生成式对抗网络 (GAN) 的原理和各种变种类型的介绍

C. Lu

2018 年 3 月 5 日

1 生成式对抗网络

1.1 引言

生成式对抗网络 (generative adversarial network, GAN)[2] 是基于可微生成器网络的一种生成式数据建模方法。

生成式对抗网络基于博弈论场景，其中生成器网络必须与对手竞争。生成器网络直接产生样本 $\boldsymbol{x} = g(\boldsymbol{z}; \boldsymbol{\theta}^{(g)})$。其对手，**判别器网络** (discriminator network) 试图区分从训练数据抽取的样本和生成器中抽取的样本。判别器发出由 $d(\boldsymbol{x}; \boldsymbol{\theta}^{(d)})$ 给出的概率值，指示 $\boldsymbol{x}$ 是真实训练样本而不是从模型中抽取的伪造样本的概率, GAN 的结构如图 1.1。

形式化表示生成式对抗网络中学习的最简单的方式是零和游戏，其中 $v(\boldsymbol{\theta}^{(g)}, \boldsymbol{\theta}^{(d)})$ 确定判别器的收益。生成器接收 $-v(\boldsymbol{\theta}^{(g)}, \boldsymbol{\theta}^{(d)})$ 作为它自己的收益。在学习期间，每个玩家尝试最大化自己的收益，因此收敛在

$$g^* = \arg\min_g \max_d \ v(g, d) \tag{1}$$

v 的默认选择是

$$v(\boldsymbol{\theta}^{(g)}, \boldsymbol{\theta}^{(d)}) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \log d(\boldsymbol{x}) + \mathbb{E}_{\boldsymbol{x} \sim p_{\text{model}}} \log(1 - d(\boldsymbol{x})) \tag{2}$$

这驱使判别器试图学习将样品正确地分类为真的或者伪造的。同时，生成器试图欺骗分类器以让其相信样本是真实的。在收敛时，生成器的样本与实际数据不可区分，并且判别器处处都输出 $\frac{1}{2}$。然后就可以丢弃判别器。

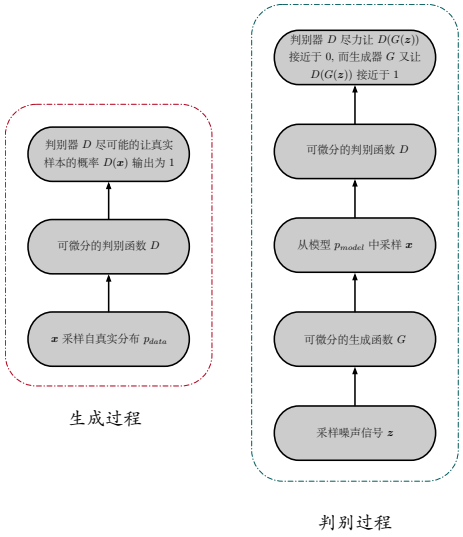

图 1: GAN 的结构示意图

设计 GAN 的主要动机是学习过程既不需要近似推断，也不需要配分函数梯度的近似。当 $\max_d v(g,d)$ 在 $\boldsymbol{\theta}^{(g)}$ 中时凸的（例如，在概率密度函数的空间中直接执行优化的情况）时，该过程保证收敛并且是渐近一致的。

1.2 生成式对抗网络的训练算法

生成式对抗网络的训练式生成器与判别器互相博弈的过程。生成器与判别器交替使用最优化算法 (如：梯度下降算法) 来最大化各自的价值函数。在训练初期，判别器的能力较弱，无法正确区分出真实样本和伪造样本；此时，可以认为设立一个超参数 k, 训练 k 轮判别器后，再进行生成器的训练。具体算法流程如算法 1所示。

1.3 生成式对抗网络的理论依据

[2] 中证明了式 (1) 的最优解为

$$p_g = p_{data}$$

即生成模型能够很好的代表了真实数据的分布。

Algorithm 1 生成式对抗网络的随机梯度下降训练算法。判别器训练的循环次数 k 是人为指定的超参数。($k \geq 1$)

for 训练迭代次数 **do**
  for k **do**
    从噪声的先验分布 $p_g(\boldsymbol{z})$ 中采样 m 个噪声样本 $\{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(m)}\}$
    从数据的真实分布 $p_{\text{data}}(\boldsymbol{x})$ 中采样 m 个真实数据样本 $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}\}$
    通过梯度上升法来更新判别器的参数，梯度由以下公式给出：

