生成式对抗网络研究进展

王万良，李卓蓉
（浙江工业大学计算机科学与技术学院，浙江 杭州 310024）

摘 要：生成式对抗网络（GAN, generative adversarial network）对生成式模型的发展具有深远意义，自提出后立刻受到人工智能学术界和工业界的广泛研究与高度关注，随着深度学习的技术发展，生成式对抗模型在理论和应用上得到不断推进。首先，阐述生成对抗模型的研究背景与意义，然后，详细论述生成式对抗网络在建模、架构、训练和性能评估方面的研究进展及其具体应用现状，最后，进行分析与总结，指出生成式对抗网络研究中亟待解决的问题以及未来的研究方向。

关键词：深度学习；生成式对抗网络；卷积神经网络；自动编码器；对抗训练
中图分类号：TP183
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2018032

Advances in generative adversarial network

WANG Wanliang, LI Zhuorong
College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310024, China

Abstract: Generative adversarial network (GAN) have swiftly become the focus of considerable research in generative models soon after its emergence, whose academic research and industry applications have yielded a stream of further progress along with the remarkable achievements of deep learning. A broad survey of the recent advances in generative adversarial network was provided. Firstly, the research background and motivation of GAN was introduced. Then the recent theoretical advances of GAN on modeling, architectures, training and evaluation metrics were reviewed. Its state-of-the-art applications and the extensively used open source tools for GAN were introduced. Finally, issues that require urgent solutions and works that deserve further investigation were discussed.

Key words: deep learning, generative adversarial network, convolutional neural network, auto-encoder, adversarial training

1 引言

近年来，深度学习在计算机视觉[1,2]、自然语言处理[3,4]、语音[5]等多个应用领域中都取得了突破性进展，其动机在于建立能够模拟人类大脑神经连接结构的模型，在处理图像、文本和声音等高维信号时，通过组合低层特征形成更加抽象的高层表示、属性类别或特征，进而对数据进行层次化表达[6~8]。深度学习的模型可大致分为判别式模型和生成式模型，目前，深度学习取得的成果主要集中在判别式模型，即将一个高维的感官输入映射为一个类别标签[9,10]，这些成果主要归功于反向传播（BP, back propagation）算法[11]和 Dropout 算法[12,13]对模型的训练。著名物理学家 Richard 指出，要想真正理解一样东西，我们必须能够把它创造出来。因此，要想令机器理解现实世界，并基于此进行推理与创造，从而实现真正的人工智能，必须使机器能够通过观测现实世界的样本，学习其内在统计规律，并基于此生成类似样本，这种能够反映数据内在概率分布规律并生成全新数据的模型为生成式模型。

收稿日期：2017-05-24；**修回日期**：2018-01-17
基金项目：国家自然科学基金资助项目（No.61379123）
Foundation Item: The National Natural Science Foundation of China (No.61379123)

然而，相对判别式模型来说，生成式模型的研究进展较为缓慢，究其原因主要是较高的计算复杂度。典型的生成式模型往往涉及最大似然估计、马尔可夫链方法、近似法等[14]。受限玻尔兹曼机（RBM, restricted Boltzmann machine）[15]及其扩展模型（如深度置信网络[16]、深度玻尔兹曼机[17]）采用最大似然估计法，即令该参数下模型所表示的分布尽可能拟合训练数据的经验分布。最直接的方法是利用梯度上升法求得对数似然函数最大值，但由于样本分布未知且包含归一化函数（也称配分函数）而无法给出参数梯度的解析解，替代方法是基于采样构建以数据分布为平稳分布的马尔可夫链，以获得满足数据分布的样本，然后利用蒙特卡罗迭代对梯度进行近似，这种方法计算复杂。变分自编码器（VAE, variational autoencoder）[18]采用近似法，其性能优劣取决于近似分布的好坏，而该近似分布的假设需要一定的先验知识，此外，由于受变分类方法的局限，VAE 对概率分布的估计是有偏的，在学习过程中对目标函数下界而不是目标函数进行逼近。PixelRNN[19]是自回归模型的一种，将图像生成问题转化为像素序列预测学习问题，假设每个像素的取值只依赖于空间中某种意义的近邻，通过给定的像素对每个像素的条件分布进行建模，采样效率较低。上述生成式模型的复杂训练使之只能生成 MNIST[20]和 CIFAR-10[21]等简单数据集的图片，并不适用于较大尺度的复杂图像。

生成式模型是一个极具挑战的机器学习问题，主要体现在以下 2 点。首先，对真实世界进行建模需要大量先验知识，建模的好坏直接影响生成式模型的性能；其次，真实世界的数据往往非常复杂，拟合模型所需计算量往往非常庞大，甚至难以承受。针对上述两大困难，Goodfellow 等[22]提出一种新型生成式模型——生成式对抗网络（GAN, generative adversarial network），开创性地使用对抗训练机制对 2 个神经网络进行训练，并可使用随机梯度下降（SGD, stochastic gradient descent）实现优化。这避免了反复应用马尔可夫链学习机制带来的配分函数计算，不需变分下限也不需近似推断，从而大大提高了应用效率[23]。尽管 GAN 从提出至今不过两年半时间，但关注和研究热度急速上升，并已从学术界延伸至工业界，Google、OpenAI、Facebook 和 Twitter 等知名人工智能企业纷纷投入大量精力研究和拓展 GAN 的应用[24~27]。目前，GAN 已成功应用于图像生成[28~30]和视频生成[31,32]领域，此外，若干研究工作[33~35]已成功将 GAN 应用在强化学习中。

本文论述了 GAN 在建模、架构、训练和性能评估方面的最新研究进展及其具体应用现状，最后进行分析与总结，指出生成式对抗网络研究中亟待解决的问题。

2 生成式对抗网络

2.1 基本思想

受博弈论中二元零和博弈的启发，GAN 的框架中包含一对相互对抗的模型：判别器和生成器。判别器的目的是正确区分真实数据和生成数据，从而最大化判别准确率；生成器则是尽可能逼近真实数据的潜在分布。为了在博弈中胜出，二者需不断提高各自的判别能力和生成能力，优化的目标就是寻找二者间的纳什均衡。GAN 示意[36]如图 1 所示，生成器（点划线框内的多层感知机）的输入是一个来自常见概率分布的随机噪声矢量 $\boldsymbol{z}$，输出是计算机生成的伪数据；判别器（虚线框内的多层感知机）的输入是图片 x（x 可能采样于真实数据，也可能采样于生成数据），输出是一个标量，用来代表 x 是真实图片的概率，即当判别器认为 x 是真实图片时输出 1，反之输出 0[22]。判别器和生成器不断优化，当判别器无法正确区分数据来源时，可以认为生成器捕捉到真实数据样本的分布。

2.2 标准模型

1) 极大极小博弈

生成器和判别器可以是任意可微函数，因此，可以利用随机梯度下降法（SGD）进行优化，而采用 SGD 的前提是建立一个目标函数来判断和监视学习的效果。由于判别器是一个二分类模型，因此，可用交叉熵表示其目标函数，即

$$J(D)=-\frac{1}{2}\mathrm{E}_{x\sim p_{\mathrm{data}}(x)}\left[\log D(x)\right]-\frac{1}{2}\mathrm{E}_{z\sim p_z(z)}\left[\log(1-D(G(z)))\right] \tag{1}$$

其中，E 是求期望，G 和 D 分别表示生成器与判别器的可微函数，x 是真实数据样本，$\boldsymbol{z}$ 是随机噪声矢量，$G(\boldsymbol{z})$是判别器的生成数据。式(1)第一项表示 D 判断出 x 是真实数据的情况，第二项则表示 D 判别出数据是由生成器 G 将噪声矢量 $\boldsymbol{z}$ 映射而成的生成数据。由于 G 与 D 进行二元零和博弈，因此，生

