ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

A Survey of Transformers in Video Prediction

Weichen Ji*

State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China, 100024
**Corresponding author: 2022200810J4001@cuc.edu.cn*

Abstract: *Transformer is an encoder-decoder architecture based on the self-attention mechanism, which can effectively obtain global information and has shown great strength in the construction of long-distance dependencies. In recent years, Transformer has become the mainstream architecture of Natural Language Processing (NLP). Inspired by Transformer's success in NLP, researchers have gradually applied it to video processing tasks, one of which is video prediction. The essence of video prediction tasks is to generate future frames based on past ones. The model in this task needs to have strong sequence modeling capabilities, so the application and fusion of Transformer has become a major direction. This paper sorts out the application of Transformer in video prediction, lists typical models and analyzes improvement ideas, and finally summarizes and looks forward to its development.*

Keywords: *Transformer, Self-attention, Video prediction, Computer Vision, Survey*

1. Introduction

Transformer [1] is a deep neural network based on the self-attention mechanism. It was proposed by Google in 2017. It was first used in the field of NLP and has shown outstanding performance in tasks such as machine translation and sequence modeling. In 2018, Devlin et al. proposed Bert [2], a pre-training language model based on Transformer. At that time, Bert achieved state-of-the-art (SOTA) in 11 natural language processing tasks, which is amazing. Transformer has become the mainstream deep learning model in the field of NLP in a short period of time.

Inspired by the powerful functions of Transformer, researchers have begun to apply it to the field of Computer Vision (CV) [3], [4], [5]. The traditional Convolution Neural Networks (CNN) have a limited receptive field and are difficult to perceive global information. In this case,Transformer-based models with global perception capabilities gradually show their advantages. In 2018, Image Transformer [6] applied Transformer to the field of CV for the first time. Since then, the object detection model DETR [7] and image classification model ViT [8] have shown the great potential. So far, visual Transformer has been widely used in image classification, object detection, semantic segmentation and video tasks, etc. Among them, the processing of video data can be regarded as the problem of studying image frame sequences, and the global perception ability of Transformer is of great benefit to understanding the motion trajectory and dynamic appearance changes of video data. Given the sequential nature of video and the redundancy introduced by the temporal dimension, long-term modeling with image-based or NLP-based designs is not sufficient and we need a separate study on this branch. At present, the application of Transformer in video processing tasks is gradually increasing. However, there are few papers that classify and summarize this branch in detail. They often only talk about video processing in general when summarizing image processing applications, ignoring the challenges brought by the highly redundant spatial-temporal visual features of video to modeling. In addition, the small amount of existing research on Transformer in video processing tasks is often limited to video classification tasks, lacking a summary of video prediction.

This paper summarizes typical algorithms for applying Transformer in video prediction, and analyzesimprovement ideas. Section 2 introduces the basic principles and structure of Transformer; Section 3 sorts out the basic tasks of video prediction and the application of Transformer in this field, and also lists typical models and analyzes improvement ideas; finally, Section 4 summarizes the application of Transformer and looks forward to its future development.

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

2. Principle and Structure of Transformer

Overall, the Transformer consists of encoding and decoding components. The encoding component includes a 6-layer encoder, and the decoding component also includes a 6-layer decoder. Each encoder consists of a multi-head self-attention layer and a feed-forward neural network layer; and each decoder has 3 layers, the first and third layers are similar to the encoder, while the second layer is a cross-attention layer whose Key and Value inputs come from the corresponding outputs of the last encoder. This chapter will introduce the main modules inTransformer one by one in figure 1.

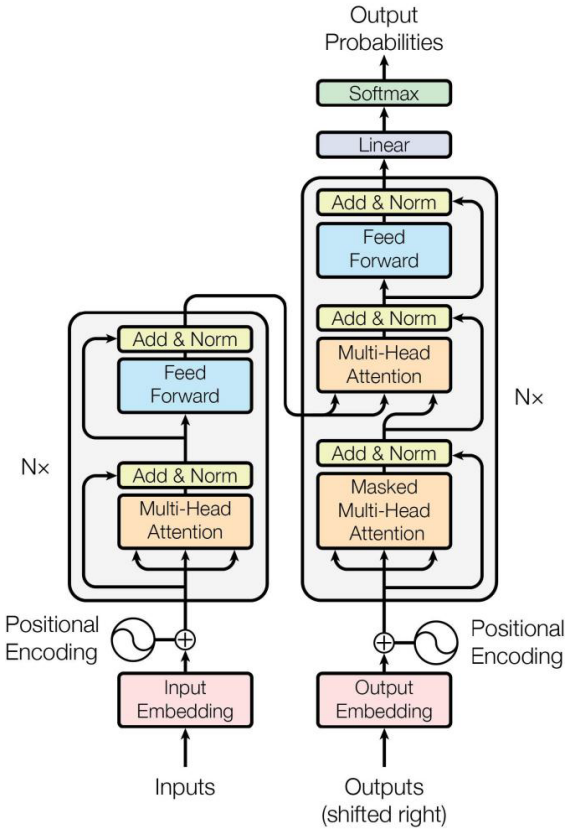

Figure 1: The overall structure of Transformer

2.1. Multi-head Self-attention Mechanism

The information that people receive at every moment is massive and complex, which far exceeds the processing capacity of the human brain. Therefore, when processing information, people will focus on the information that needs attention and filter other irrelevant external information. The emergence of the traditional attention mechanism stems from people's ability to process external information andthe self-attention mechanism is an improvement based on it. Q (Query), K (Key) and V (Value) of the self-attention mechanism come from the same set of elements, which is the most obvious difference between the two. Since the input of the neural network is many vectors of different sizes, and there is a certain relationship between different input vectors, the proposal of the self-attention mechanism can make the model fully notice the correlation between the inputs.

