A REVIEW OF VIDEO PREDICTION BASED ON RECURRENT NEURAL NETWORKS

Haoyu WU[1], Lei YU[2] and Qiuyue DENG[3]

[1]Master of Science, College of Information and Communication Engineering, Harbin Engineering University
(145 Nantong street, Nangang District, Harbin150001, China)
E-mail: why1024@hrbeu.edu.cn

[2]Associate Professor, College of Information and Communication Engineering, Harbin Engineering University
(145 Nantong street, Nangang District, Harbin150001, China)
E-mail: yulei@hrbeu.edu.cn

[3]Master of Science, College of Information and Communication Engineering, Harbin Engineering University
(145 Nantong street, Nangang District, Harbin150001, China)
E-mail: doreenyue@hrbeu.edu.cn

Recurrent Neural Network (RNN) as a Convolutional Neural Networks (CNN) improves the time dimension and is widely used in sequence information processing with the development of computer vision research. Video prediction is one of the classical tasks in sequence information processing. It requires the model to learn the characteristics of the first few images and infer the next few images. RNN has a circular network structure, which makes the current output information of the network affect the next output information. This feature enables RNN to process sequence information more effectively. Videos usually contain multiple pictures of time series information, RNN is a reasonable choice for video prediction. With the advancement of video prediction research in recent years, the application of various improved RNNs in video prediction is also developing. This paper summarizes the common research directions of video prediction, introduces the research progress of RNN in the field of video prediction in recent years, and discusses the opportunities and challenges of RNN in the field of video prediction.

Keywords: Video prediction, RNN, sequence information processing

1. INTRODUCTION

Video prediction is a challenging task in computer vision. Video prediction not only requires the model to have the ability to infer the details of the video frame, but also requires the model to restore the corresponding image. With the continuous application of deep learning in this task, the research of video prediction has made great progress. Learning through massive video data representation[1)] has wide application value in unmanned driving, weather prediction, robot decision-making, traffic flow prediction and other fields.

2. VIDEO PREDICTION

Video prediction refers to giving a continuous video frame X_1、X_2、X_3……X_n, constructing a model can accurately predict X_{n+1} or X_{n+2}、X_{n+3}……X_{n+t}(t is the number of frames to be predicted)[2)]. In video prediction tasks, there is no need to label the data extra, so video prediction is an unsupervised category. In video prediction tasks, public datasets include KTH[3)]、Moving-MNIST[4)]、UCF-101[5)]. In addition, evaluating the quality of image generation is also an important process, which can effectively reflect the quality of the model by evaluating the effect of generating future frames. The commonly used measures to assess video quality are mean square error (MSE), peak signal to noise ratio (PSNR), structural similarity index (SSIM).

3. RNN FOR VIDEO PREDICTION

Compared with CNN[6], RNN[7] is a neural network that can process sequence data. Its state travels through its own network and can process any length of sequence. RNN is closer to the structure of biological neural network than feed forward neural network[8]. Long Short-Term Memory (LSTM)[9] is also a variant of RNN, which optimizes the disappearance of gradients in RNN itself.

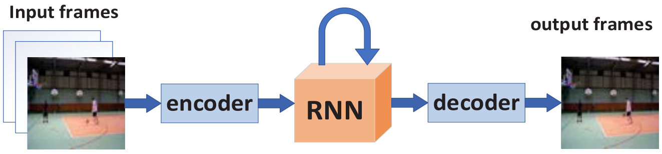

Fig.1 Recurrent neural network structures in video prediction

The common structure of video prediction is shown in **Figure 1**, based on which many meaningful models have been proposed. The main research directions include direct pixel synthesis, motion coding with content separation, etc. Direct pixel synthesis is the most common research method. Direct pixel synthesis is an end-to-end model. This chapter will focus on these research directions.

(1) Direct pixel synthesis

Direct pixel synthesis attempts to directly predict future pixel strength for video prediction models without any explicit modeling of scene dynamics. The early model was proposed by Convolution LSTM (ConvLSTM) by Xingjian Shi et al.[4] to solve rainfall forecasting. This model improves the deficiency of spatial correlation in the traditional fully connected long short-term memory network (FC-LSTM), making ConvLSTM achieve better results than LSTM in acquiring spatial-temporal relationships.

