Training for High-fidelity Few-shot Image Synthesis Based On FastGAN

Jining Chen 30920211154145
Peiying Zeng 30920211154174
Zhenzhen Wang 30920211154169

School of Informatics XiaMen University
Xiamen, China

Abstract

Generative Adversarial Networks (GANs) have shown promising results in generating realistic images, however the state-of-the-art GANs usually require a lager amount of training data and expensive computational cost which limited their application in real-life scenarios. FastGAN is a cutting-edge GAN designed specifically for training from scratch using small datasets with low computation budget. In this work, we tried several optimization methods designed for GAN to further improve the performance of FastGAN. Specially, in order to avoid overfitting of the network, we apply addtional two forms of data augmentation methods offset and cutout to ensure data diversity. Besides, we modify the generator of the network to improve the performance by introducing an modified attention mechanism. Finally, we make further modification via the use of relativistic versions of hinge loss function in the discriminator. Experiment results on three datasets covering different image domains show that there is a slight performance gain when applying relativistic loss on FastGAN, while the other two techniques do not play a positive role.

Introduction

Generative Adversarial Networks(GANs)(Goodfellow et al. 2014) have become a rapidly heating-up research topic in recent years. GANs consist of two models: a generator which tries to capture the distribution of true examples for new data example generation and a discriminator distinguishing generated examples from the real ones as accurately as possible. Both of these networks play a min-max game where one is trying to outsmart the other. The adversarial idea successfully allows GANs to generate more realistic and vivid images, which is much superior to traditional methods.

The quality and realism of synthetic images has seen tremendous improvement in recent years thanks to the breakthroughs in GANs. Now state-of-the-art GANs, e.g., StyleGAN2(Karras et al. 2020) or BigGAN(Brock, Donahue, and Simonyan 2019), can synthesis fake images that can't be easily recognized by human or even by machines. However a massive amount of training data and computation resources are required for these excellent GANs to be well trained. For instance, training with StyleGAN2 on the Flickr Faces High Quality (FFHQ) dataset to generate 1024x1024 realistic images of faces requires 3 days and 22 hours with 8 tesla-v100 GPUs. With one card, it would be 27days and 23 hours.

In real-life scenarios, there are not enough mount of data available for training in many cases, such as an artist's masterpieces, a particular person's images. Aside from this, due to hardware conditions of our team and time constraints, we are interesting in training GANs with small datasets and limited computation resources while ensuring high quality.

A newly proposed lightweight GAN structure called FastGAN(Liu et al. 2021) is designed specifically for training from scratch using small datasets. The skip-layer channel-wise excitation mechanism (SLE) in generator and a self-supervised regularization on the discriminator significantly boost the synthesis performance of FastGAN. Notably, the model converges from scratch within 24 hours of training on a single RTX-2080 GPU.

In the paper of FastGAN, we found that the evaluation index Frechet Inception Distance(FID)(Heusel et al. 2017) of different datasets in the experiment part remain relatively high and there is some space for reduction, which is what we want to work on. Aiming at the improvement of FastGAN's image generation quality, we do modifications on three different aspects: input, generator structure, discriminator loss. Contributions are summarized as follow:

1. We extend data augmentation methods into four types from color, translation to color, translation, cutout, offset to increase data diversity.

2. we add residual attention layers in generator, and replace the attention mechanism in the layer with an variant called Efficient Attention.

3. We replace default hinge loss of discriminator with a relativistic version which estimates the probability that the given real data is more realistic than randomly sampled fake data.

Related Work

High-Quality Image Generation

BigGAN(Brock, Donahue, and Simonyan 2019) applies orthogonal regularization to the generator and uses a simple truncation trick to eventually produce high quality images. However, the parameters required for training are huge and the hardware requirements are correspondingly high. In order to alleviate the problem that the increased model parameters lead to more rigid gradient flow, the multi-scale gradient GAN structure(Karnewar and Wang 2020) is proposed.

Inevitably, many methods to improve the quality of the generated image add further computational costs.

Training Stability Improvement

The problems of vanishing gradient and mode collapse have severely restricted GANs, development, so many works have been proposed to stabilize its training. Wasserstein GAN(Arjovsky, Chintala, and Bottou 2017) replaces the Jensen-Shannon divergence in ordinary GAN with Wasserstein distance, which makes the training more stable and theoretically solves the problem of mode collapse and vanishing gradient. However, the need to carefully balance the training degree of G and D is a challenge that exists in WGAN. WGAN-GP (Gulrajani et al. 2017) is further proposed to solve this problem, but it requires a greater time cost to make the training converge. Based on few-shot learning, DiffAugment(Zhao et al. 2020) imposes various types of differentiable augmentations on both real and fake samples, effectively stabilizes training, and leads to better convergence.

