OVERVIEW OF RESEARCH PROGRESS OF GENERATIVE ADVERSARIAL NETWORKS IN VIDEO PREDICTION

Haicheng HU[1], Lei YU[2]and Haoyu WU[3]

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

Recent years, as an emerging unsupervised learning algorithm, GAN has received extensive attention from many researchers, become a current research hotspot. The video prediction, that the first few frames of the given model video, it can accurately predict the next few frames of video. With the excellent performance of GAN in the field of generative models, it also plays an important role in the field of video prediction. At present, GAN has shown its powerful performance in the field of computer vision. Based on, this article mainly introduces GAN in the field of computer vision, especially in the field of video prediction. Summarized several typical GAN network models for video prediction, and discussed the future research hotspots and development prospects of GAN in the field of video prediction.

Key words: Generative Adversarial Network (GAN); Video prediction; Machine learning; Unsupervised learning.

1. INTRODUCTION

GAN stands for Generative Adversarial Network[1]. Generative Adversarial Network (GAN) as a generative model, has gradually become the current hot research direction. The essence of GAN is an unsupervised model composed of a generator and a discriminator (Discriminator). Because of its strong generation ability and unique confrontation- training ideas, GAN can meet image processing, language processing, needs in many areas such as network security. Help with image generation, image fusion[2], machine translation, speech generation[3], text classification[4], and other issues.

The first chapter of this article introduces the application of GAN in video prediction, the second chapter summarizes the previous work.

2. APPLICATION OF GENERATIVE ADVERSARIAL NETWORK (GAN) IN VIDEO PREDICTION

The proposal of Generative Adversarial Network, provides a new solution to problems in the field of machine learning, its unique training

method and outstanding performance, has attracted the attention of many researchers. Because of its irreplaceable advantages in the field of computer vision, especially image generation, so many scholars learn from this method of confrontation training for video prediction. This chapter will introduce the application of GAN in video prediction from two directions: Improv-ed model based on encoder-decoder and improved model based on original GAN.

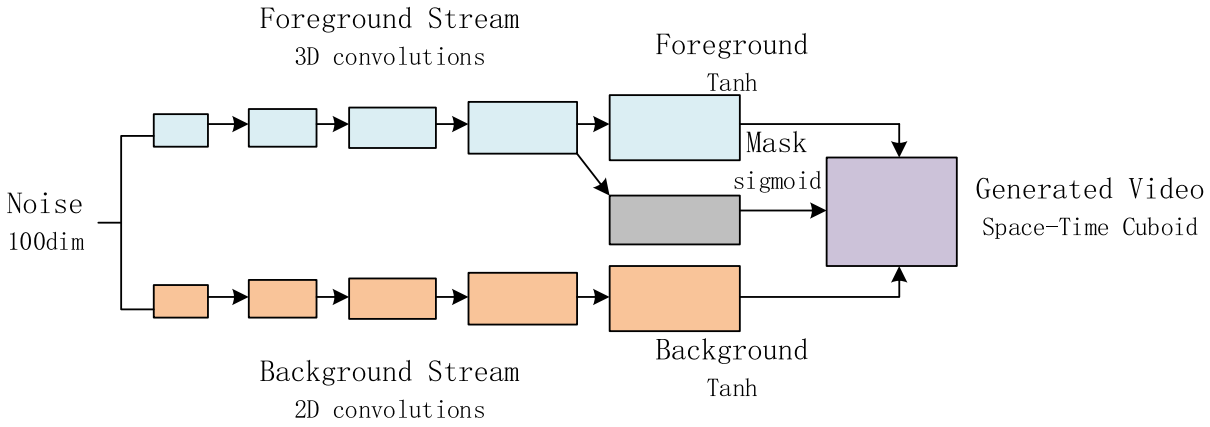

Fig.1 Generator network model

(1) Improved model based on encoder-decoder

Carl Vondrick[5] and others proposed a GAN-based network model, Can perform video prediction and video generation tasks at the same time. That is, using some unlabeled videos to train the model to solve the prediction problem and the task of video generation at the same time.

