Teaching StyleGAN to Read: Improving Text-to-image Synthesis with U2C Transfer Learning

Vinicius Gomes Pereira , Jônatas Wehrmann

Pontifical Catholic University of Rio de Janeiro

Introduction

We introduce Unconditional-to-Conditional Transfer Learning (U2C transfer) which allows the use of pre-training information from an unconditional network to generate text-conditioned images. It is able to leverage pre-trained models that generate human faces and adapt them to synthesize high-quality images of Birds and Flowers for instance. This method not only stabilizes the training process but also makes the training convergence faster. U2C transfer is tested in the standard StyleGAN2 unconditional models by finetuning them in two widely used datasets. The resulting model, a Text-Conditioned StyleGAN2 (TC-StyleGAN2), effortlessly bests prior work results in a few hundred iterations. TC-StyleGAN2 also makes use of two data augmentation mechanisms to help preventing overfitting: (i) a textual augmentation technique, and (ii) adaptive discriminator augmentation (ADA) which automatically regulates the strength of the image transformations. We show that our model is capable of image editing by doing arithmetic operations on the text embedding information and interpolation in the latent intermediate space of the Mapping Network.

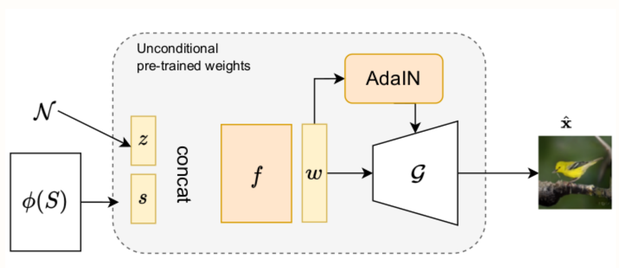

Figure 1: Overall architecture of TC-StyleGAN2. generator

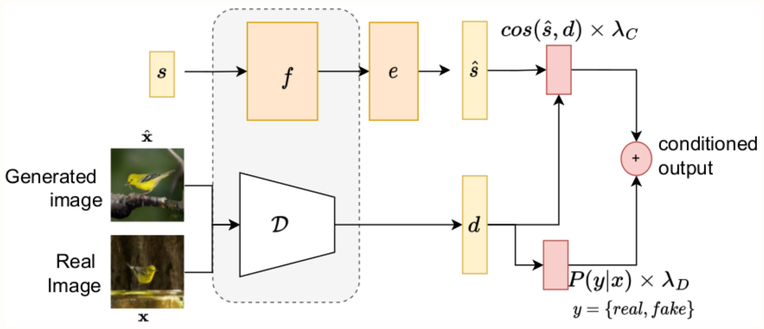

Figure 2: Overall architecture of TC-StyleGAN2. discriminator

Figure 3: Images synthesized by the proposed approach and the ground truth.

Overall architecture

Figure 1 and 2 show an overall view of TC-StyleGAN2. We reuse the maximum number of pre-trained weights as possible, making only indispensable changes to the architecture. We made two main modifications to the original model so as to introduce natural language information while still being able to reuse the pre-trained weights: (i) The dimensions of the noise vector $\mathbf{z} \in \mathcal{Z}^{512}$ are split in half, and we concatenate text information in the other half of the vector; and (ii) we enforce the discriminator latent representation vector $\mathbf{d}$ to have high cosine similarity to the sentence embedding $\mathbf{s}$, obtained by inputting the original condition vector into a new trained network implemented in parallel, used as the image synthesis condition. Such score is also used as additional information for the discriminator prediction. Both modifications are important so the generator can synthesize a plethora of different images, due to the sampling $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$, though respecting the condition fixed on the other part of the input vector. By adding the cosine constraint in the discriminator, it becomes able to penalize when generated images are not correspondent to the original sentence.

Results

In Figure 4, training from scratch results in more training instability after a few iterations, and may cause divergence. Each proposed component in TC-StyleGAN2 is quite important and helpful not only for improving generalization but also for accelerate and stabilize the training procedure. Using ADA the FID will take longer to improve but model gets far more robust to overfitting and unstability. Using the complete approach, the transfer-learning results in state-of-the-art FID values very quickly.

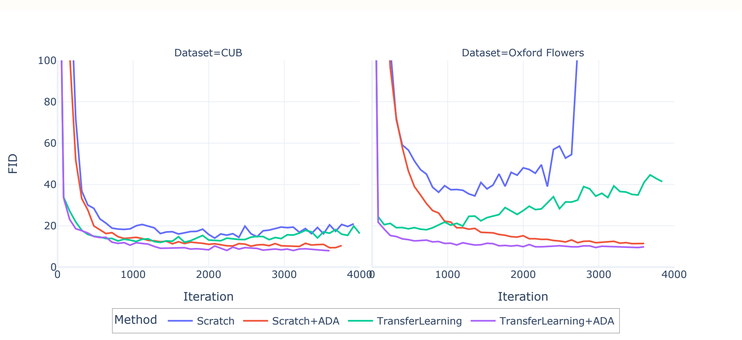

Figure 4: FID for various generator designs (lower is better)

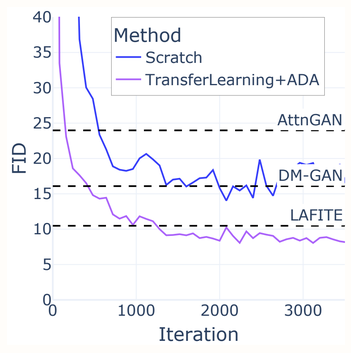

Figure 5: TC-StyleGAN2 outperforms most of prior work in a few hundred iterations.