YunboWang et al.[10] proposed the PredRNN. If future frames need to be predicted, the model needs to remember as much historical detail as possible. Therefore, the authors propose a new cyclic architecture that allows cross-layer interaction of memory states belonging to different LSTMs. As a key component of PredRNN, the author designed a new space-time LSTM (ST-LSTM) unit, shown in **Figure 2**. It models the representation of space and time in a single memory unit and transmits memory vertically and horizontally. PredRNN achieved the best predictions of the time on three video datasets. It is a general module framework for predictive learning and is not limited to video prediction.

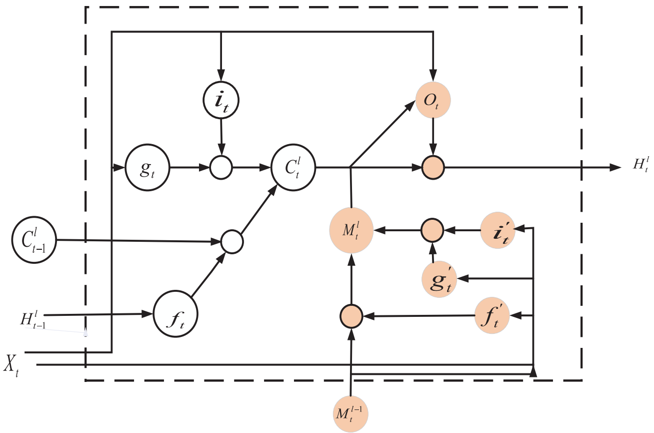

Fig.2 ST-LSTM

After that, Yunbo Wang et al.[11] improved it and proposed PredRNN++.The author constructs a new LSTM structure (Causal LSTM) by cascading the double memory. A gradient highway unit work is also proposed to solve the problem of gradient disappearance, and the structure and Causal LSTM are seamlessly connected. To solve the non-stationary information prediction in video information, Yunbo Wang et al.[12] then proposed the Memory In Memory (MIM) model. This model focuses on the stationary and non-stationary problems in model prediction. For the stationary and non-stationary information in the data, the main difficulty is the non-stationary prediction, because non-stationary information is basically irregular, MIM model has achieved good performance in the processing of non-stationary information.

Meanwhile, 3D convolution[13] is beginning to be used in video processing3D convolution adds a third time dimension to 2D convolution, which enables the convolution process not only to focus on the spatial characteristics of single-frame data, but also to extract time information between multiple-frame data. Therefore, the fusion of 3D convolution and LSTM is also a research idea. However, because the working mechanisms of the two modules are quite different, the direct integration effect is not excellent. Yunbo Wang et al.[14] proposed the strong memory network eidetic 3d LSTM (E3D-LSTM) after MIM. Replacing gate update operation in original LSTM with 3D convolution enables LSTM to extract short-term dependent representation and motion features not only in time but also in space, thus combining the two networks at a deeper mechanism level. In addition, the introduction of self-attention mechanism[15] in LSTM further enhances LSTM's long-term memory and makes it more aware of the effects of long-distance information.

Zhihui Lin et al.[16] extended the idea of direct pixel synthesis that Yunbo Wang et al. improved from within the network and proposed Self-Attention ConvLSTM. The basic model is the combination of self-attention and ConvLSTM. In addition, a memory module is introduced, where both memory and hidden state h are aggregated through self attention, and a gated structure like LSTM is constructed, which is valid in multiple datasets.

In addition to the internal structure improvement of RNN, great progress has been made in network structure improvement. Wei Yu et al.[17] proposed a two-way convolution structure, CrevNet. This model builds a network of two-way encoding and two-way decoding and draws on the idea of similar residual[18]. It is not only superior to other models in predicting results, but also has great advantages in occupying resources compared with other structures.

Beibei Jin et al.[19] proposed Spatial Wavelet Analysis Modules (S-WAMs) and Temporal Analysis Modules (T-WAM) starting with the frequency domain characteristics of the image. The main idea is to use the wavelet transform to obtain low-frequency and high-frequency information in time and space domains respectively, and to integrate it into the network, to solve the problem of inaccurate spatial dimension details and temporal dimension motion information in video prediction tasks. The model captures different frequency information in video.

Vincent Le Guen et al.[20] proposed PhyDNet to construct a model that predicts the laws of objective physical motion, as shown in **Figure 3**. The main idea is to try to build a physical constraint model with deep networks by using convolution simulation bias and moment loss as a supervisor to learn physical information to supplement existing networks.