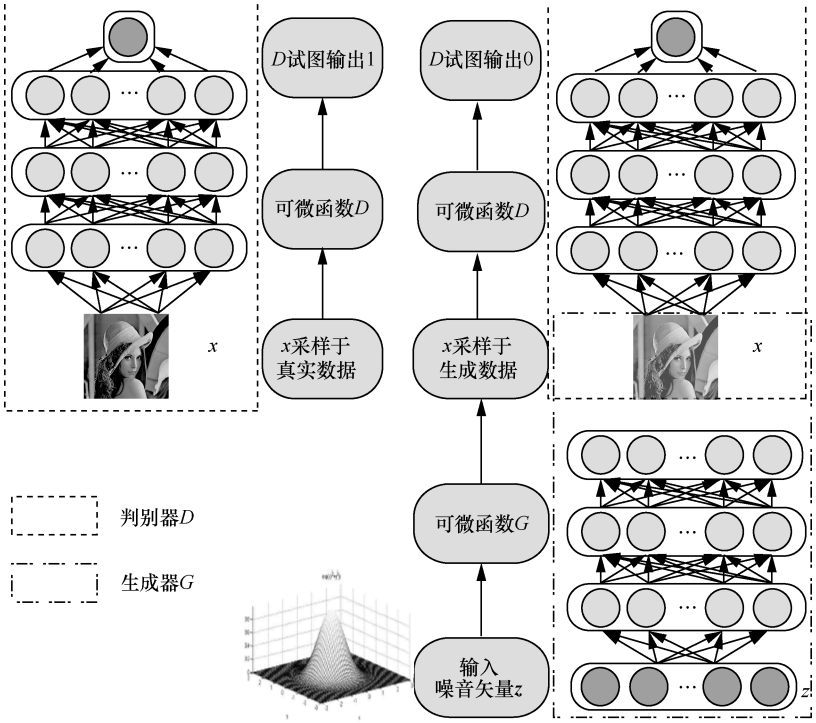


图 1　GAN 示意

成器 G 的目标函数 $J(G)=-J(D)$。因此，GAN 的优化问题可描述为如下极大极小博弈问题。

$$\min_G \max_D V(D,G)=\mathrm{E}_{x\sim p_{\mathrm{data}}(x)}\left[\log D(x)\right]+\mathrm{E}_{z\sim p_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \quad (2)$$

由于在训练初期缺乏足够训练，G 所生成的数据不够逼真，因此，D 很容易就能将生成数据与真实数据区分开来，导致 G 得不到足够梯度。因此，文献[22]提出，通过最大化 $\log D(G(z))$而不是最小化 $\log(1-D(G(z))$来训练 G 是一个更好的策略。

2) 非饱和博弈

为了解决生成器的弱梯度问题，除了采用文献[22]的方法外，还可以把极大极小博弈替换成非饱和博弈，即

$$\begin{cases} J(D)=-\dfrac{1}{2}\mathrm{E}_{x\sim p_{\mathrm{data}}(x)}\left[\log D(x)\right]-\dfrac{1}{2}\mathrm{E}_{z\sim p_z(z)}\left[\log(1-D(G(z)))\right] \\ J(G)=-\dfrac{1}{2}\mathrm{E}_{z\sim p_z(z)}\log D(G(z)) \end{cases} \quad (3)$$

换言之，G 用自己的伪装能力来表示自己的目标函数，而不是简单直接地取 $J(D)$的相反数。从而均衡不再完全由价值函数 min max $V(G,D)$决定，即使 D 准确地拒绝了所有生成样本，G 仍可以继续学习。

2.3　泛化模型

Goodfellow 等[22]从博弈论的角度阐释了 GAN 的思想，即 GAN 的训练目标是使生成器 G 与判别器 D 达到纳什均衡，此时，生成模型 G 产生的数据分布完全拟合真实数据分布。若从信息论角度理解，GAN 所最小化的实际上是真实数据分布和生成分布之间的 Jensen-Shannon 散度。Goodfellow[37]认为 Kullback-Leibler 散度比 Jensen-Shannon 散度更适用于 GAN 的目标函数构建，Sønderby 等[38]和 Kim 等[39]基于 Kullback-Leibler 散度对 GAN 进行建模，通过最小化两者之间的交叉熵进行训练。文献[40]对此进行拓展，提出的 f-GAN 将基于 Jensen-Shannon 散度的 GAN 建模泛化为基于 f-散度的优化目标，从而将 Kullback-Leibler 等经典散度度量也包含在 f-散度中。

2.4　网络结构实现

在生成器 G 和判别器 D 的网络结构方面，朴素生成式对抗网络[22]通过多层感知机（MLP, multi-layer perceptron）来实现。由于卷积神经网络（CNN, convolutional neural network）较 MLP 有更好的抽象能力，DCGAN[28]将朴素生成式对抗网络的 MLP 结构替换为 CNN 结构，考虑到传统 CNN 所包含的池化层并不可微，DCGAN 用步进卷积网络（strided convolution）及其转置结构分别实现判

别器 D 和生成器 G，用于训练过程的空间降采样和升采样。该研究工作提出了 GAN 架构下的一种具体且有效的实现方式和经验指导，成为后续许多理论研究和应用研究的基础。另外，朴素 GAN 的定义域为实数且生成器 G 和判别器 D 均可微，这样设计是为了根据 D 的梯度信息对生成数据进行微调，从而提高生成数据质量。然而，当数据是离散时此方式并不可行，这也是在自然语言处理中应用生成对抗网络的主要障碍。为了生成离散序列，TextGAN[41]和 SeqGAN[42]等模型往往通过循环神经网络实现判别器 D，通过 CNN 实现生成器 G。

3 GAN 的架构

3.1 条件生成式对抗网络

GAN 的最大优点体现在其对抗训练方式通过对 $p(x)$直接采样来逼近真实样本，利用反向传播即可获得梯度而不需复杂的马尔可夫链和推断过程，从而大大简化了计算。然而，文献[22]使用 z 作为先验，但生成式模型如何利用这个先验却是无法控制的。换言之，GAN 的学习模式过于自由而导致 GAN 的训练过程和结果都不可控。为了提高 GAN 的稳定性，Mirza 等[43]提出条件生成式对抗网络（cGAN, conditional GAN），将条件变量 y 作为模型的附加信息以约束生成过程，这种条件变量可以是类别标签甚至还可以是不同模态的数据。GAN 的架构如图 2 所示。图 2(a)是 cGAN 的概念图，可以看出，cGAN 在朴素 GAN 的基础上将条件变量 y 与 z 同时输入生成器 G 中，在判别器 D 中，真实样本 x 和条件变量 y 同时作为判别函数的输入。因此，cGAN 的目标函数在朴素 GAN 的基础上进一步改写为

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim p_{\text{data}}(x)}\left[\log D\left(x|y\right)\right] + \mathrm{E}_{z\sim p_z(z)}\left[\log\left(1-D\left(G\left(z|y\right)\right)\right)\right] \quad (4)$$

cGAN 需要同时对 z 和条件变量 y 进行采样，其中，对随机噪声采样是简单的，但生成条件变量则需多加考虑。最常见的一种方法是直接从训练数据中获取条件变量，例如，文献[43]的条件变量采用的是类别标签 y，其同时作为生成器和判别器的附加输入层。然而，这种情况下生成器可能会记住这些训练样本从而达到虚假的最优。针对这个问题，Gauthier[44]提出在训练过程中基于训练样本的条件变量值构造核密度估计（也称为帕尔森窗口估计），对条件变量进行随机采样。文献[43]和文献[44]都是以类别标签作为条件变量，与之不同，LAPGAN[45]和 GRAN[32]的条件变量是上一级所生成的图片，利用前一步得到的生成结果进行训练，相当于将复杂分布的建模问题转化为一系列简单子问题，从而问题得以大大简化。金字塔中每一级都通过使用 cGAN 来训练一个单独的生成性卷积网络 G_i，以避免模型过拟合，这是 LAPGAN 的显著特点和最大优势。StackGAN[46,47]本质上也是一种 cGAN，基于前一层所生成的分辨率较低图片及文字信息生成分辨率较高的图片。

递归?

