Fast Fréchet Inception Distance

Frederik Hvilshøj

May 5, 2021

Data-Intensive Systems Group, Aarhus University

Paper URL

Slides URL

Today (45 minutes):

1. Comparing Generative Models

2. Metrics

3. Fast Fréchet Inception Distance

4. What is it Good For? ♬

Comparing Generative Models

Which one is better?

"Real"

$$x_1, \ldots, x_n \sim P_{\text{real}}$$

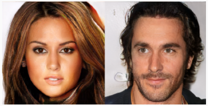
Which one is better?

"Real"

$x_1, \dots, x_n \sim P_{\text{real}}$

"Fake 1"

$x_1^{(1)}, \dots, x_n^{(1)} \sim P_{G_1}$



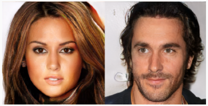
Which one is better?

“Real”

$x_1, \dots, x_n \sim P_{\mathrm{real}}$

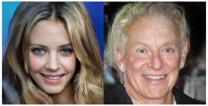
“Fake 1”

$x_1^{(1)}, \dots, x_n^{(1)} \sim P_{G_1}$



“Fake 2”

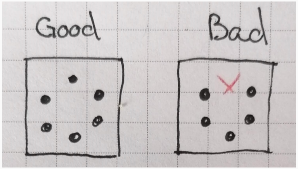
$x_1^{(2)}, \dots, x_n^{(2)} \sim P_{G_2}$



Desiderata

1. Fast

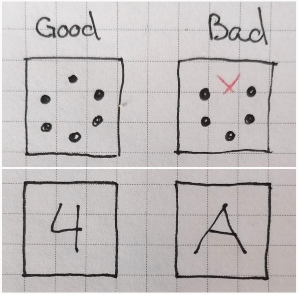
Desiderata



1. Fast
2. Diversity

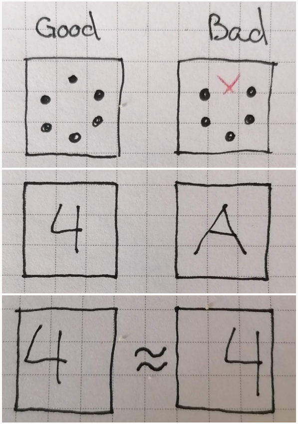
Desiderata

1. Fast
2. Diversity
3. Classifiable



Desiderata

1. Fast
2. Diversity
3. Classifiable
4. Translation invariant



Metrics

First Idea: Inception Score [1]

[1] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X., 2016. Improved techniques for training gans. arXiv preprint arXiv:1606.03498.

Inception Score

Idea: Use Inception network $N(x) = p(y|x)$ to "analyze" the generated data.

Inception Score

Idea: Use Inception network $N(x) = p(y|x)$ to "analyze" the generated data.

Easily classifiable If N is confident in predictions $\Rightarrow$ better sample quality.

Inception Score

Idea: Use Inception network $N(x) = p(y|x)$ to "analyze" the generated data.

Easily classifiable If N is confident in predictions $\Rightarrow$ better sample quality.

Diversity If all the classes are represented, the samples are diverse.

Inception Score

Easily classifiable If N is confident in predictions $\Rightarrow$ better sample quality.

Diversity If all the classes are represented, the samples are diverse.

(1)

(2)

Inception Score

Easily classifiable If N is confident in predictions $\Rightarrow$ better sample quality.

Diversity If all the classes are represented, the samples are diverse.

$$IS(X) = \exp\left\{ \quad \mathbb{E}_x \left[\quad \mathrm{KL}\left(p(y \mid x) \| p(y)\right) \quad \right] \quad \right\} \tag{1}$$

$$\tag{2}$$

Inception Score

Easily classifiable If N is confident in predictions $\Rightarrow$ better sample quality.

Diversity If all the classes are represented, the samples are diverse.

$$IS(X) = \exp\{ \mathbb{E}_x [\mathrm{KL}(p(y \mid x) \| p(y))] \} \tag{1}$$
$$= \exp\{ \mathrm{H}(y) - \mathbb{E}_x [\mathrm{H}(y|x)] \} \tag{2}$$

Inception Score

Easily classifiable If N is confident in predictions $\Rightarrow$ better sample quality.

