Conditional Video Generation

Kasra Sehhat
AI Researcher
Sharif University Of Technology, Tehran, Iran
kasra.sehhat@sharif.edu

Abstract

This paper comprises 4 parts which are as follows: the first part is related to the literature review and previous works done in conditional image generation and a brief description of algorithms used generally. In the second part, a method is proposed for solving this challenge. In the third part, the advantages of the proposed method are pointed out. In the fourth part, the estimation of RD has been shown. The fifth part is related to the GitHub link which codes of the second problem of this project are pushed there. Finally, the last part is the list of references used in this study.

1. Review of Previous Works at Conditional Video Generation

Videos are created to express emotion, exchange information, and share experiences. Video synthesis has intrigued researchers for a long time. Despite the rapid progress driven by advances in visual synthesis, most existing studies focus on improving the quality of the frames and the transitions between them, while little progress has been made in generating longer videos. This study was conducted on conditional video generation. The first part is related to a brief literature review. Afterward, an idea was proposed for achieving this goal. Early works, e.g., 2 cDNA [5] and PredRNN [18], leverage deterministic methods to directly predict the next frame via CNNs or RNNs. However, these deterministic models are unable to capture the stochastic temporal patterns and synthesize coherent complex scenes.

Generative models, especially Generative Adversarial Networks [7] (GANs), begin to dominate the area as they can perform unconditional or class-conditional video synthesis without the first frames. VGAN [17] is the first one to use GAN for video generation. It decomposes video to a static background and a moving foreground and then generates them with 2D and 3D convolutional networks respectively. TGAN [13] proposes to separately generate the temporal latent variables and spatial information, and MoCoGAN [14] similarly decomposes the latent space into context and motion subspaces. DIGAN [21] applies implicit neural representations for video encoding. Recently, text-to-video generation emerges as a promising direction. The framework of VQVAE [15] and autoregressive transformers [16] quickly become the mainstream method [19]. Ho et al. [8] proposes a video diffusion model along with a gradient method recently for text-to-video generation. Moreover, most of these models are not publicly available [9].

Autoregressive transformers, e.g., DALL-E [11] and CogView [2], have revolutionized text-to-image generation recently. It is natural to investigate the potential of autoregressive transformers in text-to-video generation. Previous works followed this basic framework [6], e.g., VideoGPT [20], verifying its superiority over GAN-based methods [1], but are still far from satisfactory [9]. This study demonstrates how combining the effectiveness of the inductive bias of CNNs with the expressivity of transformers enables them to model and thereby synthesize high-resolution images. Also, it shows how to (i) use CNNs to learn a context-rich vocabulary of image constituents, and in turn (ii) utilize transformers to efficiently model their composition within high-resolution images [4] This study presents a large-scale pre-trained text-to-video generative model, CogVideo, which is of 9.4 billion parameters and trained on 5.4 million text-video pairs.

We build CogVideo based on a pre-trained text-to-image model, CogView2 [3], to inherit the knowledge learned from the text-image pretraining. We propose multi-frame-rate hierarchical training to ensure the alignment between text and its temporal counterparts in the video [9]. We tackle the problem of long video generation. Building upon the recent advances of VQGAN [10] for high-resolution image generation, we first develop a baseline by extending the 2D-VQGAN to 3D (2D space and 1D time) for modeling videos. This naively extended method, however, fails to produce high-quality, coherent long videos. Our work investigates the model design and identifies simple changes that significantly improve the capability to generate long

videos of thousands of frames without quality degradation when conditioning on no or weak information. Our core insights lie in 1) removing the undesired dependence on time from VQGAN and 2) enabling the transformer to capture long-range temporal dependence [6]. Learning useful representations without supervision remains a key challenge in machine learning. This paper proposes a simple yet powerful generative model that learns such discrete representations. This model, the Vector Quantized-Variational Autoencoder (VQ-VAE), differs from VAEs in two key ways: the encoder network outputs discrete, rather than continuous, codes; and the prior is learned rather than static. To learn a discrete latent representation, this paper incorporates ideas from vector quantization (VQ).

Using the VQ method allows the model to circumvent issues of posterior collapse where the latents are ignored when they are paired with a powerful autoregressive decoder. typically observed in the VAE framework. Pairing these representations with an autoregressive prior, the model can generate high-quality images, videos, and speech as well as doing high-quality speaker conversion and unsupervised learning of phonemes, providing further evidence of the utility of the learned representations [15]. In this study, they scale and enhance the autoregressive priors used in VQ-VAE to generate synthetic samples of much higher coherence and fidelity than possible before. Also, they use simple feed-forward encoder and decoder networks, making our model an attractive candidate for applications where the encoding and/or decoding speed is critical. Additionally, VQ-VAE requires sampling an autoregressive model only in the compressed latent space, which is an order of magnitude faster than sampling in the pixel space, especially for large images. We demonstrate that a multi-scale hierarchical organization of VQ-VAE, augmented with powerful priors over the latent codes, can generate samples with quality that rivals that of state of art Generative Adversarial Networks on multifaceted datasets such as ImageNet, while not suffering from GAN's known shortcomings such as mode collapse and lack of diversity [12].

