LARGE SCALE GAN TRAINING FOR HIGH FIDELITY NATURAL IMAGE SYNTHESIS

ICLR 2019

박성현

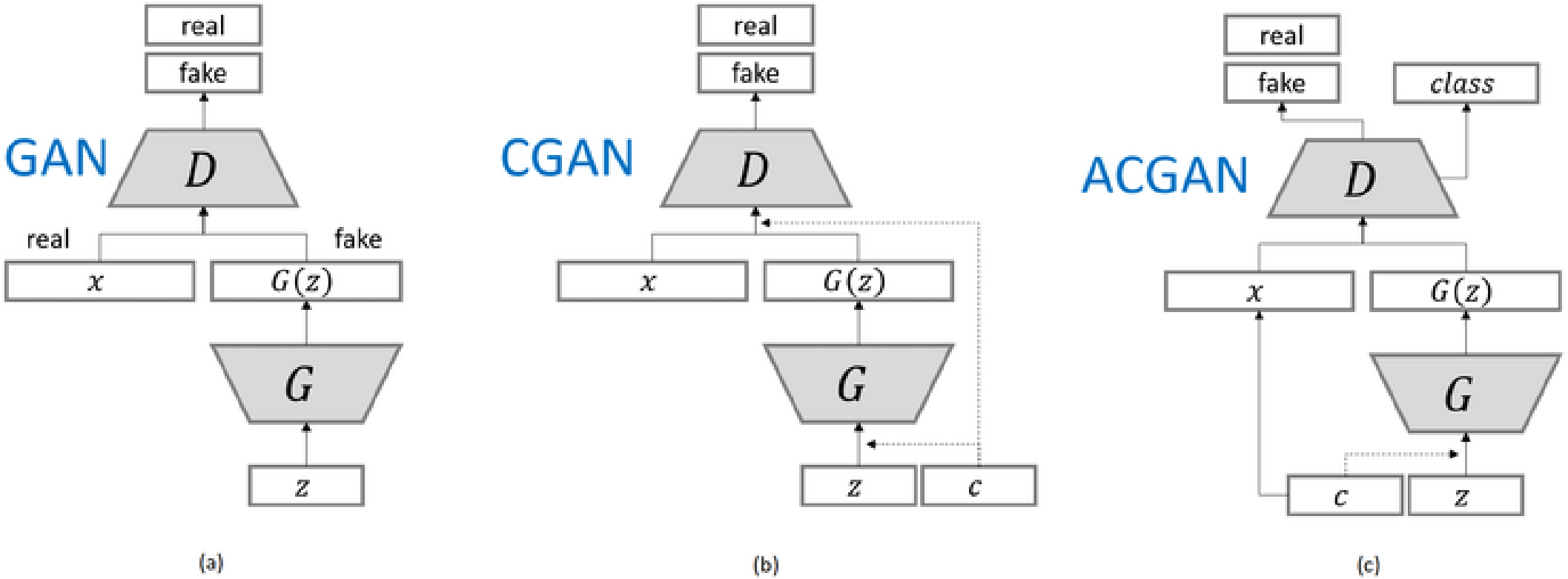
DAVIAN

Data and Visual Analytics Lab

Motivation

- ***Scalability***: As the authors discovered that GANs benefit dramatically from *scaling*, they introduced two architectural changes to improve scalability.
- ***Robustness***: The orthogonal regularization applied to the generator makes the model amenable to the *"truncation trick"* so that fine control of the trade-offs between fidelity and variety is possible by truncating the latent space.
- ***Stability***: The authors discovered and characterized instabilities specific to large-scale GANs, and devised solutions to minimize the instabilities
  — although these involved a relatively high trade-off on performance.

Background – Class-conditional GANs



Scaling Up GANs

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)

Table 1: Fréchet Inception Distance (FID, lower is better) and Inception Score (IS, higher is better) for ablations of our proposed modifications. *Batch* is batch size, *Param* is total number of parameters, *Ch.* is the channel multiplier representing the number of units in each layer, *Shared* is using shared embeddings, *Skip-z* is using skip connections from the latent to multiple layers, *Ortho.* is Orthogonal Regularization, and *Itr* indicates if the setting is stable to 10^6 iterations, or it collapses at the given iteration. Other than rows 1-4, results are computed across 8 random initializations.

