文章编号:2095-6134(2022)04-0551-10

基于互信息约束的生成对抗网络分类模型*

胡兵兵[1,2,3†],唐华[1,2,3],吴幼龙[1]

(1 上海科技大学信息科学与技术学院, 上海 201210; 2 中国科学院上海微系统与信息技术研究所, 上海 200050; 3 中国科学院大学, 北京 100049)

(2020 年 4 月 27 日收稿; 2020 年 8 月 10 日收修改稿)

Hu B B, Tang H, Wu Y L. Classification models based on generative adversarial networks with mutual information regularization[J]. Journal of University of Chinese Academy of Sciences, 2022, 39(4): 551-560. DOI: 10.7523/j.ucas.2020.0037.

摘 要 传统的机器学习方法需要大量的含标注数据集来训练模型,并且容易引发过拟合,而生成对抗网络可以无监督地进行训练。此外,互信息约束能够让模型生成指定类别的数据,可用于扩充数据集。提出 InfoCatGAN 和 C-InfoGAN 两种模型,前者在 CatGAN 的基础上增加了互信息约束,使得生成的图片更加逼真;后者使用 InfoGAN 模型中的辅助网络 Q 做分类,能够在生成高质量图片的同时,达到较好的分类准确率。二者均能通过隐变量控制生成图片的类别,这对数据增强具有一定意义。另外,在加入少量标签信息之后,模型的准确率能有所提升。

关键词 生成对抗网络;无监督学习;半监督学习;互信息

中图分类号:U283.4 **文献标志码**:A **DOI**:10.7523/j.ucas.2020.0037

Classification models based on generative adversarial networks with mutual information regularization

HU Bingbing[1, 2, 3], TANG Hua[1, 2, 3], WU Youlong[1]

(1 *School of Information Science and Technology, ShanghaiTech University, Shanghai* 201210, *China*;
2 *Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Science, Shanghai* 200050, *China*;
3 *University of Chinese Academy of Science, Beijing* 100049, *China*)

Abstract This paper studies classification models based on generative adversarial networks with mutual information regularization. Traditional machine learning methods rely on a large number of labeled datasets, which are scarce in practice, to train the model and can easily overfit to spurious correlations in the data; while generating adversarial networks can be trained in an unsupervised manner. In addition, mutual information constraint allows the model to generate data of a specified category, which can be used to expand the data set. This paper proposes the InfoCatGAN and C-InfoGAN classification models. The former adds the mutual information term to CatGAN model in order to generate images of higher visual fidelity; the latter uses the InfoGAN model for classification, which can ensure the quality of the generated images and provide a mentionable classification accuracy. Additionally, both two models can control the category of generated images

* 国家自然科学基金(61901267)和上海市浦江人才计划(18PJ1408500)资助

† 通信作者,E-mail: hubb@shanghaitech.edu.cn

through latent variables, which has a certain significance for data augmentation. Moreover, after adding a small amount of label information, the accuracy of the model can be improved.

Keywords GANs; unsupervised learning; semi-supervised learning; mutual information

分类问题一直是机器学习领域经久不衰的话题。目前有监督分类方法已经相对成熟,其中不少方法在某些数据集上已经达到非常高的准确率。近年来,深度学习社区的活跃研究已经催生出许多使用深度神经网络去做分类的成功案例[1-3]。这些方法均需要经过以下 3 个过程:数据压缩,特征提取和模型预测。这些过程往往依赖于大量的数据标注,但现实生活中标注好的数据十分稀缺。因此,无监督和半监督学习顺势兴起。在无监督学习中,数据的分布 $p(\boldsymbol{x})$ 与条件分布 $p(y|\boldsymbol{x})$ 有一定的联系,其中 $\boldsymbol{x}$ 表示数据,$y \in \{1, \cdots, K\} \triangleq [K]$ 表示未知的数据标签。不同于有监督学习,无监督学习中标签信息 $p(y)$ 无法直接获得,因此只能利用数据的结构特征推断训练样本的标签。作为无监督学习家族的重要一员,无监督分类通常建模为聚类问题,并且已经具有一些经典的方法:K-means、Gaussian mixture model、density estimation,这些方法均是针对数据分布进行建模。此外,一些判别式方法比如 maximum margin clustering (MMC)[4]、regularized information maximization (RIM)[5],则是将数据划分到某个类别,无须估计数据分布。尽管判别式方法更为直接,但是它们容易受一些虚假相关性的影响而产生过拟合[6]。当与深度神经网络这种拟合能力很强的模型相结合的时候,过拟合现象尤为显著。随着深度学习领域崛起[7-9],越来越多的学者使用深度模型研究无监督或半监督学习。这些方法通常是训练一个生成式模型,比如波尔兹曼机[10-11]、前馈神经网络[12-13]以及自编码器[14-15],通过重建输入样本学习数据特征,刻画数据分布。这类方法避免了因直接划分数据而产生的过拟合问题,但是在重建训练样本的过程中没有额外的约束,所以会保留原始数据的所有信息,这和训练分类器的目标相背①。

生成对抗网络(generative adversarial network, GAN)[16]是最近非常热门的研究课题之一。相较于纯生成式模型,GAN 训练生成器的同时,还训练一个判别器,通过二者对抗使得生成器学习到真实数据分布并生成较为逼真的数据。InfoGAN[17]通过最大化隐变量和生成图片之前的互信息,能够学习到数据的局部特征,从而调控生成图片的样式。CatGAN[6]利用生成对抗网络模型,将生成式方法和判别式方法相结合,在 MNIST[18]和 CIFAR-10[19]上均取得了十分可观的分类准确度。Li 等[20]指出,良好的分类准确率和良好的生成效果互不相容,进而提出具有 3 个模块的 GAN 模型。EnhancedTGAN[21]在 TripleGAN 的基础上额外增加一个分类器,并重新设计目标函数,达到了更好的效果。由于增加了分类专用网络,所以基于 TripleGAN 的模型无法进行无监督学习。