In the video generation task, this article proposes a new modeling idea: Model the foreground and background separately in the video. The background does not change in most cases, the background information is the same in multiple frames of video. Therefore, the background modeling can ignore the dimension of "time", model only as an image. Prospect information is complex and important, involving the movement of people and objects in the video, it is related to time and space dimensions. After stripping the background information from the video, prospect information will become easier to model. For example, in a table tennis game, only the two baffles and the position of the table tennis ball are in motion, the rest of the position is the background. The network structure specifically used in Generator is shown in Figure 1. The network generates the foreground and background separately, then combines the foreground and background with a Mask weighting. The drop network uses 2D Transpose Conv to model the background, the above network uses 3D Transpose Conv to model the foreground and Mask, 3D convolution can extract both temporal and spatial information simultaneously, its role is equivalent to ConvLSTM network, but it is simpler in implementation. The background image is expanded into 3D and the foreground information is weighted to obtain the final video output.

Added sparse prior loss to training $\lambda\left|m(z)\right|_1$, that is, the mask should be close to 0, the image generated at this time uses background information. Additionally, Discriminator net- work uses ordinary 3D convolution to build a two-classification network. Visualize the in- formation generated by the foreground and background in the experiment, the experimental results show that the model can automatically summarize the stable and unchanging parts of the scene as the back- ground, summarize the part of the movement into the foreground.

In the video prediction task of this network,

video prediction should generate subsequent pictures based on the pictures of the previous frames, therefore, the input of the network cannot be random noise. Must add an Encoder part to the Generator network, take the known first few frames as input, encode as a latent code by Encoder, and the sampled noise code as input at the same time. Simultaneously, adding a loss, the first frame picture of the generated Video is required to be equal to the input picture.

Liang[6] and others proposed future streaming embedded video prediction based on Dual Motion GAN. They proposed the Daul Motion Generative Adversarial Net (D-MGAN) architecture, the dual learning mec-hanism makes the future frames explicitly predicted by the model consistent with the pixel stream in the video. A closed loop is formed between basic video frame prediction and dual optical flow prediction, the feedback information between the two tasks enables the prediction tasks to promote each other. To make the predicted frame and stre-am information more realistic, the paper al-so proposes a dual training method to ensur-e that the predicted optical flow can help t-he network to reason, makes the predicted future frames more realistic, simultaneously, the future frame prediction task also makes the predicted optical flow information more realistic. The DMGAN in this article also u-ses a probabilistic motion encoder (based o-n variational autoencoder, VAE) to deal wit-h the uncertainty of the natural motion of pixels in different positions. Experiments sh-ow that this article has achieved SOTA perf-ormance on video frames and optical flow prediction tasks, at the same time, the method in this article has good generalization ab-ility in various data set scenarios, demonstrates its superiority in unsupervised video r-epresentation-learning tasks.

The DMGAN proposed in this article is mainly composed of the following three modules:

a) Probabilistic Motion Encoder, used to capture the uncertainty of motion appearing in different positions, and can generate potential representations of movement, used as input for two generators.

b) Future frame generator, video frames used to predict the future, mainly through two aspects to evaluate, evaluate the accuracy of the frame through the frame discriminator; Evaluate the accuracy of the optical flow generated between the previous frame and the predicted frame through the optical flow discriminator.

c) Optical flow generator, optical flow information used to predict the future, it also evaluates through two aspects, evaluate the accuracy of the predicted optical flow through the optical flow discriminator; Combine the predicted optical flow information with the previous frame to get the future frame through the flow-warping layer, and input the frame to the frame discriminator.

In the experimental section, this paper first makes an experimental comparison on video prediction tasks, include prediction of next frame and prediction of continuous multiple frames. Then, each module of the model was analyzed by ablation experiments. Additionally, this is still predicting light flow, the generalization ability of the proposed model is demonstrated in the tasks of optical flow estimation and unsupervised representation learning.

(2) Improved model based on the original GAN

Yong-Hoon[7] and others proposed video frame prediction technology based on retrosrpective CycleGAN. The proposed network consists of a generator and two discriminators (frame discriminator and sequence discrimi- nator).

Where the generator can predict not only future frames but also past ones, even if prediction frames are included in the input sequence.

The frame discriminator is used to determine whether the input frames are generated or real, sequence discriminator is used to determine whether the input sequence contains generated frames, false prediction if any, otherwise the prediction is true.