Scaling Up GANs

- ***Increasing the batch size*** by a factor of 8 improves the state-of-the-art IS by 46%
- Increase the number of channels in each layer by 50%. This leads to a further IS improvement of 21%, which authors posit is due to the increased capacity of the model relative to the complexity of the dataset.

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)

Scaling Up GANs

- Class embeddings c used for the conditional BatchNorm layers in G contain a large number of weights.
- Instead of having a separate layer for each embedding, use a ***shared embedding***, which is linearly projected to each layer's gains and biases.
- This reduces computation and memory costs and improves training speeds by 37%.

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(±1.18)
1024	64	81.5	✗	✗	✗	1000	14.88	63.03(±1.42)
2048	64	81.5	✗	✗	✗	732	12.39	76.85(±3.83)
2048	96	173.5	✗	✗	✗	295(±18)	9.54(±0.62)	92.98(±4.27)
2048	96	160.6	✓	✗	✗	185(±11)	9.18(±0.13)	94.94(±1.32)
2048	96	158.3	✓	✓	✗	152(±7)	8.73(±0.45)	98.76(±2.84)
2048	96	158.3	✓	✓	✓	165(±13)	8.51(±0.32)	99.31(±2.10)
2048	64	71.3	✓	✓	✓	371(±7)	10.48(±0.10)	86.90(±0.61)

Scaling Up GANs

- Add direct skip connections (***skip-z***) from the noise vector z to multiple layers of G rather than just the initial layer.
- The intuition behind this design is to allow G to use the latent space to directly influence features at different resolutions and levels of hierarchy.
- Skip-z provides a modest performance improvement of around 4%, and improves training speed by a further 18%

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(±1.18)
1024	64	81.5	✗	✗	✗	1000	14.88	63.03(±1.42)
2048	64	81.5	✗	✗	✗	732	12.39	76.85(±3.83)
2048	96	173.5	✗	✗	✗	295(±18)	9.54(±0.62)	92.98(±4.27)
2048	96	160.6	✓	✗	✗	185(±11)	9.18(±0.13)	94.94(±1.32)
2048	96	158.3	✓	✓	✗	152(±7)	8.73(±0.45)	98.76(±2.84)
2048	96	158.3	✓	✓	✓	165(±13)	8.51(±0.32)	99.31(±2.10)
2048	64	71.3	✓	✓	✓	371(±7)	10.48(±0.10)	86.90(±0.61)

