天津科技大学学报
Journal of Tianjin University of Science & Technology
ISSN 1672-6510,CN 12-1355/N

《天津科技大学学报》网络首发论文

题目：基于监督对比正则化项的信息蒸馏生成对抗网络
作者：陈亚瑞，王晓捷，李晴，刘浩天，史艳翠，赵婷婷
DOI：10.13364/j.issn.1672-6510.20240023
收稿日期：2024-02-18
网络首发日期：2024-09-30
引用格式：陈亚瑞，王晓捷，李晴，刘浩天，史艳翠，赵婷婷．基于监督对比正则化项的信息蒸馏生成对抗网络[J/OL]．天津科技大学学报.
https://doi.org/10.13364/j.issn.1672-6510.20240023

网络首发时间：2024-09-30 14:13:55
网络首发地址：https://link.cnki.net/urlid/12.1355.N.20240930.0940.003

DOI：10.13364/j.issn.1672-6510.20240023

基于监督对比正则化项的信息蒸馏生成对抗网络

陈亚瑞，王晓捷，李　晴，刘浩天，史艳翠，赵婷婷
(天津科技大学人工智能学院，天津 300457)

摘　要：传统生成对抗网络主要通过最大化解耦表示和生成数据之间的互信息来学习解耦表示，较少分析解耦表示各维度之间的独立性。本文提出一种基于监督对比正则化项的信息蒸馏生成对抗网络(information distillation generative adversarial net based on supervised contrastive regularization, IDGAN-SC)。首先，IDGAN-SC 模型利用 β-VAE 模型学习解耦表示空间，约束解耦表示空间和生成模型之间具有强相关性；然后，通过最大化解耦隐向量和生成数据之间的互信息对模型进行解耦表示学习，并进一步利用监督对比正则化项的对比分类信息增强解耦隐变量各维度之间的独立性。在 dSprites、MNIST、CelebA 数据集上，分别从定性和定量的角度设计了对比实验，实验结果表明相比已有的生成对抗网络的解耦性能，IDGAN-SC 模型具有较强的解耦能力并具有明显的解耦效果。

关键词：生成对抗网络；变分自编码器；解耦表示；对比学习
中图分类号：TP18　　**文献标志码：**A

Information Distillation Generative Adversarial Net Based on Supervised Contrastive Regularization

CHEN Yarui，WANG Xiaojie，LI Qing，LIU Haotian，SHI Yancui，ZHAO Tingting
(College of Artificial Intelligence, Tianjin University of Science & Technology, Tianjin 300457, China)

Abstract：For generative adversarial net (GAN), traditional methods mainly disentangle the latent presentation based on the mutual information between the disentangled representation and the generated data. However, those methods rarely analyze the independence among dimensions of the latent vector. This paper proposes an information distillation generative adversarial net based on supervised contrastive regularization (IDGAN-SC). The IDGAN-SC model firstly learns disentangled representation space through training β-VAE, which enforces strong correlation between the disentangled representation space and the generative model. Then, the model constructs the disentangled structure by maximizing the mutual information between the disentangled latent vectors and the generated data. Furthermore, the model utilizes the contrastive classification information of the supervised contrastive regularization to enhance the independence between dimensions of the latent vectors. This paper performs quantitative and qualitative experiments on the dSprites, MNIST, and CelebA datasets. Experiments show that IDGAN-SC significantly outperforms current disentanglement methods based on the disentanglement metrics.

Key words：generative adversarial net; variational auto-encoder; disentangled representation; contrastive learning

从表示学习的角度来看，真实数据是由物理含义可解释的生成因子通过复杂的方式相互耦合产生

收稿日期：2024-02-18；**修回日期：**2024-05-08
基金项目：国家自然科学基金项目(61976156)
作者简介：陈亚瑞(1982—)，女，河北邢台人，教授，yrchen@tust.edu.cn

的。例如，在 MNIST 数据集中，一张数字图像由宽度、风格、粗细、角度 4 个生成因子相互耦合而成。近年来，深度模型能从数据中自适应地进行学习，在语音识别[1–2]、自然语言处理[3–4]、人脸识别[5–6]、目标检测[7–8]等领域取得了突破性进展。这些模型更多依赖于复杂的网络结构，人们无法探知模型从数据中学习到哪些知识，从而难以对模型的决策进行解释和预判。为了解决这些问题，通过对深度模型添加额外的约束，使网络能够学习到可解释的数据表示显得尤为重要[9]。解耦表示学习旨在对数据中的生成因子进行建模，使某生成因子的变化只引起数据在某一特征上的变化，而其他特征保持不变[10]。例如，人脸的特征有表情、发型等，如果模型成功学习到人脸的解耦表示，可以通过改变解耦隐向量某一维度来改变人脸的表情，而该人脸的其他特征保持不变。解耦表示学习有助于完成监督学习、强化学习、迁移学习等任务。因此，学习数据的解耦表示对人工智能具有重要影响。

无监督方式的解耦表示学习网络主要基于变分自编码(variational auto-encoder, VAE)[11]模型和生成对抗网络(generative adversarial net, GAN)[12]模型。VAE 模型从极大似然的角度对真实数据进行表征建模，有效结合了深层神经网络和变分贝叶斯方法，是经典的深度生成模型之一。β-VAE 模型通过在 VAE 模型优化目标的 KL 项上施加大于 1 的惩罚系数 β，鼓励模型学习独立的生成因子，是探索无监督解耦表示学习的开始[13]。相较于 VAE 模型，GAN 模型通过对抗训练的方式训练模型的生成能力，不需要显示设计数据的分布模型，更注重数据的生成过程。InfoGAN 模型以 GAN 模型为基础，通过最大化解耦隐向量和生成数据之间的互信息对模型进行解耦表示学习，有效提升了图像生成质量[14]。InfoGAN-CR 模型通过改变解耦隐向量某一维度的值(其他维度取值不变)生成一组样本，增加判别模型来预测这组样本由解耦隐向量的哪一维度变化产生的，使用交叉熵损失函数优化判别模型的预测能力[15]。每行表示遍历解耦隐向量时图像的变化如图 1 所示，模型通过改变解耦隐向量第一维度生成样本 $\boldsymbol{x}$、$\boldsymbol{x}_1^1$、$\boldsymbol{x}_2^1$，改变解耦隐向量第二维度生成样本 $\boldsymbol{x}$、$\boldsymbol{x}_1^2$、$\boldsymbol{x}_2^2$。第一行中椭圆的大小发生明显变化，属于大小类别，第二行椭圆的旋转角度发生明显变化，属于旋转类别。在模型解耦性能较差时，生成样本的大小和旋转会同时发生变化，这会影响到 InfoGAN-CR 模型中类别标签的准确性。判别模型将该类别标签作为监督信息，利用交叉熵损失函数优化模型参数。错误的标签信息会降低判别模型的预测能力。因此，样本标签的质量对 InfoGAN-CR 模型的解耦性能至关重要。

