生成对抗网络评价指标

目前在很多GAN网络生成图片的论文中都会出现两个评价指标：Inception Score(IS)和Frechet Inception Distance(FID)

Inception Score

- 评价一个生成模型，我们主要考虑个体和整体两个方面：1）单一样本的独特性；2）多样本间的差异性。生成的图片不清晰，说明生成模型性能不好；生成的图片清晰了，但是多样性不足，只能生成固定的几种类别的图片，说明生成模型出现了模型坍塌（mode collapse）
- **单一样本的独特性**：把生成的图片x输入到Inception V3网络中，将输出1000维的向量y，向量的每个维度的值代表图片属于某一类的概率。我们希望它属于某一类的概率应该非常大，即生成的图片更像是某一类，而不是模棱两可，转化成数学描述就是$p(y|x)$的熵应该很小（熵越小代表确定性足够大）
- **多样本间的差异性**：如果一个生成模型能够生成足够多样的图片，那么它生成的图片在各个类别中的分布应该是平均的，假设生成了10000张图片，那么最理想的情况是，1000个类中每个类都生成了10张图片，转化成数学描述是，生成图片在所有类别概率的边缘分布$p(y)$熵很大（均匀分布）。具体计算时，可以先用生成器生成的N张图片，然后用公式（1）的经验分布来代替：

$$\hat{p}(y) = \frac{1}{N}\sum_{i=1}^{N} p(y|x^{(i)}) \tag{1}$$

- 总和上面两个方面，Inception Score的公式为：

$$IS(G) = exp(E_{x\sim p_g} D_{KL}(p(y|x)||p(y))) \tag{2}$$

  其中，
  - $x \sim p_g$表示从生成器中采样
  - $p(y|x)$：把生成的图片x输入到Inception V3中，得到的一个1000维的向量y，也就是该图片属于各个类别的概率分布。IS提出者假设，对于足够清晰的图片，这个向量的某个维度的值应该非常大，而其余的维度非常小（也就是概率密度图非常尖）
  - $p(y)$：N个生成的图片，每个生成的图片都输入到Inception V3网络中，各自得到一个自己的的概率分布向量，把这些向量求一个平均，得到生成器生成的图片全体在所有类别上的边缘分布，公式（1）
  - D_{KL}：对$p(y|x)$和$p(y)$求KL散度，KL散度离散形式的公式如下：

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P_{(i)}}{Q_{(i)}} \tag{3}$$

- KL散度用以衡量两个概率分布的距离，它是非负的，值越大说明两个概率分布越不像
- 我们的期望是$p(y|x^{(i)})$的某个维度的值很大，而$p(y)$总体均匀，因此要把$p(y|x^{(i)})$放在竖线的左边
- 只要$p(y|x^{(i)})$和$p(y)$足够大，就能证明生成模型足够好。因为前者是一个很尖锐的分布，后者是一个均匀分布，这两个分布的距离本来就应该很大
- 为了美观，对公式（2）进行改写

$$IS(G) = exp(\frac{1}{N}\sum_{i=1}^{N} D_{KL}(p(y|x^{(i)})||\hat{p}(y))) \tag{4}$$

Inception Score的局限性

1、Inception Score对神经网络的内部权重非常敏感

作者利用TensorFlow、Pytorch和Keras等不同框架下预训练的Inception V3，计算同一个数据库CIFAR-10的Inception Score，发现尽管不同框架下预训练的网络达到同样的分类精度，但是由于其内部权重的微小不同，导致Inception Score有很大的变化

2、计算Inception Score的方式不对

去掉为了美观而加上的exp，直接解释为互信息，改进后的Inception Score公式为：

$$IS(G) = \frac{1}{N}\sum_{i=1}^{N} D_{KL}(p(y|x^{(i)})||\hat{p}(y)) \tag{5}$$

Inception Score存在的问题

Inception Score是基于两个假设：

1、越真实的图片，输入预训练的Inception V3的飞累的结果越明确，输出的概率分布图越尖锐

2、生成的图片多样性越强，那么类别的边缘分布就越平均，边缘分布的概率函数图像越平整

Inception Score计算这两个概率分布的散度来衡量模型的表现

这两个假设存在问题：

1、对于第一个假设，若某一物体所属的类别在分类网络并不存在，那么它的分布函数是否依然尖锐

2、对于第二个假设，假设模型在每个类上都生成了50张图片，那么生成图片的类别的边缘分布是严格 均匀的，但是如果这50张图片是一模一样的，依然是mode collapse，Inception Score无法检测这种 情况

出现这些问题的原因是计算Inception Score时只考虑了生成样本，没有考虑真实数据，即Inception Score无法反映真实数据和生成样本之间的距离。

Frechet Inception Distance

预训练好的神经网络的顶层可以提取图片的高级信息，常常被用来衡来差异，如Perceptual loss

Frechet Inception Score计算的事真实图片和生成图片在feature层面的距离，公式如下：

$$FID = \|\mu_r - \mu_g\|^2 + Tr(\sum_r + \sum_g - 2(\sum_r \sum_g)^{1/2}) \tag{6}$$

μ_r：真实图片的特征均值

μ_r：生成图片的特征均值

Tr：对角线元素之和

$\sum_r$：真实图片的特征的协方差矩阵

$\sum_g$：生成图片的特征的协方差矩阵

FID值越低代表生成图像质量和多样性越好

参考

- 全面解析Inception Score原理及其局限性
- Frechet Inception Distance