Diversity If all the classes are represented, the samples are diverse.

$$IS(X) = \exp\left\{ \quad \mathbb{E}_x \left[\quad \mathrm{KL}\left(p(y \mid x) \| p(y)\right) \quad \right] \quad \right\} \tag{1}$$

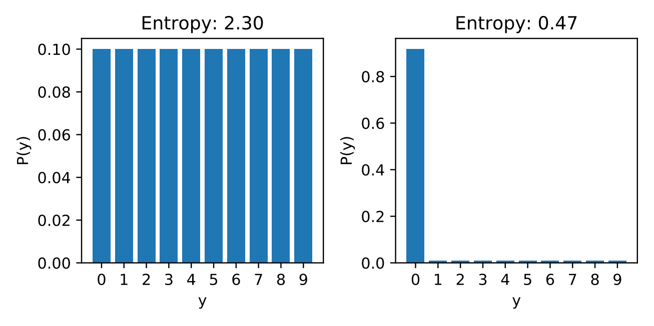
$$= \exp\left\{ \quad \mathrm{H}(y) - \mathbb{E}_x \left[\mathrm{H}(y|x)\right] \quad \right\} \tag{2}$$

Inception Score

$$IS(X) = \exp\{\ \mathrm{H}(y) \quad - \quad \mathbb{E}_x[\,\mathrm{H}(y|x)\,]\ \}$$

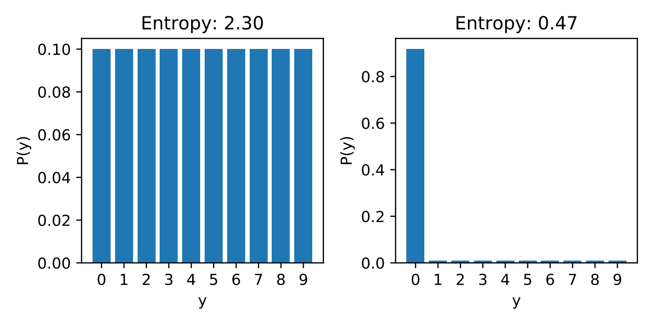
Inception Score

$$IS(X) = \exp\{\ \mathrm{H}(y) - \mathbb{E}_x[\,\mathrm{H}(y|x)\,]\ \}$$



Inception Score

$$IS(X) = \exp\{ \underbrace{\mathrm{H}(y)}_{\text{Maximize}} - \mathbb{E}_x[\underbrace{\mathrm{H}(y|x)}_{\text{Minimize}}] \} \tag{2}$$



Inception Score

Correlates with human judgement but doesn't take P_d into account!

Second Idea: Fréchet Inception Distance [2]

[2] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. arXiv preprint arXiv:1706.08500.

Fréchet Inception Distance

Idea: *Compare Inception network encodings between P_d and P_f.*

Fréchet Inception Distance

Idea: *Compare Inception network encodings between P_d and P_f.*

Given means μ_{r}, μ_{f} and covariances Σ_r, Σ_f of Inception encodings, the Fréchet Inception Distance (FID) is defined as

Fréchet Inception Distance

Idea: Compare Inception network encodings between P_d and P_f.

Given means μ_r, μ_f and covariances Σ_r, Σ_f of Inception encodings, the Fréchet Inception Distance (FID) is defined as

$$\mathrm{FID}(X_r, X_f) = W_2^2\left(\mathcal{N}\{\mu_r, \Sigma_r\}, \mathcal{N}\{\mu_f, \Sigma_f\}\right) \tag{3}$$

$$\tag{4}$$

Fréchet Inception Distance

Idea: Compare Inception network encodings between P_d and P_f.

Given means μ_r, μ_f and covariances Σ_r, Σ_f of Inception encodings, the Fréchet Inception Distance (FID) is defined as

$$\mathrm{FID}(X_r, X_f) = W_2^2\left(\mathcal{N}\{\mu_r, \Sigma_r\}, \mathcal{N}\{\mu_f, \Sigma_f\}\right) \tag{3}$$

$$= \underbrace{\|\mu_r - \mu_f\|_2^2}_{\mathcal{O}(d)} + \qquad + \qquad - 2 \tag{4}$$

Fréchet Inception Distance

Idea: Compare Inception network encodings between P_d and P_f.

