第 38 卷　第 4 期
2023 年 8 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 38　No. 4
Aug. 2023

DOI:10.13364/j.issn.1672-6510.20220257

基于生成对抗网络的变分自编码器解耦合

张贤坤，赵亚婷，丁文强，张翼英
（天津科技大学人工智能学院，天津 300457）

摘　要：深度生成模型从观测数据中学习到潜在因素，然后通过潜在因素生成目标，在人工智能领域受到广泛关注。现有深度生成模型学习的潜在因素往往是耦合的，无法让潜在因素每一维控制所得数据的不同特征，即无法单独改变某一特征而不影响其他特征。为此，在 β–变分自编码器（beta-variational autoencoder，β-VAE）的基础上，结合生成对抗网络（generative adversarial networks，GAN），提出基于生成对抗网络的变分自编码器（beta-variational autoencoder based on generative adversarial network，β-GVAE）模型。该模型是一种改进的 β-VAE，通过引入生成对抗网络约束 β-VAE 中损失函数的 KL 项（Kullback-Leibler divergence），促进模型的解耦合。在数据集 CelebA、3D Chairs 和 dSprites 上进行对比实验，结果表明 β-GVAE 不仅具有更好的解耦合表示，同时生成的图像具有更好的视觉效果。

关键词：解耦合；β–变分自编码器；生成对抗网络；深度生成模型
中图分类号：TP399　　**文献标志码**：A　　**文章编号**：1672-6510（2023）04-0062-07

Decoupling of Variational Autoencoder Based on Generative Adversarial Network

ZHANG Xiankun，ZHAO Yating，DING Wenqiang，ZHANG Yiying
（College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：Deep generative models learn latent factors from observational data，and then generate targets through latent factors，which have received extensive attention in the field of artificial intelligence. The latent factors learned by the existing deep generative models are often coupled，and each dimension of the latent factors cannot control different characteristics of the obtained data，that is，it is impossible to change a certain characteristic independently without affecting other characteristics. Therefore，beta-variational autoencoder（β-VAE） based on generative adversarial network（β-GVAE） is proposed based on β-VAE and combined with generative adversarial networks（GAN）. This model is an improved β-VAE，which promotes the decoupling of the model by introducing a generative adversarial network to constrain the KL divergence of the loss function in β-VAE. By designing comparative experiments on three datasets，CelebA，3D Chairs and dSprites，it is proved that β-GVAE not only has better decoupled representation，but also the generated images has better visual effects.

Key words：decoupling；beta-variational autoencoder；generative adversarial network；deep generative models

近年来，解耦合表示学习[1]引起了机器学习界的广泛关注。解耦合表示学习的目的是得到解耦合的潜在因素，这种解耦合的潜在因素从观测数据中学习得到，潜在因素的维度之间相互独立，每个维度控制一种特征的生成，彼此之间互不影响[2-6]。解耦合表示学习具有一定的优势：当它用于下游任务时，可以提高预测性能，降低样本复杂度，提供可解释性[7-8]，提高公平性，并已被确定为克服深度学习[9]中快捷学习[10-11]的一种方法。

解耦合一直没有一个标准的定义，每个人对解耦合的具体定义可能都不完全相同，但其所表达的解耦合的含义是相同的，都可以通过一个例子来解释。例

收稿日期：2022–11–14；**修回日期**：2023–01–15
基金项目：天津市科技计划项目（22KPXMRC00210）
作者简介：张贤坤（1970—），男，安徽芜湖人，教授；**通信作者**：赵亚婷，硕士研究生，zhaoyating@mail.tust.edu.cn