Chen 等[48]提出的 InfoGAN 是条件生成式对抗网络中的另一重要模型。同样地，InfoGAN 的生成器 G 的输入包含条件变量，不同的是，这个条件变量是从噪声矢量中拆分出来的结构化隐变量。朴素 GAN 利用其唯一的输入信号 z 能生成与真实数据相一致的模型分布，然而人们并不清楚它究竟是如何将 z 的具体维度与数据的有效语义特征对应起来的，因此，也就无法通过控制 z 以生成期望的数据，针对这个问题，InfoGAN 通过从噪声矢量中拆分出结构化的隐变量，进而使生成过程具备可控性且生成结果具备可解释性。具体地，InfoGAN 将朴素 GAN 中的 z 进行拆解，从而 InfoGAN 中输入的先验变量可拆分为：1）一组用于表示数据语义特征的结构化隐变量，用 c 表示这部分具有可解释性的先验，例如，对于 MNIST 数据集来说，可用 $c_1,c_2,\cdots,c_L$ 表示光照方向、笔画粗细和字体的倾斜角度等；2）不能再压缩的、无法描述的非结构化噪声矢量 z，将 z 和 c 同时输入生成器，如图 2(b)所示。根据信息论，互信息 I(x;y)度量了 y 的信息对 x 不确定性的减少量，因此，为了学习重要的语义特征，可通过最大化隐变量 c 和生成分布 $G(z,c)$的互信息 I(c;$G(z,c)$)使生成过程中的重要特征在生成过程中得到充分学习。InfoGAN 的价值函数为

$$\min_G \max_D V_{\mathrm{I}}(D,G) = V(D,G) - \lambda \mathrm{I}(c;G(z,c)) \quad (5)$$

利用结构化隐变量 c 的可解释性，InfoGAN 能控制生成样本在某个特定语义维度的变化，从而使生成器能生成更符合真实样本的结果。通过引入变分分布逼近真实样本分布，并与互信息下限的优化进行交替迭代，从而实现具体优化。

3.2 双向生成式对抗网络

GAN 通过将简单的隐变量分布映射至任意复

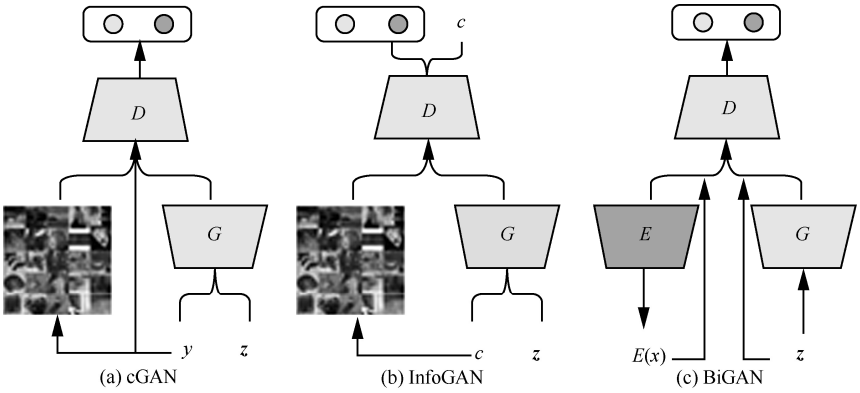


图 2　GAN 的架构

杂的数据分布来生成令人信服的自然图像[28,45]，这表明 GAN 的生成器能对隐空间中的数据分布进行语义线性化，学习到隐空间中数据的良好特征表达。然而，GAN 缺少一种有效的推断机制，未能学习从数据空间映射至隐空间的逆映射[49]。为解决这个问题，Donahue 等[50]和 Dumoulin 等[51]将单向的 GAN 变为双向的 GAN，从而既能进行有效推断又保证了生成样本质量。Donahue 等[50]提出的双向生成式对抗网络（BiGAN, Bidirectional GAN），除了学习标准 GAN 中的生成器 G 和判别器 D 外，还学习了一个将数据映射至隐式表达的编码器 E，如图 2(c)所示。Dumoulin 等[51]提出的 ALI 与 BiGAN 一样，通过编码器 E 学习到的特征表达有助于训练判别器 D，通过同时训练编码器和解码器以迷惑生成器，使判别器难以区分究竟是真实样本及其编码还是生成样本及其隐变量，换言之，双向生成对抗网络的生成器判别的是联合样本（x, z）而不是样本 x。

3.3　自编码生成式对抗网络

基于变分自动编码器的生成模型[15]能学习一种可以用于半监督学习或图像修复等辅助任务的近似推断机制，但该方法的最大似然训练模式会使生成样本比较模糊[52]。而基于 GAN 的生成模型虽然生成样本质量更优，但缺少一种有效的推断机制[53]。Larsen 等[54]将 VAE 和 GAN 并入同一个无监督生成模型中，当将编码器和解码器看作一个生成模型整体时，这个生成模型和判别器构成了扩展的生成式对抗模型；若将解码器和判别器看作一个整体时，这个整体相当于解码器，并与编码器共同构成了扩展的自编码器，因此，该模型结合了 GAN 和 VAE 的优点。Che 等[55]在此基础上提出将 VAE 的重构误差作为遗失模式的正则项，进而提高 GAN 的稳定性和生成样本质量。对抗自编码器（AAE, adversarial autoencoder）[56]将利用自编码器得到的重构误差和对抗训练得到的隐变量与目标分布之间的误差进行组合，从而既能通过自编码器进行推断又能得到结构化的隐变量。

3.4　组合生成式对抗网络

通过对朴素 GAN 进行堆叠、平行或相互反馈，来调整 D 和 G 的组合方式。Wang 等[57]提出 GAN 的自组合和级联组合，前者对经过不同迭代次数的同一模型进行组合，既充分利用模型组合的效果又可避免其带来过多额外计算，后者将多个不同的 cGAN 进行级联，通过门函数将未被充分利用的训练数据传入下一个 GAN 进行重复使用，如图 3(a)所示。Liu 等[58]提出的 CoGAN 包含一对 GAN，每个 GAN 负责生成一个领域的图片，如图 3(b)所示。在训练过程中，共享生成器低层和判别器高层的参数，共享的参数使这一对 GAN 所生成的图片相似，其余不共享的参数使每个 GAN 所生成的图片不完全相同。Im 等[59]提出生成式对抗的平行化 GAP，即不让判别器与固定且唯一的生成器进行对抗训练，而是同时训练几组 GAN，并令每个判别器周期性地与其他 GAN 的生成器进行对抗训练，如图 3(c)所示。GAP 适用于 GAN 的任何扩展模型，因此，可将 GPU 分配给不同的 GAN 衍生模型（如 DCGAN 和 LAPGAN）进行并行计算。并行对抗训练能增加判别器所处理的模式数量，从而有效避免模式坍塌问题，因此，可将 GAP 视为正则化手段。Zhu 等[60]提出的 CycleGAN 包含 2 个判别器 D_x 和 D_y，用于鼓励图片在 2 种不同风格之间的迁移。Li 等[61]提出 TripleGAN，在生成器 G 和判别器 D 的基础上额外增加一个分类器 C，G 和 C 的目的都是使 D 难辨真假，C 的引入避免了判别器 D 既需判别生成样本又需对生成样本进行分类。

