DAY 5: ADVANCED GENERATIVE ADVERSARIAL NETWORKS (GANS)

2021-05-07 | Mehdi Cherti | Cross Sectional Team Deep Learning, Helmholtz AI @ JSC

JÜLICH
Forschungszentrum

GENERATIVE MODELS

- Impressive progress in last years, algorithmic improvements coupled with **large scale training** and **large models**

(Source: `https://bit.ly/3azTV7J`)

GENERATIVE MODELING BENEFIT FROM SCALING: BIGGAN

- BigGAN (Brock, Donahue, and Simonyan 2019) was the first architecture to scale to ImageNet-1K
- Trained on high resolution up to 512x512, and have **high diversity and high quality samples**

GENERATIVE MODELING BENEFIT FROM SCALING: BIGGAN

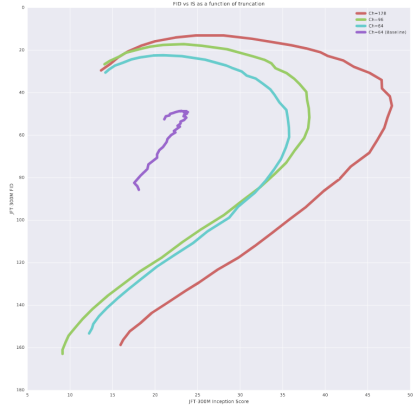
- Model much larger than previous works: scaling width and and depth
- Batch size (up to 2048) much bigger than previous works
- Benefit of scaling the model: reach better performance in **fewer iterations**

Batch	Ch.	Param (M)	Shared	Skip-z	Ortho.	Itr $\times 10^3$	FID	IS
256	64	81.5	SA-GAN Baseline			1000	18.65	52.52
512	64	81.5	✗	✗	✗	1000	15.30	58.77($\pm$1.18)
1024	64	81.5	✗	✗	✗	1000	14.88	63.03($\pm$1.42)
2048	64	81.5	✗	✗	✗	732	12.39	76.85($\pm$3.83)
2048	96	173.5	✗	✗	✗	295($\pm$18)	9.54($\pm$0.62)	92.98($\pm$4.27)
2048	96	160.6	✓	✗	✗	185($\pm$11)	9.18($\pm$0.13)	94.94($\pm$1.32)
2048	96	158.3	✓	✓	✗	152($\pm$7)	8.73($\pm$0.45)	98.76($\pm$2.84)
2048	96	158.3	✓	✓	✓	165($\pm$13)	8.51($\pm$0.32)	99.31($\pm$2.10)
2048	64	71.3	✓	✓	✓	371($\pm$7)	10.48($\pm$0.10)	86.90($\pm$0.61)

Ch.	Param (M)	Shared	Skip-z	Ortho.	FID	IS	(min FID) / IS	FID / (max IS)
64	317.1	✗	✗	✗	48.38	23.27	48.6/23.1	49.1/23.9
64	99.4	✓	✓	✓	23.48	24.78	22.4/21.0	60.9/35.8

GENERATIVE MODELING BENEFIT FROM SCALING: BIGGAN

- 512x512 resolution model trained on 512 TPUs (TPU v3 pod)
- Training takes between 24 hours and 48 hours for most models
- Results as model size is increased:



GENERATIVE MODELING BENEFIT FROM SCALING: BIGGAN

- a lot of tuning is necessary in experimentation, before finding the good range of hyper-parameters

We performed various hyperparameter sweeps in this work:

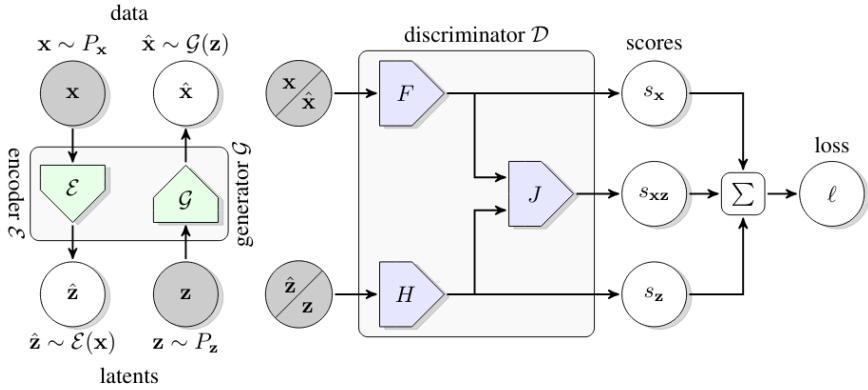
- We swept the Cartesian product of the learning rates for each network through [10^{-5}, $5 \cdot 10^{-5}$, 10^{-4}, $2 \cdot 10^{-4}$, $4 \cdot 10^{-4}$, $8 \cdot 10^{-4}$, 10^{-3}], and initially found that the SA-GAN settings (**G**'s learning rate 10^{-4}, **D**'s learning rate $4 \cdot 10^{-4}$) were optimal at lower batch sizes; we did not repeat this sweep at higher batch sizes but did try halving and doubling the learning rate, arriving at the halved settings used for our experiments.
- We swept the R1 gradient penalty strength through [10^{-3}, 10^{-2}, 10^{-1}, 0.5, 1, 2, 3, 5, 10]. We find that the strength of the penalty correlates negatively with performance, but that settings above 0.5 impart training stability.
- We swept the keep probabilities for DropOut in the final layer of **D** through [0.5, 0.6, 0.7, 0.8, 0.9, 0.95]. We find that DropOut has a similar stabilizing effect to R1 but also degrades performance.
- We swept **D**'s Adam β_1 parameter through [0.1, 0.2, 0.3, 0.4, 0.5] and found it to have a light regularization effect similar to DropOut, but not to significantly improve results. Higher β_1 terms in either network crippled training.
- We swept the strength of the modified Orthogonal Regularization penalty in **G** through [10^{-5}, $5 \cdot 10^{-5}$, 10^{-4}, $5 \cdot 10^{-4}$, 10^{-3}, 10^{-2}], and selected 10^{-4}.

REPRESENTATION LEARNING BENEFIT FROM SCALING: BIGBIGAN

- BigBiGAN(Donahue and Simonyan 2019) asked the following question: can GANs learn a useful general representation from unlabeled data ?
- Can we learn high level concepts that we can exploit for downstream tasks ?

REPRESENTATION LEARNING BENEFIT FROM SCALING: BIGBIGAN

- Similar to BigGAN but in addition to discriminator and generator, we also have an **encoder**
- Three networks to optimize simultanously

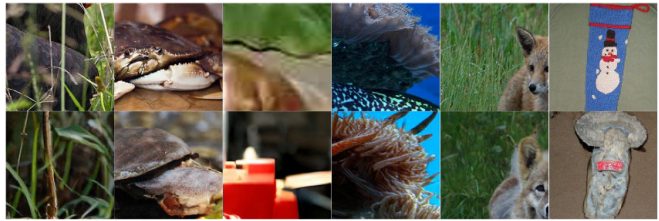


REPRESENTATION LEARNING BENEFIT FROM SCALING: BIGBIGAN

- Training on ImageNet-1K up to 256x256 resolution, completely unsupervised (no conditioning)
- Training on 32 to 512 TPU cores
- Batch size of 2048 similar to BigGAN
- Architecture similar to BigGAN for **generator** and **discriminator**, for **encoder** architecture is based on ResNet-50

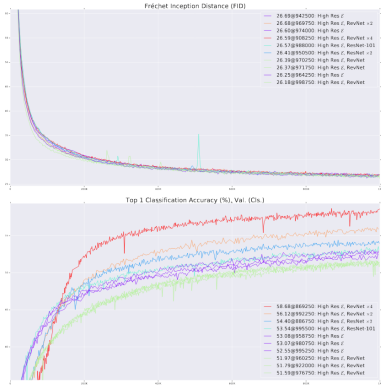
REPRESENTATION LEARNING BENEFIT FROM SCALING: BIGBIGAN

- Learned representation focus on high-level semantic details



REPRESENTATION LEARNING BENEFIT FROM SCALING: BIGBIGAN

- Better image modeling (FID) translates to better performance in downstream task (supervised classification)
- Bigger models perform **better** in downstream task (supervised classification)



