Deep Generative Models
Lecture 10

Roman Isachenko

Moscow Institute of Physics and Technology

2020

Evaluation of likelihood-free models

How to evaluate generative models?

Likelihood-based models

- Split data to train/val/test.
- Fit model on the train part.
- Tune hyperparameters on the validation part.
- Evaluate generalization by reporting likelihoods on the test set.

Not all models have tractable likelihoods

- VAE: compare ELBO values.
- GAN: ???

Evaluation of likelihood-free models

Let take some pretrained image classification model to get the conditional label distribution $p(y|\mathbf{x})$ (e.g. ImageNet classifier).

What do we want from samples?

- **Sharpness**

Low sharpness

High sharpness

The conditional distribution $p(y|\mathbf{x})$ should have low entropy (each image $\mathbf{x}$ should have distinctly recognizable object).

- **Diversity**

Low diversity

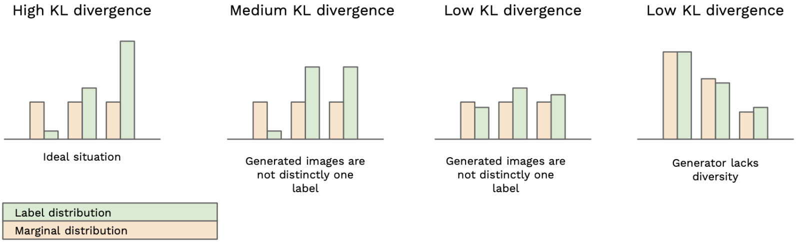
High diversity

The marginal distribution $p(y) = \int p(y|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ should have high entropy (there should be as many classes generated as possible).

Evaluation of likelihood-free models

What do we want from samples?

- **Sharpness.** The conditional distribution $p(y|\mathbf{x})$ should have low entropy (each image $\mathbf{x}$ should have distinctly recognizable object).
- **Diversity.** The marginal distribution $p(y) = \int p(y|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ should have high entropy (there should be as many classes generated as possible).



https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a

Evaluation of likelihood-free models

What do we want from samples?

- Sharpness $\Rightarrow$ low $H(y|\mathbf{x}) = -\sum_y \int_{\mathbf{x}} p(y, \mathbf{x}) \log p(y|\mathbf{x}) d\mathbf{x}$.
- Diversity $\Rightarrow$ high $H(y) = -\sum_y p(y) \log p(y)$.

Inception Score

$$
\begin{aligned}
IS &= \exp(H(y) - H(y|\mathbf{x})) \\
&= \exp\left(-\sum_y p(y) \log p(y) + \sum_y \int_{\mathbf{x}} p(y, \mathbf{x}) \log p(y|\mathbf{x}) d\mathbf{x} \right) \\
&= \exp\left(\sum_y \int_{\mathbf{x}} p(y, \mathbf{x}) \log \frac{p(y|\mathbf{x})}{p(y)} d\mathbf{x} \right) \\
&= \exp\left(\mathbb{E}_{\mathbf{x}} \sum_y p(y|\mathbf{x}) \log \frac{p(y|\mathbf{x})}{p(y)} \right) = \exp\left(\mathbb{E}_{\mathbf{x}} KL(p(y|\mathbf{x}) || p(y)) \right)
\end{aligned}
$$

---

https://arxiv.org/abs/1606.03498

Evaluation of likelihood-free models

Inception Score

$$IS = \exp\left(\mathbb{E}_{\mathbf{x}} KL(p(y|\mathbf{x}) || p(y))\right)$$

IS limitations

- Inception score depends on the quality of the pretrained classifier $p(y|\mathbf{x})$.
- If generator produces images with a different set of labels from the classifier training set, IS will be low.
- If the generator produces one image per class, the IS will be perfect (there is no measure of intra-class diversity).
- IS only require samples from the generator and do not take into account the desired data distribution $\pi(\mathbf{x})$ directly (only implicitly via a classifier).