Given means μ_{r}, μ_{f} and covariances Σ_r, Σ_f of Inception encodings, the Fréchet Inception Distance (FID) is defined as

$$\mathrm{FID}(X_r, X_f) = W_2^2\left(\mathcal{N}\{\mu_r, \Sigma_r\}, \mathcal{N}\{\mu_f, \Sigma_f\}\right) \tag{3}$$

$$= \underbrace{\|\mu_r - \mu_f\|_2^2}_{\mathcal{O}(d)} + \underbrace{\mathrm{Tr}[\Sigma_r]}_{\mathcal{O}(d)} + \underbrace{\mathrm{Tr}[\Sigma_f]}_{\mathcal{O}(d)} - 2 \tag{4}$$

Fréchet Inception Distance

Idea: Compare Inception network encodings between P_d and P_f.

Given means μ_r, μ_f and covariances Σ_r, Σ_f of Inception encodings, the Fréchet Inception Distance (FID) is defined as

$$\mathrm{FID}(X_r, X_f) = W_2^2\left(\mathcal{N}\{\mu_r, \Sigma_r\}, \mathcal{N}\{\mu_f, \Sigma_f\}\right) \tag{3}$$

$$= \underbrace{\|\mu_r - \mu_f\|_2^2}_{\mathcal{O}(d)} + \underbrace{\mathrm{Tr}[\Sigma_r]}_{\mathcal{O}(d)} + \underbrace{\mathrm{Tr}[\Sigma_f]}_{\mathcal{O}(d)} - 2\mathrm{Tr}[\underbrace{\sqrt{\Sigma_r \Sigma_f}}_{\mathcal{O}(d^3)}] \tag{4}$$

Fréchet Inception Distance

Idea: Compare Inception network encodings between P_d and P_f.

Given means μ_r, μ_f and covariances Σ_r, Σ_f of Inception encodings, the Fréchet Inception Distance (FID) is defined as

$$\mathrm{FID}(X_r, X_f) = W_2^2\left(\mathcal{N}\{\mu_r, \Sigma_r\}, \mathcal{N}\{\mu_f, \Sigma_f\}\right) \tag{3}$$

$$= \underbrace{\|\mu_r - \mu_f\|_2^2}_{\mathcal{O}(d)} + \underbrace{\mathrm{Tr}[\Sigma_r]}_{\mathcal{O}(d)} + \underbrace{\mathrm{Tr}[\Sigma_f]}_{\mathcal{O}(d)} - 2\mathrm{Tr}[\underbrace{\sqrt{\Sigma_r \Sigma_f}}_{\mathcal{O}(d^3)}] \tag{4}$$

Today this is state-of-the-art.

Computing $\mathrm{Tr}[\sqrt{\Sigma_r \Sigma_f}]$

Previous method explicitely computes $\sqrt{\Sigma_r \Sigma_f}$ and then computes the trace:

Computing $\mathrm{Tr}[\sqrt{\Sigma_r \Sigma_f}]$

Previous method explicitely computes $\sqrt{\Sigma_r \Sigma_f}$ and then computes the trace:

Input : Σ_r, Σ_f
Output: $\mathrm{Tr}[C] = \mathrm{Tr}\left[\sqrt{\Sigma_r \Sigma_f}\right]$

Computing $\mathrm{Tr}[\sqrt{\Sigma_r \Sigma_f}]$

Previous method explicitely computes $\sqrt{\Sigma_r \Sigma_f}$ and then computes the trace:

Input : Σ_r, Σ_f

Output: $\mathrm{Tr}[C] = \mathrm{Tr}\left[\sqrt{\Sigma_r \Sigma_f}\right]$

1 $Q, V \leftarrow$ *SchurDecompose*(A) ; /* $QVQ^\intercal = A$ */

Computing $\mathrm{Tr}[\sqrt{\Sigma_r \Sigma_f}]$

Previous method explicitely computes $\sqrt{\Sigma_r \Sigma_f}$ and then computes the trace:

Input : Σ_r, Σ_f

Output: $\mathrm{Tr}[C] = \mathrm{Tr}\left[\sqrt{\Sigma_r \Sigma_f}\right]$

$Q, V \leftarrow$ ***SchurDecompose***(A); /* $QVQ^\intercal = A$ */

$U \leftarrow$ ***TriangSqrt***(V); /* $V = U^2$ */

Computing $\mathrm{Tr}[\sqrt{\Sigma_r \Sigma_f}]$

Previous method explicitely computes $\sqrt{\Sigma_r \Sigma_f}$ and then computes the trace:

Input : Σ_r, Σ_f

Output: $\mathrm{Tr}[C] = \mathrm{Tr}\left[\sqrt{\Sigma_r \Sigma_f}\right]$

```
Q, V ← SchurDecompose(A) ;        /* QVQᵀ = A */
U ← TriangSqrt(V) ;               /* V = U² */
C ← QUQᵀ ;                        /* C = √(Σ_r Σ_f) */
```

Computing $\mathrm{Tr}[\sqrt{\Sigma_r \Sigma_f}]$

Previous method explicitely computes $\sqrt{\Sigma_r \Sigma_f}$ and then computes the trace:

Input : Σ_r, Σ_f

Output: $\mathrm{Tr}[C] = \mathrm{Tr}\left[\sqrt{\Sigma_r \Sigma_f}\right]$

$Q, V \leftarrow$ `SchurDecompose(`A`)` ; /* $QVQ^\intercal = A$ */

$U \leftarrow$ `TriangSqrt(`V`)` ; /* $V = U^2$ */

$C \leftarrow QUQ^\intercal$; /* $C = \sqrt{\Sigma_r \Sigma_f}$ */

return $\mathrm{Tr}[C]$;

Computing $\mathrm{Tr}[\sqrt{\Sigma_r \Sigma_f}]$

Previous method explicitely computes $\sqrt{\Sigma_r \Sigma_f}$ and then computes the trace:

Input : Σ_r, Σ_f
Output: $\mathrm{Tr}[C] = \mathrm{Tr}\left[\sqrt{\Sigma_r \Sigma_f}\right]$

1 $Q, V \leftarrow$ *SchurDecompose(A)*; /* $QVQ^\intercal = A$ */
2 $U \leftarrow$ *TriangSqrt(V)*; /* $V = U^2$ */
3 $C \leftarrow QUQ^\intercal$; /* $C = \sqrt{\Sigma_r \Sigma_f}$ */
4 **return** $\mathrm{Tr}[C]$;

Line [1-3] each takes cubic time!

Fast Fréchet Inception Distance

Idea 3: Don't compute $\mathrm{Tr}\left[\sqrt{\Sigma_r \Sigma_f}\right]$, use *eigenvalues* instead.[3]

[3] Mathiasen, A. and Hvilshøj, F., 2020. Fast Fréchet Inception Distance. arXiv preprint arXiv:2009.14075.

Lemmas

Lemma 1
$\mathrm{Tr}[\sqrt{A}] = \sum_i |\sqrt{\lambda_i(A)}|$. [4]

[4] There are some nuances here, please refer to paper for full details.

Lemmas

Lemma 1
$\mathrm{Tr}[\sqrt{A}] = \sum_i |\sqrt{\lambda_i(A)}|$. [4]

Lemma 2
Computing eigenvalues of $d \times d$ matrix A takes $\mathcal{O}(d^3)$ time. (similar time to compute $\sqrt{A}$)

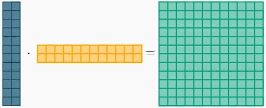
[4] There are some nuances here, please refer to paper for full details.

Lemma 3
The nonzero eigenvalues of AB are equal to those of BA, as long as the products are square. [5]

[5] Nakatsukasa, Y., 2019. The low-rank eigenvalue problem. arXiv preprint arXiv:1905.11490.