如对于人脸数据，可能得到的解耦合的潜在因素有十维，第一维控制肤色，第二维控制头发的长度，第三维控制眼睛的大小；如果调整第一维，保留其他维度不变，就可以生成同一个人脸不同肤色的图像。典型的解耦合表示学习方法主要有三大类。第一类是基于变分自编码器（variational autoencoder，VAE）[12-13]，使用特定分布的随机化向量作为输入并生成相应的数据，不使用判别器而是使用编码器估计特定分布，促进模型学习可分离的潜在变量表示，从而达到解耦合的效果，但该类方法未考虑到真实世界的复杂语义信息，一般只能应用在简单数据集进行解耦表征学习。第二类是基于生成对抗网络（generative adversarial networks，GAN）[14-15]，通过对抗方式训练生成器和判别器，生成器用于生成尽可能逼真的假样本，判别器则尽可能准确地区分真假样本。该类方法能够处理复杂场景大规模数据集以及数据流信息的解耦合，然而生成对抗网络存在训练不稳定、模式崩溃和梯度消失等问题。第三类是基于主成分分析（principal components analysis，PCA）[16-17]，利用线性投影将高维数据映射到低维空间中并尽可能保留最大的信息量。目前利用 PCA 算法提取特征主要应用在人脸识别领域，在复杂的人脸识别算法中可以得到较好的解耦合效果，而对于较为简单的数据集，使用 PCA 算法进行解耦表征学习与前面两种方法得到的结果相差不大，反而显得有些浪费资源空间。

通过学习这些方法观察到，VAE 的重构能力很高但解耦合效果很差，当对 VAE 损失函数中的 KL（Kullback-Leibler divergence）项增大权重时，可以让模型产生较好的解耦合效果。基于此，β–变分自编码器（beta-variational autoencoder，β-VAE）[18]在 VAE 的基础上对 VAE 损失函数中的 KL 项加以限制，这样尽管模型重构能力有所下降，但解耦合能力有一定程度的提高。β-VAE 中参数 β 的人为设置导致模型过于死板，缺少灵活性，而生成对抗网络恰好可以解决此问题。因此，本文在 β-VAE 的基础上结合 GAN，提出基于生成对抗网络的变分自编码器（beta-variational autoencoder based on generative adversarial network，β-GVAE）模型，引入生成对抗网络进一步对 KL 项进行限制。这种限制可以让神经网络自主训练学习，让 KL 项中的估计分布 $p(z|x)$ 更接近真实分布 $p(z)$，既可以学习到网络中隐含的内容，也避免了人为对损失函数 KL 项限定的主观性，增加了模型的灵活性。

本文首先介绍 VAE[12]、β-VAE[18]以及 GAN[14]，其次对提出的网络框架 β-GVAE 进行详细介绍。该方法用生成对抗网络进一步约束 β-VAE 损失函数中的 KL 项，使模型具有更好的解耦合表示。与此同时，生成对抗网络还会优化模型的生成数据，使生成数据具有更好的视觉效果。最后结合实验验证该生成模型在给定参数的情况下能够增加推理模型的表示能力，解耦合性能更好，可以有效提升图像的生成效果。

1　背　景

1.1　变分自编码器

变分自编码器（VAE）是由推理模型（又称编码器）和生成模型（又称解码器）组成，其中推理模型是通过多层神经网络将真实数据 x 编码为一个低维隐变量 z，生成模型是将隐变量 z 通过多层神经网络还原映射到高维度数据空间[19]。变分自编码模型如图 1 所示，其中：白色节点 z 表示隐变量，灰色节点 x 表示可观测量，节点之间的有向线段表示变量之间的依赖关系；$q_\varphi(z|x)$ 与虚线部分为推理过程，$p_\theta(x|z)$ 与实线部分为生成过程，θ 和 φ 为相关过程的参数；方框表示该过程可以重复出现，例如在该模型中基于隐变量可以重复生成数据样本，观测到 N 条数据，则该过程重复出现了 N 次[20-21]。