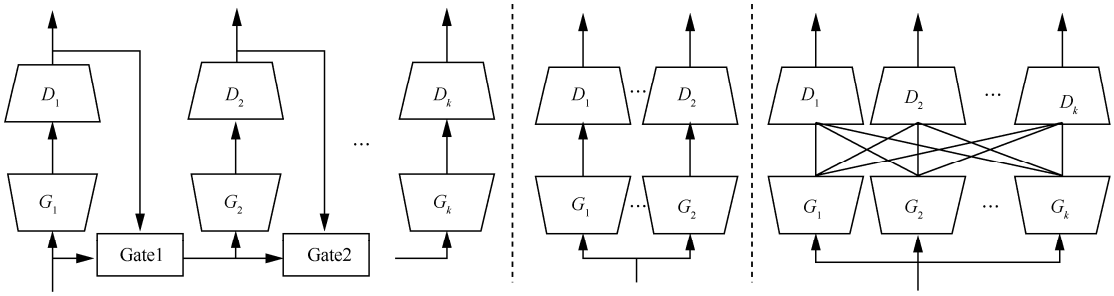


图 3　GAN 的不同组合方式

4　训练与评估

4.1　训练机制

GAN 的训练可采用交替优化方法，即固定 G 的参数以更新 D 的参数，然后固定 D 的参数以更新 G 的参数。Goodfellow 等[22]指出，当生成器 G 固定时，存在唯一的最优判别器 $D^*=\frac{p_{\text{data}}}{p_g+p_{\text{data}}}$；当 $p_g=p_{\text{data}}$ 时，生成器 G 达到最优，此时 $D^*=0.5$，即最优判别器 D^*无法区分真实数据与生成数据。在训练过程中如何平衡生成器 G 和判别器 D 非常关键，理论上，当固定生成器的参数时，判别器的参数不断更新直至最优，然而在实际应用中，往往判别器 D 的参数更新 k 次后，生成器 G 的参数才更新一次。

GAN 的判别器 D 和生成器 G 都是可微函数，因此，可用随机梯度下降进行训练。在判别器 D 接近最优时，生成器 G 的损失函数 $\mathrm{E}_z[\log(1-D(G(z))]$ 实质上是最小化真实分布和生成分布之间的 Jensen-Shannon 散度。然而，当真实分布和生成分布的支撑集是高维空间中的低维流形时，这 2 个分布的重叠部分测度为零的概率为 1，此时，Jensen-Shannon 散度是常数 log2，导致训练梯度消失。针对这个问题，Arjovsky 等[62]提出 Wasserstein-GAN（WGAN），用 Wasserstein 距离代替 Jensen-Shannon 散度来衡量真实分布和生成分布的距离。由于 Wasserstein 距离较 Jensen-Shannon 散度具有更佳的平滑性，解决了梯度消失问题，因此，理论上 WGAN 彻底解决了训练不稳定问题。而且，Wasserstein 距离的连续性和可微性能够提供持续的梯度信息，因此，WGAN 不要求生成器 G 与判别器 D 之间严格平衡。WGAN 中用一个 Lipschitz 连续性约束对应着 GAN 判别器 D 的批评函数 f，在如何进行 Lipschitz 约束的问题上，WGAN 采用了权值截断，然而，带有权值截断的优化器会在一个比 1-Lipschitz 小的空间中搜索判别器，导致判别器偏向非常简单的函数，此外，截断后梯度在反向传播过程中会出现梯度消失或弥散。针对这个问题，Gulrajani 等[63]提出用梯度惩罚代替权值截断来进行 Lipschitz 约束，以获得更快的收敛速度和更高的生成样本质量。此外，朴素 GAN 没有对生成分布做任何假设，要想拟合任意分布必须给判别器 D 引入无限建模能力，而这容易导致过拟合。Qi 等[64]对判别器的无限建模能力进行约束，通过将损失函数限定在满足 Lipschitz 连续性约束的函数类上并使用（真实样本，生成样本）这样的成对统计量来学习批评函数 f，迫使两者之间必须配合，从而实现建模能力的按需分配。

上述研究工作的共同之处在于梯度信息是一阶的，Metz 等[65]提出在训练时对判别目标函数进行展开优化，即生成器当前决策是基于判别器因该决策而采取的后续 k 个决策而生成的，这个方法在本质上是用二阶甚至高阶梯度指导生成器的训练。在标准 GAN 的训练中，生成器和判别器的参数更新都是在其他模型参数固定的前提下对自身参数使用梯度下降法进行更新，而该研究工作中提出的代理损失函数使得模型参数基于其他模型参数的变化而进行更新，具体地，在对生成器使用梯度下降法进行参数更新后，基于判别器的后续 k 步参数更新再去调整生成器的梯度。这种额外的信息能使生成器的概率分布更加平均，从而判别器的下一步不易坍塌至某一个点，但由于对判别目标函数的展开优化涉及二阶甚至高阶梯度，因而计算非常复杂。

4.2　训练技巧

GAN 的训练技巧研究大大加快了 GAN 的研究和应用进展，目前，研究工作[24]是围绕训练技巧展

开的，此外，也有很多研究工作[28,38]提出了针对具体训练问题的技巧，本节将对常用的训练技巧进行简单介绍。

仅以最大化判别器的输出为优化目标容易导致判别器过训练，Salimans 等[24]提出特征匹配（feature matching），将判别器学到的特征作为生成器附加信息，以使生成样本尽量匹配这些统计特征，此时判别器的作用是指出能有效区分真假数据的特征。与其他机器学习算法一样，GAN 所得到的标签最好是经过平滑处理的，例如，将用 0 和 1 表示类别的离散标签替换为更加平滑的 0.1 和 0.9。标签平滑最近被证明能有效降低对抗样本对神经网络的干扰[66]，对于 GAN 而言，标签平滑能避免判别器向生成器传递过大的梯度信号，从而防止算法坍塌至极端样本。经标签平滑处理的最优判别器从原来的 $D^*=\frac{p_{\text{data}}}{p_{\text{data}}+p_{\text{model}}}$ 变为 $D^*=\frac{\alpha p_{\text{data}}+\beta p_{\text{model}}}{p_{\text{data}}+p_{\text{model}}}$，然而，这带来了新问题：由于 p_{model} 出现在分子中，当 p_{data} 接近 0 而 p_{model} 较大时，p_{model} 所生成的样本便不会趋向真实样本。因此，Salimans 等[24]提出单边标签平滑，仅对真实样本的标签进行平滑，而生成样本的标签始终设为 0。然而，Sønderby 等[38]认为标签平滑在贝叶斯最优分类器不唯一时并不奏效，进而提出在样本中而不是标签中加入噪声。加入噪声后贝叶斯最优判别器是唯一的，由于训练分布变宽，判别器不易过拟合，从而可对判别器进行更多次训练。

Ioffe 等[67]提出的批归一化（BN, batch normalization）每次取一批而不是单独一个数据进行归一化，从而使数据变得更加集中，利用批归一化是 GAN 的常用训练技巧之一。例如，Springenberg 等[68]在判别器的所有层以及生成器除最后一层外的所有层中均采用了批归一化，使激活值产生边界，有效防止了生成器的模式震荡并改善了判别器的泛化性能；此外，批归一化在 DCGAN[28]中的使用明显改善了网络的优化。然而，批归一化难免使网络的输出高度依赖于与输入数据 x 位于同一批的其余数据，那么，当批内数据过于相似时，对生成器的输入进行批归一化会导致生成图片内出现强相关。针对这个问题，Salimans 等[24]提出“参照批归一化”，即取一批固定数据作为参照数据集，待处理的输入数据依据参照数据集的均值和标准差进行批归一化，而这种方法的缺陷在于归一化效果依赖于参照数据集的选取。鉴于此，进一步提出“虚拟批归一化”[24]，在对输入数据进行归一化时，将输入数据加入参照数据集中形成新的数据集——虚拟数据集，对此数据集进行批归一化处理，能有效避免生成数据与参照数据过于相似。由于虚拟批归一化需对 2 批数据进行前馈计算，开销较大，故只在生成器中使用。