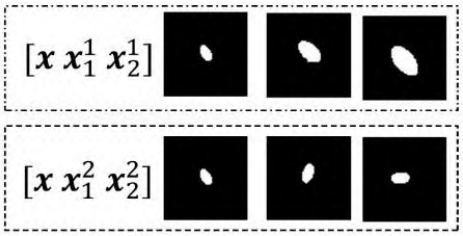

图 1　每行表示遍历解耦隐向量时图像的变化

Fig. 1　Each row shows how the image changes when traversing disentangled latent vectors

对比学习是一种自监督学习方法，通过比较样本之间的差异，相似样本对应的表示向量距离更近，不相似样本对应的表示向量距离更远。对比学习被广泛应用于图像识别、文本分类等任务[16–18]。监督对比学习方法进一步利用标签信息，将与锚点样本属于同一类别的样本归为正样本，将其他类别的样本归为负样本，有效约束相同类别样本的表示向量在特征空间更紧密[19]。

本文受到对比学习思想的启发，构建监督对比正则化项，实现解耦隐向量的强解耦效果，提出一种基于监督对比正则化项的信息蒸馏生成对抗网络(information distillation generative adversarial net based on supervised contrastive regularization, IDGAN-SC)。模型采用两阶段训练的方法，首先，模型利用 β-VAE 模型学习解耦表示空间，并蒸馏该表示空间作为生成对抗网络的解耦隐向量输入空间；然后，模型利用监督对比正则化项增强互信息式生成对抗网络的解耦表示能力。在 dSprites 数据集、MNIST 数据集、CelebA 数据集上设计对比实验，从定性和定量上分析 IDGAN-SC 模型的解耦能力和解耦效果。

1　相关工作

无监督解耦表示学习方法主要基于 VAE 和 GAN 等深度生成模型。基于 VAE 的模型主要通过增加归纳偏好，促使模型学习数据内部生成因子的解耦表示。基于 GAN 的模型通过最大化解耦隐向量和生成数据之间的互信息对模型进行解耦表示学习，能够生成更高质量的图像，但是相较于基于 VAE 的模型，这类模型的解耦性能较差。

VAE 模型有效结合了变分贝叶斯方法和深层神经网络。基于 VAE 模型的解耦表示学习策略主要通过设计各种正则化约束，使隐向量各维度满足统计独立性，包括对 KL 项看作整体进行约束[13, 20–21]、对基于隐向量聚合后验概率分布的 KL 距离进行约束[22–24]、对隐向量聚合后验概率分布的 TC 项进行约束[25–26]等。β-VAE 模型通过在 VAE 模型优化目标的 KL 项上增加大于 1 的超参数 β，使模型学习独立的生成因子，β-VAE 模型损失函数为

$$\mathcal{L}(\theta,\phi)=E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})-\beta D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}\mid\boldsymbol{x})\,\|\,p(\boldsymbol{z})) \tag{1}$$

式中：右边第一项表示重构误差；第二项表示概率分布 $q_\phi(\boldsymbol{z}\mid\boldsymbol{x})$ 与 $p(\boldsymbol{z})$ 之间的KL散度距离，用来约束模型的隐空间；$q_\phi(\boldsymbol{z}\mid\boldsymbol{x})$ 是变分推理模型，参数为 ϕ；$p_\theta(\boldsymbol{x}\mid\boldsymbol{z})$ 是生成模型，参数为 θ；当 $\beta=1$ 时，为原始的VAE模型。

与 VAE 模型从变分贝叶斯方法的角度对真实数据的生成分布进行建模不同，GAN 模型采用对抗训练的方式训练模型的生成能力。生成模型以随机噪声 $\boldsymbol{z}$ 作为输入生成样本，判别模型辨别样本是来自真实数据还是生成模型。由于 GAN 模型对噪声 $\boldsymbol{z}$ 没有施加任何限制，所以 $\boldsymbol{z}$ 的每一维度没有明显的语义特征。InfoGAN 将噪声分解为两部分：一部分是噪声 $\boldsymbol{z}$ 为模型提供信息能够生成样本，另一部分是代表数据生成因子的解耦隐向量 $\boldsymbol{c}$。InfoGAN 通过最大化隐向量 $\boldsymbol{c}$ 和生成数据 $\mathrm{G}(\boldsymbol{z},\boldsymbol{c})$ 之间的互信息，使隐向量 $\boldsymbol{c}$ 具有解耦效果。InfoGAN 缺乏对解耦隐向量的有效推理机制，针对生成因子复杂性高的场景解耦效果较差。为了解决该问题，IB-GAN 模型和 ID-GAN 模型在 InfoGAN 模型的框架中，采用 VAE 模型和 GAN 模型相结合的方式，引入具有解耦表示学习能力的网络学习解耦隐向量[27–28]。ID-GAN 模型利用基于 VAE 的方法学习数据的解耦表示，然后将该解耦表示和随机噪声作为生成模型的输入生成样本，并最大化生成样本和解耦表示间的互信息。该方法不仅提升 InfoGAN 模型的解耦性能，而且能够生成高质量的图像。

