Creating Videos conditioned on Music using Generative Networks

Jacinth David
jacinthdavid@cs.umass.edu

Abstract

In this project, I have attempted to create music video using Generative models. In particular, I have put together conditional GANs, TGAN and wavenet to generate videos. To avoid the training issues with GAN, I have extracted a simplified dataset of videos with music to test the network

1. Introduction

Music videos accompany a piece of music and are an art-form on their own right. Well created videos match their songs in many respects but I have begun by trying to teach the network rhythm or timing. I also aim for my network to be used by musicians. I hope they will be able to create entertaining videos for the promotion of their songs.

1.1. Framework

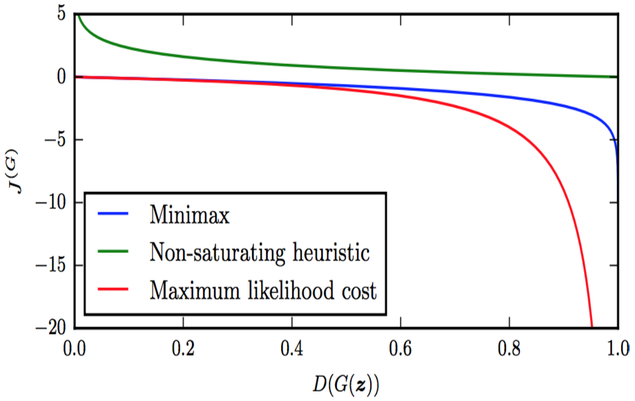
I have used conditional GAN[1]. The authors used the conditional GAN to caption images. Specifically, the netowrk generated captions conditioned on images. I replaced the LSTM for captioning with TGAN[3] to create videos. TGAN generates video tensors, which is differentiable, hence, I did not have to use policy gradients like in the original paper to train with non-differentiable captions. I also replaced the convolutional neural network used for image embeddings with wavenet[4]. This network generated music embeddings for me.

2. Background

Generative adversarial networks generate samples that seem like they come from the true distribution [2]. They achieve this by reaching the approximate nash equilibrium of a two-player game. The generator creates fake samples, while the discriminator learns to classify them as fake by comparing them to real examples from the training dataset. The minimax loss function is shown as follows.

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{x\sim pdata(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$

GAN's, when training converges, generate realistic samples as compared to other generative models, like autoencoders. However, in this formulation, the gradient is saturated for the generator in the beginning as shown in figure



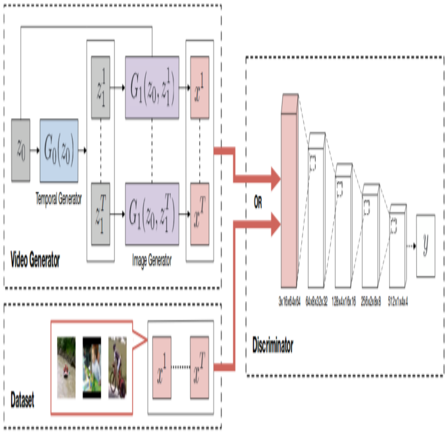
GANS also face the problem of mode-dropping. Furthermore, because of the minimax objective, training converges to a saddle point rather than local minima. To alleviate these problems, many improvements have been proposed. One such improvement is the wasserstein GAN, which minimizes the earth mover's distance instead of the Jensen-shannon divergence.

3. Approach

I extracted a videos dataset, where a metronome ticks along to some tempo. This was obtained from youtube. After that, I used ffmpeg to crop the video to 12 fps, 96*128 pixels, for 3 seconds each. Once extracted, I converted to numpy tensors of (36,96,128,3) dimension using the skvideo module of python. For music, I used the pre-trained model (almost 1 GB) trained on the nsynth dataset provided by deepmind.

As an aside, it was not easy to find it at all. I actually found it by accident while searching for the solution to some other bug. The pre-trained model does not show up on google and I wasted almost 2 days trying to train wavenet.

The music embedding is concatenated with the latent 100*100 dimension code. This is randomly sampled and TGAN converts it into video. The generated video is fed into the discriminator network along with the minibatch sample from the dataset. Finally, i take the dot product of the output of the discriminator and the music embedding. This dot product, after being passed through a logistic transform becomes the objective function for the network. The generator and discriminator are similar to the TGAN architecture, which is implemented in the chainer framework.



4. Experiment

5. Experiment

6. Conclusion

I learnt about GANs and generative models through this project and I was get the setup necessary to train the model

6.1. Possible Future work

A related problem is setting music to videos. It’s applications can be for promotional videos or highlight videos, in sports and other fields. I intend to explore this problem in future work.

References

[1] B. Dai et al. Towards Diverse and Natural Image Descriptions via a Conditional GAN.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets, 2014.

[3] Masaki Saito et al. Temporal Generative Adversarial Nets with Singular Value Clipping.

[4] Oord et al. WAVENET: A GENERATIVE MODEL FOR RAW AUDIO.