本文将 InfoGAN 和 CatGAN 相结合,提出 InfoCatGAN 模型。CatGAN 只关注分类精度,仅仅将判别器作为提取特征的工具,以致生成的图片不够逼真。InfoGAN 可以指定生成图片的特征,对分类有指导作用。两者结合,InfoCatGAN 能够通过超参数 λ 的设置,实现分类准确率和生成数据逼真度的折中,即当 λ 较小时,分类准确度较高,但生成图片质量较差;当 λ 较高时,生成图片质量较高,分类准确率较低。为了简化模型,同时避免超参数不确定性所带来的影响,本文基于 InfoGAN 提出 Classifier InfoGAN(C-InfoGAN),该模型可以在牺牲少量的分类准确率的情况下,获得更高的生成质量。二者均可以对生成图片的类别进行调控,此外 C-InfoGAN 能够对图片局部特性进行调整,如改变字体粗细、倾斜度等(见图 1),这对指定特征的数据补足有较大意义。与 TripleGAN 和 EnhancedTGAN 相比,本文提出的基于互信息约束的模型支持无监督分类,且能够调节生成图片的局部特征,与此同时还具有更强的可解释性。

1 生成对抗网络

生成对抗网络由 Goodfellow 等[16]在 2014 年提出,在该模型中,他们训练一个生成器 G —给定噪声生成虚假数据,和一个判别器 D —给定输

① 在训练分类器时,通常只希望保留和分类目标相关的信息,从而使得模型对其他不重要的信息更加鲁棒。

(a) 隐变量c_1调控生成数字的粗细　(b) 隐变量c_2调控生成数字的角度　(c) 隐变量c_1调控生成图片的亮度

图 1　隐变量对生成图片的调控

Fig. 1　The impact of variations of latent codes

入判别其真假。训练过程可以类比为两个玩家博弈:判别器读取一个数据希望能够分别真假,而生成器希望生成以假乱真的数据从而让判别器判定为真。

在实际应用中,生成器和判别器通常实现为可微的深度神经网络。设 $\mathcal{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$ 为真实数据集,其中 $\boldsymbol{x}_i=(x_{i1},\cdots,x_{in})\in\mathbf{R}^n$,$\boldsymbol{z}=(z_1,\cdots,z_m)\in\mathbf{R}^m$ 为按分布 p_z 采样的隐空间噪声,其中 N 为样本个数,n 为单个样本的维度,m 为噪声维度。可以将生成器描述为 $G:\mathbf{R}^m\mapsto\mathbf{R}^n$,将判别器描述为 $D:\mathbf{R}^n\mapsto(0,1)$,其中 $D(\boldsymbol{x})$ 表示 $\boldsymbol{x}$ 来自真实数据分布的概率。对于给定的 G,训练 D 使得对于真实数据 $\boldsymbol{x}$,$D(\boldsymbol{x})$ 接近于 1;对于虚假数据 $\tilde{\boldsymbol{x}}=G(\boldsymbol{z})$,$D(\tilde{\boldsymbol{x}})$ 接近于 0。当 D 训练至最优,固定 D 训练 G 以降低判别器对于虚假数据的区分精度。当生成器对应的概率分布 p_g 与真实数据的分布 p_{data} 完美契合的时候,D 无法分别真假,对于所有输入都输出 0.5 的概率。综上所述,GAN 的目标函数如下

$$\min_G \max_D V_{\text{GAN}}(D,G)=\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D(\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{z}\sim p_z}[\log(1-D(G(\boldsymbol{z})))]. \tag{1}$$

1.1　InfoGAN

原始的 GAN 没有对输入噪声 $\boldsymbol{z}$ 做任何限制,这使得生成器在生成虚假数据时没有指向性,以至于生成的数据高度耦合,数据特征难以解释。InfoGAN 将噪声分解为两部分:一部分依然是无结构的噪声 $\boldsymbol{z}\sim p_z$,为模型提供足够的容量;另一部分作为隐变量 $\boldsymbol{c}\sim p_c$,用于学习数据的特殊语义。InfoGAN 最大的创新是引入互信息约束,通过最大化隐变量 $\boldsymbol{c}$ 与生成数据 $\tilde{\boldsymbol{x}}=G(\boldsymbol{z},\boldsymbol{c})$ 之间的互信息

$$I(\boldsymbol{c};\tilde{\boldsymbol{x}})=H(\boldsymbol{c})-H(\boldsymbol{c}\mid\tilde{\boldsymbol{x}}), \tag{2}$$

将隐变量绑定到数据的某些特征,$H(\cdot)$ 表示 Shannon 熵。在信息论中,互信息 $I(X;Y)$ 用来衡量在观测到随机变量 X 之后,随机变量 Y 的不确定性的减少量。互信息越大说明两个变量之间的关系越紧密,反之互信息为 0,则说明变量间相互独立。InfoGAN 将互信息作为正则项加入其目标函数

$$\begin{aligned}\min_{G,Q}\max_D \quad & V_{\text{InfoGAN}}(G,D,Q,\lambda)=\\ & V_{\text{GAN}}(G,D)-\lambda I(\boldsymbol{c};\tilde{\boldsymbol{x}}),\\ I(\boldsymbol{c};\tilde{\boldsymbol{x}}) \quad & =\mathbb{E}_{p(\boldsymbol{c},\tilde{\boldsymbol{x}})}\left[\log\frac{Q(\boldsymbol{c}\mid\tilde{\boldsymbol{x}})}{p_c(\boldsymbol{c})}\right]+\\ & \mathbb{E}_{p_g}[D_{\text{KL}}(p(\boldsymbol{c}\mid\tilde{\boldsymbol{x}})\parallel Q(\boldsymbol{c}\mid\tilde{\boldsymbol{x}}))]\\ & \geqslant\mathbb{E}_{p(\boldsymbol{c},\tilde{\boldsymbol{x}})}[\log Q(\boldsymbol{c}\mid\tilde{\boldsymbol{x}})]+H(\boldsymbol{c})\\ & \triangleq L_I(Q(\boldsymbol{c}\mid\tilde{\boldsymbol{x}})),\end{aligned} \tag{3}$$

其中:Q 是辅助网络用于估计后验概率 $P(\boldsymbol{c}\mid\boldsymbol{x})$,$\lambda$ 是正则化系数,D_{KL} 表示 Kullback-Leibler 距离用于衡量两个概率分布间的差异。而由于在实现中互信息难以计算,故采用其变分下界 L_I 代替[22],其中 $H(\boldsymbol{c})$ 在训练过程中视为常量,在实现中可以略去,模型结构见图 2。

