Deep Learning

Collin Abidi
Jarod Vickers
Justin Long
Nick Buck

Dance Learning

January 30, 2020

WHAT DO WE PROPOSE TO DO

Inspired by the “Stick Figure Story Generation” idea in the Project Proposal Examples document, we plan on training a deep network using a Generative Adversarial Network (GAN) to create a unique stick figure dance sequence to any audio clip.

WHAT HAVE OTHERS ATTEMPTED

To the best of our knowledge, there are no other attempts at using GANs to generate synthetic dance sequence. However, there are many examples of attempts at generating other synthetic behavior that we will take inspiration from. The best example of a similar approach is the Speech2Gesture model.

The Speech2Gesture model was trained to produce sequential data that learns the conversational style of talk show hosts given an audio signal of the host talking. We plan on potentially using transfer learning to model this approach on dance sequences.

Sequential GANs are an interesting extension to the typical GAN structure. SeqGANs extend the usual GAN approach to the spatio-temporal domain.

GANs have also been utilized to create illustrations of humans from stick figure poses. The model was trained using regular images and stick figure poses created by the developers. The model then took stick figure poses as an input, and successfully generated images of characters in said poses. Although this example does not include motion or dancing, it does show that stick figures can be incorporated into a GAN architecture.

WHY OUR PROPOSAL IS INTERESTING

We will be attempting a novel approach that fuses work on GANs with deep learning methods that convert audio to conversational gestures with application to a whole new domain; dancing.

WHY IS IT CHALLENGING

This project is challenging because it both extends the Conversational Gesture work to the domain of dancing and will have to use GANs with sequencing (as generated videos will be in the spatio-temporal domain).

WHY IS IT IMPORTANT

This project is important because it not only proves the functionality of GANs but showcases its ability in an interesting way. If we can create dances that make sense to the human eye while also ‘fooling’ another network we can add work that proves the power of GANS.

WHAT DATA DO WE PLAN TO USE

We plan on using the following datasets:

1. **Speech2Gesture** (http://people.eecs.berkeley.edu/~shiry/speech2gesture/)
2. **Let’s Dance** (https://www.cc.gatech.edu/cpl/projects/dance/)
3. **YouTube-8M** (https://research.google.com/youtube8m/)

HIGH-LEVEL IDEA

We will use a Generative Adversarial Network to train a deep neural network to create highly-realistic sequences of stick figures dancing. We will follow the typical GAN approach of generator vs. discriminator, but will utilize more state-of-the-art sequence GANs that can extend GANs to the temporal dimension.

We will train the GAN using stick figure sequences from the **Let’s Dance** dataset as the “true” data (sequences in the dataset have ground-truth stick figures along with original frames and optical flow).

HOW IS OUR METHOD NOVEL

Typical GANs simply use a random noise distribution as input to the Generative network. However, since we want to use audio signal as an input, we will be developing a method to incorporate the audio signal as a part of the training procedure.

There are many examples of GANs that synthesize fake audio as an output, but, to the best of our (limited) knowledge, there are none that use audio signal as an input to generate spatio-temporal data.

HOW WE WILL EVALUATE USING METRICS AND BASELINES

The metric/evaluation technique is built-in to the GAN. The adversarial part of the GAN will be the evaluator; once the adversarial network cannot discriminate between fake and real data, we will know that our generative network has converged.

CONSERVATIVE MILESTONE SCHEDULE

Early February: Datasets properly formatted for transfer learning from Speech2Gesture model to get familiar with data and methods.

Late February: Classical GAN example completed and understood.

Early April: Sequential GAN model completed for usage with generating synthetic dances without audio signal input.

Mid April: Sequential GAN implemented without audio signal as an input to the system. Model able to train fully and generate synthetic dance sequences.

AMBITIOUS MILESTONE SCHEDULE

Early February: Datasets properly formatted for transfer learning from Speech2Gesture model to get familiar with data and methods.

Late February: Sequential GAN model completed for usage with generating synthetic dances without audio signal input.

Late March: Sequential GAN model completed for usage with generating synthetic dances without audio signal input.

Mid April: Conditional Sequential GAN implemented with audio signal as an input to the system. Model able to train fully and generate synthetic dance sequences given any audio signal.

References

[1] Let's Dance https://www.cc.gatech.edu/cpl/projects/dance/

[2] Speech2Gesture: Learning Individual Styles of Conversational Gesture

http://people.eecs.berkeley.edu/~shiry/projects/speech2gesture/

[3] YouTube-8M https://research.google.com/youtube8m/

[4] SeqGAN https://medium.com/prathena/seqgan-gans-for-sequence-generation-2099a85baed0

[5] Conditional GANs from Scratch
https://machinelearningmastery.com/how-to-develop-a-conditional-generative-adversarial-network-from-scratch/

[6] Conditional GANs https://arxiv.org/abs/1411.1784

[7] Understanding GANs
https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29

[8] Generation of Character Illustrations from Stick Figures Using a Modification of Generative Adversarial Network https://ieeexplore.ieee.org/document/8377853