechanism to replace the Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) coding, and has achieved great results in the field of NLP. Later, Transformer researchers also turned their eyes to the CV field. For example, Liu et al. [2] proposed TransCnn based on attention mechanism and CNN, and Dosovitskiy et al [3]. proposed VIT (Transformer in Version) specifically applied in the vision field. Transformer This breaks the boundary between NLP and CV to some extent. However, from the multimodal perspective, Transformer is still insufficient for the fusion of CV and NLP. Although the attention mechanism of Transformer can theoretically span various modal information, the potential of Transformer in multimodal information has not been fully explored. On the one hand, by reviewing the working mechanism of attention mechanism, encoder and decoder, the unity of Transformer for various types of data in multimodal will be further revealed; on the other hand, by comparing the typical learning network in a single mode (such as CNN in CV) and Transformer, it can also get whether the differences of various data in multimodal mode and how to solve these differences is a very valuable research direction in the future.

2. Transformer mechanism

2.1. Word embedding and location encoding

2.1.1. Word embedding. Word embedding is a kind of processing for the input in the Transformer model, which has a special meaning for the model. Word embedding is an extension of word encoding. Word coding is a necessary means for NLP models to process the input sentences. Simple One-Hot encoding gives each word in the input sentence a unique identity to identify the position of the current words in the vocabulary. As a result, there is a unique and unmanned identification between words. According to the One-Hot encoding rule, the length of the word vector is equal to the size of the vocabulary, and each word vector consists of one and the remaining zero. The disadvantage of this is obvious, as the word vector length expands as the vocabulary expands and contains a lot of redundant information. Thus, the word embedding came into being [4]. The practice of word embedding is to create a weight matrix, and the number of columns of the weight matrix is written as d_model. Multiplying the word vector with the weight matrix yields the word embedding vector, with the dimension of the word embedding vector equal to d_model. The parameters of the weight matrix will be learned by backpropagation. In Transformer, the input to Encoder consists of both word embeddings and place coding, and next it will discuss place coding [5].

2.1.2. Positional encoding. Position encoding is where Transformer differs from other classical models. Since Transformer uses parallel rather than time-order based serial processing, positions should be encoded to represent the information of words in the sentence. There are two ways of location coding, using machine learning or defining a function. The classical Transformer uses sine and cosine coding, that is, the predefined function for position coding, and the calculation formula is shown in Equation (1) and (2).

$$PE(pos, 2i) = Sin\left(\frac{pos}{10000^{2i/d\,mod\,el}}\right) \tag{1}$$

$$PE(pos, 2i{+}1) = Cos\left(\frac{pos}{10000^{2i/d\,mod\,el}}\right) \tag{2}$$

It is worth noting that many subsequent improved versions have abandoned this approach and adopted machine learning for location coding. Location coding allows Transformer to process all information in parallel, rather than having to wait for the next node like RNN, which greatly improves the computational efficiency of the model and makes massively parallel computing possible [6].

2.2. Attention mechanism

2.2.1. Self-attention. In the structure of Transformer, the self-attention mechanism can be read as a special multi-head attention mechanism. Vaswani et al [1]. proposed three attention elements, Key (K), Value (V), and Query (Q), whose parameter matrix can be recorded as Wk, Wv, Wq. Equation (3) describes the way that attention is calculated.

$$Attention\left(Q,K,V()\,max\left(\frac{QK^T}{\sqrt{d_k}}\right)\right) \tag{3}$$

Each word vector in Transformer has a Query matrix corresponding to the Key matrix of the other word vectors, and in the case of the self-attention mechanism, the Query, Value and Key corresponding to each word vector are invariant. The structure of the self-attention mechanism is shown in Figure 1.

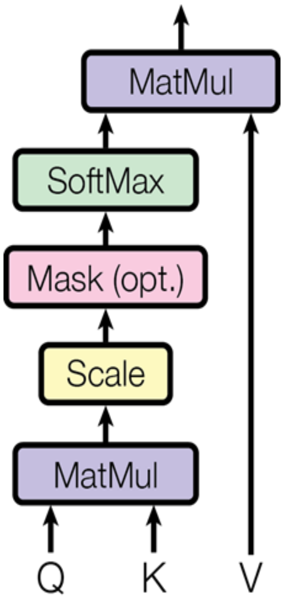

Figure 1. Self attention architecture diagram [1].

