

生成对抗网络研究综述

王正龙¹, 张保稳^{1,2}

(1. 上海交通大学网络安全技术研究院, 上海 200240;
2. 上海市信息安全综合管理技术研究重点实验室, 上海 200240)

摘要: 首先介绍了生成对抗网络基本理论、应用场景和研究现状, 并列举了其亟待改进的问题。围绕针对提升模型训练效率、提升生成样本质量和降低模式崩溃现象发生可能性3类问题的解决, 从模型结构和训练过程两大改进方向和7个细分维度, 对近年来生成对抗网络的主要研究工作、改进机理和特点进行了归纳和总结, 并结合3方面对其未来的研究方向进行了探讨。

关键词: 生成对抗网络; 生成模型; 深度学习; 模式崩溃; 分布距离度量; 神经网络鲁棒性

中图分类号: TP183

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2021080

Survey of generative adversarial network

WANG Zhenglong¹, ZHANG Baowen^{1,2}

1. Institute of Cyber Science and Technology, Shanghai Jiao Tong University, Shanghai 200240, China
2. Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai 200240, China

Abstract: Firstly, the basic theory, application scenarios and current state of research of GAN (generative adversarial network) were introduced, and the problems need to be improved were listed. Then, recent research, improvement mechanism and model features in 2 categories and 7 subcategories revolved around 3 points (improving model training efficiency, improving the quality of generated samples, and reducing the possibility of model collapse) were generalized and summarized. Finally, 3 future research directions were discussed.

Keywords: generative adversarial network, generative model, deep learning, mode collapse, distribution similarity measurement, robustness of artificial neural network

1 引言

近年来, 随着人工智能的蓬勃发展, 用于解决人工智能问题的神经网络生成算法成为研究热

点。生成模型学习样本概率分布, 并从分布中生成新的数据, 以此衡量研究者对于高维概率分布的操控能力。此类模型可用于数据增强、图像翻译等领域。传统生成模型通常基于马尔可夫链、

收稿日期: 2020-12-31; 修回日期: 2021-06-10

通信作者: 张保稳, zhangbw@sjtu.edu.cn

基金项目: 国家重点研发计划 (2020YFB1807504, 2020YFB1807500)

Foundation Item: The National Key R&D Program of China (2020YFB1807504, 2020YFB1807500)

论文引用格式: 王正龙, 张保稳. 生成对抗网络研究综述[J]. 网络与信息安全学报, 2021, 7(4): 68-85.

WANG Z L, ZHANG B W. Survey of generative adversarial network[J]. Chinese Journal of Network and Information Security, 2021, 7(4): 68-85.

最大似然及近似推理,其代表模型有RBM^[1](restricted Boltzmann machines)及其衍生模型如DBN^[2](deep belief network)、DBM^[3](deep Boltzmann machines)、VAE^[4](variational auto-encoder)等,此类方法计算复杂且生成效果有限。

2014年,Goodfellow等提出了基于博弈论的生成模型——生成对抗网络^[5](GAN, generative adversarial network),它使用两个神经网络进行对抗训练,通过反向传播更新网络权值,易于计算且效果显著。该模型一经提出就引起神经网络领域研究人员的关注,对其进行了大量的研究和改进。

作为一种生成模型,GAN及其衍生模型常用于为深度学习中的数据增强、数据预处理方法生成样本,并在图像处理、生物医学、网络与信息安全等领域有着广泛的应用场景。在图像处理领域中,Koo等^[6]使用深度卷积生成对抗网络^[7](DCGAN, deep convolutional GAN)为黑白照片着色;CycleGAN^[8]通过学习两类样本空间的双向映射实现图像风格迁移;CVAE-GAN^[9]能够在不改变样本分类结果的前提下生成细粒度修改的新样本。在生物医学领域,Schlegl等^[10]使用WGAN^[11](wasserstein GAN)检测病变的视网膜光学相干断层扫描成像数据;Wolterink等^[12]使用WGAN合成血管的几何形状,用于对冠状动脉心脏CT血管造影图进行补充。在网络与信息安全领域,刘西蒙等^[13]总结了3种基于GAN的对抗样本问题防御策略,这3种策略都通过GAN重建应用系统的输入样本,以排除输入样本中可能存在的微小扰动带来的影响;张煜等^[14]提出了一种基于GAN的文本序列数据集脱敏模型,该模型无须针对隐私属性精心设计处理规则即可实现数据集脱敏;Beaulieu-Jones等^[15]提出一种在训练GAN模型时添加随机噪声的方法以避免模型从病人的就医数据中学习到个人隐私的方法,郭鹏等^[16]在文献[15]的基础上,使用自适应梯度裁剪阈值进行了优化,提高了生成图像的清晰度和多样性;王旭东等^[17]提出了一种基于SN-GAN^[18](spectral normalization GAN)的人脸图像补全方法,用于在人脸被遮挡时尽量保留人脸的身份特

征并对身份校验。

随着GAN的应用场景日渐复杂,研究者需要进一步深入研究GAN理论,以指导其设计过程,并解决GAN模型常见的模式崩溃、训练困难等问题。本文对既有GAN的研究工作及其评价方法进行归纳与总结,并在此基础上探讨了其未来发展的方向。

2 GAN的基本模型及工作机理

GAN的基本模型的主要思想是令两个神经网络不断进行二元极大极小博弈,在此过程中模型逐渐习得真实样本分布。一般而言,当两个网络的对抗达到纳什均衡后,认为训练完成。

图1给出了GAN的基本模型,生成器网络(记为 G)的输入为来自隐空间(记为 p_z)的随机变量(记为 z),输出生成样本,其训练目标是提高生成样本与真实样本的相似度,使其无法被判别器(记为 D)网络区分,即令生成样本(记为 p_g)分布与真实样本(记为 p_{data})分布尽量相同^[5]。 D 的输入为真实样本(记为 x)或生成样本(记为 x'),输出判别结果,其训练目标是分辨真实样本与生成样本。判别结果用于计算目标函数,通过反向传播更新网络权值。在对抗训练的过程中, D 判别真假样本的能力逐渐提高,而 G 为了欺骗 D ,生成样本逐渐趋近于真实样本,最终使整个模型生成质量较好的新数据。

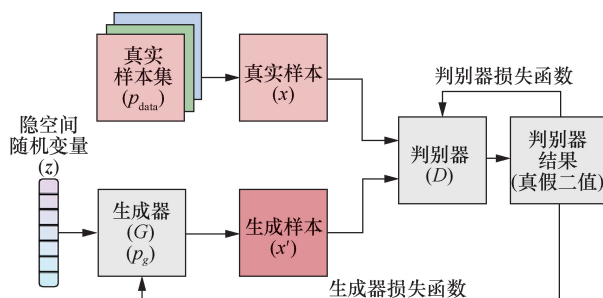


图1 GAN的基本模型
Figure 1 The basic model of GAN

如前所述,训练目标可以概括为:最小化 p_g 与 p_{data} 之间的距离,最大化 D 辨别样本的准确率,即 $D(x)$ 值趋于1且 $D(x')$ 值趋于0。由此可得到目标函数表达式

$$\min_G \max_D E_{x \sim p_{\text{data}}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

其本质为最小化 p_g 和 p_{data} 的 KLD (Kullback-Leibler divergence)。KLD 用于衡量两种不同分布之间的差异, 定义为

$$\text{KL}(p_{\text{data}} \| p_g) = E_{x \sim p_{\text{data}}(x)} \left[\log \frac{p_{\text{data}}(x)}{p_g(x)} \right] = \int p_{\text{data}}(x) \log \frac{p_{\text{data}}(x)}{p_g(x)} dx \quad (2)$$

由式(2)易知 KLD 不对称, 可通过数学技巧将其变为对称, 即 JSD (Jensen Shannon divergence), 定义为

$$\text{JS}(p_{\text{data}} \| p_g) = \frac{1}{2} \text{KL} \left(p_{\text{data}} \| \frac{p_{\text{data}} + p_g}{2} \right) + \frac{1}{2} \text{KL} \left(p_g \| \frac{p_{\text{data}} + p_g}{2} \right) \quad (3)$$

对抗训练过程由两个神经网络交替进行, 每一步均先训练 D , 具体步骤为选取大小为 m 的一批真实样本和一批隐变量, 隐变量通过 G 得到生成样本, 保持 G 权值不变, 利用随机梯度上升法, 通过式(4)计算并更新 D 网络的权值; 随后训练 G , 具体步骤为选取一批隐变量通过 G 得到生成样本, 保持 D 权值不变, 利用随机梯度下降法通过式(5)计算并更新 G 网络的权值。对抗训练的纳什均衡为 $p_g = p_{\text{data}}$, 此时达到全局最优解, D 至多以 50% 的概率随机猜测样本属于真实样本或生成样本。

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m [\log D(x^i) + \log(1 - D(G(z^i)))] \quad (4)$$

$$-\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log(1 - D(G(z^i))) \quad (5)$$

GAN 的基本模型具有以下优点: 计算过程使用反向传播而不需要马尔可夫链, 回避了近似计算困难的概率难题, 此外, 训练时无须推断随机变量。它与 DBN 相比能更快生成样本, 与 VAE 相比有更好的生成效果。

同时, GAN 的基本模型有许多亟待改进之

处。在模型训练过程中, 存在模式崩溃问题^[19]: 生成器从具有多种模式的训练集中仅学习到单个或有限的模式, 而错失对其他模式的学习, 从而无法生成具有多样性的样本。若想解决模式崩溃问题, 就需要对 GAN 模型加以改进, 指导生成器学习真实样本集中包含的多种模式, 或避免生成器仅学习有限种模式。此外, 使用随机向量作为生成器的输入缺乏语义和可操作性, 无法控制模型生成具有指定特征的样本; 模型使用到的神经网络本身可以被进一步改造和优化; 训练过程中易出现梯度消失、训练不稳定、神经网络难以收敛等问题。上述问题导致模型的训练难度较大、训练时间较长。

当前对 GAN 理论的主流研究工作大部分围绕上述问题的优化处理和对模型进行研究改进。本文分别按照模型结构的改进、模型训练过程的改进两个方向, 从输入输出、生成器、判别器、多模块组合、模型交叉、分布距离度量、梯度计算过程 7 个维度 (如表 1 所示) 对近年来 GAN 主要研究工作、改进机理和特点进行归纳和总结。

3 基于模型结构改进的 GAN 衍生模型

图 1 给出的 GAN 的基本模型由输入输出、生成器、判别器三大模块按一定原理或思想组合而成。因此, 对模型结构进行改进的切入点可以划分为输入输出、生成器、判别器、模型的模块结构和模块结构的组合思想 5 部分, 本节按照这 5 个部分对现有 GAN 研究进行归纳分类与介绍。

