TTVGAN: Teach Machine to be Creative, Video Generation from text using GAN

BMVC 2018 Submission # 1

Abstract

Visual Information or visual representation can be captured by human brain very easily over the textual data. Using recent advancement in adversarial learning and deep learning, we are able to generate images and videos from algorithm at some level. Using this as a prior we proposed the framework called Text to Video Generative Adversarial Network (TTVGAN) for generation of video as per given text input sentence containing a simple subject, object, and verb formation. The proposed framework generates the video using three key content from English sentence subject, Object, Verb (SVO Triplet) and convert it to action, actor and scene than using this input triplet, framework will generate the video. Video generation is the main part of framework which is accomplished by sampling a video frames from small video clips and adding noise sequentially over time using recurrent neural network into it. Now for better supervision over the generated content of the video we use the above said SVO triplet as condition parameter in adversarial training. With this, we also use context features from sentence to generate visually perfect video. Experimental result on the Human action Dataset (KTH dataset) and custom animation dataset are quite impressive with actions which are significantly different than each other. Our Proposed framework is able to generate small video with length ranging from 2 seconds to 5 seconds from modeled train data very well. This is the initial version of the work in which we tried to achieve simple small videos for easily described actions like walking, jumping, waving of hand etc.

1 Introduction

Generative Adversarial Network and its siblings like Autoencoders recently received an increasing amount of attention, not only because they provide the reason to learn feature representations in unsupervised or semi-supervised method that can leverage all image data available on Internet for learning but also because by using that kind of network structure we can create novel unseen images which further can be used in various another task.

1.1 Recent Works in Generative Models

Work Like "StackGAN" Zhang *et al*. [9] tried to generate images through text description very well. In their work authors proposed the GAN (Generative Adversarial Network) network with 2 stages and one RNN (Recurrent Neural Network network) for learning the sentence which is specifically used for extracting the features of the image to be generated. In "CVAE-GAN" Bao *et al*. [1] network the authors propose the generative network which

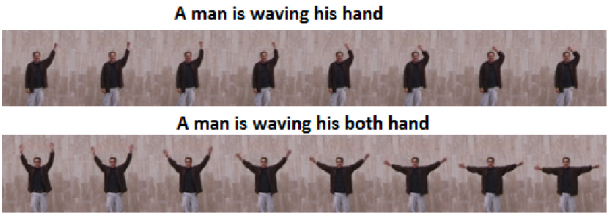

Figure 1: output for TTVGAN Framework for sentence (1) A man is waving his hand and (2) A man is waving his both hands

is a combination of "Conditional GAN" Odena *et al.* [5] and "CVAE (Conditional Variational Autoencoders)" Kingma *et al.* [4]. By fusing this both network authors achieve the visually correct images with as high resolution as 256 x 256. In this network authors use the condition as an extra parameter to force the network to learn only one particular class at a time. Recently a few attempts were also made to generate the video using GAN (Generative Adversarial Network) which are Generating videos with scene dynamics [7] Carl Vonrick *et al.*, and MoCoGAN [6] Sergey Tulyakov *et al.* In [6] authors propose the network in which they achieve the good variable.length video clip using the decomposition of motion and content from the video clips of train data. So, they proposed the framework which generates the video frames sequentially in time with help of deep generative network to produce a random vector of an image. Now Random vector is the important part in their network. For motion sampling, they used the recurrent neural network.

1.2 Video Generation Challenges

Transit from a generation of an image to generation of video using neural network is quite a difficult job though the generated video data has only one extra dimension which is time.Because of this extra time parameter, it's harder to achieve output video sequence to be same as input sequences. Challenges with video generation are, 1) video is a saptio-temporal element and it includes various objects, people, and different action etc.so,generative model should learn the physical representation of the objects as well as appearance of the same and if model fail to learn properly than it may happen that generated video is messed up in terms of action or objects.2) Time dimension of the video comes with significant amount of variation e.g. when person is walking on the forest we need to keep in mind the context and speed of walking of the person at every move the tree should move in back direction when person move in forward direction or when we want to show the conversation between the human it's more difficult than previous example.3) most important thing is we humans are sensitive to motions very much so to generate visually convincing video is a big challenge.

