Recursive Stacked Adversarial Network for Conditional Video Generation

Shujon Naha
Department of Computer Science
Indiana University
Bloomington, IN
snaha@iu.edu

Khandokar Nayem
Department of Computer Science
Indiana University
Bloomington, IN
knayem@iu.edu

Md. Lisul
Department of Computer Science
Indiana University
Bloomington, IN
islammdl@indiana.edu

Abstract

Generating video frames based on a pre-condition is a challenging problem and requires understanding of per frame contents and visual dynamics and their relevacies to the pre-condition. In this project, we propose a novel recurrent stacked adversarial network based model to generate video frames based on a given pre-condition. In our knowledge, this is the first work to address the problem of conditional video generation using adversarial network.

1 Introduction

Generative adversarial networks have been shown incredible results to produce images from pre-conditions such as text, attributes etc [1,4]. In these works, random noises incorporated with a semantic vector representation of the prec-condition is given as input to the generator network to produce images relevant to the pre-condition. The discriminator network then learns to distinguish between the images generated by the generator and the real image from the database. A min-max learning algorithm is used to train these both models where the generator tries to continously fool the discriminator by producing better images similar to the original one and the discriminator learns to make the job harder for the generative network by getting better at distinguishing real and fake images. Most of the times, the generative network is a convolutional neural network which produces image from a single vector by using several deconvolution steps. The discriminator network is also a convolutional neural network which takes the image from the generator network output and the corresponding original image from the database and tells the similarity between the generated image and the real image.

Generative adversarial networks have been also used for predicting future frames from a video sequence and generate videos with scene dynamics [5,6]. In this works, multiple frames are combined together and 3D convolution is used in the domains of space and time to predict the next frames. These works have shown the capability of adversarial networks to capture the scene dynamics although for small temporal intervals.

In this project, we address the problem of generating videos based on pre-conditions such as action classes, fMRI signals and sentence descriptions using adversarial network. Generating videos based on pre-conditions pose a unique set of challenges than the conditional image generation and unconditional video generation problem. In our case, each of the frames will be generated based on

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Figure 1: Generating video sequence based on a given pre-condition (sentence description).

the previous frames and the given precondition such as in Figure 1. The numbers of previous frames can vary from zero to a maximum number. Thus the usual approach of using 3D convolution will not be applicable in our case. Moreover, we need to make the precondition available to the system at each time frame so the whole generated video is consistent with the pre-condition.

2 Our Approach

In our problem, we are given a pre-condition either in the form of a action class name, fMRI signal or textual description and we need to generate a sequence of video frames which will be coherent with the given pre-condition. We will first discuss how can we generate a high resolution frame independently based on the pre-condition using a stacked adversarial network. Then we will discuss how we can make a recurrent fully convolutional network so we can propagate the context to generate the next frame. Finally, we will describe our recurrent stacked adversarial network to generate videos based on pre-conditions.

