Script2Scene: Automatic Generation of complex visual scene from a textual script

Ajay Narayanan
TCS Research and Innovation
Bangalore, India
ajay.narayanan@tcs.com

***Abstract*—Generating videos from a textual descriptions (such as a script) is a non-trivial task due to the complexity of temporal relationship between image frames in a video.A complex visual scene is one which contains more than one non-trivial interaction between one or more subjects. Although, GANs have been successfully used to create short coherent videos from textual descriptions, it is still very challenging to generate realistic complex scenes from textual scripts. This proposal aims to understand the extend to which current SOTA models are able to generate realistic videos of complex scenes. Also, this aims to create a new pipeline for the generation of complex scenes from textual scripts.**

***Index Terms*—vision and language, script parsing, generative adversarial networks**

I. Introduction

Automatic generation of scripted video has tremendous applications in a wide variety of different fields like video editing, video game generation and computer aided design. However, many modern and creative works require a huge amount of human intervention through various digital graphical design tools. Thus, a system that can create a visual representation of complex interactions described through a textual script could help increase the accessibility of digital media creation to the larger public. This proposal takes a step in that direction by conceptualising an text to video generation task.

Specifically, the focus of this proposal is complex scene generation from descriptive scripts. Compared to image generation [3]–[6], the problem of video generation is more complex because a video is a coherent series of images that should follow strong spatial and temporal dependencies. Generation of video from text is even more complicated due to difficulties in aligning the textual information to the video frames. Thus, the already existing text-to-image models cannot be reused for this problem.

Recently, some exciting work has been proposed to address the challenge of video generation from text. Li et al. [2] proposed a two-stage VAE based generator to yield 'gist' of the video, where gist is an image with information about the background and the object layout. However, since they mostly neglect the relations between consecutive frames, the video tends to be incoherent at times. On the other hand, Pan et al. [7] proposed a novel discriminator architecture that considers the temporal relation between frames. However, in this method, the model fails to identify subtle semantics of the text, which is crucial to getting the details right. To handle both these problems, Chen et al. [1] proposed a bottom-up GAN architecture that handles both the temporal relation between frames and the semantics of the text to produce short videos. However, this method only takes into consideration very simple interactions that involve only a single subject.

This proposal advocates the generation of a complex visual scene, typically involving more than 1 interacting between more than 1 subjects. This proposal aims to investigate the performance of the above models when provided with a more complex scene. This problem will also investigate how to include long-term attention to the textual semantics and how temporal relations between the frames can be maintained for complex interactions. To make this problem more concrete, we assume that the textual script has a semantic structure, which is most commonly seen in screenplays.

II. Problem Statement

The objective of the problem is the automatic generation of a visual scene involving complex interactions among subjects, all of which are described in a textual script. This problem is closely related to Video Generation from Text problem but is more complicated as the generative model must correctly model the interactions between subjects in the script. The main problem can be divided into sub-problems based on the complexity. There are many open research problems involved in this due to fact that it is a relatively less explored area in generative computer vision.

Some of the key problems this proposal wishes to address are: 1) How to parse a textual script to a context vector that encompasses the various interactions in the scene? 2) How to infer knowledge from the structure(semantics) of the textual script? 3) Design of a GAN or VAE architecture for converting the context vector to a video. 4) Whether the train the different modules individually or jointly? 5) Design of loss functions and the methodology for training.

III. Preliminary Literature Review

A. Video Generation

Video generation is a topic that has been studied extensively by the research community recently. [8] proposes

a generative adversarial network for video with a spatio-temporal convolutional architecture that untangles the scene's foreground from the background. In an effort to learn the semantic representation of unlabeled videos, [9] exploits two different types of generators: a temporal generator and an image generator, which transform a single latent variable to a video. [10] proposes a framework to generate videos by decomposing motion and content in a unsupervised manner.

Many similar works focus on generating a video conditioned on a static image(frame) as input, such as [11]–[14]. A combination of using a CNN for generation and frames and a sequence-to-sequence model like an RNN for frame prediction is used to achieve greater performance in works like [15]–[19]. Although these models have made wondeful strides in generation of unconditioned videos and videos conditioned on a input image, this proposal intends to generate a video of a complex scene conditioned on a textual script.

B. Video Generation from Text

Video generation from text aims to produce a videa that is semantically aligned to a given text or caption. A method to generate a video from text by combining a Variational Autoencoder with Attention mechanisms was proposed by [20]. [21] improves on this by incorporating short-term and long-term dependencies among image frames in a incremental manner. Recently, Li et al. [2] proposed a two-stage VAE based generator to yield 'gist' of the video, where gist is an image with information about the background and the object layout. Pan et al. [7] proposed a novel discriminator architecture that considers the temporal relation between frames. Chen et al. [1] proposed a bottom-up GAN architecture that handles both the temporal relation between frames and the semantics of the text to produce short videos. Unlike these methods which aim to generate videos of simple scenes from captions, this proposal intends to create a architecture able to produce viable visual representation of complex scenes.

IV. Challenges

On the ground level, this problem involves 2 major components:

1) Script parsing module

2) Video generation module

I believe that both these modules pose a lot of open technical challenges. In the Script Parsing module, there is a need to parse the textual data into a form that can be used by the video generation module. Concepts pertaining to Natural language processing would be needed to convert the script into an intermediate representation. Here, it must be noted that the structure of the script provides semantic meaning to the actual text as well. Also, association of a dialogue to a subject is crucial to the video coherence.

