Long-Text-to-Video-GAN

Ayman Talkani, Anand Bhojan

Long-text-to-video-GAN

Ayman Talkani and Anand Bhojan

School of Computing, National University of Singapore, COM1, 13, Computing Dr, Singapore 117417

Abstract. While there have been several works regarding the task of video generation from short text[1][2][3], which tend to focus more on the continuity of the generated images or frames, there has been very little attention drawn towards the task of story visualization[4], which attempts to generate dynamic scenes and characters described in a large amount of detail in a multi-para input text. We therefore propose our own novel take on this task which attempts to compile these dynamic scenes into a larger video, while also improving the scores of the current state of the art models in story visualization and video generation respectively. We intend to do this by making use of semantic disentangling connections[5] in between our generators in order to maintain global consistency between consecutive images, as well as ensuring similarity between the video re-description and the input text, thus leading to a higher image quality. Once these images are generated, we make use of a depth-aware video interpolation framework[6] in order to generate the remaining non-existing frames of the video in between the generated images. We then evaluate our model on the CLEVR-SV and Pororo-SV datasets for the story visualization task, and the UCF-101 dataset to measure the accuracy of the video generated. This way, we intend to outperform existing state-of-the-art models significantly.

Keywords: Generating Adversarial Networks · Text-to-Video

1 Introduction

Recently, there has been an explosive influx of research on generating models. These include GANs[7] and VAEs[8], which are currently the best performing models in this area. These models have been utilized in many tasks from the generation of new unseen images or text, to generating multiple output images from the given input text. While producing sequences of images from natural language is a daunting task by itself, very little work has been done in order to generate coherent sequences of images for multi-paragraph sentences as input[4], and virtually no work has attempted the generation of video from this long text. We thus propose our own take on a Story visualization GAN which attempts to successfully achieve the above task as shown in Figure 1. In order to attain this result, we must ensure that the images generated consistently and coherently

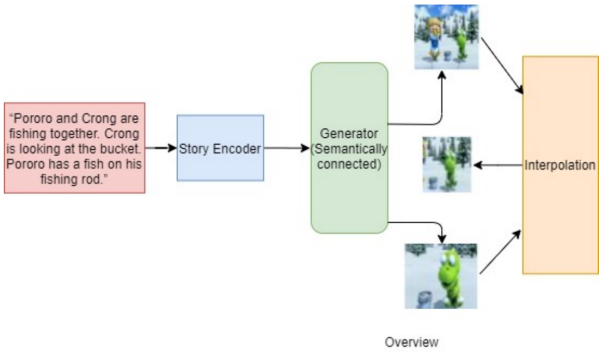

Fig. 1. Overview of model with example

depict the whole story described in the input. This task is highly related to text-to-image generation, a topic that has been extensively researched[5]. However, one of the most conspicuous flaws with these types of models is their inherent lack of consistency. As human descriptions are highly subjective and diverse in their expressions, we frequently attain different images for the same input text. [5] attempts to fix this problem by making use the SD-GAN, which makes use of a Siamese structure along with a contrastive loss in order to distill semantic commons from texts for a more consistent output with similar text. Inspired by this approach, we propose a novel GAN that generates sequential images by making use of a context encoder in order to generate sequential images of the different visual scenes, while ensuring that the images maintain semantic consistency. Once we have attained the relevant scenes, we make use of a depth-aware video interpolation[6]method in order to develop additional frames, thus creating a longer and more dynamic video from the input text.

We summarize our contributions as follows:
(1)We propose a novel GAN architecture for the story visualization task that makes use of semantic connections in order to output a cohesive output from our input.
(2)With the use of our novel GAN that makes use of Attention networks[9] and Video captioning in order to ensure similarity between this generated caption and the input text, we attempt to attain a higher visual quality that the current state-of-the-art models.
(3)We propose a novel task where we convert the sequence of images from a multi-paragraph text into a long video involving dynamic scene changes. We experiment on the CLEVR-SV and Pororo-SV datasets in hopes of outperforming current state of the art models in story visualization, and the UCF-101 dataset in order to test the accuracy of our video generation. We further add annotations to the CUB dataset in order to test our model on similar text in order to test it's inter-class and intra-class variability, similar to [5].