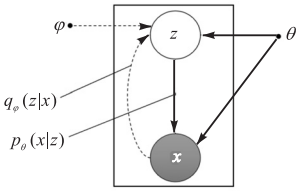


图 1　变分自编码模型

Fig. 1　Variational autoencoder model

VAE 假设高维数据 x 是由低维隐变量 z 生成，其生成模型表示形式为

$$p(x,z)=p(z)p_\theta(x|z) \tag{1}$$

其中：$p(z)=N(z;0,\boldsymbol{I})$ 表示隐向量先验概率分布，一般为标准多元高斯分布，$\boldsymbol{I}$ 表示单位矩阵；$p_\theta(x|z)=N(x;\mu_\theta,\sigma_\theta^2\boldsymbol{I})$ 表示条件概率分布，θ 为生成模型神经网络参数。VAE 中的数据生成过程为：先从先验分布 $p(z)$ 中采样隐向量 z，然后将 z 输入到条件概率分布 $p_\theta(x|z)$ 中生成数据 x[22]。

对于上述生成模型，难以精确计算边缘概率分布 $p(x)$ 和后验分布 $p(z|x)$，VAE 则通过引入变分推理

模型 $q_\varphi(z|x)$ 近似后验分布 $p(z|x)$，将推理问题转化为优化问题，其中 φ 为推理模型中神经网络参数。对于单样本点 x，VAE 的证据下界（ELBO）为

$$L_{\text{ELBO}} = E_{q_\varphi(z|x)}\left[\ln p_\theta(x|z)\right] - D_{\text{KL}}\left(q_\varphi(z|x)\|p(z)\right) \tag{2}$$

其中：式（2）右边第一项表示重构误差，第二项 KL 散度用来约束 VAE 的隐空间。

此时，变分自编码模型的目标是通过随机梯度下降算法[23]学习到最优的模型参数 φ 和 θ，使证据下界最大，即

$$\theta,\varphi = \arg\max_{\theta,\varphi} L(\theta,\varphi) \tag{3}$$

1.2　β-VAE

β-VAE 是对变分自编码器的改进，它为原始的 VAE 目标引入了一个可调的超参数 β，β-VAE 的损失函数为

$$L(\theta,\varphi;x,z,\beta) = E_{q_\varphi(z|x)}\left[\ln p_\theta(x|z)\right] - \beta D_{\text{KL}}\left(q_\varphi(z|x)\|p(z)\right) \tag{4}$$

其中：θ 为生成模型神经网络的网络参数，φ 为推理模型神经网络的网络参数，x 为观测样本，z 为隐变量，β 为限制 KL 项的超参数，$q_\varphi(z|x)$ 为 β-VAE 的编码器部分，$p_\theta(x|z)$ 为 β-VAE 的解码器部分。

选择良好的 β 值（通常是 $\beta>1$）会导致更多的解耦合的潜在表示 z。当 $\beta=1$ 时，β-VAE 模型将等同于原来的 VAE 框架。β-VAE 实际上是对原始 VAE 损失函数的第二项 $D_{\text{KL}}\left(q_\varphi(z|x)\|p(z)\right)$ 施加更强的约束，让 $q_\varphi(z|x)$ 和标准高斯分布 $p(z)$ 更加接近，从而获得解耦合的能力，并且仍然可以很好地重建样本 x。

与此同时，较好的解耦合往往导致重建效果不好，较好的重建效果往往导致解耦合的能力不好。因此，鼓励解耦合所必需的更高的 β 值通常需要在 β-VAE 重建的保真度与其潜在代码 z 的解耦合性质之间权衡。

1.3　生成对抗网络

生成对抗网络（GAN）[14]受启发于博弈论中的二人零和博弈理论，其独特的对抗训练思想能生成高质量的样本，具有比传统机器学习算法更加强大的特征学习和特征表达能力。

GAN 的网络结构由生成网络和判别网络两部分组成，模型结构如图 2 所示。生成器 G 接收随机变量 z，生成假样本数据 $G(z)$。生成器的目的是尽量使生成的样本和真实样本一样。判别器 D 的输入由两部分组成，分别是真实数据 x 和生成器生成的数据 $G(z)$，其输出通常是一个概率值，表示 D 认定输入是真实分布的概率，若输入来自真实数据，则输出 1，否则输出 0。判别器的输出会反馈给 G，用于指导 G 的训练。理想情况下 D 无法判别输入数据是来自真实数据 x 还是生成数据 $G(z)$，即 D 每次的输出概率值都为 1/2（相当于随机猜），此时模型达到最优。在实际应用中，生成网络和判别网络通常用深层神经网络实现。

