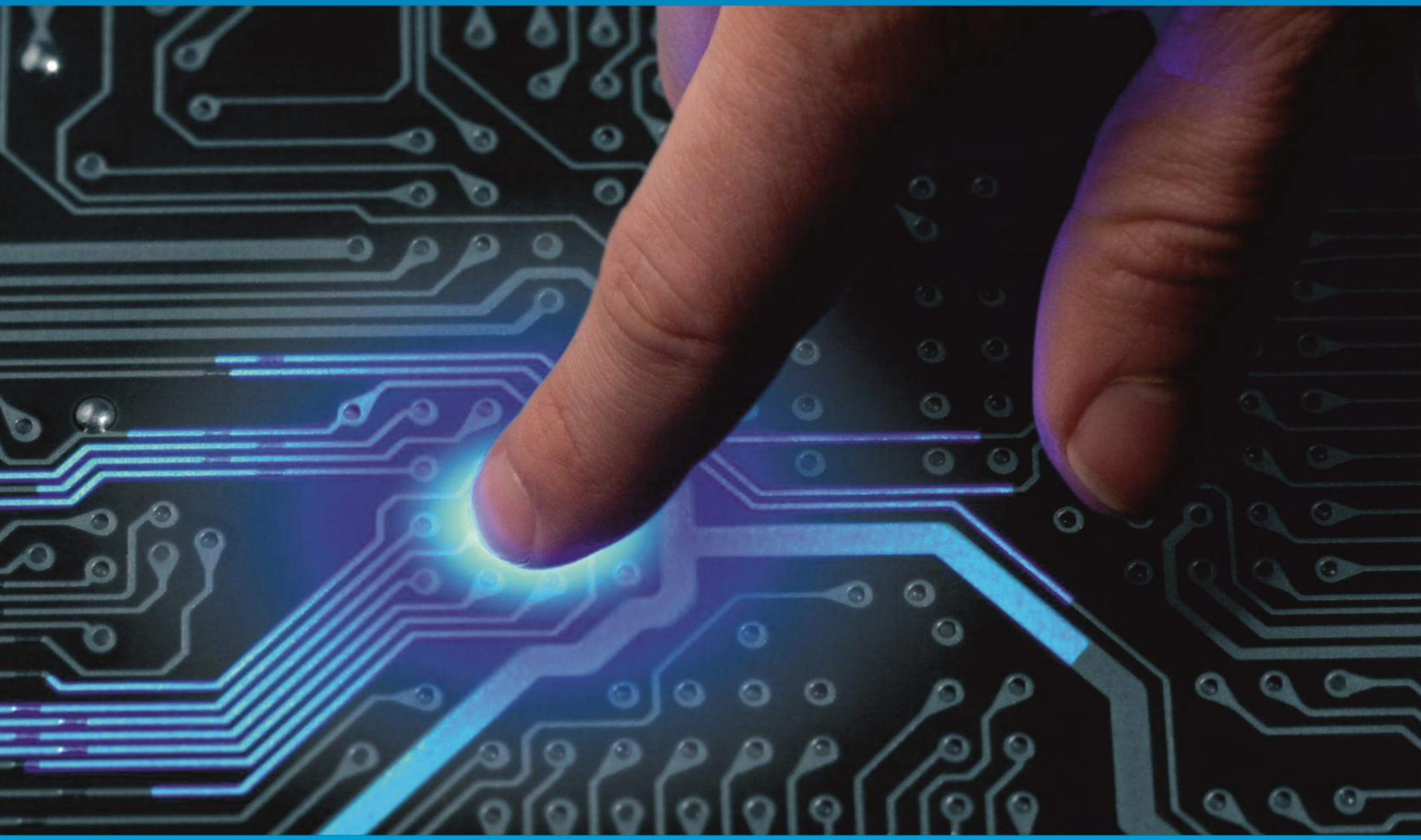
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 4, April 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

9940 572 462

6381 907 438

ijircce@gmail.com

www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904166 |

Scripted Video Generation Using Text

Teena Varma[1], Vishakha Naik [2], Revati Nashte [3]

Assistant Professor, Department of Computer Engineering, Xavier Institute of Engineering, Mahim, Mumbai, Maharashtra, India[1]

Student, Department of Computer Engineering, Xavier Institute of Engineering, Mahim, Mumbai, Maharashtra, India[2]