4.3　模型评估

常见的生成式模型评价指标有平均对数似然、核密度估计和生成样本的视觉保真度[37,52]，这些方法分别适用于不同的生成式模型，而对 GAN 目前仍没有一个标准的定量评估指标。文献[22,43]通过帕尔森窗口法对 GAN 进行评估，帕尔森窗口法是一种非参数的密度函数估计方法，既不需利用样本分布的先验知识，也不需对样本分布作任何假设，是一种从样本出发研究数据分布的方法。然而当数据的维度很高时，即便大量的样本也不能保证通过帕尔森窗口估计可逼近模型的真实分布，样本维数越高，采用帕尔森窗口估计的效果越差。文献[45]提出了人工检视，通过 AMT（amazon mechanical turk）平台让人类标注者判断所见图片是真实样本还是生成样本。这种情况下，标注者充当着判别器的角色，而生成器是经过训练的 GAN，当标注者获得反馈信息时，判别的准确性会极大地提高。

人工检视的问题在于成本高昂和主观性强，为降低人工检视所需的实验成本，Salimans 等[24]提出一种与人工检视高度相关的替代方法将人工检视过程自动化，由于该评价方式是基于 Inception 模型[69]的，因而取名为 Inception 得分。基于 Inception 得分的强分类器能以较高置信度生成优质样本，然而仅当样本足够多时，Inception 得分才能有效评价生成样本的多样性。Che 等[55]进一步指出，假设一个生成器能生成很好的样本而这些样本都是同一种模式的，这种情况下，尽管生成器发生了模式坍塌，但它依然能够获得很高的 Inception 得分。因此，对于有标签的数据集，Che 等[55]提出一种“MODE 得分”来同时评价视觉保真度和样本多样性。

文献[28]提出基于分类性能对模型进行评估，这种方法最突出的问题是评估结果高度依赖于分类器的选择。例如，文献[28]中采用最近邻分类器，而欧氏距离对图像来说并不是一种很好的相似性度量。Im 等[59]提出一种针对 GAN 的评估方法 GAM（generative adversarial metric），令 2 组 GAN 互相

竞争、互为评委。尽管 GAM 是一个有效的评估标准，但是 GAM 要求相互比较的判别器在留存测试数据集的误差率不相上下，然而，对于进行并行对抗训练的模型，其判别器的泛化性会有明显提升，致使并行训练与非并行训练的模型之间错误率差别较大，从而无法使用 GAM 对模型进行评估，鉴于此，Im 等[59]进一步提出了 GAM Ⅱ，去除 GAM 的上述限制，仅度量这些判别器的平均（或者最差）错误率。

5 GAN 的应用

GAN 作为一种生成式模型，最直接的应用就是数据生成，即对真实数据进行建模并生成与真实数据分布一致的数据样本[14]，如图像、视频、语音、自然语言文本等。此外，GAN 还可用于机器学习中的半监督学习。本节将从计算机视觉、语言与语音、半监督学习以及其他领域对 GAN 的应用进行介绍。

5.1 计算机视觉领域

目前，GAN 应用最成功的领域是计算机视觉，包括图像和视频生成，如图像翻译[30,60,70~72]、图像超分辨率[26]、图像修复[73]、图像上色[74]、人脸图像编辑[75~80]以及视频生成[31,32]等。

文献[60,70,72]将 GAN 应用于图像翻译，例如，根据轮廓图像生成照片、根据白天图像生成对应夜景等，如图 4 所示[70]。Zhu 等[71]进一步将图像翻译拓展使多模态图像翻译，大大增加了生成图像的多样性，如图 5 所示。除了从二维图像到二维图像的翻译外，Gadelha 等[30]提出的 PrGAN 能够以一种完全无监督的训练方式将给定的一系列甚至是一张 2D 图像翻译为该物体的 3D 体素形状和深度信息。

Ledig 等[26]提出一个用于超分辨率的生成式对抗网络 SRGAN，该模型的目标函数由对抗损失函数和内容损失函数共同构成，其中，对抗损失函数通过训练判别器区分真实图片和由生成器进行超分辨重构的图片，从而能够学习自然图片的流形结构，通过峰值信噪比和结构相似性等指标对重建图像进行评估，结果表明 SRGAN 的效果比现有最先进的采用深度残差网络优化均方差更接近高分辨率原图。Pathak 等[73]将 cGAN[43]应用到图像修复，以图像缺失部分的周边像素为条件训练生成式模型，生成完整的修复图像，利用对抗思想训练判别器对真实样本和修复样本进行判断，经对抗训练后，生成器所生成的修复图像与遮挡区块周边是连贯的，而且是符合语义的，如图 6 所示[73]。人脸图像去遮挡是图像复原的延伸应用，Zhao 等[81]训练判别器区分真实无遮挡人脸图像和基于有遮挡图像而复原的人脸图像，能有效移除人脸图像中的遮挡物并用于人脸识别。文献[75~80]将 GAN 应用于人脸图片编辑。GAN 除了能够生成高质量的自然图像（例如手写字体[22]、卧室[28,82]、人眼[83]和人脸[84]等）外，还能生成抽象的艺术作品[85]。

(a) 根据轮廓图像生成照片　　(b) 根据白天图像生成对应夜景

图 4 图像翻译

(a) 输入　　(b) 多模态输出

图 5 多模态图像翻译

图 6　图像修复

Mathieu 等[31]最先提出将对抗训练应用于视频预测，即生成器根据前面一系列帧生成视频最后一帧，判别器对该帧进行判断。除最后一帧外的所有帧都是真实的图片，这样的好处是判别器能有效地利用时间维度的信息，同时也有助于使生成的帧与前面的所有帧保持一致。实验结果表明，通过对抗训练生成的帧比其他算法（基于 l_1 或 l_2 损失）更加清晰。由于该模型是完全可微的，因此，可在精调后用于其他任务。与光流预测进行结合或将下一帧预测相关应用中的光流算法替换为生成对抗训练，有望进一步改善应用效果。最近，Vondrick 等[32]利用 GAN 在视频生成中取得了突破性进展，能生成 32 帧（标准电影每秒 24 帧）分辨率为 64×64 的逼真视频，描绘的内容包括沙滩、高尔夫球场、火车站以及新生儿，20%的 AMT 标记员认为这些生成视频是真实视频。Vondrick 等[28]在 DCGAN 的基础上提出“双流架构”，双流分别是移动的前景流和静止的背景流，其中，前景流是一个时空卷积神经网络，而背景流则是一个空间卷积神经网络。前景流相比背景流多了一个时间维度，这是为了让前景移动而背景静止，双流之间相互独立，这一架构迫使生成器在前景对象移动时对静止背景进行渲染。此外，该研究工作还能从静态照片中生成多帧视频，首先，识别静态图片的对象，然后，生成 32 帧的视频，这些生成视频中对象的动作非常合乎常理。这种对动作的预测能力是机器未来融入人类生活的关键，因为这使机器能辨别什么动作于人于己都是没有伤害的。此前的模型都是逐帧创建场景的，这意味着信息被分成很多块，从而不可避免地带来较大误差，而该研究工作则是同时预测所有帧，当然，一次生成所有帧在使预测更加精确的同时也带来了复杂的计算，在长视频中此问题尤为突出。