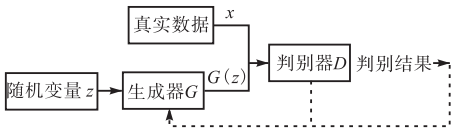


图 2　GAN 网络模型结构示意图

Fig. 2　Structure diagram of GAN network model

GAN 的思想来自博弈论中的二人零和博弈理论，生成器和判别器可以看成是博弈中的两个玩家。在模型训练的过程中，生成器和判别器会各自更新自身的参数使损失最小，通过不断迭代优化，最终达到纳什均衡状态[24]，此时模型达到最优。GAN 的目标函数定义为

$$\min_G \max_D V(D,G) = E_{x\sim p_{\text{data}}(x)}\left[\log D(x)\right] + E_{z\sim p_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \tag{5}$$

2　基于生成对抗网络的变分自编码器

2.1　模型结构

β-GVAE 的网络结构图如图 3 所示，主要由两部分组成，上半部分为 β-VAE，下半部分为 GAN 的判别器。

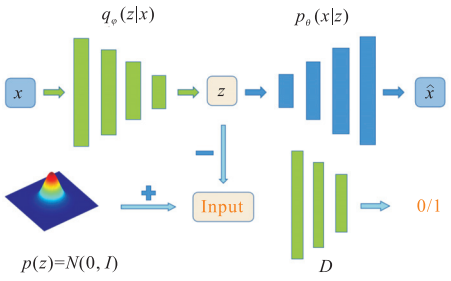


图 3　β-GVAE 的网络结构

Fig. 3　Network structure of β-GVAE

$q_\varphi(z|x)$ 为 β-GVAE 的编码器，同时相当于 GAN 中的生成器。$p_\theta(x|z)$ 为 β-GVAE 的解码器，θ 和 φ 为生成模型神经网络的网络参数和推理模型神经网络的网络参数。D 为 β-GVAE 中 GAN 部分的判别

器，由推理模型 $q_\varphi(z|x)$ 得到的隐变量 z 作为负样本，由 $p(z)=N(0,I)$ 中采样得到的隐变量 z 作为正样本，将正负样本一同输入到判别器 D 中，若输入来自正样本，则输出 1，否则输出 0。

2.2　损失函数

β-GVAE 的损失函数是在 β-VAE 的基础上，通过引入了 GAN 约束 β-VAE 中损失函数的 KL 项，促进模型的解耦合。模型 β-GVAE 中 GAN 的损失函数表示形式为

$$\min_G \max_D V(D,G)=E_{x\sim p_{\text{data}}(x)}\left[\log D(x)\right]+E_{z\sim p(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \tag{6}$$

其中：G 代表生成器，其本质为模型 β-GVAE 中 β-VAE 的编码器，D 代表判别器；$z\sim p(z)$ 代表从标准正态分布 $p(z)=N(0,I)$ 中采样得到的隐变量 z 作为正样本，$x\sim p_{\text{data}}(x)$ 代表从样本数据中采样样本 x，然后将样本 x 输入到生成器 G 中，得到生成的隐变量 z 作为负样本。在训练时，将正负样本输入到判别器 D 中，若判别器 D 的输入来自正样本，则输出 1，否则输出 0。

β-GVAE 的损失函数表示形式为

$$L(\theta,\varphi,\delta;x,z,\beta,\gamma)=E_{q_\varphi(z|x)}\left[\ln p_\theta(x|z)\right]-\beta D_{\text{KL}}\left(q_\varphi(z|x)\|p(z)\right)+\gamma D\left(q_\varphi(z|x)\right) \tag{7}$$

其中：θ 为生成模型神经网络的网络参数，φ 为推理模型神经网络的网络参数，δ 为 GAN 中判别器的神经网络参数，x 为观测样本，z 为隐变量，β 为限制 KL 项的超参数，γ 为判别器的超参数，$q_\varphi(z|x)$ 为 β-GVAE 的编码器部分，$p_\theta(x|z)$ 为 β-GVAE 的解码器部分，D 为判别器。