此外，InfoGAN 模型缺少对隐向量 $\boldsymbol{c}$ 的独立性约束。完全耦合的隐向量可能使互信息 $I(\boldsymbol{c},\mathrm{G}(\boldsymbol{z},\boldsymbol{c}))$ 的值无限大。这表明，最大化互信息项不一定能够使模型获得良好的解耦性能[15]。InfoGAN-CR 模型在 InfoGAN 模型的框架中设计正则化项，通过各维度独立解耦进一步提升模型解耦性能。InfoGAN-CR 模型通过改变解耦隐向量某一维度的值(其他维度不变)生成一对样本，增加判别模型预测该样本由解耦隐向量的哪一维度变化产生的，使用交叉熵损失函数训练判别模型的判别能力。交叉熵损失函数为

$$\begin{aligned}\min_{\mathrm{G,Q}}\max_{\mathrm{D}}\mathcal{L}_{\text{InfoGAN-CR}}(\mathrm{D,G,Q})=\\ \mathcal{L}_{\text{Adv}}(\mathrm{D,G})-\lambda\mathcal{L}_{\text{info}}(\mathrm{G,Q})-\alpha I(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{d})\end{aligned} \tag{2}$$

式中：生成模型 G 生成接近真实数据的样本，降低判别模型对假样本的分辨能力；判别模型 D 辨别样本是来自真实数据还是生成模型；判别模型 Q 用于计算隐向量和生成数据间的互信息，通常判别模型 D 和判别模型 Q 除最后一层外，使用相同的深度神经网络。公式右边第一项表示 GAN 模型的对抗损失函数；第二项表示 InfoGAN 的互信息项， 模型选取解耦隐向量的某一维度，并将该维度的索引作为类别标签 d，然后在该维度采样 2 次，其他维度采样一次后生成一对样本 $(\boldsymbol{x}_1,\boldsymbol{x}_2)$；第三项表示样本对 $(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 与 d 之间的互信息；λ、α 是正则化系数。

InfoGAN-CR 通过最大化样本 $(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 和类别标签 d 之间的互信息对模型各维度独立解耦。然而，在模型解耦性能较差时，样本 $(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 中多个生成因子会同时发生变化，这会影响类别标签 d 的准确性。

2 本文模型

针对解耦表示学习问题，本文提出 IDGAN-SC 模型。首先，IDGAN-SC 模型利用 β-VAE 模型学习解耦表示空间，并蒸馏该表示空间作为生成对抗网络的解耦隐向量输入空间[28]。然后，利用最大化解耦隐向量和生成数据之间的互信息对模型进行解耦表示学习，并进一步利用监督对比正则化项的对比分类信息增强解耦隐变量各维度之间的独立性。最后，本文给出了模型的算法表示。

2.1 整体框架

IDGAN-SC 模型蒸馏 β-VAE 模型学习到的解耦表示空间作为生成模型隐向量的输入空间，使隐向量 $\boldsymbol{c}$ 具有解耦因子的信息，而不仅仅是服从简单的高斯先验分布。IDGAN-SC 模型利用监督对比正则化项进一步提升模型的解耦性能。与 InfoGAN-CR 相比，IDGAN-SC 模型提升生成模型前期的解耦能力，从而提高监督对比正则项中正负样本标签的准确性。监督对比正则化项的对比分类信息增强对解耦隐变量各维度之间的独立性约束。原始 IDGAN 模型虽然能够提升 GAN 模型的解耦性能，但其解耦能力仍受限于基于 VAE 模型的选择。本文提出的模型利用监督对比正则化项强化模型的解耦效果。

模型架构如图 2 所示。IDGAN-SC 包括两部分：第一部分采用具有解耦表示能力的 β-VAE 模型进行蒸馏得到解耦表示空间；在第二部分中，生成模型 G 以随机噪声和 β-VAE 模型中学习到的解耦隐向量作为输入生成样本，判别模型 D 使用交叉熵损失函数辨别生成样本的真假，判别模型 Q 采用互信息损失约束解耦隐向量与生成样本间的相关性，对比分类模型 E 提取样本的表示向量并使用监督对比正则化项对其进行约束，使相同类别样本的表示向量更紧密。

首先，IDGAN-SC 使用随机梯度下降方法训练 β-VAE 模型学习数据的解耦表示。训练结束后，β-VAE 的推理模型对数据 $\boldsymbol{x}$ 进行表示推理，得到解耦隐向量 $q(\boldsymbol{c}\mid\boldsymbol{x})$。然后，IDGAN-SC 模型蒸馏 β-VAE 模型的解耦表示空间作为生成模型的解耦隐向量输入空间。