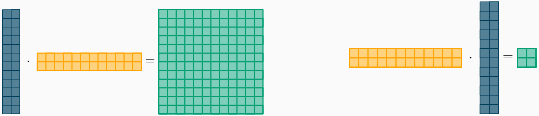
Lemmas

Lemma 3
The nonzero eigenvalues of AB are equal to those of BA, as long as the products are square. [5]



[5] Nakatsukasa, Y., 2019. The low-rank eigenvalue problem. arXiv preprint arXiv:1905.11490.

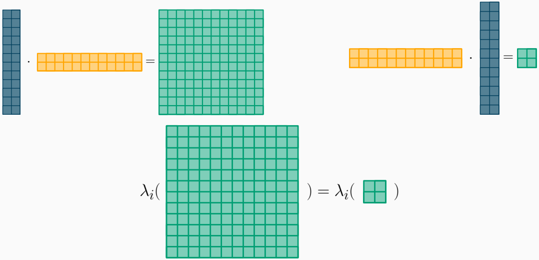
Lemma 3
The nonzero eigenvalues of AB are equal to those of BA, as long as the products are square. [5]



[5] Nakatsukasa, Y., 2019. The low-rank eigenvalue problem. arXiv preprint arXiv:1905.11490.

Lemmas

Lemma 3
The nonzero eigenvalues of AB are equal to those of BA, as long as the products are square. [5]



[5] Nakatsukasa, Y., 2019. The low-rank eigenvalue problem. arXiv preprint arXiv:1905.11490.

Fast Fréchet Inception Distance

High level idea: *Construct "small" matric M such that $\lambda_i(M)$ satisfy $\sum_i |\sqrt{\lambda_i(M)}| = \mathrm{Tr}[\sqrt{\Sigma_r \Sigma_f}]$. When M is sufficiently small, computing eigenvalues will be faster than computing $\sqrt{\Sigma_r \Sigma_f}$ explicitly.*

Fast Fréchet Inception Distance

Stack the m fake encoded samples into a $d \times m$ matrix X_f.

Fast Fréchet Inception Distance

Stack the m fake encoded samples into a $d \times m$ matrix X_f.

$$\Sigma_f = C_f C_f^\mathsf{T} \quad \text{where } C_r = \frac{1}{\sqrt{n-1}}\left(X_r - \mu_r \mathbf{1}_n\right) \tag{5}$$

Stack the m fake encoded samples into a $d \times m$ matrix X_f.

$$\Sigma_f = C_f C_f^\intercal \quad \text{where } C_r = \frac{1}{\sqrt{n-1}}\left(X_r - \mu_r \mathbf{1}_n\right) \tag{5}$$

Then

$$\Sigma_r \Sigma_f = \Sigma_r C_f C_f^\intercal \tag{6}$$

Fast Fréchet Inception Distance

$$\Sigma_r \Sigma_f = \Sigma_r C_f C_f^\intercal \tag{6}$$

Fast Fréchet Inception Distance

$$\Sigma_r \Sigma_f = \Sigma_r C_f C_f^\intercal \tag{6}$$

Using Lemma 3:

$$\lambda_i(\underbrace{\Sigma_r C_f C_f^\intercal}_{d \times d}) = \lambda_i(\underbrace{C_f^\intercal \Sigma_r C_f}_{m \times m}) \tag{7}$$

Fast Fréchet Inception Distance

$$\Sigma_r \Sigma_f = \Sigma_r C_f C_f^\mathsf{T} \tag{6}$$

Using Lemma 3:

$$\lambda_i(\underbrace{\Sigma_r C_f C_f^\mathsf{T}}_{d \times d}) = \lambda_i(\underbrace{C_f^\mathsf{T} \Sigma_r C_f}_{m \times m}) \tag{7}$$

Eigenvalue computations go from $\mathcal{O}(d^3)$ to $\mathcal{O}(m^3)$ (Lemma 2).

$$\Sigma_r \Sigma_f = \Sigma_r C_f C_f^{\mathsf{T}} \tag{6}$$

Using Lemma 3:

$$\lambda_i(\underbrace{\Sigma_r C_f C_f^{\mathsf{T}}}_{d \times d}) = \lambda_i(\underbrace{C_f^{\mathsf{T}} \Sigma_r C_f}_{m \times m}) \tag{7}$$

Eigenvalue computations go from $\mathcal{O}(d^3)$ to $\mathcal{O}(m^3)$ (Lemma 2).