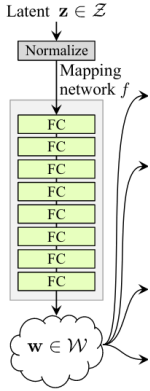
INNOVATIONS IN ARCHITECTURE: STYLEGAN2

- StyleGAN/StyleGAN2 (Karras et al. 2020) introduces a novel way to structure the generator architecture
- It decomposes the latent space into **high level attributes** (encoding concepts such pose and identity) and **stochastic variation** (e.g., to handle freckles and hair)

(a) Generated image (b) Stochastic variation

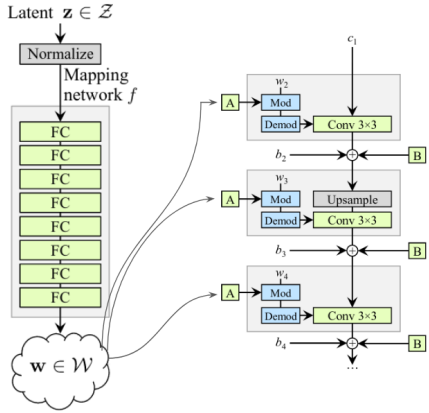
INNOVATIONS IN ARCHITECTURE: STYLEGAN2

- The latent $\boldsymbol{z}$ is converted to $\boldsymbol{w} = f(\boldsymbol{z})$ using 8 fully connected layers f
- The $\boldsymbol{w}$ vector is then mapped to a style vector using a fully connected network for each resolution. One style vector per resolution.



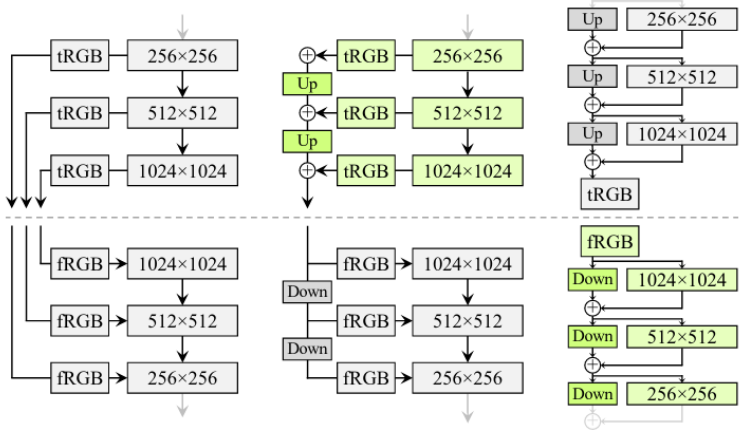
INNOVATIONS IN ARCHITECTURE: STYLEGAN2

- One block per resolution, each block upsample by 2
- Each resolution block is affected by its dedicated style vector *A* and noise *B*



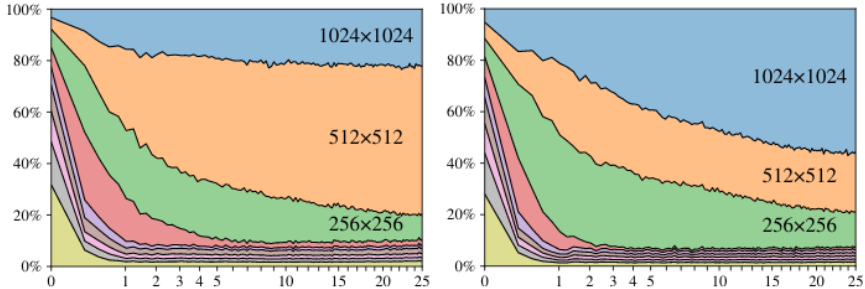
INNOVATIONS IN ARCHITECTURE: STYLEGAN2

- No need progressive generation like in ProGAN (Karras et al. 2018)
- The generated image RGBs are a sum of RGBs from each resolution outputs, everything is learned simultanously