3.1 基于输入输出改进的 GAN 模型

基于输入输出的改进主要是指从 G 的输入端和 D 的输出端着手进行改进。在 GAN 的基本模型中, G 的输入为隐空间上的随机变量, 因此对其改进主要从隐空间与隐变量这两点展开, 改进隐变量的目的是使其更好地控制生成样本的细节, 而改进隐空间则是为了更好地区分不同的生成模式。 D 输出的判别结果是真假二分类, 可以配合目标函数将其调整为多分类或去除神经网络的 Softmax 层直接输出特征向量, 进而优化训练过程、实现半监督学习等效果。

表 1 本文讨论的现有主要的 GAN 衍生模型
Table 1 An overview of GAN variants discussed in this paper

改进维度	改进方法	衍生模型
输入	基于隐变量的改进	cGAN ^[20] 、BiCoGAN ^[21] 、IcGAN ^[22] 、InfoGAN ^[23]
	基于隐空间的改进	DeLiGAN ^[24] 、NEMGAN ^[25]
	其他输入改进方法	FCGAN ^[26]
输出	多分类输出	SGAN ^[27] 、AC-GAN ^[28]
	特征输出	InfoGAN ^[23]
生成器	集成方法	AdaGAN ^[29]
	样本模式区分	MADGAN ^[30] 、MGAN ^[31]
	其他改进方法	MPMGAN ^[32]
判别器	集成方法	PacGAN ^[33] 、GMAN ^[34] 、DropoutGAN ^[35]
	样本模式区分	D2GAN ^[36] 、StabilizingGAN ^[37]
	其他改进方法	EBGAN ^[38] 、BEGAN ^[39]
多模块组合	多阶段模型	GRAN ^[40] 、StackGAN ^[41] 、ProgressGAN ^[42]
	辅助模块模型	TripleGAN ^[43] 、ControlGAN ^[44]
	其他改进方法	SGAN ^[45] 、MemoryGAN ^[46]
模型交叉	神经网络结构替换	DCGAN ^[7] 、CapsuleGAN ^[47]
	复合生成模型	VAEGAN ^[48] 、DEGAN ^[49] 、AAE ^[50] 、BiGAN ^[51] 、ALi ^[52] 、MatAN ^[53]
	其他领域思想	SAGAN ^[54] 、KDGAN ^[55] 、IRGAN ^[56] 、LapGAN ^[57] 、QuGAN ^[58] 、BayesianGAN ^[59]
分布距离度量	基于 f-散度的方法	f-GAN ^[60] 、LSGAN ^[61] 、EBGAN ^[38]
	基于 IPM 的方法	WGAN ^[11] 、WGAN-GP ^[62] 、WGAN-LP ^[63] 、BWGAN ^[64] 、CT-GAN ^[65] 、RWGAN ^[66] 、CramerGAN ^[67] 、SN-GAN ^[18] 、WGAN-div ^[68] 、McGAN ^[69] 、GMMN ^[70] 、MMDGAN ^[71] 、FisherGAN ^[72]
	其他方法	IGAN ^[73] 、MIX+GAN ^[74] 、OT-GAN ^[75] 、MMGAN ^[76]
梯度计算过程		MAGAN ^[77] 、LSGAN ^[61] 、SoftmaxGAN ^[78] 、RGAN ^[79]

cGAN^[20] (conditional GAN) 在 G 与 D 的输入中额外加入类别标签或其他有助于样本生成的附加信息 (记为 c)，从而使 GAN 可以生成符合特定条件的数据。cGAN 模型如图 2 所示。

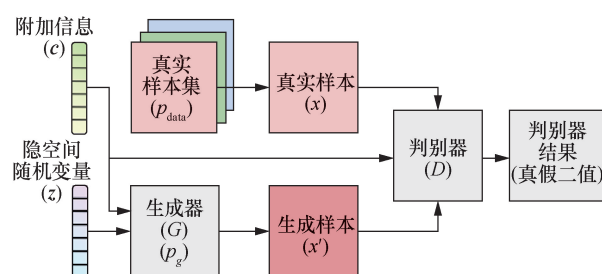


图 2 cGAN 模型
Figure 2 Model of cGAN

BiCoGAN^[21]模型的提出者认为文献[20]的输入 z 与 c 相互纠缠，因此增加了一个编码器 (记为 E) 用于学习从判别器输出到生成器两个输入的逆映射，从而更精确地编码 c ，以提升模型表现。如图 3 所示，将 z 与 c 的拼接 (记为 $\hat{z}$) 输入生成器得到输出 $G(\hat{z})$ ，将真实样本 x 输入编码器得到输出 $E(x)$ ，判别器接收 $G(\hat{z})$ 或 $[x, E(x)]$ 作为输入，判定该输入来自生成器或为真实数据的某一类。由于真实样本 x 具有的标签可视为 c ，而 $E(x)$ 又可以被拆分为 z' 与 c' ，因此使 c 与 c' 尽可能接近，也成为模型训练的目标，从而使编码器学习逆映射。文中提出使用 EFL (extrinsic factor loss) 衡量两个分布 p_c 与 $p_{c'}$ 的距离，并提出如式(6)所示的目标函数。

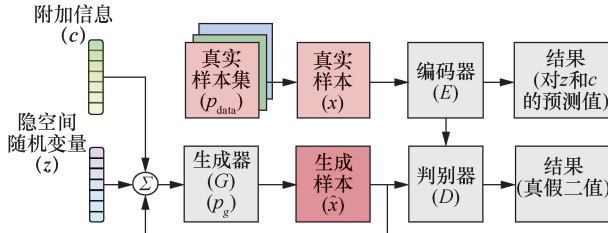


图 3 BiCoGAN 模型
Figure 3 Model of BiCoGAN

$$\min_{G,E} \max_D E_{x \sim p_{data(x)}} [\log D(E(x), x)] +$$

$$E_{z \sim p_{z(z)}} [\log(1 - D(\hat{z}, G\hat{z}))] +$$

$$\gamma E_{(x,c) \sim p_{data(x,c)}} [EFL(c, E_c(x))]$$
(6)

IcGAN^[22] (invertible conditional GAN) 以文献[20]为基础, 增加了两个预训练的编码器 E_z 和 E_y , E_z 用于生成隐空间中的随机变量 z , E_y 用于生成原始条件 y , 通过将 y 修改成 y' 作为 cGAN 的输入条件, 从而控制合成图像的细节 (如图 4 所示)。本文提出了 3 种从分布中进行采样获得 y' 的方法: 当 y 为二进制向量时, 可通过 KDE (kernel density estimation) 拟合分布并进行采样; 当 y 为实向量时, 可选取训练集的标签向量进行直接插值; 当某个条件并不是在所有训练集中表现唯一时, 可直接对 p_{data} 进行采样。

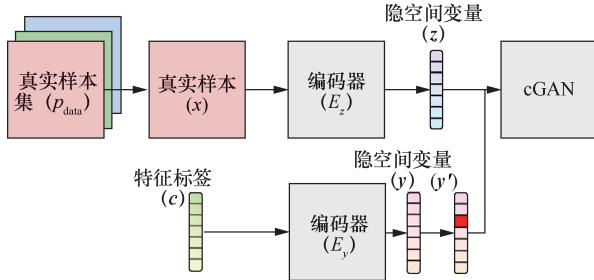


图 4 IcGAN 模型
Figure 4 Model of IcGAN

InfoGAN^[23] (模型如图 5 所示) 将 G 的输入拆分成随机变量 z 与隐编码 c , D 的输出包括真假二分类和去掉 Softmax 层的特征向量输出两项, 以 c 与特征向量间的互信息作为约束, 从而使 c 能够表示生成样本的不同特征维度, 为网络增加了可解释性。

DeLiGAN^[24] 适用于训练数据规模小、种类多的场景, DeLiGAN 模型如图 6 所示。Gurumurthy 等^[24] 提出使用 GMM (Gaussian mixture model)

对隐空间进行参数化, 再随机选择一个高斯分量进行重参数化, 从指定的高斯分布中获取样本, 但模型使用 GMM 是一种简化假设, 限制了其逼近更复杂分布的能力。

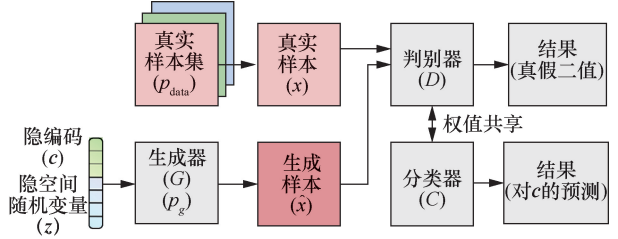


图 5 InfoGAN 模型
Figure 5 Model of InfoGAN

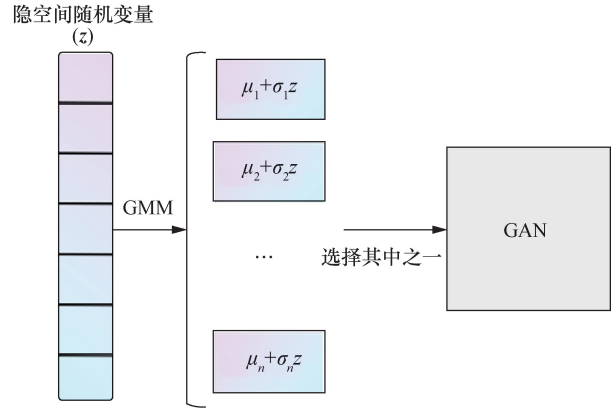


图 6 DeLiGAN 模型
Figure 6 Model of DeLiGAN

NEMGAN (noise engineered mode matching GAN)^[25] 的提出者提出一种能够在训练集存在数据不均衡情况下表现较好的模式匹配策略, 根据生成样本训练出其在隐空间中的对应表示, 得到潜在模式的先验分布, 从而将生成样本的多种模式进行分离, 并且与真实样本的模式进行匹配, 保证了生成样本中包含多个真实样本的模式, 以缓解模式崩溃问题。

FCGAN (fully conditional GAN)^[26] 以文献[20]为基础, 将额外信息 c 连接到神经网络的每一层, 一定程度上提升了有条件生成样本时的生成样本质量, 但该模型在 c 较为复杂或大向量的场景中运算效率低。

SGAN (semi-supervised learning GAN)^[27] 是一种能够为数据集重建标签信息的半监督模型, 其模型如图 7 所示。它将 D 改进为分类器与判别器的结合体, D 的输出包含 N 类真实样本和

