生成模型读书笔记九

2021年1月11日 21:55

GAN的evaluation，各种GAN模型，以及训练tricks

a. GAN的evaluation

GAN不像之前的方法，如VAE，对于每个测试输入我们都有明确的目标输出，因此它的评估很困难，也是研究方向之一。对于一个GAN模型，我们希望能够评估：

(1)生成样本的质量，够不够真实；

(2)生成样本的多样性，会不会有mode collapse；

(3)对于z，生成器能够学习出有明确意义的从z到x的映射，通过在z空间插值等方法，能够得到某种意义下连续的x样本。

常用的方法有结合标签分类从分类结果评估或者计算生成和真实分布的一些统计量，这里列出几个常用的评估指标。

i. Inception Score(IS)

在有标签的情况下，把生成的样本拿去用一个inception network分类，得到样本标签后验概率p(y|x)。IS想要从两个方面来衡量GAN模型：生成样本的质量和生成样本多样性。首先如果一个样本是有意义的，高质量的，那么它应该会被比较确定地分为某一类，也即后验分布p(y|x)的熵值比较低。另一方面，一个好的GAN应该生成多样性的样本，也就是能够生成多种种类的样本，即边缘分布p(y)=$\int \mathrm{p}(y|\mathrm{x}=G(z))dz$的熵值较高。将这两个要求结合起来，就有了下式

$$\text{Inception Score} = \exp(E_x KL(\mathrm{p}(\mathrm{y}|\mathrm{x}) \parallel \mathrm{p}(\mathrm{y})))$$

IS的值越大意味着模型可以生成高质量的样本并且这些样本具有多样性（把上式的KL距离展开，第一项就是负的p(y|x))的熵，如上所述，p(y|x)的熵越低越好，所以负的熵越大越好，第二项展开并计算Ex期望之后可以得到p(y)的熵，也是越大越好，所以IS越大就意味着这两个指标都比较大。

IS的问题：类内发生的mode collapse无法探测；IS的高低会受到图片像素影响；

ii. IS的一些改进版本

1) Mode Score： $\exp(E_x KL(\mathrm{p}(\mathrm{y}|\mathrm{x}) \parallel \mathrm{p}^*(\mathrm{y})) - KL(p(y) \parallel p^*(y)))$，其中$\mathrm{p}^*(y)$是真实分布的类别概率，与IS相比多考虑了真实数据集的标签信息
2) Modified Inception Score: $\exp(E_{x_i}[E_{x_j}[\mathrm{KL}(\mathrm{p}(\mathrm{y}|x_i) \parallel p(x_j))]])$，其中$\mathrm{x}_i, x_j$是同一类别的样本，相比IS考虑了类内多样性
3) AM Score：$KL(p^*(y) \parallel p(y)) - E_x[\mathrm{H}(\mathrm{y}|\mathrm{x})]$，其中$\mathrm{p}^*(y)$是真实分布的类别概率，考虑了真实数据集的类别分布(在IS中被假设均匀分布)

iii. Frechet Inception Distance(FID)

同样地，也是利用一个inception network来提取特征，把从生成样本和真实样本得到的特征$\phi(\mathrm{x})$看成两个高斯分布，并且计算这两个分布的均值μ_{data}, μ_g和方差C_{data}, C_g，得到下式

$$\mathrm{FID}(\mathrm{p}_{data}, p_g) = \|\mu_{\mathrm{data}} - \mu_g\| \mathrm{tr}\left(C_{data} + C_g - 2(C_{data}C_g)^{\frac{1}{2}}\right)$$

FID计算了生成分布和真实分布之间的距离，FID越小说明模型越好。FID同样也是对样

本质量和多样性的衡量。假设模型生成了不真实的样本，那么它的均值就会和真实分布的均值相差很远，FID值就会大；假设模型缺乏多样性，比如极端情况下只能生成一个样本，那么它的方差就是0，FID也会比较大。

iv. 1-NN classifier

将n个真实样本和n个生成样本组合成一个数据集，每次从中选择一个样本作为测试集，剩余的2n-1个作为训练集训练一个1-nn二分类器，二分类器用来分类是真实样本还是生成样本，重复2n次，计算2n次的平均分类正确率。如果生成分布和真实分布一样，则分类器相当于随机猜测，正确率会在50%左右；如果生成分布和真实分布相差很大，那么二分类器可以轻易地将它们分开，正确率会很高；如果生成分布只是在简单地记忆训练样本，那么分类器会分错，正确率很低，接近0。这种方法计算实现都很简单，不包含其他预训练的网络或者需要调节的超参数。

v. GANtrain and GANtest

用真实样本组成一个数据集S_{t}，一个验证集S_{v}，用生成样本组成一个数据集S_g。(1)在S_{t}上训练分类器，S_{v}上计算准确率，记为GANbase；(2)在S_g上训练分类器，S_{v}上计算准确率，记为GANtrain；(3)在S_{t}上训练分类器，S_{g}上计算准确率，记为GANtest。