2. Proposed Method

In this study, we propose a network to generate high-resolution videos created under special conditions which look natural and indistinguishable from real videos. For this purpose, we tried to combine the effectiveness of the inductive bias of CNNs with the expressivity of transformers enabling them to model and thereby synthesize high-resolution video. This study suggests a huge network comprised of two main subnetworks that would be trained in different stages. The first part which is the most important part of the main network is the hierarchical VQGAN. This part of the network is comprised of four subnetworks and two code books. Subnetworks include an encoder, decoder, comparator, and discriminator. The whole of them which constitute VQGAN is used to create a code book and is also used for the discrete representation of the image and to tokenize the image into discrete tokens and the second part which is comprised of the decoder is used to generate high-quality images. The comparator and discriminator are used just in the training phase.

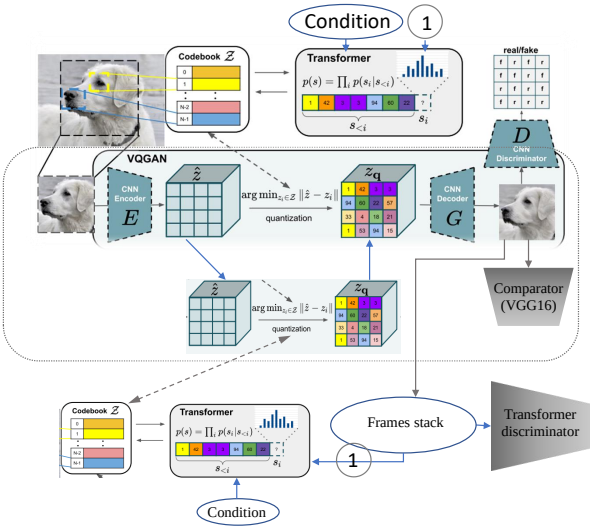

Figure 1. Different parts of network. N successive frames are fed to discriminator transformer to be recognized as real or fake video. Also, they are fed to generator transformers to generate discrete representation of next frame

Also, to get the frames of high quality, we use a hierarchical algorithm that creates code books for a different level of representation. This method leads to preserving features of the image at different levels. Backpropagation through the non-differentiable quantization operation is achieved by a straight-through gradient estimator, which simply copies the gradients from the decoder to the encoder, such that the model and codebook can be trained end-to-end. The generated image by this part will be fed to the comparator and discriminator. Comparator is used to evaluate generated image in terms of perceptual error which is comprised of feature extractor like VGG16. A discriminator is used to recognize each patch of an image as real or fake. After this step, we freeze this vqgan part and there will not be any changes in this network till the end.

The second part of this huge network is comprised of transformers. These networks are designed to generate discrete representations at different levels for the next frame of video. Therefore, the inputs for these transformers would be the condition and previous frames of video. At the beginning of these transformers, we used positional encoders to encode the place of each token. Because in the image construction task each token is affected by surrounding tokens.

To train these transformers, we have data comprised of conditions, previous frames, and also a desire frame wanted to be constructed. At first, we have to feed the desire frame to the vqgan to get the discrete representation of this image at different levels. These discrete representations are considered as outputs of these transformers and the previous frames and condition would be the input of these transformer networks. In such a way, designed transformers learn long-range interactions on sequential data and would be able to generate next from previous frames and mentioned conditions. Up to now there is nothing to check the video if it looks real or not. For this purpose, at the end of the complete network, there is a discriminator in order to recognize created video as real or fake and control the intensity of changes within successive frames. The input of this discriminator would be the five successive frames of generated video. The structure of this discriminator would be the encoder part of transformers. Therefore, we can use pre-trained networks instead of this part and fine-tune it on video frames.

3. Advantage

Advantages of this network include combining the effectiveness of the inductive bias of CNNs with the expressivity of transformers enabling them to model and thereby synthesize high-resolution videos. This approach could be applied to different conditional synthesis tasks, where both non-spatial information and spatial information can be controlled.

4. Timeline

The exact time required for this project completely depends on the accuracy and innovation required for this project. But I estimate just a rough timeline as Table 1.

Table 1. Timeline Table

Task	Required time
Precise literature review	20-30 days
Ideation	10-15 days
Coding and solving challenges	10-15 days
Extra time for unpredicted problem	15 days
Prepare presentation	5 days

5. GitHub Link for Problem 2

https://github.com/kasrasehat/Conditional-video-generation.git

References

[1] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019. 1

[2] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 1

[3] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 1

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1

[5] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016. 1

[6] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. *arXiv preprint arXiv:2204.03638*, 2022. 1, 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[8] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 1

[9] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 1

[10] Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan. Transformation-based adversarial video prediction on large-scale data. *arXiv preprint arXiv:2003.04035*, 2020. 1

[11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[12] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[13] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2839, 2017. 1

[14] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 1526–1535, 2018. 1

[15] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1, 2

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[17] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016. 1

[18] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017. 1

[19] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 1

[20] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 1

[21] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022. 1