艾庆忠, 秦国旭, 王加昌,等. 基于对比标准化流的多模态生成模型研究[J]. 智能计算机与应用,2026,16(4):153-158.
DOI:10.20169/j.issn.2095-2163.26011602

基于对比标准化流的多模态生成模型研究

艾庆忠[1,2], 秦国旭[1,2], 王加昌[1,2], 庞 勃[1,2], 万静意[1,2], 马邓勇[1,2], 唐 雷[1,2]
(1 中国核动力研究设计院 核反应堆技术全国重点实验室, 成都 610213; 2 中国核动力研究设计院, 成都 610213)

摘 要: 人类的智能很大程度体现为对多感官信息的快速感知能力。然而,利用多模态数据进行学习在深度学习领域仍是一个极具挑战性的问题。作为一种新颖的多模态数据学习范式,多模态变分自编码器(VAE)近年来备受关注,并取得了令人瞩目的成果。然而,由于现有多模态变分自编码器模型在共享潜在空间中的对齐约束或受限的后验设计,导致其在语义连贯性和生成质量方面存在不足。为此,本研究提出了一种新的多模态深度生成模型,名为基于对比标准化流的变分自编码器,简称为 CNF-MVAE(Contrastive Normalizing Flow-based Multimodal Variational AutoEncoder)。在 CNF-MVAE 中,首先引入了基于可逆标准化流的"元"潜在空间,该空间是与特定模态相关潜在空间相分离的高层次语义空间。随后,采用对比学习实现在元潜空间中样本级别的对齐。CNF-MVAE 能够在显著提升语义连贯性的同时,几乎不降低生成质量。在基准多模态数据集上的实验结果表明,CNF-MVAE 在不同任务中均实现了优于其他当前先进多模态变分自编码器的竞争力提升,充分验证了模型的有效性。

关键词: 多模态生成模型; 对比学习; 变分自编码器
中图分类号: TP181 **文献标志码**: A **文章编号**: 2095-2163(2026)04-0153-06

Research on multimodal generation models based on contrastive normalizing flows

AI Qingzhong[1,2], QIN Guoxu[1,2], WANG Jiachang[1,2], PANG Bo[1,2], WAN Jingyi[1,2], MA Dengyong[1,2], TANG Lei[1,2]
(1 National Key Laboratory of Nuclear Reactor Technology, Nuclear Power Institute of China, Chengdu 610213, China;
2 Nuclear Power Institute of China, Chengdu 610213, China)

Abstract: Human intelligence is largely manifested as the ability to rapidly perceive multisensory information. However, learning from multimodal data remains a highly challenging problem in deep learning. As a novel paradigm for multimodal data learning, multimodal variational autoencoder (VAE) have attracted significant attention in recent years and achieved remarkable results. Nevertheless, existing multimodal VAE suffer from limitations in semantic coherence and generation quality due to alignment constraints or restricted posterior designs in shared latent spaces. To address these issues, this study proposes a novel multimodal deep generative model, named Contrastive Normalizing Flow-based Multimodal Variational Autoencoder (CNF-MVAE). In CNF-MVAE, a high-level semantic space—the "meta" latent space—is introduced, decoupled from modality-specific latent spaces through invertible normalizing flows. Subsequently, contrastive learning is employed to achieve sample-level alignment within the meta latent space. CNF-MVAE significantly enhances semantic coherence while preserving generation quality. Experimental results on benchmark multimodal datasets demonstrate that CNF-MVAE consistently outperforms other state-of-the-art multimodal VAE across various tasks, validating its effectiveness.

Key words: multimodal generative models; contrastive learning; variational autoencoders

0 引 言

人类通过多种感官信息,能够对事物形成可靠的认知。例如,提到"小狗"的概念时,其外形和吠叫声都会在人们的意识中浮现。因此,如何模拟人脑对多模态数据进行有效学习,成为一项既至关重要又充满挑战的机器学习任务,由此催生的研究被称为多模态学习[1]。高效的多模态学习依赖于数据信息的完整性,因此有监督方法已得到广泛且成功的应用[2-3]。然而,标注数据往往非常耗时且成本高昂,尤其是在处理多种类型数据时这一问题更为突出。因此在本研究中采用无监督方法,其中生成模型作为一种代表性策略,因其能够学习多模态数据的联合分布而备受关注。在种类繁多的生成模