Student, Department of Computer Engineering, Xavier Institute of Engineering, Mahim, Mumbai, Maharashtra, India[3]

ABSTRACT: Generating movies from textual content has validated to be a tremendous project for present generative fashions. We address this problem through schooling a conditional generative version to extract each static and dynamic records from textual content. This is manifested in a hybrid framework, using a Variational Autoencoder (VAE) and a Generative Adversarial Network (GAN). The static features, called "gist," are used to comic strip textual content-conditioned heritage shadeation and item format structure. Dynamic features are taken into consideration through remodeling enter textual content into an photo filter. To attain a huge quantity of statistics for schooling the deep-learning version, we increase a technique to robotically create a matched textual content-video corpus from publicly to be had on-line movies. Experimental effects display that the proposed framework generates manageable and numerous movies, even as correctly reflecting the enter textual content records. It notably outperforms baseline fashions that without delay adapt textual content-to-photo technology procedures to provide movies. Performance is evaluated each visually and through adapting the inception rating used to assess photo technology in GANs.

KEYWORDS: Generative Adversarial networks, Video generation, Sementic alignment, Temporal coherence, Video captioning .

I. INTRODUCTION

Vision is one of the very important ways in which people experience, interacts with, understand, and learn about the world around them. Intelligent machine systems that can generate scripted videos for human users has tremendous potential application, such as video editing, video games, and computerized design. Unfortunately, many modern and creative works are now generated or edited with the help of digital graphic design tools. The complexity of such tools may lead to inaccessibility issues, particularly for the people with insufficient technical knowledge or resources. That's why, a system that has the ability to follow the speech- or the text-based instructions and after it perform a corresponding video editing method could improve accessibility substantially. These benefits can easily extend to other domains of video generation such as gaming, animation, creating visual teaching material, etc. In this paper, we take a step in this exciting research direction by the text to video generation task. Specifically, we focus on video generation from text, which aims to generate a video semantically aligned with some given descriptive scripts. The video generation task is much more difficult since the video is often a complex sequence of many frames which should follow strong spatial and temporal dependencies. More critically, generating video conditioned on given text is even more complicated due to the requirement of semantic alignment between video and text at both frame and video levels. Thus, although there are already a lot of existing models for text-to-image generation, simply using the image generator to synthesise videos may incur poor performance . In scripted video generation from text, there are two imprtant challenges: 1) semantic alignment between given text and video content; 2) realistic video generation with temporal coherence across frames.

II. RELATED WORK

A. Video Generation

A spatial-temporal 3-D deconvolution-primarily based totally GANs is seasoned posed for unconditioned video era in To analyze the semantic illustration of unlabeled motion pictures, Saito et al. layout extraordinary generators (a temporal generator and an photograph generator), which sequentially remodel a unmarried latent variable right into a video. Tulyakov et al. advocate a framework to generate video with the aid of using decomposing movement and content material in an unmonitored manner. On the opposite hand, many programs consciousness on producing video conditioned on a sta ble photograph (body), such as. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are broadly used and gain the remarkable overall performance in those programs . The not unusualplace thread of those works is to apply CNN for encoding every body after which to use an series-to-series version for body prediction. Instead of unconditioned or photograph conditioned motion pictures era, we consciousness on producing motion pictures conditioned on text, that is greater hard because of the requirement of semantic alignment among motion pictures and herbal languages.

B. Video Generation Conditioned on Caption

Video technology conditioned on textual content ambitions to synthesise a video that is semantically aligned with the given descriptive sentence, together with a caption or script. Mittal et al. devise a way to generate video from textual content via way of means of combining a Variational Autoencoder with a Recurrent Attention Mechanism, which captures the temporally established collection of frames. Marwah et al. advocate an advanced version to incorporate the long-time period and short-time period dependencies among frames and generate video in a incremental manner. Most recently, Li et al. advocate a two-level VAE-primarily based totally generator to generate a 'gist' of the video from enter textual content, in which the gist is an photograph that offers the historical past shade and object layout. The content material and movement of the video are then generated via way of means of conditioning at the gist. Meanwhile, because of the achievement of Generative Adversarial Networks (GANs) , Pan et al. do not forget the temporal coherence throughout frames and the semantic matching among textual content the complete video with a cautiously designed discriminator. Besides the body coherence and video-textual content semantic matching taken into consideration in the above methods, we additionally expand a bottom-up mechanism to make certain the spatial-temporal coherence and semantic matching among textual content and video at a couple of levels, such as region, body and video.