The self-attention layer defines three learnable weight matrices W^Q, W^K and W^V. The input sequence is projected onto these three weight matrices to obtain the query vector Q, the keyvector K, and the value vector V. The calculation process is shown as:

$$Q = XW^Q, K = XW^K, V = XW^V. \tag{1}$$

Then we calculate the dot product of Q and K and divide it by $\sqrt{d_k}$, where d_k is the dimension of K; finally, we use the softmax operation to normalize the calculation result into a probability distribution, and then multiply it by V to get the final result. The calculation formula of self-attention can be described as:

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

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

On the basis of the self-attention mechanism,Transformer proposes a multi-head self-attention mechanism, which divides the sequence X into *h* heads in the channel dimension, and each head uses different learnable weights to generate different {Q, K, V} groups.The calculation process is as follows:

$$\begin{aligned} &Q_{\text{i}}=XW_i^Q, K_i=XW_i^K, Vi=XW_i^V \\ &Z_i=\text{Attention}(Q_i,K_i,V_i),\ \ \text{i}=1,2,\cdots,h \\ &\text{MultiHead}(Q,K,V)=\text{Concat}(Z_1,Z_2,\cdots,Z_h)W^O \end{aligned}, \tag{3}$$

where W_i^Q, W_i^K and W_i^V are learnable weight matrices, *h* is the number of attention heads, Z_i indicates that the output of each attention head is consistent, and W^O is the output projection matrix. Using the multi-head self-attention mechanism we can form multiple subspaces, and can map the input to different spaces, so that the model can learncorrelation relationships of different angles of the input data, and improve the performance of the self-attention layer.

2.2. Position Encoding

Since the self-attention module in the Transformer cannot obtain the order information of the input data, and the relationship between the data is affected by the order information, position encoding is added to the model to embed an additional position vector for the input word vector. The obtained result is finally input to the encoder. The encoding method is as follows:

$$\begin{aligned} PE_{(pos,2i)}&=\sin(pos/10000^{2i/d_{\text{model}}}) \\ PE_{(pos,2i+1)}&=\cos(pos/10000^{2i/d_{\text{model}}}) \end{aligned} \tag{4}$$

Where *pos* represents the position of each element in the sequence, and d_{model} represents the dimension of the position code.

2.3. Feed-forward Neural Network and Layer Normalization

In Transformer, the encoder and decoder modules not only contain a self-attention layer, but also a feed-forward neural network consisting of two linear layers with a ReLU activation layer in between. The calculation process is shown in formula (5):

$$\text{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2. \tag{5}$$

One detail in the encoder and decoder structure is that each sublayer (attention layer and feed-forward neural network layer) has a residual connection, and then performs a Layer Normalization operation, which is calculated as follows:

$$\text{SubLayerOutput}=\text{LayerNorm}(x+\text{SubLayer}(x)) \tag{6}$$

3. Transformer Application in Video Prediction Tasks

3.1. Overview of Video Prediction Tasks

The ability to predict, anticipate and reason about future events is the essence of intelligence [9] and one of the main goals of decision-making systems. The video prediction task refers to the problem of generating a video from its past frames [10], [11], which we can define as follows:

Let $X_{\text{t}}\in\mathbf{R}^{w\times h\times c}$ be the t-th frame of the n-frame video sequence $X=(X_{t-n},\cdots,X_{t-1},X_t)$, where w, h, and c represent the width, height, and channel number of this frame respectively, and the goal is to predict future video frames $Y=(\hat{Y}_{t+1},\hat{Y}_{t+2},\cdots,\hat{Y}_{t+m})$ based on the input sequence.

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

At present, video prediction has been applied in many research fields, such as automobile automatic driving [12], precipitation forecast [13], anomaly detection [14], etc. Many types of models have been used to solve video prediction problems, such as Generative Adversarial Network (GAN) [15], [16], Recurrent Neural Network (RNN) [17], [18], [19], [20], [21] and Variational Auto Encoder (VAE) [22], [23], [24].At present, many models that achieve SOTA effects in the field of video prediction are based on ConvLSTM [13], and researchers are also working to improve performance by developing more and more complex ConvLSTM-based models. However, these networks have some inherent problems that RNNs have, such as slow training speed, error accumulation, and gradient disappearance problems.

As researchers have applied Transformer to the field of computer vision and achieved gratifying results, the application cases of Transformer in the field of video prediction have gradually emerged. Of course, in the process of application, researchers will inevitably find new problems.For example, the cost of applying Transformer to high-dimensional visual feature calculation is very high. Next, this article will sort out application cases of Transformer in video prediction tasks and analyze the improvement ideas.