Reduction In computational costs

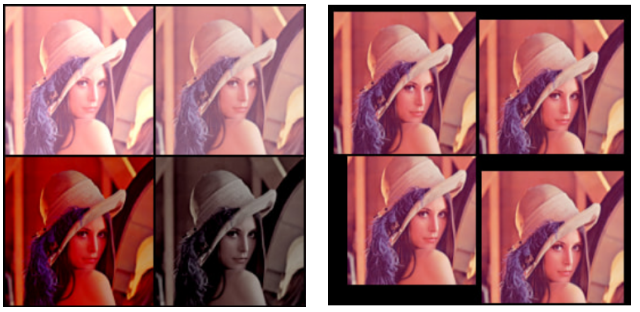
Expensive computational costs and a large amount of training data limit the use of many improved models in practical applications. Faced with such problem, few-shot learning can be a good way to reduce the computational costs if it can generate high quality images as well. DiffAugment(Zhao et al. 2020) can generate high-fidelity images using only 100 images without pre-training. A newly proposed light-weight GAN structure called FastGAN (Liu et al. 2021) gains superior quality on 1024 × 1024 resolution, even with less than 100 training samples.

Method

Data Augmentation

Data augmentation is essential for GAN to work effectively in a low data setting. In addition to the augmentation in the original FastGAN, we adopt more diverse data augmentation methods to implicitly increase the size of the datasets, thus enhancing the robustness and generalization of FastGAN.

To be specific, the augmentation types is set to color and translation by default, where color means randomly changing brightness, saturation and contrast of the input images, translation is randomly moving images on the canvas with black background. We extend this line of work by adding two types of augmentations called offset and cutout(DeVries and Taylor 2017). The offset operation randomly moves image by x and y-axis with repeating image, while the cutout augmentation works by creating random black boxes on the image. Figure 1 illustrate all four types of augmentations visually.



(a) color (b) translation

(c) offset (d) cutout

Figure 1: Illustration of four types of data augmentations

Attention In Generator

Attention mechanism, originally proposed in the field of natural language processing, has now been shown to be effective in computer vision tasks.A recent paper(Yu et al. 2021) investigated various attention mechanism in image generation models, and showed that the network performance benefits from the engagement of attention. We do architectural modification by introducing a residual version of self-attention layers (Yu et al. 2021) to get a better result.Furthermore, we replace the attention module in residual attention layer with an variant with linear complexity called Efficient Attention (Shen et al. 2021) to avoid excessive extra computational consumption.

In details, let $X \in \mathbb{R}^{h\times w\times c}$ be the input tensor to a convolutional layer in the original architecture. The feature map X pass through 1x1 convolutional kernel to get query tensor $Q(X) \in \mathbb{R}^{h\times w\times c}$. And we obtain key and values tensors $K(X), V(X) \in \mathbb{R}^{h\times w\times c}$ separately using 3x3 depthwise convolution with a padding of 1. For computing vision data, the module then flattens all three tensors by combining h axis and w axis to form $Q(x), K(x), V(x) \in \mathbb{R}^{n\times d_k}$. Instead of interpreting the keys as n features vectors in $\mathbb{R}_{d_k}$, the module regards them as d_k single-channel feature maps. Efficient attention uses each of these feature maps as a weighting over all position and aggregates the value features through weighted summation to form a global context vector.The name reflects the fact that the vector does not correspond to a specific position, but is a global description of the input features, as Figure 2 shows.

The following equation characterizes the efficient attention mechanism :

$$attn(Q(X), K(X), V(X)) = \rho_q(Q(X))\rho_k(K(X))^T V(X) \quad (1)$$

where ρ_q, ρ_k are normalization function for query and key features,respectively.The implementation of two normaliza-