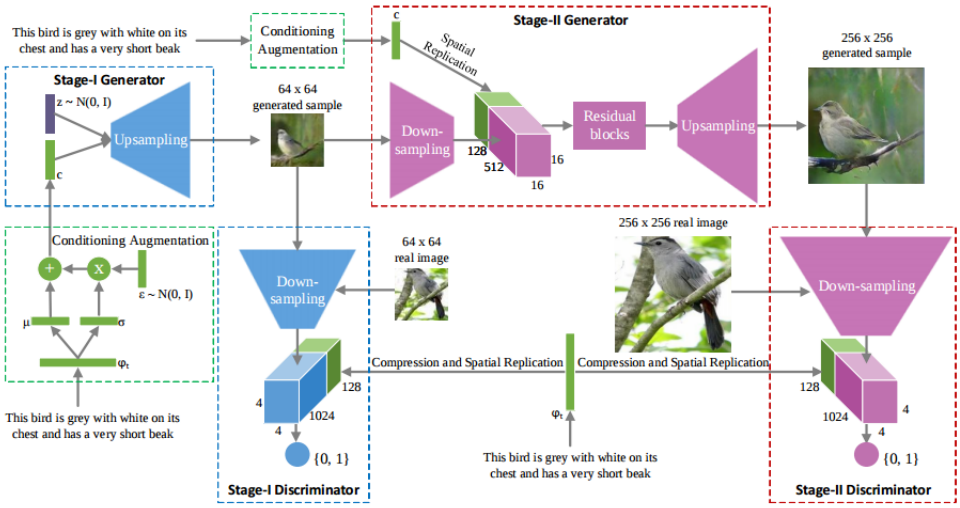

Figure 2: The architecture of the StackGAN from [1]. The Stage-I generator draws a low resolution image by sketching rough shape and basic colors of the object from the given text and painting the background from a random noise vector. The Stage-II generator generates a high resolution image with photo-realistic details by conditioning on both the Stage-I result and the text again.

2.1 Stacked Adversarial Network

To generate a video we first need to learn to generate a single frame from the pre-condition. We have adapted the StackGAN adversarial network model [1] for this problem, as it can generate photo-realistic images from sentence descriptions. This model first generates an image based on the encoded text and a random vector and at first stage. Then the generated image in the first stage is used as the input with the encoded text vector to the second stage adversarial network to generate a photo-realistic image. The architecture of the model can be seen in Figure 2.

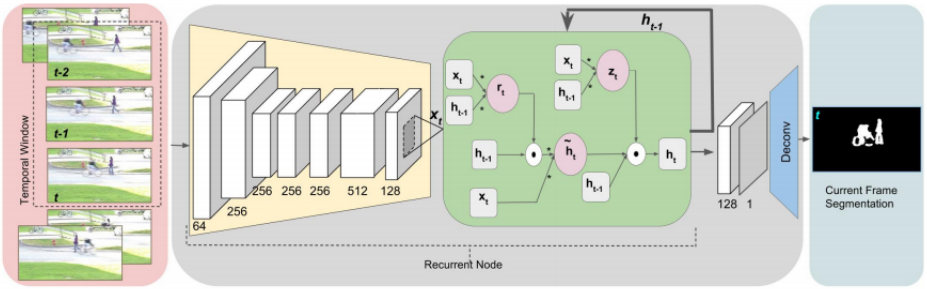

Figure 3: The architecture of RFC-VGG from [2]. Images are fed frame by frame into a recurrent FCN. A Conv-GRU layer is applied on the feature maps produced by the preceding network at each frame. The output of this layer goes to one more convolutional layer to generate heat maps. Finally, a deconvolution layer up-samples the heat map to the desired spatial size.

2.1.1 Recurrent Fully Convolutional Network

Now to generate a video we need to pass the contextual information from previous frames to the current frame so the frames are coherent to each other. For the discriminator network we can do that using a regular LSTM network. The discriminator network generates a vector from the input images which can be transfered to the next instance of the discriminator. But ~~as~~ the generator network is mostly a fully convolutional neural network it is not straightforward to create a recursive model for the generator. We have considered the recurrent fully convolutional network [2] to solve this problem. The model uses convolutional gated recurrent units which establishes recurrent connections between the colvolutional layers. This model preserves the spatial information while passing context to the next LSTM unit and reduces the number oflearned parameters as well. The model is described in Figure 3.