2 Related Work

Generative Adversarial Networks[7], Variational Auto-Encoders[8] and several other generating models have made significant strides in many tasks, including text-to-image generation[5], video-generation[1][2][4][3], style transfer and many more tasks. The task of story visualization falls into this task of generating networks, but has distinct aspects to it. It aims to generate sequential images for dynamic scenes without focusing on continuous frames between the images.
One of the most relevant topics to this topic is text-to image generating networks[5] which can generate high resolution images through the use of cascaded generators[10], Attention-networks[9], Re-descriptions[11] and so on. The task of story visualization[4] also attempts to understand longer and more complex input text. For example, this has been explored in dialogue-to-image generation, where the input is a complete dialogue session rather than a single sentence [12]. In our particular model, we prioritize output consistency[5]. We will further modify this SD-GAN in order to attain a higher video quality.
Another important task related to story visualization is video generating networks, especially that of text-to video[1][2][4][3] or image to video generators. Models in this domain tend to be focused on a smooth motion transition across successive video frames. A trajectory, skeleton or simple landmark is used in existing work, to help model the motion feature [13][14][15]. In the task of story visualization however, the whole story sets the static features and each input sentence encodes dynamic features. While conditional video generation has only one input, the story visualization task has sequential, evolving inputs; and while images in this task aim to visualize a story through discrete and often with different scene views, we will attempt to make these frames continuous with the help of interpolation.
We will also be making use of video frame interpolation in order to attain a smooth transition between the dynamic scene changes based on the input text. While there have been several implementations in this field[16][17], we will be making use of [6], which makes use of depth estimation and optical flow in order to attain a relatively high performance. While existing research aims to increase the frame rate of video, we intend to utilize this model in order to generate more sequential images between our key images, thus leading to smoother video generation.
Inspired by [11], we also attempt to make use of re-descriptions for our outputs in order to compare it with the input text. However, as the output is a video in this case, we cannot make use of standard image captioning. We will thus make use of recent advancements in video captioning, particularly multi modal dense video captioning[18] in order to achieve this task.

3 Our Method

In this section, we go over the basic pipeline utilized in this model. We make use of advancements in text-to-image GANs[11][9][10][5]in order to significantly

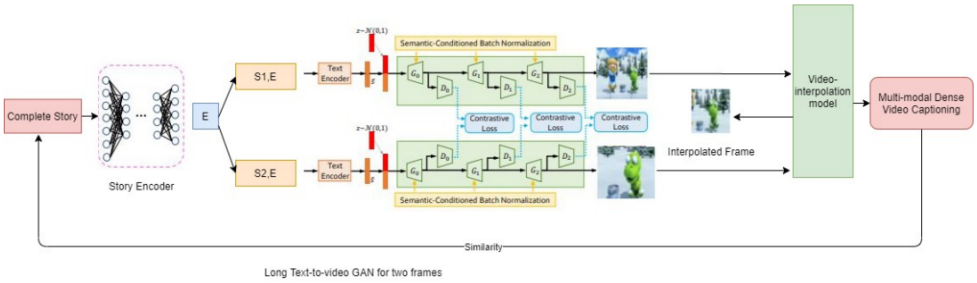

Fig. 2. Long-text-to-video-GAN. Note that the discriminators here refer to both the image and story discriminator as described in [4]

enhance the quality of the video produced. We further make use of multi-modal dense video-captioning[18] along with the the comparison of the text used in the STREAM module from [11] in order to attain a higher video quality. We also add semantic connections similar to [5] in the current state-of-the-art models in order to implicitly disentangle semantics from the text description. This way, we can make sure that the sequentially generated images remain semantically consistent for similar text throughout the generating process. While this text-to-image GAN can help the model visualize the different dynamic scenes used in the story visualization task[1], it does not generate enough frames for a smooth transition between these scenes due to a lack of semantic consistency between their images, leading to drastic changes between the scenes. We attempt to ameliorate this effect with the help of the semantic connections from [5], along with video interpolation[16][17][6]. Through this pipeline, we are able to attain a smooth video from the given text, while also ensuring it's cohesiveness. We will then train this model in the CLEVRSV and Pororo-SV datasets for the story visualization task, and the UCF-101 dataset for the video generation task, and compare the results with state of the art models in terms of quality, overall relevance and consistency.