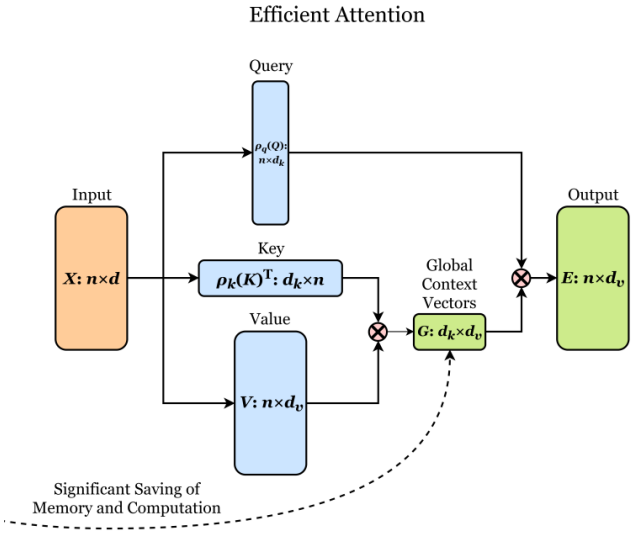


Figure 2: Illustration of the architecture of efficient attention. Each box represents an input, output, or intermediate matrix. Above it is the name of the corresponding matrix, and inside are the variable name and the size of the matrix. ρ_q, ρ_k are the normalizers on Q, K,respectively. n, d, d_k, d_v are the input size and the dimensionalities of the inputs,the keys, and the values,respectively. $\bigotimes$ denotes matrix multiplication.

tion methods are

$$\begin{aligned} \text{Scaling:} \quad & \rho_q(Y) = \rho_k(Y) = \frac{Y}{\sqrt{n}} \\ \text{Softmax:} \quad & \rho_q(Y) = \sigma_{row}(Y), \\ & \rho_k(Y) = \sigma_{col}(Y) \end{aligned} \tag{2}$$

where σ_{row},σ_{col} denote applying the softmax function along each row or column of matrix Y,respectively.

When using scaling normalization, efficient attention is mathematically equivalent to conventional dot-product attention,while the softmax variant of efficient attention is a close approximation of that variant of dot-product attention.For formal proof, refer to (DeVries and Taylor 2017).

We choose the softmax version to get the self attention result, and reshape it to $h \times w \times c$. Finally, it adds the resultant features to the input features to form a residual structure.

$$\begin{aligned} \overline{O}_{self} &= attn(Q(x), K(x), V(x)) \\ O_{self} &= \overline{O}_{self} + X \end{aligned} \tag{3}$$

Relativistic Discriminator Loss

The original GAN loss function can cause the GAN to get stuck in the early stages of GAN training when the discriminator's job is very easy. Several different variations of loss have been proposed to improve training stability. For computational efficiency, FastGAN employs the hinge version of the adversarial loss.

$$L_D = -\mathbb{E}_{x \sim I_{real}}[min(0, -1 + D(x))] - \mathbb{E}_{\hat{x} \sim G(z)}[min(0, -1 - D(\hat{x}))] + L_{recons} \tag{4}$$

$$L_G = -\mathbb{E}_{z \sim N}[D(G(z)] \tag{5}$$

Inspired by (Jolicoeur-Martineau 2018), we do a slight change of the original hinge loss by using a "relativistic discriminator" which estimate the probability that the given real data is more realistic than a randomly sampled fake data.The relativistic discriminator loss functions can be formulated by redefined $D(x)$ and $D(\hat{x})$ in equation (4):

$$D(x) = D(x) - \mathbb{E}_{\hat{x} \sim G(z)} D(\hat{x}) \tag{6}$$

$$D(\hat{x}) = D(\hat{x}) - \mathbb{E}_{x \sim I_{real}} D(x) \tag{7}$$

The intuition of this approach is that when generator is trained good enough to fool the discriminator, fake samples may appear to be more realistic than real samples, both real and fake samples are being classified as real by the discriminator.In that case, GAN completely ignores the prior knowledge that half of the mini-batch samples are fake. The discriminator should assign a higher probability of being fake to real samples rather than classify all samples are real.

Experiment

Datasets

For convenience, we conduct experiments on three datasets with multiple content categories provided by the author of FastGAN. On 256x256 resolution,we test on Animal-Face Dog. On 512x512 resolution, we test on anime face and art painting . These datasets are designed to cover images with different characteristics: photo realistic, graphic-illustration, and art-like images. Datasets used in the paper can be found at https://drive.google.com/drive/folders/1nCpr84nKkrs9-aVMET5h8gqFbUYJRPLR. The details of the datasets we experiment on are presented in Table 1.Sample images of the datasets are given by Figure 3.

	AnimalFace-Dog	Anime Face	Art paintings
Resolution	256x256	512x512	512x512
Image Number	389	100	1000

Table 1: Information about the training sets

Metrics

We use Fréchet Inception Distance (FID)(Heusel et al. 2017) to measure the models' synthesis performance,which is the golden standard measuring the overall semantic realism of the synthesized images.

We let G generate 5000 images and compute FID between the synthesized images and the whole training set. We save the checkpoints every 10k iterations during training and report the best FID from the checkpoints.

