Video Generation with intended motion using GAN

JiHong Park, SangHun Jung, SeungYoon Kim
2012210046, 2013210017, 2013210023
Department of Computer Science
Korea University
parkjh824@korea.ac.kr, wjdtkdgjs888@korea.ac.kr, dskym@korea.ac.kr

Abstract

Convolutional neural network(CNN) and Generative adversarial neural network(GAN) have recently shown outstanding performance in various fields. In particular, GAN has shown good results in the field of image and it is expanding its fields into voice, text, and video. Our goal is to generate a video with intended action, for a given behavioral condition. To achieve this, we first used Video GAN which generates video, among several GAN models. In addition, we add conditional GAN model to generate conditional video, and construct our new GAN model to allow VGAN to generate action corresponding to a given condition. Then Our new model, called Conditional Video GAN, generates Video for specific condition. Although there is no standard metric to measure video generation quantitatively, but video is generated that can distinguish some behavior by our model. Using our model, it is expected to be able to apply to short video production and simple emoticon generation.

1 Introduction

1.1 Overview

Recently one of the most active area of research is about GAN models. GAN is composed of Generator which generates images as similar as possible to original image, and Discriminator which distinguish generated image and original image as possible. GAN model is trained by competing each other. In early days, GAN outperforms only for generating images like transforming images and generating high resolution images. However in recent, GAN model is applied to various fields like text, voice, and video. Using Video GAN model published by MIT, our object is generating video with intended text with specific motion like boxing, running, and walking. To make Video GAN generate video with condition, we combined Video GAN with Conditional GAN which concatenate latent code Z and vector of condition in generator and discriminator. The data set used for experiment is videos of people with 6 actions. Our GAN model receives the condition of desired behavior among the six actions, then it generates the video that performs the action corresponding to the condition. This report is structured as follows. First, we describe the VGAN, generates Video, and introduce the Conditional GAN that gives the specific condition to the GAN model. Then, we propose and explain the conditional Video GAN we used to produce the expected results. Finally, we show the generated videos, compare it to originals, and present the conclusion.

1.2 Related Work

To make our model for generating video with motion text, we combined two GANs, Video GAN of Generating Videos with dynamic scene[1] and Conditional GAN of Conditional Generative Adversarial Nets[2].

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

1.2.1 Video GAN Model

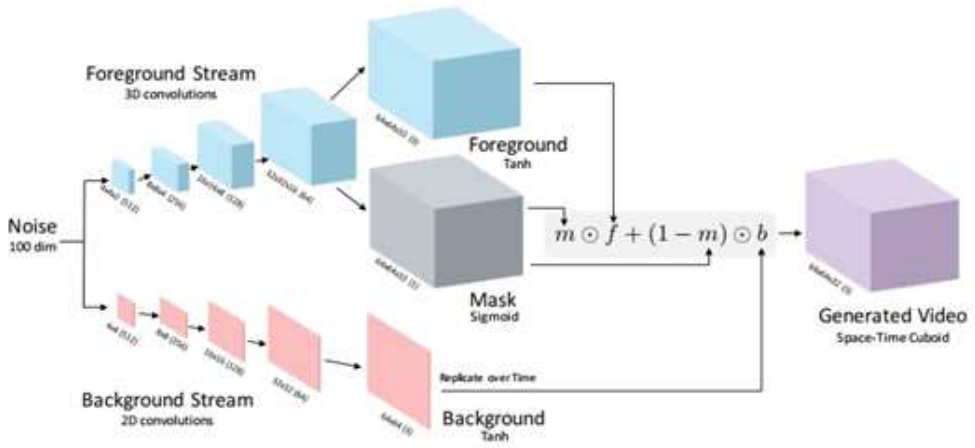

Figure 1: Video GAN Generator

To Generate Video, we used 32 images of 64 * 64 pixel and 3 channels per video. For Generator, we implemented with two main streams, Foreground stream and Background stream. Given latent code Z, Foreground stream generates object for video and mask values for deciding which one to use between background and object per pixel. Foreground network is composed of 3D de-Convolution layer, Batch-Norm layer, and non-linear unit(ReLU) and after splitted, Foreground generates object using 3D de-Convolution and tanh unit, and Mask generates mask using 3D de-Convolution and sigmoid unit. Background stream generates background image(2D) for video and stream is composed of 2D de-Convolutional layer, Batch-Norm, and non-linear unit(ReLU) and at last layer, we used tanh unit for generating image.