作者简介: 艾庆忠(1995—),男,博士,主要研究方向:深度生成模型。Email:aqz1995@163.com。

收稿日期: 2026-01-16

型中,变分自编码器[4-5]凭借其快速推理与生成的能力脱颖而出,并在数据降维[6]、学习表示[7]以及数据生成[8]等领域取得了显著成效。更为重要的是,变分自编码器在多模态数据上的可扩展性使其成为最受欢迎的基础模型之一,由此衍生出的模型被称为多模态变分自编码器。近年来,沿着这一研究方向涌现出许多先进的模型,例如 CADA-VAE[9]、MVAE[10]、MMVAE[11]、MoPoE-VAE[12]等。

现有多模态变分自编码器模型中存在若干有待改进的问题。其一是其生成质量较差[13],与单模态变分自编码器相比,现有的多模态变分自编码器在生成质量上始终表现欠佳。其二,先前的多模态变分自编码器还面临多模态样本之间联合与交叉语义一致性不足的问题,这成为这些模型应用的一大障碍。上述问题的主要根源可归结为以下两个方面。首先,各模态共享的潜在空间限制了模型的性能。不同模态的数据维度和数据类型通常各不相同,因此需要为不同模态的数据分别构建特定的潜在空间。然而,现有模型倾向于使用共享的潜在空间,从而制约了单模态模型的表现,换句话说,这损害了模型的生成质量。其次,不同模态直接在共享潜在空间中进行语义对齐,反而不利于单模态生成的效果。为了提升语义一致性,一些模型,如 AVAE[14],在共享潜在空间中引入了模态对齐约束。然而,这种做法会破坏各模态内部特有的潜在空间结构,进而影响生成质量。

针对上述缺陷,本文引入“元”潜在空间,旨在构建一个与特定模态相关的潜在空间相分离的高层次语义空间,如图 1 所示。通过可逆标准化流,将“元”潜在空间与各个特定模态的潜在空间连接起来,实现双向映射。随后,在“元”潜在空间中通过对比约束进行语义对齐,从而提出了一种全新的多模态深度生成模型,命名为基于对比标准化流的变分自编码器(CNF-MVAE)。在 CNF-MVAE 中,模态对齐约束从模型特定的潜在空间转移到了“元”潜在空间。这样的话,特定模态潜在空间的空间结构便不会受到其他约束的干扰,从而确保了各模态生成样本的质量。此外,在“元”潜在空间中引入实例级对比学习,进一步保证了生成的多模态样本之间的语义一致性。最后,在基准多模态数据集上对 CNF-MVAE 进行了测试,实验结果表明,CNF-MVAE 在不同任务中均表现出色,包括语义一致性、生成质量和潜在表示。尤其值得一提的是,与其它先进的多模态变分自编码器相比,CNF-MVAE 能够在保持良好生成质量的同时,实现最佳的语义一致性。

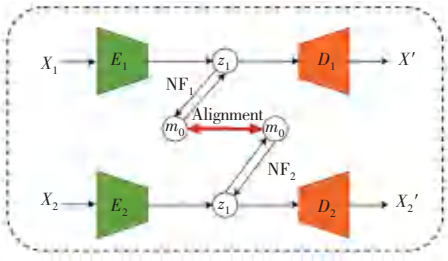


图 1 基于对比标准化流的多模态生成模型示意图

Fig. 1 Diagram of a multimodal generative model based on contrastive normalized flows