一类生成样本, 共有 $N+1$ 类。向模型输入无标签的样本且判别器将其分类为真实样本时, 可以将判别器的输出作为该样本的标签。

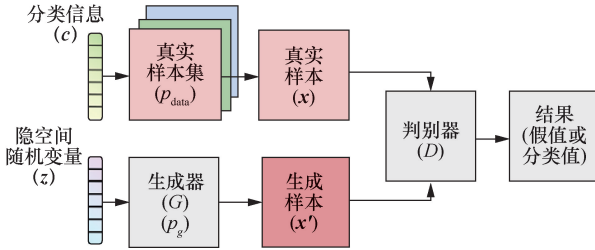


图7 SGAN 模型
Figure7 Model of SGAN

AC-GAN (auxiliary classifier GAN)^[28]同时具备文献[20]和文献[27]的特点, G 输入随机变量与分类信息 c , D 输出样本为假和分类概率, 该方法能够在有条件生成样本时输出生成样本所属的类别。

3.2 基于生成器改进的 GAN 模型

基于生成器进行改进的工作, 旨在提高生成样本质量与避免模式崩溃问题, 使模型能够生成多种类的样本, 且同一种类内的样本具有多样性。改进的思路包括: 使用集成学习 (ensemble learning) 的思想综合多个弱生成器所学习到的模式、设计每个生成器专注于学习特定模式多生成器架构, 从而使模型整体包含多个模式, 使用多智能体系统的思想使多个生成器之间产生竞争与合作的关系等。

AdaGAN 模型^[29]的提出者提出了一种融入集成学习思想的迭代训练算法。在单步迭代过程中, 根据训练样本与混合权值得到一个弱生成器, 该弱生成器与上一轮迭代得到的弱生成器加权混合, 得到本次迭代结果。若干轮迭代以后, 生成器综合了多个弱生成器各自学习到的模式, 缓解了模式缺失导致的模式崩溃问题, 并能够生成出质量较好的样本。但是, 混合多个生成器网络导致输入的隐空间不连续, 不能像基本 GAN 模型那样通过插值法得到新的隐变量。

MADGAN (multi-agent diverse GAN)^[30]由多个生成器和一个判别器组成, 其模型如图 8 所示。其中, 判别器负责判断输入样本是真实样本还是生成样本, 若为生成样本则判断它是由哪一个生成器所生成的。每个生成器专注于学习特定模式, 模型使多个生成器各自学习到的 p_{g_n} 混合

来近似 p_{data} , 模型最终得到的生成样本来自多个学习到了不同模式的生成器, 显式地保证了生成样本的多样性, 缓解了模式崩溃问题。

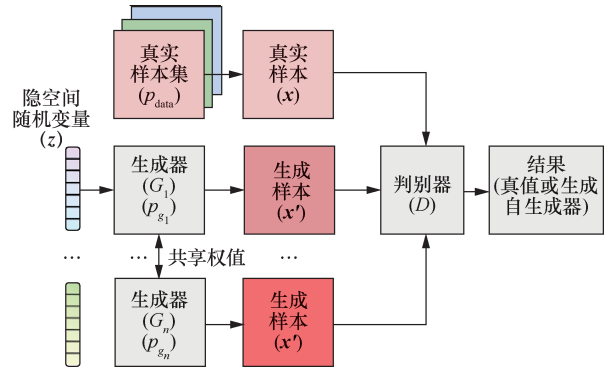


图8 MADGAN 模型
Figure 8 Model of MADGAN

MGAN^[31]缓解模式崩溃问题的思路与文献[30]类似, 其模型如图 9 所示。该模型设计了一种与判别器权值共享但去除掉 Softmax 层的分类器, 用于承担判断生成样本所属生成器的功能, 判别器仅负责判别样本为真实样本还是生成样本。

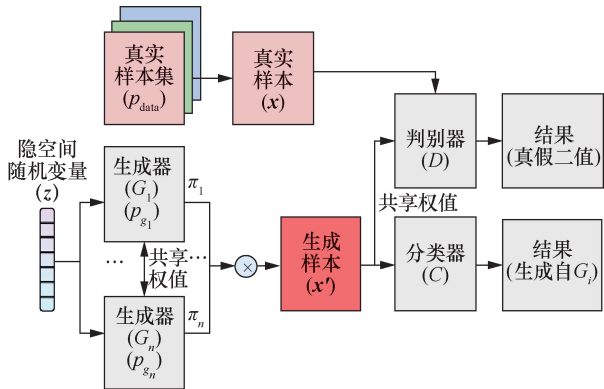


图9 MGAN 模型
Figure 9 Model of MGAN

文献[32]提出一种引入消息传递机制的多生成器 MPMGAN (message passing multi-agent GAN) 模型, 如图 10 所示。生成器输出作为传递给其他生成器的消息。在消息共享机制的作用下, 所有生成器都有合作目标、竞争目标两种目标。合作目标鼓励其他生成器的生成样本优于自身的生成样本; 竞争目标促使自身的生成样本优于其他生成器的生成样本。两种目标共同作用使生成样本质量得以优化。

3.3 基于判别器改进的 GAN 模型

GAN 模型训练过程中, 最初的生成样本质量

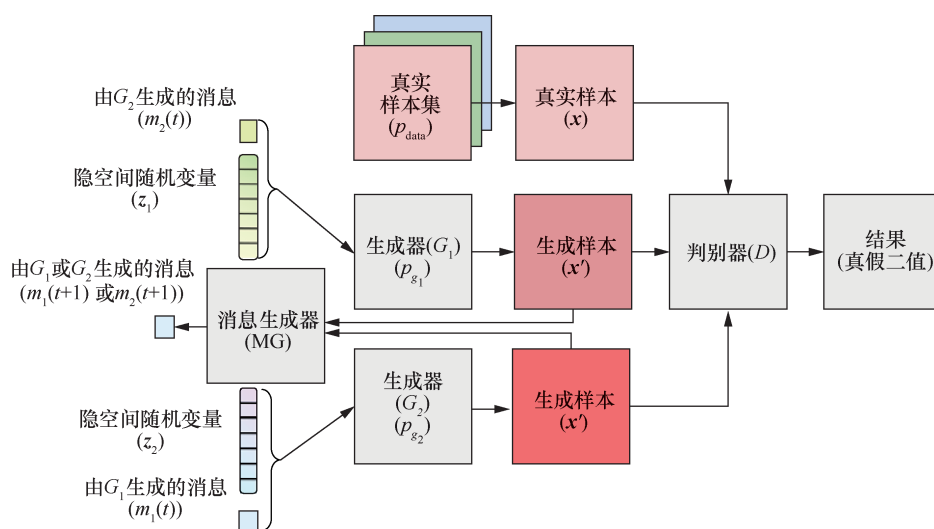


图 10 MPMGAN 模型
Figure 10 Model of MPMGAN

较差, 判别器可以简单地区分样本, 这导致生成器初始训练速度慢。改进判别器, 使其符合生成器当前能力有助于加快训练, 使其识别多种模式可以缓解模式崩溃问题。改进思路包括使单一判别器能识别出更多模式, 以及使多个判别器中的每个判别器专注于识别特定模式等。PacGAN 模型如图 11 所示。

PacGAN^[33]将同一类的多个样本“打包”后一起输入判别器, 以此保证每次判别器输入的样本都具有多样性。由于判别器每次接受输入时都能感知到样本的多样性, 生成器试图欺骗判别器时, 需要保证生成样本的多样性, 这有助于缓解模式崩溃问题。

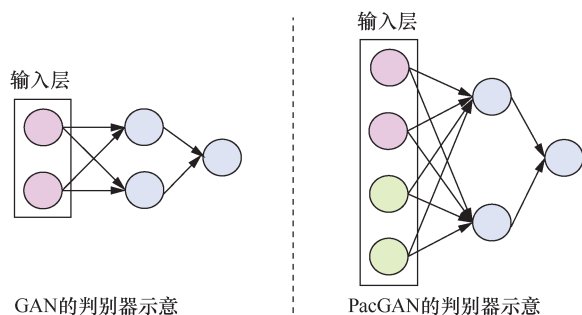


图 11 PacGAN 模型
Figure 11 Model of PacGAN

GMAN (generative multi-adversarial networks)模型^[34]的提出者认为过度改进判别器会使目标函数过于苛刻, 反而抑制生成器学习, 因此

提出一种结合集成学习的方法, 通过设置多个判别器, 生成器从多判别器聚合结果中学习, 从而使网络加速收敛。GMAN 模型如图 12 所示。

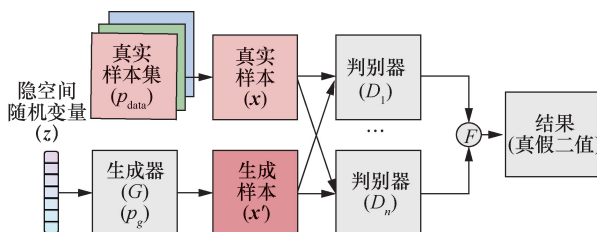


图 12 GMAN 模型
Figure 12 Model of GMAN

DropoutGAN^[35]设置了一组判别器, 在每批样本训练结束时, 以一定概率删除该结果, 将剩余结果聚合后反馈到生成器, 以此使生成器不局限于欺骗特定判别器。DropoutGAN 模型的提出者认为模式崩溃问题是生成器对特定判别器或静态集成判别器的过度拟合, 即生成器学习到了使判别器输出真值的特殊条件而非学习到了样本模式, 而该模型的结构中, 判别器集合是动态变化的, 生成器无法学习到欺骗判别器的特殊条件, 从而使生成器学习多种样本模式, 有助于缓解模式崩溃问题。DropoutGAN 模型如图 13 所示。

D2GAN (dual discriminator GAN)^[36]设置了两个判别器 D_1 、 D_2 , 分别使用正向 KL 散度及逆向 KL 散度, 以充分利用二者互补的统计特性。其中 D_1 通过正确判定样本来自真实样本分布获

