Lip movement Synthesis from Text

By - Shishir Mathur
Guided By - Prof. Vinay P Namboodiri
Generative Adversarial Networks (GANs)

GANs were first introduced by a group of researchers at the University of Montreal lead by Ian Goodfellow.

The main idea behind a GAN is to have two competing neural network models. One takes noise as input and generates samples (and so is called the generator). The other model (called the discriminator) receives samples from both the generator and the training data, and has to be able to distinguish between the two sources.
These two networks play a continuous game, where the generator is learning to produce more and more realistic samples, and the discriminator is learning to get better and better at distinguishing generated data from real data.

Why are GANs interesting

GANs have been primarily applied to modelling natural images.
They can understand the underlying image semantic and can help in inpainting tasks.

Source: Semantic Image Inpainting with Perceptual and Contextual Losses
Using a contextual Recurrent Neural Network over a GAN it can be trained to generate the next image in a sequence of images

Source: Contextual RNN-GANs for Abstract Reasoning Diagram Generation
They can also be trained for Super Resolution to get photorealistic images

AND..... Many Many More.......
VideoGAN

Based on Generating Videos with Scene Dynamics by Carl Vondrik et.al.

This approach uses the basic GAN framework on video samples by upsampling the noise using volumetric convolutions to generate video samples rather than just images.

Source: http://carlvondrick.com/tinyvideo/
Generative Adversarial Text to Image Synthesis

Based on the paper by Scott Reed

It generates realistic images based on text using a text embedding in the generator and discriminator to ensure conditional generation.
<table>
<thead>
<tr>
<th>GT</th>
<th>GAN</th>
<th>GAN - CLS</th>
<th>GAN - INT</th>
<th>GAN - INT - CLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>an all black bird with a distinct thick, rounded bill.</td>
<td>this small bird has a yellow breast, brown crown, and black superciliary</td>
<td>a tiny bird, with a tiny beak, tarsus and feet, a blue crown, blue coverts, and black cheek patch</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Aim of the work

The aim of my work is to generate lip movement given a text.

There has been a lot of recent work in Lip Reading like Lip Reading in the Wild by Chung et.al., LipNet by Yannis et.al.(Oxford). But the reverse was not very well explored using the existing Neural Network Techniques.
I am working on Video Generation by VideoGANs conditioned on the text data to generate lip movement corresponding to the word being said.

Using the VideoGAN framework I am conditioning it to the text embedding as done by Scott Reed in his paper Text to Image Synthesis.
The present dataset I am using for the task is The GRID audiovisual sentence corpus.

The corpus consists of high-quality audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers.

Advantages of GRID Corpus

- Good Quality video with a high frame rate to extract lip features.
- Word captioning accurate to the millisecond.

Source: http://spandh.dcs.shef.ac.uk/gridcorpus/
Preprocessing Methodology

I am generating just the lip region through the GANs. Since all datasets have different faces the only thing that remains invariant across different video samples is the lip movement for word utterances.

Using DLib face keypoints extraction we were able to extract the lip regions in all the video frames.

The video frames were then paired with the word extractions to create 32 frames long videos of lip movement.
Network Structure

Used a basic DCGAN structure with volumetric convolutions for video generation.

Used Skip Connections in both Generator and Discriminator for better visualizations and better Discriminator learning.

Text encoding is appended to the noise vector in Generator and replicated and appended at a pre final layer in the discriminator.

I have added a Dropout layer before every convolution and deconvolution for better learning.
Training Methodology

Experimented with different training methods for GANs like WGANs, Conditional GANs etc. Finally decided upon the training methodology similar to Scott Reed.

- Train the Discriminator with the real video and the corresponding text embedding. (Label TRUE)
- Train the Discriminator with the fake video and the same text embedding. (Label FALSE)
- Train the Discriminator with the generated video and the corresponding text embedding. (Label FALSE)
This format of training allows for a more conditional training of GANs by training the discriminator to not only detect the fake video sample but also tie the video sample with the text embeddings and also detect if the video and the embedding are not tied.

But here since the discriminator learns more than the generator in each timestep over fitting of the discriminator is possible which leads to bad generator.

To avoid that we have scaled down the Discriminator error and added multiple Dropout layers.
Results
Future Work

Working on multiple ways of embedding the text data for better generation.

Tweaking the Network changing Batch-Norm Layers and Dropout layers for better video generation.

Stabilizing the generator and discriminator for better sample generation.
Some Advice if you work on GANs

- Try to stick to the general network structure, we can increase the number of layers or the number of kernels or adding skip connections but changing it too much from the basic DCGAN structure leads to worse results.
- Normalize inputs
- Use Spherical Noise input
- Use Batch Norm, LeakyReLU, DropOut Layers with ADAM optimizer.
- Use soft and noisy labels for training.
- Don’t try to over train the Discriminator or Generator.
- Keep looking up at new GAN implementations for inspirations
Thank You!!