1.3 Introduction to Our Work TTVGAN

With Kept all above in mind we propose a TTVGAN network which will generate the robust video clip from the small text description given as input. To achieve this we use all above discuss concepts as a base. In our work first, we create the sentence understanding network to understand the input sentence and creation of condition based on that. This network will

output an SVO Triplet (Sub. Obj. Verb) as well as sentence context and scene understanding vectors. Now, this input will be used as a conditional parameter to train the GAN network for image generation based on [1]. To generate video we use the concept proposed in [6] and replace the image and video generative model to the more robust framework like CVAE-GAN. With this, we use the GRU (Gated Recurrent Unit) to add the random noise in each output frame to generate next output frame from that. We use the two datasets KTH Human Action dataset and one Custom made Animation dataset to evaluate the performance of the network we created. Through this evaluation, we observe that the proposed network is able to generate visually correct and very good resolution video from the given input sentence. With this, our network is also configurable for different input contexts as well e.g. the network can be configured for animated story generation by providing training data of characters, objects, and the scene in the story. The network will adopt it easily and even it will adopt the sentence understanding for new input scenario.

1.4 Related Work

Over Past few Years, video blogs as well as showing Video-based content over Textual content is rising. This happens because audio-visual content will convey more information with less effort and effectiveness. But video creation is not the ease job, it will require the creativity to convey the right message to the right audience effectively. Some of the platforms on the internet are doing this it can be classified into two categories 1) Algorithm based: video generation using algorithmic logic and programming 2) AI based: video generation using AI-based technology.

In algorithm based platform some examples are: Animoto, Shakr, Typito, Adobe Spark whereas in AI Based platform some examples are: GilaStudio [2] and Wibbitz [8].

GLIA Studio

It is an AI Based Platform.GliaStudio [2] is developed by Taiwan based AI company and it is proprietary platform of the company. This tool is made for advertising and to convert news articles into short video stories. The approach used by the company includes the following: Content Analysis, Material Mapping, Video Creation, and Video Optimization for social media sharing and marketing.

Wibbitz

It is an AI Based Platform.Wibbitz [8] uses the similar approach like GliaStudio [2] to generate videos. But here system will recommend different videos to the user as well as user can search based on a story after that platform will help user to create a video in very less time. Even user can add human voice over in video.

2 Database Creation

We created a custom dataset for our work. In this dataset, we downloaded the various animated videos from the YouTube as well as GIF from various website performing actions like walk, run, jump, and cycling. We collect around 200 videos of various length after this, all videos are trimmed to 3 sec at 30 FPS and arranged side by side so we get 18,400 solo

frames for Image network and 200 videos converted to images 200 x 90 x image-size for video network in total and one sentence pattern dataset to learn various content and scene from the given input sentence.

As stated problem is video generation but our underlying architecture is based on GAN which contains the multiple CNN networks and CNN network can able to process images only not the video. So, to convert the input video into CNN understandable rank-4 matrix we used the OpenCV and extract all frames of the video. To create the input set for training we follow the below described process: 1)Load the input video using OpenCV 2)Extract each frame of video and saved into list 3)Two different variable is crated one for set of images (Rank-4 Matrix) and another for videos (Rank- 5 matrix). 4)Image matrix contains the whole set of frames whereas Video Matrix contains only set of sampled frame for each video.

3 The TTVGAN Framework

The TTVGAN network generates the video from the text input which involves mainly an action, actor and a scene. Many phases are involved in this process. It is being divided into six different phases and each phase plays its own important role in the process. Without one phase the other would be irrelevant. Mainly the input sentence to be processed to identify the action, actor and scene involved and based on this information the GAN network has to be used to generate the video as described. In this section first we discuss the approach which we use for solving the problem of text to video generation after that we will discuss the network architecture and execution part.

3.1 Approach to solve the problem

Below Fig. 2 is describing the approach used in the solving of end-to-end problem. It has gone through the whole process for creating a video. The phases itself gives a vivid idea about the process undertaken behind each action. Major time was spent in dataset creation and fine tuning the model for getting a video in best resolution.

3.1.1 Data Collection

In this section we created the custom dataset with animated character performing action like walking, running, jumping and cycling. For doing this first we downloaded videos from YouTube after that we extracted video frames and arrange the 3sec frames from all video side by side. Currently we have around 200 such videos.

3.1.2 Text Parsing Model

As data collection is over we started with the model for parsing the text input received from user into the form by which training and testing network requires it.Basically while training we require the SVO Triplet as well as context and scene as feature condition vector to the GAN network similar like [1] but fine-tuned and changed according to our requirement. While testing we only require the SVO Triplet for retrieval of video based on that.