C. Conditional Generative Networks

Two of the maximum famous deep generative fashions are the Variational Autoencoder (VAE) (Kingma and Welling 2013) and the Generative Adversarial Network (GAN) (Goodfellow et al. 2014). A VAE is discovered with the aid of using maximizing the variational decrease sure of the remark even as encouraging the approximate (variational) posterior distribution of the hidden latent variables to be near the earlier distribution. The GAN framework is based on a minimax sport among a "generator" and a "discriminator." The generator synthesizes information while the discriminator seeks to differentiate among actual and generated information. In multi-modal situations, GAN empirically indicates benefits over the VAE framework (Goodfellow et al. 2014). In order to construct relationships among textual content and videos, it is important to construct conditionally generative fashions, which has obtained enormous current attention. In particular, (Mirza and Osindero 2014) proposed a conditional GAN version for textual content-to-picture generation. The conditional information turned into given to each the generator and discriminator with the aid of using concatenating a function vector to the enter and the generated picture. Conditional generative fashions had been extended in numerous directions. (Elman Mansimov and Salakhutdinov 2016) generates photos from captions with an RNN version using "attention" at the textual content. (Liu and Tuzel 2016; Zhu et al. 2017) proposed conditional GAN fashions for either fashion or area switch learning. However, those strategies targeted on switch from picture to picture. Converting those strategies to use to textual content and picture pairs is non-trivial.

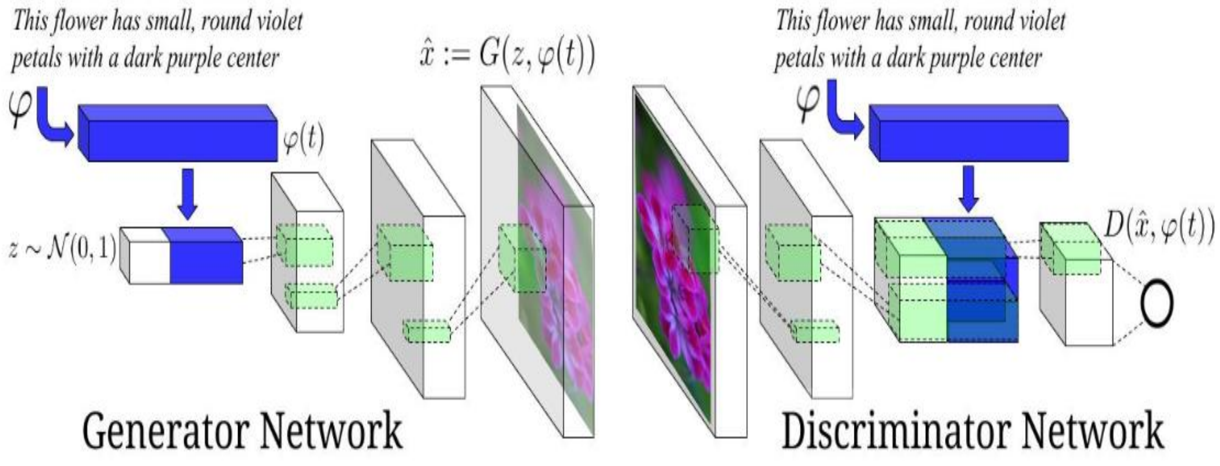

This flower has small, round violet petals with a dark purple center
$\hat{x} := G(z, \varphi(t))$
This flower has small, round violet petals with a dark purple center
φ
$\varphi(t)$
φ
$z \sim \mathcal{N}(0,1)$
$D(\hat{x}, \varphi(t))$
Generator Network
Discriminator Network

Fig. 1.This fig shows how Generator Network and Discriminator Network Works.