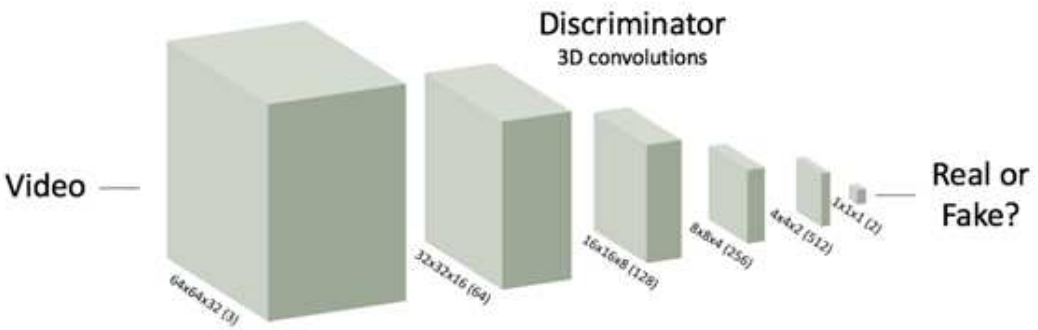

Figure 2: Video GAN Discriminator

For Discriminator, we used 3D Convolutions for distinguishing Generated Video and Real Video. Discriminator is composed of 3D Convolution layers, Batch-Norm layer, and non linear unit(Leaky ReLU).

1.2.2 Conditional GAN Model

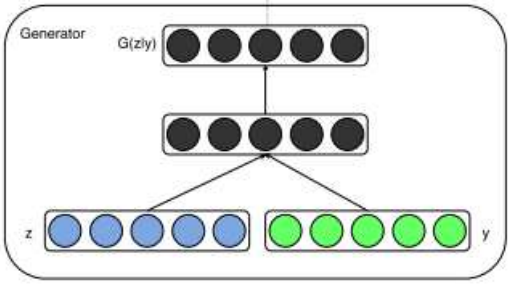

Figure 3: Conditional GAN Generator

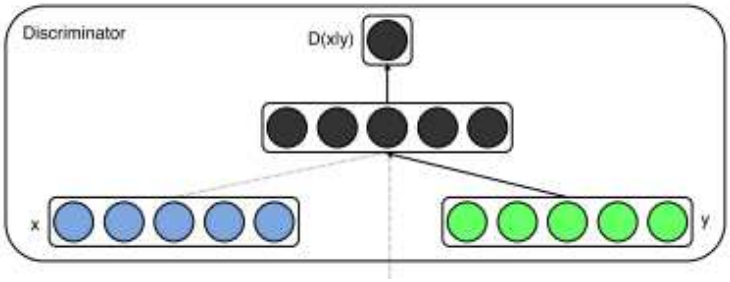

Figure 4: Conditional GAN Discriminator

To Generate intended Video, we used Conditional GAN. Instead of latent code Z, CGAN concatenated Z and 1-hot Vector for Generator input. For discriminator, 1-hot vector is concatenated with real-video value.

2 Model

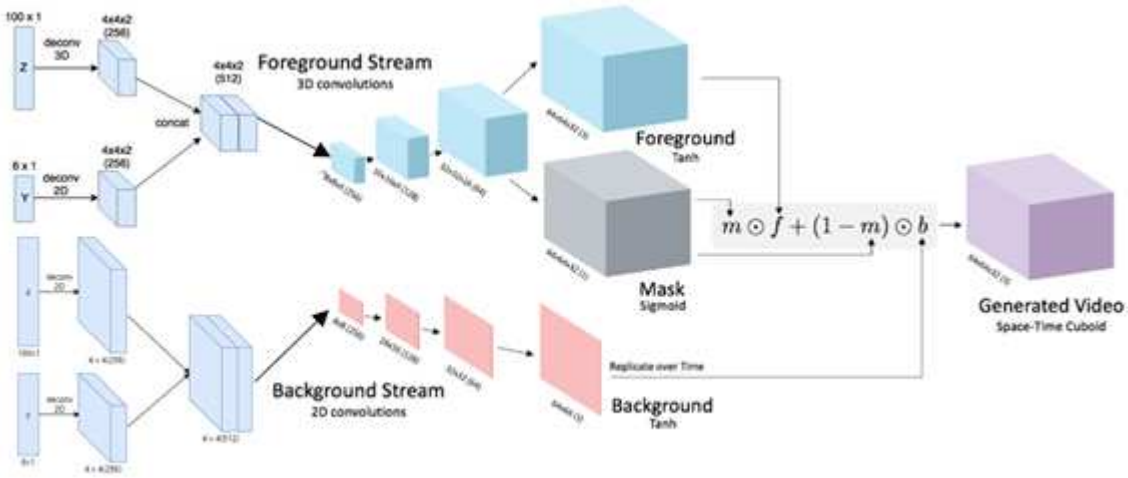

Figure 5: Conditional Video GAN Generator

Our model is combined version of Video GAN and Conditional GAN. To concatenate latent code Z and y vector, we used 3D de-Convolutions and 2D de-Convolutions to make both 4*4*2(256)

dimension and 4*4(256) dimension and entered two tensor into Foreground stream and Background stream each. Following process is same as VGAN.

