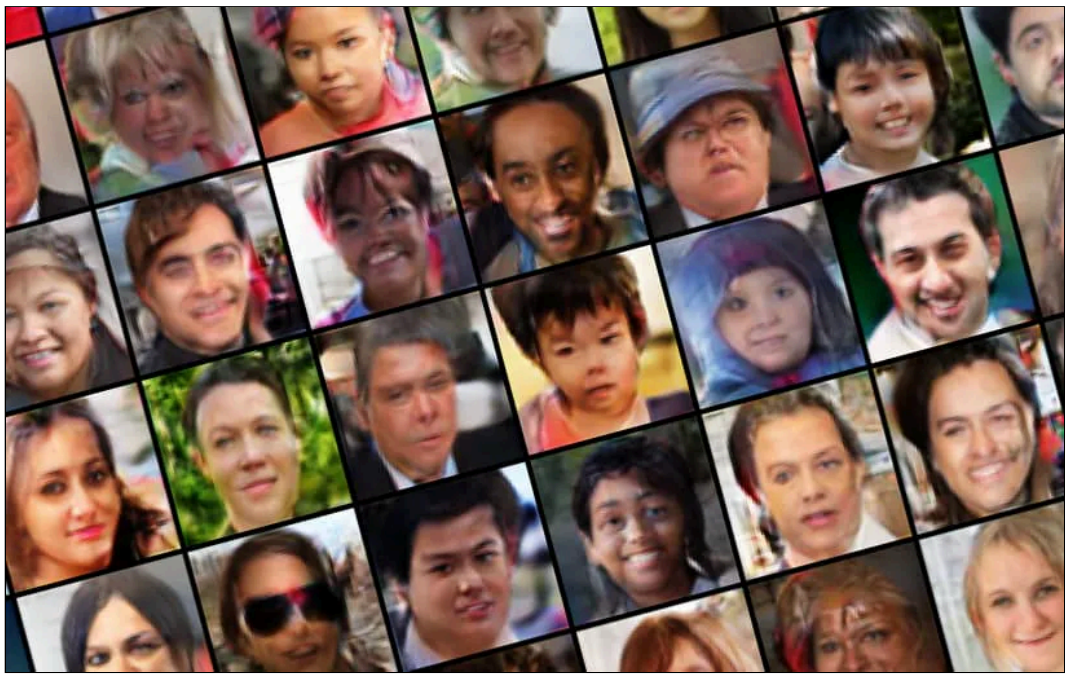
Image Synthesis Sector Has Adopted a Flawed Metric, Research Claims

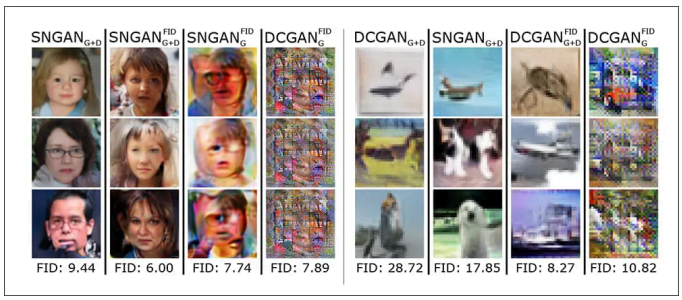
Originally published at https://www.unite.ai/image-synthesis-sector-has-adopted-a-flawed-metric-research-claims/



2021 has been a year of unprecedented progress and a furious pace-of-publication in the image synthesis sector, offering a stream of new innovations and improvements in technologies that are capable of reproducing human personalities through neural rendering, deepfakes, and a host of novel approaches.

However, researchers from Germany now claim that the standard used to automatically judge the realism of synthetic images is fatally flawed; and that the hundreds, even thousands of researchers around the world that rely on it to cut the cost of expensive human-based results evaluation may be heading down a blind alley.

In order to demonstrate how the standard, *Fréchet Inception Distance (FID)*, does not measure up to human standards for evaluating images, the researchers deployed their own GANs, optimized to FID (now a common metric). They found that FID is following its own obsessions, based on underlying code with a very different remit to that of image synthesis, and that it routinely fails to achieve a 'human' standard of discernment:



FID scores (lower is better) for images generated by various models using standard datasets and architectures. The researchers of the new paper pose the question 'Would you agree with these rankings?'. Source: https://openreview.net/pdf?id=mLG96UpmbYz

In addition to its assertion that FID is not fit for its intended task, the paper further suggests that 'obvious' remedies, such as switching out its internal engine for competing engines, will simply swap one set of biases for another. The authors suggest that it now falls to new research initiatives to develop better metrics to assess 'authenticity' in synthetically-generated photos.

The paper is titled *Internalized Biases in Fréchet Inception Distance*, and comes from Steffen Jung at the Max Planck Institute for Informatics at Saarland, and Margret Keuper, Professor for Visual Computing at the University of Siegen.

The Search for a Scoring System for Image Synthesis

As the new research notes, progress in image synthesis frameworks, such as GANs and encoder/decoder architectures, has outpaced methods by which the results of such systems can be judged. Besides being expensive and therefore difficult to scale, human evaluation of the output of these systems does not offer an empirical and reproducible method of assessment.

Therefore a number of metric frameworks have emerged, including *Inception Score* (IS), featured in the 2016 paper *Improved Techniques for Training GANs*, co-authored by GAN inventor, Ian Goodfellow.