本文通过在 β-VAE 中引入 GAN 增强 β-VAE 中损失函数 KL 项的约束力，促进 β-GVAE 模型的解耦合。

2.3　超参数优化

在模型的训练过程中，超参数的选择至关重要，它可以使模型发挥更好的性能，得到最优的结果。在训练模型 β-GVAE 的过程中，参考 β-VAE 模型[18]中参数 β 的选择，综合考虑本文模型的超参数 β 和 γ，最终得到最优的超参数并进行训练。超参数的选择见表 1。

表 1　超参数的选择

Tab. 1　Hyperparameter selection

指标	超参数选择						
	$\beta=2,\gamma=2$	$\beta=3,\gamma=3$	$\beta=4,\gamma=1$	$\beta=4,\gamma=4$	$\beta=4,\gamma=10$	$\beta=4,\gamma=15$	$\beta=4,\gamma=20$
重构误差	**228.511**	236.819	242.275	244.180	230.867	243.808	243.951
KL 散度	20.363	17.997	**16.073**	16.197	16.235	16.097	16.079

注：加粗数字表示最优的超分指标值。

由表 1 可知：当 β=2、γ=2 时，重构误差取得最优值，为 228.511；当 β=4、γ=1 时，KL 散度值取得最优值，为 16.073。由于重构误差和 KL 散度成反比，即重构误差小时 KL 散度大，重构误差大时 KL 散度小，因此选择 β=4、γ=10 作为模型 β-GVAE 训练时的超参数。此时，模型 β-GVAE 的重构误差和 KL 散度都可以取得较好的结果。

3　实　验

设计对比实验，验证 β-GVAE 模型的解耦合和生成能力。具体包括两个实验：实验 1，在 CelebA 数据集上设计对比实验，计算模型 VAE、β-VAE 和 β-GVAE 的重构误差和 KL 散度，进行精度对比；实验 2，展示模型 VAE、β-VAE 和 β-GVAE 在数据集 CelebA、3D Chairs 和 dSprites 上的生成图像，对比模型的解耦效果和图像生成效果。

3.1　实验设置

本实验是在 NVIDIA GeForce RTX 3090 GPU 上训练所有网络，训练数据来自 CelebA、3D Chairs 和 dSprites 数据集。CelebA 是名人人脸属性数据集，其包含 10 177 个名人身份的 202 599 张人脸图片，由香港中文大学开放提供，广泛用于与人脸识别相关的计算机视觉训练任务。3D Chairs 是一个三维椅子的数据集，其中包括各式各样的椅子，关于椅子的潜在因素有 width、size、azimuth、leg style、back height 等。dSprites 是一个二维形状的数据集，由 6 个与地面实况无关的潜在因素程序生成。这些因素是 dSprites 的 color、shape、scale、rotation、x 位置和 y 位置。这些潜在的所有可能组合只出现一次，总共生成 737 280 张图像。

3.2　定量对比

本实验对比了模型 VAE、β-VAE 和 β-GVAE 在数据集 CelebA 上生成图像的重构误差和 KL 散度。

实验中的重构误差主要体现模型生成图像的能力，重构误差越小，生成图像效果越佳；KL 散度主要体现模型的解耦合能力，KL 散度越小，说明解耦合能力越好。实验结果如表 2、图 4 和图 5 所示。

表 2　定量结果实验对比

Tab. 2　Experimental comparison of quantiative results

模型	重构误差	KL 散度
VAE	**227.243**	24.878
β-VAE	247.686	16.528
β-GVAE	230.867	**16.235**