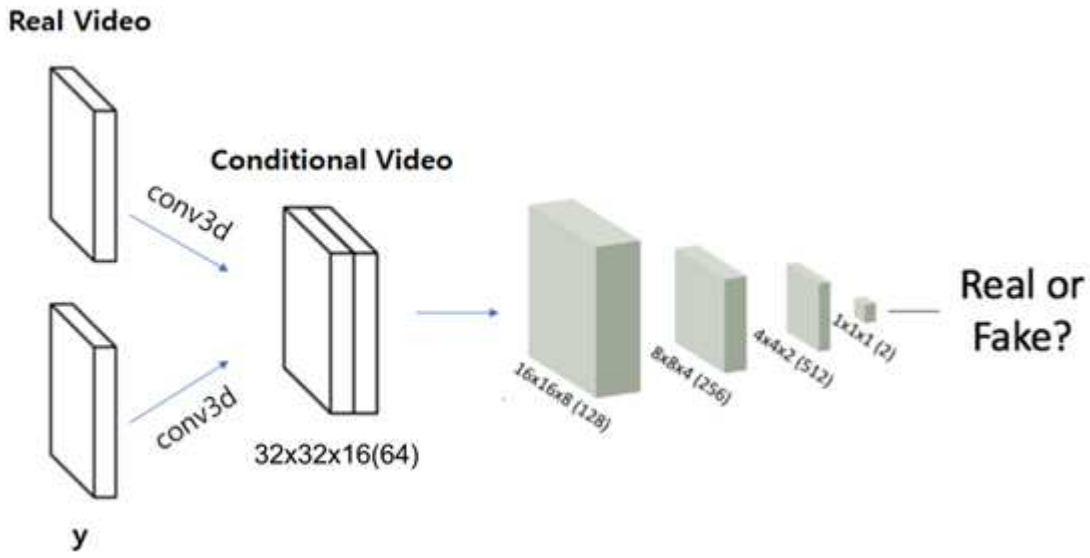

Figure 6: Conditional Video GAN Discrimiator

For Discriminator, to concatenate video and y vector, we used 3D Convolutions to make 32*32*16(32) each. After concatenation process is same as VGAN. We used Adam optimizer, Binary Cross Entropy for Loss Function and set learning rate and batch size as 0.0002 and 64. We run our model for 800 epochs.

3 Experiments

3.1 Data Set

Figure 7: Data Set

We needed video dataset with labeled actions. We use Human Activity Video Datasets containing six types of human actions(boxing, hand clapping, hand waving, jogging, walking and running). We have 3093 videos and split them into 32 frames each video. The spatial resolution of our data set is 160 x 120 pixels, but we transformed it into 64 x 64 pixels. We represented human actions to one hot vector.

4 Result

Figure 8: Boxing, HandClapping, HandWaving Result

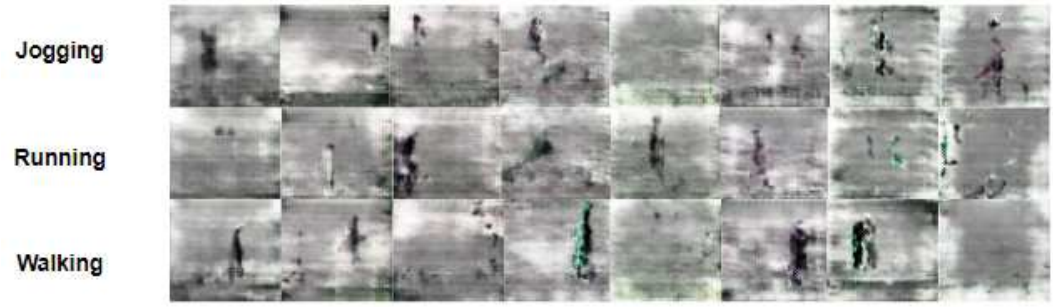

Figure 9: Jogging, Running, Walking Result

We show real videos and fake videos which our model generate. We show fake videos which our model generate every 100 epoch. If we combined one hot vector to our model(walking + boxing), we expected to make video which represents person who is boxing and walking same time. But our result represents one person who is boxing and the other person who is walking separately.

5 Conclusion

We add a condition to the Video GAN model in the existing paper and make Conditional Video GAN model which generate video we want to behave. Current model does not generate good result due to a lack of data set, learning iteration and stabilization and the number of actions which our model generate is very low. However, if you have a large number of video containing many actions later on, We think that there will be many applications, such as generating emoticons or simple video when you need.

6 Future Work

To generate perfect videos we want, we should find more dataset and revise our model. We found Stack GAN model which revised conditional GAN's drawbacks, we apply this model to ours to make object act many motions. And we do stabilization on frames as data preprocessing.

References

[1] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016. 1.2

[2] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1.2