1.2　CatGAN

与原始的 GAN 不同,CatGAN 基于合理的假设重新设计了目标函数,并将判别器扩展为多类别分类器。考虑如下无监督分类问题,假设真实数据集 $\mathcal{X}$ 有 K 个类别,CatGAN 训练一个判别器 $D:\mathbf{R}^n\mapsto(0,1)^K$,给定一个数据 $\boldsymbol{x}$,$D(\boldsymbol{x})$ 给出该数

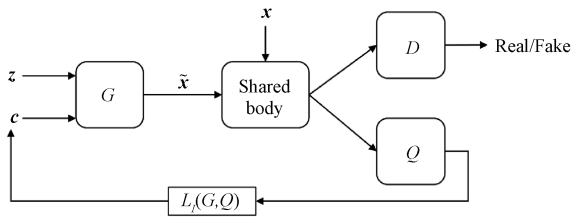


图 2 InfoGAN 结构示意

Fig. 2 The architechture of InfoGAN

据属于每个类别的概率 $p(y \mid \boldsymbol{x})$，且 $\sum_{k=1}^{K} p(y=k \mid \boldsymbol{x})=1$。CatGAN 的判别器损失函数 L_D^{cat} 和生成器损失函数 L_G^{cat} 形式如下：

$$\begin{aligned} L_D^{\text{cat}} &= -H_x(p(y)) + \mathbb{E}_{x\sim p_{\text{data}}}[H(p(y\mid \boldsymbol{x}))] - \\ &\quad \mathbb{E}_{\tilde{x}\sim p_g}[H(p(y\mid \tilde{\boldsymbol{x}}))], \\ L_G^{\text{cat}} &= -H_G(p(y)) + \mathbb{E}_{\tilde{x}\sim p_g}[H(p(y\mid \tilde{\boldsymbol{x}}))]. \end{aligned} \tag{4}$$

式中各项的计算方式请参考文献[6]，模型结构见图 3。

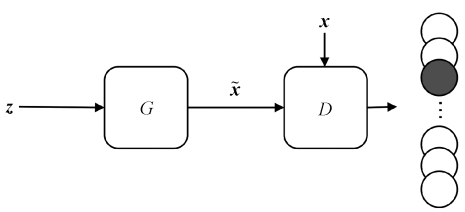


图 3 CatGAN 结构示意

Fig. 3 The architecture of CatGAN

2 InfoCatGAN

2.1 无监督分类方法

在训训练概率分类模型的过程中，通过优化条件熵可以将分类边界调整到更自然的位置（数据分散区域）[23]，因此 CatGAN 使用条件熵作为判别器判断真假数据的依据。但是，使用熵作为目标函数的一个缺点是没有类别指向性（K 个类别中任意一个都可以使 $p(y \mid \boldsymbol{x})$ 呈单峰分布）。对于一个分类器，理想的情况是对于给定输入 $\boldsymbol{x}$，有且仅有一个 $k \in [K]$，$p(y=k \mid \boldsymbol{x})$ 能够到达最大，而对于任意 $k' \neq k$，$p(y=k' \mid \boldsymbol{x})$ 均很小。然而问题在于训练数据集没有标注，每个数据样本对应的标签无从获得。

对于上述问题，本文从 InfoGAN 中获得启发，提出 InfoCatGAN 模型。InfoGAN 将输入噪声划分为 $\boldsymbol{z}$ 和 $\boldsymbol{c}$，实际上是对隐空间的结构进行人为划分。一部分提供模型的容量，使得模型具有足够的自由度去学习数据的细节（高度耦合的特征）；一部分提供隐变量，用于在学习过程中绑定到数据的显著特征（如：MNIST 中的数字类别、笔画粗细、角度）。模型的核心思想如下：通过在隐空间构造一维隐变量 c，在训练过程中将生成数据的类别标签与之绑定，使得可以通过 c 来控制生成数据的类别。CatGAN 对 GAN 的扩展主要在于改变了判别器的输出结构：为所有真实数据分配一个类别标签而对于虚假数据则保持一个不确定的状态。类似地，生成器应该致力于生成某个具体类别的数据而不是仅仅生成足够逼真的图片。

下面给出 InfoCatGAN 的损失函数：设 $\boldsymbol{x} \in \mathcal{X}$ 为一个真实数据样本，$\tilde{\boldsymbol{x}} = G(\boldsymbol{z}, c)$ 为一个生成数据，其中 $\boldsymbol{z} \sim p_z$ 为噪声，$c \sim p_c$ 为隐变量。为了简单起见，这里只考虑 c 为一维离散随机变量，p_c 为离散均匀分布。生成器 $G = G(\boldsymbol{z}, c; \theta_G)$ 和判别器 $D = D(\boldsymbol{x}; \theta_D)$ 均为可微深度神经网络，其中 θ_G，θ_D 分别为生成器和判别器的参数①。通过在 D 网络的最后一层做 Softmax 变换，可以直接将 $D(\boldsymbol{x})$ 作为条件概率 $p(y \mid \boldsymbol{x})$ 的估计。注意到式(4)可以重写为

$$\begin{aligned} L_D^{\text{cat}} &= -I(\boldsymbol{x}; y) - \mathbb{E}_{\tilde{x}\sim p_g}[H(p(y\mid \tilde{\boldsymbol{x}}))], \\ L_G^{\text{cat}} &= -I(\tilde{\boldsymbol{x}}; y), \end{aligned} \tag{5}$$

其中：$\boldsymbol{x} \sim p_{\text{data}}$，$\tilde{\boldsymbol{x}} \sim p_g$ 分别表示真实数据和虚假数据对应的随机变量，y 表示未知标签对应的随机变量。可以看到，CatGAN 其实是在优化数据与标签之间的互信息。互信息是常用的变量间相关性的衡量标准，所以用它作为生成器损失函数的正则项，由此得到 InfoCatGAN 的损失函数如下

$$\begin{aligned} L_D &= L_D^{\text{cat}}, \\ L_G &= L_G^{\text{cat}} - \lambda_1 I(c; \tilde{\boldsymbol{x}}), \end{aligned} \tag{6}$$

其中 λ_1 为正则系数，可知当 $\lambda_1 = 0$ 时，InfoCatGAN 退化为 CatGAN，模型结构见图 4。图中 D 的输出为 $P(y \mid \cdot)$。在训练生成器的时候，将判别器的输出 $P(y \mid \tilde{\boldsymbol{x}})$ 和隐变量 c 通过某种度量 $d(\cdot, \cdot)$ 建立联系使得条件概率的峰值与 c 的