2.2.2. Multi-head attention. Transformer differs from other models in its working principle. Compared to the method entirely based on feature vector calculation, this calculation method creates different weight matrices when calculating the attention of each node, allowing the model to obtain more information from the context [7-8]. Figure 2 shows that when the long head attention mechanism is different from the self attention mechanism, Q, K, and V are split to obtain different attention information and ultimately assembled. The splicing process is given by equation (4) as [1].

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head_1,.....,headh_h)W^o \\ where\quad head_i &= Attention(QW_i{}^Q,KW_i{}^K,VW_i{}^V) \end{aligned} \tag{4}$$

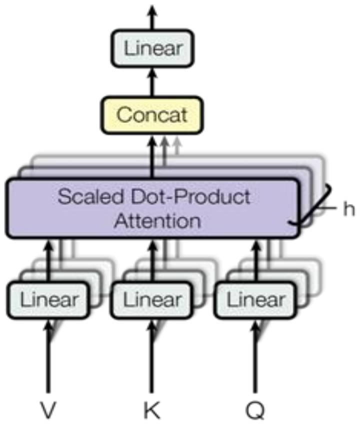

Figure 2. Multi head attention architecture diagram [1].

2.3. ResNet module

He et al. [4] first proposed resnet in 2014, the main response is in the big model network layer increase and gradient descent problem, in order to avoid this problem, what introduced the concept of residual, in each node delay will be residual calculation, if the subsequent network tends to make the residual larger, the model will allow to skip the node directly for the subsequent calculation [9].

The idea of residue is applied in Transformer, and an Add & Norm model is introduced into the model, which uses residual links to avoid gradient descent [10].

2.4. Encoder with the decoder block

The Transformer proposed by Vaswani et al. adopts the classical Encoder and Decoder structures, as shown in Figure Figure 3.

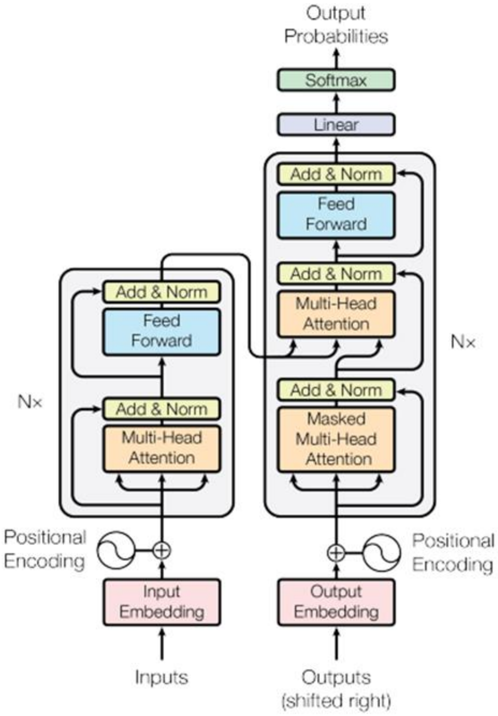

Figure 3. The overall architecture of the Transformer [1].

Encoder Is encoded into easy to learn vector forms, which successively constitute the input of the encoder through word embedding, position encoding, multi-head attention and residual link. A serial link is used between the encoder and the encoder, with the output of each encoder being the input of the next encoder. The form of the decoder is similar to the encoder, but the concept of mask is introduced, which is to simply allow the model to predict the following text through the previous decoded information, so as to avoid the model obtaining the predicted information from the later text, so as to ensure that the prediction is purely based on the previous text. The mask is a lower triangle matrix to block future information for the model.

3. Evolution and extension

3.1. Transformer in the CV field

Transformer It was originally used in translation in NLP, but gradually expanded to the field of image processing by researchers.

3.1.1. TransCnn. Liu et al. [5] proposed TransCnn in 2021. A module called the hierarchical MHSA (H-MHSA) is proposed, and it can be conveniently inserted into the CNN network. This CNN network, which combines the attention mechanism, is named TransCnn and has the advantage of both CNN and Transformer. TransCnn Is a pure visual model, which mainly introduces the concept of hierarchical convolution to reduce the training time and the number of parameters on a large scale. Compared with VIT, the operation efficiency is greatly improved without losing too much accuracy.