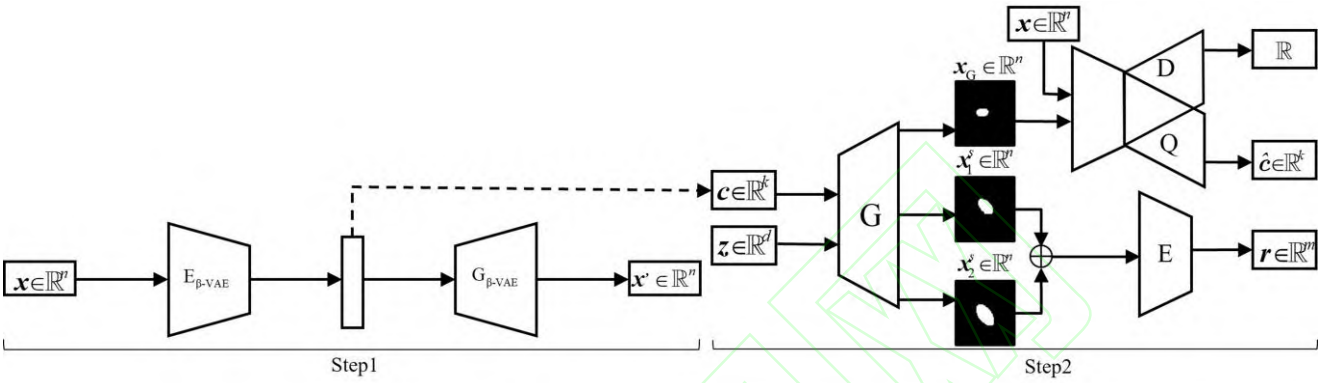

图 2　模型架构图

Fig. 2　Architecture of model

生成模型 G 的输入包括两部分，一部分是普通噪声向量 $\boldsymbol{z}\in\mathbb{R}^d$，该向量一般服从高斯分布；另一部分是解耦隐向量 $\boldsymbol{c}\in\mathbb{R}^k$，该解耦隐向量 $\boldsymbol{c}$ 服从 $q(\boldsymbol{c}\mid\boldsymbol{x})$。生成模型 G 以噪声向量 $\boldsymbol{z}$ 和解耦隐向量 $\boldsymbol{c}$ 的采样为输入生成样本 $\boldsymbol{x}_{\mathrm{G}}$。判别模型 D 采用交叉熵损失判别样本是来自真实数据 $\boldsymbol{x}$ 还是生成数据 $\boldsymbol{x}_{\mathrm{G}}$，判别模型 Q 采用互信息损失约束解耦隐向量 $\boldsymbol{c}$ 与生成数据 $\boldsymbol{x}_{\mathrm{G}}$ 之间的相关性。对比分类模型 E 以一对图像作为输入并提取特征。模型的损失函数为

$$\min_{\mathrm{G,Q,E}}\max_{\mathrm{D}}\mathcal{L}_{\mathrm{IDGAN\_SC}}(\mathrm{G,D,Q,E})=\mathcal{L}_{\mathrm{Adv}}(\mathrm{D,G})-\lambda\mathcal{L}_{\mathrm{info}}(\mathrm{Q,G})-\alpha\mathcal{L}_{\mathrm{SC}}(\mathrm{E,G}) \tag{3}$$

式(3)右侧第一项表示模型的对抗损失，用于判别样本是来自真实数据还是生成模型，具体形式为

$$\mathcal{L}_{\mathrm{Adv}}(\mathrm{D,G})=E_{\boldsymbol{x}\sim P_{\mathrm{data}}}[\log(\mathrm{D}(\boldsymbol{x}))]+E_{\boldsymbol{x}_{\mathrm{G}}\sim\mathrm{G}(\boldsymbol{z},\boldsymbol{c})}[\log(1-\mathrm{D}(\boldsymbol{x}_{\mathrm{G}}))] \tag{4}$$

式(3)右侧第二项表示模型的互信息项，用于最大化解耦隐藏向量 $\boldsymbol{c}$ 与生成数据 $\boldsymbol{x}_{\mathrm{G}}$ 间的互信息，具体形式为

$$\mathcal{L}_{\mathrm{info}}(\mathrm{Q,G})=E_{\boldsymbol{c}\sim q(\boldsymbol{c}\mid\boldsymbol{x}),\boldsymbol{x}_{\mathrm{G}}\sim\mathrm{G}(\boldsymbol{z},\boldsymbol{c})}[\log\mathrm{Q}(\boldsymbol{c}\mid\boldsymbol{x}_{\mathrm{G}})]+H(\boldsymbol{c}) \tag{5}$$

式(3)右侧第三项表示监督对比正则化项，用于区分生成因子维度变化。

2.2　监督对比正则化项

IDGAN-SC 模型通过改变解耦隐向量某一维度的值(其他维度取值不变)生成一对样本，将该维度作为这对样本的类别标签。监督对比正则化项将类别相同的样本作为正样本，其他类别的样本作为负样本，约束正样本与锚点样本在表示空间的距离远远小于负样本与锚点样本的距离，从而增强解耦隐变量各维度之间的独立性。

IDGAN-SC 模型首先随机选取解耦隐向量某一维度 s 并将其作为类别标签，在解耦隐向量第 s 维采样两次而其他维度只采样一次得到 $\boldsymbol{c}_1^s$ 和 $\boldsymbol{c}_2^s$，将噪声 $\boldsymbol{z}$ 和解耦隐向量 $\boldsymbol{c}_1^s$ 作为生成模型 G 的输入生成样本 $\boldsymbol{x}_1^s$，将噪声 $\boldsymbol{z}$ 和解耦隐向量 $\boldsymbol{c}_2^s$ 作为输入生成样本 $\boldsymbol{x}_2^s$，样本对 $(\boldsymbol{x}_1^s,\boldsymbol{x}_2^s)$ 属于类别 s。IDGAN-SC 将样本对 $(\boldsymbol{x}_1^s,\boldsymbol{x}_2^s)$ 作为对比分类模型 E 的输入得到表示向量 $\boldsymbol{r}^s=\mathrm{E}(\boldsymbol{x}_1^s,\boldsymbol{x}_2^s)$。按照上述方法，IDGAN-SC 模型通过选取解耦隐向量不同的维度生成 m 组样本对并得到对应的表示向量，记作 $\{\boldsymbol{r}_i^t;i\in I,t\in S\}$，其中 $I=\{1,\ldots,m\}$ 表示样本索引的集合，S 表示类别标签集合。

监督对比正则化项在集合 I 中随机选取一个样本 $\boldsymbol{r}_i^s$，将与锚点类别相同的样本作为正样本，将其他类别的样本作为负样本，最小化正样本与锚点样本在表示空间的距离，最大化负样本与锚点样本的距离，具体形式为

$$\mathcal{L}_{\mathrm{SC}}(\mathrm{E},\mathrm{G})=\sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\boldsymbol{r}_i^s\bullet\boldsymbol{r}_p^s/\tau_1)}{\sum_{j\in A(i)}\exp(\boldsymbol{r}_i^s\bullet\boldsymbol{r}_j^t/\tau_2)} \tag{6}$$