Finally due to Lemma 1:

$$\mathrm{Tr}\left[\sqrt{\Sigma_r \Sigma_f}\right] = \sum_{i=1}^{m-1} \left|\sqrt{\lambda_i(C_f^{\mathsf{T}} \Sigma_r C_f)}\right| \tag{8}$$

Overall, we get runningtime

$$\mathrm{FID} = \underbrace{\|\mu_r - \mu_f\|_2^2}_{\mathcal{O}(d)} + \underbrace{\mathrm{Tr}[\Sigma_r + \Sigma_f]}_{\mathcal{O}(d)} - 2\sum_{i=1}^{m-1} |\underbrace{\sqrt{\lambda_i(C_f^T \Sigma_r C_f)}}_{\mathcal{O}(d^2 m + m^3)}| \quad (9)$$

What is it Good For? 🎵

The Greater Perspective

$$\mathrm{FID} = \underbrace{\|\mu_r - \mu_f\|_2^2 + \mathrm{Tr}[\Sigma_r + \Sigma_f]}_{\mathcal{O}(d)} - 2\sum_{i=1}^{m-1} \Big| \underbrace{\sqrt{\lambda_i(C_f^{\mathsf{T}}\Sigma_r C_f)}}_{\mathcal{O}(d^2 m + m^3)} \Big| \quad (9)$$

The Greater Perspective

$$\mathrm{FID} = \underbrace{\|\mu_r - \mu_f\|_2^2 + \mathrm{Tr}[\Sigma_r + \Sigma_f]}_{\mathcal{O}(d)} - 2\sum_{i=1}^{m-1} \Big|\underbrace{\sqrt{\lambda_i(C_f^\mathsf{T}\Sigma_r C_f)}}_{\mathcal{O}(d^2 m + m^3)}\Big| \quad (9)$$

During training, we typically have $n \gg d \gg m$

The Greater Perspective

$$\mathrm{FID} = \underbrace{\|\mu_r - \mu_f\|_2^2 + \mathrm{Tr}[\Sigma_r + \Sigma_f]}_{\mathcal{O}(d)} - 2 \sum_{i=1}^{m-1} \left| \underbrace{\sqrt{\lambda_i(C_f^\top \Sigma_r C_f)}}_{\mathcal{O}(d^2 m + m^3)} \right| \quad (9)$$

During training, we typically have $n \gg d \gg m$

Example 4
For GANs on ImageNet, test size (n) is 10 000, encodings (d) are 2048, and batch size (m) is typically 128.

The Greater Perspective

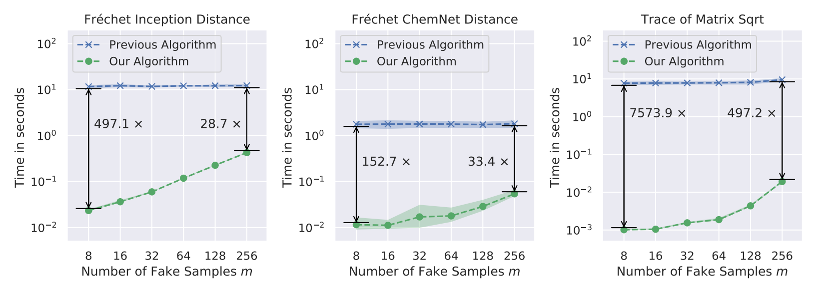
$$\mathrm{FID} = \underbrace{\|\mu_r - \mu_f\|_2^2 + \mathrm{Tr}[\Sigma_r + \Sigma_f]}_{\mathcal{O}(d)} - 2 \sum_{i=1}^{m-1} \left| \underbrace{\sqrt{\lambda_i(C_f^\top \Sigma_r C_f)}}_{\mathcal{O}(d^2 m + m^3)} \right| \quad (9)$$

During training, we typically have $n \gg d \gg m$

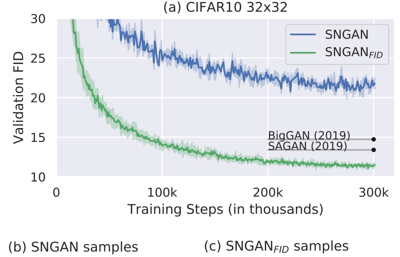
Example 4
For GANs on ImageNet, test size (n) is 10 000, encodings (d) are 2048, and batch size (m) is typically 128.