$$\nabla_{\boldsymbol{\theta}_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right]$$

  end for
  从噪声的先验分布 $p_g(\boldsymbol{z})$ 中采样 m 个噪声样本 $\{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(m)}\}$
  通过梯度下降法来更新判别器的参数，梯度由以下公式给出

$$\nabla_{\boldsymbol{\theta}_g} \frac{1}{m} \sum_{i=1}^{m} \left[\log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right]$$

end for

2 深度卷积生成式对抗网络

在实践中，由神经网络表示的 g 和 d 以及 $\max_d v(g,d)$ 不凸时，GAN 中的学习可能是困难的。[3] 认为不收敛可能会引起 GAN 的欠拟合问题。稳定的 GAN 学习仍然是一个开放问题。幸运的是，当仔细选择模型架构和超参数时，GAN 的学习效果很好。[9] 设计了一个深度卷积生成式对抗网络 (DCGAN), 在图像合成的任务上表现非常好，并表明其潜在的表示空间能捕获到变化的重要因素。图 2 展示了生成器生成的图像实例。

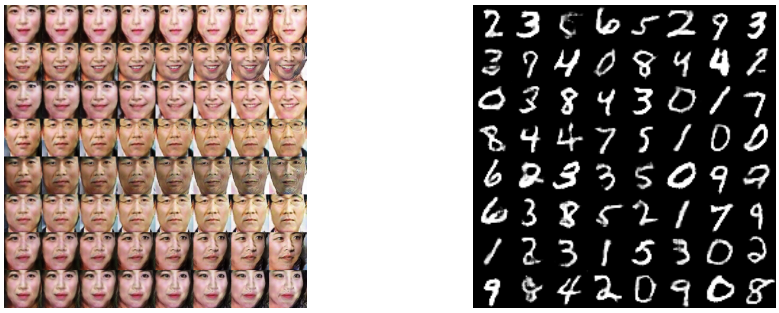

图 2: 在亚洲人脸数据集 (左) 和手写数字数据集 (右) 上训练后，由 DCGAN 生成的图像

2.1 深度卷积生成式对抗网络架构

深度卷积生成式对抗网络 (DCGAN) 采用卷积神经网络 (CNN) 作为生成器和判别器。与原始的 CNN 架构不同，DCGAN 主要做了以下几点重要的改动：

移除池化函数 将所有的池化函数 (如：MaxPooling) 全部替换为固定步长的卷积函数。这样做的目标是为了让神经网络自己去学习如何采样。

移除全连接层 移除 CNN 中的全连接层。对于判别器来说，最后一层卷积输出被展平成一个向量；对于生成器，噪声信号 $\boldsymbol{z}$ 采样自一个均匀分布，乘以一个变换矩阵 $\boldsymbol{W}$ 之后变成一个高维向量，然后将其重新调整为一个 4 维的张量。具体结构如图 2.1所示。

批标准化 在生成器与判别器的网络中加入批标准化层 [5]。批标准化是一个自适应的重参数化的方法，可以用来训练非常深的模型。如果直接将批标准化应用到模型的每一层中，会导致生成样本和模型的不稳定性，所以，在 DCGAN 中，生成器的输出层和判别器的输入层没有加入批标准化。

激活函数的改动 生成器中使用 ReLU[8] 激活函数，输出层是用 Tanh 激活函数；与原始 GAN 使用的 Maxout[4] 激活函数不同，判别器使用 LeakyReLU[10] 会有更好的效果，而且在更深的模型效果更显著。

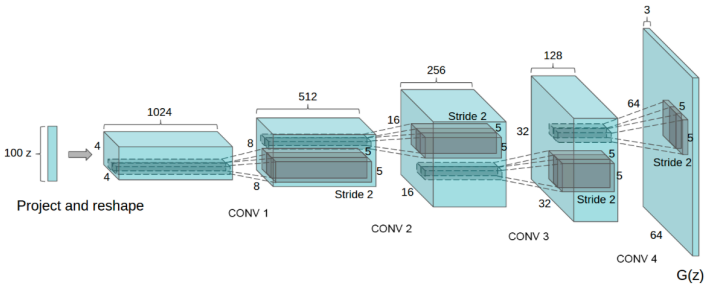

图 3: 用于 LSUN 数据集的 DCGAN 生成器的结构示意图

2.2 DCGAN 中特征表示的向量运算

生成器从采样噪声 $\boldsymbol{z}$ 中经过卷积运算生成特定的图像样本，可以把噪声信号 $\boldsymbol{z}$ 认为是生成图片的低维表示，$\boldsymbol{z}$ 的每个分量代表图像的某种特征。例如，对于人脸图像，$\boldsymbol{z}$ 的某个分量代表了性别，其他某个分量代表了是否带眼镜。与词向量类似，可以进行算数上的加减操作，如:

$$\boldsymbol{v}(\text{国王}) - \boldsymbol{v}(\text{男人}) + \boldsymbol{v}(\text{女人}) = \boldsymbol{v}(\text{王后})$$

