자가 학습을 통한 비디오 생성 성능 향상 연구

홍기범, 문철현, 변혜란
연세대학교

(cha2068, cheolhyunmun, hrbyun)@yonsei.ac.kr

A Study on Video Generation with Self-supervised Learning

Hong Kibeom, Mun Cheolhyun, Byun Hyeran
Yonsei Univ.

(cha2068, cheolhyunmun, hrbyun)@yonsei.ac.kr

요 약

Training the video generative models is even more sophisticated than image generative ones due to the complexity of dimension on videos. To solve this problem, we propose the novel framework to generate plausible videos with self-supervised learning. To this end, we introduce the Arrow of Time (AoT) as an additional self-superivsory task with generative adversarial networks. Finally, our networks train not only the video generation but also the direction of video plays. Our methods achieve the improvement of performance for video generation with three baselines.

Ⅰ. Introduction

Though recent studies of image generations produce real-like images with or without guidance, generating a video is more sophisticated than generating a single image since the dimension of videos is more complex. Previous methods [1,2,3] employ the GAN [4] framework for video generation. Nevertheless, they still struggle from the difficulty in generating plausible videos. To overcome this difficulty, we introduce the novel framework that critics the discriminators with better understanding about the properties of video, time. Furthermore, we explore categorical video generation with recent techniques in image generation, namely conditional batch normalization [5], spectral normalization [6], projection discriminator [7], and mode-seeking regularizer [8]. Our extensive experiments demonstrate that our framework improve the video generation performance based on three baselines.

Ⅱ. Method

Our intuition is to teach discriminators and generators to have sense of Arrow of Time (AoT) as humans do. We propose the classifying AoT as an auxiliary task for self-supervised learning and explain how the generators benefit from it.

We first induce our discriminator to distinguish real and fake videos as well as the direction of video plays (i.e., forward, and backward videos). We expect our generator G to learn that the generated videos should run forward in time as an inductive bias. We note that G only produces forward videos, but D receives both forward and backward generated videos by reversal, to encourage them to resemble real videos. We apply our novel framework on previous three baselines [1,2,3].

Furthermore, we add recent techniques in conditional image generation to gain improved performance for categorical video generation. First, we extend conditional batch normalization [5] to embed class label into the generator well. Second, the vanilla discriminator is replaced by the projection discriminator [7]. Lastly, we employ the spectral normalization layers [6] at each layer of discriminators and use mode-seeking loss [8] for stable adversarial training. We utilize three datasets: Weizmann action, UCF-sports and UCF-101 for training and evaluation.

Ⅲ. Conclusion

We proposed the self-supervised video discriminator with arrow of time. Our video discriminator not only critic the generated samples but also learn the properties of videos without additional labeled data. By doing so, our networks generate the various and realistic videos. As a results, our proposed framework improves the performance of video generation over all baselines and all datasets in terms of IS and FVD metrics.

Also, we have succeeded in generating conditional videos based on the projection method to mitigate intra-class mode dropping. Specifically, our categorical video generative method achieves the state-of-the-art IS of 28.97 and the FVD value of 26.68 on UCF 101 which is a benchmark for video generation.

ACKNOWLEDGMENT

"This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2021-2016-0-00464) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)"

Reference

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

[2] 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.

[3] 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.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, DavidWarde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672- 2680, 2014.

[5] Harm De Vries, Florian Strub, J´er´emie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In Advances in Neural Information Processing Systems, pages 6594- 6604, 2017.

[6] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In International Conference on Learning Representations, 2018.

[7] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In International Conference on Learning Representations, 2018.

[8] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1429- 1437, 2019.