ISSN: 2641-2039

Global Journal of Engineering Sciences

DOI: 10.33552/GJES.2022.10.000738

Iris Publishers

Mini Review

A Mini-Review on Current Difficulties of Self-attention Generative Adversarial Networks

Cheng Xu*

School of Computer Science, University College Dublin, Dublin, Ireland

***Corresponding author:** Cheng Xu, School of Computer Science, University College Dublin, Dublin, Ireland.

Received Date: November 11, 2022

Published Date: November 23, 2022

Abstract

With the rapid development of Vision Transformer, its application in the field of Generative Adversarial Networks (GAN) is becoming more and more obvious. However, based on the current research situation in this field, Transformer-based GAN has achieved better performance than traditional convolution-based GAN in some cases, but there is still a clear gap between them. In addition, there are several problems that can be witnessed in the diversity of generated content. In this research, I will review recent research with self-attention generative adversarial model and present some understandings.

Keywords: Transformer; Self-attention; GAN; Generative model

Abbreviations: Generative Adversarial Networks (GAN); Variational Auto-encoder (VAE); Convolutional Neural Network (CNN); Natural Language Processing (NLP); Vision Transformer (ViT); Conditional GAN (CGAN); Wasserstein GAN (WGAN); Multilayer Perceptron (MLP); Fre chet Inception Distance (FID); Inception Score (IS)

Introduction

Generative model and discriminative model are two basic models in machine learning, but the research of generative model is much lower than that of discriminative model, the reason is that the application scene of discriminative model is more abundant, and its effect can be achieved immediately. The research on Generative Adversarial Networks (GAN), one of the most mainstream generative models, has exploded exponentially since it was first brought into the public domain by Ian Goodfellow et al. [1] in the year 2014. The advantage is that there is no need to employ any Markov chains or unrolled approximate inference networks in training and generation process, so it is less expensive to generate a clear image than other methods. The performance of the original GAN was not good enough that the generated content was blurry, but it can be regarded as an alternative generative model in addition to Variational Auto-encoder (VAE) [2,3], PixelCNN [4] and so on.

Since the self-attention mechanism was proposed by [5], it has shone brilliantly in the field of Natural Language Processing (NLP) in the form of Transformer. The solutions, based on autoregressive language modeling in GPT [6-8] and masked autoencoding in BERT [9]. In the field of vision, which is unified by convolutional neural networks [10-13], inspired by the robust development of self-attention in NLP, there is also a lot of work to introduce self-attention to visual tasks [14-17], but the work that makes Transformer favored in the field of vision is [18, 19], the performance of Vision Transformer (ViT) model is better than that of state-of-the-art convolution neural network in some cases [20-22].

Transformer-Based GAN

Generative adversarial networks

Generative Adversarial Networks were developed by Ian Goodfellow, et al. [1] in the year 2014. GANs belong to the class of Generative models [23]. GANs are based on the min-max, zero-sum game theory. For this, GANs consist of two neural networks: one is the Generator and the other is the Discriminator. The goal of the Generator is to learn to generate fake sample distribution to deceive the Discriminator whereas the goal of the Discriminator is to learn to distinguish between real and fake distribution generated by the Generator. The Generator tries to minimize the following function while the Discriminator tries to maximize it. The Minimax loss is given as,

$$Min_G Max_D f(D,G) = \mathrm{E}_X\left[\log(D(X))\right] + E_Z\left[\log(1 - D(G(Z)))\right]$$

Here, $\mathbb{E}_x$ is the expected value over all real data samples, D(x) is the probability estimate of the Discriminator if x is real, $G(z)$ is the output of the Generator for a given random noise vector z as input, $D(G(z))$ is the Discriminator's probability estimate if the fake generated sample is real, $\mathbb{E}_z$ is the expected value over all random inputs to the Generator.

Aided by the rapid and numerous follow-up related research, the potential of GANs were gradually released. Some of the more prominent work, such as Conditional GAN (CGAN) [24], GAN just generates samples randomly, and CGAN makes it possible to generate content according to some fixed directories; Wasserstein GAN (WGAN) [25], WGAN use their delicate mathematical derivation to improved stability of training process and get rid of several problems, e.g., mode collapse. In most cases, the purpose of people using GANs is to generate near-real data, such as text and images. In addition, a lot of exciting work has been born in recent years [26-31].

Visual transformer

The original transformer was built for NLP [5], where the multi-head self-attention and feedforward MLP layer are stacked to capture the long-term correlation between words. Its popularity among computer vision tasks rises recently [32-34]. The core of a transformer is the self-attention mechanism, which characterizes the dependencies between any two distant tokens. With the first pure Transformer-based visual backbone network ViT proposed by [18], the number of studies on A has increased significantly. The main reason is that it does not use the convolution structure, but achieved the performance beyond the traditional convolution neural network, and compared with the convolution neural network, it has less computational cost in the training of the model with the same parameters. However, it is also shown in the original paper that this Transformer-based visual model requires a huge amount of data to be fed in order to achieve excellent performance. In the case of less training data, the result is not as good as the traditional convolution neural network, e.g., [35,36].