3.2. Transformer Application and Improvement in Video Prediction

Through the retrieval and sorting of major journals and conferences, this paper summarizes the typical Transformer-based video prediction models that have emerged in recent years, and analyzes their improvement ideas as table 1.

Table 1: Typical Transformer-based models in video prediction.

Method	Design Focus	Publication Time
Video Transformer[25]	3D block-local self-attention	2019
Latent Video Transformer[31]	latent space video generation	2020
ConvTransformer[42]	multi-head convolutional self-attention layer	2020
TransAnomaly[48]	U-Net + ViViT	2021
VideoGPT[35]	VQ-VAE + GPT	2021
Nüwa[40]	multimodal pre-training	2022
MaskViT[30]	masked vision modeling; window attention	2022
VPTR[29]	spatial local self-attention; autoregressive and non-autoregressive model design	2022

3.2.1. Local self-attention

Weissenborn et al. proposed Video Transformer [25] in 2019, which is the first time Transformer has been applied to video generation and prediction tasks. In this model, in order to reduce the spatial complexity of the self-attention mechanism, the researchers used block-local self-attention to extend the image generation method proposed by Parmar et al. [26] and Chen et al. [27] to the calculation of 3D volume.The basic idea of this method is similar to the convolutional network, which extracts features by moving the convolution kernel. The self-attention mechanism of the original Transformer allows each element in a sequence containing N_p elements to be connected to other elements through the Attention matrix $A \in \mathbf{R}^{N_p \times N_p}$. Since the complexity grows quadratically with the image size, videos usually consist of hundreds of thousands or more pixels, so the order of magnitude of complexity becomes large for video. So researchers divide the video into smaller non-overlapping 3D patches to apply local self-attention. In addition, in order to further reduce the memory requirements of the model, the researchers also borrowed the subscale method [28] proposed by Menick et al. and extended it to video processing in figure 2.

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

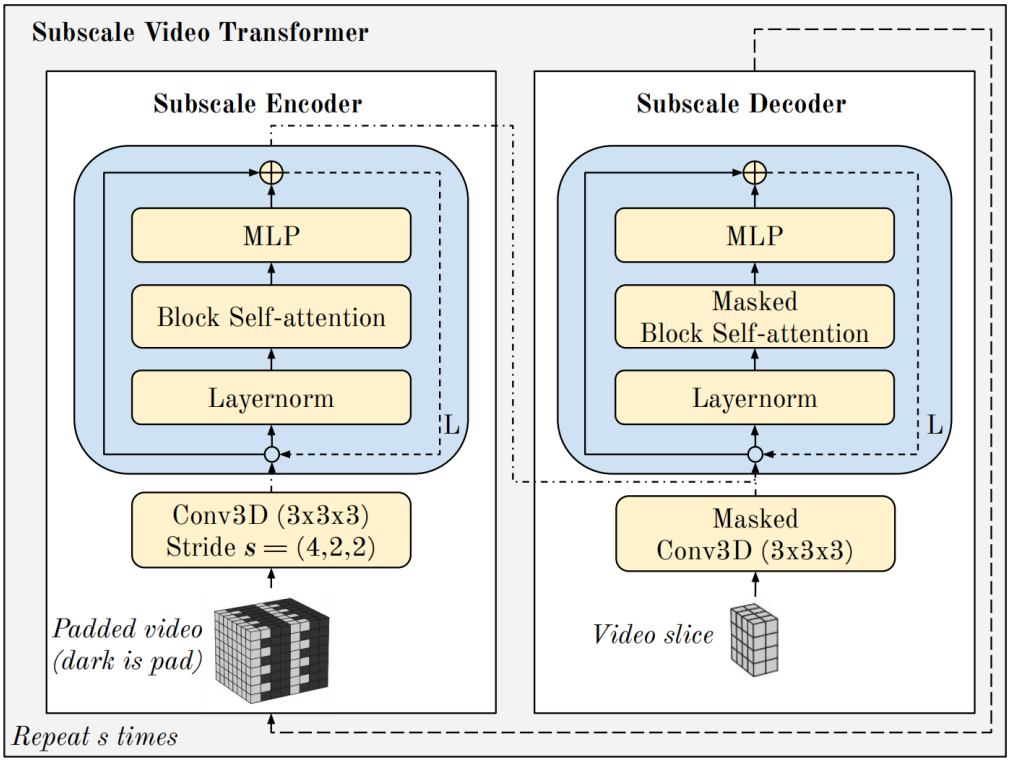

Figure 2: Structure of Video Transformer

The VPTR model was proposed by Ye et al. In order to reduce the complexity of Transformer and make it suitable for high-dimensional video representation learning, this model only applies attention to local spatial blocks and separates spatial and temporal attention. The local spatial multi-head self-attention layer proposed in this model is shared by frames with different time steps. First we define the video sequence $Z \in \mathbf{R}^{N\times T\times H\times W\times d_{\text{model}}}$, and divide it into $P=\frac{HW}{K^2}$ local patches $\{Z_1, Z_2, \cdots, Z_p\}$ evenly along the width and height dimensions, each patch size is $K\times K$, so $Z_p \in \mathbf{R}^{(NT)\times K^2\times d_{\text{model}}}$. Each multi-head self-attention formula is as follows:

$$\text{MHSA}(Z_p)=\text{Concat}[head(Z_p)_1,\cdots head(Z_p)_h]. \tag{7}$$

The calculation formula of head is shown in formula (8).