In the video generation module, main challenges involve ways to produce image frames that are consistent with the script or screenplay. Also, components like emotion, pose, dialogue delivery etc pose a huge challenge as most of the generative algorithms today are not really equipped to deal with such complexities. One other complication arises when there are multiple subjects in the scene and multiple interactions among these subjects. Video generation techniques so far, to the best of my knowledge, do not handle such intricate interactions between multiple subjects.

V. Milestones

Since the scope of this research problem is broad and given that not much research has gone into such problems, I feel it is important to divide the tasks into different technical milestone. A brief description for the technical milestones involved in this project is shown below:

1) **Environment and subject generation:** Most scripts or screenplays start by providing a description of the environment and the subjects involved in the scene. It usually acts as the starting frame of the scene. Creating this starting frame by parsing the script will be the first technical milestone. Note that the environment also usually has some information about the start position of the subjects as well.

2) **Single Subject Interaction generation in the created environment:** The 2nd major milestone in this problem will be generating short video clips of the interaction between the subject and the environment. Note that this does involves only single subject interactions. A simple example could be "The red bearded paced around in the main hallway. Hearing a knock in the front door, he rushed to open it". Here, this sentence describes a simple interaction of a subject with the environment. Parsing this information from the script and generation of a series of images which shows this interaction will be the second milestone.

3) **Interaction between multiple subjects in the simulated environment:** The next milestone would be the generation of simple interactions between multiple subjects. These may involve the show of emotion, pose, actions and even dialogues. This will purely focus on the interaction among the subjects alone and leave out the interactions of individual subjects with the environment.

4) **Temporal continuity and complex interactions:** Temporal continuity is a key component of any video, there needs an elegant and logical connectivity between frames. Also, there should be a logical connectivity in the interactions of the subjects among themselves and their environment.

At each stage of these milestones, the complexity of the Script parsing module increases. The goal is to produce an end-to-end framework that coverts the textual knowledge into a visual representation.

References

[1] Q. Chen, Q. Wu, J. Chen, Q. Wu, A. van den Hengel and M. Tan, "Scripted Video Generation With a Bottom-Up Generative Adversarial Network," in IEEE Transactions on Image Processing, vol. 29, pp. 7454-7467, 2020, doi: 10.1109/TIP.2020.3003227.

[2] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, Aniruddha Kembhavi; Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 598-613

[3] . Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016,arXiv:1605.05396.[Online]. Available: http://arxiv.org/abs/1605.05396

[4] T. Xuet al., "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," inProc. IEEE/CVF Conf.Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1316–1324.

[5] H. Zhanget al., "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," inProc. IEEE Int. Conf.Comput. Vis. (ICCV), Oct. 2017, pp. 5907–5915.

[6] H. Zhanget al., "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," 2017,arXiv:1710.10916. [Online].Available: http://arxiv.org/abs/1710.10916

[7] Y. Pan, Z. Qiu, T. Yao, H. Li, and T. Mei, "To create what you tell:Generating videos from captions," in Proc. ACM Multimedia Conf.(MM), 2017, pp. 1789–1798.

[8] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in Proc. Adv. Neural Inf. Process. Syst., 2016,pp. 613–621

[9] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in Proc. IEEE Int. Conf. Comput. Vis.(ICCV), Oct. 2017, pp. 2830–2839.

[10] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN:Decomposing motion and content for video generation," in Proc.IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018,pp. 1526–1535.

[11] Y.-W. Chao, J. Yang, B. Price, S. Cohen, and J. Deng, "Forecasting human dynamics from static images," in Proc. IEEE Conf. Comput. Vis.Pattern Recognit. (CVPR), Jul. 2017, pp. 548–556.

[12] B. Chen, W. Wang, and J. Wang, "Video imagination from a single image with transformation generation," in Proc. Thematic Workshops ACM Multimedia-Thematic Workshops.NewYork,NY,USA:ACM,2017, pp. 358–366.

[13] V. Vukoti c, S.-L. Pintea, C. Raymond, G. Gravier, and J. C. Van Gemert,"One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network," in Proc. Int. Conf. Image Anal.Process.Springer, 2017, pp. 140–151.

[14] . Walker, C. Doersch, A. Gupta, and M. Hebert, "An uncertain future:Forecasting from static images using variational autoencoders," in Proc.Eur. Conf. Comp. Vis., 2016, pp. 835–851

[15] X. Jia, B. D. Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 667–675.

[16] N. Kalchbrenneret al., "Video pixel networks," inProc. Int. Conf. Mach.Learn., 2017, pp. 1–16.

[17] J. van Amersfoort, A. Kannan,M. Ranzato, A. Szlam, D. Tran, and S. Chintala, "Transformation-based models of video sequences," 2017,arXiv:1701.08435. [Online]. Available: http://arxiv.org/abs/1701.08435

[18] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," inProc. Int. Conf.Learn. Represent., 2017, pp. 1–22.

[19] W. Xiong, W. Luo, L. Ma, W. Liu, and J. Luo, "Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,Jun. 2018, pp. 2364–2373.

[20] G. Mittal, T. Marwah, and V. N. Balasubramanian, "SyncDRAW:Automatic video generation using deep recurrent attentive architectures,"in Proc. ACM Multimedia Conf. (MM), 2017, pp. 1096–1104.

[21] T. Marwah, G. Mittal, and V. N. Balasubramanian, "Attentive semantic video generation using captions," in Proc. IEEE Int. Conf. Comput. Vis.(ICCV), Oct. 2017, pp. 1426–1434.