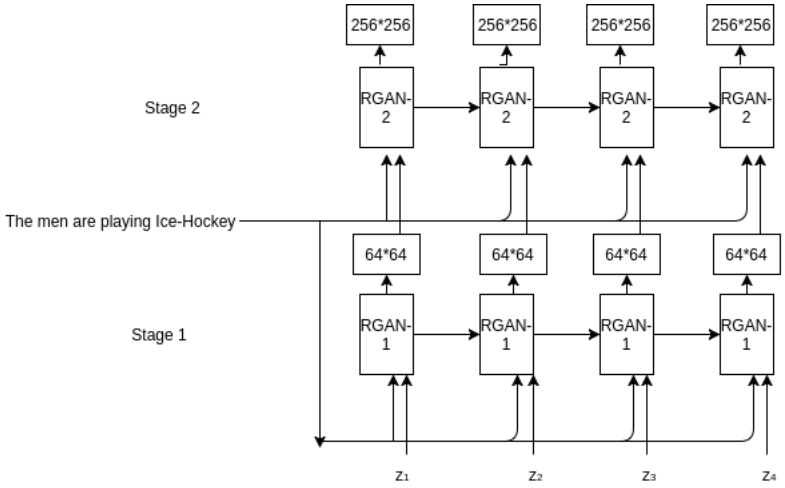

Figure 4: The proposed Recurrent Stacked Generative network (RSGAN). The recurrent adversarial network modules at the first stage (RGAN-1) take the encoded pre-condition and a random vector z_t and then produces a low resolution (64x64) size image. The modules at stage-2 (RGAN-2) takes the generated image at stage-1 and the encoded pre-condition to generate a high resolution image (256x256).

2.1.2 Recurrent Stacked Adversarial Network

Now, we have the models to generate individual images from the precondition and also we can connect this individual adversarial networks using recurrent connection. We combine these two models and propose the Recurrent Stacked Generative network (RSGAN) for conditional video

generation. The model expands both in the temporal and spatial dimension. Each module in the first stage takes the encoded pre-condition and a random vector as input and passes a contextual matrix to the next module to generate a low resolution frame sequence. Then the modules in the second stage takes the encoded pre-condition and the low resolution output from the previous stage to generate a high resolution video sequence. The model is described in Figure 4.

2.1.3 Training

The model is trained using back propagation through time (bptt) and the usual adversarial training to train the recurrent generators and discriminators.

3 Experiments

To evaluate our model we will conduct experiments on three video datasets.

3.1 UCF-101 Dataset

The UCF-101 [7] dataset contains short videos of 101 action classes. We will get the word-vector representation of the action class names and use those as the pre-conditions to generate the videos.

3.2 VIM-2 Dataset

VIM-2 dataset [8] contains 7200 seconds of training and 540 seconds of test fMRI BOLD (blood oxygen level) signals and corresponding 15 frames of stimulation for each second. We will use the fMRI signals from the early visual area (V1, V2, V3, V3a, V3b) as our precondition to reproduce the video stimulations.

3.3 Large Scale Movie Description Dataset

Finally, we will use the Large Scale Movie Description Dataset [9] which contains 4-5 seconds videos from movies and their textual descriptions. We will use the Skip-Thought vector to encode the sentence descriptions and use them as pre-conditions.

4 Evaluation

As this is the first work on conditional video generation, we do not have a baseline or established evaluation criteria yet. We are still working on a reasonable evaluation criteria. Human evaluation can be an effective approach here.

5 Conclusion

We have proposed a novel adversarial network based model to generate videos based on a given condition. Our model can generate video frames which are coherent and consistent to the given condition.

References

[1] Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X. & Metaxas, D. 2016. *StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. arXiv preprint arXiv:1612.03242.*

[2] Valipour, S., Siam, M., Jagersand, M., & Ray, N. (2016). *Recurrent Fully Convolutional Networks for Video Segmentation. arXiv preprint arXiv:1606.00487.*

[3] Mogren, O. (2016). *C-RNN-GAN: Continuous recurrent neural networks with adversarial training. arXiv preprint arXiv:1611.09904.*

[4] Yan, X., Yang, J., Sohn, K., & Lee, H. (2016, October). *Attribute2image: Conditional image generation from visual attributes. In European Conference on Computer Vision (pp. 776-791). Springer International Publishing.*

[5] Mathieu, M., Couprie, C., & LeCun, Y. (2015). *Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440.*

[6] Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). *Generating videos with scene dynamics. In Advances In Neural Information Processing Systems (pp. 613-621).*

[7] Soomro, K., Zamir, A. R., & Shah, M. (2012).*UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild., CRCV-TR-12-01*

[8] Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). *Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology, 21(19), 1641-1646.*

[9] Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., ... & Schiele, B. (2016). *Movie Description. arXiv preprint arXiv:1605.03705.*