INNOVATIONS IN ARCHITECTURE: STYLEGAN2

- Larger configuration network renders high resolution details better



(a) StyleGAN-sized (config E) (b) Large networks (config F)

INNOVATIONS IN ARCHITECTURE: STYLEGAN2

- Distributed training on 8 V100 GPUs
- Reduce training time from 70 days with 1 GPU to 10 days with 8 GPUs

Configuration	Resolution	Total kimg	1 GPU	2 GPUs	4 GPUs	8 GPUs	GPU mem
`config-f`	1024×1024	25000	69d 23h	36d 4h	18d 14h	9d 18h	13.3 GB
`config-f`	1024×1024	10000	27d 23h	14d 11h	7d 10h	3d 22h	13.3 GB
`config-e`	1024×1024	25000	35d 11h	18d 15h	9d 15h	5d 6h	8.6 GB
`config-e`	1024×1024	10000	14d 4h	7d 11h	3d 20h	2d 3h	8.6 GB
`config-f`	256×256	25000	32d 13h	16d 23h	8d 21h	4d 18h	6.4 GB
`config-f`	256×256	10000	13d 0h	6d 19h	3d 13h	1d 22h	6.4 GB

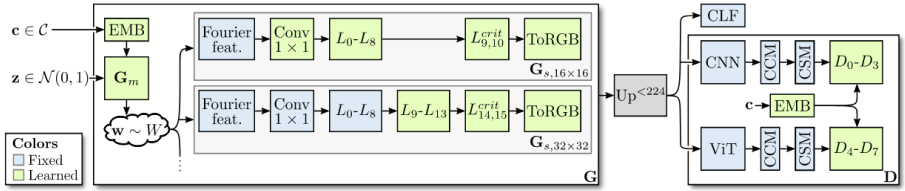
INNOVATIONS IN ARCHITECTURE: STYLEGAN2

- One should not forget the cost of exploration as well: 51 GPU years was required in total

Item	GPU years (Volta)	Electricity (MWh)
Initial exploration	20.25	58.94
Paper exploration	13.71	31.49
FFHQ config F	0.23	0.68
Other runs in paper	7.20	16.77
Backup runs left out	4.73	12.08
Video, figures, etc.	0.31	0.82
Public release	4.62	10.82
Total	51.05	131.61

STYLEGAN-XL: SCALING STYLEGAN TO LARGE DIVERSE DATASETS

- Up until now, StyleGAN models had difficulties with large diverse datasets such as ImageNet
- By combining different techniques, StyleGAN-XL (Sauer, Schwarz, and Geiger 2022) could achieve state of the art results on ImageNet for the first time



STYLEGAN-XL: SCALING STYLEGAN TO LARGE DIVERSE DATASETS

- They leverage several recent techniques to improve sample quality
- In particular, they exploit the rich representation of several pre-trained models (supervised and self-supervised)

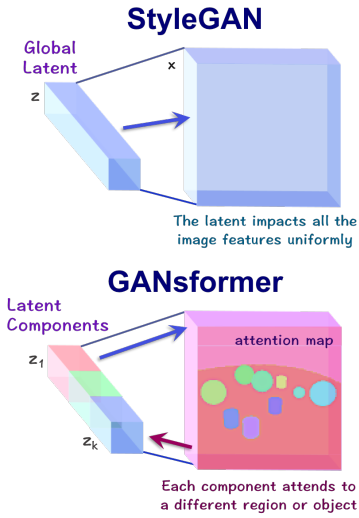
Configuration		FID ↓	IS ↑
A	StyleGAN3	53.57	15.30
B	+ Projected GAN & small **z**	22.98	57.62
C	+ Pretrained embeddings	20.91	35.79
D	+ Progressive growing	19.51	35.74
E	+ ViT & CNN as $\mathbf{F}_{1,2}$	12.43	56.72
F	+ CLF guidance (StyleGAN-XL)	**12.24**	**86.21**

INNOVATIONS IN ARCHITECTURE: GANSFORMER

- StyleGAN2 have been shown to have difficulties with datasets with a lot of diversity, e.g., complex scenes with multiple objects
- This is possibly attributed to the fact that one global latent controls all the styles simultanously
- GANsFormer (Hudson and Zitnick 2021) is a new architecture, based on StyleGAN2, where they have multiple latents and use transformers to integrate information from the latents into the image

