Metrics of Image Generation

References:

- Wikipedia - Fréchet Inception Distance ⭐
- Wikipedia - Fréchet Distance
- Wikipedia - Inception Score ⭐
- GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium
- IS、FID、PPL，GAN网络评估指标

FID

Definition

The **Fréchet inception distance (FID)** is a metric used to assess the quality of images created by a generative model, like a generative adversarial network (GAN) or a diffusion model.

FID is an extension of **Fréchet distance**: For two multivariate normal distributions $p = \mathcal{N}(\mu_p, \Sigma_p)$ and $q = \mathcal{N}(\mu_q, \Sigma_q)$, the Fréchet distance is given by:

$$\mathrm{FD}(p, q) = \|\mu_p - \mu_q\|_2^2 + \mathrm{tr}(\Sigma_p + \Sigma_q - 2(\Sigma_p \Sigma_q)^{1/2})$$

This is also known as **Wasserstein-2 distance**.

For two groups of samples $X \in \mathbb{R}^{N\times d}$ and $Y \in \mathbb{R}^{N\times d}$, we can first transform them into some **feature vectors**, denoted as $\tilde{X} \in \mathbb{R}^{N\times f}$ and $\tilde{Y} \in \mathbb{R}^{N\times f}$. Assume the feature vectors have a normal distribution $\tilde{X} \sim \mathcal{N}(\mu_{\tilde{X}}, \Sigma_{\tilde{X}})$ and $\tilde{Y} \sim \mathcal{N}(\mu_{\tilde{Y}}, \Sigma_{\tilde{Y}})$, we can use their Fréchet distance to measure the similarity between the two groups of samples.

FID uses pretrained **Inception V3 network** as the feature extractor. The Inception V3 network is a deep convolutional neural network trained on ImageNet dataset. It has a penultimate layer of dimension 2048, and the output of this layer is often used as the feature vector of generated images.

Calculation

1. **Data Preparation**: Prepare a large number of images (typically 10000), $X \in \mathbb{R}^{N\times d}$. Resize the images to shape $3 \times 299 \times 299$ and normalize them to range $[0, 1]$.
2. **Feature Extraction**: Input the images to the Inception V3 network, and take the penultimate layer output $\tilde{X} \in \mathbb{R}^{N\times 2048}$, $\tilde{Y} \in \mathbb{R}^{N\times 2048}$ as the extracted features.
3. **Fréchet Distance Calculation**: Calculate the mean and covariance of $\tilde{X}$ and $\tilde{Y}$ as $\mu_X, \Sigma_X, \mu_Y, \Sigma_Y$, and compute the FID as:

$$\mathrm{FID}(X, Y) = \|\mu_X - \mu_Y\|_2^2 + \mathrm{tr}\left(\Sigma_X + \Sigma_Y - 2(\Sigma_X \Sigma_Y)^{1/2}\right)$$

Some State-of-the-art Results

Model/Dataset	CIFAR10 Unconditional	CelebA 64×64
NCSN	25.32	25.30
NCSN++	10.87	10.23
DDPM	3.17	3.26
DDIM	4.04	3.51

Inception Score

Definition

The **Inception Score (IS)** is a metric used to evaluate the quality and diversity of images created by a genearative model, such as a Generative Adversarial Network (GAN). Similar to FID, IS leverages the class predictions from a pretrained Inception v3 model to assess two aspects:

1. **Quality**: Each generated image should belong to a recognizable class, resulting in a low-entropy distribution $p(y|x)$ for the image.
2. **Diversity**: The set of generated images should cover a wide variety of classes, leading to a high entropy marginal distribution $p(y)$ over all classes.

Calculation

1. **Data Preparation**: Prepare a large number of images (typically 10000 or 50000), $X \in \mathbb{R}^{N\times d}$. Resize the images to shape $3 \times 299 \times 299$ and normalize them to range $[0, 1]$.
2. **Feature Extraction**: Feed the images to the Inception v3 network, and take output $\tilde{X} \in \mathbb{R}^{N\times 1000}$ as the extracted features.
3. **Class Probabilities**: For each image x, obtain the class probability distribution $p(y|x)$ from the Inception v3's output (a softmax over 1000 ImageNet classes).
4. **Marginal Distribution**: Compute the marginal distribution $p(y)$ by averaging all $p(y|x)$ across the generated images.
5. **KL Divergence**: Calculate the KL divergence between $p(y|x)$ and $p(y)$ for each image. The IS is the exponential of the expected KL divergence:

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

Interpretation:

- A high IS indicates both high quality (each image is confidently classified) and diversity (images cover many classes).
- **Limitations**:
  - IS does not compare generated images to real data; it only evaluates the generated distribution internally.
  - It may favor models that generate "ImageNet-like" images, as the Inception v3 is pretrained on ImageNet.
  - Overfitting can lead to deceptively high IS if the model reproduces training set statistics without true diversity.