① 为简便起见，在无歧义的情况下通常省略网络参数。

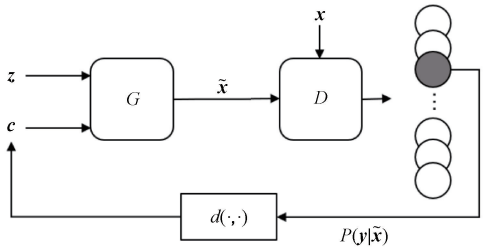


图 4　InfoCatGAN 模型结构

Fig. 4　The architecture of InfoCatGAN

取值对应。参考式(3), $I(c;\tilde{\boldsymbol{x}})$ 可以放缩为 $\mathbb{E}_{p(c,\tilde{x})}[\log p(c\mid\tilde{x})]$, 在实现中通常使用交叉熵

$$CE[\boldsymbol{c},p(\boldsymbol{c}\mid\tilde{\boldsymbol{x}})]=-\sum_{i=1}^{K}c_i\log p(c=c_i\mid\tilde{\boldsymbol{x}}) \tag{7}$$

来优化此项,这里的 $\boldsymbol{c}\in\mathbf{R}^K$ 是隐变量 $\boldsymbol{c}$ 经过 one-hot 编码之后的向量, $p(c\mid\tilde{\boldsymbol{x}})$ 可以用 $D(\tilde{x})$ 来近似。

2.2　半监督分类方法

作为 CatGAN 的扩展,InfoCatGAN 可以很自然地适用于半监督的情况。假设 $\mathcal{X}^L=\{\boldsymbol{x}_i^L\}_{i=1}^m$ 为 m 个有标签的样本, $\boldsymbol{y}_i^L\in\mathbf{R}^K$ 为经过 one-hot 编码之后的标签向量。对于有标签的样本, $D(\boldsymbol{x}^L)$ 的分布信息可以明确获得,所以可以通过计算 $\boldsymbol{y}^L$ 和 $p(\boldsymbol{y}\mid\boldsymbol{x}^L)$ 之间的交叉熵

$$CE[\boldsymbol{y}^L,p(\boldsymbol{y}\mid\boldsymbol{x}^L)]=-\sum_{i=1}^{K}y_i\log p(y=y_i\mid\boldsymbol{x}^L), \tag{8}$$

辅助判别器做出更精确的判断。半监督版本的 InfoCatGAN 损失函数如下

$$L_D^L=L_D+\mathbb{E}_{(\boldsymbol{x}^L,\boldsymbol{y}^L)\sim\mathcal{X}^L}[CE[\boldsymbol{y}^L,p(\boldsymbol{y}\mid\boldsymbol{x}^L)]], \tag{9}$$

生成器的损失函数同式(6): $L_G^L=L_G$.

3　C-InfoGAN

InfoCatGAN 无法同时获得较高的准确率和生成质量,只能通过正则系数 λ_1 实现二者的性能折中。考虑到 InfoGAN 模型中的隐变量可以较好地绑定到数据的类别特征,而且生成的图片较为逼真,本文提出 C-InfoGAN 模型,旨在保证生成质量的前提下,尽可能提高分类准确率。

3.1　无监督分类方法

InfoGAN 能够做到无监督地学习数据类别的特征,并且可以通过隐变量控制生成数据的类别,这为分类任务提供了基础。InfoGAN 中使用一个辅助的 Q 网络来估计后验概率 $P(\boldsymbol{c}\mid\boldsymbol{x})$, 如果隐变量 $\boldsymbol{c}=(c,c_1,c_2,\cdots)$ 中的 c 能够学习到数据的类别特征,则可以利用 $Q(c\mid\boldsymbol{x})$ 作为一个概率分类器。具体来说,本文在 InfoGAN 的目标函数上添加一个正则项 $L(c,\hat{c})$, 其中 $\hat{c}=Q(c\mid\tilde{\boldsymbol{x}})\in\mathbf{R}^K$ 是 Q 网络的输出。称这个分类模型为 C-InfoGAN(CIG),其目标函数如下

$$\min_{G,Q}\max_{D}V_{\mathrm{CIG}}(G,D,Q,\lambda_1,\lambda_2)=V_{\mathrm{InfoGAN}}(G,D,Q,\lambda_1)+\lambda_2L(c,Q(c\mid\tilde{\boldsymbol{x}})), \tag{10}$$

其中: λ_2 是正则化系数, $L(c,\hat{c})=L(c,Q(c\mid\tilde{\boldsymbol{x}}))$ 在实现中一般采用交叉熵,参见式(8),模型结构见图 5。无监督情况下,生成数据 $\tilde{\boldsymbol{x}}$ 和真实数据 $\boldsymbol{x}$ 参与训练,通过和 D 共享部分结构, Q 网络可以将 GAN 模型学习到的特征加以利用,实现分类任务。

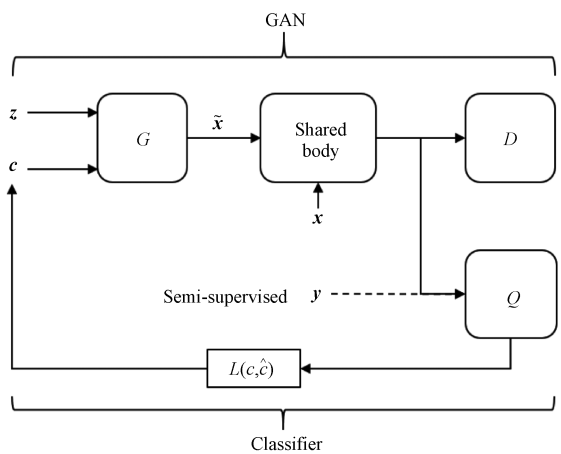


图 5　C-InfoGAN 模型结构

Fig. 5　The architecture of C-InfoGAN

3.2　半监督分类方法

当拥有少量标签信息时,C-InfoGAN 可以利用这些标签进一步提升分类准确率和生成效果。同时将隐变量 c 直接绑定到真实的标签,实现精准调控。针对少量标注信息,文献[24]提出将隐变量 c 进一步分解为无监督部分 c_{us}, 负责捕捉大量无标注数据的潜在特征;和有监督部分 c_{ss}, 负责捕捉已有标签 y。同时他们设置了两组隐变量对应的先验分布,以及对应的辅助网络 Q_{us} 和 Q_{ss},使用隐变量 c_{ss} 和辅助网络 Q_{ss} 专门处理那部分有标注信息。本文直接将标签信息加入 Q 网络,先用真实数据和标签训练,接着用生成数据和虚假标签(即隐变量 c)来训练。这样可以使真实标签的信息流入隐变量 c 中,即用真实标签指导 c 绑定到正确的类别特征。经过实践发现,使用上述方法也能达到同样的效果,而且模型更为简单。