Evaluation of transformer-based GAN

One of the most difficult aspects of GAN training is assessing their performance, or determining how well a model approximates a data distribution. In terms of theory and applications, significant progress has been made, with a large number of GAN variants now available. However there has been relatively little effort put into evaluating GANs, and there are still gaps in quantitative assessment methods. From the extensive work on evaluation of GANs [37,38], the general evaluation indicators of GANs are the use of Fre chet Inception Distance (FID [39]) and Inception Score (IS [40]). In order to understand the performance of the model more intuitively, some benchmark data sets are usually used for experiments. The most common are CIFAR-10 [41], CelebA [42], ImageNet [43], etc.

Conclusion

Along with the rapid development of Transformer, some work has been done to introduce this self-attention mechanism into the architecture of GAN, and achieved state-of-the-art results. [20,44,45]. Using Transformer to build a GAN model in addition to the advantages of using numerous tricks in the NLP domain, etc., more current difficulties can be summarized as follows:

1. Judging from the latest work, Transformer-based GAN needs to have a deeper understanding of semantic information [44-47].
2. Transformer-based GAN requires stronger attention mechanisms, especially those tailored to the characteristics of generative models, not only performance, but also efficiency, e.g. [48-50].
3. Introduce pre-training into the training of Transformer-based GAN, e.g. [51].
4. Conditional image generation in Transformer-based GAN [26,52].
5. The lack of diversity is a problem in Transformer-based GAN [53-58].

Acknowledgement

None.

Conflict of Interest

The author declares no conflict of interest.

References

1. I J Goodfellow, J Pouget Abadie, M Mirza, B Xu, D Warde Farley, et al. (2014) Generative adversarial nets. In NIPS.
2. D P Kingma, M Welling (2014) Auto-encoding variational bayes. CoRR, abs/1312.6114.
3. D P Kingma, S Mohamed, D J Rezende, M Welling (2014) Semi-supervised learning with deep generative models. ArXiv, abs/1406.5298.
4. A van den Oord, N Kalchbrenner, L Espeholt, K Kavukcuoglu, O Vinyals, et al. (2016) Conditional image generation with pixelcnn decoders. In NIPS.
5. A Vaswani, N M Shazeer, N Parmar, J Uszkoreit, L Jones, et al. (2017) Attention is all you need. ArXiv, abs/1706.03762.
6. A Radford, K Narasimhan (2018) Improving language understanding by generative pre-training.
7. A Radford, J Wu, R Child, D Luan, D Amodei, et al. (2019) Language models are unsupervised multitask learners.
8. T B Brown, B Mann, N Ryder, M Subbiah, J Kaplan, et al. (2020) Language models are few-shot learners. ArXiv, abs/2005.14165.