式中：$P(i)=\{\boldsymbol{r}_p^s;p\neq i,s\}$ 表示锚点样本 $\boldsymbol{r}_i^s$ 的正样本集合；$A(i)=\{\boldsymbol{r}_j^t;j\neq i,t\in S\}$ 表示 I 中除锚点外的其他样本；• 代表表示向量的内积。对数函数中的分子表示样本 $\boldsymbol{r}_i^s$ 与正样本在表示空间中的距离，分母表示样本 $\boldsymbol{r}_i^s$ 与集合 I 中其他样本的距离。

监督对比正则化项能够有效利用标签信息，将相同类别的样本作为正样本，约束正样本得到更加紧密的表示向量；将不同类别的样本作为负样本，扩大负样本间表示向量的距离。本文在实验中设置解耦隐向量的采样次数，使 IDGAN-SC 模型能够生成任意数量的正负样本。

2.3 模型训练

模型训练过程分为两个阶段，模型算法详见算法 1。

算法 1

输入：模型超参数 $\{\beta,\lambda,\alpha,l,m,\tau_1,\tau_2\}$ 和训练数据集合 D，其中 $\boldsymbol{x}\in D$。

输出：模型 β-VAE，生成模型 G，判别模型 D，判别模型 Q 和对比分类模型 E 的参数。

/* 阶段一：训练 β-VAE 模型*/

1. 随机初始化 β-VAE 模型参数

Repeat:

2. 随机抽取大小为 l 的训练数据，记为 D_{batch}。
3. β-VAE 模型以 D_{batch} 作为模型的输入，使用优化目标函数(1)更新模型参数。

until 满足收敛条件

/* 阶段二：训练生成模型 G，判别模型 D，判别模型 Q，和对比分类模型 E */

4. 随机初始化生成模型 G、判别模型 D、判别模型 Q、对比分类模型 E 的参数。

Repeat:

5. 随机抽取大小为 l 的训练数据，记为 D'_{batch}。
6. 固定 β-VAE 模型参数不变，提取 $\boldsymbol{x}\in D'_{\mathrm{batch}}$ 的解耦隐向量 $\boldsymbol{c}$，IDGAN-SC 模型以 $\boldsymbol{z}$ 和 $\boldsymbol{c}$ 作为输入生成样本，按 2.2 中的方法构造 m 组正负样本，使用优化目标函数(3)更新生成模型 G、判别模型 D、判别模型 Q、对比分类模型 E 的参数。

until 满足收敛条件

3 实 验

设计对比实验验证 IDGAN-SC 模型的解耦能力。采用目前主流的用于定量评估模型解耦性能的 dSprites 数据集[13]，以及定性评估模型解耦性能的 MNIST 数据集、CelebA 数据集。dSprites 数据集包含 5 个生成因子的标签：形状、大小、旋转、x 坐标和 y 坐标。MNIST 是标准的手写数字识别数据集，包含数字的风格、粗细、角度、宽度生成因子。CelebA 数据集是人脸属性数据集，包含有人脸表情、角度、背景等生成因子。数据集的详细信息见表 1。

表 1 数据集的详细信息

Tab. 1 Detailed Information of Datasets

数据集名称	训练集数量	数据类型
dSprites	737280	64×64×1 灰度图像
MNIST	70000	28×28×1 灰度图像
CelebA	202599	64×64×3 彩色图像

对比实验的模型包括 FactorVAE 模型[25]，InfoGAN 模型[14]、InfoGAN-CR 模型[15]与 IDGAN-SC 模型。解耦性能评价指标包括 β-VAE 度量[13]、分离属性可预测性(separated attribute predictability，SAP)度量[23]、解耦性/完整性/信息性(disentanglement/completeness/informativeness，DCI)度量[29]、FactorVAE 度量[25]、互信息间隔(mutual information gap，MIG)度量[26]。

3.1 监督对比正则化项对解耦性能的影响

在 dSprites 数据集上，分析监督对比正则化项对模型解耦性能的影响，在相同实验设置下对比 InfoGAN 模型、InfoGAN-CR 模型及 IDGAN-SC 模型的解耦性能，采用 FactorVAE 度量作为评价指标。IDGAN-SC 的生成模型服从高斯先验分布。

实验使用 Adam 优化器进行训练优化，将训练过程分为两个步骤：第一步是 InfoGAN 模型训练阶段，设置迭代次数为 25 并保存模型参数；第二步分别继续训练 InfoGAN 模型、InfoGAN-CR 模型、IDGAN-SC 模型，见表 2。实验设置批大小为 64，生成模型学习率为 0.0001，判别模型和对比分类模型学习率为 0.0002，隐向量 $\boldsymbol{c}$ 的维度为 5，噪声 $\boldsymbol{z}$ 的维度为 5。在监督对比正则化项中，实验设置正负样本数量 m 为 384、λ=0.05、α=2、τ_1=0.07、τ_2=0.1、r 维度 128。对比分类模型 E 设置类别数量与隐向量 $\boldsymbol{c}$ 的维度一致，然后按照 2.2 节的方法随机构造 384 对样本，并得到维度为 128 的表示向量。将 10 次 FactorVAE 度量的平均值作为解耦分数。InfoGAN 模型、InfoGAN-CR 模型、IDGAN-SC 模型在第二步的训练结果如图 3 所示。

表 2 监督对比正则化项消融实验设置

Tab. 2　Ablation experiment settings of supervised contrastive regularization

模型	第一步	第二步
InfoGAN	InfoGAN	无正则化项
InfoGAN-CR	InfoGAN	InfoGAN-CR 正则化项
IDGAN-SC (不使用 β-VAE)	InfoGAN	监督对比正则化项