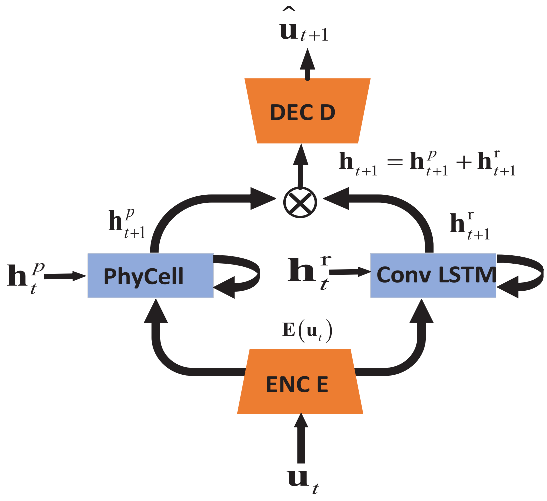

Fig.3 PhyDNet

(2) Motion coding with separated contents

The idea of motion coding with content separation is since in most video data, the background generally shows a static state or little movement, and the moving part is only the smaller position in the image. Therefore, separating the moving part from the background, coding and predicting the moving part and fusing it with the background can get better prediction results and reduce the number of parameters of the model.

Finn et al.[21] proposed a Convolutional Dynamic Neural Advection (CDNA) model. In CDNA model, the appearance and background information of an object can be obtained directly in the previous frame without requiring model storage. This model combines the appearance information of the previous frame with the predicted motion. The result can better predict the future multi-step video sequence or even objects that you have not seen in training. CDNA model is a good attempt to combine video prediction with robotics.

Villegas et al.[22] proposed the Motion-Content network (MCnet). When a motion encoder captures temporal dynamics in a sequence of image differences, the content encoder extracts meaningful spatial features from the last observed RGB frames. The network then calculates the motion content characteristics of the input decoder to predict the next frame. The model uses adversarial loss to optimize the output image quality.

(3) Reflections on loss function

In video prediction models based on recurrent neural network architecture, the mean square error (MSE) is used as a loss function in most cases. However, in the actual training, minimizing the mean square error as a loss function will result in excessive pursuit of the minimum mean when correcting pixel point values after reverse propagation, which will result in blurring. Mathieu et al.[23] proposed three strategies to solve this problem: multiscale structure, adversarial training, and gradient difference loss function of images. The gradient differential loss function is shown in Formula (1).

$$\begin{aligned}\mathcal{L}_{gdl}(X,Y) &= L_{gdl}(\hat{Y},Y) \\ &= \sum_{i,j} || Y_{i,j} - Y_{i-1,j} | - | \hat{Y}_{i,j} - \hat{Y}_{i-1,j} ||^{\alpha} \\ &+ || Y_{i,j-1} - Y_{i,j} | - | \hat{Y}_{i,j-1} - \hat{Y}_{i,j} ||^{\alpha}\end{aligned} \quad (1)$$

4. SUMMARY AND PROSPECT

Video prediction, as a new research direction, not only challenges the learning ability of the model's features, but also has higher requirements on the model's generation ability. Therefore, in recent years, people have proposed a variety of new methods to complete the task of video prediction, which have achieved good results and promoted the development of RNN, We summarize it in **Table 1**.

Table 1 summary.

Methods	Main models
Direct pixel synthesis	ConvLSTM, PredRNN, etc.
Motion coding with separated contents	CDNA, MCnet, etc
Reflections on loss Function	Gradient Difference Loss, etc.

The existing methods also have some deficiencies that need to be further studied. First, the loss function in the model is not rich enough. In RNN and variant LSTM, MSE is mainly used for model training, while MSE results in later training models that blur images in pursuit of lower loss values. Although gradient differential loss is introduced later, there is still no in-depth study to improve the loss function, and there is no loss function that fits the video prediction.

Then, the RNN architecture itself is more focused on extracting video sequence features, which is not good at generating video frames. Although the early network architecture is not mature, its generation results can basically reflect the movement trend, the difference is basically reflected in the quality of image generation. Generative Adversarial Networks[24] calculates the loss by generating and discriminating two networks, which can produce higher quality output images. It is possible to obtain better video prediction results when applied to RNN structures.

Finally, as a new research direction, video prediction only exists in academia at present and has not been started in practical application. Therefore, combining video prediction with industrial applications is also a challenging direction.

Video prediction learning is a powerful means to understand and model the dynamics of natural scenes, and it is also an important breakthrough point for unsupervised learning. Although there are still many challenges and difficulties, with the increasing computing resources, it will also be conducive to the development of video prediction research.

