Internalized Biases in Fréchet Inception Distance

Steffen Jung, Margret Keuper

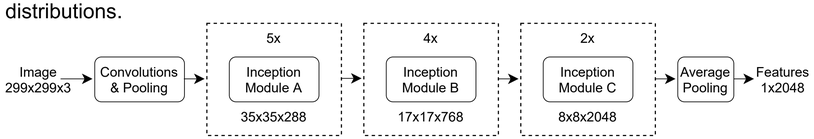
max planck institut informatik

UNIVERSITÄT SIEGEN

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation / Contribution

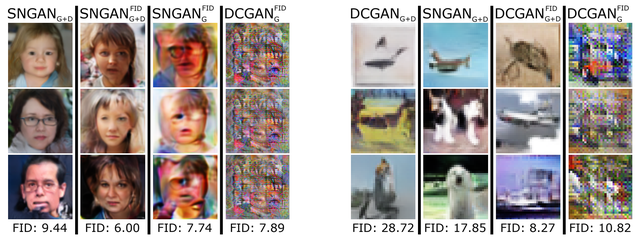
Fréchet Inception Distance (FID) [1] is a commonly used metric to decide if one image generating model is superior to another. To compute FID between two image datasets (datasets denoted by $\mathbf{X}_i$), image features are extracted by sampling images from those datasets and feeding them into the pretrained Inception v3 (ϕ) [2] image classification network. FID is then computed via the Wasserstein-2 distance between both feature distributions.



$$\mathrm{FID} = ||\mu_1 - \mu_2||_2^2 + tr(\Sigma_1) + tr(\Sigma_2) - 2 \cdot tr(\sqrt{\Sigma_1 \Sigma_2})$$

$$\mu_i = \frac{1}{|\mathbf{X}_i|} \sum_{x \in \mathbf{X}_i} \phi(x) \qquad \Sigma_i = \frac{(\phi(\mathbf{X}_i) - \mu_i)^T (\phi(\mathbf{X}_i) - \mu_i)}{n-1}$$

Here, we show images generated by different models trained on FFHQ (**left**) and CIFAR10 (**right**) with their respective FIDs (lower is better). Would you agree with these rankings?



Our contributions are as follows:

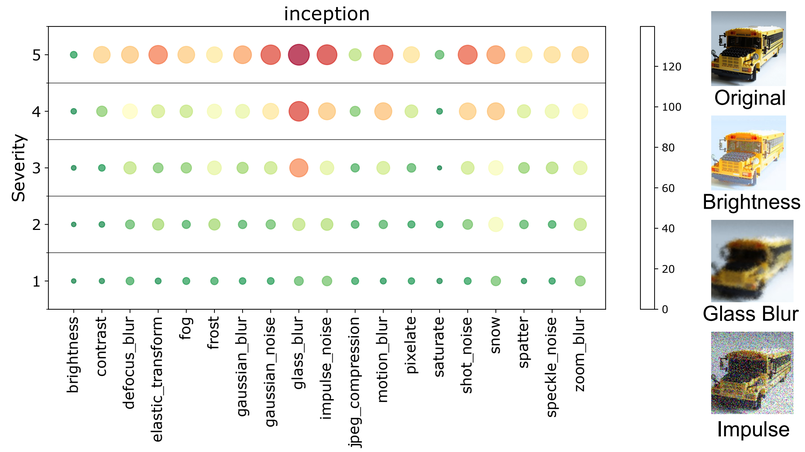
1) We investigate the bias of FID inherited by the feature extractor Inception v3.
2) We show that minimizing FID does not always imply improved image quality / fidelity.
3) We show that exchanging Inception v3 with other common image classification networks does not solve the problem.

1) Biases in Inception v3

FID is related to model robustness: A common way to equip image classification networks with robustness against distributional shifts is to augment the training data. Inception v3 was trained with the following augmentation pipeline: a) cropping with random scale and aspect ratio, b) random flipping horizontally, and c) random color distortions in terms of hue, saturation, brightness, and contrast. On ImageNet-C we can uncover which types of corruptions influence the feature distribution provided by Inception v3 (and therefore FID) to which degree. Here, we make the following observations:

a) Inception v3 is mostly robust to deviations in brightness, saturation, and contrast up to a certain degree.
b) Inception v3 is sensitive to noise that acts on the high frequency spectrum [3].

These observations lead us to the assumption that Inception v3 has a bias towards extracting features based on edges and textures rather than color and intensity information.



Takeaway: When used as ranking metric, FID might prefer generative models reproducing textures well.

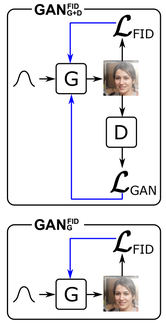
2) Reinforcing the Bias: Minimizing FID

If FID is a metric that aligns with human judgement, we can assume that the visual appearance of images generated by any GAN model should improve if we optimize the generator by minimizing FID. Can this assumption hold?