[9] 中给出了类似的实验结果，如图 4 所示。

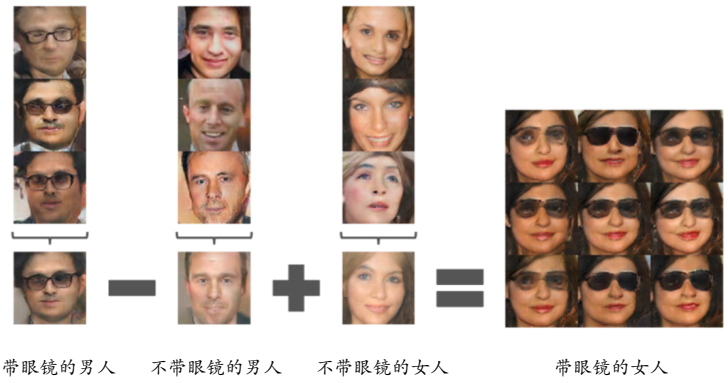

图 4: CGAN 中特征表示的向量运算

3 条件生成式对抗网络

条件生成式对抗网络 (Conditional GAN, 简称 CGAN) 是在原始生成式对抗网络的生成器和判别器中都加入条件信息 $\boldsymbol{y}$，（例如，在手写数字识别的问题中，$\boldsymbol{y}$ 可能是特定的数字标签；在图像生成的问题中，$\boldsymbol{y}$ 可以是对要生成图像的特定描述)，控制生成模型生成满足特定条件的样本。

在原始 GAN 的生成器中，输入的噪声信号 $\boldsymbol{z}$ 带代表了要生成图像的低维表示；同样的，也可以将要加入的附加信息进行编码，得到附加信息的向量表示 $\boldsymbol{y}$, 之后，将 $\boldsymbol{z}$ 与 $\boldsymbol{y}$ 一同输入到生成器中，以产生特定条件限制下的样本。对于判别器也是类似的操作，同时将样本 $\boldsymbol{x}$（来自真实分布 p_{data} 或者模型的分布 p_{model}）与条件信息 $\boldsymbol{y}$ 输入进行判别。

条件生成式对抗网络的目标函数如公式 3所示：

$$V(\boldsymbol{\theta}^{(G)}, \boldsymbol{\theta}^{(D)}) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}} \log D\left(\boldsymbol{x}|\boldsymbol{y}\right) + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})} \left[\log\left(1 - D\left(G\left(\boldsymbol{z}|\boldsymbol{y}\right)\right)\right)\right] \quad (3)$$

图 3简要的展示的使用神经网络的条件生成对抗网络。

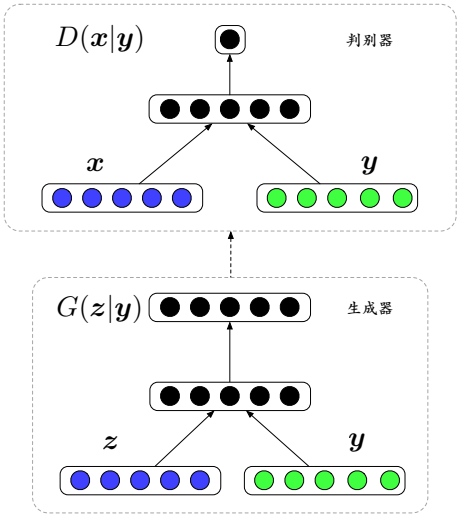

图 5: 条件生成对抗网络

3.1 条件生成对抗网络的实例

3.1.1 手写数字生成

3.1.2 基于多模态的图像自动标注

3.2 条件对抗网络之图像到图像的转换

4 基于能量的生成式对抗网络

参考文献

[1] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Ian J Goodfellow. On distinguishability criteria for estimating generative models. *arXiv preprint arXiv:1412.6515*, 2014.

[4] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456, 2015.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[7] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[8] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[9] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[10] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[11] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.