注：加粗数字表示最优的超分指标值。

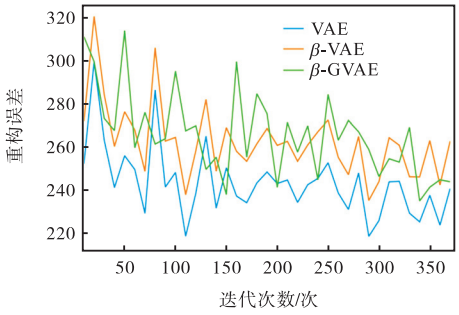


图 4　不同模型的重构误差对比

Fig. 4　Comparison of reconstruction errors between different models

由表 2、图 4 和图 5 可知，VAE 具有最小的重构误差，但模型 β-GVAE 的重构误差与 VAE 相差不大，低于 β-VAE。β-GVAE 模型不仅具有较低的重构误差，还具有最低的 KL 散度，这说明 β-GVAE 的解耦合能力最好。综上所述，本文模型 β-GVAE 不仅具有更好的解耦合能力，而且生成的数据具有较好的视觉效果，远远超过 β-VAE 模型。

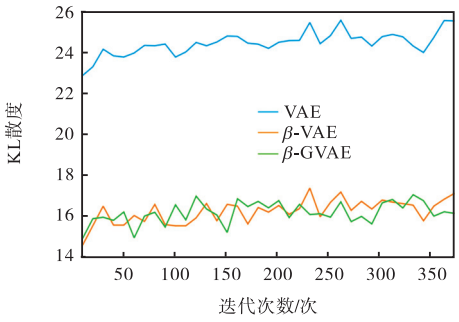


图 5　不同模型的 KL 散度对比

Fig. 5　Comparison of KL divergence among different models

3.3　定性对比

为了进一步展示 3 个模型 VAE、β-VAE 和 β-GVAE 的图像生成效果和解耦合效果，对比了 3 个模型在 CelebA、3D Chairs 和 dSprites 数据集上的生成图像，如图 6—图 8 所示。

(a) VAE　　(b) β-VAE　　(c) β-GAVE

图 6　不同模型在 CelebA 数据集上生成图像的结果

Fig. 6　Results of generated images on the CelebA dataset using different models

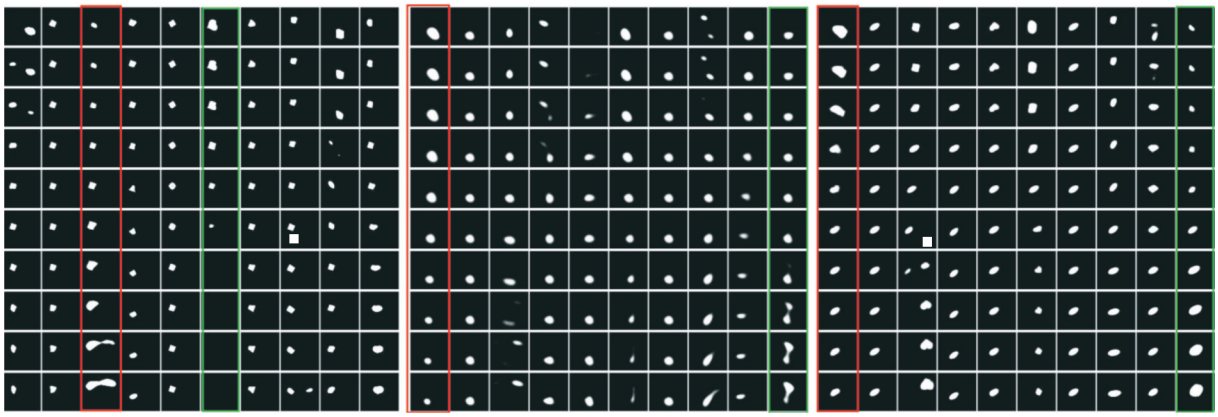
从图 6—图 8 可以观察到模型 VAE、β-VAE 和 β-GVAE 在 3 个数据集上的生成图像效果。由于图像较多，在每组图像中选取两组进行解释说明，两组图像分别用红框和绿框标记。