Model architecture of BigGAN

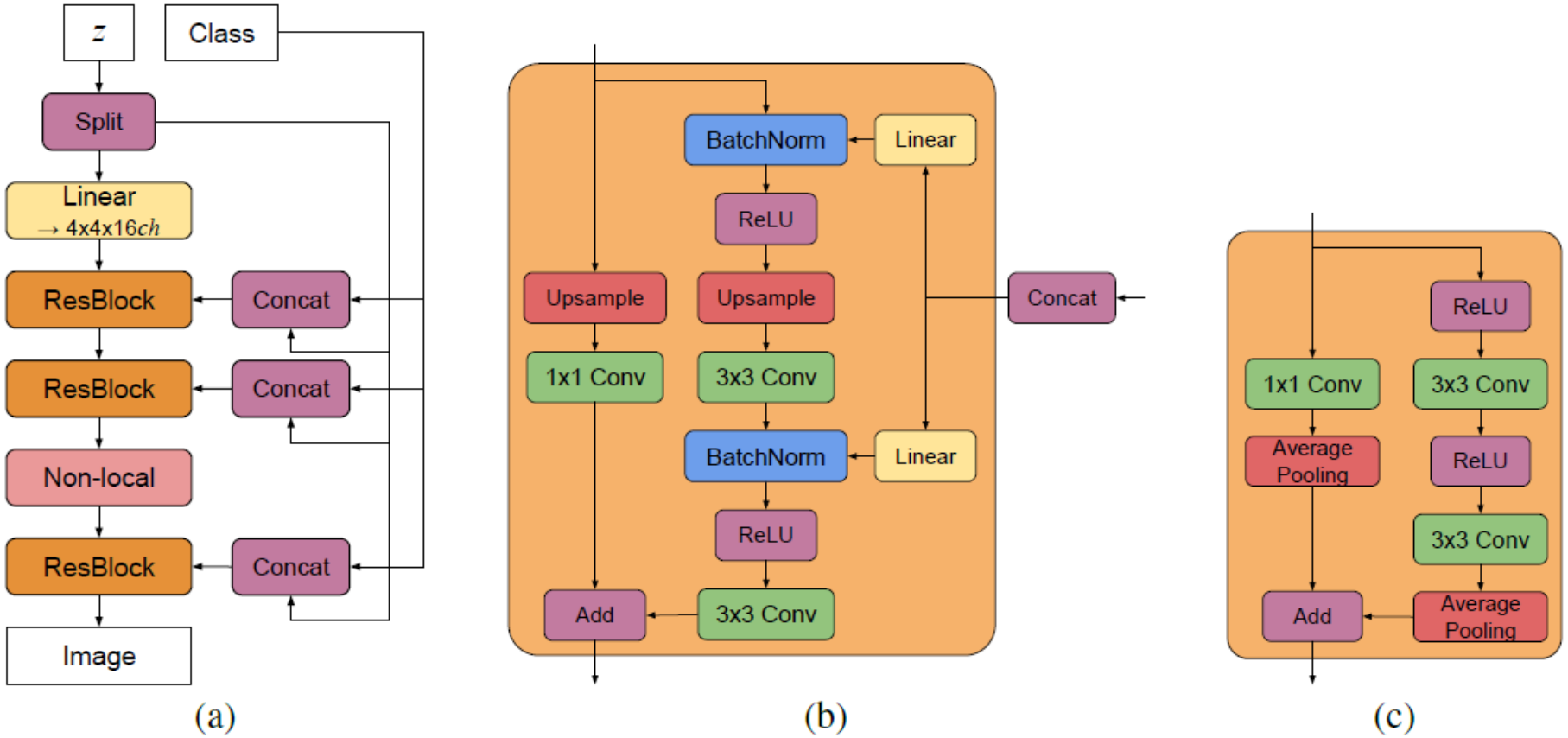


Figure 15: (a) A typical architectural layout for BigGAN's **G**; details are in the following tables. (b) A Residual Block (*ResBlock up*) in BigGAN's **G**. (c) A Residual Block (*ResBlock down*) in BigGAN's **D**.

Truncation Trick

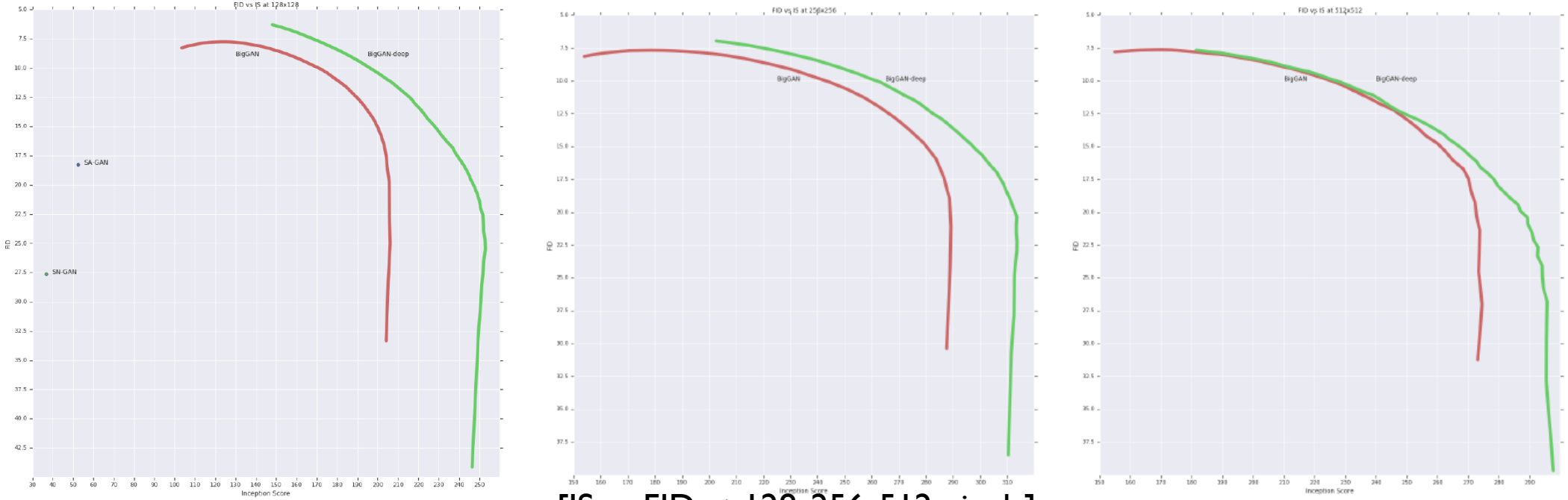
- Taking a model trained with $z \sim N(0, I)$ and sampling z from a ***truncated normal*** immediately provides a boost to IS and FID.
- ***Truncation Trick*** : truncating a z vector by resampling the values with magnitude above a chosen threshold lead to improvement in individual sample the values with magnitude above a chosen threshold leads to improvement in individual sample quality at the cost of reduction in overall sample variety.
- ***Truncation Trick*** allows fine-grained, post-hoc selection of the trade-off between sample quality and variety for a given G.

