Project Report
TellGAN: Motion Prediction and Video Generation From Natural Text

Gunhee Lee, Jacob Morton
20172811, 20172327
Computer Science and Engineering, POSTECH
victorleee@postech.ac.kr

1. Abstract

We propose a Tell Generative Adversarial Network (TellGAN) which, given an initial starting frame with landmarks and text that describes an action, will generate plausible video frames of the given target performing the action. We call it TellGAN, because we tell the target what to do. Our method is a generic end-to-end approach that predicts motion and generates video frames. We divide the task into two parts, landmark trajectory prediction from a given word and generating a video frames from the predicted landmarks and initial target position. We also propose a novel use of an LSTM to predict motion over a various number of frames, to model motions at different speeds. We chose mouth motion prediction for a given word as an application for our model. Although this application is related to lip reading (video-to-text), a well researched area, it appears that text-to-video face motion has not been approached directly before. Among the several datasets used for lip reading, we used the GRID Corpus[1].

2. Introduction

Motion Prediction is a well researched area with several existing approaches that predict landmarks using RNN and LSTM networks. Although these networks are able to predict motions given an action, they currently can only learn and predict actions at a constant rate of change, basically learning an average representation. This is a problem in real life, when actions are performed at different rates by different individuals and times. Our model can learn and predict motions performed at various speeds. To our knowledge this problem hasn't been tackled yet. We are able to train an LSTM motion predictor to learn an 'oversampled' representation of the action learned at a higher rate than available from the ground truth. We use a weak loss for predictions with missing ground truths and rely on the dataset to provide various sequence length actions to correctly train the network. We chose a lib-reading dataset to provide a large number of actions with varying rates of speed to train and test our model on text-driven animation.

Text-Driven animation is not a new topic, but the current approaches use text-to-speech generators to drive animation and are not typically generic. There has been some recent independent techniques developed concurrently with ours that use Generative Adversarial Networks (GAN) for generating video frames, but again use audio as a driving factor. Text-to-speech techniques offer several advantages as there are plenty of previous works and is an easier task to map continuous wave forms rather than discrete words to the motions of the mouth during speech. Our proposed method will skip the text-to-speech generation by training a GAN and an LSTM with the aid of lip-reading techniques. Visual only lip reading is a heavily research topic and is essentially a video-to-text and action classification problem. Our proposed approach is the inverse to lip reading, a Text-to-Video and action generation problem. Generating compelling speech motions from words is a challenging problem. Humans are good at recognizing fine facial movements and can intemperate words from mouth motions. Also most motion prediction techniques don't demonstrate predictions of the fine motions required to model speech.

Generative adversarial networks have improved in recent years. There also has been research which relates generative adversarial network and natural language processing. Text-to-image generation is one of the most obvious examples of it. Using a GAN to Generate video or images from changing landmarks is currently an active area of research. We use a UNet architecture that takes as input an initial frame and desired landmark positions to change the image to the desired state. The results are of a high degree of realism, where the line between a fake image and real image is blurred.

Compared with the normal image generation model, our model receives the input image. The contemporary state of the art network called AttentionGAN (Tao Xu et al.) can only change the image as a whole to meet the sentence direction, which is not desirable. For example, the minor

[1] http://spandh.dcs.shef.ac.uk/gridcorpus/

changes that occure between frames would result in a completely different image. Our model would make the image more controllable than the previous approach. Also our model has a temporal component, as we propose generating a series of images that correspond spoken motions derived from the given input to make a video sequence. This has several challenges, among which are maintaining temporal coherency and preventing errors cascading through future generated frames. Previous work on text-to-video using GAN exists[2]. We will not use this method, as we will not have an end frame nor do we wish to use simultaneous generation of frame sequences. Instead we will opt for simplicity with a sequential approach that divides the task into separate steps, motion prediciton from action and image generation from landmarks.

3. Dataset

Since our approach is related to lip-reading, we can use many of the available datasets. We will choose to use the GRID Coprus. The dataset contains 33 individuals each speaking 1000 sentences. Each sentence is composed of 6 words. Video, audio, and timestamped transcripts (word level) are included. The recorded subjects are static in position, uniformly lit, with no expression. Instead of using a large vocabulary range (51 words), GRID tries cover more phonemes instead[3]. The smallest unit of speach is the phoneme. A word can easily converted into a set of phonemes. A viseme is the visual component of the phoneme, where there is a many to one phoneme-to-viseme mapping. The english language is comprised of 45 phonemes which are mapped to 17 visemes according to the Fisher phoneme-to-viseme mapping[4].