由图 6 可知：在 CelebA 数据集上，VAE 模型红框中的生成图像有背景颜色和性别两个维度的转变，绿框中有背景颜色、人脸角度、头发颜色 3 个维度的转变，而 β-VAE 模型和 β-GVAE 模型红框中的生成图像分别有人脸角度和头发颜色仅 1 个维度的转变，绿框中都只有背景颜色和性别两个维度的转变。这表明 β-VAE 模型和 β-GVAE 模型相比较于 VAE 有更好的解耦合能力，且 β-GVAE 模型生成图像的质量看起来比 β-VAE 模型的高。

由图 7 可知：在 dSprites 数据集上，VAE 模型红框中的生成图像有 shape 和 scale 2 个维度的转变，但是生成了较差的 shape，绿框中也有 shape 和 scale 2

个维度的转变，但最后没有生成图像；β-VAE 模型的生成图像红框中和绿框中分别有 scale 和 shape 各 1 个维度的转变，但最后会生成较差的 shape；β-GVAE 模型生成图像红框中和绿框中也同样各有 1 个维度的转变。这表明 β-GVAE 模型解耦合能力较好，且生成图像的质量更高。

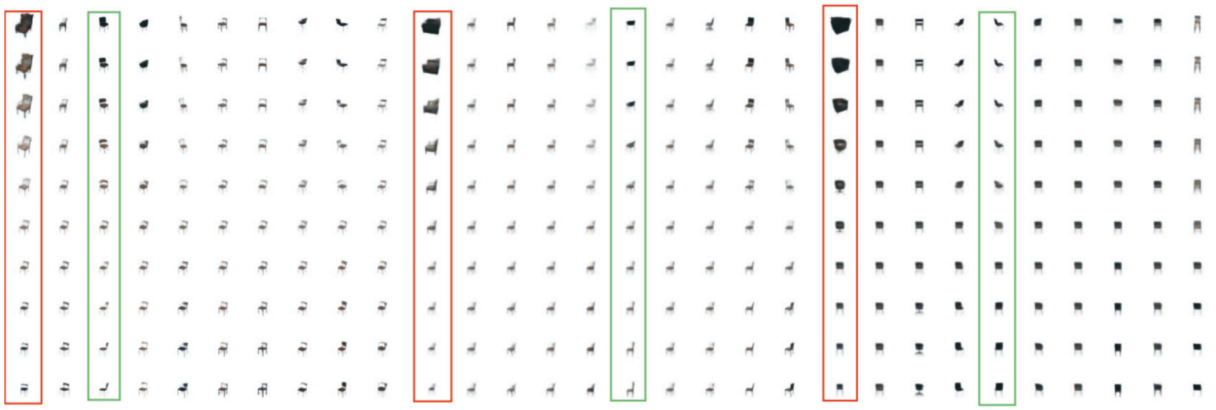
由图 8 可知：在 3D Chairs 数据集上，观察红框与绿框中椅子的 size 和 rotation 两个维度的变化，可以看出 β-GVAE 模型与 VAE 模型、β-VAE 模型具有类似的解耦合能力，但最后生成图像的质量看起来比 VAE 模型和 β-VAE 模型的高。



(a) VAE　　(b) β-VAE　　(c) β-GAVE

图 7　不同模型在 dSprites 数据集上生成图像的结果

Fig. 7　Results of generated images on the dSprites dataset using different models



(a) VAE　　(b) β-VAE　　(c) β-GAVE

图 8　3D Chairs 生成结果

Fig. 8　Results of generated images on the 3D Chairs dataset using different models

综上所述，本文所提出的 β-GVAE 模型不仅具有更好的解耦合能力，而且生成的图像具有不错的视觉效果，效果远远超过 VAE 模型与 β-VAE 模型。

4　结　语

本文在 β-VAE 模型的基础上，结合生成对抗网络，提出了一种新的网络框架 β-GVAE，将生成对抗网络中的生成器看作 β-VAE 模型中的编码器，判别器作为其解码器，进而可以约束损失函数的 KL 项，促进模型的解耦合，提高生成数据的视觉效果。在真实数据集上的实验结果表明，本文算法与经典模型 VAE 和 β-VAE 相比，解耦合能力更强，生成图像效果更好，模型表现更为突出。

解耦表示学习的研究尚处在起步阶段，仍然面临着一些挑战，未来解耦表示学习的研究应该更加着力于对归纳偏好、无监督或者自监督学习的探索以及设计不同模型实现的解耦程度的量化标准。

参考文献：

[1]　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.