使用和 2.2 节中类似的方法,给出半监督 C-InfoGAN(ss-CIG)的目标函数如下

$$\min_{G,Q}\max_{D} V_{\text{ss-CIG}}(G,D,Q,\lambda_1,\lambda_2) = V_{\text{CIG}}(G,D,Q,\lambda_1,\lambda_2) + \mathbb{E}_{(\boldsymbol{x}^L,\boldsymbol{y}^L)\sim\mathcal{X}^L}[CE[\boldsymbol{y}^L,Q(\boldsymbol{y}\mid\boldsymbol{x}^L)]], \tag{11}$$

模型结构参见图 5。在半监督情况下,一部分真实标签 y 会直接被 Q 网络利用,以得到更好的效果。优化 Q 网络的输出 $\hat{c}$ 和隐变量 c 构成的损失函数 $L(c,\hat{c})$ 来增加 Q 的分类准确率。

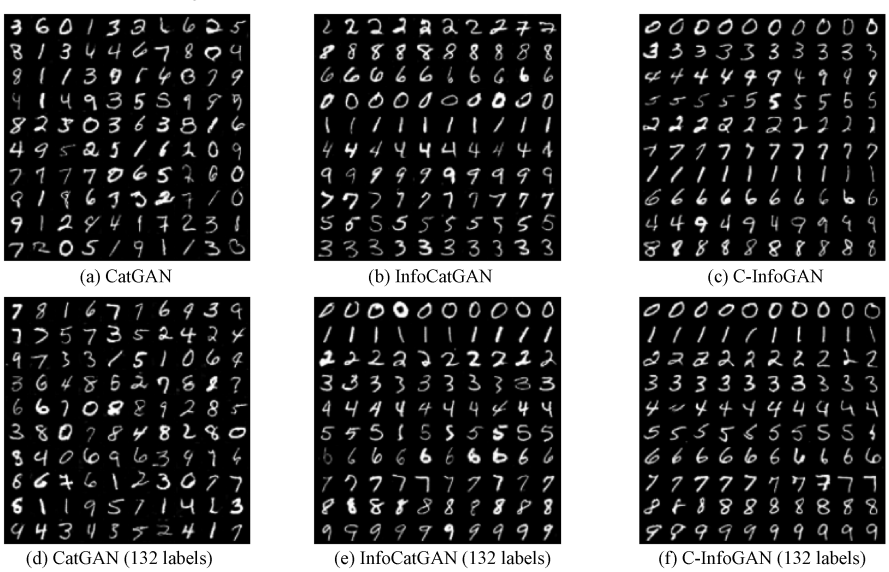
4 实验结果与分析

在所有实验中,本文考察两个指标:分类准确率和图片生成质量。对于分类准确率,计算模型预测值并不像一般分类器那样直接。隐变量虽然可以学习到数据类别的特征,但是其取值并不和真实标签正确对应(例如 $c=1$ 可能对应生成真实标签 2 的数据),因此无法直接使用隐变量的取值作为模型的预测值,必须将隐变量的取值与真实标签之间做一个映射。对于这个问题,本文采取与文献[6]相同的做法:在测试集上选取一批样本计算模型在这批数据上的预测值。模型为每一个数据分配一个虚假标签 $l_i, i\in[K]$,然后将预测值和真实标签对比:将虚假标签落入最多的真实标签的取值作为该虚假标签的取值。比如在所有 10 个被分类为虚假标签 l_3 的样本中,有 9 个真实标签为类别‘7’,则将虚假标签 l_3 映射到真实类别‘7’。对于图片生成质量,本文采用 Fréchet inception distance (FID)[25]进行衡量①,相较于 Inception Score[26]只考虑生成数据,FID 还利用了真实数据,因此更能反映生成数据和真实数据的差异。FID 越小代表生成的图片和真实图片越接近,生成质量越好。

4.1 MNIST

MNIST 是常用的衡量生成式模型的数据集,它包含了 60 000 张手写数字图片,并且附有类别标签。

图 6(a)和 6(b)是在无监督情况下 CatGAN 和 InfoCatGAN 的生成效果,其中每一行对应隐变量 c 的一个取值,从 0 到 9。可以看到,InfoCatGAN 的生成效果略高于 CatGAN,并且每一行基本是一种数字类别,对应隐变量的不同取值。半监督情况下有类似的结果,不同的是在少量标签信息的辅助下,InfoCatGAN 可以将隐变量 c 和真实标签正确绑定,例如,$c=1$ 对应生成数字‘1’,见图 6(e)。CatGAN 生成的图片质量较差,原因在于其目标函数是为了分类而设计的。生成器的作用只是为了判别器能够更加鲁棒,如 2.1 节所述,从式(4)中可以看到,G 的目标函数只有条件熵,无法针对性地生成图片,从而会降低生成图片的质量。而 InfoCatGAN 由于增加了隐变量 c,并在训练过程中有意识地将生成数据的类别与之绑定,所以生成的图片质量较好。



(a) CatGAN (b) InfoCatGAN (c) C-InfoGAN

(d) CatGAN (132 labels) (e) InfoCatGAN (132 labels) (f) C-InfoGAN (132 labels)

图 6 模型在 MNIST 上的生成效果

Fig. 6 Generated images on MNIST

① FID 一般用于彩色图片,而 MNIST 数据集是单通道的灰度图片,本文将单通道复制 3 份形成 RGB 彩色图片计算其 FID 值。

图 6(c)和 6(f)给出了无监督和半监督情况下 C-InfoGAN 的生成结果。从图中可以看出无监督情况下,模型已经达到了很好的生成效果,隐变量 c 基本可以控制生成图片的类别,但是仍有部分类别未能精确控制(图 6(c));在半监督情况下,隐变量达到了精确的绑定,每一行对应生成一种类别的数字,而且顺序和真实标签是对应的。另外从图 1 可以看出,C-InfoGAN 模型不仅可以生成指定类别的图片,并且可以通过额外的隐变量调节图片局部特征,如手写数字的粗细、角度等,这对指定特征的数据补足具有一定意义。