得奖励, D_2 则通过正确判定样本来自生成样本分布获得奖励。生成器同时欺骗两个判别器, 以此来提升生成样本的质量。D2GAN 模型如图 14 所示。

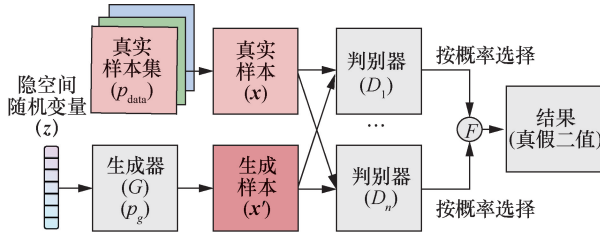


图 13 DropoutGAN 模型
Figure 13 Model of Dropout GAN

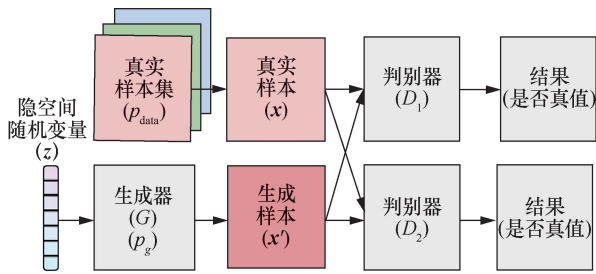


图 14 D2GAN 模型
Figure 14 Model of D2GAN

StabilizingGAN 模型^[37]的提出者认为真实样本在空间中集中分布, 而生成样本初始时在空间中分散分布, 导致训练初期判别器能够准确判断出几乎所有生成样本, 产生无效梯度, 使生成器训练缓慢。因此, 他们提出同时训练一组视角受限的判别器, 每个判别器都专注于空间中的一部分投影, 生成器逐渐满足所有判别器的限制, 以此稳定训练, 提升生成样本质量。

在 EBGAN (energy-based GAN) 模型^[38] (如图 15 所示) 中引入了能量函数的方法, 事物间差异越大能量越高, 故而真实分布附近样本具有较低能量。其研究者设计了一个由编码器和解码器构成的判别器, 使用 MSE (mean square error) 衡量生成样本与真实样本的差异并作为能量函数, 生成器目标为生成最小化能量的生成样本。BEGAN^[39] (boundary equilibrium GAN) 使用自编码器替代文献[38]中的判别器。

3.4 基于多模块组合改进的 GAN 模型

除了更好地拟合真实样本分布之外, 提升网络收敛的速度、提高生成图片的清晰度、将其应用在半监督学习上等同样是 GAN 模型改进的方

向。这类研究工作通过调整模块结构, 对不同的影响因素加以优化处理, 使模型达到特定目的。

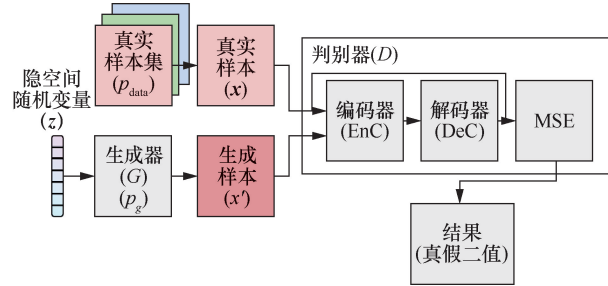


图 15 EBGAN 模型
Figure 15 Model of EBGAN

GRAN^[40] (generative recurrent adversarial networks) 是一种递归生成模型, 它反复生成以上一状态为条件的输出, 最终得到更符合人类视觉的生成样本。

StackGAN^[41]以文献[20]为基础构建了一种两阶段模型 (如图 16 所示)。它将文本描述作为额外信息, 阶段一生成较低分辨率的图像并输出至阶段二, 阶段二输出较高分辨率的图像, 从而提高生成图像的分辨率。

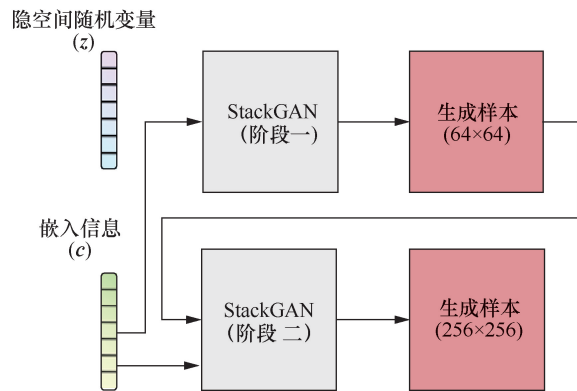


图 16 StackGAN 模型
Figure 16 Model of StackGAN

ProgressGAN 模型^[42]的提出者认为小尺度图像能够保证多样性且细节不丢失, 他们使用多个且逐渐增大的 WGAN-GP^[62]网络, 逐步训练最终生成高清图像。

TripleGAN^[43]通过增加一个分类器网络, 为真实样本生成标签, 生成器为真实标签生成样本, 判别器判别接收的样本标签对是否为有真实标签的真实样本, 从而同时训练出效果较好的分类器和生成器, 将 GAN 的能力扩展到

可以为无标签样本打标签。TripleGAN 模型如图 17 所示。

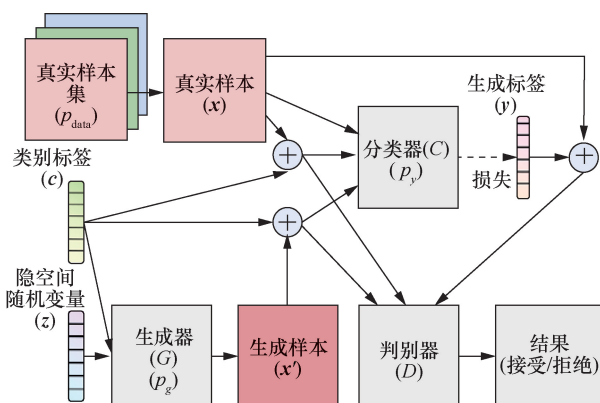


图 17 TripleGAN 模型
Figure 17 Model of TripleGAN

ControlGAN 模型^[44]的提出者认为文献[20]中的判别器同时承担了真实样本分类与判别真假样本两个任务, 因此将其拆分为独立的分类器和判别器, 从而在有条件生成样本时更细粒度地控制生成样本的特征。ControlGAN 模型如图 18 所示。

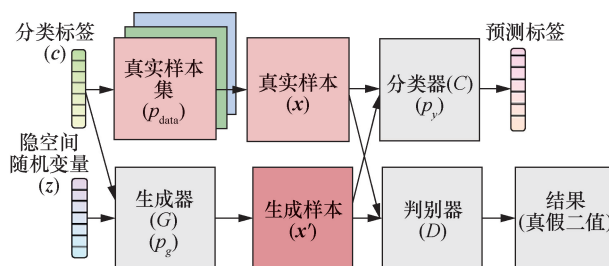


图 18 ControlGAN 模型
Figure 18 Model of ControlGAN

SGAN^[45] (several local pairs GAN) 使用若干组局部网络对和一组全局网络对, 每组网络对有一个生成器与一个判别器。局部网络对使用固定的配对网络进行训练, 不同局部网络对之间没有信息交互, 全局网络利用局部网络进行训练。由于每一个局部网络对都可以学到一种模式, 在使用局部网络对更新全局网络对后, 能够保证全局网络对综合了多种模式, 从而缓解模式崩溃问题。SGAN 模型如图 19 所示。

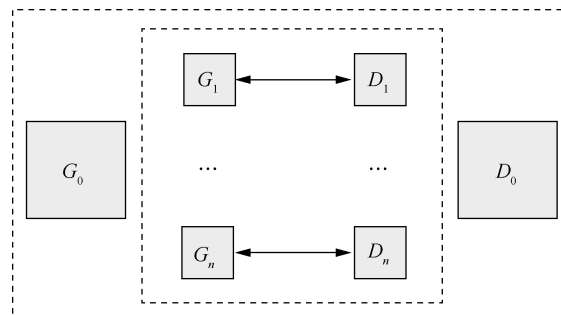


图 19 SGAN 模型
Figure 19 Model of SGAN

3.5 基于模型交叉思想改进的 GAN 模型

结合其他生成模型思想及其他领域思想对 GAN 模型进行改进,同样可以起到优化模型表现或拓展模型应用场景的效果。

DCGAN^[7] 使用去除池化层的 CNN (convolutional neural network) 替代基本 GAN 模型中的多层感知机 (如图 20 所示), 并使用全局池化层替代全连接层以减少计算量, 以提高生成样本的质量, 优化训练不稳定的问题。

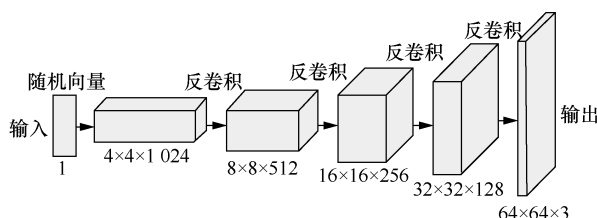


图 20 DCGAN 模型中的 CNN
Figure 20 CNN of DCGAN model

CapsuleGAN^[47]使用胶囊网络作为判别器的框架（如图 21 所示）。胶囊网络可以用于替代神经元，将节点输出由一个值转变为一个向量，神经元用于检测某个特定模式，而胶囊网络可以检测某个种类的模式，以此提高判别器的泛化能力，从而提高生成样本质量。

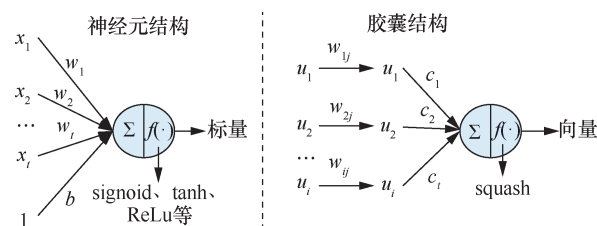


图 21 CapsuleGAN 的基本原理
Figure 21 Basic theory of CapsuleGAN

VAEGAN^[48]利用 GAN 来提高 VAE 生成样本的质量。其观点是: 在 VAE 中, 编码器将真实分布编码到隐空间, 而解码器将隐空间恢复为真实分布。单独解码器即可用作生成模型, 但生成样本质量较差, 因此再将其输入判别器中。

DEGAN (decoder-encoder GAN) 模型^[49]的提出者认为输入的随机变量服从高斯分布, 因此生成器需将整个高斯分布映射到图像, 无法反映真实样本分布。因此借鉴 VAE 的思想, 在 GAN 中加入预训练的编码器与解码器, 将随机变量映射为含有真实样本分布信息的变量, 再传递给 GAN, 从而加速收敛并提高生成质量。

AAE (adversarial auto-encoder)^[50]通过在 AE (auto-encoder) 的隐藏层中增加对抗的思想来结合 AE 与 GAN。判别器通过判断数据是来自隐藏层还是真实样本, 使编码器的分布向真实样本分布靠近。

BiGAN^[51]使用编码器来提取真实样本特征, 使用解码器来模仿生成器, 并使用判别器来辨别特征样本对是来自编码器还是解码器, 最终使编码方式和解码方式趋近于互逆, 从而使随机变量与真实数据形成映射。ALi^[52]和 BiGAN 本质相同, 二者仅有细微区别。BiGAN 模型如图 22 所示。