4. Network Architecture

Our network is shown in Figure 1. The basic network architecture is a combination of two different networks. This network is a combination of Geometry-Contrastive GAN network(Qiao et al, arXiv, 2018) and Segmentation from Natural Language Expressions network(Hu et al, ECCV, 2016).

For training, the natural sentence would be made directly by the label from paired data. The form of the sentence given to change the input image is limited. Something like Change the color of the person to yellow or Rotate the head 40 degrees clockwise would be the most obvious example of the sentence.

The natural language is tokenized and changed into a word vector through a word embedding matrix. Then we use Long-Short Term Memory (LSTM) network to encode sequence data. At the end of the text sequence, we use the hidden state as the final encoded vector representation of the sentence.

On the other hand, the input image goes into the encoder part of the generator. This module can be a VGG type, ResNet type, or any other network. After it passes through the encoder part, the output is now turned into a high-level abstract feature. This feature is merged with the encoded expression from the natural language. Similar to the merging mechanism proposed in Hu et al, we concatenate the final encoded vector representation to the abstract feature at each location.

After the visual and language information is merged, the generator now up-samples the image. The upsampling network is not fixed yet. The decoder part will make the fake image from the input image.

Now, The discriminator network discriminates the generator's fake images from the real image. The real image contains the target pose and color values. So that discriminator network is only given the image as an input. The discriminator computes the adversarial loss. Here we added another loss, the identity loss. The fake image should look almost like a real image which has the correct pose or color information.

There are more things to improve with this network. This network can be improved using the region proposal network (RPN). This would select what and where to be changed. Affine transformation in the feature level could be added to the generator. This transformation module could help this image to be rotated (roll), scaled, and translated. The modification power would improve further if we improve this network.

References

[2] Video Generation From Text (Li, 2017)

[3] An audio-visual corpus for speech perception and automatic speech recognition (Cooke, 2006)

[4] Confusions among visually perceived consonants (Fisher, 1968)

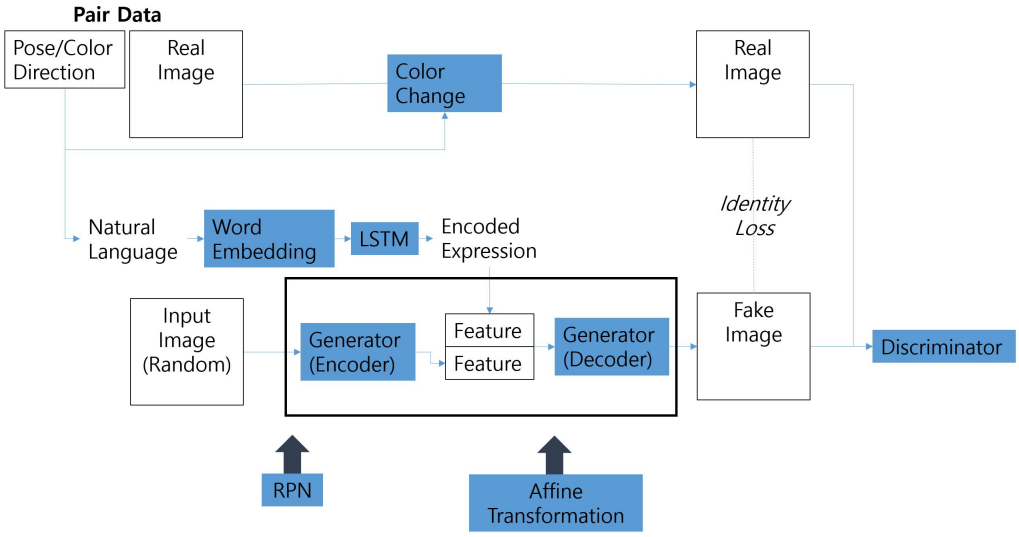

Figure 1. The whole flow of our network.