(a) (b)

Figure 2: (a) The effects of increasing truncation. From left to right, the threshold is set to 2, 1, 0.5, 0.04. (b) Saturation artifacts from applying truncation to a poorly conditioned model.

FID vs IS

- As IS does not penalize lack of variety in class-conditional models, reducing the truncation threshold leads to a direct increase in IS(analogous to precision).
- FID penalizes lack of variety(analogous to recall) but also rewards precision, so initially see a moderate improvement in FID, but as truncation approaches zero and variety diminishes, the FID sharply drops.



[IS vs FID at 128, 256, 512 pixels]

Inception Score

- ***Inception Score***
  - Use Inception network to classify the generated images and predict $P(y|x)$
  - $P(y|x)$ – where y is the label and x is the generated data.
  - $P(y)$ - marginal probability

$$\int_z p(y|x = G(z))dz$$

  - Inception Score

$$IS(G) = \exp(E_{x\sim p_a} D_{KL}(p(y|x) \,||\, p(y)))$$

- ***Frechet Inception Distance (FID)***
  - Use Inception network to extract features from an intermediate layer. Then model the data distribution for these features using a multivariate Gaussian distribution with mean μ and covariance Σ. (real images x / generated images g)

$$FID(x, g) = \|\mu_x - \mu_g\|_2^2 + Tr(\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{\frac{1}{2}})$$

Orthogonal Regularization

- The distribution shift caused by sampling with different latents than those seen in training is problematic for many models. Some of larger models are not amenable to truncation, producing saturation artifacts.
- To counteract this, seek to enforce amenability to truncation by conditioning G to be smooth, so that the full space of z will map to good output samples.
  → *Orthogonal Regularization*
- *Orthogonal Regularization* (Brock et al., 2017)

$$R_\beta(W) = \beta \|W^\top W - I\|_{\mathrm{F}}^2,$$

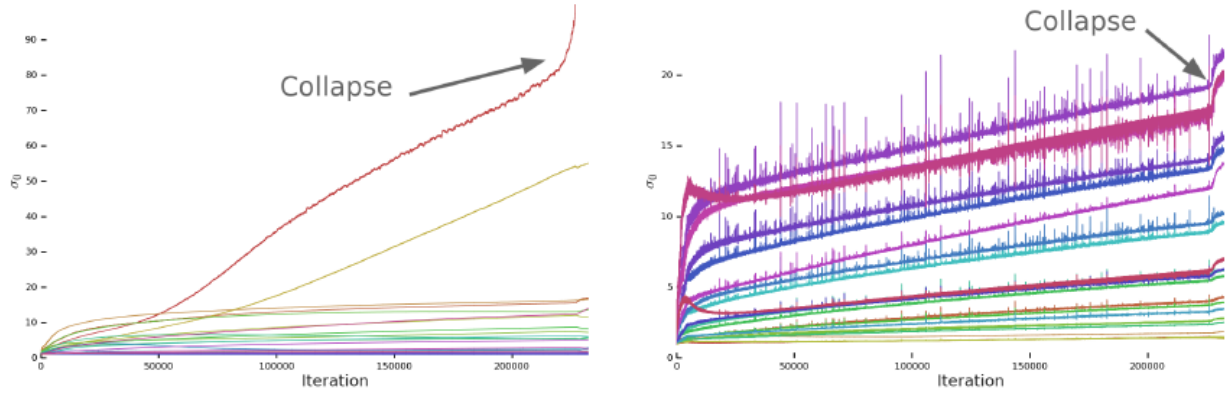
- *Orthogonal Regularization* (BigGAN version)

$$R_\beta(W) = \beta \|W^\top W \odot (1 - I)\|_{\mathrm{F}}^2,$$

- Without Orthogonal Regularization, only 16% of models are amenable to truncation, compared to 60% when trained with Orthogonal Regularization.