5.2　语言与语音领域

相对于在计算机视觉领域的应用，GAN 在语言处理领域的报道较少。这是因为图像和视频数据的取值是连续的，可直接应用梯度下降对可微的生成器和判别器进行训练，而语言生成模型中的音节、字母和单词等都是离散值，这类离散输出的模型难以直接应用基于梯度的生成式对抗网络。为使模型适用于文本生成，TextGAN[41]采用一些技巧对离散变量进行处理，例如，采用光滑近似来逼近 LSTM 的离散输出，并在生成器训练过程中采用特征匹配技术[24]。由于 LSTM 的参数明显多于 CNN 的参数个数而更难训练，TextGAN 的判别器仅在生成器多次更新后才进行一次更新。Yu 等[42]提出的 SeqGAN 借鉴强化学习处理离散输出问题，将判别器输出的误差视为强化学习中的奖赏值，并将生成器的训练过程看作强化学习中的决策过程，应用于诗句、演讲文本以及音乐生成。Li 等[86]和 Kusner 等[87]分别将 GAN 应用于开放式对话文本生成和上下文无关语法（CFG, context-free grammar）。相比前述的从图像到图像的转换，从文本到图像的转换困难得多，因为以文本描述为条件的图像分布往往是高度多模态的，即符合同样文本描述的生成图像之间差别可能很大。另一方面，虽然从图像生成文本也面临着同样问题，但由于文本能按照一定语法规则分解，因此，从图像生成文本是一个比从文本生成图像更容易定义的预测问题。Reed 等[29]利用这个特点，通过 GAN 的生成器和判别器分别进行文本到图像、图像到文本的转换，二者经过对抗训练后能够生成以假乱真的图像，例如，根据文本“这只小鸟有着小小的鸟喙、胫骨和双足，蓝色的冠部和覆羽以及黑色的脸颊”生成图 7 所示的图片[29]。此外，通过对输入变量进行可解释的拆分，能改变图像的风格、角度和背景。当然，目前所合成的图像尺寸依然较小，该研究的下一步工作是尝试合成像素更高的图像和增加文本所描述的特征数量。

5.3　半监督学习

GAN 强大的表征能力使之能够生成与真实数据分布相一致的数据，因此，可用于解决训练数据不足时的学习问题，有效缓解基于深度学习的解决方案对训练数据量的需求。此外，尽管 GAN 作为一种无监督学习方法被提出，但可广泛应用于半监

图 7　根据文本描述生成图像

督学习[88]过程中无标签数据对模型的预训练[14]。GAN 的判别器 D 实际上是一个二分类的分类器，区分样本是真实样本还是生成样本。Springenberg[68]和 Salimans 等[24]结合文献[89]先对样本进行聚类，然后，通过计算有标签数据的预测类别分布和真实类别分布之间的交叉熵进行半监督学习，将朴素 GAN 的判别器从二元分类器扩展为多类别分类器，从而输出 Softmax 分类结果而不是图片来自真实样本的概率。此外，Odena[90]提出的 Semi-GAN 和 AC-GAN[27]也是GAN在多分类问题上的成功应用。文献[68]指出，将经过训练的判别器作为一个通用特征提取器用于多分类问题，只需结合少量标签信息便可达到令人满意的分类效果，例如，在 MNIST 数据集上对每一个类别仅用 10 个有标签样本就能达到 98.61%的分类准确率，这一结果已经非常接近使用全部（60 000 个）有标签样本所能得到的最佳结果（99.52%）。

5.4　其他领域

Santana 等[91]利用 GAN 辅助自动驾驶，首先，生成与真实交通场景图像分布一致的图像，然后，训练一个基于循环神经网络的转移模型来预测下一个交通场景。Wu 等[92]提出对抗神经机器翻译，将神经机器翻译（NMT, neural machine translation）作为 GAN 的生成器，采用策略梯度方法训练判别器，通过最小化人类翻译和神经机器翻译的差别生成高质量的翻译。Schlegl 等[93]将 GAN 用于医学图像的异常检测，通过学习健康数据集的特征能抽象出病变特征，例如，能够检测到测试样本中的视网膜积液，而这在训练样本集中并没有出现过。Hu 等[94]基于 GAN 生成具有对抗性的病毒代码样本，用于恶意软件检测的训练。Chidambaram 等[95]提出一个 GAN 的扩展模型，并将其作为风格迁移算子，用判别器对生成器进行正则化，并通过国际象棋实验证明该模型的有效性。

6　总结与展望

6.1　GAN 的优点

GAN 的最大优势在于不需对生成分布进行显式表达，既避免了传统生成式模型中计算复杂的马尔可夫链采样和推断，也没有复杂的变分下限，从而在大大降低训练难度的同时，提高了训练效率。GAN 提供了一个极具柔性的架构，可针对不同任务设计损失函数，增加了模型设计的自由度。依赖数据自然性解释的传统生成式模型难以适用于概率密度不可计算的情形，而 GAN 由于巧妙的内部对抗机制依然适用。此外，结合无监督的 GAN 训练和有监督的分类或回归任务，能产生一个简单而有效的半监督学习方法。

6.2　GAN 的缺点

尽管 GAN 解决了已有生成式模型存在的普遍

问题，但同时也带来了新的问题，最突出的是训练过程不稳定。GAN 的目标函数所优化的实质是真实分布与生成分布之间的 Jensen-Shannon 散度，当二者具有极小重叠甚至没有重叠时，Jensen-Shannon 散度是常数，从而导致优化梯度消失。而且，GAN 对多样性不足和准确性不足的惩罚并不平衡，导致生成器倾向生成重复但会被判别器认为真实的少数几种甚至一种样本，而不是丰富多样但有可能被判别器拒绝的样本，即模式坍塌（mode collapse）。此外，GAN 因其神经网络结构，可解释性较差，可微的设计使之仅适用于连续数据，从而导致自然语言等离散数据应用 GAN 的障碍。

6.3　GAN 的研究展望

1) 克服模式坍塌

模式坍塌是指 GAN 生成样本的模式总是集中在少数几个甚至单一模式上，这导致数据生成结果缺乏多样性[24]。因此，如何增加生成样本多样性是亟待研究的内容：通过模型组合（如并行或级联）对多个 GAN 的生成样本模式进行组合；利用推断机制保证样本空间与隐变量空间的对应性，从而保证生成器尽可能多地覆盖真实样本空间的所有模式；将有效的多样性度量加入损失函数中，从而指导模型训练等。

2) 标准的评价指标

对于生成模型这个研究领域来说，一个突出问题是缺乏公认的定量评价指标，对于 GAN 来说也是如此。生成样本的质量优劣仍依赖于主观判断，而对于常用的客观评价指标，如平均对数似然，核密度估计和生成样本的视觉保真度之间互不依赖且分别适用于不同类型的生成模型，即使对相同类型的生成模型，当应用对象不同时采用不同评估标准也可能导致差别较大的训练效果。因此，如何对 GAN 进行评估以及如何将 GAN 与其他类型的生成模型进行比较是亟待解决的问题。

3) 生成过程的可解释性

早期研究工作着眼于模型的输出而忽视了模型内部运作方式和产生输出的过程，解释 GAN 是如何在无监督方式下“理解”图像和视频等数据的研究工作至今鲜有报道。通过可视化手段解释模型内部运作机理能更好地指导模型训练，如通过反卷积操作将生成过程可视化，或激活某些中间层的特征以表征和推断更高层次的特征。相信深度学习的研究突破将为解决此问题提供新颖思路及技术手段。此外，通过增加从图像空间到隐变量空间的推断过程，从而将隐变量的属性分离，也是使生成过程可解释的有效手段。

4) 半监督学习

GAN 作为一种无监督学习方法被提出，可以对无标签数据进行特征学习。尽管实际应用中难以获得海量的标签数据，但获得少量标签数据往往是可能的，实际应用结果表明，少量标签数据即能大大提高 GAN 的表现。因此，如何充分利用有限的标签数据或对无标签数据自动添加标签，是 GAN 的理论研究中具有广阔研究前景的方向之一。

