Generative Adversarial Network in Video

Sreehari Guruprasad
Computer Science
Washington State University
Pullman WA USA
sreehari.guruprasad@wsu.edu

ABSTRACT

This article talks about Video Generative adversarial networks which is a type of generative adversarial network abbreviated as GAN. This survey paper identifies the methods and implementation of ideas from various other papers. A comprehensive review what each paper aims to achieve has been done in this literature.

CCS CONCEPTS

• GAN in Video Generation • Generator • Discriminator • DIGAN

KEYWORDS

Generative Adversarial Network, GAN, Conditional video generation, Unconditional video generation, Generator, Discriminator, INR based DIGAN

1. INTRODUCTION

Generative Adversarial Network also knows as GAN is a deep neural network framework that can learn from a collection of training data and produce new data that shares the same properties as the training data. GAN is a framework often using in generative AI for unsupervised learning.

A wide range of data kinds, including text, photos, music, and even 3D models, have been generated by GANs. They have the power to completely transform a wide range of sectors, such as manufacturing, entertainment, healthcare and many more.

1.1 GAN Architecture:

Generative Adversarial Network consists of two neural network- Generator and a discriminator.[4]

Generator learns to generate data that is close to real data. It is responsible for producing realistic fake data which serves as an input for the discriminator. Discriminator is a neural network that differentiates real data from fake data. This means that discriminator is responsible for determining how close the generated data is to the real data.

Generator and discriminator are trained together in a zero sum game. The generator tries to generate a fake data such that the discriminator cannot distinguish between real and fake data but the discriminator tries to distinguish the data and penalize the generator if there is difference between real data and fake data. The below diagram shows the architecture of the GANs.

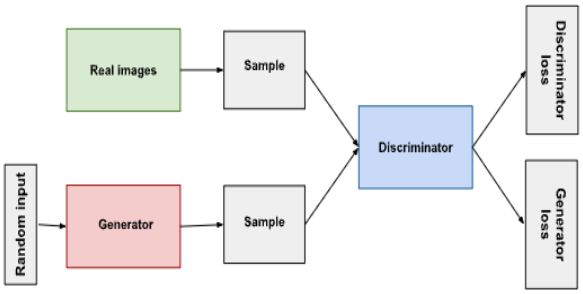

Figure 1: GAN architecture showing the working of generator and discriminator.

1.2 GAN in video:

Generative Adversarial Network can be used in the field of video generation just like mentioned in the previous section.

Generator in the GAN takes a random noise vector as input and outputs a video which is sequence of images. Discriminator on the other hand tries to distinguish between fake frames and real frames and penalizes the generator. This process iteratively improves the performance of both the neural networks.

Applications of GAN in video generation are:

1. It can used in video inpainting which editing a video like filling in missing parts of the video.
2. Style of one video can be transferred to another video.

Some of the challenges of using GAN in video generation are:

1. VGANs are computationally expensive since it needs to be trained on large amount of video data.
2. VGAN need to capture complex information from the video data due to which generating long and realistic video is a very hard problem.
3. As there are no single metric to accurately measure the realism of a video, Evaluating the quality of a video is difficult.

These challenges are addressed by the papers mentioned in the literature survey section by using new architecture and solutions.

2. LITERATURE SURVEY

This section describes briefly about all the literature papers surveyed while writing this paper.

2.1 Generating Videos with Scene Dynamics- Carl Vondrick, Hamed Pirsiavash, Antonio Torralba[1]

A generative model for videos that learns from a lot of unlabeled video data is shown in the paper. To distinguish the foreground from the background, the model makes use of a spatio-temporal convolutional architecture in a generative adversarial network. With little guidance, the model can produce videos with realistic dynamics and motions and can also learn features that are helpful for tasks involving video recognition. Applications of the model for representation learning and future prediction are also covered in the article.

Methodology: This paper discusses the use of generator network and discriminator network to generate videos. Generator network is used to produce a video whereas a discriminator network is used to distinguish synthesized video and real video.

Generator network is a low dimensional latent code. There are two different network architecture: One stream architecture and two stream architecture.

a. One Stream Architecture: A five layer 3-D convolution network of $4 \times 4 \times 4$ convolutions with a stride of 2, except for the first layer which uses $2 \times 4 \times 4$ convolutions (time × width × height) has been used in this paper.
b. Two Stream Architecture: Two stream architecture is used to generate video which has a static background and moving foreground. As shown in the figure below, two stream architecture combines two independent stream of inputs to generate a video.

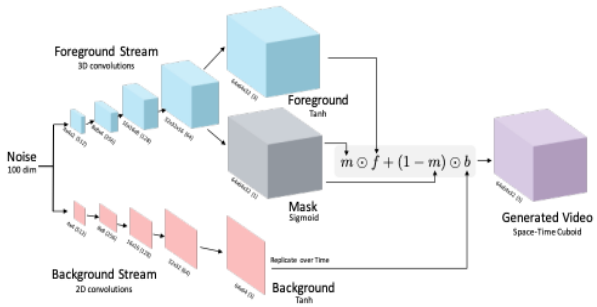

Figure 2: This figure depicts the two stream architecture of generator network

Discriminator Network is used for two tasks: distinguishing between synthesized video and a realistic scene as well as recognizing realistic scene between frames.