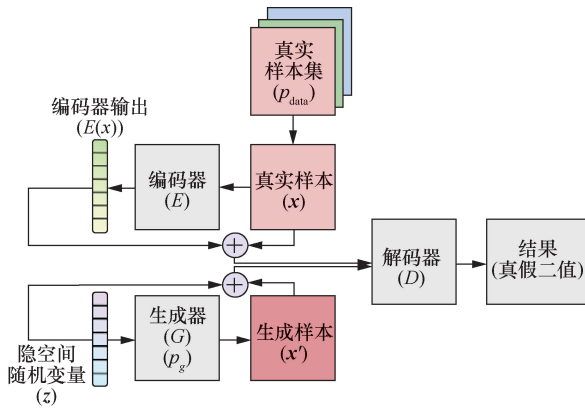


图 22 BiGAN 模型
Figure 22 Model of BiGAN

MatAN (matching adversarial network)^[53]使用孪生网络替换判别器, 以将正确标签考虑在生成器目标函数中。孪生网络用于衡量真实数据与生成数据的相似度。该方法对加快生成器训练有效。

SAGAN (self-attention GAN) 模型^[54]的提出

者认为 GAN 在合成结构约束少的种类上表现较好, 但难以捕捉复杂的模式, 通过在网络中引入自注意力机制以解决该问题。

KDGAN^[55]运用 KD (knowledge distillation) 的思想, 模型包含作为学生网络的轻量分类器、大型复杂教师网络及判别器, 其中, 分类器和教师网络都生成标签, 二者通过互相蒸馏输出学习彼此的知识, 最终可训练得到表现较好的轻量级分类器。

IRGAN^[56]利用 GAN 将 IR (information retrieval) 领域中的生成式检索模型与判别式检索模型相结合, 对于生成器采用基于策略梯度的强化学习来训练, 从而在典型的信息检索任务中取得较好的表现。IRGAN 模型如图 23 所示。

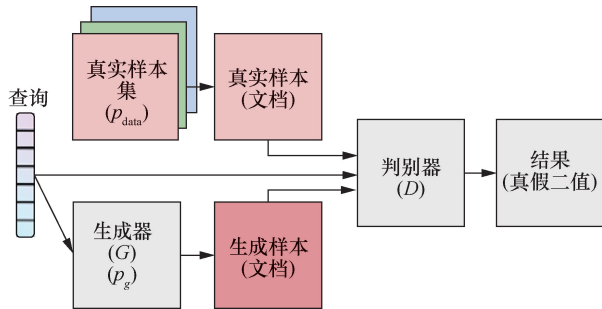


图 23 IRGAN 模型
Figure 23 Model of IRGAN

LapGAN^[57]使用了图像处理领域的思想, 同时使用三组 cGAN^[20], 按照高斯金字塔的模式对图像逐级下采样训练网络, 按照拉普拉斯金字塔的模式对图像逐级上采样, 从而达到从模糊图像中重构高像素图像的目的。

QuGAN^[58]将 GAN 的思想与量子计算的思想相结合, 将生成器类比生成线路, 判别器类比判别线路, 生成线路尽可能模仿真实线路的波函数, 判别线路尽可能仅通过对辅助比特的测量来确定输入的波函数来自生成线路还是真实线路。

BayesianGAN 模型^[59]的提出者认为 GAN 隐式学习分布的方法难以显式建模, 因此提出使用随机梯度哈密顿蒙特卡洛方法来边际化两个神经网络的权值, 从而使数据表示具有可解释性。

4 基于训练过程改进的 GAN 衍生模型

GAN 模型的训练目标是使生成器学习到一

个与真实样本分布尽可能相似的生成样本分布, 需要有一个合理的样本分布距离度量方法用于衡量两个分布的相似程度。此外, 神经网络的训练过程是使用梯度下降等方法使神经网络损失函数值越来越小的过程, 因此优化损失函数的计算也是改进 GAN 模型训练过程的一个方向。在这两个方面加以研究, 有助于改善 GAN 模型训练过程遇到的不稳定、难收敛等问题, 提升模型表现。

4.1 基于分布距离度量改进的 GAN 模型

目前应用于 GAN 中的分布距离度量有两大主流方法: 基于 f-散度, 选择合理的 $f(x)$ 制造不同的 GAN; 基于 IPM (integral probability metric) 框架, 选择合理的函数空间制造不同的 GAN。

f-GAN 模型^[60]的提出者认为 KLD 是分布距离度量的一种特殊情况, 并提出了其推广方式。给定任意两个分布 P 、 Q , $p(x)$ 、 $q(x)$ 分别为其绝对连续的密度函数, f 是定义域为 χ 的下半连续凸函数且满足 $f(1)=0$, 定义 f-散度为

$$D_f(P \| Q) = \int_{\chi} q(x) f\left(\frac{p(x)}{q(x)}\right) dx \quad (7)$$

借助凸函数的 Fenchel 共轭函数可以将 f-散度定义 GAN 的目标函数表示为

$$\min_G \max_D E_{x \sim p_{\text{data}}(x)} [D(x)] + E_{z \sim p_{\text{G}}(z)} [\log(1 - D(G(z)))] \quad (8)$$

由此, 可以选择满足不同条件的 f 函数来训练 GAN 模型, 且 GAN 的基本模型可以被认为是 f-GAN 的一个特例, 其他满足 f-GAN 要求的模型包括 LSGAN^[61] (least squares GAN) 中使用的 Person χ^2 散度、EBGAN^[38] 中使用的总体方差等。

IPM^[80]是一种度量两个分布之间距离的方法, 它首先规定一个实值函数集, 再寻找集合中使式(9)中两个分布差异最大的函数, 将最大差异值定义为分布距离。通过式(8)计算得到的值是真实值而非概率值。

$$D_f(p_{\text{data}} \| p_g) = \sup_{f \in F} \{E_{x \sim p_{\text{data}}(x)} [f(x)] - E_{x \sim p_g(x)} [f(x)]\} \quad (9)$$

WGAN 模型^[11]的提出者指出 JSD 在生成分

布与真实分布无交集时会引起生成器梯度消失, 当判别器表现足够好时, KLD 与 JSD 会给出不一致的结果, 影响网络收敛, 因此使用 Wasserstein 距离来取代原有衡量方法, 其定义如式(10)所示, 其中 $\Pi(p_{\text{data}}, p_g)$ 指所有边缘分布为 p_{data} 和 p_g 的联合概率分布。

$$W(p_{\text{data}}, p_g) = \inf_{\gamma \in \Pi(p_{\text{data}}, p_g)} E_{(x,y) \sim \gamma} [\|x - y\|] \quad (10)$$

文献[11]进一步提出, 判别器需要满足 1-Lipschitz 条件以保证其优化目标有界。其思想为: 如果局部最小点附近梯度范围较小, 那么细微扰动对输出影响较低, 模型泛化能力较好, 也较稳定。Lipschitz 条件的定义如式(11)所示。

$$\|D(x'_1) - D(x'_2)\| \leq K \|x'_1 - x'_2\| \quad (11)$$

为满足该条件, WGAN 使用截断以限制判别器网络权值取值范围, 这使得判别器网络中大部分权值落在边界上, 此时生成器无法获得有效的梯度并进行学习, 这可能引发模式崩溃。WGAN-GP^[62]使用式(12)定义的 L2 梯度惩罚项替代截断, 使限制不过于极端。

$$GP|_{x'} : E_{x'} \left[\left(\|\nabla_{x'} D(x')\|_2 - 1 \right)^2 \right] \quad (12)$$

文献[63]提出了如式(13)所示的新梯度惩罚项对 WGAN-GP 进行改进, 即 WGAN-LP, 它可以稳定训练, 降低网络对梯度惩罚项的敏感性。

$$LP|_{x'} : E_{x'} \left[\left(\max \{0, \|\nabla_{x'} D(x')\|_2 - 1\} \right)^2 \right] \quad (13)$$

BWGAN (Banach WGAN) 模型^[64]的提出者认为 WGAN-GP 必须使用 L2 范数使其失去在特定应用场景下选择其他范数的能力, 因此对梯度惩罚项进行泛化, 使模型适用范围从希尔伯特空间推广到巴拿赫空间。

CT-GAN 模型^[65]的提出者认为 WGAN-GP 的梯度惩罚项在训练迭代次数有限的情况下仅在判别器输入样本附近有效, 且当两分布距离较远时梯度惩罚项不能保证真实样本分布的连续性, 因此在其目标函数上增加一项用于约束来自真实样本输入产生的梯度, 其定义如式(14)所示。

$$CT|_{x_1, x_2} : E_x \left[\max(0, \frac{\|D(x_1) - D(x_2)\|}{\|x_1 - x_2\|} - K) \right] \quad (14)$$

文献[66]提出介于 WGAN 与 WGAN-GP 间的折中方案,即 RWGAN (relaxed WGAN),它同时使用非对称截断与梯度惩罚项,说明了该散度在 DRO (distribution ally robust optimization) 和 RPO (robust portfolio optimization) 问题上有较好的表现。

CramerGAN 模型^[67]的提出者认为使用 Wasserstein 距离训练 GAN 时会产生有偏样本梯度,可能导致优化到错误的最小值,因此提出了可以作为分布距离无偏估计量的 Cramer 距离作为其替代方法。

文献[18]提出使用谱范数来约束神经网络权值最大变化范围,具体做法是使用谱归一化方法,即判别器网络的权值矩阵 A 除以 $A^T A$ 最大特征值平方根可以使判别器网络满足 Lipschitz 条件。

WGAN-div^[68]利用文献[81]中给出的结论,得出一种既无须满足 Lipschitz 条件又保持 Wasserstein 距离性质的方法,将其称为 Wasserstein 散度。

McGAN^[69] (mean and covariance WGAN) 在 IPM 基础上结合了 MMD (maximum mean discrepancy) 的思想, MMD 可以在 RKHS (reproducing kernel Hilbert space) 中度量分布差异。文献[67]将分布距离定义为具有权值矩阵 ω 的生成器网络分别接受真实样本和生成样本输入后,所输出均值差异的最大值,并在此基础上进一步同时考虑方差特征。

GMMN (generative moment matching networks) 模型^[70]的提出者认为 GAN 的极小极大问题难以优化,因此提出直接以 MMD 的平方作为生成器目标函数,其定义如式(15)所示。

$$D_{\text{MMD}^2} = \left\| \frac{1}{N} \sum_{i=1}^N \phi(x_{\text{data}}) - \frac{1}{M} \sum_{j=1}^M \phi(x_g) \right\|^2 \quad (15)$$

其中, ϕ 是一个再生核变换,它可以将样本从原空间映射到目标空间,文献[68]使用高斯核函数。最小化该目标函数可以使生成分布的统计量尽可能与真实分布相似,该方法不再需要判别器网络。模型使用一个多元均匀分布上的随机采样作为输入,将其经过多层神经网络后的输出作为生成样本。