$$head(Z_p)_i=\text{SoftMax}[\frac{(Z_p^Q W_i^Q)(Z_p^K W_i^K)}{\sqrt{d_{\text{model}}/h}}]Z_p W_i^V, \tag{8}$$

where Z_p^Q, Z_p^K refer to query and key, and W_i^Q, W_i^K, W_i^V are the linear projection matrix of query, key and value of each head.

Although the local self-attention mechanism is computationally friendly, it lacks rich context information, so combining local self-attention and global self-attention is an improvement idea. Gupta et al. proposed MaskViT [30], which uses windows to calculate self-attention in video prediction tasks. The window self-attention of this model is based on two types of non-overlapping configurations. The first one is spatial window (SW): attention is limited to all tokens within a subframe of size $1\times h\times w$; the second type is spatiotemporal window (STW): attention is limited to a 3D window of size $T\times h'\times w'$.

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

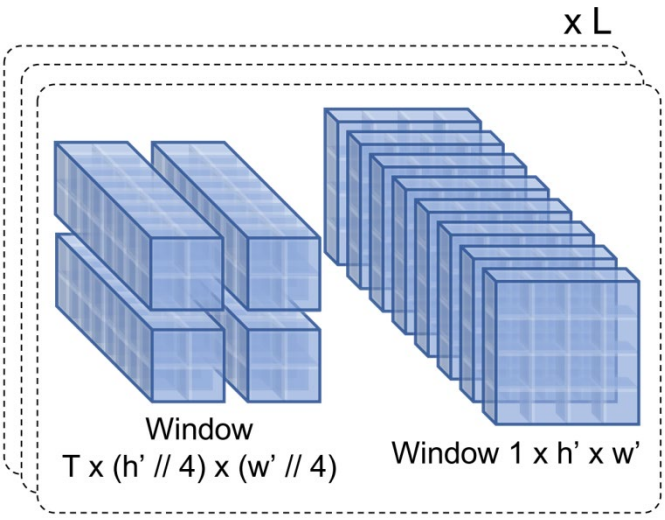

Figure 3: Bidirectional Window Transformer

The model sequentially stacks two types of window configurations, obtaining local and global interactions in a single block repeated L times for figure3. Furthermore, they found that a small window size of $h' = w' = 4$ is sufficient to build a good video prediction model while significantly reducing memory requirements.

3.2.2. Using Transformer in latent space

Based on the Video Transformer model mentioned in the previous chapter, Rakhimov et al. proposed the Latent Video Transformer [31], which uses VQ-VAE [32] as a frame autoencoder and Video Transformer as an autoregressive generation model. What needs to be added is that the main difference between VQ-VAE and VAE [33] is that the former discretizes the output of the encoding module through vector quantization, which makes it easier to model the prior. What needs to be added is that the main difference between VQ-VAE and VAE is that the former discretizes the output of the encoding module through vector quantization, which makes it easier to model the prior. For VQ-VAE, which uses PixelCNN [34] to learn priors, discrete encoding only needs to simply use softmax operation for multi-classification. Therefore, the main improvement of Latent Video Transformer based on Video Transformer is to use frame autoencoder to learn discrete potential representation, and then transfer the task of video modeling from pixel space to potential space, which reduces GPU memory consumption and speed up inference.

Yan et al. believed that video contains a large amount of spatial and temporal redundancy, and proposed to perform noise reduction downsampling encoding on high-resolution input to eliminate redundancy, and then perform autoregressive modeling in the downsampled latent space without spatiotemporal redundancy, so as to achieve good results. Based on this, VideoGPT [35] is proposed, which combines the architectures of VQ-VAE and Image-GPT. In order to reduce the dimensionality, MaskViT mentioned in the previous section trains a frame autoencoder for a single video frame, so as to represent the video as a sequence of discrete tokens grids, and each video frame is individually labeled into a tokens grid of size 16×16 regardless of the original resolution. In practice they used VQ-GAN [36], which improved VQ-VAE by adding adversarial resistance [37] and perceptual loss [38], [39].

Similarly, Wu et al. proposed Nüwa [40], which also uses VQ-GAN to encode video frames, but the model simply uses 2D VQ-GAN to encode each frame of video instead of extending the convolution of the VQ-VAE encoder from 2D to 3D and training video-specific representations like VideoGPT. Such operations make Nüwa applicable to a variety of visual synthesis tasks, including image and video tasks, forming a unified multi-modal pre-training model.