表 1 给出了无监督和半监督情况下的分类准确率①和 FID。从表中看出,InfoCatGAN 的分类准确率虽略低于 CatGAN,但在图像生成质量上 InfoCatGAN 均一致高于 CatGAN,这说明增加互信息约束可以提高图像的生成质量。相较于 CatGAN 模型,C-InfoGAN 模型可以获得更高的准确率和生成质量,而且隐变量的绑定效果也更好。而在无监督情况下,C-InfoGAN 在保证生成质量的前提下,仍然能够达到 87.59% 的分类准确率。这是因为 InfoGAN 模型使用的是一个辅助网络 Q 来做类别绑定和分类任务,训练过程中并没有判别器做过多约束,所以无论如何调整分类网络或更改分类约束,也不会对生成效果产生很大影响。这使得模型可以进一步利用生成的图片和标签扩充数据集,以达到更进一步的性能提升。

表 1　分类准确率对比

Table 1　Model accuracy

Model	分类准确率/%/FID			
	MNIST		FashionMNIST	
	0 label	132 labels	0 label	100 labels
CatGAN	**89.18**/37.83	**98.22**/21.36	67.38/69.40	71.07/67.88
InfoCatGAN	87.76/12.61	96.89/**5.99**	**70.61**/41.17	75.24/44.36
C-InfoGAN	87.59/**3.07**	95.85/9.44	69.57/**11.88**	**75.40/15.99**

表 2 给出了正则系数 λ_1 的不同取值对于半监督 InfoCatGAN 的影响。从表中可以看出,当系数较小时,分类准确率较高,但生成图片的质量非常差;当系数较大时,生成的图片效果很好,但分类准确率有所降低。通过调节参数 λ_1,可以实现生成效果和分类准确率之间的折中。实验使用的默认值是 $\lambda_1 = 0.9$,当 λ_1 减小时,生成图片的质量开始下降,同时分类准确率也会相应增加;当 $\lambda_1 = 0$ 时,InfoCatGAN 退化为 CatGAN。

表 2　正则系数对于 InfoCatGAN 的性能影响

Table 2　The effect of regularizer to InfoCatGAN

λ_1	准确率/%	FID
0.0	98.22	21.36
0.2	97.55	18.13
0.4	97.14	17.38
0.9	96.89	5.99

4.2　FashionMNIST

FashionMNIST[27] 是一个类似 MNIST 的数据集,二者拥有同样的图像大小,同样的类别数目。但是相对于 MNIST,FashionMNIST 拥有更复杂的图像结构,以及更难获得非常高的分类准确率,所以对模型更具有检验性。

表 1 给出了模型在 FashionMNIST 的数值结果。从表中可以看出,在无监督条件下,InfoCatGAN 较 CatGAN 在分类准确率和生成质量上均有所提升,C-InfoGAN 在一定程度上兼顾二者,不仅生成质量最优,而且具有相对较高的分类准确率,此外其模型复杂度也较低。在半监督条件下,C-InfoGAN 在两个方面均体现出优势,分类准确率达到 75.40%,FID 为 15.99,生成效果见图 7(f)。

图 7 给出了所有模型的生成结果,其中每一行对应隐变量 c 的一个取值。值得一提的是,加入互信息约束的半监督版本(图 7(e)、7(f))的模型从上往下每一行都对应同一个类别,并且顺序和训练数据的真实标签正确对应。这说明隐变量正确绑定到类别特征,并且可以精准调控生成图片的类别。

4.3　收敛速度分析

本文提出的两个模型在原理上都属于正则化生成对抗网络,与原先的两个模型 CatGAN 和 InfoGAN 相比,增加的计算复杂度较小。由于

① 表中有关 CatGAN 的数据来自本文复现的结果,与文献[5]有所差距。

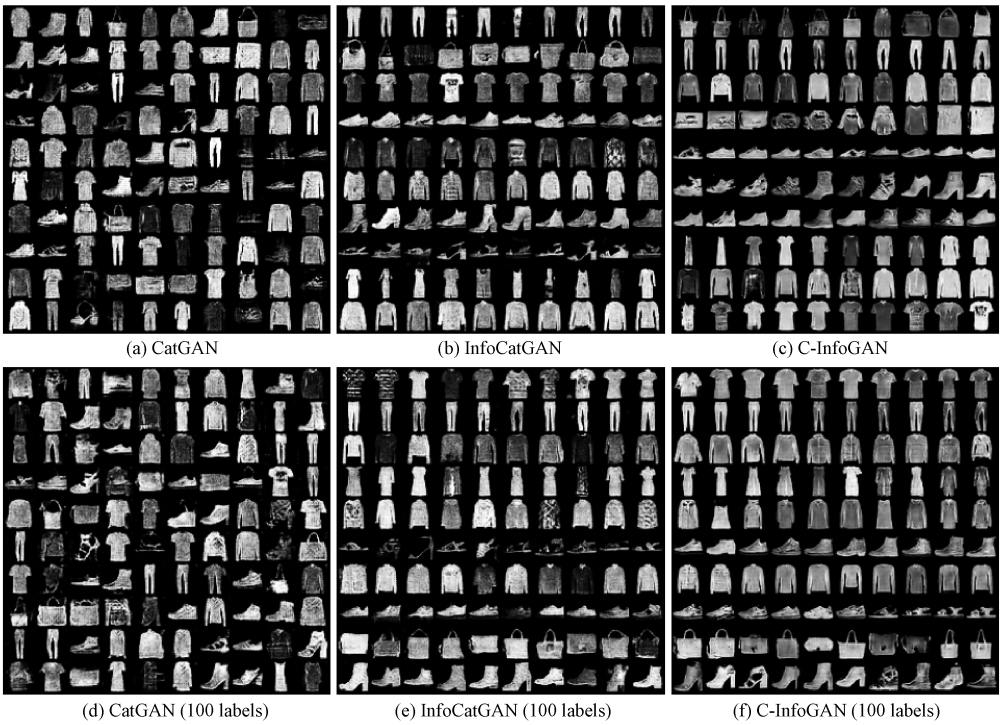


图 7　模型在 FashionMNIST 上的生成效果

Fig. 7　Generated images on FashionMNIST

GAN 的训练方式特殊,训练的过程是生成器和判别器的对抗,因此目前没有一个统一的评判收敛性的标准。针对 InfoCatGAN 和 C-InfoGAN 两种模型,本文分别用条件熵损失(即判别器输出的概率分布对应的熵)以及互信息损失(实际采用交叉熵估计,详见 3.1 节)作为模型收敛的佐证,见图 8。