1 方法

1.1 基础知识

在介绍所提出的模型之前,首先简要介绍一些预备知识,包括变分自编码器和标准化流[15-16]。这两者均属于生成模型的范畴。设数据集 $D=\{x_i\}_{i=1}^{N}$ 包含 N 个独立同分布的样本,其证据分布为 $\tilde{q}(x)$,生成模型的目标是通过如下形式的分布来拟合该数据集。

$$p(x)=\int p(z)p(x\mid z)\mathrm{d}z \tag{1}$$

其中,$p(z)$ 作为先验分布,通常采用标准高斯分布。$p(x\mid z)$ 描述了生成过程,这一过程因不同的生成模型而异。例如,变分自编码器采用高斯分布,而在生成对抗网络和基于标准流的模型中,则采用狄拉克分布。在优化过程中,理想情况下可以最大化期望函数 $E_{x\sim\tilde{q}}[\log p(x)]$,或者等价地,最小化 KL(Kullback-Leibler,KL)散度 $(\tilde{q}(x)\parallel p(x))$。然而,由于需要知道真实的后验分布 $p(z\mid x)$,证据的显式计算是不可行的。因此,研究人员采用了各种技巧,从而催生了多种不同的生成模型。在本工作中重点探讨其中两种:变分自编码器和标准化流模型。

变分自编码器通过引入变分后验分布 $q(z\mid x)$ 来近似未知的真实后验分布 $p(z\mid x)$,从而解决这一问题。这种近似通过 KL 散度进行约束,具体如下:

$$\mathrm{KL}[q(z\mid x)\parallel p(z\mid x)]=$$
$$Eq(z\mid x)[\log q(z\mid x)-\log p(z\mid x)] \tag{2}$$

通过应用贝叶斯法则,变分自编码器的目标函数可以形式化为证据下界(ELBO),其定义如下式:

$$L_{\mathrm{ELBO}}=E_{q(z\mid x)}[\log p(x\mid z)]-\mathrm{KL}[q(z\mid x)\parallel p(z)] \tag{3}$$

其中，第一项是重构误差，第二项 KL 散度是正则化项。

标准化流又称为基于流的模型，通过精心设计的模型架构来计算积分。因此，标准化流能够直接最大化对数似然函数。具体而言，基于流的模型 f 被构建为一种可逆变换，其将观测数据 x 映射到标准高斯潜变量 z，即 $z=f(x)$。为了确保模型的可逆性和可计算性，该模型 f 由一系列简单的可逆流逐层堆叠而成，形式为 $f(x)=f_1\circ f_2\circ\cdots\circ f_L(x)$，其中每个 f 都具有可计算的逆和可计算的雅可比行列式。在训练过程中，可直接最大化对数似然，因为该模型的概率密度易于计算，其定义如下：

$$\log p(x)=\log p(z)+\sum_{i=1}^{L}\log\left|\det\frac{\partial f_i}{\partial f_{i-1}}\right| \tag{4}$$

标准化流的采样可通过计算 $f^{-1}(z)=f_L^{-1}\circ f_{L-1}^{-1}\circ\cdots\circ f_1^{-1}$ 来实现，其中 z 服从标准高斯分布 $N(0,I)$。

1.2　基于对比标准化流的变分自编码器

在本文中，提出了一种新颖的多模态深度生成模型 CNF-MVAE，旨在提升多模态变分自编码器的语义一致性、生成质量和潜在表示能力。为此，首先为每种单模态变分自编码器引入了一个“元”潜在空间，该空间是一个高层次语义空间，通过可逆标准化流与各模型特有的潜在空间相连接。随后，在“元”潜在空间中采用对比约束，以实现实例级别的对齐。通过这种方式，能够在不破坏各模态特有潜在空间结构的前提下，确保不同模态之间的语义对齐，从而跨模态地保障更优质的生成效果。此外，在“元”潜在空间中使用对比对齐还能够促进更优的潜在空间表示。为了更清晰地介绍 CNF-MVAE，首先利用标准化流为单模态变分自编码器引入“元”潜在空间；其次通过对比约束在元空间中对齐各模态。

1.2.1　基于标准化流的元空间构造

为了将对齐约束从特定于模型的潜在空间中分离出来，以获得更优质的生成效果，首先利用可逆标准化流为单模态变分自编码器引入了“元”潜在空间。

可以直接变分自编码器的 ELBO 损失函数入手，得到：

$$L_{\mathrm{ELBO}}=E_{q(z|x)}[\log p(x|z)]-\mathrm{KL}[q(z|x)\,\|\,p(z)]=-\int q(z|x)\log\frac{q(z|x)}{p(x|z)p(z)}\mathrm{d}z \tag{5}$$