3.2.3. Combination of convolution and Transformer

Convolutional Neural Networks (CNN) can effectively model the spatial structure of images [41] and perform well when a large number of labeled training samples are available.However, when applied to video tasks, due to the interference of realistic factors such as object deformation and displacement, scene brightness changes, and the limited receptive field, which is more suitable for short-range inter-frame dependency modeling, the performance of convolutional neural networks in video

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

prediction has been greatly limited.So researchers began to try to combine the CNN structure with the RNN structure, giving full play to the temporal relationship modeling ability of RNN and the spatial feature extraction ability of CNN. For example, ConvLSTM and its variant models have achieved good results. Inspired by this idea, scholars such as Liu proposed ConvTransformer [42], a network combining CNN and Transformer, which is more parallel than methods based on ConvLSTM, and has achieved better results in video prediction and video frame insertion tasks. The core part of the network is a multi-head convolutional self-attention layer, which is used to learn long-range spatial and temporal dependencies of sequences in Figure 4.

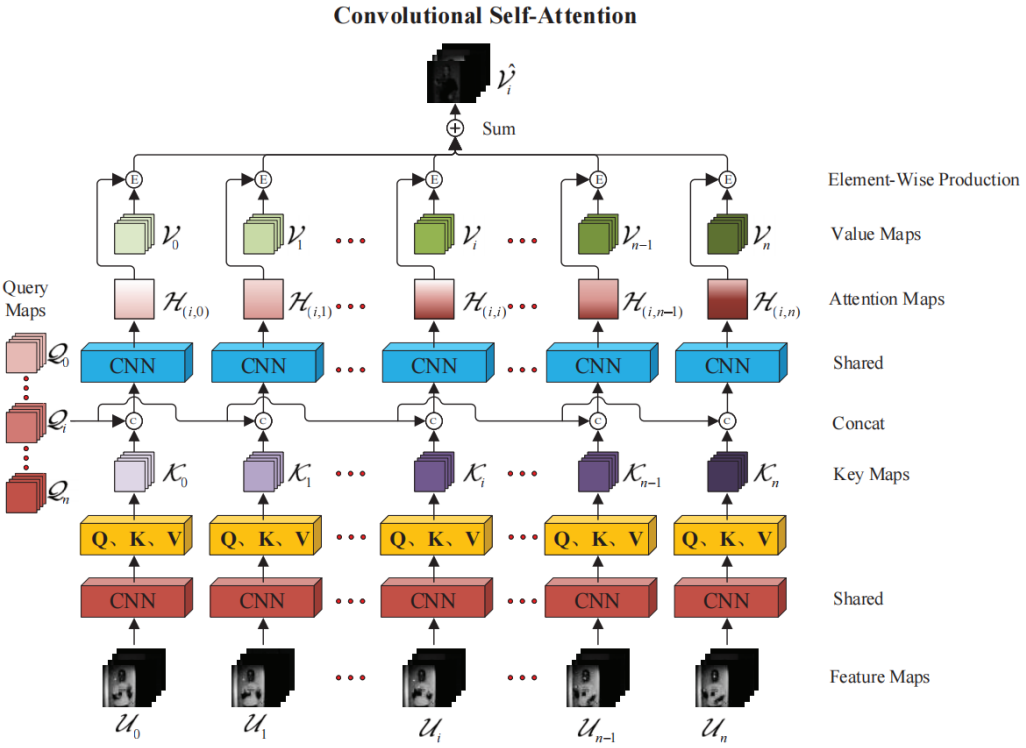

Figure 4: Structure of Convolutional Self-Attention Layers

ConvTransformer consists of five main modules, namely feature embedding module, position encoding module, encoder module, decoder module and synthetic feed-forward network, where each encoding layer and decoding layer has a multi-head convolutional self-attention layer.

Specifically, the implementation path of the convolutional self-attention layer is as follows:

First we define the input $U = \{U_0, U_1, \cdots, U_n\}, U_i \in \mathbf{R}^{H \times W \times d_{\text{model}}}$, and then each video frame goes through a convolutional subnetwork with an output channel number of 3. Each channel represents the query, key and value of this frame respectively. The output expression is

$$U' = \{\{Q_0, K_0, V_0\}, \{Q_1, K_1, V_1\}, \cdots, \{Q_n, K_n, V_n\}\}, \{Q_i, K_i, V_i\} \in \mathbf{R}^{H \times W \times 3}.$$

After obtaining U', we need to calculate the attention map $H_{(i,j)} \in \mathbf{R}^{H \times W \times 1}$ between the i-th frame and the j-th frame. The calculation method is to add Q_{i} and K_{j} and send the result to the convolutional sub-network to get the output. After all inter-frame relationships are calculated, for the i-th frame, its attention map is

$$H_{(i)} = \{H_{(i,0)}, H_{(i,1)}, \cdots, H_{(i,\text{n})}\}, H_{(i,k)} \in \mathbf{R}^{\text{H} \times \text{W} \times 1}.$$

Then we perform concat and softmax operations on the third dimension to get the attention map $H_i \in \mathbf{R}^{H \times W \times n}$ of the i-th frame:

$$H_{(i)} = \text{SoftMax}(H_{(i)})_{\text{dim}}, \text{dim} = 3. \tag{9}$$

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

In the last step, we perform point multiplication and accumulation of $H_{(i,j)}$ and V_j to finally get the output of the i-th frame through the self-attention mechanism:

$$V_{(i)} = \sum_{j=1}^{n} H_{(i,j)} V_j \,. \tag{10}$$

The multi-head convolutional self-attention mechanism is expressed as formula (11):

$$\text{MultiHead}(\hat{V}_i) = \text{Concat}(\hat{V}_{i_1}, \cdots, \hat{V}_{i_h}) \,. \tag{11}$$

In the end, the researchers compared ConvTransformer with DVF [43] and MCnet [44] on the three data sets of UCF101 [45], Adobe240fps [46], and Video90K [47]. The comparison of PSNR and SSIM values in the results shows that ConvTransformer has advantages in accurate pixel value prediction of videos.