3.1.2. Vision in transformer. VIT was proposed by Alexey et al. [6] in 2020, mainly showing the application of Transformer model in the field of vision. The article innovatively proposed an improved Transformer version purely for image processing. The core process of VIT includes four main parts: make patches, patch embedding, location coding and MLP classification processing.

VIT adopts pure Transformer structure, uses the image as patch for graph embedding, and adds position coding to patch. VIT is closer to the Transformer applied to the NLP domain proposed in [1], they all use embeddings, encode positions, and compute the input from the high-dimensional space to the low-dimensional space.TransCnn Cnn network to obtain local information through Cnn network, in ImageNet is better than the traditional Cnn network.

3.1.3. Swin. Liu et al. [5] proposed another improved version of VIT called Hierarchical Vision Transformer using Shifted Windows in 2021. Instead of VIT, the multi-head parallel attention mechanism in the Swim network will not be conducted for the entire Global, but only within each window. To solve the problem that windows and windows can not transmit information effectively, Liu et al. proposed the window based self-attention(W-MSA) mechanism to solve it. Swing needs no global attention computation, which is better than VIT.

3.2. Transformer in the NLP field

Since 2017, the attention mechanism represented by Transformer has continuously guided the establishment and evolution of large-scale language models. In the large-scale language model, Transfer Text to Text Transformer (T5) continues the encoder-decoder structure, and Bidirectional Encoder Representations from Transformer (BERT) adopts the pure encoder structure.

3.2.1. BERT. Dosovitskiy et al. proposed the Bidirectional Encoder Representations from Transformers model in 2019 to find an improved version of Transformer with good pan-Chinese capability and easy pre-training. BERT is a pure encoder structure, and Segment Embedding [6] is added at the input to allow the model to learn richer corpus information based on the context semantics. The highlight of BERT is the pre-training with a pure encoder, and the fine-tuning of the output after determining a specific task.

3.2.2. GPT. Radford et al. [7] constructed the prototype of GPT based on the decoder structure in Transformerr. Similar to the decoder [1], GPT cannot understand the semantics based on the context like BERT, but predicts the below purely based on the above. The subsequent GPT work adopted a larger scale of parameters, and the number of GPT 3 parameters reached a staggering 175 billion.

3.2.3. T5. Transfer Text to Text Transformer aims to build a unified model across all NLP tasks. In the T5 model, the Encoder Decoder structure is still preserved, but when combined, it becomes a block. The original sine position encoding was replaced by learning to obtain it.

3.3. Transformer with multimode

The attention mechanism contained in Transformer is considered to have a powerful ability to process multi-state information [8], and the fusion of multimodal information can be completed by relying on the model itself. Before the VIT proposal, it was widely believed that CNN networks must be relied on for obtaining image information in the image processing field. After the VIT proposal, some CNN networks incorporating self-attention mechanism were still proposed [5]. Nevertheless, work like VIT and SWIN strongly proves to people that Transformer, as the first model proposed in the NLP field, can directly process image information without a large range of changes to the model. There are well-established examples of Transformer-based multimodal studies, such as Unit proposed by Hu et al. [9] aims to construct a unified input paradigm to allow Transformer to simultaneously accept images or text as input.

Through word embedding, it can build a text-based attention mechanism. The way of graph embedding is also discussed in this article 3.1.2. Theoretically, through various different types of embeddings, it is able to integrate different types of inputs into the attentional framework of Transformer. Compared with the traditional multimodal technology focusing on cross-domain data fusion, the Multimodal Transformer (MulT) proposed by Tsai et al. [10] allows us not to focus on how to extract features of different types of input, but to directly learn from unaligned text, speech or images through the attention mechanism.

Since the tedious process of feature fusion of data in different fields, it will be necessary to reveal the unity between computer vision, natural semantic processing or other tasks in the processing of the data learned by the model. Shi introduce the concept of global context to image tampering detection, and attribute such tasks to multimodality rather than simple visual tasks, thus achieving improvements in recognition accuracy and robustness. The attention mechanism has natural advantages in the work of multiple data fusion, and is expected to meet the demand for unified modeling of multiple data in the multimodal field.

4. Prospects and discussions

4.1. Problems exist at this stage

Although Transformer has been widely considered to have good cross-domain information processing power, the training difficulty, cost and number of the model should be a concern. In particular, in order to achieve unity on the model, sometimes the accuracy or efficiency of the computation will have to be reduced. Taking the visual task as an example, although VIT was regarded as a milestone to no longer restrict Transformer to the NLP task, after comparing the performance of the model on ImageNet, it found that VIT was weaker than the TransCnn in both the accuracy of the classification task.