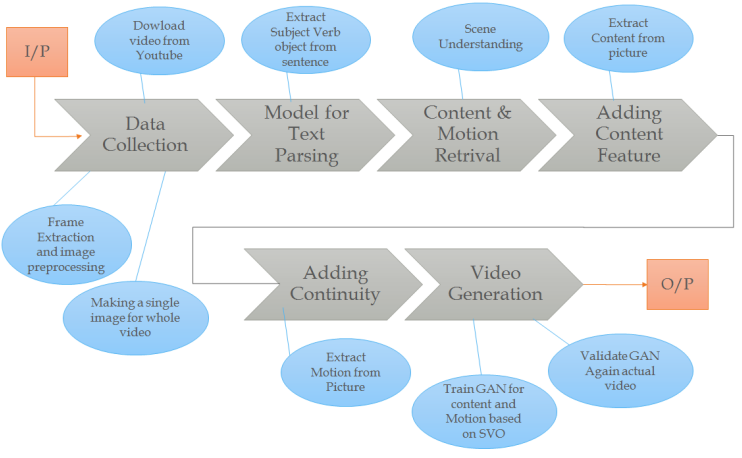

Figure 2: Approach to solve the problem TTVGAN

3.1.3 Content and Motion Retrieval

Here we train the Discriminator part of GAN network separately to understand image representation and video representation separately and created Generator part with the video generator network which also include 2 layer of GRU for random noise generation and sequential frame generation. Basically in this particular part of the work the GAN network is trained to learn: the content (video characters), Action performed by the character over time and scene or context of the video.

3.1.4 Adding Continuity

So, as described in previous section to achieve the continuous video sequence generation of any length we need continuity or time factor as one parameter in the generator network. Here, recurrent neural network (RNN) help us as it is meant to learn the long sequences. Input to the network is randomly sampled video sequences from the original sequences. This help in better understanding of actual video clip at whole.

3.1.5 Video Generation

By reaching this point in flow the model is trained over input sequence to generate the videos. So, the only requirement is to parse the input sentence for testing of a model and after parsing of it we can query the model to generate the video of given input sentence. The model will generate the video mainly based on SVO extracted from the given input sentence. Currently, we are able to generate visually correct videos at 128 x 128 resolution with four classes.

4 Text Parsing Model in detail

The Text Parsing model is the heart of the network without it the whole process will be a mess. To make a text parsing model we used different approach in different phase i.e. text

parsing model is different at the time of training and testing phase. To parse the text and convert it to condition vector which can be given as input to GAN is a challenging task. Here, we use the concept from NLP and Neural Network to build algorithm for training and for testing we have used the simple SVO extraction using NLP.

```
## extraction of subject object verb ##

  for chunk in doc.noun_chunks:
    if (chunk.root.dep_ == "nsubj"):
      sub = chunk.text
      verb = chunk.root.head.text

    elif (chunk.root.dep_ == "conj"):
      secondary_sub = chunk.text

    elif (chunk.root.dep_ == "pobj" or chunk.root.
        dep_  == "pobj"):
      obj = chunk.text
    remove_unnecessary(sub,verb,obj);
```

Figure 3: Pseudo code for text parsing model

The algorithm used during training phase is set to output a condition vector $<$ S, V, O, C, S_{ce} $>$, here S = Subject, V = Action, O = Object set, C = Context Vector, S_{ce} = Scene features. This algorithm is a combination of NLP Techniques and Neural Network. Here S, V, and O are retrieved using the NLP Dependency parser called spaCy [3] and C and S_{ce} is retrieved using classification of sentence into two different classes namely context and scene. In Fig. 3 the pseduo is defined for extraction of SVO. basically in pseudo code, we extracted the noun-verb-noun pairs and Using spaCy [3] we retrive the subject, action and object with tags endings subj, obj, and root.

To extract C and S_{ce} from given input sentence we are classifying given sentence with two predictor variable context and scene.Created the database of various possible context and scene features with appropriate sentence pattern than this sentence feature will be converted to word-vectors using word2vec technique to build classifier model on top of it. We built the randomforest classifier on top of it to classify context and scene separately. As an output of both process we get the vector < S, V, O, C, Sce > described above. Now, this vector is used during the training process.