比较GANtrain和GANbase，如果生成分布丢失了一些mode，或者样本不够真实，分类器学习不到有效的特征，就会导致分类器的效果不好，GANtrain<GANbase；比较GANbase和GANtest，如果GANtest非常高，那么说明生成器仅仅在记忆训练数据，如果GANtest非常低，说明生成的样本质量不高，无法很好的分类。

vi. Quality measures

在这类方法中，会通过计算一下样本的统计量，直接比较样本的质量，而不像之前的方法一样借助其他模型来比较。比如在图像生成的任务中，会设计统计量来比较样本的亮度，对比度，结构，峰值信噪比，锐度等，可以仿照这个方法根据自己的任务设计一些样本质量的评估指标。

评估目前还是一个难题，因为人类对生成质量的感受很难用这些评估指标来表达，很多时候评估指标好的样本在人类看来不一定好，指标低的也不一定不好。

b. 各种各样的GAN

GAN的对抗思想很令人惊艳，但是它的训练困难，原始GAN生成不可控等问题，导致了在实际应用中作用并不大。于是人们近年提出了各种各样的GAN变体，这一节介绍两个方面的变体以及一些具有代表性的模型。

i. 基于提升GAN训练的变体

1) 提出新的Loss function

a) WGAN, WGAN-GP: 上文已经着重介绍过，也是最受欢迎的GAN训练方法之一

b) LSGAN：least square GAN。LSGAN旨在解决原始GAN第一种loss形式中的梯度消失问题，它指出原始GAN的判别器对于远离决策面的样本惩罚很小，所以导致了梯度消失。所以LSGAN提出用least square loss来代替原始GAN中的交叉熵loss。

类似的还有替换成hinge loss，softmax cross entropy等损失函数的，把分类函数中能用的损失函数都试了一遍。

c) EBGAN：energy-based GAN，使用energy function做D的损失函数，生成的假样本赋高energy值，真样本赋低energy值

d) SN-GAN：spetral normalized GAN. 使用了一种叫spetral normalization的方法来对weights做normalization，使得它满足Lipschitz连续(和WGAN-GP相比呢？)，使得判别器D更加稳定，如BigGAN里也用到了
e) f-gan：探索了用不同divergence来衡量生成分布和真实分布距离的效果

2) G和D网络的多种实现

在原始GAN中，G和D都是MLP，其实根据任务的不同需求，他们还可以是其他网络结构

a) Laplacian GAN：使用Laplacian金字塔的结构去逐级生成高分辨率高质量的图片(styleGAN貌似也是这样的)。类似的还有SinGAN/InGAN，都是从一张图片学习生成，学习不同尺度的patch distribution(可以去了解下)
b) DCGAN：全CNN网络结构，没有pooling层，除了G的最后一层和D的第一层以外每一层都使用batch normalization
c) styleGAN：后面再详细讲

3) 多个G，多个D

a) MGAN：多个generator的结合，类似混合概率模型，以避免mode collapse问题
b) D2GAN，MoCoGAN，AGED：多个discriminator的结合，每个D都用于做不同的判别任务，如MoCoGAN中一个D负责判别单独的frame，一个D负责判别整个video

4) 解决mode collapse问题的变体

ii. 基于可控生成，隐变量学习的变体

单从随机噪声生成样本的GAN在实际应用中是没什么意义的，所以在GAN框架提出后，人们又想尽办法将这个框架变成一些有意义的模型。研究集中在几个方面：希望可以控制生成；希望可以知道噪声的对于生成样本的意义，可以找出z和x之间的明确关系，这样GAN也可以像VAE一样用于representation learning了；G和D能不能随意组合，loss能不能随意加，这样GAN就可以进行一些多样的任务了。