4.2. Future outlook

For future Transformer on multimodality will focus on two aspects. On the one hand, the optimization of the number of participants, the difficulty of training and the training method. Multimode often means more information and computational amount. For the large number of parameters, SWIN inspires us that it can solve the problem of parameter number by optimizing the block method and proposing a new window interaction mode. Sometimes Transformer does not require multi-head attention processing of the global. Reasonable pre-training is also shown to help compress the training cost, requiring only refined model modification in specific scenarios; compression techniques for large models such as model pruning and distillation will help Transformer to achieve lightweight. The current model compression is mostly aimed at the general model. How to achieve better model compression for the attention mechanism is expected to become the future research direction.

On the other hand, how to prove that Transformer outperforms conventional models in specific fields, these works do not necessarily need to deny the superiority of the original model, just as the CNN does in the CV field. But it need to find out how to incorporate the original model into multimodal models such as MulT through embedding, such as exploring how to better integrate the advantages of graph embedding and convolution kernel.multimodal means more application scenarios and different training data. In the field of image [2], excellent research results based on Transformer have emerged in, text and speech. How to absorb the strengths of these models, but also take into account the universality of the unified multimodal model in the training stage and downstream tasks, is expected to become the most valuable research direction for Transformer in the future.

5. Conclusion

In general, since Transformer was proposed in 2017, countless studies have shown that Transformer has great potential in natural language processing, computer vision and speech processing, and its attention mechanism gives us the opportunity to form a unified learning model. By embedding, it is able to map different types of sources to an approximate multimodal space, and multi-head attention operations to the correlation of the vectors to each other. For data from different types of sources, where there is often a granularity size gap, and studying how to map them to a unified vector space is expected to find us a unified applicable multimodal model. In large-scale semantic models, the success of GPT makes us see the importance of pre-training, and studying how to fine pre-trained models will also become a promising research direction. This also has implications for the multimodal work. In order to solve the problem that the model with strong generalization ability is inferior to the typical network in this scenario (for example, TransCnn is still proved to be better than VIT in the image field), pre-training and fine tuning are a new path. After getting the pre-trained output through the unified multimodal model (such as Mult), by judging the different task types (simple image processing, word processing or image generation text), combined with different scenes in each type of task, will make the unified model also has a strong performance in the downstream task.

In the future, the search and establishment of a unified large model will become a research hotspot. Multimodal will be truly treated as a whole in subsequent work, rather than simply feature fusion. We can imagine that it will be possible to locate specific video frames through text input, generate videos from text, and automatically generate posters for music. In order to achieve these exciting goals, exploring the unified relationship between pre-training and fine tuning, modeling high-dimensional vector space that can accommodate more data with different granularity, or developing hardware that can support large-scale attention network training will be widely studied in the academic community.

References

[1] Vaswani A., Shazeer N M., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. (2017) Attention is All you Need. Energies, 89(1): 56-59.

[2] Liu Y., Sun G., Yu Q., Zhang L. (2021) Transformer in Convolutional Neural Networks. Energy, 67: 210-218.

[3] Dosovitskiy A., Beyer L., Kolesnikov A. (2020) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Energies, 20: 205-213.

[4] He K., Zhang X., Ren S., Sun J. (2015) Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 78: 770-778.

[5] Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B. (2021) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 78: 9992-10002.

[6] Devlin J., Chang M., Lee K., Toutanova K. (2019) BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding. Nature Machine Intelligence, 78(2): 773-778.

[7] Kwon S., Go B H., Lee J H. (2020) A text-based visual context modulation neural model for multimodal machine translation. Pattern Recognition Letters, 20: 201-216.

[8] Song T., Dai H., Wang S. (2022) TransCluster: A Cell-Type Identification Method for single-cell RNA-Seq data using deep learning based on transformer. Frontiers in genetics, 13: 1038919.

[9] Technology M., Xiangtan H., Sha H C. (2010) Novel Converter Transformer Application in Power System Harmonic Suppression of Mine. Electrical Automation, 20: 110-118.

[10] Shi Z., Chen H., Zhang D., Shen X. (2012) Pre-training-driven multimodal boundary perception vision Transformer. Journal of Software, 67(1): 34-38.