5 TTVGAN Network Architecture

TTVGAN network architecture is based on GAN network but it is not a single network. It is a combination of 5 different networks which are Recurrent neural network R_n, Image Generative Network G_I, Image Discriminative Network D_I, Video Discriminator Network D_V and Text Parsing Network T_C. Image Generator Network G_I is the output network of the whole framework. So, it generates the Image-sequence by sequentially mapping the Context Feature vector (describe in above section) to latent space of videos and to add time factor in each sequential frame GRU is used. As a result we get correct video as output. During

Training we tried to lower the image difference between generated frame and actual frame. Architecture is described below:

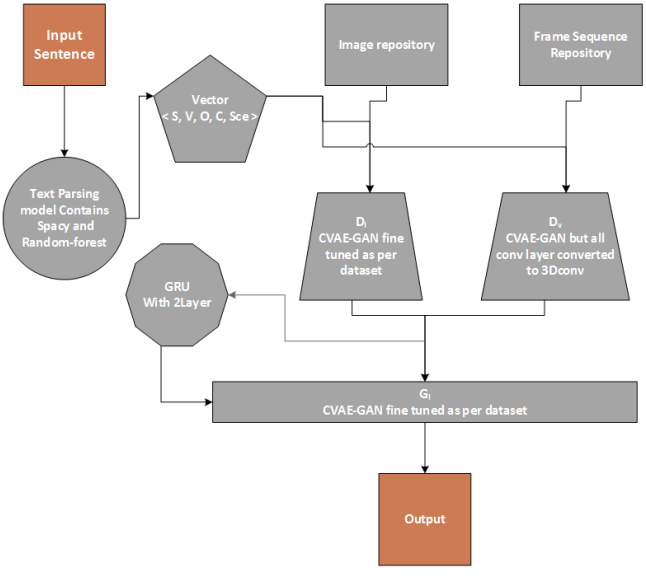

Figure 4: TTVGAN Nework Architecture

as described in Fig. 4 there are multiple networks in the architecture. Here, I am describing significance of each in brief. Flow start from the *InputSentence* block this must be given by the user. Now, this input will be parsed in *TextParsing* block and output is condition feature vector. *ImageRepository* contains all frames from the video it is a matrix of shape (n,h,w,c) where n = number of images, h= height, w = width, and c= channel. *VideoRepository* contains the video sequence in terms of image matrices shape of this matrix is (m,n,h,w,c) where m = number of frame sampled rest are same. Now, this input will go to discriminator networks. Image Discriminator network D_I will take input of Feature Condition vector and Images from the image repository and Video Discriminator network D_V take input from Feature Condition Vector and Video sequence from video repository. Now output of this network will be used for authentication when Image Generator network G_I will generate the image with this GRU will also use as input to add time delta.

Image Discriminator D_I

Image Discriminator network is inspired by the CVAE-GAN [1]. So CVAE-GAN [1] is a combination of VAE and Conditional GAN. In our work, we use it for learning the image representation based on action class, context, and scene coming from Feature Condition vector. In our work Image Discriminator's role is to learn the features of each action class with this we have scene and context feature in terms of latent space which will be given as input to VAE part of CVAE-GAN. By making this kind of GAN network we are able to generate the accurate images for the class given in condition i.e. It will learn as per Feature

condition vector to authenticate generated high-quality images in terms of GAN with the resolution of 128 x 128 or 256 x 256.

Video Discriminator D_V

Video Discriminator network is inspired by the spatio temporal CNN learning.Here, we are using 3D Convolution to learn the extra *Z* dimension in terms of time representation of the video. We have created an extra parameter to sample n frames from the video at a time and then we learn all the frames at a time. This process is carried out for learning the representation of that particular video sequence. The 3_{rd} dimension is important to learn for feature learning of video frames over time in an accurate way.

Image Generator G_I and role of GRU

Image generator network is also taken for CVAE-GAN generator network but here we changed it to allow random noise from the GRU layers. The task of Generator network is to generate the images. Here, for video generation still type of images are not at all useful we require the moving images which should change with the time, therefore, GRU network comes in to picture. Gated Recurrent Unit is a type of RNN.This is best fit to learning the time part of the video frames so input to GRU is features learned from the D_V network and GRU will generate the noise part or time difference part to be used for next frame generation in Image generator network.