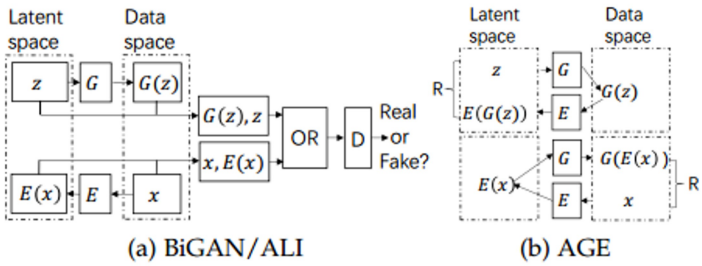
1) 输入条件实现可控的生成

a) Conditional GAN：给生成器和判别器一个条件，如样本的类别标签，噪声z与样本的label拼接在一起作为G和D的输入，希望生成指定类别的样本，这是最简单的CGAN；也可以用于学习某种源域到目标域的映射，不需要噪声z，G的输入是源域，输出是目标域，源域和目标域的真假样本组合一起作为D输入，以判别样本是否真实以及是否与源域对应，代表性的模型有pix2pix；更进一步，判别器除了判断真假以及标签样本匹配以外，再添加一个多分类器输出样本类别，高质量的样本才能被识别出类别，能够进一步通过标签提升生成质量。

2) 与autoencoder的结合

因为想要学习隐变量，因此考虑是否可以利用autoencoder的encoder学习从生成样本到隐变量z的映射

a) GAN与encoder结合：加一个encoder，学习从x/G(z)到z的映射，隐变量与样本拼接一起作为D的输入，代表性的模型有：BiGAN，ALI，如下方左图所示



也有人仿照VAE的思路，用高斯或者混合高斯模型来拟合隐变量

b) GAN与encoder循环结合：单单添加一个encoder，如上图，虽然可以模拟到隐空间与样本空间之间映射，但是Encoder只作用于真实样本x，Generator只作用于随机噪声z，无法保证他们是互相对应的逆映射。因此有人提出了ALI，重构循环，如上方右图所示，计算隐空间和样本空间循环变换之后的重构误差。类似的模型：CycleGAN(cycleGAN可以证明是变分推断的一个特殊例子，有时间了解一下)； StarGAN

c) GAN与autoencoder的结合： DRGAN(有时间了解下)

3) 对隐变量进行限制

a) InfoGAN：将输入噪声分为两部分，一部分是不可压缩的噪声，另一部分是指定某种意义的隐变量，计算指定隐变量和生成样本的互信息，通过最大化两者的互信息，希望指定的隐变量包含指定的特征。

c. GAN训练的一些技巧(持续补充)

i. 有label的数据就用上label

ii. Virtual batch normalization（之后的章节会讲一讲各种normalization的作用）
BN方法有个副作用：当minibatch比较小的时候，会产生震荡。
一个优化的方法：reference batch normalization。事先运行网络，采样一些样本作为不变的reference样本，每次做normalization的时候就用这些reference样本来计算均值和方差。为了避免对renference样本的过拟合，又有人提出了当前样本和reference的组合来计算均值方差，就称为virtual batch normalization。

iii. Minibatch features
用于解决mode collapse。判别器将一个样本分别与一minibatch的生成样本和一minibatch的真实样本做比较，通过对比样本与两者的距离，可以知道这个样本和其他生成样本之间是否过于相似。
另一个解决mode collapse的方法是unrolled GAN。

iv. 输入要归一化，噪声输入最好从N(0,1)中采样

v. 如果模型基于CNN，decode时尽量用upsample+conv2d代替transposeconv

vi. 关于normalization，bn在分类任务上不错，但是生成任务推荐使用其他normalization方法，如参数化的instance, layer, 非参的pixel，还有大杂烩switchable

vii. 推荐使用多个discriminator，多个scale，结构相同参数不同

viii. Spectral normalization有时候比gradient penalty有效

ix. 可以给G和D设置不同的学习率，让D更快收敛

x. 尽量不要用会导致梯度稀疏的计算，如ReLU, MaxPool

xi. 使用soft的label，如将label从0，1改成0.2，0.8等等

xii. 使用混合的模型，如KL+GAN，VAE+GAN

xiii. 可以对D和G使用不同的优化器，如G用Adam，D用SGD
xiv. 更早地发现训练失败：D loss很快收敛到0；检查梯度的norm，太大会崩；正常来说，D的loss不会变化太大，且单调减小
xv. 往D的输入添加noise，并且随训练逐步减小；G的每一层可以添加一点高斯noise(或者使用dropout)
xvi. 时不时输出生成样本看一看