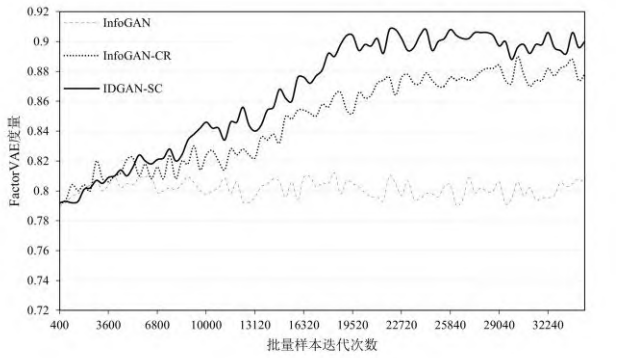

图 3　训练过程中不同方法的性能趋势

Fig. 3　Performance trends of different methods during training

由于图 3 可知，InfoGAN 模型的解耦分数在训练过程中稳定在 0.8 左右。InfoGAN-CR 模型和 IDGAN-SC 模型通过添加正则化项提升模型的解耦分数，在迭代次数为 22000 左右时解耦性能趋于稳定。InfoGAN-CR 模型的解耦分数达到 0.88 左右，IDGAN-SC 模型进一步提高解耦分数至 0.90 左右。相较于 InfoGAN，InfoGAN-CR 模型和 IDGAN-SC 模型的解耦分数分别提高了 0.08 和 0.10。

实验结果表明 InfoGAN-CR 模型通过添加正则化项明显提升模型的解耦性能。IDGAN-SC 监督对比正则化项强化解耦隐向量每个维度之间的独立性，进一步提升解耦效果且效果优于 InfoGAN-CR 模型。

3.2　模型解耦性能的对比实验

IDGAN-SC 模型使用 Adam 训练优化，设置 β-VAE 模型批大小为 64，迭代次数为 50，学习率为 0.0002，参数 β 为 8，生成模型学习率为 0.0001，判别模型和对比分类模型学习率为 0.0002，批大小为 64，迭代次数为 28，解耦隐向量 $\boldsymbol{c}$ 的维度为 10，噪声 $\boldsymbol{z}$ 的维度为 5。在监督对比正则化项中，实验设置正负样本数量 m 为 384、λ=0.05、α=2、τ_1=0.07、τ_2=0.1、r 维度 128。对比分类模型 E 设置类别数量与隐向量 $\boldsymbol{c}$ 的维度一致，然后按照 2.2 的方法随机构造 384 对样本，并得到维度为 128 的表示向量。本节设置 IDGAN-SC(不使用 β-VAE)模型的参数与 3.1 节实验基本一致，差异在于隐向量 $\boldsymbol{c}$ 维度的设置。由于本实验设置 β-VAE 的隐向量维度为 10，为了消除不同维度的隐向量对实验结果的影响，本节设置 IDGAN-SC(使用 β-VAE)和 IDGAN-SC(不使用 β-VAE)的隐向量 $\boldsymbol{c}$ 维度均为 10。在相同参数设置下设计实验：IDGAN-SC(使用 β-VAE)利用 β-VAE 模型学习解耦表示，IDGAN-SC(不使用 β-VAE)模型的解耦隐向量服从高斯分布。本实验取 50 次模型训练的平均值作为结果，其他模型的指标数据来源于 InfoGAN-CR[15]。不同模型在 dSprites 数据集解耦分数的比较结果见表 3。

表 3　不同模型在 dSprites 数据集解耦分数的比较结果

Tab. 3　Comparisons of the disentanglement scores of different models in the dSprites dataset

模型	FactorVAE 度量	DCI 度量	SAP 度量	MIG 度量	β-VAE 度量
FactorVAE (40.0)	0.82±0.10	0.74±0.10	0.56±0.00	0.43±0.01	0.84±0.01
InfoGAN	0.82±0.10	0.60±0.02	0.41±0.02	0.22±0.01	0.87±0.01
InfoGAN-CR	0.88±0.01	0.71±0.10	0.58±0.01	0.37±0.01	0.95±0.01
IDGAN-SC (不使用 β-VAE)	0.90±0.01	0.72±0.10	0.59±0.01	0.38±0.01	0.96±0.02
IDGAN-SC (使用 β-VAE)	0.91±0.02	0.75±0.10	0.60±0.01	0.42±0.02	0.96±0.01

从表 3 可知，IDGAN-SC(使用 β-VAE)和 IDGAN-SC(不使用 β-VAE)在 FactorVAE 度量、DCI 度量、SAP 度量、β-VAE 度量评价指标中的解耦分数显著高于其他模型。在 MIG 评价指标中，IDGAN-SC(使用 β-VAE)模型的解耦分数比 FactorVAE 模型低 0.01 左右。相较于 IDGAN-SC(不使用 β-VAE)模型，IDGAN-SC(使用 β-VAE)利用 β-VAE 模型学习解耦表示空间，进一步提升模型的解耦分数。

通过实验分析可知，IDGAN-SC(不使用 β-VAE)引入监督对比正则化项后，模型在解耦性能上有明显的提升。这也表明监督对比正则化项强化

解耦隐向量每个维度之间的独立性，使模型解耦效果更优。IDGAN-SC(使用 β-VAE)模型使用 β-VAE 模型学习解耦表示空间的方式进一步提高模型的解耦性能。

3.3 定性评估模型解耦性能

3.3.1 dSprites 数据集

在 dSprites 数据集上，设置实验参数与 3.2 节中相同。IDGAN-SC 模型通过遍历隐向量某一维度的值(其他维度不变)生成一组样本，结果如图 4 所示。

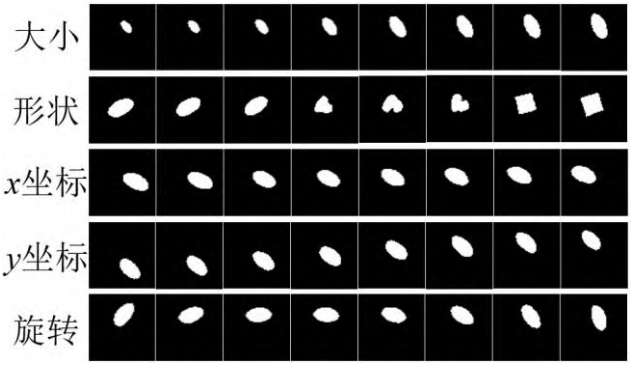

图 4 IDGAN-SC 在 dSprites 数据集上解耦效果

Fig. 4 Disentangled generative effect of IDGAN-SC on the dSprites dataset

由图 4 可知：IDGAN-SC 模型分别控制大小、形状、x 坐标、y 坐标和旋转 5 个生成因子发生变化。实验结果表明在 dSprites 数据集中，IDGAN-SC 模型具有较强的解耦能力。