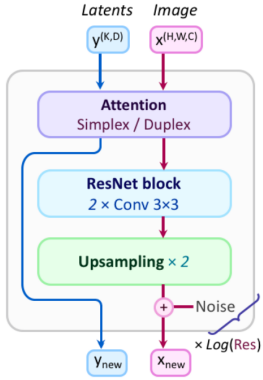
INNOVATIONS IN ARCHITECTURE: GANSFORMER

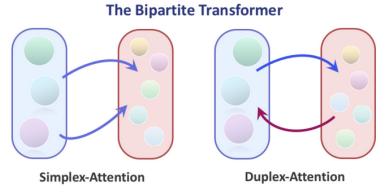
- we have multiple latents instead of a single one that globally controls the image



INNOVATIONS IN ARCHITECTURE: GANSFORMER

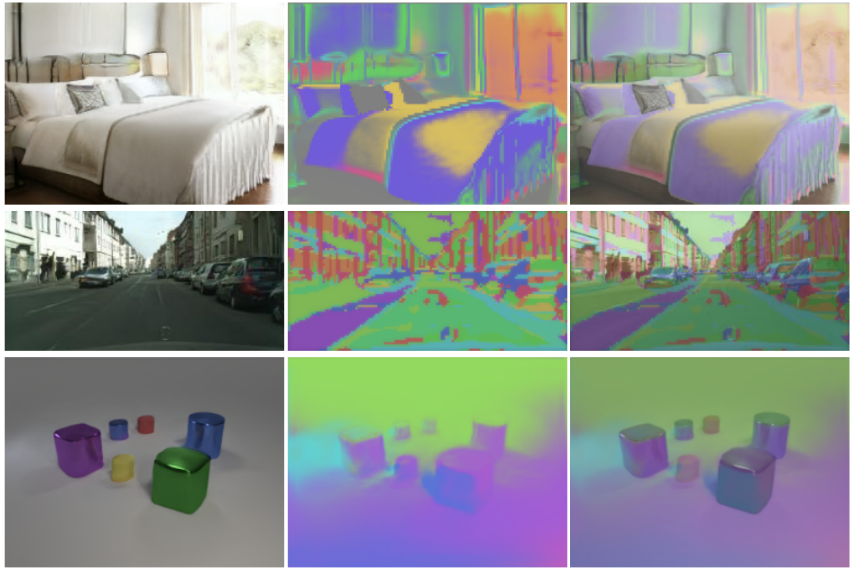
- To integrate information from the latents Y into the image X, they use a transformer architecture
- To make the transformer efficient, they use a bipartite structure, where connections are made between image features and latents only





INNOVATIONS IN ARCHITECTURE: GANSFORMER

- Different latents specialize in different aspects of the image



SUMMARY

- We have seen different architectures proposed in the literature
- The GANs are in general costly to train, especially with larger resolutions and for large datasets.
- Distributed training helps to make training faster

REFERENCES I

Brock, Andrew, Jeff Donahue, and Karen Simonyan. 2019. “Large Scale Gan Training for High Fidelity Natural Image Synthesis.” http://arxiv.org/abs/1809.11096.

Donahue, Jeff, and Karen Simonyan. 2019. “Large Scale Adversarial Representation Learning.” http://arxiv.org/abs/1907.02544.

Hudson, Drew A, and C Lawrence Zitnick. 2021. “Generative Adversarial Transformers.” arXiv Preprint arXiv:2103.01209.

Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. “Progressive Growing of Gans for Improved Quality, Stability, and Variation.” http://arxiv.org/abs/1710.10196.

REFERENCES II

Karras, Tero, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. “Training Generative Adversarial Networks with Limited Data.” http://arxiv.org/abs/2006.06676.

Sauer, Axel, Katja Schwarz, and Andreas Geiger. 2022. “StyleGAN-XI: Scaling Stylegan to Large Diverse Datasets.” arXiv Preprint arXiv:2202.00273.