This document is a preliminary project proposal for the course CS1699 - Deep Learning at the University of Pittsburgh in Spring 2020.

Visual Storytelling with Stick Figures

Haoyue Cui, Ziyi Gong, Sijia Rong, Xingchen Zhao

(Last names in alphabetical order)

Abstract

Human movement is often accompanied by stories. Given a noise vector, we would like to generate a reasonable sequence of fluent movements that tells a story. Here we proposed a model that combines two generative adversarial network (GAN) [4] to create videos of stick figures. We first use keypoints extracted from videos of human behaviors to train a generator yielding sets of fluent body movements, and then train a choreographer that scales, displaces and rotates each set as a whole such that the stick figures interact with each other and tell a story.

1 Introduction

Currently, visual stories are used in many applications such as advertisements, social network memes, and artistic creations, yet professional knowledge and experience are required for people to produce coherent and reasonable visual stories. It is interesting to see how capable machines are to generate visual stories from scratch. Moreover, a good generative model for visual storytelling are useful to provide novel ideas for designing, enhance freedom in video games, etc.

However, learning to generate a meaningful and smooth video accompanied by a story is challenging, because storytelling entails high cognitive ability. Previous works on story visualization have been focusing on generating videos or sequences of images to describe an input story written in multiple sentences [3][6][7][8]. Textual semantic embedding makes

the task feasible, yet the semantic embedding models used possibly lack some latent features or add unnecessary constraints.

Nevertheless, story visualization might not always be correlated to textual stories. We would like to see if a trained model can generate stories without texts as its cues. We propose a method of generating visual stories with stick-figures randomly by combining two generative adversarial network (GAN) [4].

2 Method

2.1 Model Structure

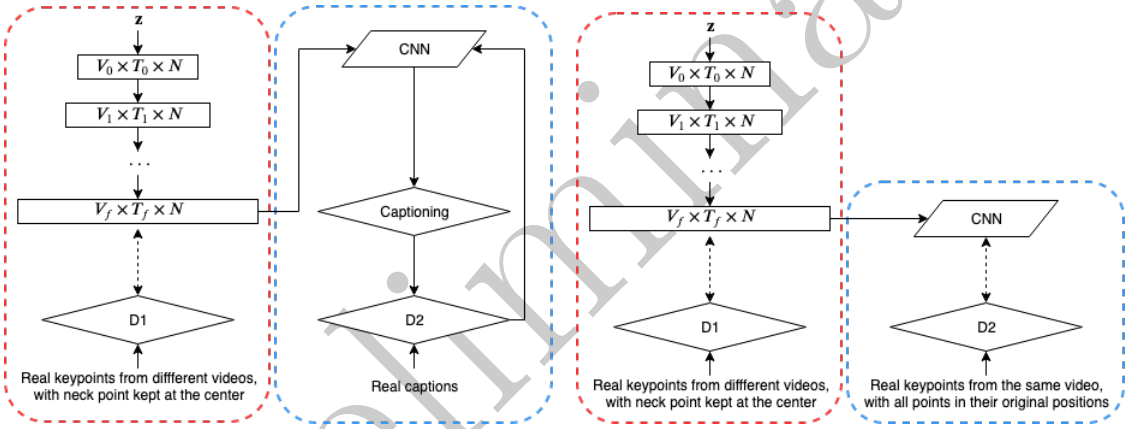

Figure 1: General structure of our frameworks.

Our method can be divided into two parts, fluent movement generation (Fig.1, red) and choreography (Fig.1, blue), each seeking to perform minimax optimization like a typical GAN:

$$\min_{\Theta_g} \max_{\Theta_d} \left\{ \mathbb{E}_{\mathbf{x} \sim \mathbf{X}} [\log D_{\Theta_d}(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim \mathbf{Z}} [\log(1 - D_{\Theta_d}(G_{\Theta_g}(\mathbf{z})))] \right\}$$

where $\theta_c, c \in \{g, d\}$ are the weight sets, x is a vector of real keypoints, z is a vector of noise, and G and D are the generator and discriminator respectively.