6 Results

We conducted experiments to evaluate the proposed framework. In order to evaluate the work we check the ability of proposed framework in 1) generating videos of the same object and same scene as describe in input sentence 2) Quality of the video generated. Evaluating generative models, especially on generating visual outputs, is a challenging task, since all the popular metrics are subject to flaws. Hence, we reported experiment results on the datasets where we could use reliable performance metrics.

we used the following dataset in our experiment:

- **Custom Animation Dataset**: Custom dataset made from various Youtube videos described.
- **Human action video generation**: We used the Weizmann Action database, containing 81 videos of 9 people performing 9 actions including jumping-jack and waving-hands. We resized the videos to have a resolution of 96 x 96.

Fig. 5 is the visualization result here we are showing the 24 frames of output video side by side as video can't be added to PDF:

7 Evaluation of result

To evaluate our work we conducted the experiment with random user at our organization. We gave them the video output and associated sentence with it now task of user is to label it as 0 or 1. 0 for wrong output and 1 for correct output. We experimented with 5 user with 20

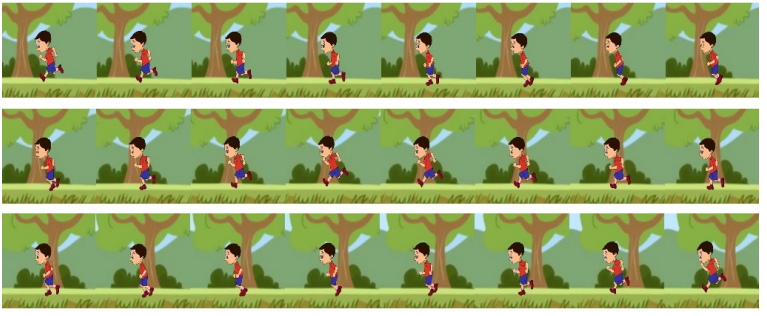

Figure 5: Output for sentence: A boy is running in jungle

videos to test. We set the system in such a way that it will not generate the output for unseen sentence. Currently we have 4 action class walking, running, jumping, and cycling based on it below are the results:

Sentence	User - 1	User - 2	User - 3	User - 4	User - 5
a man is walking	1	1	1	1	1
a girl is running	1	0	1	0	1
a old man is jumping with rope	1	1	1	0	0
a woman is cycling	1	0	0	0	1
a man is cycling on road	1	1	1	1	0
a girls is walking	1	1	1	1	1
a man is cycling on road	1	1	1	0	1
a man is running in jungle	1	1	1	0	0
a boy is running on seashore	1	0	0	1	1
a boy is jumping	0	1	0	0	1
an old man is walking with stick	1	1	1	0	1
an old man is walking	1	1	1	1	1
an old woman is walking	1	1	1	1	1
an old man is running in street	0	1	0	0	1
an woman is walking in park	0	1	1	0	0
a man is jumping	1	1	1	1	1
a man is running	0	1	0	1	1
a woman is running on the road	0	1	1	0	0
a young girl is walking on road	1	1	1	0	1
a old man is cycling in street	1	1	0	0	0
	Total Example		Correct		Correct %
Class Walking	7		6		85.71429
Class Running	6		4		66.66667
Class Jumping	3		2		66.66667
Class Cycling	4		2		50
			Avg Correct		67.2619
Thresold is set to 60% i.e. if 3 out 5 user says it is correct means the video is ok as per sentence					

Table 1: Result Evaluation sheet

8 Conclusions & Future Enhancement

In this research work, we presented the TTVGAN framework for video generation by using a simple input sentence. TTVGAN is able to learn the sentence representation required for the Generative model and able to generate the video with any number of frames. This will enable the new type of generative network which is a fuse of Recurrent network and Generative network.

In future, this network should be able to learn the complex action and scene. To achieve this we need to create robust discriminators in the network as well as we can add one more separate network to learn the various types of complexity. So if we achieve this we can create the videos for multiple use cases like kids story animation, video news from the script,advertisement etc.

References

[1] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: Fine-grained image generation through asymmetric training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2745–2754, 2017.

[2] GLIACLOUD. GLIA Studio working description, 2015. URL https://www.gliacloud.com/en/gliastudio/.

[3] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, 2015.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[5] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pages 2642–2651, 2017.

[6] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. *arXiv preprint arXiv:1707.04993*, 2017.

[7] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016.

[8] Wibbitz. Wibbitz home page, 2015. URL http://www.wibbitz.com/.

[9] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE Int. Conf. Comput. Vision (ICCV)*, pages 5907–5915, 2017.