MMDGAN^[71]基于 GMMN 进行改进,引入了 GAN 的对抗学习思想,将核变换由固定的高斯核

函数转换为神经网络对抗学习得出的核函数,从而改进了 GMMN 生成样本的质量。

FisherGAN^[72]在 IPM 基础上结合了 Fisher 判别分析思想,即利用投影技术降维并得到同一类组内偏差与不同类组间偏差,使用凸优化方法寻找使组内偏差最小化、组间偏差最大化的平面以分割不同类别。该方法在判别器的二阶矩上引入数据相关约束以区分两种分布,提升了 GAN 在半监督学习上的效果。

IGAN (improved GAN)^[73]采用最大平均偏差的思想,提出了新的目标函数,使生成样本和真实样本在通过判别器网络时中间层的特征尽可能相同,还将样本以小批量的方式输入判别器进行训练,防止其快速收敛到特定的点,此外,在损失函数中增加一项用以使梯度持续向均衡点更新,对标签设置参数进行平滑处理,通过这些方法的综合使用提升 GAN 在半监督学习上的效果。

MIX+GAN 模型^[74]的提出者认为 JSD 与 Wasserstein 距离均无法较好地衡量模型泛化能力,因此提出一种基于判别器的神经网络距离用于衡量模型泛化能力,并说明了使用该距离训练更稳定。

OT (optimal transport) 理论提供了一种衡量分布距离的方法,旨在找到一种在给定损失函数上的传输方案,使一种概率测度转移为另一种概率测度的总成本最小。OT-GAN^[75]将最优传输中的 Sinkhorn 算法与文献[67]相结合,提出了最小批处理能量距离。文献[75]利用该距离使用判别器将生成样本与真实样本的特征在隐空间中对齐,并对特征的距离进行最小化。

MMGAN^[76] (manifold matching GAN) 引入流形学习思想,即高维数据是低维数据含有冗余的映射,样本作为高维数据,通常可设法在低维空间唯一表示。文献[74]将真实分布与生成分布各作为一个流形,判别器网络不输出一个值,而是直接输出样本的向量表示,训练目的是使两个流形尽可能地匹配。

4.2 基于梯度计算过程改进的 GAN 模型

使用梯度下降等方法训练神经网络的过程是使用损失函数值越来越小的过程,因此优化损失函数的

计算使其在训练过程中稳定而不产生梯度消失等问题,可以提升模型收敛速度,提高生成样本质量。

MAGAN (margin adaptation GAN) 模型^[77]的提出者基于 EBGAN 中引入合页损失函数从而忽略高能量的生成样本这一思想,进一步提出自适应合页损失函数,即引入损失函数裕度,并根据能量预期自动调节裕度,以此提高模型稳定性。

LSGAN 模型^[61]的提出者认为判别器使用 sigmoid 函数可能导致训练过程中梯度消失,因此提出使用最小二乘损失函数加以解决。

SoftmaxGAN 模型^[78]的提出者认为 Softmax 损失函数梯度始终非零,不会造成判别器训练过程中梯度消失的问题,因此可以用于替代基本 GAN 模型中判别器网络使用的二值交叉熵损失函数。

RGAN (relativistic GAN) 模型^[79]的提出者认为判别器未将“输入样本一半为真实样本,一半为生成样本”作为先验知识加以利用,即未考虑真实样本对训练的影响,生成器应当既能提升生成样本被判别为真实样本的概率,又能降低真实样本被判别为真实的概率。文献[77]通过重新定义判别器的损失函数提高了生成样本的质量和训练稳定性。记判别器原始输出为 $C(\mathbf{x})$, 激活函数为 a , 则有

$$D(x') = a(C(x_{\text{data}}) - C(x_g)) \quad (16)$$

5 GAN 模型的评价指标

GAN 模型的评价指标用于定性或定量评价 GAN 模型的生成效果。文献[82]认为好的评价指标应偏向具有如下特点:生成样本与真实样本相似;生成样本在类内、类间保持多样化;模型在隐空间采样可控;对改变样本语义的失真和变化敏感。同时指标自身应该具有的特点包括:有明确的值域且值的大小能够反映对模型较好或较差的评价、对样本数量的需求低、计算复杂度低等。

IS (inception scores) 模型^[73]的提出者认为质量较高的生成样本更容易被明确地分类,且生成样本在各个类中均匀分布时样本多样性较好。以使用 Inception-v3^[83]作为分类器为例,预训练好的模型接收一幅图像作为输入,输出一个 1 000 维向量,每一维值为输入属于某一类别的概率。样本质量越高,输出向量中的某一个值越趋近于 1,

而其他值越趋近于 0,样本多样性越好,样本在不同类上的分布越趋向均匀分布。IS 指标可定义为式(17),其值越大说明模型越好。

$$IS(G) = \exp(E_{x \sim p_g} D_{\text{KL}}(p(y|x) \| p(y))) \quad (17)$$

IS 有一些衍生指标。MS^[84] (mode score) 与 IS 类似,但考虑了真实样本上标签的先验分布。m-IS^[24] (modified inception score) 在 IS 基础上通过对划分到同一类的样本计算交叉熵来对类内样本多样性进行衡量。AM Score^[85] 在 IS 基础上考虑真实样本数据分布不均匀的情况,要求最小化生成样本与真实样本的标签分布的 KLD。

文献[86]详细探讨了 IS 指标的局限性。从适用范围来说,它要求分类模型与生成模型在同一数据集上训练,且分类模型参数的轻微变动会影响计算结果;从计算过程来说,计算经验分布的数据量过少会影响计算结果;从意义上来说,IS 的两个基本观点不够合理,因此 IS 值较大不能代表生成模型较好,即单独使用 IS 作为评价指标是不足的。

FID (Fréchet Inception distance)^[87]的主要思想是:既然预训练网络模型可以提取样本特征信息,那么分别提取真实样本与生成样本特征信息,假设特征符合多元高斯分布,再计算分布间 Fréchet 距离,距离近则生成图片质量较高,多样性较好。记真实样本和生成样本特征均值分别为 μ_{data} 和 μ_g , Σ_{data} 和 Σ_g 分别为两者的协方差矩阵, FID 的定义如下。

$$FID = \|\mu_{\text{data}} - \mu_g\|^2 + \text{tr} \left(\Sigma_{\text{data}} + \Sigma_g - 2(\Sigma_{\text{data}} \Sigma_g)^{\frac{1}{2}} \right) \quad (18)$$

FID 值越小,说明生成模型表现越好。FID 比 IS 更合理,避免了许多 IS 的缺陷,在实际应用中较为有效,但其多元高斯分布的假设在实际中不成立。

1-NN 分类器 (1-nearest neighbor classifier)^[88]的做法是为真实样本打上正标签,为生成样本打上负标签,分类器每次判断一对图像,计算保留正标签样本的准确率,准确率越靠近 50%,说明分布距离越近,则 GAN 的效果越好。

GAN-train & GAN-test^[89]用于评价生成多种

类样本的 GAN, 它包含 3 个指标: GAN-train 是使用生成样本训练分类器网络, 再使用该网络对真实样本进行分类得到的准确率, 用于衡量生成样本的多样性; GAN-test 是使用真实样本训练分类器网络, 再使用该网络对生成样本进行分类得到的准确率, 用于衡量生成样本的真实性; GAN-base 是用真实样本训练分类器网络, 再使用该网络对真实样本进行分类得到的准确率, 当 GAN 的效果较好时, GAN-train、GAN-test 与 GAN-base 的值接近。

NRDS^[90] (normalized relative discriminative score) 使用带有正标签的真实样本和带有负标签的生成样本训练分类器网络, 认为只要在样本上训练足够多轮, 生成器总可以将真实样本和生成样本区分开, 且训练的轮数越多, 认为生成样本与真实样本越接近。

总而言之, 目前 GAN 的评价指标尚未形成统一的标准, 不少评价指标的假设基于人的主观感受提出, 缺乏严谨的理论依据。所以, 在对 GAN 模型进行评价时, 通常根据评价的侧重点, 选取更适合的指标, 或者自行设计新的评价指标。

6 GAN 的挑战和研究方向

目前, 虽然在 GAN 的研究方面取得了很大的进展, 但仍存在一些复杂的问题需要研究和解决。

6.1 模式崩溃问题

尽管现有研究在解决模式崩溃问题上进行了很多尝试, 也取得了一些进展, 但如何解决模式崩溃问题依然是 GAN 面临的主要挑战。

针对 GAN 发生模式崩溃的原因, 已有一些研究工作尝试给予解释: 文献[91]将生成器视为一个 N 维流形的参数化描述, 当流形上某点的切线空间维数小于 N , 导致在该点沿一些方向进行变化时, 数据的变化无效, 因此生成器会产生单一的数据; 文献[92]基于最优传输理论, 认为生成器将隐空间的分布映射为流形上的分布是一个传输映射, 它具有间断点, 是非连续映射, 但神经网络目前仅能近似连续映射, 从而导致生成无意义结果并引发模式崩溃; 文

献[93]认为当模式崩溃发生时, 判别器网络权值矩阵的奇异值急剧减小, 可从该问题入手解决模式崩溃问题。

与普通神经网络训练过程相比, GAN 模型中存在生成器 G 与判别器 D 之间的博弈机制, 这使得 GAN 模式崩溃问题变得复杂。总而言之, GAN 模式崩溃问题研究工作尚处于起步阶段, 研究出发角度多样, 未形成一个统一的框架来解释该问题。今后的工作如果能从 GAN 的博弈机制出发, 将生成器和判别器两方面的相关因素综合起来, 会有助于该问题的解决。

6.2 训练集样本的影响

神经网络的表现主要取决于模型自身的特点, 以及训练使用的真实样本集。同样, GAN 模型的训练学习的质量也受制于训练样本集的影响。

一方面, 样本集的自身内在数据分布情况可能会影响 GAN 的训练效率和生成质量。例如, 文献[94]在样本集上定义了类内距离集与类间距离集, 并依此提出基于距离的可分性指数, 用于量化样本可分性, 并指出当不同种类样本按相同分布混合时最难以区分, 使用这种样本集进行有监督学习时很难使模型有较好表现。这对于 GAN 的样本生成质量评价指标设计具有借鉴意义。

另一方面, GAN 模型的一大特点是学习真实样本分布, 因此需要足够多真实样本进行训练才能有较好表现, 研究如何使用小规模训练集得到较好的 GAN 模型是具有挑战 and 意义的。GAN 模型对训练集质量也有较高要求, 而高质量的数据集往往难以获得, 因此研究哪些数据会影响模型表现, 如何规避低质量样本带来的负面影响, 以降低对训练集质量的高要求, 成为今后的研究方向。