The principal intention of our Temporal GANs conditioning on Captions (TGANs-C) is to layout a generative version with the cappotential of synthesizing a temporal coherent body sequence semantically aligned with the given caption. The education of TGANs-C is achieved through optimizing the generator network and discriminator network (video and body discriminators which concurrently decide artificial or actual and semantically mismatched or matched with the caption for video and body) in a -participant minimax recreation mechanism. Moreover, the temporal coherence earlier is moreover integrated into TGANs-C to provide temporally coherent body sequence in exclusive schemes. Therefore, the general objective feature of TGANs-C consists of 3 components, i.e., video-degree matching-conscious loss to accurate the label of actual or artificial video and align video with matched caption, body-degree matching-conscious loss to similarly beautify the image truth and semantic alignment with the conditioning caption for everybody, and temporal coherence loss (i.e., temporal coherence constraint loss/temporal coherence adverse loss) to take advantage of the temporal coherence among consecutive frames in unconditional/conditional scheme.

A. Generative Adversarial Networks

The fundamental generative opposed networks (GANs) consists of networks: a generator community G that captures the facts distribution for synthesizing picture and a discriminator community D that distinguishes actual photos from artificial ones. In particular, the generator community G takes a latent variable z randomly sampled from a ordinary distribution as enter and produces a artificial picture $xsyn = G$ (z). The discriminator community D takes an picture x as enter stochastically chosen (with identical possibility) from actual photos or artificial ones via G and produces a possibility distribution $P\ (S|x) = D\ (x)$ over the 2 picture sources (i.e., artificial or actual). As proposed in [5], the entire GANs may be educated in a -participant minimax game. Concretely, given an picture instance x, the discriminator community D is educated to decrease the opposed loss, i.e., maximizing the log-likelihood of assigning accurate supply to this instance: $la(x) = -I(S{=}real)\ \log\ (P(S = real|x)) -(1 - I(S{=}real))\ \log\ (1 - P(S = real|x))$, (1) wherein the indicator feature Icondition = 1 if circumstance is true; in any other case Icondition = 0. Meanwhile, the generator community G is educated to maximise the opposed loss in Eq.(1), targeting for maximally fooling the discriminator community D with its generated artificial photos

B. Conditional Generative Network

Two of the maximum famous deep generative fashions are the Variational Autoencoder (VAE) (Kingma and Welling 2013) and the Generative Adversarial Network (GAN) (Goodfellow et al. 2014). A VAE is found out with the aid of using maximizing the variational decrease sure of the remark at the same time as encouraging the approximate (variational) posterior distribution of the hidden latent variables to be near the earlier distribution. The GAN framework is based on a minimax sport among a “generator” and a “discriminator.” The generator synthesizes information while the discriminator seeks to differentiate among actual and generated information. In multi-modal situations, GAN empirically indicates benefits over the VAE framework (Goodfellow et al. 2014). In order to construct relationships among textual content and videos, it is important to construct conditionally generative fashions, which has obtained great current attention. In particular, (Mirza and Osindero 2014) proposed a conditional GAN version for textual content-to-picture generation. The conditional information become given to each the generator and discriminator with the aid of using concatenating a characteristic vector to the enter and the generated picture. Conditional generative fashions had been extended in numerous directions. (Elman Mansimov and Salakhutdinov 2016) generates pics from captions with an RNN version using “attention” at the textual content. (Liu and Tuzel 2016; Zhu et al. 2017) proposed conditional GAN fashions for either fashion or area switch learning. However, those techniques targeted on switch from picture to picture. Converting those techniques to use to textual content and picture pairs is non-trivial.

III. Pseudo code

Algorithm 1 The education of Temporal GANs conditioning on Captions (TGANs-C)

```
Step 1: Given the variety of most education new release $T$.
Step 2: for $t = 1$ to $T$ do
Step 3: Fetch enter batch with sampled video-sentence pairs $\{(\mathcal{S}, vreal+)\}$.
Step 4: for Each video-sentence pair $(\mathcal{S}, vreal+)$ do
Step 5: Get the random noise variable z $\sim \mathcal{N}(0, 1)$.
Step 6: Produce the artificial video $vsyn+ = G$ (z, S) circumstance on the caption $\mathcal{S}$ throgh the generator community $G$.
Step 7: Randomly pick out one actual video $vreal-$ defined with the aid of using a distinctive caption from $\mathcal{S}$.
Step 8: Give up for
```