[2] LOCATELLO F, TSCHANNEN M, BAUER S, et al. Disentangling factors of variations using few labels [C]//ICLR. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2020.

[3] DITTADI A, TRÄUBLE F, LOCATELLO F, et al. On the transfer of disentangled representations in realistic settings[C]//ICLR. Proceedings of the 9th International Conference on Learning Representations. Addis Ababa: ICLR, 2021.

[4] SCHANNEN M, BACHEM O, LUCIC M. Recent advances in autoencoder-based representation learning [EB/OL]. [2021–03–30]. https://arxiv.org/abs/1812.05069.

[5] SHU R, CHEN Y N, KUMAR A, et al. Weakly supervised disentanglement with guarantees[C]//ICLR. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2020.

[6] KIM H, SHIN S, JANG J, et al. Counterfactual fairness with disentangled causal effect variational autoencoder [C]//AAAI. Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2021: 8128–8136.

[7] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12: 2493–2537.

[8] TANG Y, TANG Y, ZHU Y, et al. A disentangled generative model for disease decomposition in chest X-rays via normal image synthesis[J]. Medical image analysis, 2021, 67: 101839.

[9] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504–507.

[10] GEIRHOS R, JACOBSEN J H, MICHAELIS C, et al. Shortcut learning in deep neural networks[J]. Nature machine intelligence, 2020, 2(11): 665–673.

[11] MINDERER M, BACHEM O, HOULSBY N, et al. Automatic shortcut removal for self-supervised representational learning[C]//JMLR. Proceedings of the 37th International Conference on Machine Learning. San Diego: JMLR, 2020: 6927–6937.

[12] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB/OL]. [2021–03–30]. https://arxiv.org/abs/1312.6114.

[13] DOERSCH C. Tutorial on variational autoencoders [EB/OL]. [2021–03–30]. http://arxiv.org/abs/1606.05908.

[14] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//NIPS. Proceedings of the 27th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 139–144.

[15] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. [2021–03–30]. http://arxiv.org/abs/1511.06434.

[16] 吴翊, 李永生, 胡庆军. 应用数理统计[M]. 长沙: 国防科技大学出版社, 1995.

[17] 梅长林, 周家良. 实用统计方法[M]. 北京: 科学出版社, 2006: 53–65.

[18] HIGGINS I, MATTHEY L, PAL A, et al. β-VAE: learning basic visual concepts with a constrained variational framework[EB/OL]. [2021–03–30]. https://openreview.net/forum?id=Sy2fzU9gl.

[19] 翟正利, 梁振明, 周炜, 等. 变分自编码器模型综述[J]. 计算机工程与应用, 2019, 55(3): 1–9.

[20] 杨晨曦, 左劼, 孙频捷. 基于自编码器的零样本学习方法研究进展[J]. 现代计算机, 2020(1): 48–52.

[21] PATACCHIOLA M, FOX-ROBERTS P, ROSTEN E. Yautoencoders: disentangling latent representations via sequential encoding[J]. Pattern recognition letters, 2020, 140: 59–65.

[22] LOCATELLO F, BAUER S, LUCIC M, et al. Challenging common assumptions in the unsupervised learning of disentangled representations[C]//JMLR. Proceedings of the 36th International Conference on Machine Learning. New York: JMLR, 2019: 4114–4124.

[23] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic backpropagation and approximate inference in deep generative models[EB/OL]. [2021–03–30]. http://www.arxiv.org/pdf/1401.4082.pdf.

[24] 林懿伦, 戴星原, 李力, 等. 人工智能研究的新前线: 生成式对抗网络[J]. 自动化学报, 2018, 44(5): 775–792.

责任编辑: 郎婧