	AnimalFace-Dog	Anime Face	Art paintings
Baseline	53.80	59.19	46.19
Baseline+Data Augmentation	65.63	59.51	51.41
Baseline + Attention	101.88	/	69.53
Baseline + Relativistic discriminator loss	**50.54**	**58.87**	**45.31**

Table 2: FID comparison on few-sample datasets (50k iters). "/" stands for non-convergence

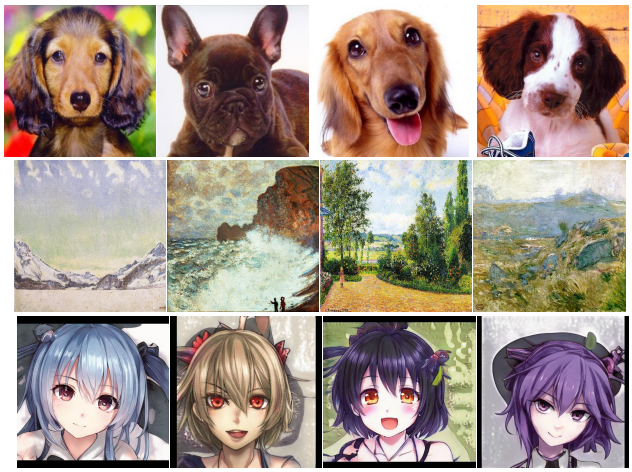


Figure 3: Sample images of the three training sets

Figure 4: 256x256 generated images of AnimalFace-dog. Each row represents a model-generated image.1: Baseline, 2: Baseline+Data Augmentation, 3: Baseline + Attention, 4: Baseline + Relativistic discriminator loss.

Performance Comparison

FastGAN has been shown to achieve superior performance on the few-shot datasets compared to state-of-the-art(SOTA) large volume model StyleGAN2. For the purpose of further improving the model's performance,we made small modification to each of the three parts of the model: augment of input data, attention in generator, relativistic discriminator. Experiments are conducted with a single nvidia Titan-XP GPU card. Due to the limitations of the hardware, we only add new attention layers after the 4x4, 8x8, and 16x16 resolution convolution layers of the generator. Experiment results are shown in Table 2.

From Table 2, we observe that applying data augmentation does not improve the model performance,which is not in line with our perception, while using a relativistic discriminator generally improve data generation quality.

We also can draw the conclusion that our modified residual attention layer brings unstable performance of FastGAN. The experiment results either become worse or even do not converge.We infer that it could be one or more of the following 4 reasons. First, code implementation is incorrect, which is inconsistent with the paper's approach. Second, the residual attention structure (Yu et al. 2021) can only works well with the corresponding proposed attention mechanism, not efficient attention. Besides, the residual attention structure is adopted not only in the generator, but also in the discriminator in (Yu et al. 2021). Only when both are used, can the performance improve. Finally, we just followed the FastGAN authors' original settings for the experiments while did not make any hyperparameter adjustments in order to saving time.

Qualitative results are presented in Figure 4-6. The quality of the images generated by the baseline + attention model is significantly worse than that of the other models, while the other models make little difference visually, as the difference in FID metrics is not very large.

Conclusion

In this paper, we apply three techniques to baseline model FastGAN to further improve image generation quality. On three datasets with a diverse content variation, we show that hat there is a slight performance gain when applying relativistic loss on FastGAN, while the other two techniques do not play a positive. We hope our work can provide new study perspectives on few-shot image synthesis for future research.

Figure 5: 512x512 generated images of Art Paintings. Each row represents a model-generated image. 1: Baseline, 2: Baseline+Data Augmentation, 3: Baseline + Attention, 4: Baseline + Relativistic discriminator loss.

Figure 6: 512x512 generated images of Anime Face. Each row represents a model-generated image. 1: Baseline, 2: Baseline+Data Augmentation, 3: Baseline + Attention, 4: Baseline + Relativistic discriminator loss.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. arXiv:1701.07875.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. 2017. Improved Training of Wasserstein GANs. arXiv:1704.00028.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Jolicoeur-Martineau, A. 2018. The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734*.

Karnewar, A.; and Wang, O. 2020. MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks. arXiv:1903.06048.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

Liu, B.; Zhu, Y.; Song, K.; and Elgammal, A. 2021. Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis. In *International Conference on Learning Representations*.

Shen, Z.; Zhang, M.; Zhao, H.; Yi, S.; and Li, H. 2021. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3531–3539.

Yu, N.; Liu, G.; Dundar, A.; Tao, A.; Catanzaro, B.; Davis, L. S.; and Fritz, M. 2021. Dual contrastive loss and attention for gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6731–6742.

Zhao, S.; Liu, Z.; Lin, J.; Zhu, J.-Y.; and Han, S. 2020. Differentiable Augmentation for Data-Efficient GAN Training. arXiv:2006.10738.