3.2.4. Autoregressive, partial autoregressive and non-autoregressive Transformer

In Section 3.2.1 we mentioned that Ye et al. proposed VPTR, a Transformer-based video prediction model that only applies attention to local spatial blocks in order to reduce complexity. In this study, the researchers also introduced three variants of VPTR—fully autoregressive VPTR (VPTR-FAR), partially autoregressive VPTR (VPTR-PAR) and non-autoregressive VPTR (VPTR-NAR), and compared their performance.

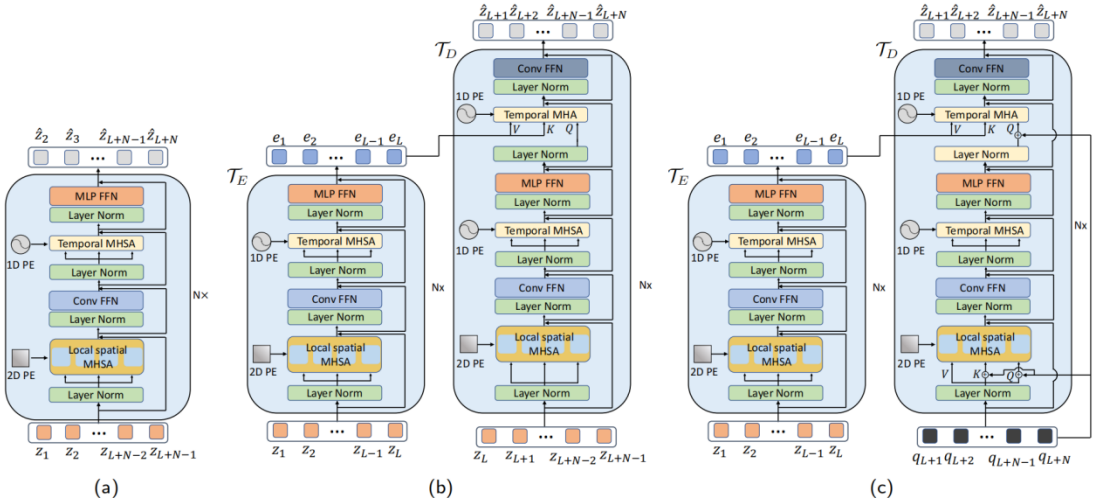

Figure5: Three VPTR variant model structures, where (a), (b) and (c) are VPTR-FAR, VPTR-PAR, and VPTR-NAR, respectively.

Fully autoregressive VPTR predicts the next frame conditionally on all previous frames, which is the most common mode of video prediction SOTA models. In terms of structure, it is only composed of multiple VidHRFormer blocks (Figure 5) stacked. In this mode researchers model the following distribution:

$$p(x_1, \cdots, x_L, \cdots, x_{L+N}) = \prod_{t=1}^{L+N} p(x_t \mid x_{t-1}, \cdots, x_1) \,. \tag{12}$$

The second mode is partial autoregressive VPTR, which includes a Transformer encoder and a decoder, where the encoder encodes all past frame features. Compared with full autoregressive VPTR, it only decomposes the probability distribution of future frames into a product of a series of conditional distributions, so this model is named partial autoregressive VPTR. In this mode researchers model the following distribution:

$$p(x_{L+N}, \cdots, x_{L+1} \mid x_L, \cdots, x_1) = \prod_{t=L+1}^{L+N} p(x_t \mid x_{t-1}, \cdots, x_{L+1}, x_L, \cdots, x_1) \,. \tag{13}$$

In order to reduce the cumulative error of prediction and improve the speed of inference, the researchers also proposed a non-autoregressive variant named VPTR-NAR, which directly simulates

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

the following distribution:

$$p(x_{L+N},\cdots,x_{L+1}\,|\,x_L,,x_1)\,. \tag{14}$$

Comparing the prediction performance of the three models, the researchers found that the partial autoregressive model has similar performance to the full autoregressive model, but the former has faster inference speed; comparing the two autoregressive variants with the non-autoregressive variant , the first two variants achieve better PSNR and SSIM than the latter, but due to the accumulation of errors introduced by loop inference, autoregressive variants have a faster performance drop than the non-autoregressive variant.

4. Summary and Prospect

This paper introduces the principle and basic structure of Transformer, and sorts out the application of Transformer in the current hot field of video prediction, and summarizes the improvements and innovation directions of the Transformer-based video prediction models, such as using local self-attention to reduce memory requirements, transferring the application scene of Transformers to latent space to reduce memory consumption and speed up inference, integrating convolution operations into the multi-head self-attention mechanism to enhance the ability to extract features, and considering the respective advantages and disadvantages of autoregressive models and non-autoregressive models to use Transformers according to scenarios, etc. The improvement ideas above have important reference significance for future scientific research. It is not difficult to find that many improvements are aimed at reducing the complexity of the model, which is mainly due to the quadratic growth relationship between the complexity of the self-attention mechanism and the image scale.