此外, 在降低训练集样本数量需求方面已有一些研究。文献[95]通过迁移学习, 在预训练的生成器网络和判别器网络上使用适当的样本进行微调, 但样本严重不足或样本与预训练数据区别较大时效果不佳。文献[96]认为网络权值的奇异值与生成样本的语义有关, 因此通过对网络权值进行奇异值分解, 微调预训练模型的奇异值来达到使用较少样本训练的目的。文献[97]在 GAN 上

使用元学习, 在小样本训练问题上取得了一定的效果。文献[98]使用重建损失和三元组损失改造 GAN 的损失函数, 从而将自监督学习的思想引入 GAN 中, 在小样本训练问题上取得了一些效果。文献[99]设计了自适应判别器增强方法, 将训练集所需数据缩小为原有的 1/10~1/20。

对于降低训练集样本质量需求的研究已有一些研究。文献[100]提出构建噪声标签转移模型, 将其合并文献[28]的分类器中, 从而排除有监督学习场景下噪声的影响。NRGAN^[101]在模型中设置了图像生成器和噪声生成器, 分别用以学习真实样本中的数据分布和噪声分布, 从而在无须预知噪声分布的情况下从有噪训练集中生成无噪样本。

目前, 有关训练集样本对 GAN 的影响的研究仍处于初期, 缩小训练集规模往往导致对复杂模式支持较差, 而降低训练集样本质量需求则伴随着过多假设, 如文献[100]设置了标签的噪声转移矩阵, 文献[101]引入训练集的噪声是被指定分布的。后续工作应进一步研究产生这些限制的原因, 并以此为指导使其应用场景更符合真实情况。

6.3 与模型鲁棒性问题研究的交叉

神经网络鲁棒性反映当输入数据集上出现微小扰动后, 模型依然能在输出端表现出抗干扰的能力^[102]。GAN 的研究与人工神经网络鲁棒性的研究相辅相成, 密切相关。一方面, GAN 使用对抗样本对网络模型进行训练, 有助于提升模型的鲁棒性^[103]。另一方面, 神经网络鲁棒性的相关研究与 GAN 的改进存在内在联系, 如文献[104]提出深度神经网络经过对抗训练后损失在峰值附近更加平滑, 以及在 CNN 中使用 Lipschitz 条件可以使模型同时具有较好的鲁棒性与准确性, 这些方法与文献[11]、文献[62]、文献[63]在思想上相似, 因此该领域的相关研究对于 GAN 的改进有一定的参考借鉴价值, 特别是在生成对抗样本质量的评价和生成器的目标研究方面。

文献[105]提出一种度量鲁棒性的方法, 从对抗频度和对抗严重程度两方面描述神经网络在数据集上的鲁棒性。其中对抗频度反映数据集上对抗性扰动发生的可能性, 对抗严重程度反映扰动发生时导致输出偏离的程度。该方法在 GAN 生

成对抗样本数据集质量的评价层面具有借鉴价值, 并对生成器的训练具有指导意义。文献[106]提出一种基于符号线性松弛的神经网络安全性分析方法, 把对抗性扰动当作安全属性违反的一种约束特例来处理, 其框架可以定义 5 种不同的安全属性约束, 针对对抗性扰动的结果进行细化。这些工作有助于 GAN 生成器设计目标的分类研究。

7 结束语

作为近期人工智能领域的热点问题之一, GAN 模型的研究可以提升神经网络的鲁棒性和安全性。本文将 GAN 模型的研究总结归纳两大方向(基于模型结构的改进和基于训练过程的改进), 分别从输入输出、生成器、判别器、多模块组合、模型交叉、分布距离度量、梯度计算过程 7 个维度对近年来 GAN 的主要研究工作、改进机理和特点进行了归纳总结, 并研究了现有用于评价 GAN 模型生成样本质量的指标, 发现现有 GAN 模型的改进主要围绕提升模型训练效率、降低模式崩溃现象发生概率和提升生成样本质量 3 类问题的解决加以展开。在此基础上, 本文从模式崩溃问题解决、训练集数据集影响以及对鲁棒性问题研究交叉 3 方面, 探讨了 GAN 研究可能遇到的问题与挑战, 提出了未来的研究方向。

参考文献:

- [1] SMOLENSKY P. Information processing in dynamical systems: Foundations of harmony theory[R]. 1986.
- [2] HINTON G, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006 18(7): 1527-1554.
- [3] SALAKHUTDINOV R, HINTON G. Deep boltzmann machines[C]//Artificial Intelligence and Statistics. 2009: 448-455.
- [4] KINGMA D P, WELING M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.
- [5] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. 2014: 2672-2680.
- [6] KOO S. Automatic colorization with deep convolutional generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 212-217.
- [7] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

- [8] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2223-2232.
- [9] BAO J, CHEN D, WEN F, et al. CVAE-GAN: fine-grained image generation through asymmetric training[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2745-2754.
- [10] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery[C]//International Conference on Information Processing in Medical Imaging. 2017: 146-157.
- [11] ARJOVSKY M, CHINTALA S, BOTTOU L. WassersteinGAN[J]. arXiv preprint arXiv:1701.07875, 2017.
- [12] WOLTERINK J M, LEINER T, ISGUM I. Blood vessel geometry synthesis using generative adversarial networks[J]. arXiv preprint arXiv:1804.04381, 2018.
- [13] 刘西蒙, 谢乐辉, 王耀鹏, 等. 深度学习中的对抗攻击与防御[J]. 网络与信息安全学报, 2020(5): 36-53.
- LIU X M, XIE L H, WANG Y P, et al. Adversarial attacks and defenses in deep learning[J]. Chinese Journal of Network and Information Security, 2020, 6(5): 36-53.
- [14] 张煜, 吕锡香, 邹宇聪, 等. 基于生成对抗网络的文本序列数据集脱敏[J]. 网络与信息安全学报, 2020(4): 109-119.
- ZHANG Y, LYU X X, ZOU Y C, et al. Differentially private sequence generative adversarial networks for data privacy masking[J]. Chinese Journal of Network and Information Security, 2020, 6(4): 109-119.
- [15] BEAULIEU-JONES B K, WU Z S, WILLIAMS C, et al. Privacy-preserving generative deep neural networks support clinical data sharing[J]. Circulation: Cardiovascular Quality and Outcomes, 2019, 12(7): e005122.
- [16] 郭鹏, 钟尚平, 陈开志, 等. 差分隐私 GAN 梯度裁剪阈值的自适应选取方法[J]. 网络与信息安全学报, 2018, 4(5): 10-20.
- GUO P, ZHONG S P, CHEN K J, et al. Adaptive selection method of differential privacy GAN gradient clipping thresholds[J]. Chinese Journal of Network and Information Security, 2018, 4(5): 10-20.
- [17] 王旭东, 卫红权, 高超, 等. 身份保持约束下的人脸图像补全[J]. 网络与信息安全学报, 2018, 4(8): 71-76.
- WANG X D, WEI H Q, GAO C, et al. Identity preserving face completion with generative adversarial networks[J]. Chinese Journal of Network and Information Security, 2018, 4(8): 71-76.
- [18] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks[J]. arXiv preprint arXiv:1802.05957, 2018.
- [19] GOODFELLOW I. NIPS 2016 tutorial: generative adversarial networks[J]. arXiv preprint arXiv:1701.00160, 2016.
- [20] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.
- [21] JAISWAL A, ABDALMAGEED W, WU Y, et al. Bidirectional conditional generative adversarial networks[C]//Asian Conference on Computer Vision. 2018: 216-232.
- [22] PERARNAU G, VAN DE WEIJER J, RADUCANU B, et al. Invertible conditional GANs for image editing[J]. arXiv preprint arXiv:1611.06355, 2016.
- [23] CHEN X, DUAN Y, HOUTHOOFT R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[C]//Advances in Neural Information Processing Systems. 2016: 2172-2180.
- [24] GURUMURTHY S, KIRAN SARVADEVABHATLA R, VENKATESH BABU R. DeLiGAN: Generative adversarial networks for diverse and limited data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 166-174.
- [25] MISHRA D, JAYENDRAN A, SRIVASTAVA V, et al. Mode matching in GANs through latent space learning and inversion[J]. arXiv preprint arXiv:1811.03692, 2018.
- [26] KWAK H, ZHANG B T. Ways of conditioning generative adversarial networks[J]. arXiv preprint arXiv:1611.01455, 2016.
- [27] ODENA A. Semi-supervised learning with generative adversarial networks[J]. arXiv preprint arXiv:1606.01583, 2016.
- [28] ODENA A, OLAH C, SHLENS J. Conditional image synthesis with auxiliary classifier GANs[C]//International Conference on Machine Learning. 2017: 2642-2651.
- [29] TOLSTIKHIN I O, GELLY S, BOUSQUET O, et al. AdaGAN: boosting generative models[C]//Advances in Neural Information Processing Systems. 2017: 5424-5433.
- [30] GHOSH A, KULHARIA V, NAMBOODIRI V P, et al. Multi-agent diverse generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8513-8521.
- [31] HOANG Q, NGUYEN T D, LE T, et al. MGAN: training generative adversarial nets with multiple generators[C]//International Conference on Learning Representations. 2018.
- [32] GHOSH A, KULHARIA V, NAMBOODIRI V. Message passing multi-agent GANs[J]. arXiv preprint arXiv:1612.01294, 2016.
- [33] LIN Z, KHETAN A, FANTI G, et al. PacGAN: The power of two samples in generative adversarial networks[J]. IEEE Journal on Selected Areas in Information Theory, 2020, 1(1): 324-335.
- [34] DURUGKAR I, GEMP I, MAHADEVAN S. Generative multi-adversarial networks[J]. arXiv preprint arXiv:1611.01673, 2016.
- [35] MORDIDO G, YANG H, MEINEL C. Dropout-GAN: learning from a dynamic ensemble of discriminators[J]. arXiv preprint arXiv:1807.11346, 2018.
- [36] NGUYEN T, LE T, VU H, et al. Dual discriminator generative adversarial nets[C]//Advances in Neural Information Processing Systems. 2017: 2670-2680.
- [37] NEYSHABUR B, BHOJANAPALLI S, CHAKRABARTI A. Stabilizing GAN training with multiple random projections[J]. arXiv preprint arXiv:1705.07831, 2017.
- [38] ZHAO J, MATHIEU M, LECUN Y. Energy-based generative adversarial network[J]. arXiv preprint arXiv:1609.03126, 2016.
- [39] BERTHELOT D, SCHUMM T, METZ L. BeGAN: boundary equilibrium generative adversarial networks[J]. arXiv preprint arXiv:1703.10717, 2017.
- [40] IM D J, KIM C D, JIANG H, et al. Generating images with recurrent adversarial networks[J]. arXiv preprint arXiv:1602.05110, 2016.
- [41] ZHANG H, XU T, LI H, et al. StackGAN: text to photo-realistic

