Fréchet Inception Distance

Jul 15, 2018

> *TL;DR: FID measures the distance between the Inception-v3 activation distributions for generated and real samples*

I often run into things that are new to me (but not explained in that particular paper) when I'm reading. A recent example is Fréchet Inception Distance (FID), a method for **measuring the quality of generated image samples**.

Inception Score

Before FID, the Inception Score (IS) was the original method for measuring the quality of generated samples. (Salimans et al., 2016) proposed applying an Inception-v3 network pre-trained on ImageNet to generated samples and then comparing the conditional label distribution with the marginal label distribution:

$$\text{IS} = \exp\left(\mathbb{E}_{x\sim p_g} D_{KL}(p(y|x)||p(y))\right)$$

Ideally, the generator should:

1. Generate images ***with meaningful objects***, so that the conditional label distribution $p(y|x)$ is *low entropy*.
2. Generate ***diverse*** images, so that the marginal label distribution $p(y) = \int_x p(y|x)p_g(x)$ is *high entropy*.

Higher scores are better, corresponding to a larger KL-divergence between the two distributions.

Fréchet Inception Distance

The FID is supposed to improve on the IS by actually *comparing* the statistics of generated samples to real samples, instead of evaluating generated samples in a vacuum.[1] (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017) propose using the Fréchet distance between two multivariate Gaussians,

$$\text{FID} = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}),$$

where $X_r \sim \mathcal{N}(\mu_r, \Sigma_r)$ and $X_g \sim \mathcal{N}(\mu_g, \Sigma_g)$ are the 2048-dimensional activations of the Inception-v3 pool3 layer for real and generated samples respectively.[2]

Lower FID is better, corresponding to more similar real and generated samples as measured by the distance between their activation distributions.

Measuring progress

It's helpful to look at some of the IS and FID scores that have been reported to get a feel for what good/bad scores look like and see how different models compare. These scores are for unsupervised models on CIFAR-10:[3]

Paper	IS	FID
Real CIFAR-10 data (Salimans et al., 2016)	11.24	–
Unsupervised representation learning with deep convolutional generative adversarial networks (DCGAN) (Radford, Metz, & Chintala, 2015)	6.16[4]	37.1[5]
Conditional image generation with PixelCNN decoders (van den Oord et al., 2016)	4.60[6]	65.9[6]
Adversarially learned inference (ALI) (Dumoulin et al., 2016)	5.34[7]	–
Improved techniques for training GANs (Salimans et al., 2016)	6.86	–
Improving generative adversarial networks with denoising		

feature matching (Warde-Farley & Bengio, 2016)	7.72	–
Learning to generate samples from noise through infusion training (Bordes, Honari, & Vincent, 2017)	4.62	–
BEGAN: Boundary equilibrium generative adversarial networks (Berthelot, Schumm, & Metz, 2017)	5.62	–
MMD GAN: Towards deeper understanding of moment matching network (Li, Chang, Cheng, Yang, & Póczos, 2017)	6.17	–
Improved training of Wasserstein GANs (Gulrajani, Ahmed, Arjovsky, Dumoulin, & Courville, 2017)	7.86	–
Coulomb GANs: Provably optimal Nash equilibrium via potential fields (Unterthiner et al., 2017)	–	27.3
GANs trained by a two time-scale update rule converge to a local Nash equilibrium (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017)	–	24.8
Autoregressive quantile networks for generative modeling (AIQN) (Ostrovski, Dabney, & Munos, 2018)	5.29	49.5
Spectral normalization for generative adversarial networks (SN-GAN) (Miyato, Kataoka, Koyama, & Yoshida, 2018)	8.22	21.7
Learning implicit generative models with the method of learned moments (Ravuri, Mohamed, Rosca, & Vinyals, 2018)	7.90	18.9

There are no universally agreed-upon performance metrics for unsupervised learning, and people have already pointed out many shortcomings of these Inception-based methods (Barratt & Sharma, 2018). Until something better comes along though, they're going to show up in every paper so it's worth knowing what they are.