Let's use FID for optimizations!

Performance



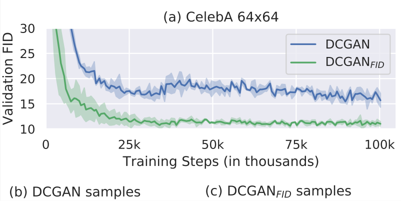
Minimizing FID



GAN

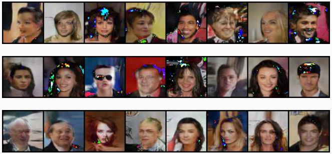
GAN_{FID}

Minimizing FID

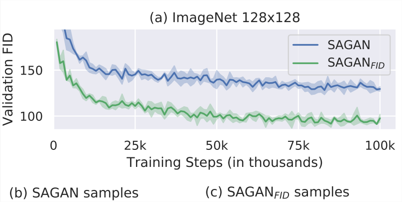


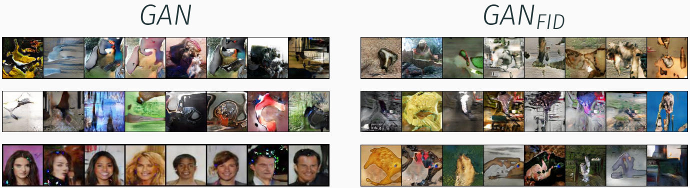
GAN

GAN_{FID}



Minimizing FID



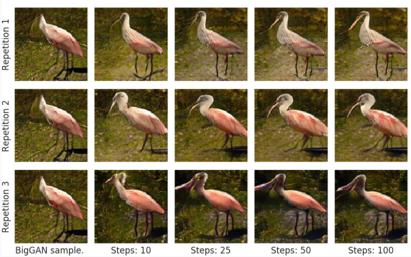


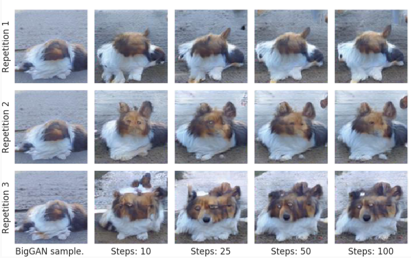
Can FID Loss Improve Generated Images?

What will happen if we just optimize for FID?

Can FID Loss Improve Generated Images?

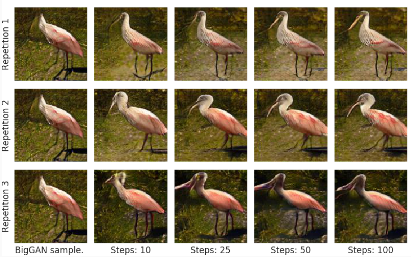
What will happen if we just optimize for FID?

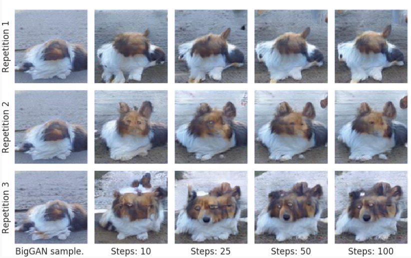




Can FID Loss Improve Generated Images?

What will happen if we just optimize for FID?





$$\underbrace{\left\|\mu_r - \mu_f\right\|_2^2}_{\text{mean difference}} + \underbrace{\mathrm{Tr}\left[\Sigma_r\right] + \mathrm{Tr}\left[\Sigma_f\right] - 2\mathrm{Tr}\left[\sqrt{\Sigma_r \Sigma_f}\right]}_{\text{covariance difference}} \tag{10}$$