Implementation: The authors of the paper have experimented the generation of video using 5000 hours of filtered unlabeled videos which consisted of: Golf course, Babies, beaches and train stations. Various stabilization techniques has been incorporated to keep the background stable and capture the moving foreground.

Figure 3: Figure depicting the use of background and foreground information from independent streams and merging it to generate a video.

Even though this method incorporated in the paper successfully learns to disentangle the foreground from the background, the object of motion in the foreground is often not clear. This method was not successful in generating long video but this method shows promise in generation of short videos. This paper has shown progress in the domain of video generation by using GAN for the same instead of autoencoder.

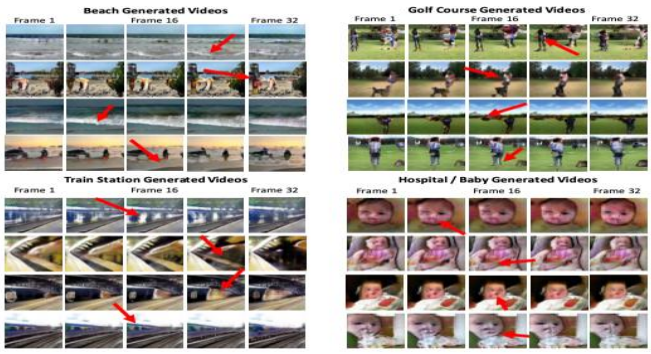

Figure 4: Example of video generation using two-stream model.

Achievement: DIGAN can be trained on 128 frame movies of 128×128 resolution, which is 80 frames longer than the 48 frames of the previous state-of-the-art approach. It also increases the previous state-of-the-art FVD score on UCF-101 by 30.7%.

2.2 Generating Videos with Dynamics-Aware Implicit Generative Adversarial Networks - Sihyun Yu , Jihoon Tack, Sangwoo Mo, Hyunsu Kim , Junho Kim , Jung-Woo Ha , Jinwoo Shin [2]

This paper proposes dynamics aware implicit generative adversarial networks which makes use of the implicit neural representations that mitigates various challenges in video generation. This is referred to as DIGAN in the paper.

Methodology: This paper uses INR based DIGAN. This means that they used image INR to generate videos by adding another dimension to it. To explain it elaborately, Image INR uses two space dimensions(x,y). Time dimension is added to this image INR to generate videos. But this alone does not result in a good video generator. Hence, they make use of generators and discriminator in conjunction with the video INR.

Generators synthesis a parameter of INR called content parameter φ in the first layer. Using the output of this first layer another parameter is incorporated called as motion parameter w_t . The

output of the INR is a continuous trajectory with respect to time. This means that the output can be seen as continuous spectrum of images over time frames say at t=0,1,2,3 etc. To increase the efficiency of video generation the following methods are incorporated:

1. The above-mentioned time frames are taken with smaller time frequency than space frequencies.
2. Sampling motion diversity vector in addition to the content latent vector.
3. Applying a nonlinear mapping on top of motion features at time t.

Discriminators: An efficient of use of discriminators is proposed in this paper where 2D convolution network is used instead of a 3D convolution network. The discriminators used in this paper distinguishes the real video from the synthesized one by comparing two arbitrary frames created by video INR instead of the whole sequence. This addresses the computational issues of a conventional discriminator. The below picture depicts the architecture of generator and discriminator in the paper.

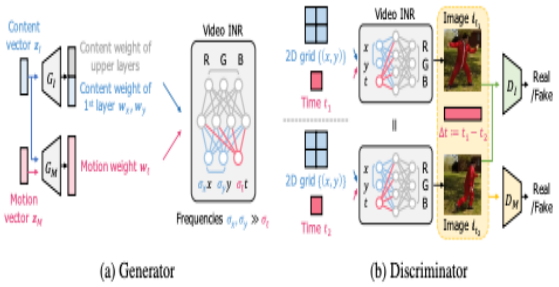

Figure 5: Depicts the generator and discriminator architecture used in the DIGAN. This architecture consists of two discriminators that distinguishes between real and fake images and motions.

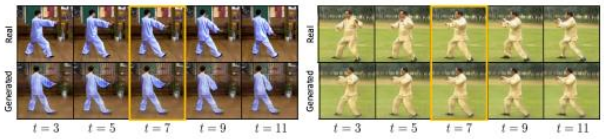

Figure 6: Examples of forward and backward predicting using DIGAN.

Achievements: The paper achieves to excel at the following applications:

1. Since the DIGAN uses INR which is not computationally expensive, it can be used in successfully generating long videos. 128 frame videos have been successfully generated as per mentioned in this literature.
2. Time and space interpolation and extrapolation
3. DIGAN can provide samples of any duration, allowing it to concurrently calculate the full video at once or deduce previous or interim frames from subsequent frames.