Step 9: Obtain all the actual-artificial tuple $\{vsyn+, vreal+, vreal-\}$ with the corresponding caption $\mathcal{S}$, denoted as $\mathcal{T}$ in total.
Step 10: Compute video-degree matching-conscious loss through Eq. (3).
Step 11: Compute frame-degree matching-conscious loss through Eq. (4).
Step 12: -Scheme 1: TGANs-C-C
Step 13: Compute temporal coherence constraint loss through Eq. (6).
Step 14: Update the discriminator community D w.r.t loss in Eq. (8).
Step 15: Update the generator community G w.r.t loss in Eq. (10).
Step 16: -Scheme 2: TGANs-C-A
Step 17: Compute temporal coherence hostile loss through Eq. (7).
Step 18: Update the discriminator community D w.r.t loss in Eq. (9).
Step 19: Update the generator community G w.r.t loss in Eq. (11).
Step 20: Give up for

IV. Simulation Results

We examine and evaluate our proposed BoGAN version with numerous trendy approaches, on synthetic text-to-video datasets and real-global datasets , with both quantitative and qualitative assessment metrics. A detailed ablation examine is accomplished to check the contribution of each version aspect and a human examine is accomplished to examine the reality, relevance and coherence of the generated videos. We ultimately examine the generalisation cappotential of our proposed version.

1. Dataset Creation
   Because there's no popular publicly to be had textual content-to-video era dataset, we advocate a manner to down load motion pictures with matching textual content description. This approach is comparable inconcept to the approach in that became used to create a large-scale video-type dataset. Retrieving huge numbers of motion pictures from YouTube is easy; however, computerized curation of this dataset isn't always as straightforward. The data-series procedure we've taken into consideration proceeds as follows. For every keyword, we first collected a hard and fast of motion pictures collectively with their identify, description, length and tags from YouTube. The dataset became then wiped clean with the aid of using outlier-elimination techniques. Specifically, the approach of became used to get the ten maximum frequent tags for the set of video. The fine of the chosen tags is similarly assured with the aid of using matching them to the phrases in current classes in ImageNet and ActionBank . These datasets help make sure that the chosen tags have visually detectable objects and actions. Only motion pictures with at the least 3 of the chosen tags have been included. Other necessities include (i) the length of the video ought to be in the variety of 10 to 400 seconds, (ii) the identify and outline ought to be in English, and (iii) the identify ought to have greater than 4 meaningful phrases after doing away with numbers and forestall phrases

2. Video Processing
   Current video-technology strategies simplest cope with clean dynamic changes. A unexpected alternate of shot or fast-changing history introduces complicated non-linearities between frames, inflicting present fashions to fail. Therefore, every video is reduce and simplest certified clips are used for the training The clips have been certified as follows. Each video makes use of a sampling price of 25 frames in step with second. SIFT key factors are extracted for every frame, and the RANSAC set of rules determines whether or not continuous frames have sufficient key-factor overlap. This step guarantees clean motions withinside the history and objects.

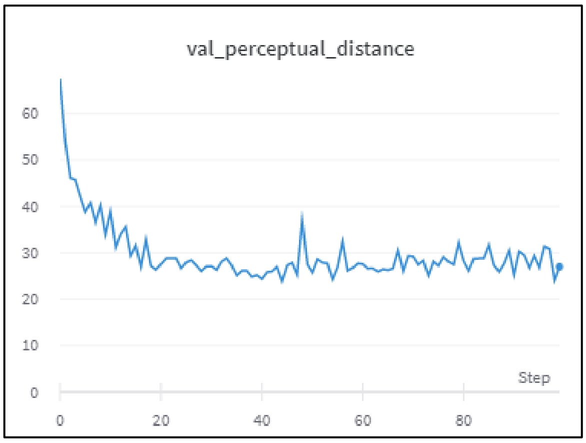


Fig.2. Value Perceptual distance between frames.