The discrediting of the IS score as a broadly applicable metric for multiple GAN networks in 2018 led to the widespread adoption of FID in the GAN image synthesis community. However, like Inception Score, FID is based on Google's Inception v3 image classification network (IV3).

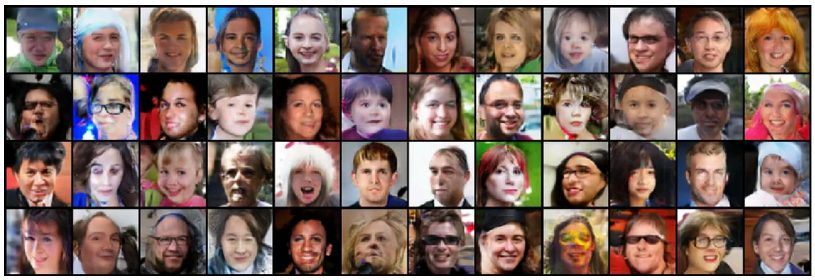
The authors of the new paper argue that Fréchet Inception Distance propagates damaging biases in IV3, leading to unreliable classification of image quality.

Since FID can be incorporated into a machine learning framework as a discriminator (an embedded 'judge' that decides if the GAN is doing well, or should 'try again'), it needs to accurately represent the standards that a human would apply when evaluating the images.

Fréchet Inception Distance

FID compares how features are distributed across the training dataset used to create a GAN (or similar functionality) model, and the results of that system.

Therefore, if a GAN framework is trained on 10,000 images of (for example) celebrities, FID compares the original (real) images to the fake images produced by the GAN. The lower the FID score, the nearer the GAN has gotten to 'photorealistic' images, according to FID's criteria.



From the paper, results of a GAN trained on FFHQ64, a subset of NVIDIA's very popular FFHQ dataset. Here, though the FID score is a wonderfully low 5.38, the convincing to the average human.

The problem, the authors contend, is that Inception v3, whose assumptions power Fréchet Inception Distance, is not looking in the right places – at least, not when considering the task at hand.

Inception V3 is trained on the ImageNet object recognition challenge, a task that is arguably at odds with the way that the aims of image synthesis have evolved in recent years. IV3 challenges the robustness of a model by performing data augmentation: it flips images randomly, crops them to a random scale between 8-100%, changes the aspect ratio (in a range from 3/4 to 4/3), and randomly injects color distortions relating to brightness, saturation, and contrast.

The Germany-based researchers have found that IV3 has a tendency to favor the extraction of edges and textures, rather than color and intensity information, which would be more meaningful indices of authenticity for synthetic images; and that its original purpose of object detection has therefore been inappropriately sequestered for an unsuitable task. The authors state*:

> *'[Inception v3] has a bias towards extracting features based on edges and textures rather than color and intensity information. This aligns with its augmentation pipeline that introduces color distortions, but keeps high frequency information intact (in contrast to, for example, augmentation with Gaussian blur).*
>
> *'Consequently, FID inherits this bias.* ***When used as ranking metric, generative models reproducing textures well might be preferred over models that reproduce color distributions well****.'*

Data and Method

To test their hypothesis, the authors trained two GAN architectures, DCGAN and SNGAN, on NVIDIA's FFHQ human face dataset, downsampled to 64^2 image resolution, with the derived dataset called FFHQ64.

Three GAN training procedures were pursued: GAN G+D, a standard discriminator-based network; GAN FID|G+D, where FID performs as an additional discriminator; and GAN FID|G. where the GAN is entirely powered by the rolling FID score.

Technically, the authors note, FID loss should stabilize the training, and potentially even be able to *completely substitute* the discriminator (as it does in #3, GAN FID|G), while outputting human-pleasing results.

In practice, the results are rather different, with – the authors hypothesize – the FID-assisted models 'overfitting' on the wrong metrics. The researchers note:

> *'We hypothesize that the generator learns to produce unsuitable features to match the training data distribution. This observation becomes more severe in the case of [GAN FID|G]. Here, we notice that the missing discriminator leads to spatially incoherent feature distributions. For example [SNGAN FID|G] adds mostly single eyes and aligns facial characteristics in a daunting manner.'*

Examples of faces produced by SNGAN FID|G.

The authors conclude*:

> *'While human annotators would surely prefer images produced by SNGAN D+G over SNGAN FID|G (in cases where data fidelity is preferred over art), we see that this is not reflected by FID.* ***Hence, FID is not aligned with human perception****.*
>
> *'We argue that discriminative features provided by image classification networks are not sufficient to provide the basis of a meaningful metric.'*

No Easy Alternatives

The authors also found that swapping Inception V3 for a similar engine did not alleviate the problem. In substituting IV3 with 'an extensive choice of different classification networks', which were tested against ImageNet-C (a subset of ImageNet designed to benchmark commonly-generated corruptions and perturbations in output images from image synthesis frameworks), the researchers could not substantially improve their results:

> *'[Biases] present in Inception v3 are also widely present in other classification networks. Additionally, we see that different networks would produce different rankings in-between corruption types.'*

The authors conclude the paper with the hope that ongoing research will develop a 'humanly-aligned and unbiased metric' capable of enabling a fairer rank for image generator architectures.

** Authors' emphasis.*

First published 2oth December 2021, 1pm GMT+2.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More