Footnotes

1. Implementations are available for both TF and PyTorch. ↩
2. Other feature layers of the Inception-v3 network can also be used, with different dimensionalities. Note that at least d samples are needed to estimate the Gaussian statistics for d-dimensional features. ↩
3. It's hard to tell if different papers are using consistent methodology for computing IS and FID (e.g., some use 5k samples for FID while others use 50k), and some papers report different numbers for the same model. I tried to pick the best score reported by each paper for their own method where possible and yolo'd the rest. ↩
4. Reported by (Huang, Li, Poursaeed, Hopcroft, & Belongie, 2017). ↩
5. Reported by (Ostrovski, Dabney, & Munos, 2018). ↩
6. Reported by (Ostrovski, Dabney, & Munos, 2018). ↩ ↩[2]
7. Reported by (Warde-Farley & Bengio, 2016). ↩

References

1. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *Advances in Neural Information Processing Systems* (pp. 2234–2242).
2. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems* (pp. 6626–6637).
3. Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *ArXiv Preprint ArXiv:1511.06434*.
4. van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., & others.

(2016). Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems* (pp. 4790–4798).
5. Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., & Courville, A. (2016). Adversarially learned inference. *ArXiv Preprint ArXiv:1606.00704*.
6. Warde-Farley, D., & Bengio, Y. (2016). Improving generative adversarial networks with denoising feature matching.
7. Bordes, F., Honari, S., & Vincent, P. (2017). Learning to generate samples from noise through infusion training. *ArXiv Preprint ArXiv:1703.06975*.
8. Berthelot, D., Schumm, T., & Metz, L. (2017). BEGAN: boundary equilibrium generative adversarial networks. *ArXiv Preprint ArXiv:1703.10717*.
9. Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., & Póczos, B. (2017). MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems* (pp. 2203–2213).
10. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems* (pp. 5767–5777).
11. Unterthiner, T., Nessler, B., Klambauer, G., Heusel, M., Ramsauer, H., & Hochreiter, S. (2017). Coulomb gans: Provably optimal nash equilibria via potential fields. *ArXiv Preprint ArXiv:1708.08819*.
12. Ostrovski, G., Dabney, W., & Munos, R. (2018). Autoregressive Quantile Networks for Generative Modeling. *ArXiv Preprint ArXiv:1806.05575*.
13. Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *ArXiv Preprint ArXiv:1802.05957*.
14. Ravuri, S., Mohamed, S., Rosca, M., & Vinyals, O. (2018). Learning Implicit Generative Models with the Method of Learned Moments. *ArXiv Preprint ArXiv:1806.11006*.
15. Barratt, S., & Sharma, R. (2018). A Note on the Inception Score. *ArXiv Preprint ArXiv:1801.01973*.
16. Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., & Belongie, S. (2017). Stacked generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 2).

Recommend 1 Tweet Share Sort by Best

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Israr Bacha • a year ago

hi neal!
i want to evaluate image to image translation with GAN models like cycleGAN,BicycleGAN and pix2pix etc. i download pretrained models for them, now is it like to run the test on i.e horse2zebra dataset and put the generated zebra images and real zebra images in two folders and then calculate FID?. what are the necessary condition for these images? is these images will be from test set or train set?

Reply • Share ›

Nitin Bansal • 2 years ago

Hi Neal!

Thanks Indeed for great Details! The Blog was really easy to follow!
Neal, I was working on Similar areas, where I need to check the quality of the Image Generated by the GANs. I went through the GITHUB link of the code for FID and IS.

I had a small query though. When I load the Inception model, I find the pretrained model is based on an Imagenet Dataset. Now supposing that I am doing a task based on CIFAR10 Dataset on GAN. Does it make sense to use Inception model pretrained on Imagenet ? As it has 1000 activations in the last layers, and even the size of kernels used also would be larger, Since the size of Imagenet Image would be bigger than CIFAR10/100.

Any comments or sugggestion on this point.

Regards,
Nitin Bansal

Reply • Share ›

Neal Jean Mod → Nitin Bansal • 2 years ago

Neal Jean Mod ↗ Nitin Bansal • 2 years ago

Hi Nitin, I think it's standard to use the Inception model trained on ImageNet and also to resize your images to match - check out this TF implementation:

https://github.com/tensorfl...

Reply • Share ›

Nitin Bansal ↗ Neal Jean • 2 years ago

Thank you for the prompt reply Neal!

I will definitely go through this code. Thanks for the pointer.
Regards,
Nitin Bansal

Neal Jean

Neal Jean
nealjean@stanford.edu

nealjean

nealjean

nealjean1

PhD student, Stanford AI Lab