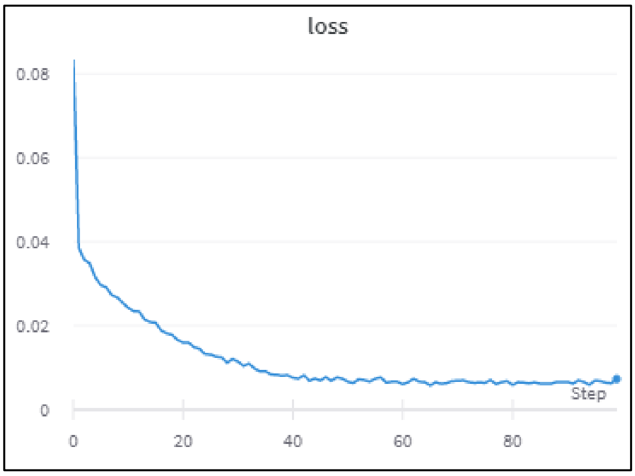


Fig. 3. Loss for Video Frame Prediction

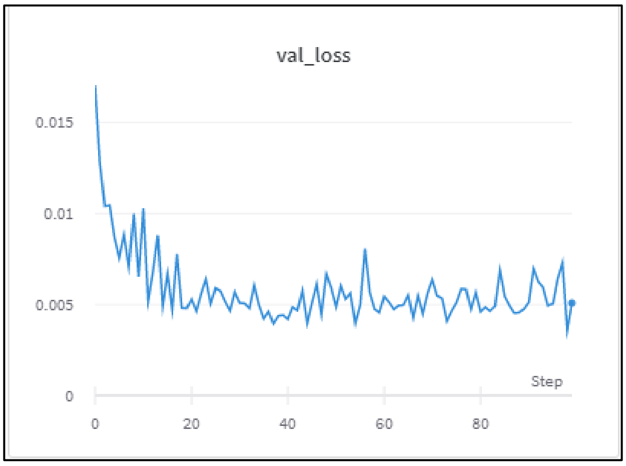


Fig. 4. Value Loss in Frames

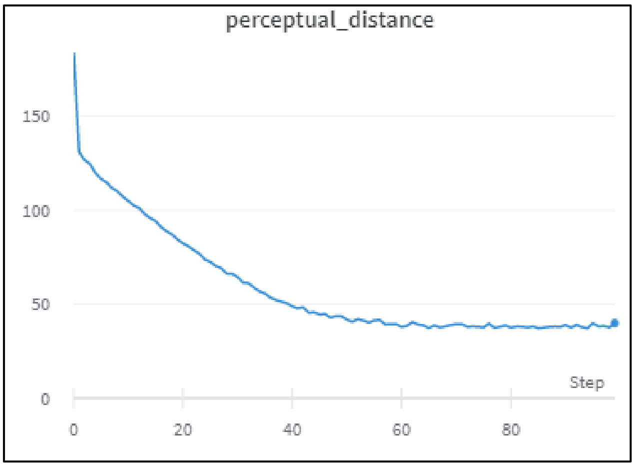


Fig 5. Perceptual Distance in Frame

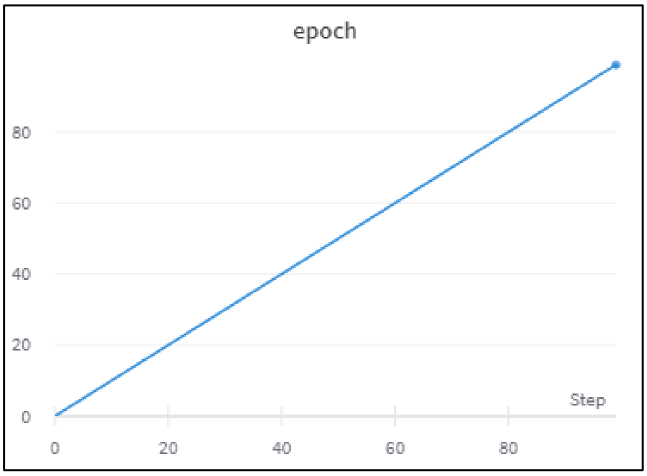


Fig. 6.Epoch in between the video Frames

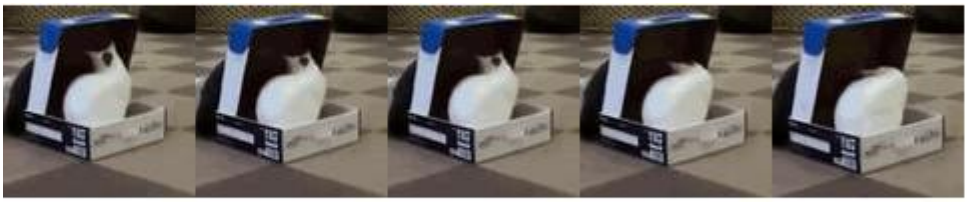


Fig.7.Cat is hiding inside the box.