Table 1 shows results for our main approach, TC-StyleGAN2, as well as results from prior work. We achieve 7.48 FID versus 8.02 for Lightweight ManiGAN in CUB dataset. TC-StyleGAN2 results present a relative improvement of roughly 40% when compared to LAFITE. Compared to the previously reported FID results for Oxford-102, once again results of TC-StyleGAN2 are second to none.

	FID ↓		IS ↑	
Methods	CUB	Oxford-102	CUB	Oxford 102
StackGAN++	15.30	48.68	4.04 ± 0.05	3.26 ± 0.01
AttnGAN	23.98	-	4.36 ± 0.03	-
DM-GAN	16.09	-	4.75 ± .07	-
RATGAN	13.91	-	5.36 ± 0.20	**4.09**
ET2I	11.17	16.47	4.23 ± 0.05	3.71 ± 0.06
Lightweight ManiGAN	8.02	-	-	-
LAFITE	10.48	-	5.97	
TC-StyleGAN2 (Ours)	**7.49**	**9.47**	**5.99 ± 0.20**	3.84 ± 0.15

Table 1: Comparison of TC-StyleGAN2 against state-of-the-art models.

Standard TC-StyleGAN2 incarnation employs U2C transfer, Condition Augmentation, and ADA. Table 2 provides an ablation study that shows the impact of each component in the generator design of TC-StyleGAN2. The models trained from scratch had the worst results, and sometimes the training procedure diverged. ADA clearly causes strong reductions of **33.05%** and **66.99%** in the FID scores. It is very clear as well that U2C transfer does bring important improvements to the results. Textual augmentation improved performance on the CUB dataset.

	CUB		Oxford-102	
Methods	FID ↓	KID (x 10^3) ↓	FID ↓	KID (x 10^3) ↓
From Scratch	14,04	5,55	34,44	22,87
+ ADA	9,40	4,11	11,37	3,51
+ U2C transfer	8,02	2,25	**9,47**	**2,11**
+ Conditional Augmentation	7,53	**2,07**	10,13	2,87
+ λ_D Tuning	**7,49**	2,14	9,85	2,41

Table 2: FID and KID for various generator designs (lower is better)

Image Editing

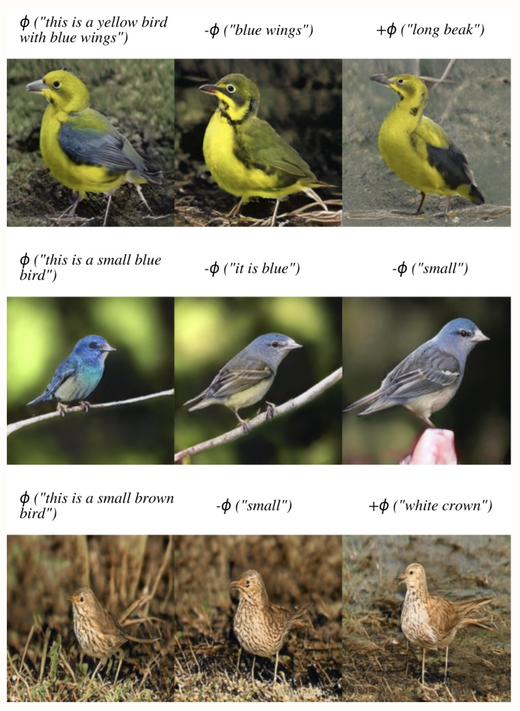

Figure 6: Arithmetic in the text-embedding space which enables natural language-based image-editing.

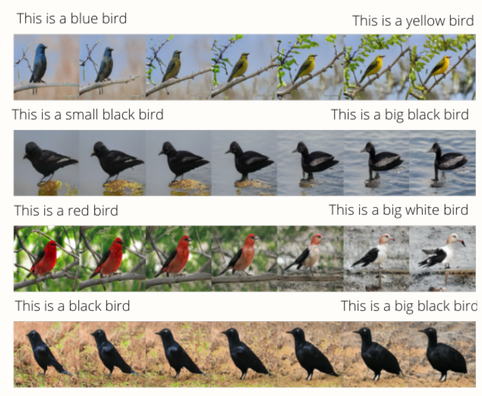

Figure 7: Images generated by linear interpolation in the intermediate space $\mathcal{W}$ (left to right)

Conclusion

We proposed a simple, yet very effective transfer-learning approach for training text-conditioned GANs, namely Unconditional-to-Conditional Transfer Learning (U2C transfer). By using such an approach we were able to modify the unconditional architecture of StyleGAN2 to allow text-conditioned image synthesis, which we called Text-Conditioned StyleGAN2 (TC-StyleGAN2). We also added stronger augmentation recipes and strategies, which allowed us to train reasonably large models in very small datasets. Such a method effortlessly outperformed previous state-of-the-art models by large margins in terms of FID in widely used benchmarks. We have shown that pre-training information of an unconditional model trained in a different and more diverse dataset is really helpful when training in smaller datasets. TC-StyleGAN2 took only a few hundred iterations to top most of the prior work. The learning procedure was much more stable when used the proposed strategy. We show that our model is capable of image editing by doing arithmetic operations on the text embedding information and interpolation in the latent intermediate space of the Mapping Network.