3.1 Long-text-to-video-GAN

Our text to video GAN is motivated by the image re-descriptions utilized in [11] so as to guide the multistage cascaded generator[9] to produce more accurate images with relatively scarce data. We will also make use of the contrastive loss from [5] in order to maintain semantic consistency between the generated scenes. These semantic connections between the discriminators ensure image consistency throughout the generation process. This way, we can produce high quality sequential images that are relevant in both global, as well as sentence levels. We will then make use of multi-modal dense video captioning in order to achieve our novel task of long-text to video interpolation.

3.2 Video Interpolation model

On passing through our GAN, we attain sequential images for the many dynamic scenes described in the input. However, we lack smooth transitions between these images. In order to solve this, we make use of video interpolation in order to interpolate in between the generated scenes and increase the frame-rate, thus leading to a much more smooth and realistic video. While there are many different kinds of video interpolating models available[16][17], we will be making use of the depth based frame interpolation[6] as it is the current state of-the-art, making use of optical flows and depth maps in order to achieve this result.

3.3 Objective functions

We shall train our Generator, Discriminator and text re-description on the loss functions described in the basic architecture of [11]. However, we will be making use of multi-modal dense video captioning in this loss instead of the standard image-captioning model used in [11]. We will also be adding the contrastive loss from [5] in order to ensure inter-class diversity and intra-class similarity. The loss for our generator will thus look like this:

$$L_g = L_{mg} + \lambda * L_{stream} - E_z * [D_w[I_i]] + L_{contr} \tag{1}$$

Where the last term is Wasserstein loss, L_{mg} is the generator loss without the image captioning part and I_i is the generated image from the distribution pi in the i_{th} stage.The first term of L_{mg} is the visual realism adversarial loss, which is used to distinguish whether the image is visually real or fake, while the second term is the text-image paired semantic consistency adversarial loss, which is used to determine whether the underlying image and sentence semantics are consistent. L_{stream} refers to the loss from our text-similarity score from [11] and lambda is a loss weight to handle the importance of adversarial loss and the text-semantic reconstruction loss.
Finally, in order to train our model for the story visualization task, we will be making use of the CLEVR-SV and Pororo-SV datasets for the story visualization task, and the UCF-101 dataset for the text-to-video task, as these are used by the current state of the art models. By making use of our superior generators, as well as video interpolation, we hope to surpass it's performance in visual quality, as well as relevance and consistence.

4 Experiments

While this project is still a work in progress, we have made some progress regarding this task that we would like to document here. As mentioned earlier, we intend to test our model out on both the story-visualization task, as well as the video-synthesis task. For now, we have prioritized working on the story visualization task as we primarily need to generate sequential images that are

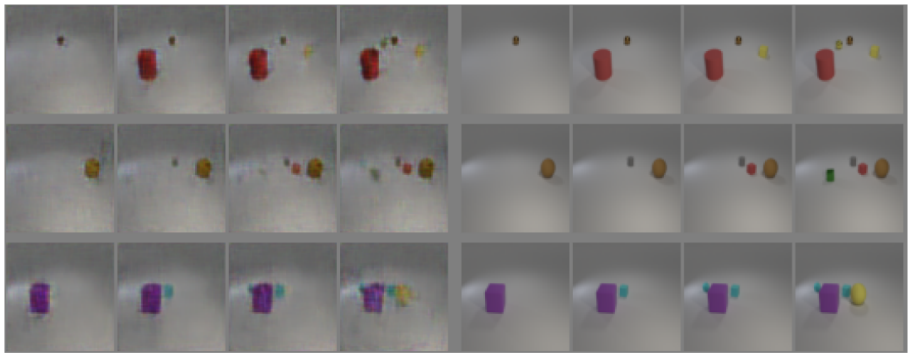

Fig. 3. Story visualization task on CLEVR dataset. Our generated images take the first five columns on the left, and ground truth takes the next five columns on the right

similar to each other, as this will make it an easier task to interpolate between the sequential images and generate a dynamic video from it.

