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

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

***Abstract*—**

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

I. 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.

II. Existing work

The confluence of computer vision and natural language processing has given rise to several interesting research problems like VQA, visual grounding. In generative computer vision, a lot of work has been done to generate images based on natural language captions. Models like StackGANs have been largely successful in achieving this. Generation of images is a much easier task when compared to generation of videos because of the temporal relationships in videos. This task of video generation also has been widely explored, but the basic assumption that most people make is that the first frame and the last frame are already known. Another line of work that has been going that has gone into automatic video generation is a probabilistic prediction of future frames, sometimes conditioned by some attributes.

When it comes to video generation from text, the problem involves understanding the text and generation of frames which visualizes the textual information. An interesting work by Li et.al [1], is b far, the best attempt at this problem. They suggest a hybrid architecture consisting of both a VAE and GANs to extract both static and dynamic information from textual and produce a series of frames that visualizing it. Although an excellent attempt, the work only considers quite simple and disjointed textual information. The research proposal takes into consideration complex scripts or screenplays of a scene and intents to produce a video representation of that scene. To the best of my knowledge, this is a novel application of generative computer vision combined with natural language understanding.

III. Open technical and research 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.

IV. Technical 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