Firstly, the movement generator (Fig.1 red, top half) is trained to map a latent space with dimension (V_0, T_0, N) to a sample space with dimension (V_f, T_f, N), i.e. generating N

groups of V_f keypoint positions moving over T_f frames. Notice that in the tensor, each set of keypoints is kept at the center, i.e. the neck point is always $(0, 0)$ while others are changing. The generated tensor is forward to a discriminator $D1$ and another CNN (Fig.1 blue, top half). $D1$ tries to label each set of generated keypoints as fake movement and each set of keypoints extracted from different, real videos as real movement. The second CNN, the "choreographer," tries to find the location, scale, and rotating angle of each generated keypoint set as a whole, so as to create interactions between different stick figures. The outcome of the choreographer is captioned with a pre-trained video captioning model, and trained with a discriminator $D2$ which distinguishes real and fake captions (Fig.1a). This approach relies on the captioning model and might be hard to train and converge.
An alternative method (Fig.1b) is to directly train the choreographer with another $D2$ which takes real keypoint positions from the same video. This time, the real keypoint sets are not centered, because the relative locations and scales between different stick figures are considered in the training. The choreographer's network architecture should be carefully designed in order to successfully learn the transformations while preserving the variety of generated stories.

2.2 Data Collection and Preprocessing

Data are intended to be collected from Human Actions and Scenes Dataset [10] and YouTube-8M Segments Dataset [1], two large datasets that provide labelled video scenes suitable for joint keypoint extraction. Moreover, the video clips are from movies, games, and YouTube channels whose captions could be enriched by the descriptions of the original sources, with proper use of data gathering methods. Joint keypoints will be detected from the video scenes using OpenPose [2][13], and those where keypoints are unable to detect will be removed.

2.3 Evaluation Metrics in Consideration

2.3.1 Inception Score

Inception score (IS) is one of the most commonly used metrics to evaluate generative models[12]. In general, it evaluates both the variety and quality of the generated samples, i.e. the label distribution $p(y)$ and the label distribution given generated samples $p(y \mid \mathbf{x})$. The inception score is given by

$$\mathrm{IS}(\mathbf{X}) = \exp[\mathbb{E}_{\mathbf{x}\sim\mathbf{X}} D_{KL}(p(y \mid \mathbf{x}) \| p(y))]$$

2.3.2 Fréchet Inception Distance

Fréchet Inception Distance (FID) is another widely chosen metrics which measures the distance between the feature spaces of generated samples and real samples using a feature extractor such as CNN [5]. Let $\mu_g, \Sigma_g, \mu_r, \Sigma_r$ be the mean and covariance among features of generated samples and real samples respectively, FID is calculated as

$$d^2 = \|\mu_r - \mu_g\|_2^2 + \mathrm{Trace}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}})$$

2.3.3 Real-World Human Ratings Evaluation

Real-world human ratings evaluation is also a good way for evaluating the final output of the model in general, since human beings can provide easily-understood and informative feedback. Questions should be open-ended, such as:

1. Can you easily infer what are happening in these generated stick figure videos?
2. Rate the fluency and authenticity of those videos.

2.3.4 Evaluation based on Captioning

With a pre-trained video captioning model, the general comprehensibility of the stories generated can also be evaluated using their corresponding captions. Potential evaluation metrics for those captions include BLEU [11], ROUGE [9] and CIDEr [14].