---

https://arxiv.org/abs/1801.01973

Evaluation of likelihood-free models

Theorem

If $\pi(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\theta})$ has moment generation functions then

$$\pi(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta}) \Leftrightarrow \mathbb{E}_\pi \mathbf{x}^k = \mathbb{E}_p \mathbf{x}^k, \quad \forall k \geq 1.$$

This is intractable to calculate all moments.

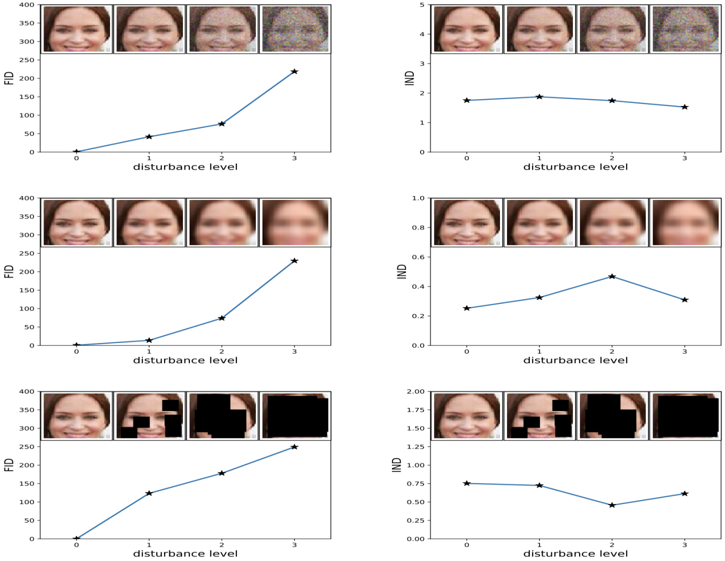
Frechet Inception Distance

$$D^2(\pi, p) = \|\mathbf{m}_\pi - \mathbf{m}_p\|_2^2 + \mathrm{Tr}\left(\mathbf{C}_\pi + \mathbf{C}_p - 2\sqrt{\mathbf{C}_\pi \mathbf{C}_p}\right)$$

- $\mathbf{m}_\pi$, $\mathbf{C}_\pi$ are mean vector and covariance matrix of feature representations for real samples from $\pi(\mathbf{x})$
- $\mathbf{m}_p$, $\mathbf{C}_p$ are mean vector and covariance matrix of feature representations for generated samples from $p(\mathbf{x}|\boldsymbol{\theta})$.
- Representations are output of intermediate layer from pretrained classification model.

https://arxiv.org/abs/1706.08500

Evaluation of likelihood-free models



https://arxiv.org/abs/1706.08500

Evaluation of likelihood-free models

Frechet Inception Distance

$$D^2(\pi, p) = \|\mathbf{m}_\pi - \mathbf{m}_p\|_2^2 + \mathrm{Tr}\left(\mathbf{C}_\pi + \mathbf{C}_p - 2\sqrt{\mathbf{C}_\pi \mathbf{C}_p}\right)$$

FID limitations

- FID depends on the pretrained classification model.
- FID needs a large samples size for evaluation.
- Calculation of FID is slow.
- FID extimates only two sample moments.

---

https://arxiv.org/abs/1706.08500

Summary

- Wasserstein GAN uses Kantorovich-Rubinstein duality to estimate Wasserstein distance.
- Gradient Penalty proposes the regularizer to enforce Lipschitzness.
- Spectral normalization is a weight normalization technique to enforce Lipshitzness.
- f-divergence family is a unified framework for divergence minimization.
- Inception Score and Frechet Inception Distance are the common metrics for GAN evaluation.

References

- A Note on the Inception Score
  https://arxiv.org/abs/1801.01973
  Summary: Inception Score is not an ideal metric.
- GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium
  https://arxiv.org/abs/1706.08500
  Summary: Frechet inception distance was proposed for GAN evaluation.