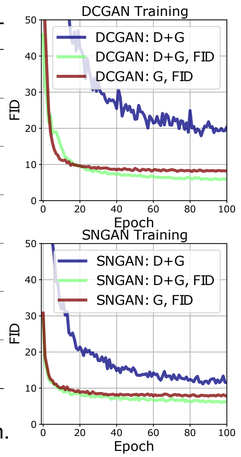
To verify, we train GANs in three training procedures:
1) G+D: Vanilla GAN training
2) G+D (FID): Adds FID as additional loss for generator
3) G (FID): Drops discriminator network

We make the following observations:

a) Adding FID loss "stabilizes" training in terms of FID.
b) Subjectively, image quality between G+D and G+D (FID) is similar.
c) Only minimizing FID leads to spatially incoherent feature distributions, but improved FID metric.



Data	Resolution	Model	FID↓	Images
FFHQ	64^2	DCGAN$_{G+D}$	14.86	
		DCGAN$^{FID}_{G+D}$	5.38	
		DCGAN$^{FID}_{G}$	7.89	
		DCGAN-Up$^{FID}_{G}$	7.61	
		SNGAN$_{G+D}$	9.44	
		SNGAN$^{FID}_{G+D}$	6.00	
		SNGAN$^{FID}_{G}$	7.74	

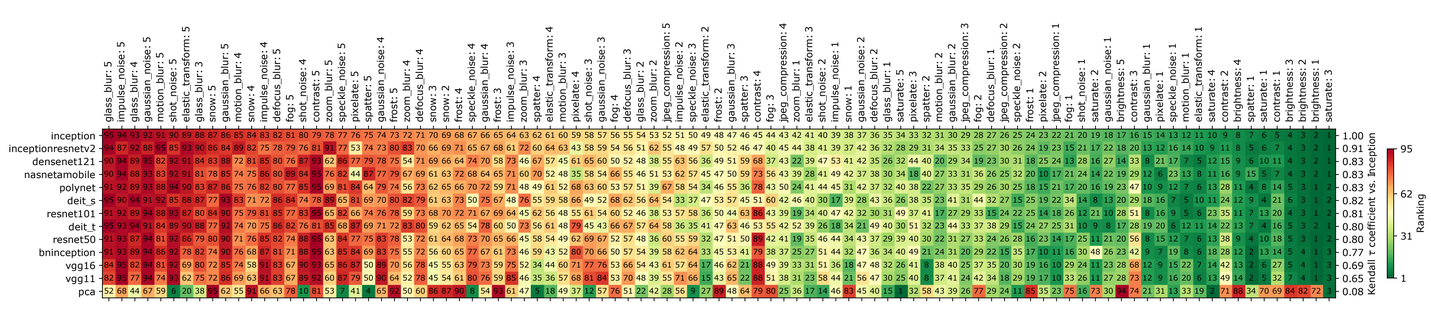


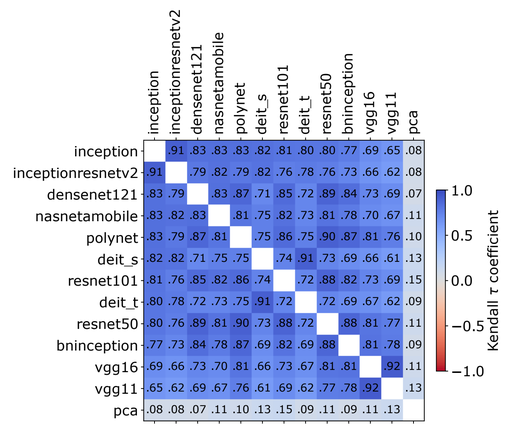
Takeaway: FID is not aligned with human perception.

3) Exchanging the Backbone: Can we just use another network?

We test this hypothesis by computing FIDs between different corruptions and severities for a range of different classification networks trained on ImageNet. We observe that exchanging the network changes the ranking we measure between different corruptions, however, biases present in Inception v3 are also present in other classification networks. We additionally show the ranking that PCA features produce.

Takeaway: Discriminative features provided by image classification networks might not be sufficient to provide the basis of a meaningful metric.





Conclusion

We showed that the way Inception v3 was trained enforced it to become robust to certain corruptions while sensitive to others. These biases influence the ranking produced by FID. Additionally, we showed that generators trained with FID as loss produce images with substantially improved FIDs, but worse visual appearance. With further research, we hope to close the gap towards a humanly aligned and unbiased metric that enables fair rankings.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In NIPS, 2017.

[2] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In CVPR, 2016.

[3] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A Fourier Perspective on Model Robustness in Computer Vision. In NIPS, 2019.