TransGAN: Do Two Transformers Make One Strong GAN?

Azfar Khoja and Alexander Peplowski

Université de Montréal

Introduction

In this project we evaluate and extend TransGAN [1], a GAN framework which eliminates the use of convolutional operations, substituting them with a transformer architecture.
Can transformers provide the spatial coherency in structure, color and texture needed for generating images?

TransGAN Model Architecture

Generator: Pixels are input as tokens to each transformer block with iterative upsampling of the feature map resolution and decreasing the embedding dimension at each stage, keeping the feature size constant.

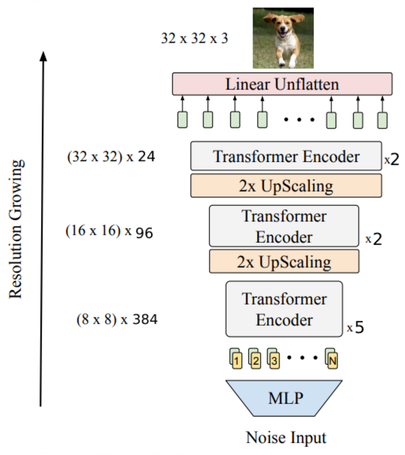

Figure: **TransGAN-S Generator**. Modified from [1]

Discriminator: Vision Transformer (ViT) [2] with 8x8 image patches as input tokens. Output from [CLS] token used to classify real or fake.

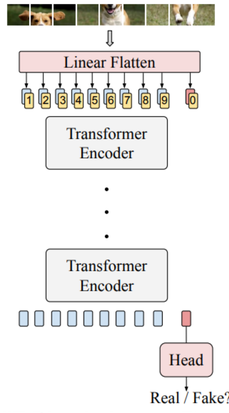

Figure: **ViT Discriminator**. Figure from [1]

Training Tricks

Data Augmentation: DiffAugment [3] (translation, cutout, color) and horizontal reflection are applied to real images for the Discriminator.
Co-Training Task: Provide a low resolution input of the real image at Stage 2, use the upsampled output to calculate MSE loss for the generator.

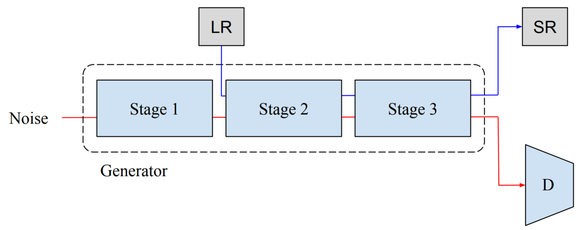

Figure: **Super-Resolution Co-Training Task**. From [1]

Initialization for Self Attention: Encode inductive image biases by introducing locality aware initialization for the attention mask early during training

Figure: **Attention Initialization**. Left: illustration from [1]. Right, our illustration of the actual implementation

Parameter Averaging: Maintains an exponential moving average of model parameter values. At inference time, use the parameter values from the moving average for model evaluation. Used for TransGAN, but this trick [4] is not mentioned in the paper. Parameter averaging is performed as per the equation:

$$\overline{\theta}_t = \alpha\theta_t + (1-\alpha)\overline{\theta}_{t-1}$$

Conditional Loss: Based on ACGAN[5], we add a conditional log-likelihood term to the discriminator and the generator loss. The conditional loss is defined as:

$$\mathbb{E}[\log\ p(y=Y|X_{real})] + \mathbb{E}[\log\ p(y=Y|X_{fake})]$$

Experimental Setup

Reproduce TransGAN-S: with WGAN-GP loss [6].
Ablation study: We perform an ablation study using each of the training "tricks" proposed by the authors and by using our implementation of conditional loss. We limit the training effort to 150 generator epochs for each experiment to reduce compute cost.

Results

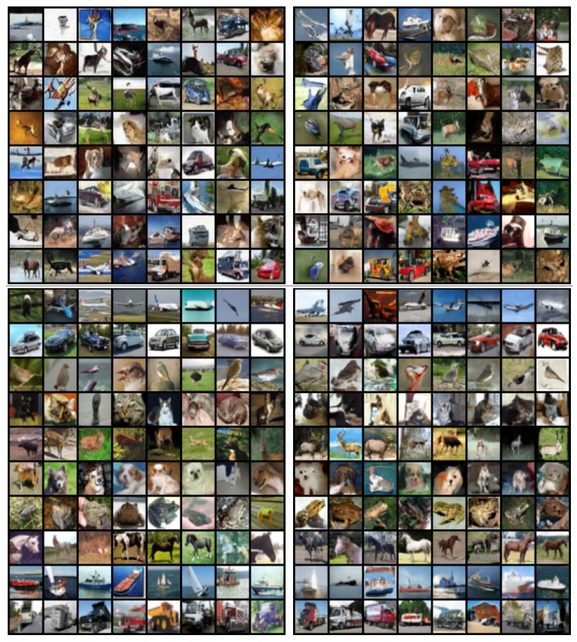

Figure: **Final Model Results** Clockwise from top-left: 1) Jiang et al's [1] TransGAN-XL on CIFAR, at FID = 11.07; 2) Our unconditional implementation of TransGAN-S, at FID = 22.59; 3) Our conditional implementation of TransGAN-S at FID = 32.16; 4) Training images from CIFAR-10 dataset.

	FID Score		
Configuration	Theirs	Ours[2]	Ours[3]
Vanilla TransGAN-S	41.41	77.84	-
Above + Data Augmentation	19.85	40.78	-
Above + Co-Training Task	19.12	40.10	-
Above + Attention Masking	**18.58**	37.70	-
Above + Parameter Averaging	**18.58**[1]	33.49	**22.59**
Above + Conditional Loss	-	34.52	-

Table: Ablation study of TransGAN-S training configurations using CIFAR-10. Note 1: No ablation study on parameter averaging in original paper. Note 2: With training limited to 150 generator epochs. Note 3: Without training limit.

Conclusions

- All tricks proposed improve performance and data augmentation provides the most benefit.
- Transformer architectures need much more data to provide spatial coherency required for images.
- TransGAN with an extra conditional constraint successfully generates images for the given class.
- We obtain similar results (in terms of FID score) as the authors with fewer compute resources.
- Some items were left out from the original paper such as parameter averaging and the attention initialization strategy.
- We propose that with enough compute for a thorough hyperparameter search (not all hyperparameters are published), the results can be better reproduced!

Credits

We follow the work of [1]. Our contributions include:

- Development of a new version of the (unpublished) TransGAN training loop.
- Validation of the training tricks and ablation study for the TransGAN-S model.
- Extension of the TransGAN model by adding conditional loss.

References

[1] Yifan Jiang, Shiyu Chang, et al.
Transgan: Two transformers can make one strong gan, 2021.

[2] Alexey Dosovitskiy, Lucas Beyer, et al.
An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[3] Shengyu Zhao, Zhijian Liu, et al.
Differentiable augmentation for data-efficient gan training, 2020.

[4] Tim Salimans, Ian J. Goodfellow, et al.
Improved techniques for training gans, 2016.

[5] Augustus Odena, Christopher Olah, et al.
Conditional image synthesis with auxiliary classifier gans, 2017.

[6] Ishaan Gulrajani, Faruk Ahmed, et al.
Improved training of wasserstein gans, 2017.

[7] Martin Arjovsky, Soumith Chintala, et al.
Wasserstein gan, 2017.

Université de Montréal Mila