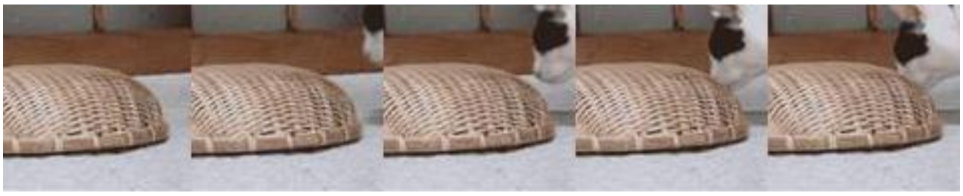


Fig. 8.Cat is trying to smell.

V. CONCLUSION AND FUTURE WORK

Video technology from textual content is difficult because of the intrinsic complexity. In this paper, we've proposed a unique BottomUp Generative Adversarial Network (BoGAN) to make certain the realism of the generated video and attain the specified multiscale semantic alignment. Specifically, to make certain the coherence among generated frames and the semantic healthy among a video and a language description, we've devised a bottom-up optimisation mechanism that consists of 3 levels, from local to global. The proposed approach outperforms its competitors at the benchmarks, which demonstrates the electricity of the structure we've described. In phrases of human study, our proposed approach additionally plays higher than the competitors, that is a long way extra indicative of the cost of our approach.

This paper proposes a framework for producing video from textual content the use of a hybrid VAE-GAN framework. To the high-satisfactory of our knowledge, this paintings proposes the primary a hit framework for video era from textual content. The intermediate gistgeneration step significantly allows implement the static background of video from enter textual content. The proposed Text2Filter allows seize dynamic movement data from textual content. In the future, we plan to construct a extra effective video generator with the aid of using producing human pose or skeleton features, a good way to similarly improve the visible first-rate of generated human activity videos.

REFERENCES

1. S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” 2016, arXiv:1605.05396. [Online]. Available: http://arxiv.org/abs/1605.05396
2. Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K. 2015. Sequence to sequence-video to text. In IEEE ICCV.
3. Vondrick, C., and Torralba, A. 2017. Generating the future with adversarial transformers. In CVPR..
4. DilipKumar S. M. and Vijaya Kumar B. P. ,’Energy-Aware Multicast Routing in MANETs: A Genetic Algorithm Approach’, *International Journal of* Computer *Science and Information Security* (IJCSIS), Vol. 2, 2009.
5. Y. Pan, Z. Qiu, T. Yao, H. Li, and T. Mei, “To create what you tell: Generating videos from captions,” in Proc. ACM Multimedia Conf. (MM), 2017, pp. 1789–1798.
6. T.-C. Wang et al., “Video-to-video synthesis,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 1–14.
7. C. Vondrick, H. Pirsiavash, and A. Torralba, “Generating videos with scene dynamics,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 613–621.
8. Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.
9. C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, “Highresolution image inpainting using multi-scale neural patch synthesis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6721–6729.
10. Ye, G.; Li, Y.; Xu, H.; Liu, D.; and Chang, S.-F. 2015. Eventnet: A large scale structured concept library for complex event detection in video. In ACM Int. Conf. on Multimedia
11. Walker, J.; Doersch, C.; Gupta, A.; and Hebert, M. 2016. An uncertain future: Forecasting from static images using variational autoencoders. In ECCV.

12. N. C. Rakotonirina and A. Rasoanaivo, “ESRGAN+: Further improving enhanced super-resolution generative adversarial network,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 3637–3641.

BIOGRAPHY

Vishakha Vijay Naik, Student, Completed Diploma in Information Technology from Thakur Polytechnic, Mumbai, India in academic year 2017-18 .Currently pursuing BE (Bachelor Of Engineering and Technology) in Computer Engineering from Xavier Institute Of Engineering (affiliated to Mumbai University)., Mumbai, India.

Revati Vijay Nashte, Student, Completed Diploma in Information Technology from Goverment Polytechnic, Mumbai, India in academic year 2017-18 .Currently pursuing BE (Bachelor Of Engineering and Technology) in Computer Engineering from Xavier Institute Of Engineering (affiliated to Mumbai University)., Mumbai, India.

Teena Varma, Assistant Professor, Completed B.E. in Electronics, Completed M.E. in Computer Engineering, Working in XIE since 2007, Lecturer in Computer Engineering Department (2007-2009), Assistant Professor in Computer Engineering Department(2009 till now), Mumbai, India.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com

Scan to save the contact details

www.ijircce.com