5) 与其他模型的融合

从应用实例可发现，融合能量函数的 GAN[25,39,96]在判别器的建模和训练方法选取上具备较高的柔性，除了通常所使用的二值分类器外，LeCun[97]所呈现的一系列基于能量的损失函数都能结合到 EBGAN 中，利用吉布斯分布可将能量转化为概率，因此，这个方向具有广阔的研究前景，后续研究可考虑结合 GAN 与那些能提供概率密度的深度生成器[33]，例如，自回归模型或采用可逆变换的模型，这种方法能提供更加稳定的训练、更加好的生成器以及更加广泛的应用（如自然语言处理）。其次，目前已有研究主要是 GAN 与 VAE、EBM 和 RL 的融合，而与其他深度模型（如 LSTM/BLSTM 和 RBM/DBN）或非深度模型融合的研究工作鲜有报道，是值得关注的研究方向之一。此外，强化学习与深度学习相结合在单一任务的处理上展现了夺目成效，因此，融合强化学习与 GAN，并用于跨任务学习将有力推进 AI 应用发展。

6) 拓展应用领域

在应用范围方面，尽管 GAN 比主流的基于最大似然训练的生成式模型能生成更加清晰和合理的图像，但仍存在生成图像噪声较多、对象不稳定以及训练图像的类别较为单一等亟待改善的问题；而在场景预测和视频生成方面，可尝试通过序列化和局部损失函数等方式提高训练样本尺度和生成视频时间维度，并通过最大化利用深度学习所取得的理论研究成果（如残差网络）降低视频生成的计算复杂度，从而将 GAN 拓展至基于视频生成的应用，如视频理解、动态场景标记和行为预测等。目前 GAN 的应用成果集中在图像和视频生成领域，然而 GAN 作为一种生成性深度学习框架，天然具备在自然语言处理和语音合成等方面的优良特性

和潜力，因此，GAN 的应用领域有着极大拓展空间。

7　结束语

本文概述了生成式对抗网络的研究背景并阐述了其基本原理，在此基础上围绕其重要架构、训练方法以及评价方式等方面对 GAN 的研究进展进行了论述，总结了当前研究存在的问题并指出未来的工作展望。

参考文献：

[1] LI Y, HE K, SUN J. R-fcn: object detection via region-based fully convolutional networks[C]//The Advances in Neural Information Processing Systems. 2016: 379-387.

[2] HONG S, ROH B, KIM K H, et al. PVANet: lightweight deep neural networks for real-time object detection[J]. arXiv: arXiv1611.08588, 2016.

[3] LI X, QIN T, YANG J, et al. LightRNN: memory and computation-efficient recurrent neural networks[J]. arXiv: arXiv1610.09893, 2016.

[4] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks[J]. arXiv: arXiv1609.03499, 2016.

[5] OORD A V D, DIELEMAN S, ZEN H, et al. WaveNet: a generative model for raw audio[J]. arXiv: arXiv1609.03499, 2016.

[6] BENGIO Y. Learning deep architectures for AI[J]. Foundations & Trends® in Machine Learning, 2009, 2(1):1-127.

[7] 王万良. 人工智能及其应用（第三版）[M]. 北京: 高等教育出版社, 2016.
WANG W L. Artificial intelligence: principles and applications (third edition)[M]. Beijing: Higher Education Press, 2016.

[8] 周昌令, 栾兴龙, 肖建国. 基于深度学习的域名查询行为向量空间嵌入[J]. 通信学报, 2016, 37(3): 165-174.
ZHOU C L，LUAN X L, XIAO J G. Vector space embedding of DNS query behaviors by deep learning[J]. Journal on Communications, 2016, 37(3): 165-174.

[9] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[10] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//The International Conference on Neural Information Processing Systems. 2012: 1097-1105.

[11] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[12] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4): 212-223.

[13] LIN M, CHEN Q, YAN S. Network in network[J]. arXiv: arXiv1312.4400, 2013.

[14] 王坤峰, 苟超, 段艳杰, 等. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报, 2017, 43(3): 321-332.
WANG K F, GOU C, DUAN Y J, et al. Generative adversarial networks: the state of the art and beyond[J]. ACTA Automatica Sinica, 2017, 43(3): 321-332.

[15] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic backpropagation and approximate inference in deep generative models[J]. Eprint Arxiv, 2014: 1278-1286.

[16] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 1989, 18(7): 1527-1554.

[17] SALAKHUTDINOV R, HINTON G. Deep boltzmann machines[J]. Journal of Machine Learning Research, 2009, 5(2): 1967-2006.

[18] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. arXiv: arXiv1312.6114, 2013.

[19] OORD A V D, KALCHBRENNER N, KAVUKCUOGLU K. Pixel recurrent neural networks[C]//The International Conference on Machine Learning, 2016: 1747-1756.

[20] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[21] KRIZHEVSKY A, HINTON G E. Learning multiple layers of features from tiny images[R]. University of Toronto, Technical Report, 2009.

[22] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//International Conference on Neural Information Processing Systems. 2014: 2672-2680.

[23] GOODFELLOW I. Generative adversarial networks[J]. arXiv: arXiv 1701.00160, 2017.

[24] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training gans[J]. arXiv: arXiv1606.03498,2016.

[25] ZHAO J, MATHIEU M, LECUN Y. Energy-based generative adversarial network[J]. arXiv: arXiv 1609.03126, 2016.

[26] LEDIG C, THEIS L, HUSZAR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[J]. arXiv: arXiv1609.04802, 2016.

[27] ODENA A, OLAH C, SHLENS J. Conditional image synthesis with auxiliary classifier GANs[J]. arXiv: arXiv1610.09585, 2016.

[28] ZHU W, MIAO J, QING L, et al. Unsupervised representation learning with deep convolutional generative adversarial networks. computer science[J]. arXiv: arXiv1511.06434, 2015.

[29] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[C]//International Conference on Machine Learning, 2016: 1060-1069.

[30] GADELHA M, MAJI S, WANG R. 3D shape induction from 2D views of multiple objects[J]. arXiv: arXiv1612. 05872, 2016.

[31] MATHIEU M, COUPRIE C, LECUN Y. Deep multi-scale video prediction beyond mean square error[J]. arXiv: arXiv1511.05440, 2015.

[32] VONDRICK C, PIRSIAVASH H, TORRALBA A. Generating videos with scene dynamics[C]//Conferrence on Neural Information Processing Systems. 2016: 613-621.

[33] FINN C, CHRISTIANO P, ABBEEL P, et al. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models[J]. arXiv: arXiv1611.03852, 2016.

[34] HO J, ERMON S. Generative adversarial imitation learning[C]//Advances in Neural Information Processing Systems. 2016: 4565-4573.

[35] PFAU D, VINYALS O. Connecting generative adversarial networks and actor-critic methods[J]. arXiv: arXiv1610.01945, 2016.

[36] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[C]// Computer Vision and Pattern Recognition.