2.3 Video Generative Adversarial Networks: A review- Nuha Aldausari, Arcot Sowmya, Nadine Marcus, Gelareh Mohammadi [3]

This paper provides a comprehensive and systematic review of various GANs frameworks in the video domain. The two main types of GAN discussed in this paper is condition GAN and Unconditional GAN:

1. Unconditional GAN: Unconditional GAN means generating video without any prior information.
   a. It is more difficult to produce movies without prior knowledge since the model must capture the data distribution on its own, without assistance from the input signal that can aid in reducing the target space. Even though it can be challenging to train unconditional video GANs, some of the unconditional models have been used as the basis for conditional frameworks.

b. VGAN first GAN to generate videos. This makes use of 2 convolution neural network. 3D network to detect foreground objects and the other 2D CNN to detect static background.
c. FTGAN encounters flaws of VGAN in detecting foreground by adding optical flow with the help of progressive architecture. 2GAN's are used in the FTGAN where one GAN model generates the flow whereas the other generates the texture.
d. Motion Content GAN: Another type of Unconditional GAN that consists of a N to N RNN. It uses a combination N content vectors and N motion variables. Each combination is used to generate N images which is synthesized to generate a video.
e. Temporal Generative Adversarial Nets (TGAN): Difference between the MoCoGAN and TGAN is that the latter uses a N 2D image deconvolutional generators to produce N frames whereas the first one uses the RNN structure.
f. Dual video discriminator GAN: These GAN have the capabilities to produce 48 high quality images based on the complex datasets which can be passed to two discriminators to generate videos.

2. Conditional video generation: A conditional signal in GANs is used in a number of works to govern the processing and control modes of the generated data; An audio signal, text, semantic map, picture, or video could be the condition. There are many conditional GAN, some of the examples are:
   a. Speech to Video Synthesis: Some of the techniques used in the speech to video synthesis are:
      i. Disentangled Audio-Visual System (DAVS): In this technique the audio is decoupling the information of a person's face and the speech related information in order to fine tune the lip sync
      ii. Variational AutoEncoder(VAE): The audio is decoupled for 3 information: content, emotion and noise. A frame level discriminator and a video level discriminator collectively produce the video.
   b. Text to Video Synthesis: Temporal GAN conditioning on captions architecture is used in an LSTM based encoder to generate videos using the text-based inputs. the model comprises three discriminators: one for the video level, one for the frame level, and one for the motion level.
   c. Image to Video Synthesis: The main purpose of this task is to predict future frames based on a given frame. To extract content from motion, some models employ 3D convolutional networks; others use recurrent networks or dual networks.
   d. Video to Video Synthesis: Object animation, or retargeting motion from one object to another, is one of the main uses for this task. Certain models direct the creation process by motion strokes, optical flow, or pose estimation. Certain models can also use low-dimensional inputs, like joystick movements, to manipulate the human subject in the generated frames.

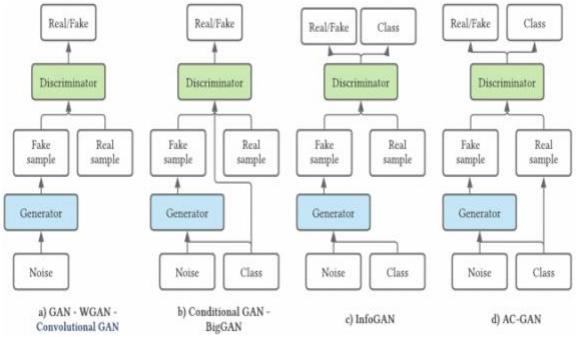

Figure 7: Architecture of various types of GAN

In conclusion, this paper discusses the possible uses and problems with the GAN in the video domain. Some of the problems mentioned in the paper are data complexity, temporal coherence, semantic consistency, evaluation metrics and ethical issues. Some future work suggested are exploiting temporal information, improving diversity and realism and addressing ethical concerns.

CONCLUSION

This paper aims to give an overview of recent advancements in the domain of video generation by using generative adversarial networks. Use of Generative adversarial networks for video generation have proven beneficial in comparison to autoencoder. There were initially many challenges to use GAN for video, but innovative methods and approaches have helped overcome these challenges. An effort to perform a comprehensive and systematic study of these innovative solution has been made in this paper by surveying multiple paper and outlining the methods implemented.

ACKNOWLEDGMENTS

I would like to thank Dr Assefaw Gebremedhin for giving me this opportunity and guiding me throughout the course with all the resources.

REFERENCES

[1] Vondrick, C., Pirsiavash, H., & Torralba, A. (2016, October 26). *Generating videos with scene dynamics*. arXiv.org. https://arxiv.org/abs/1609.02612

[2] Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.-W., & Shin, J. (2022, February 21). *Generating videos with dynamics-aware implicit generative adversarial networks*. arXiv.org. https://arxiv.org/abs/2202.10571

[3] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. 2022. Video Generative Adversarial Networks: A Review. *ACM Comput. Surv.* 55, 2, Article 30 (January 2022), 25 pages.

[4] Wood, T. (2020, July 22). *Generative Adversarial Network*. DeepAI. https://deepai.org/machine-learning-glossary-and-terms/generative-adversarial-network