Characterizing Instability

- The instabilities authors observe occur for settings which are stable at small scale, necessitating direct analysis at large scale.
- Authors found the top 3 singular values $\sigma_0, \sigma_1, \sigma_2$ of each weight matrix to be the most informative. They can be computed using the Alrnoldi iteration method, which extends the power iteration method to estimation of additional singular vectors and values.



(a) **G** (b) **D**

Figure 3: A typical plot of the first singular value σ_0 in the layers of **G** (a) and **D** (b) before Spectral Normalization. Most layers in **G** have well-behaved spectra, but without constraints a small subset grow throughout training and explode at collapse. **D**'s spectra are noisier but otherwise better-behaved. Colors from red to violet indicate increasing depth.

Characterizing Instability

- The instabilities authors observe occur for settings which are stable at small scale, necessitating direct analysis at large scale.
- Authors found the top 3 singular values $\sigma_0, \sigma_1, \sigma_2$ of each weight matrix to be the most informative. They can be computed using the Alrnoldi iteration method, which extends the power iteration method to estimation of additional singular vectors and values.
- Generator regularization:

$$W = W - \max(0, \sigma_0 - \sigma_{clamp}) v_0 u_0^\top,$$

- Discriminator regularization(zero-centered gradient penalty):

$$R_1 := \frac{\gamma}{2} \mathbb{E}_{p_{\mathcal{D}}(x)} \left[\|\nabla D(x)\|_F^2 \right].$$

Experiments: Evaluation on ImageNet

(a) 128×128 (b) 256×256 (c) 512×512 (d)

Figure 4: Samples from our BigGAN model with truncation threshold 0.5 (a-c) and an example of class leakage in a partially trained model (d).

Model	Res.	FID/IS	(min FID) / IS	FID / (valid IS)	FID / (max IS)
SN-GAN	128	27.62/36.80	N/A	N/A	N/A
SA-GAN	128	18.65/52.52	N/A	N/A	N/A
BigGAN	128	$8.7 \pm .6/98.8 \pm 3$	$7.7 \pm .2/126.5 \pm 0$	$9.6 \pm .4/166.3 \pm 1$	$25 \pm 2/206 \pm 2$
BigGAN	256	$8.7 \pm .1/142.3 \pm 2$	$7.7 \pm .1/178.0 \pm 5$	$9.3 \pm .3/233.1 \pm 1$	$25 \pm 5/291 \pm 4$
BigGAN	512	8.1/144.2	7.6/170.3	11.8/241.4	27.0/275
BigGAN-deep	128	$5.7 \pm .3/124.5 \pm 2$	$6.3 \pm .3/148.1 \pm 4$	$7.4 \pm .6/166.5 \pm 1$	$25 \pm 2/253 \pm 11$
BigGAN-deep	256	$6.9 \pm .2/171.4 \pm 2$	$7.0 \pm .1/202.6 \pm 2$	$8.1 \pm .1/232.5 \pm 2$	$27 \pm 8/317 \pm 6$
BigGAN-deep	512	7.5/152.8	7.7/181.4	11.5/241.5	39.7/298

Table 2: Evaluation of models at different resolutions. We report scores without truncation (Column 3), scores at the best FID (Column 4), scores at the IS of validation data (Column 5), and scores at the max IS (Column 6). Standard deviations are computed over at least three random initializations.

Experiments: Evaluation on JET-300M

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
96	207.9	✓	✓	✓	18.84	27.86	17.1/23.3	51.6/38.1
128	355.7	✓	✓	✓	13.75	30.61	13.0/28.0	46.2/47.8

Table 3: BigGAN results on JFT-300M at 256×256 resolution. The *FID* and *IS* columns report these scores given by the JFT-300M-trained Inception v2 classifier with noise distributed as $z \sim \mathcal{N}(0, I)$ (non-truncated). The *(min FID) / IS* and *FID / (max IS)* columns report scores at the best FID and IS from a sweep across truncated noise distributions ranging from $\sigma = 0$ to $\sigma = 2$. Images from the JFT-300M validation set have an IS of 50.88 and FID of 1.94.

Thank you!