The objective function to minimize the training process is as follows:

$$L = L_{image} + \lambda_1 L_{LoG} + \lambda_2 L_{adv}^{frame} + \lambda_3 L_{adv}^{seq}$$

It consists of two refactoring functions and two antagonistic training functions, where is $\lambda_1, \lambda_2, \lambda_3$ the non-zero weight of the loss function for different parts, training to balance parts.

The network structure of the generator and discriminator is shown in Figure 2 below:

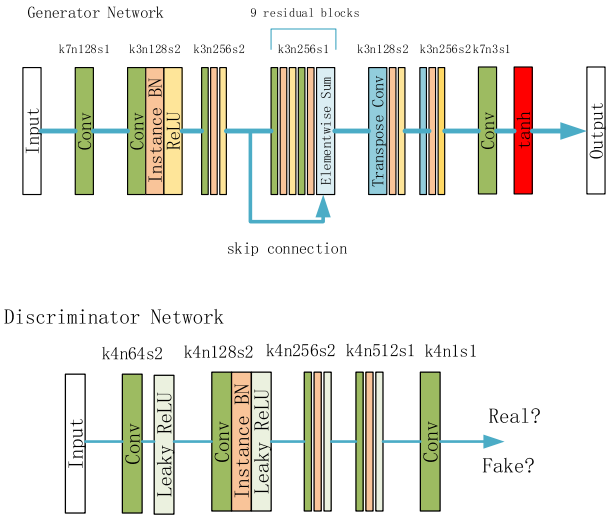

Fig.2 Generator and network structure of discriminator

Where the generator contains four convolution layers, 9 Residual Modules and 4 Transposed Convolution Layers. The discriminator network consists of five convolution layers plus LeakyReLU activation function, and the network structure of and is D_A, D_B exactly the same, it's that the number of frames entered during training is inconsistent. Additionally, instance normalization (IN)[8] is used in each layer of the generator and discriminator except the input and output layers.

To summarize, this paper presents an unsupervised learning framework, Retrospective CycleGAN, for predicting future video frames. It consists of one generator and two discriminators, during training, generators use forward and backward sequences as input and use the constraints of circular review to ensure consistency of bidirectional predictions. Additionally, this paper also introduces two discriminators for antagonistic training, frame Discriminator Used to Determi-ne the True or False of a Single Picture, sequence discriminators are used to determine whether an input sequence contains generated frames to improve the accuracy and robustness of video frame prediction in time domain, in addition, this paper verifies the superiority of the proposed method compared with the previous method through experimental comparison, and achieved the current optimal performance on the task of video prediction.

Wang[9] and others proposed a video generation model based on parsing video ap-pearance and action. The main innovation i-s to propose G3AN network structure, it h-as a THREE-STREAM generator architectur-e, mainstream coded space-time video repr-esentation, with two auxiliary streams, to re present the appearance and movement of i-ndependent generators. A self-focus mechani-sm for advanced feature maps ensures satis-factory video quality. Therefore, G3AN can generate realistic videos by following traini-ng assignments and without additional inpu-t.

Its proposed THREE-STREAM generator considers a single appearance feature (spatial flow), motion characteristics (Time Flow) and Smoothly Generated Video (Mainstream) Learning. A novel factor, space-time self-attention (FSA), is also presented. Considered the first self-attention module for video generation, the goal is to model the global space time representations and improve the quality of generated

video.

Existing video generation methods include Variational Auto Encoder (VAE), autoregression models and the most well-known method for generating antagonistic networks (GANs). Most methods contain a condition. Video Generation with additional inputs. One of the highlights of the G3AN model proposed in this paper is that it does not require any additional input.

The G3AN network architecture is shown in Figure 3. G3AN consists of three-stream generator and two-stream discriminator. Generator contains five stacked G3 modules, a decomposed self-attention (F-SA) module, and take two random noise vectors Za and Zm as input, which represent appearance and movement separately. Using two-stream discriminator architecture, it contains video stream DV and image stream DI. During training, DV accepts full video as input, DI randomly sample frames from the video.