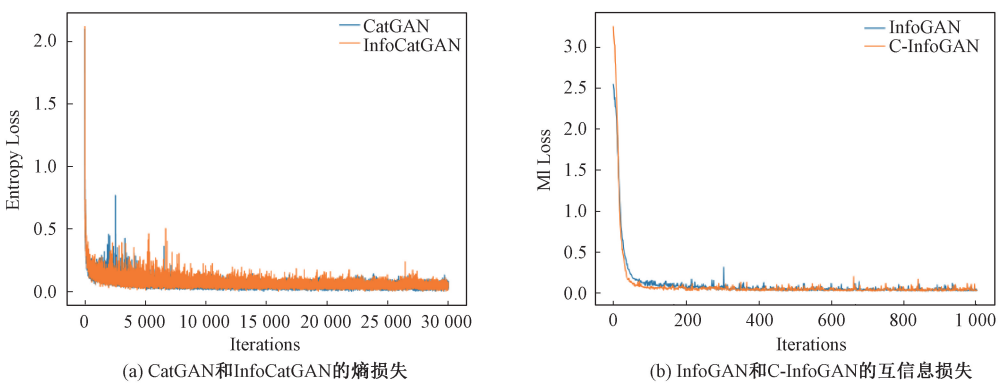


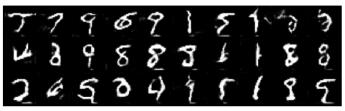
图 8　模型在 MNIST 上的收敛速度

Fig. 8　Convergence speed on MNIST

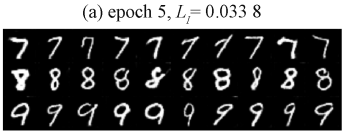
4.4　模型可解释性

从以上结果可以看出,加入互信息约束可以给模型带来许多增益,其中最为显著的是生成质量的提升。图 9 给出了 MNIST 数据集下 InfoCatGAN 在训练的不同阶段对应的生成效果。其中,L_I 是公式(3)中的互信息下界,可以看到,随着隐变量 c 和生成数据 $\tilde{\boldsymbol{x}}$ 的互信息增加,生成的图片开始具有绑定效果,并且生成的图像越来越好。

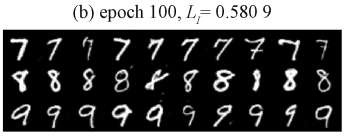
对于 InfoCatGAN 在生成质量上的增益,本文参考文献[24]从互信息的角度给出一些直观解释。由式(5)、式(6)可知,生成器 G 的优化目标是最大化 $I(\tilde{\boldsymbol{x}};y)$ 和 $I(\tilde{\boldsymbol{x}};c)$,这可以令虚假标签 c 和真实标签 y 对应;而半监督条件下判别器 D 的目标是最大化 $I(\boldsymbol{x};y)$,以及最小化真实分布和预测分布之间的交叉熵 $-\mathrm{E}_{p(y|\boldsymbol{x})}\log p(c\mid\boldsymbol{x})$,进而使得真实标签 y 中的信息流入隐变量 c。事实上,在半监督版本的 InfoCatGAN 中,本文就是采用 c 的



(a) epoch 5, L_I= 0.033 8



(b) epoch 100, L_I= 0.580 9



(c) epoch 300, L_I= 1.533 2

图 9　InfoCatGAN 不同阶段的生成效果

Fig. 9　Generated samples of different phases

后验概率 $p(c\mid\cdot)$ 作为模型的预测值,从实验结果也可以看出虚假标签和真实标签是正确对应的(图 6(e))。换句话说,模型的优化目标变为 $I(c;\boldsymbol{x})$ 和 $I(c;\tilde{\boldsymbol{x}})$。现假设:

1) $\boldsymbol{x}\leftarrow c\rightarrow\tilde{\boldsymbol{x}}$:其中 $\rightarrow$ 表示依赖性。这个假设来源于图像是由多个独立隐变量的相互作用生成的,实际中这些隐变量还包括噪声 $\boldsymbol{z}$,以及其他因素。为简单起见,这里假设只和隐变量 c 有关。

2)初始 $I(\boldsymbol{x};\tilde{\boldsymbol{x}})=0$:开始的时候,虚假数据和真实数据无关。

3) $H(c)$ 为常量:假设 c 的先验分布在训练过程中没有改变。

由文献[24], $H(c)$ 可分解为

$$H(c)=I(c;\boldsymbol{x})+I(c;\tilde{\boldsymbol{x}})+H(c\mid\boldsymbol{x},\tilde{\boldsymbol{x}})-I(\boldsymbol{x};\tilde{\boldsymbol{x}})+I(\boldsymbol{x};\tilde{\boldsymbol{x}}\mid c).$$

由假设 1),

$$H(c)=I(c;\boldsymbol{x})+I(c;\tilde{\boldsymbol{x}})+H(c\mid\boldsymbol{x},\tilde{\boldsymbol{x}})-I(\boldsymbol{x};\tilde{\boldsymbol{x}}).$$

由假设 3),

$$0=\Delta I(c;\boldsymbol{x})+\Delta I(c;\tilde{\boldsymbol{x}})+\Delta H(c\mid\boldsymbol{x},\tilde{\boldsymbol{x}})-\Delta I(\boldsymbol{x};\tilde{\boldsymbol{x}}),$$

其中 Δ 表示变化量。进一步得到以下两种情况:

$$\Delta I(c;\boldsymbol{x})+\Delta I(c;\tilde{\boldsymbol{x}})\geqslant-\Delta H(c\mid\boldsymbol{x},\tilde{\boldsymbol{x}})\Rightarrow\Delta I(\boldsymbol{x};\tilde{\boldsymbol{x}})\geqslant 0,$$

$$\Delta I(c;\boldsymbol{x})+\Delta I(c;\tilde{\boldsymbol{x}})<-\Delta H(c\mid\boldsymbol{x},\tilde{\boldsymbol{x}})\Rightarrow\Delta I(\boldsymbol{x};\tilde{\boldsymbol{x}})<0.$$