ACKNOWLEDGMENT: The work is supported by NSFC (61771155) and Fundamental Research Funds for the Central Universities. And thanks a lot to Miss Zheng Liying for her brilliant help to the research.

REFERENCES

1) Y. Bengio, A. Courville, and P. Vincent,: Representation Learning: A Review and New Perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, 2013.

2) Lingfei M., Hongliang J., Xuanpeng L.: Review of video prediction based on deep learning. Journal of intelligent systems, pp. 85-96, 2018.

3) C. Schuldt, I. Laptev, and B. Caputo: Recognizing human actions: a local SVM approach, in Proceedings of the 17th International Conference on Pattern Recognition, 2004.

4) Shi, X., Chen, Z., Wang, H et al.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. arXiv:1506.04214, 2015.

5) Soomro K., Zamir A. R., Shah M.: UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint axXiv:1202.0402, 2012.

6) Krizhevsky, A. , I. Sutskever , and G. Hinton . :ImageNet Classification with Deep Convolutional Neural Networks. NIPS Curran Associates Inc. 2012.

7) Zaremba W, Sutskever I, and Vinyals O.: Recurrent Neural Network Regularization. arXiv:1409.2329, 2014.

8) G. E. Hinton, S. Osindero, and Y.-W. Teh,: A Fast Learning Algorithm for Deep Belief Nets, Neural Computation, vol. 18, no. 7, pp. 1527-1554, 2006.

9) A. Graves: Long Short-Term Memory, Studies in Computational Intelligence, pp. 37-45, 2012.

10) Wang Y, Long M, Wang J., Z. Gao，PS. Yu: PredRNN: recurrent neural networks for predictive learning using spatiotemporal LSTMs, Proceedings of the 31st International Conference on Neural Information Processing Systems, pp, 879–888, 2017.

11) Yunbo W., Zhifeng G., Mingsheng L., Jianmin W. and Philip Y.: PredRNN++: Towards A Resolution of the Deep-in-Time Dilemma in Spatiotemporal Predictive Learning, arXiv:1804.06300, 2018.

12) Wang Y., Zhang J., Zhu H., Long M. and Yu PS.: Memory in Memory: A Predictive Neural Network for Learning Higher-Order Non-Stationarity From Spatiotemporal Dynamics, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2019.

13) D. Tran, L. Bourdev, R. Fergus, L Torresan，M Paluri.: Learning Spatiotemporal Features with 3D Convolutional Networks, in 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4489-4497, 2015.

14) Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei: Eidetic 3d LSTM: A model for video prediction and beyond, ICLR, 2019.

15) S. Woo, J. Park, J.-Y. Lee and IS Kweon.: CBAM: Convolutional Block Attention Module, Computer Vision – ECCV 2018, Lecture Notes in Computer Science, pp. 3-19, 2018.

16) Lin Z., Li M., Zheng Z., Cheng Y. and Chun Y.:.Self-Attention ConvLSTM for Spatiotemporal Prediction. Proceedings of the AAAI Conference on Artificial Intelligence , pp. 11531-11538, 2020.

17) W. Yu, Y. Lu, S. Easterbrook, and S. Fidler,: Efficient and information-preserving future frame prediction and beyond, ICLR, 2020.

18) K. He, X. Zhang, S. Ren et al.: Deep Residual Learning for Image Recognition, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

19) B. Jin, Y. Hu, Q. Tang et al.: Exploring Spatial-Temporal Multi-Frequency Analysis for High-Fidelity and Temporal-Consistency Video Prediction, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4553-4562, 2020.

20) V. Le Guen, and N. Thome, :Disentangling Physical Dynamics From Unknown Factors for Unsupervised Video Prediction, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11471-11481, 2020.

21) Finn, C., Goodfellow, I. and Levine, S.:Unsupervised Learning for Physical Interaction through Video Prediction, arXiv:1605.07157, 2016.

22) Villegas, R., Yang, J., Hong, S,. Lin., X and Lee, H.: Decomposing Motion and Content for Natural Video Sequence Prediction, arXiv:1706.08033. 2017.

23) Mathieu, M., Couprie, C. and Lecun, Y.: Deep multi-scale video prediction beyond mean square error, arXiv:1511.05440, 2015.

24) IJ. Goodfellow，J. Pouget-Abadie，M. Mirza，B. Xu，D. Warde-Farley，S. Ozair，A. Courville and Y. Bengio.: Generative adversarial networks, Communications of the ACM, vol. 63, no.

11, pp. 139-144, 2020.