3 Timeline

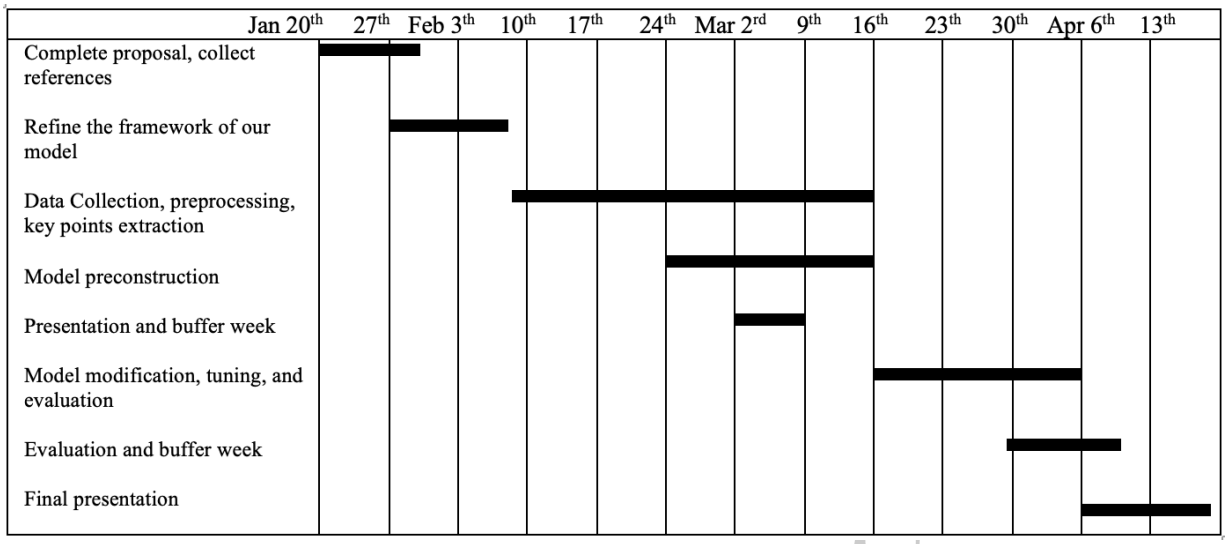

Figure 2: Gantt Chart, Timeline

References

[1] Abu-El-Haija, S. et al. "YouTube-8M: A Large-Scale Video Classification Benchmark". In: *arXiv:1609.08675*. 2016.

[2] Cao, Z. et al. "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields". In: *arXiv preprint arXiv:1812.08008*. 2018.

[3] Deng, K. et al. "IRC-GAN: Introspective Recurrent Convolutional GAN for Text-to-video Generation". In: *IJCAI*. 2019.

[4] Goodfellow, I. J. et al. "Generative Adversarial Networks". In: *arXiv e-prints*, arXiv:1406.2661 (June 2014), arXiv:1406.2661. arXiv: 1406.2661 [stat.ML].

[5] Heusel, M. et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium". In: *CoRR* abs/1706.08500 (2017). arXiv: 1706.08500. URL: http://arxiv.org/abs/1706.08500.

[6] Li, J. et al. "Video Storytelling: Textual Summaries for Events". In: *IEEE Transactions on Multimedia* 22.2 (2020). DOI: 10.1109/TMM.2019.2930041, pp. 554–565.

[7] Li, Y. et al. "StoryGAN: A Sequential Conditional GAN for Story Visualization". In: *CoRR* abs/1812.02784 (2018). arXiv: 1812.02784. URL: http://arxiv.org/abs/1812.02784.

[8] Li, Y. et al. "Video Generation From Text". In: *CoRR* abs/1710.00421 (2017). arXiv: 1710.00421. URL: http://arxiv.org/abs/1710.00421.

[9] Lin, C.-Y. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013.

[10] Marszałek, M., Laptev, I., and Schmid, C. "Actions in Context". In: *IEEE Conference on Computer Vision & Pattern Recognition*. 2009.

[11] Papineni, K. et al. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: (Oct. 2002), pp. 311–318. DOI: 10.3115/1073083.1073135.

[12] Salimans, T. et al. "Improved Techniques for Training GANs". In: *ArXiv* abs/1606.03498 (2016).

[13] Simon, T. et al. "Hand Keypoint Detection in Single Images using Multiview Bootstrapping". In: *CVPR*. 2017.

[14] Vedantam, R., Zitnick, C. L., and Parikh, D. "CIDEr: Consensus-based Image Description Evaluation". In: *CoRR* abs/1411.5726 (2014). arXiv: 1411.5726. URL: http://arxiv.org/abs/1411.5726.