注意到模型的训练目标是最大化两个互信息,所以上式左边一定为正值。由假设 2),初始时 $I(\boldsymbol{x};\tilde{\boldsymbol{x}})=0$,如果第 2 种情况发生,则会导致 $I(\boldsymbol{x};\tilde{\boldsymbol{x}})$ 变为负值,而互信息是非负的,因此第 2 种情况不会发生。于是,增加 $I(c;\boldsymbol{x})$ 和 $I(c;\tilde{\boldsymbol{x}})$ 会导致 $I(\boldsymbol{x};\tilde{\boldsymbol{x}})$ 增加,这也就说明生成图片与真实图片更为接近,即模型的生成质量较好。

5　结论

本文首先提出 InfoCatGAN 模型,它通过优化隐变量和生成数据之间的互信息,能够获得更高的生成质量,同时可以通过调节正则系数实现生成质量和分类准确率的折中。为了同时兼顾二者,又提出 C-InfoGAN 模型。实验结果表明,InfoCatGAN 可以在牺牲少量准确率的条件下提高图像的生成质量,而 C-InfoGAN 在一定程度上既可以生成高质量的图像,也能够达到可观的分类准确率,并且还可以调控生成图片的局部特征。未来的研究工作包括互信息项对于提高生成器生成效果的理论分析,如何进一步提高模型的分类准确率,以及针对复杂数据集的模型优化。

参考文献

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90. DOI: 10.1145/3065386.

[2] Taigman Y, Yang M, Ranzato M, et al. DeepFace: closing the gap to human-level performance in face verification[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. June 23-28, 2014, Columbus, OH, USA. IEEE, 2014: 1701-1708. DOI: 10.1109/CVPR.2014.220.

[3] 江璐, 赵彤, 吴敏. 基于深度卷积神经网络的指纹纹型分类算法[J]. 中国科学院大学学报, 2016, 33(6): 808-814. DOI: 10.7523/j.issn.2095-6134.2016.06.013.

[4] Xu L L, Neufeld J, Larson B, et al. Maximum margin clustering[C]//Advances in Neural Information Processing Systems 17(NIPS 2004). 2005: 1537-1544.

[5] Krause A, Perona P, Gomes R G. Discriminative clustering by regularized information maximization[C]//Advances in Neural Information Processing Systems 23 (NIPS 2010). 2010: 775-783.

[6] Springenberg J T. Unsupervised and semi-supervised learning with categorical generative adversarial networks[EB/OL]. ArXiv preprint, arXiv: 1511.06390. (2016-04-30)[2020-04-15]. https://arxiv.org/abs/1511.06390.

[7] 宋旭鸣, 沈逸飞, 石远明. 基于深度学习的智能移动边缘网络缓存[J]. 中国科学院大学学报, 2020, 37(1): 128-135. DOI: 10.7523/j.issn.2095-6134.2020.01.015.

[8] 田玮, 朱廷劭. 基于深度学习的微博用户自杀风险预测[J]. 中国科学院大学学报, 2018, 35(1): 131-136. DOI: 10.7523/j.issn.2095-6134.2018.01.018.

[9] 杨建斌, 张卫强, 刘加. 深度神经网络自适应中基于身份认证向量的归一化方法[J]. 中国科学院大学学报, 2017, 34(5): 633-639. DOI: 10.7523/j.issn.2095-6134.

2017. 05. 014.

[10] Salakhutdinov R, Hinton G. Deep boltzmann machines[J]. Journal of Machine Learning Research, 2009,5:448-455.

[11] Goodfellow I, Mirza M, Courville A, et al. Multi-prediction deep Boltzmann machines [C] // Advances in Neural Information Processing Systems 26(NIPS 2013). 2013:548-556.

[12] Bengio Y, Laufer E, Alain G, et al. Deep generative stochastic networks trainable by backprop[C]//Proceedings of the 31st International Conference on Machine Learning, PMLR, 2014,32(2):226-234.

[13] Kingma D P, Mohamed S, Rezende D J, et al. Semi-supervised learning with deep generative models [C] // Advances in Neural Information Processing Systems 27(NIPS 2014). 2014:3581-3589.

[14] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313 (5786):504-507. DOI:10. 1126/science. 1127647.

[15] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]// Proceedings of the 25th International Conference on Machine Learning (ICML). 2008: 1096-1103. DOI: 10. 1145/1390156. 1390294.

[16] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] // Advances in Neural Information Processing Systems 27(NIPS 2014). 2014:2672-2680.

[17] 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 29(NIPS 2016). 2016:2172-2180.

[18] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition [J]. Neural Computation, 1989, 1(4):541-551. DOI: 10. 1162/neco. 1989. 1. 4. 541.

[19] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images [R]. Computer Science Department, University of Toronto, Tech. 2009.

[20] Li C X, Xu T, Zhu J, et al. Triple generative adversarial nets [C]//Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017:4088-4098.

[21] Wu S, Deng G C, Li J C, et al. Enhancing TripleGAN for semi-supervised conditional instance synthesis and classification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA. IEEE, 2019: 10083-10092. DOI: 10. 1109/CVPR. 2019. 01033.

[22] Poole B, Ozair S, van der Oord A, et al. On variational bounds of mutual information[C]//Proceedings of the 36th International Conference on Machine Learning, PMLR. 2019, 97:5171-5180.

[23] Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization[C]//Advances in Neural Information Processing Systems 17(NIPS 2004). 2005:529-536.

[24] Spurr A, Aksan E, Hilliges O. Guiding InfoGAN with semi-supervision[C] // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. ECML PKDD, 2017:119134. DOI:10. 1007/978-3-319-71249-9_8.

[25] 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 30(NIPS 2017). 2017:6626-6637.

[26] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs [C] // Advances in Neural Information Processing Systems 29 (NIPS). 2016: 2234-2242.

[27] Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms[EB/OL]. ArXiv Preprint, arXiv: 1708. 07747. (2017-09-15) [2020-04-15]. https://arxiv. org/abs/1708. 07747.