3.3.2 MNIST 数据集

在 MNIST 数据集中，该实验设置参数与 dSprites 数据集一致。InfoGAN-CR 模型和 IDGAN-SC 模型通过遍历解耦隐向量观察数字粗细风格等生成因子的变化。实验选取数字 4、6、7、8 的生成结果，如图 5、图 6 所示。InfoGAN-CR 模型在控制角度因子时数字 7 的风格发生了变化，控制宽度因子时数字 7 的角度发生了变化，控制粗细因子时数字 4 的角度发生了变化。实验结果表明：在 MNIST 数据集中，InfoGAN-CR 模型通过遍历解耦隐向量控制数据的生成因子时，数字角度和其他生成因子同时发生变化；IDGAN-SC 模型学习到数字粗细、风格、角度、宽度 4 个生成因子，具有更优的解耦性能。

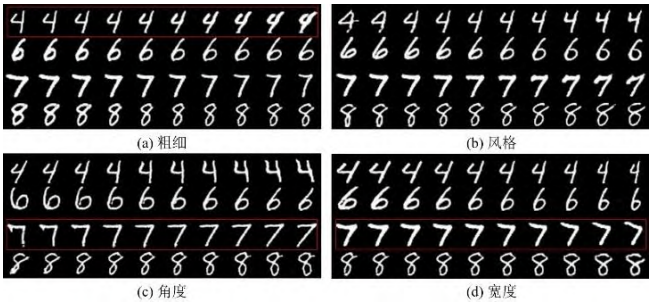

图 5 InfoGAN-CR 在 MNIST 数据集上解耦效果

Fig. 5 Disentangled generative effect of InfoGAN-CR on the MNIST dataset

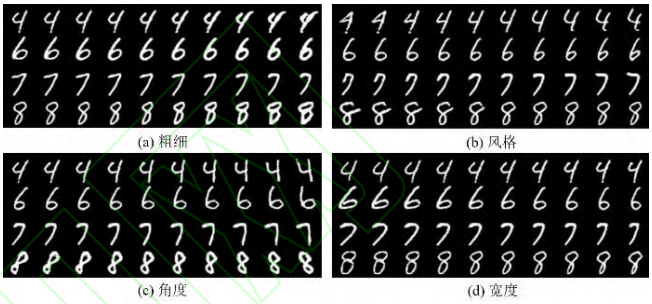

图 6 IDGAN-SC 在 MNIST 数据集上解耦效果

Fig. 6 Disentangled generative effect of IDGAN-SC on the MNIST dataset

3.3.3 CelebA 数据集

在 CelebA 数据集中，IDGAN-SC 模型使用 Adam 进行训练优化，其中 β-VAE 模型批大小为 128，迭代次数为 50，学习率为 0.0002，参数 β 为 6，生成模型批大小为 128，学习率为 0.0002，判别模型和对比分类模型学习率为 0.0004，迭代次数为 57，隐向量 $\boldsymbol{c}$ 的维度为 20，噪声 $\boldsymbol{z}$ 的维度为 256。在监督对比正则化项中，实验设置正负样本数量 m 为 768、λ=0.05、α=2、τ_1=0.07、τ_2=0.1、r 维度 128。对比分类模型 E 设置类别数量与隐向量 $\boldsymbol{c}$ 的维度一致，然后按照 2.2 节的方法随机构造 768 对样本，并得到维度为 128 的表示向量。InfoGAN-CR 模型和 IDGAN-SC 模型通过遍历解耦隐向量的方式生成图像，结果如图 7，图 8 所示。

InfoGAN-CR 模型在控制背景因子和表情因子时人脸角度发生了变化。实验结果表明：在 CelebA 数据集中，InfoGAN-CR 模型控制背景和表情生成因子时，均引起其他生成因子发生明显的变化；IDGAN-SC 模型学习到人脸背景、刘海、角度、光亮、表情 5 个生成因子，具有更优的解耦性能。

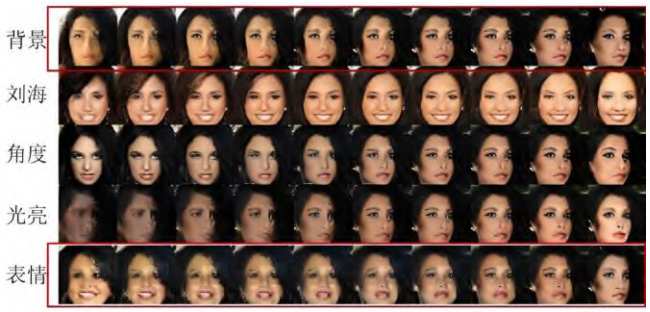

图 7　InfoGAN-CR 在 CelebA 数据集上解耦效果

Fig. 7　Disentangled generative effect of InfoGAN-CR on the CelebA dataset

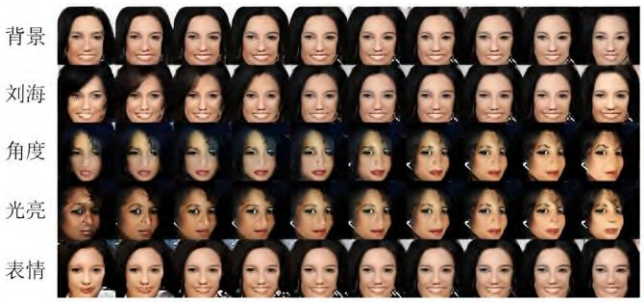

图 8　IDGAN-SC 在 CelebA 数据集上解耦效果

Fig. 8　Disentangled generative effect of IDGAN-SC on the CelebA dataset

4　结　语

本文受到对比学习思想的启发，提出基于监督对比正则化项的信息蒸馏生成对抗网络(IDGAN-SC)。该模型通过引入具有解耦表示能力的变分自编码模型进行蒸馏得到解耦表示空间，约束解耦表示空间和生成模型之间具有强相关性，并利用监督对比正则化项强化解耦隐向量每个维度之间的独立性，增强互信息式生成对抗网络的解耦表示能力。理论分析与实验结果表明，IDGAN-SC 模型具有较强的解耦能力并具有明显的解耦效果。同时本方法具有一些局限性，下一步的工作重点是探索解耦隐向量的维度对解耦性能的影响，并进一步改进模型以增强模型的鲁棒性，使其可以应用于更加复杂的数据。