9. J Devlin, M W Chang, K Lee, K Toutanova (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv, abs/1810.04805.
10. Y LeCun, B E Boser, J S Denker, D Henderson, R E Howard, et al. (1989) Backpropagation applied to handwritten zip code recognition. Neural Computation 1: 541-551.
11. A Krizhevsky, I Sutskever, G E Hinton (2012) Imagenet classification with deep convolutional neural networks. Communications of the ACM 60: 84 -90.
12. K He, X Zhang, S Ren, J Sun (2016) Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 770-778.
13. A Radford, L Metz, S Chintala (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR, abs/1511.06434.
14. X Wang, R B Girshick, A K Gupta, K He (2018) Non-local neural networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7794-7803.
15. N Carion, F Massa, G Synnaeve, N Usunier, A Kirillov, et al. (2020) End-to-end object detection with transformers. ArXiv, abs/2005.12872.
16. P Ramachandran, N Parmar, A Vaswani, I Bello, A Levskaya et al. (2019) Stand-alone self-attention in vision models. In NeurIPS.
17. H Wang, Y Zhu, B Green, H Adam, A L Yuille, et al. (2020) Axial-deeplab: Standalone axial-attention for panoptic segmentation. In ECCV.
18. A Dosovitskiy, L Beyer, A Kolesnikov, D Weissenborn, X Zhai, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv, abs/2010.11929.
19. L Yuan, Y Chen, T Wang, W Yu, Y Shi, et al. (2021) Tokens-to-token vit: Training vision transformers from scratch on imagenet. ArXiv, abs/2101.11986.
20. K He, X Chen, S Xie, Y Li, P Doll'ar, et al. (2021) Masked autoencoders are scalable vision learners. ArXiv, abs/2111.06377.
21. Z Xie, Z Zhang, Y Cao, Y Lin, J Bao, et al. (2021) Simmim: A simple framework for masked image modeling. ArXiv, abs/2111.09886.
22. Z Liu, Y Lin, Y Cao, H Hu, Y Wei, et al. (2021) Swin transformer: Hierarchical vision transformer using shifted windows. ArXiv, abs/2103.14030.
23. G Harshvardhan, M K Gourisaria, M Pandey, S S Rautaray (2020) A comprehensive survey and analysis of generative models in machine learning. Computer Science Review 38: 100285.
24. M Mirza, S Osindero (2014) Conditional generative adversarial nets. ArXiv, abs/1411.1784.
25. M Arjovsky, S Chintala, L Bottou (2017) Wasserstein generative adversarial networks. In ICML.
26. P Isola, J Y Zhu, T Zhou, A A Efros (2017) Image-to-image translation with conditional adversarial networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 5967-5976.
27. H Zhang, T Xu, H Li, S Zhang, X Wang (2017) Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV) 5908-5916.
28. J Y Zhu, T Park, P Isola, A A Efros (2017) Unpaired image-to-image translation using cycleconsistent adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV) 2242-2251.
29. T Karras, S Laine, T Aila (2019) A style-based generator architecture for generative adversarial networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 4396-4405.
30. T Chen, X Zhai, M Ritter, M Lucic, N Houlsby (2019) Self-supervised gans via auxiliary rotation loss. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 12146-12155.
31. S Zhao, Z Liu, J Lin, J Y Zhu, S Han (2020) Differentiable augmentation for data-efficient gan training. ArXiv, abs/2006.10738.
32. N Parmar, A Vaswani, J Uszkoreit, L Kaiser, N Shazeer, et al. (2018) Image transformer. In Proceedings of the 35th International Conference on Machine Learning, volume 80: 4055-4064.
33. F Yang, H Yang, J Fu, H Lu, B Guo (2020) Learning texture transformer network for image super-resolution. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 5790-5799.
34. Y Zeng, J Fu, H Chao (2020) Learning joint spatial-temporal transformations for video inpainting. ArXiv, abs/2007.10247.
35. A Kolesnikov, L Beyer, X Zhai, J Puigcerver, J Yung, et al. (2020) Big transfer (bit): General visual representation learning. In ECCV.
36. Q Xie, E H Hovy, M T Luong, Q V Le (2020) Self-training with noisy student improves imagenet classification. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 10684-10695.
37. A Borji (2019) Pros and cons of gan evaluation measures. ArXiv, abs/1802.03446.
38. A Dash, J Ye, G Wang (2021) A review of generative adversarial networks (gans) and its applications in a wide variety of disciplines - from medical to remote sensing. ArXiv, abs/2110.01442.
39. M Heusel, H Ramsauer, T Unterthiner, B Nessler, S Hochreiter (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In NIPS.
40. T Salimans, I J Goodfellow, W Zaremba, V Cheung, A Radford, et al. (2016) Improved techniques for training gans. In NIPS.
41. A Krizhevsky (2009) Learning multiple layers of features from tiny images.
42. A Coates, A Ng, H Lee (2011) An analysis of single-layer networks in unsupervised feature learning. In AISTATS.
43. J Deng, W Dong, R Socher, L J Li, K Li, et al. (2009) Imagenet: A large-scale hierarchical image database. In CVPR.
44. Y Jiang, S Chang, Z Wang (2021) Transgan: Two transformers can make one strong gan. ArXiv, abs/2102.07074.
45. K Lee, H Chang, L Jiang, H Zhang, Z Tu, et al. (2021) Vitgan: Training gans with vision transformers. ArXiv, abs/2107.04589.
46. B Wu, C Xu, X Dai, A Wan, P Zhang, et al. (2020) Visual transformers: Token-based image representation and processing for computer vision. ArXiv, abs/2006.03677.
47. Y Azzi, A Moussaoui, M T Kechadi (2020) Semantic segmentation of medical images with deep learning: Overview. In Medical Technologies Journal, volume 4.
48. X Zhu, W Su, L Lu, B Li, X Wang, J Dai (2021) Deformable detr: Deformable transformers for end-to-end object detection. ArXiv, abs/2010.04159.
49. S Wang, B Z Li, M Khabsa, H Fang, H Ma (2020) Linformer: Self-attention with linear complexity. ArXiv, abs/2006.04768.
50. K Choromanski, V Likhosherstov, D Dohan, X Song, A Gane, et al. (2021) Rethinking attention with performers. ArXiv, abs/2009.14794.
51. H Bao, L Dong, F Wei (2021) Beit: Bert pre-training of image transformers. ArXiv, abs/2106.08254.
52. C H Lin, C C Chang, Y S Chen, D C Juan, W Wei, et al. (2019) Coco-gan: Generation by parts via conditional coordinating. IEEE/CVF International Conference on Computer Vision (ICCV) 4511-4520.
53. J Xu, X Ren, J Lin, X Sun (2018) Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In EMNLP.
54. R Aralikatte, S Narayan, J Maynez, S Rothe, R T McDonald (2021) Focus attention: Promoting faithfulness and diversity in summarization. In ACL/IJCNLP.

55. Y Cao, X Wan (2020) Divgan: Towards diverse paraphrase generation via diversified generative adversarial network. In FINDINGS.

56. A Brock, J Donahue, K Simonyan (2019) Large scale gan training for high fidelity natural image synthesis. ArXiv, abs/1809.11096.

57. X Gong, S Chang, Y Jiang, Z Wang (2019) Autogan: Neural architecture search for generative adversarial networks. IEEE/CVF International Conference on Computer Vision (ICCV) 3223-3233.

58. I Gulrajani, F Ahmed, M Arjovsky, V Dumoulin, A C Courville (2017) Improved training of wasserstein gans. In NIPS.