In addition, the strong support of the Transformer structure has led to the use of Transformer-based prediction models for multiple sub-tasks of video prediction, such as anomaly detection and motion synthesis based on video prediction. Yuan et al. proposed TransAnomaly, a Transformer-based video prediction model for video outlier detection, which combines U-Net and ViViT to capture richer temporal information and more global context; Jin et al. proposed a Transformer-based baseline model ANDT for abnormal detection of UAV videos; Pan et al. added a Transformer-based controller to the key frame motion synthesis model based on a recurrent neural network for modeling root loci and velocity factors to better exploit the temporal context of video frames and enable fine-grained motion control; Bi et al. converted the online car-hailing order data into video frame data containing spatial information and time series, and proposed a prediction model called SPTformer based on the Transformer architecture to predict travel demands of online car-hailing from the scale of time and space.

Based on the current application status of Transformer in video prediction tasks, we need to discuss its development direction in the future. The most important point is that high-dimensional data computing models are still in demand. The existing Transformer video generation and prediction models still have problems such as large number of parameters and high computational complexity. Therefore, the training process of the model requires high hardware requirements, long training time, and high computational cost. Compared with NLP, the pixels in the image data need higher resolution, and Transformer is based on the calculation method of global self-attention, which makes the calculation complexity increase quadratically with the image scale, resulting in an excessively large amount of calculation. At present, there are measures such as partial attention and calculation in latent space to optimize, but continuous reduction of complexity will still be a long-term task in the future. Designing a calculation model suitable for high-dimensional data is a research hotspot. At the same time, the strategy selection of model construction should also be taken into consideration by researchers. For example, the VPTR mentioned in this paper proposed three variants based on a similar structure. It was found that the non-autoregressive model design method has advantages such as faster reasoning speed and slower performance degradation. Of course, there are also some disadvantages in this variant. These comparative studies have reference significance for future strategy selection and trade-off.

In addition, the combination of Transformer and CNN in various ways still has the value of in-depth research. ConvTransformer mentioned in this paper proposes a multi-head convolution self-attention layer, which integrates the convolution operation into the self-attention calculation, and has achieved good results in video prediction tasks. We can expect that in the future, Transformer and CNN will have a broader space for integration, such as integrating the convolution layer and pooling layer to enhance

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

the local feature extraction ability of self-attention, or combining the features of Transformer and CNN with each other through the bridge to realize feature fusion through the parallel structure, so as to give full play to their advantages.

Looking forward to the video prediction task, we can expect that a unified framework suitable for multi-tasks will continue to emerge in the future. Previous multimodal models require different processing methods for different types of data, which not only complicates the model structure, but also affects the processing effect of multiple types of data. However, Transformer's self-attention mechanism has a very powerful feature extraction ability, which is suitable for processing different types of data, converting different types of data into long sequences, and then all of them can be processed by Transformer in a unified manner, which has a great advantage over CNN-based and RNN-based models. Nüwa mentioned above has made an attempt and is suitable for generation and prediction tasks of video and image.

In general, there have been many attempts to use the Transformer structure in video prediction tasks, but it is undeniable that this attempt is still in its infancy and requires more research and improvement. This structure has shown certain advantages in the field of video generation and prediction, but it also has certain limitations due to its own structural problems, and has great potential for improvement. It is believed that the Transformer-based models will show more and more powerful generation and prediction capabilities in the future, forming a more complete multi-task fusion system.

References

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

[2] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[3] 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.

[4] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 12873-12883.

[5] Arnab A, Dehghani M, Heigold G, et al. Vivit: A video vision transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6836-6846.

[6] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer[C]//International conference on machine learning. PMLR, 2018: 4055-4064.

[7] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Springer, Cham, 2020: 213-229.

[8] 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.

[9] Ranzato M A, Szlam A, Bruna J, et al. Video (language) modeling: a baseline for generative models of natural videos[J]. arXiv preprint arXiv:1412.6604, 2014.

[10] Lotter W, Kreiman G, Cox D. Deep predictive coding networks for video prediction and unsupervised learning[J]. arXiv preprint arXiv:1605.08104, 2016.

[11] Bolte J A, Bar A, Lipinski D, et al. Towards corner case detection for autonomous driving[C]//2019 IEEE Intelligent vehicles symposium (IV). IEEE, 2019: 438-445.

[12] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[J]. Advances in neural information processing systems, 2015, 28.

[13] Liu W, Luo W, Lian D, et al. Future frame prediction for anomaly detection–a new baseline[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6536-6545.

[14] Clark A, Donahue J, Simonyan K. Adversarial video generation on complex datasets[J]. arXiv preprint arXiv:1907.06571, 2019.

[15] Tulyakov S, Liu M Y, Yang X, et al. Mocogan: Decomposing motion and content for video generation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1526-1535.