参考文献：

[1] LEE G, LI H Z. Modeling code-switch languages using bilingual parallel corpus[C]//ACL. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 860–870.

[2] CHEN X H. Simulation of English speech emotion recognition based on transfer learning and CNN neural network[J]. Journal of intelligent & fuzzy systems, 2021, 40(2): 2349-2360.

[3] TORFI A, SHIRVANI R A, KENESHLOO Y, et al. Natural language processing advancements by deep learning: a survey [EB/OL]. (2021-02-27)[2024-01-10]. https://arxiv.org/abs/2003. 01200.

[4] STOLL S, CAMGOZ N C, HADFIELD S, et al. Text2Sign: towards sign language production using neural machine translation and generative adversarial networks[J]. International journal of computer vision, 2020, 128(4): 891-908.

[5] SHI Y C, YU X, SOHN K, et al. Towards universal representation learning for deep face recognition[C]// IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. New York: IEEE, 2020: 6817–6826.

[6] NI T G, GU X Q, ZHANG C, et al. Multi-task deep metric learning with boundary discriminative information for cross-age face verification[J]. Journal of grid computing, 2020, 18: 197-210.

[7] CHEN J T, LEI B W, SONG Q Y, et al. A hierarchical graph network for 3 D object detection on point clouds[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. New York: IEEE, 2020: 392–401.

[8] 蒋弘毅,王永娟,康锦煜. 目标检测模型及其优化方法综述[J]. 自动化学报,2021,47(6):1232-1255.

[9] 文载道,王佳蕊,王小旭,等. 解耦表征学习综述[J]. 自动化学报,2022,48(2):351-374.

[10] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE Transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.

[11] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB/OL]. (2014-05-01)[2024-01-10]. https://arxiv.org/abs/1312. 6114.

[12] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[13] HIGGINS I, MATTHEY L, PAL A, et al. beta-VAE: learning basic visual concepts with a constrained variational framework[C]//ICLR. Proceedings of the 6th International Conference on Learning Representations.

Toulon: ICLR, 2017: 3514-3535.

[14] CHEN X, DUAN Y, HOUTHOOFT R, et al. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets[EB/OL]. (2016-06-12)[2024-01-10]. https://arxiv. org/abs/1606. 03657.

[15] LIN Z, THEKUMPARAMPIL K, FANTI G, et al. InfoGAN-CR and modelcentrality: self-supervised model training and selection for disentangling GANs [C]//ICML. Proceedings of the 37th International Conference on Machine Learning. Cambridge: PMLR, 2020: 6127–6139.

[16] HENAFF O. Data-efficient image recognition with contrastive predictive coding[C]//ICML. Proceedings of the 37th International Conference on Machine Learning. Cambridge: PMLR, 2020: 4182–4192.

[17] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]// ICML. Proceedings of the 37th International Conference on Machine Learning. Cambridge: PMLR, 2020: 1597–1607.

[18] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 9729–9738.

[19] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning[J]. Advances in neural information processing systems, 2020, 33: 18661-18673.

[20] BURGESS C P, HIGGINS I, PAL A, et al. Understanding disentangling in β-VAE[EB/OL]. (2018-04-10) [2024-01–10]. https://arxiv. org/abs/1804. 03599.

[21] SHAO H J, YAO S C, SUN D C, et al. Controlvae: controllable variational autoencoder[C]//ICML. Proceedings of the 37th International Conference on Machine Learning. Cambridge: PMLR, 2020: 8655–8664.

[22] MAKHZANI A, SHLENS J, JAITLY N, et al. Adversarial Auto-encoders[EB/OL]. (2016-05-25)[2024-01-10]. https://arxiv. org/abs/1511. 05644.

[23] KUMAR A, SATTIGERI P, BALAKRISHNAN A. Variational inference of disentangled latent concepts from unlabeled observations[EB/OL]. (2018-12-27)[2024-01-10]. https://arxiv. org/abs/1711. 00848.

[24] ARJOVSKY M, BOTTOU L. Towards principled methods for training generative adversarial networks[EB/OL]. (2017-01-10)[2024-01-10]. https://arxiv. org/abs/1701. 04862.

[25] KIM H, MNIH A. Disentangling by factorising[C]// ICML. Proceedings of the 35th International Conference on Machine Learning. Cambridge: PMLR, 2018: 2649–2658.

[26] CHEN R T Q, LI X C, GROSSE R B, et al. Isolating sources of disentanglement in variational autoencoders[EB/OL]. (2019-04-23)[2024-01-10]. https://arxiv. org/abs/1802. 04942.

[27] JEON I, LEE W, PYEON M, et al. Ib-gan: disentangled representation learning with information bottleneck generative adversarial networks[C]// AAAI. Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2021: 7926–7934.

[28] LEE W, KIM D, HONG S, et al. High-fidelity synthesis with disentangled representation[C]//ECCV. Proceedings of the 16th European Conference on Computer Vision. Berlin: Springer, 2020: 157–174.

[29] EASTWOOD C, WILLIAMS C K I. A framework for the quantitative evaluation of disentangled representations [C]//ICLR. Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR, 2018: 2317－2231.