While we would ideally like to use the modified PororoSV dataset from [4], we had been unable to attain this dataset until very recently. For our initial experiments, we were limited to the use of the modified CLEVR dataset as it was publicly available, and had aimed to generate sequential images from relatively long input text.

By utilizing a contrastive loss between our discriminators similar to [5], we have been able to produce consecutive images that are very similar to each other. We have added these results in Fig. 3 as shown.

5 Conclusion and Future Work

While the sequential images generated by our GAN are similar, the CLEVR dataset is not particularly dynamic in general. Therefore, we must test our architecture on a more challenging and dynamic dataset with diverse sequential images in order to truly test our model. As we have recently attained the PororoSV dataset from the authors of [4], we can now test our model out on this dynamic dataset instead, and aim to generate similar sequential images that are still able to outperform [4] in this task.

Once we are able to generate similar sequential images, we can then use video interpolating models like [6] in order to generate a dynamic video from multi-paragraph text as input. We will then attempt to further improve video quality by making use of video captioning models like [18] to improve image quality similar to [11].

6 Acknowledgment

This work is supported by the Singapore Ministry of Education Academic Research grant T1 251RES1812, “Dynamic Hybrid Real-time Rendering with Hard-

ware Accelerated Ray-tracing and Rasterization for Interactive Applications". Special thanks to the 'National Supercomputing Centre (NSCC) Singapore', for providing the computational resources required for training our architecture.

References

1. D. Kim, D. Joo, and J. Kim, "Tivgan: Text to image to video generation with step-by-step evolutionary generator," IEEE Access, vol. 8, pp. 153113–153122, 2020.
2. Y. Li, M. R. Min, D. Shen, D. E. Carlson, and L. Carin, "Video generation from text.," in AAAI, vol. 2, p. 5, 2018.
3. H. Yu, Y. Huang, L. Pi, and L. Wang, "Recurrent deconvolutional generative adversarial networks with application to text guided video generation," arXiv preprint arXiv:2008.05856, 2020.
4. Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, and J. Gao, "Storygan: A sequential conditional gan for story visualization," in Proceedings of the IEEE Conference on Computer Vision and Pattern
5. G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2327–2336, 2019.
6. W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3703–3712, 2019.
7. I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," arXiv preprint arXiv:1701.00160, 2016.
8. Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in Advances in neural information processing systems, pp. 2352–2360, 2016.
9. T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1316–1324, 2018.
10. H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in Proceedings of the IEEE international conference on computer vision, pp. 5907–5915, 2017.
11. T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1505–1514, 2019.
12. S. Sharma, D. Suhubdy, V. Michalski, S. E. Kahou, and Y. Bengio, "Chatpainter: Improving text to image generation using dialogue," arXiv preprint arXiv:1802.08216, 2018.
13. Z. Hao, X. Huang, and S. Belongie, "Controllable video generation with sparse trajectories," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7854–7863, 2018.
14. W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, "Every smile is unique: Landmark-guided diverse smile generation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7083–7092, 2018.
15. S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Efficient parametrization of multi-domain deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8119–8127, 2018.

16. S. Niklaus and F. Liu, “Context-aware synthesis for video frame interpolation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1710, 2018
17. J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, “Epicflow: Edge-preserving interpolation of correspondences for optical flow,” in Proceedings of the IEEE conference on computer vision and pattern recognition ,pp. 1164–1172, 2015.
18. V. Iashin and E. Rahtu, “Multi-modal dense video captioning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,pp. 958–959, 2020.