[16] Terwilliger A, Brazil G, Liu X. Recurrent flow-guided semantic forecasting[C]//2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019: 1703-1712.

[17] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[J]. Advances in neural information processing systems, 2015, 28.

[18] Villegas R, Yang J, Zou Y, et al. Learning to generate long-term future via hierarchical

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

prediction[C]//international conference on machine learning. PMLR, 2017: 3560-3569.
[19] Zhang J, Wang Y, Long M, et al. Z-order recurrent neural networks for video prediction[C]//2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019: 230-235.
[20] Villegas R, Pathak A, Kannan H, et al. High fidelity video prediction with large stochastic recurrent neural networks[J]. Advances in Neural Information Processing Systems, 2019, 32.
[21] Wu B, Nair S, Martin-Martin R, et al. Greedy hierarchical variational autoencoders for large-scale video prediction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2318-2328.
[22] Babaeizadeh M, Saffar M T, Nair S, et al. FitVid: Overfitting in pixel-level video prediction[J]. arXiv preprint arXiv:2106.13195, 2021.
[23] Weissenborn D, Täckström O, Uszkoreit J. Scaling autoregressive video models[J]. arXiv preprint arXiv:1906.02634, 2019.
[24] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer[C]//International conference on machine learning. PMLR, 2018: 4055-4064.
[25] Chen X, Mishra N, Rohaninejad M, et al. Pixelsnail: An improved autoregressive generative model[C]//International Conference on Machine Learning. PMLR, 2018: 864-872.
[26] Menick J, Kalchbrenner N. Generating high fidelity images with subscale pixel networks and multidimensional upscaling[J]. arXiv preprint arXiv:1812.01608, 2018.
[27] Ye X, Bilodeau G A. Video prediction by efficient transformers[J]. Image and Vision Computing, 2022: 104612.
[28] Gupta A, Tian S, Zhang Y, et al. Maskvit: Masked visual pre-training for video prediction[J]. arXiv preprint arXiv:2206.11894, 2022.
[29] Rakhimov R, Volkhonskiy D, Artemov A, et al. Latent video transformer[J]. arXiv preprint arXiv:2006.10704, 2020.
[30] Van Den Oord A, Vinyals O. Neural discrete representation learning[J]. Advances in neural information processing systems, 2017, 30.
[31] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.
[32] Zhong K, Wang Y, Pei J, et al. Super efficiency SBM-DEA and neural network for performance evaluation[J]. Information Processing & Management, 2021, 58(6): 102728.
[33] Jan N, Gwak J, Pei J, et al. Analysis of networks and digital systems by using the novel technique based on complex fuzzy soft information[J]. IEEE Transactions on Consumer Electronics, 2022, 69(2): 183-193.
[34] Yu Z, Pei J, Zhu M, et al. Multi-attribute adaptive aggregation transformer for vehicle re-identification[J]. Information Processing & Management, 2022, 59(2): 102868.
[35] Li J, Li S, Cheng L, et al. BSAS: A Blockchain-Based Trustworthy and Privacy-Preserving Speed Advisory System[J]. IEEE Transactions on Vehicular Technology, 2022, 71(11): 11421-11430
[36] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution[C]//European conference on computer vision. Springer, Cham, 2016: 694-711.
[37] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 586-595.
[38] Wu C, Liang J, Ji L, et al. Nüwa: Visual synthesis pre-training for neural visual world creation[C]//European Conference on Computer Vision. Springer, Cham, 2022: 720-736.
[39] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[40] Liu Z, Luo S, Li W, et al. Convtransformer: A convolutional transformer network for video frame synthesis[J]. arXiv preprint arXiv:2011.10185, 2020.
[41] Liu Z, Yeh R A, Tang X, et al. Video frame synthesis using deep voxel flow[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4463-4471.
[42] Villegas R, Yang J, Hong S, et al. Decomposing motion and content for natural video sequence prediction[J]. arXiv preprint arXiv:1706.08033, 2017.
[43] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.
[44] Su S, Delbracio M, Wang J, et al. Deep video deblurring for hand-held cameras[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1279-1288.
[45] Xue T, Chen B, Wu J, et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision, 2019, 127(8): 1106-1125.
[46] Yuan H, Cai Z, Zhou H, et al. TransAnomaly: Video Anomaly Detection Using Video Vision Transformer[J]. IEEE Access, 2021, 9: 123977-123986.

ISSN 2616-5775 Vol. 6, Issue 13: 35-46, DOI: 10.25236/AJCIS.2023.061306

[47] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.
[48] Arnab A, Dehghani M, Heigold G, et al. Vivit: A video vision transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6836-6846.
[49] Jin P, Mou L, Xia G S, et al. Anomaly Detection in Aerial Videos With Transformers[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-13.
[50] Pan J, Wang S, Bai J, et al. Diverse Dance Synthesis via Keyframes with Transformer Controllers[C]//Computer Graphics Forum. 2021, 40(7): 71-83.
[51] Bi S, Yuan C, Liu S, et al. Spatiotemporal Prediction of Urban Online Car-Hailing Travel Demand Based on Transformer Network[J]. Sustainability, 2022, 14(20): 13568.