- image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5907-5915.
- [42] KARRAS T, AILA T, LAINE S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. arXiv preprint arXiv:1710.10196, 2017.
- [43] LI C, XU T, ZHU J, et al. Triple generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2017, 30: 4088-4098.
- [44] LEE M, SEOK J. Controllable generative adversarial network[J]. Ieee Access, 2019, 7: 28158-28169.
- [45] CHAVDAROVA T, FLEURET F. SGAN: An alternative training of generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9407-9415.
- [46] KIM Y, KIM M, KIM G. Memorization precedes generation: Learning unsupervised gans with memory networks[J]. arXiv preprint arXiv:1803.01500, 2018.
- [47] JAISWAL A, ABDALMAGEED W, WU Y, et al. CapsuleGAN: Generative adversarial capsule network[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018.
- [48] LARSEN A B L, SØNDERBY S K, LAROCHELLE H, et al. Autoencoding beyond pixels using a learned similarity metric[C]//International Conference on Machine Learning. 2016: 1558-1566.
- [49] ZHONG G, GAO W, LIU Y, et al. Generative adversarial networks with decoder-encoder output noises[J]. Neural Networks, 2020.
- [50] MAKHZANI A, SHLENS J, JAITLEY N, et al. Adversarial autoencoders[J]. arXiv preprint arXiv:1511.05644, 2015.
- [51] DONAHUE J, KRÄHENBÜHL P, DARRELL T. Adversarial feature learning[J]. arXiv preprint arXiv:1605.09782, 2016.
- [52] DUMOULIN V, BELGHAZI I, POOLE B, et al. Adversarially learned inference[J]. arXiv preprint arXiv:1606.00704, 2016.
- [53] MÁTTYUS G, URTASUN R. Matching adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8024-8032.
- [54] ZHANG H, GOODFELLOW I, METAXAS D, et al. Self-attention generative adversarial networks[C]//International Conference on Machine Learning. 2019: 7354-7363.
- [55] WANG X, ZHANG R, SUN Y, et al. KdGAN: Knowledge distillation with generative adversarial networks[C]//Advances in Neural Information Processing Systems. 2018: 775-786.
- [56] WANG J, YU L, ZHANG W, et al. IrGAN: a minimax game for unifying generative and discriminative information retrieval models[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 515-524.
- [57] DENTON E L, CHINTALA S, FERGUS R. Deep generative image models using a laplacian pyramid of adversarial networks[C]//Advances in Neural Information Processing Systems. 2015: 1486-1494.
- [58] LLOYD S, WEEDBROOK C. Quantum generative adversarial learning[J]. Physical Review Letters, 2018, 121(4).
- [59] SAATCI Y, WILSON A G. Bayesian GAN[C]//Advances in Neural Information Processing Systems. 2017: 3622-3631.
- [60] NOWOZIN S, CSEKE B, TOMIOKA R. f-GAN: training generative neural samplers using variational divergence minimization[C]//Advances in Neural Information Processing Systems. 2016: 271-279.
- [61] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2794-2802.
- [62] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of WassersteinGANs[C]//Advances in Neural Information Processing Systems. 2017: 5767-5777.
- [63] PETZKA H, FISCHER A, LUKOVNICOV D. On the regularization of WassersteinGANs[J]. arXiv preprint arXiv:1709.08894, 2017.
- [64] ADLER J, LUNZ S. Banach WassersteinGAN[C]//Advances in Neural Information Processing Systems. 2018: 6754-6763.
- [65] WEI X, GONG B, LIU Z, et al. Improving the improved training of WassersteinGANs: a consistency term and its dual effect[J]. arXiv preprint arXiv:1803.01541, 2018.
- [66] GUO X, HONG J, LIN T, et al. Relaxed wasserstein with applications to GANs[J]. arXiv preprint arXiv:1705.07164, 2017.
- [67] BELLEMARE M G, DANIHELKA I, DABNEY W, et al. The cramer distance as a solution to biased wasserstein gradients[J]. arXiv preprint arXiv:1705.10743, 2017.
- [68] WU J, HUANG Z, THOMA J, et al. Wasserstein divergence for GANs[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 653-668.
- [69] MROUEH Y, SERCU T, GOEL V. McGAN: Mean and covariance feature matching GAN[J]. arXiv preprint arXiv:1702.08398, 2017.
- [70] LI Y, SWERSKY K, ZEMEL R. Generative moment matching networks[C]//International Conference on Machine Learning. 2015: 1718-1727.
- [71] LI C L, CHANG W C, CHENG Y, et al. MmdGAN: towards deeper understanding of moment matching network[C]//Advances in Neural Information Processing Systems. 2017: 2203-2213.
- [72] MROUEH Y, SERCU T. Fisher GAN[C]//Advances in Neural Information Processing Systems. 2017: 2513-2523.
- [73] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training GANs[C]//Advances in Neural Information Processing Systems. 2016: 2234-2242.
- [74] ARORA S, GE R, LIANG Y, et al. Generalization and equilibrium in generative adversarial nets (GANs)[J]. arXiv preprint arXiv:1703.00573, 2017.
- [75] SALIMANS T, ZHANG H, RADFORD A, et al. Improving GANs using optimal transport[J]. arXiv preprint arXiv:1803.05573, 2018.
- [76] PARK N, ANAND A, MONIZ J R A, et al. MMGAN: manifold-Matching generative adversarial networks[C]//2018 24th International Conference on Pattern Recognition (ICPR). 2018: 1343-1348.
- [77] WANG R, CULLY A, CHANG H J, et al. MaGAN: margin adaptation for generative adversarial networks[J]. arXiv preprint arXiv:1704.03817, 2017.
- [78] LIN M. SoftmaxGAN[J]. arXiv preprint arXiv:1704.06191, 2017.
- [79] JOLICOEUR-MARTINEAU A. The relativistic discriminator: a key element missing from standard GAN[J]. arXiv preprint ar-

- Xiv:1807.00734, 2018.
- [80] MÜLLER A. Integral probability metrics and their generating classes of functions[J]. *Advances in Applied Probability*, 1997: 429-443.
- [81] EVANS L C. Partial differential equations and monge-kantorovich mass transfer[J]. *Current Developments in Mathematics*, 1997, 1997(1): 65-126.
- [82] BORJI A. Pros and cons of GAN evaluation measures[J]. *Computer Vision and Image Understanding*, 2019, 179: 41-65.
- [83] SZEGEDY C, VANHOUCHE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016: 2818-2826.
- [84] CHE T, LI Y, JACOB A P, et al. Mode regularized generative adversarial networks[J]. *arXiv preprint arXiv:1612.02136*, 2016.
- [85] ZHOU Z, CAI H, RONG S, et al. Activation maximization generative adversarial nets[J]. *arXiv preprint arXiv:1703.02000*, 2017.
- [86] BARRATT S, SHARMA R. A note on the inception score[J]. *arXiv preprint arXiv:1801.01973*, 2018.
- [87] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium[C]//*Advances in Neural Information Processing Systems*. 2017: 6626-6637.
- [88] LOPEZ-PAZ D, OQUAB M. Revisiting classifier two-sample tests[J]. *arXiv preprint arXiv:1610.06545*, 2016.
- [89] SHMELKOV K, SCHMID C, ALAHARI K. How good is my GAN[C]//*Proceedings of the European Conference on Computer Vision (ECCV)*. 2018: 213-229.
- [90] ZHANG Z, SONG Y, QI H. Decoupled learning for conditional adversarial networks[C]//*2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018: 700-708.
- [91] QI G J, ZHANG L, HU H, et al. Global versus localized generative adversarial nets[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018: 1517-1525.
- [92] LEI N, GUO Y, AN D, et al. Mode collapse and regularity of optimal transportation maps[J]. *arXiv preprint arXiv:1902.02934*, 2019.
- [93] LIU K, TANG W, ZHOU F, et al. Spectral regularization for combating mode collapse in GANs[C]//*Proceedings of the IEEE International Conference on Computer Vision*. 2019: 6382-6390.
- [94] GUAN S, LOEW M, KO H. Data separability for neural network classifiers and the development of a separability index[J]. *arXiv preprint arXiv:2005.13120*, 2020.
- [95] MO S, CHO M, SHIN J. Freeze discriminator: a simple baseline for fine-tuning GANs[J]. *arXiv preprint arXiv:2002.10964*, 2020.
- [96] ROBB E, CHU W S, KUMAR A, et al. Few-shot adaptation of generative adversarial networks[J]. *arXiv preprint arXiv: 2010.11943*, 2020.
- [97] ZHANG R, CHE T, GHAHRAMANI Z, et al. MetaGAN: an adversarial approach to few-shot learning[C]//*Advances in Neural Information Processing Systems*. 2018: 2365-2374.
- [98] NGUYEN K, TODOROVIC S. A self-supervised GAN for unsupervised few-shot object recognition[J]. *arXiv preprint arXiv:2008.06982*, 2020.
- [99] KARRAS T, AITTALA M, HELLSTEN J, et al. Training generative adversarial networks with limited data[J]. *Advances in Neural Information Processing Systems*, 2020, 33.
- [100] KANEKO T, USHIKU Y, HARADA T. Label-noise robust generative adversarial networks[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019: 2467-2476.
- [101] KANEKO T, HARADA T. Noise robust generative adversarial networks[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020: 8404-8414.
- [102] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//*2nd International Conference on Learning Representations, ICLR 2014*. 2014.
- [103] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. *Stat*, 2015, 1050: 20.
- [104] BAI T, LUO J, ZHAO J. Recent advances in understanding adversarial robustness of deep neural networks[J]. *arXiv preprint arXiv:2011.01539*, 2020.
- [105] BASTANI O, IOANNOU Y, LAMPROPOULOS L, et al. Measuring neural net robustness with constraints[C]//*Advances in Neural Information Processing Systems*. 2016: 2613-2621.
- [106] WANG S, PEI K, WHITEHOUSE J, et al. Efficient formal safety analysis of neural networks[C]//*Advances in Neural Information Processing Systems*. 2018: 6367-6377.

[作者简介]



王正龙 (1997-), 男, 宁夏银川人, 上海交通大学硕士生, 主要研究方向为生成对抗网络、人工神经网络鲁棒性与安全性。



张保稳 (1975-), 男, 山东菏泽人, 上海交通大学副研究员, 主要研究方向为神经网络鲁棒性与安全性、网络信息安全本体, 以及新型网络系统安全性分析与评估。