2015: 3128-3137.
[37] GOODFELLOW I J. On distinguishability criteria for estimating generative models[J]. arXiv: arXiv1412.6515, 2014.
[38] SØNDERBY C K, CABALLERO J, THEIS L, et al. Amortised map inference for image super-resolution[J]. arXiv: arXiv1610.04490, 2016.
[39] KIM T, BENGIO Y. Deep directed generative models with energy-based probability estimation[J]. arXiv: arXiv1606.03439, 2016.
[40] 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.
[41] ZHANG Y Z, GAN Z, CARIN L. Generating text via adversarial training[C]//In Neural Information Processing Systems Workshop on Adversarial Training. 2016.
[42] YU L, ZHANG W, WANG J, et al. SeqGAN: sequence generative adversarial nets with policy gradient[J]. arXiv: arXiv1609.05473, 2016.
[43] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. Computer Science, 2014: 2672-2680.
[44] GAUTHIER J. Conditional generative adversarial nets for convolutional face generation[Z]. Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester, 2014(5): 2.
[45] DENTON E, CHINTALA S, SZLAM A, et al. Deep generative image models using a Laplacian pyramid of adversarial networks[C]// Conferrence on Neural Information Processing Systems. 2015: 1486-1494.
[46] HUANG X, LI Y, POURSAEED O, et al. Stacked generative adversarial networks[J]. arXiv: arXiv1612.04357, 2016.
[47] ZHANG H, XU T, LI H, et al. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks[J]. arXiv: arXiv1612.03242, 2016.
[48] 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.
[49] LAMB A, DUMOULIN V, COURVILLE A. Discriminative regularization for generative models[J]. arXiv: arXiv1602.03220, 2016.
[50] DONAHUE J, KRÄHENBÜHL P, DARRELL T. Adversarial feature learning[J]. arXiv: arXiv1605.09782, 2016.
[51] DUMOULIN V, BELGHAZI I, POOLE B, et al. Adversarially learned inference[J]. arXiv: arXiv1606.00704, 2016.
[52] THEIS L, OORD A, BETHGE M. A note on the evaluation of generative models[J]. arXiv: arXiv1511.01844, 2015.
[53] BROCK A, LIM T, RITCHIE JM, et al. Neural photo editing with introspective adversarial networks[J]. arXiv: arXiv1609.07093, 2016.
[54] LARSEN A B L, SØNDERBY S K, LAROCHELLE H, et al. Autoencoding beyond pixels using a learned similarity metric[J]. arXiv: arXiv1512.09300, 2015.
[55] CHE T, LI Y, JACOB A P, et al. Mode regularized generative adversarial networks[J]. arXiv: arXiv1612.02136, 2016.
[56] MAKHZANI A, SHLENS J, JAITLY N, et al. Adversarial autoencoders[J]. arXiv: arXiv1511.05644, 2015.
[57] WANG Y, ZHANG L, JOOST V D W. Ensembles of generative adversarial networks[J]. arXiv: arXiv1612.00991, 2016.
[58] LIU M Y, TUZEL O. Coupled generative adversarial networks[C]//Advances in Neural Information Processing Systems, 2016: 469-477.
[59] IM D J, MA H, KIM C D, et al. Generative adversarial parallelization[J]. arXiv: arXiv1612.04021, 2016.
[60] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[J]. arXiv: arXiv 1703.10593, 2017.
[61] LI C, XU K, ZHU J, et al. Triple generative adversarial nets[J]. arXiv: arXiv1703.02291, 2017.
[62] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[J]. arXiv: arXiv1701.07875, 2017.
[63] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein GANs[J]. arXiv: arXiv1704.00028, 2017.
[64] QI G J. Loss-sensitive generative adversarial networks on lipschitz densities[J]. arXiv: arXiv1701.06264, 2017.
[65] METZ L, POOLE B, PFAU D, et al. Unrolled generative adversarial networks[J]. arXiv: arXiv1611.02163, 2016.
[66] WARDE-FARLEY D and GOODFELLOW I. Adversarial perturbations of deep neural networks[C]//Perturbations, Optimization, and Statistics. 2016: 311.
[67] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning. 2015: 448-456.
[68] SPRINGENBERG J T. Unsupervised and semi-supervised learning with categorical generative adversarial networks[J]. arXiv: arXiv1511.06390, 2015.
[69] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//The IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.
[70] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[J]. arXiv: arXiv1611.07004, 2016.
[71] ZHU J Y, ZHANG R, PATHAK D, et al. Toward multimodal image-to-image translation[C]//Advances in Neural Information Processing Systems. 2017: 465-476.
[72] YI Z, ZHANG H, GONG PT. DualGAN: unsupervised dual learning for image-to-image translation[J]. arXiv: arXiv1704.02510, 2017.
[73] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: feature learning by inpainting[C]//The IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2536-2544.
[74] LI C, LIU H, CHEN C, et al. Alice: towards understanding adversarial learning for joint distribution matching[C]//Advances in Neural Information Processing Systems. 2017: 5501-5509.
[75] PERARNAU G, VAN DE WEIJER J, RADUCANU B, et al. Invertible conditional GANs for image editing[J]. arXiv: arXiv1611.06355, 2016.
[76] CRESWELL A, BHARATH A A. Inverting the generator of a generative adversarial network[J]. arXiv: arXiv1611.05644, 2016.
[77] ZHOU S, XIAO T, YANG Y, et al. GeneGAN: learning object transfiguration and attribute subspace from unpaired data[J]. arXiv: arXiv1705.04932, 2017.
[78] KIM T, CHA M, KIM H, et al. Learning to discover cross-domain relations with generative adversarial networks[J]. arXiv: arXiv1703.05192, 2017.
[79] WANG C, WANG C, XU C, et al. Tag disentangled generative adversarial network for object image re-rendering[C]//The Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017: 2901-2907.

[80] ANTIPOV G, BACCOUCHE M, DUGELAY JL. Face aging with conditional generative adversarial networks[J]. arXiv: arXiv1702.01983, 2017.

[81] ZHAO F, FENG J, ZHAO J, et al. Robust LSTM-autoencoders for face de-occlusion in the wild[J]. arXiv: arXiv1612.08534, 2016.

[82] YU F, SEFF A, ZHANG Y, et al. Lsun: construction of a large-scale image dataset using deep learning with humans in the loop[J]. arXiv: arXiv1506.03365, 2015.

[83] SHRIVASTAVA A, PFISTER T, TUZEL O, et al. Learning from simulated and unsupervised images through adversarial training[J]. arXiv: arXiv1612.07828, 2016.

[84] LIU Z, LUO P, WANG X, et al. Deep learning face attributes in the wild[C]//The IEEE International Conference on Computer Vision. 2015: 3730-3738.

[85] TAN WR, CHAN CS, AGUIRRE H, et al. ArtGAN: artwork synthesis with conditional categorial GANs[J]. arXiv: arXiv1702.03410, 2017.

[86] LI J, MONROE W, SHI T, et al. Adversarial learning for neural dialogue generation[J]. arXiv: arXiv1701.06547, 2017.

[87] KUSNER M J, HERNÁNDEZLOBATO J M. GANS for sequences of discrete elements with the gumbel-softmax distribution[J]. arXiv: arXiv1611.04051, 2016.

[88] DENTON E, GROSS S, FERGUS R. Semi-supervised learning with context-conditional generative adversarial networks[J]. arXiv: arXiv1611.06430, 2016.

[89] SUTSKEVER I, JOZEFOWICZ R, GREGOR K, et al. Towards principled unsupervised learning[J]. arXiv: arXiv1511.06440, 2015.

[90] ODENA A. Semi-supervised learning with generative adversarial networks[J]. arXiv: arXiv1606.01583, 2016.

[91] SANTANA E, HOTZ G. Learning a driving simulator[J]. arXiv: arXiv1608.01230, 2016.

[92] WU L, XIA Y, ZHAO L, et al. Adversarial neural machine translation[J]. arXiv: arXiv1704.06933, 2017.

[93] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery[J]. arXiv: arXiv1703.05921, 2017.

[94] HU W W, TAN Y. Generating adversarial malware examples for black-box attacks based on GAN[J]. arXiv: arXiv1702.05983, 2017.

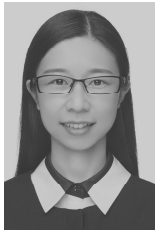
[95] CHIDAMBARAM M, QI Y J. Style transfer generative adversarial networks: learning to play chess differently[J]. arXiv: arXiv1702.06762, 2017.

[96] ZHAI S, CHENG Y, FERIS R, et al. Generative adversarial networks as variational training of energy based models[J]. arXiv: arXiv1611.01799, 2016.

[97] LECUN Y, CHOPRA S, HADSELL R, et al. A tutorial on energy-based learning[M]. Predicting Structured Data: MIT Press. 2006.

[作者简介]

王万良（1957-），男，江苏高邮人，博士，浙江工业大学教授，主要研究方向为人工智能、机器自动化、网络控制。



李卓蓉（1986-），女，广西桂林人，浙江工业大学博士生，主要研究方向为人工智能、深度学习。