Artificial Intelligence Paper Critique

Kshitij Srivastava (MT18099)
Paper Title: Video Generation From Text
Authors: Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, Lawrence Carin
Venue: AAAI 2018
Link: https://arxiv.org/pdf/1710.00421.pdf

I. SUMMARY

Generating images from text is a well studied problem, but video generation from text is not much explored. In this paper, the authors have achieved sizable results regarding the problem. Most of the video generation from text is done by generating a series of images from the given text and concatenating the frames to form a video. However, such a procedure yields poor results for this purpose, producing videos not complying with the text input. In order to produce correct results, the researchers have broken down the generation task into two components:

- A conditional variational auto-encoder(CVAE) model used to generate the static background of the video (gist).
- A GAN framework then conditions the gist and the text input to generate the content and motion of the video.

Instead of simply combining the text and gist information by concatenating their feature vectors, a set of image filters is computed from the input text and applied on the generated gist feature vector. The researchers have used YouTube videos for constructing the training data-set where the video titles and descriptions are used as the accompanying text. The procedure can be summarized as :

- Analyzing the input text and using it to generate a static gist produced by the CVAE.
- Generating image filters based on the input text and applying them on the gist to produce an encoded text-gist feature vector.
- Giving this feature vector to the GAN component to generate a related video.

A. Generating the gist

The background of a video is usually static, having small motion changes. This is the reason for using a gist. A CVAE (conditional variational auto-encoder) has been used to generate. The CVAE is trained using a pair of text and images. For images, the first frame of the video has been found to work.

$$L_{gist}(\theta_g, \phi_g; v, t) = \mathbb{E}_{q_{\phi_g}(z_g|v,y)}[log p\theta_g(v|z_g, t)] - KL(q_{\phi_g}(z_g|v,t)||p(z_g))$$

The encoder has two sub-encoder networks:

- $\eta(.)$ - applied to the video frame v.
- $\psi(.)$ - applied to the text input t.

The encoded video frame and text are combined using a linear combination layer. For testing purposes, $\eta(.)$ is ignored and only $\psi(.)$ is used. This is done to ensure the generation of a text-conditioned video. The output gist is given as an input to the video generator.

B. Video Generation

Three entangled neural networks in a GAN framework have been used here. GAN consists of two components:

- Generator: synthesizes fake samples to confuse the discriminator.
- Discriminator: tries to accurately distinguish synthetic and real samples.

Both compete in a mini-max game against each other and evolve themselves with time. GANs are quite efficient at generating images but using them for the problem at hand is quite complicated. Instead, a motion filter is first computed based on the input text, which is applied to the gist for conditioning. Unlike simply concatenating the feature sets of images from the text, this steps ensures the generation of video relevant to the input text t, having plausible motion. A text-gist vector ($\mathbf{g}_t$) is thus generated.
A random noise vector (Nv) is then added to Gt to introduce some motion diversity and detailed information. The output video from the generator is given by:

$$G(z_v = \alpha(z_v) \odot m(z_v) + (1 - \alpha(z_v)) \odot s(z_v)$$

A mask matrix $\alpha(.)$ has been used to separate the static scene from the motion. The output is a static background image, which is repeated through time to match the video dimensionality. The values given to s(.) are from an independent 2D neural network. Further details on the gist and motion parts of the video are created using text-gist vector.

C. Creating Image filters using input text

Instead of simply concatenating the gist and text encoding, computing motion-generating filter weights by utilizing the text information is a more reliable way. A 3D full-convolution layer of size $F_c \times F_t \times kx \times ky \times kz$ is used for this purpose. The text-gist filter is represented by the equation:

$$\mathbf{g}_t = Encoder(2D_{conv}(gist, f_g(t)))$$

where $f_g(t)$ is given by :

$$f_g(t) = 3D_{conv}(\psi(t))$$

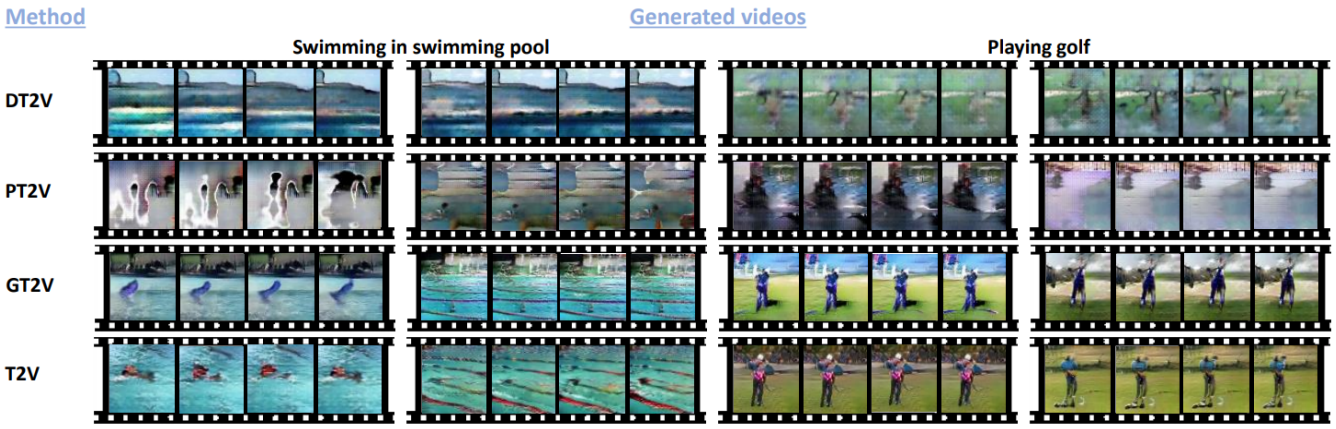

Fig. 1. Videos generated by different methods.

D. Objective Function

The objective function is:

$$L = \gamma_1 L_{CVAE} + \gamma_2 L_{GAN} + \gamma_3 L_{RECONS}$$

where γ_1, γ_2 and γ_3 are scalar weights for each loss term.

E. Methods used for video generation

- **Direct text to video generation (DT2V)**: Only encoded text and random noise are used without the generation of gist.
- **Text-to-video generation with pair information (PT2V)**: The text and video features are linearly concatenated into a pair and fed to the discriminator.
- **Text-to-video generation with gist (GT2V)**: Image filters computed form the input text are not used.
- **Video generation from text with gist and Text2Filter (T2V)**: The complete proposed model.

Figure 1 shows the samples generated by the above models. DT2V fails to generate plausible motion in the videos. Even PT2V doesn't provide much improvement over DT2V. GT2V gives the correct background but fails terribly at motion generation. T2V can be clearly seen as the best performing model here.

II. LIMITATIONS

- I think the input used for generating videos is too short. It isn't enough to provide much detail to the model for constructing detailed videos. This results in the video not having enough knowledge about the motion generation component.
- The training set consists of only 400 videos, which I believe is a bit small in terms of the output expected from the model as it directly relates to the model's understanding of the input text.
- The authors have not provided any information about the subject of the video in the input text, i.e., is the video about a human or an animal or any non-living object. Although, from the output samples it looks like a human-like form but it is no where mentioned in the paper.

III. SUGGESTIONS

- An immediate improvement would be to include more than one subjects in the scene. For this, the model should be trained on a similar data-set, having more than one subjects.
- Along with the input text, a longer description about the scene of the video might be more helpful. Like a script for the scene of a movie, itll help guide the model to produce much better, detailed videos.
- Expanding the training set to include more number of video and text pairs will help improving the model's understanding. Thus, producing better results.
- A description about the subject of the video might help in improving the results produced. Like including the movements of the subject in the scene.