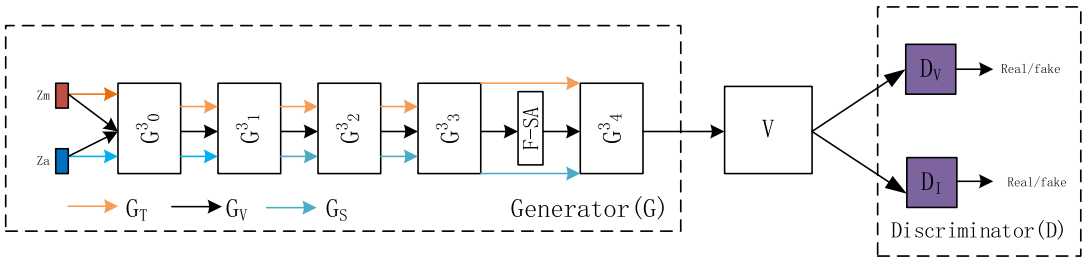

Fig.3 G3AN network structure

Its proposed F-SA module is in the ge -nerator of Figure 3. F-SA contains a tempo -ral SA (T-SA), then a space-time SA (SS A). This decomposition reduces computation -al complexity, which allows F-SA to be ap -plied on larger feature maps. According to two-stream discriminator architecture, G3AN optimizes both DV and DI. Both losses use the GAN loss function proposed by DCG-AN.

The experimental results show that the G3AN model proposed by the author performs better than other models in video generation tasks.

3. CONCLUDING REMARKS

At present, deep learning plays an increasingly important role in the field of artificial intelligence. The emergence of generation antagonism networks, because of its unique antagonistic training thought, it greatly improves the quality of video prediction generated samples and all aspects of evaluation indicators.

In the field of video prediction, improving the accuracy and prediction time of prediction video; finding ways to make the model predict more frames of video successfully are the mainstream research direction in the future. Its future development can be applied to areas that are closely related.

To people's lives, such as unmanned driving and predicting precipitation. In the future, video prediction will be applied in many fields. GAN based video prediction will be committed to improving the quality of the generated image and video, and continuously improving the performance of generator and discriminator, to achieve better performance.

ACKNOWLEDGMENT. The work is supported by NSFC (61771155) and Fundamental Research Funds for the Harbin Engi -neering University. Thanks a lot to Miss Z -heng Liying for her brilliant help to the re -search.

REFERENCE

1) GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]//Proceedings of the 2 7th Internation-al Conference on Neural Information Proce-ssing Systems, December 8-13, 2014, Mont-real, Canada. Cambridge, MA: MIT Press: 2672-2680, 2014.

2) GOODFELLOW I. NIPS 2016 tutorial: Generative adve -rsarial networks [J/OL]. ArXiv e-prints, [2021-02-22]. http s://arxiv.org/abs/1701.00160, 2016.

3) ARJOVSKY M, BOTTOU L. Towards p-rincipled meth ods for training generative ad-versarial networks [J/OL]. A rXiv e-prints, [2021-02-22]. https://arxiv.org/abs/1701.04862, 2017.

4) AGHAKHANI H, MACHIRY A, NILIZADEH S, et al. Detecting deceptive reviews using generative adversarial n etworks [C]// IEEE Security and Privacy Workshops (SP-W), 24 May, 2018, San Francisco, CA. Ne-w York: IEEE, 2018: 89-95.2018.

5) VONDRICK C, PIRSIAVASH H, TORR-ALBA A. Gen -erating videos with scene dyn-amics [J/OL]. ArXiv e-print s, [2021-02-22]. https://arxiv.org/abs/1609.02612.2016.

6) CHEN Z, WANG C, WU H, et al. DMGAN: Discrimi-native metric-based generativ-e adversarial networks [J]. K nowledge-Based Systems, 192: 105370.2020.

7) ZHU J Y, PARK T, ISOLA P, et al. Un-paired image-t o-image translation using cyc-le-consistent adversarial netw -orks [C]//2017 IEEE International Conference on Compute -r Vision (ICCV), 22-29 October, Venice, Italy. New York: IEEE, 2017: 2242-2251.2017

8) ULYANOV D, VEDALDI A, LEMPITSK-Y V. Instance normalization: The missing i-ngredient for fast stylization [J/OL]. ArXiv e-prints, [2021-02-22]. https://arxiv.org/abs/16 07.08022.2016.

9) WANG Y, BILINSKI P, BREMOND F, et al. G3AN: D isentangling appearance and motion for video generation [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13-19 June, Se-attle, WA. Ne -w York: IEEE, 2020: 5263-5272.2020.