在经典的变分自编码器中，上式中的变分后验分布 $q(z|x)$ 通常是一个具有均值函数 $\mu(x)$ 和协方差函数 $\sum(x)$ 的高斯分布，该分布充当编码器；而 $p(x|z)$ 则充当解码器。与经典变分自编码器不同的是，本文采用标准化流来构建一种功能更强大的后验分布，而非传统的高斯分布，其形式为：

$$q(z|x)=\int\delta(z-f_x(u))p(u)\mathrm{d}z \tag{6}$$

其中，$\delta(\cdot)$ 为狄拉克函数；$p(u)$ 为元潜在空间中的标准高斯分布；$f_x(u)$ 是一种标准化流，其在给定 x 的情况下对 u 是可逆的。换句话说，$f_x(u)$ 是变量 u 的标准化流，而其参数则是变量 x 的函数。随后，可得到新的 ELBO 损失函数，如下式所示：

$$L_{\mathrm{ELBO}}=-\int p(u)\log\frac{p(u)}{p(x|f_x(u))p(f_x(u))\left|\det\dfrac{\partial f_x(u)}{\partial u}\right|}\mathrm{d}u \tag{7}$$

理论上可以设计任意复杂的函数 f，以提升模型的表达能力；但对于一些常见数据而言，这种做法既浪费资源又无必要。因此，出于对实用性的考量，本文沿用了文献[17]的设定，在该设定中，$f_x(u)$ 及其对应的 $p(x|z)$ 被定义为：

$$f_x(u)=f(\mathrm{Encoder}(x)+\sigma_1 u) \tag{8}$$

$$p(x|z)=N(x|\mathrm{Decoder}(f^{-1}(z)),\sigma_2) \tag{9}$$

其中，Encoder 和 Decoder 分别代表编码器和解码器。σ_1 和 σ_2 是可训练参数，均为标量。得益于标准化流的可逆性，采样过程简单高效，具体如下：

$$u\sim p(u)\Rightarrow z=f^{-1}(u)\Rightarrow x=\mathrm{Decoder}(z) \tag{10}$$

其中，u 是元潜在空间中的随机样本，z 是对应于模型特定潜在空间中的样本。最后，便可得到带有标准化流的单模态变分自编码器的最终损失：

$$L_{\mathrm{VAE-NF}}=-\int p(u)\frac{1}{2\sigma_2^2}\|\mathrm{Decoder}(\sigma_1+\mathrm{Encoder}(x))-x\|^2+\frac{1}{2}\|f(\sigma_1 u+\mathrm{Encoder}(x))\|^2-\frac{1}{2}\|u\|^2-\log\left|\det\frac{\partial f(\mathrm{Encoder}(x)+\sigma_1 u)}{\partial u}\right|\mathrm{d}u \tag{11}$$

1.2.2　基于对比学习的语义对齐

对于一个多模态生成模型，各模态生成样本之间的语义一致性至关重要。例如，在图像-文本多模态生成任务中，随机生成的一对样本必须描述的是同一件事物（联合一致性）。此外，根据文本条件生成的图像（或反之）也应与参考图像保持相同的语义内容（跨模态一致性）。因此，不同模态之间必须施加语义对齐约束。

为此，本文采用了一种基于对比学习的实例级

对齐约束。作为一类具有代表性的表征学习范式，对比学习旨在最大化正样本对之间的相似性，同时最小化负样本对之间的相似性[18]。为了便于理解，将重点聚焦于两种模态的场景来介绍本文的实现方案，分别记为 m_1 和 m_2。关于扩展到多于两种模态的情况，将在稍后进行说明。形式上，设 $x_b = \{x_i^{m_1}, x_i^{m_2}\}_{i=1}^{N}$ 为数据集的一个小批次，其大小为 N。$\boldsymbol{u}_b = \{\boldsymbol{u}_i^{m_1}, \boldsymbol{u}_i^{m_2}\}_{i=1}^{N}$ 则是经过映射解码器 Decoder 与标准化流 f 后，在“元”潜在空间中对应的嵌入向量。随后，构建正负样本对，如图 2 所示，这与标准的对比学习略有不同。具体而言，对于某一模态的任意嵌入向量，正样本是另一模态中对应实例（相同索引）的嵌入向量，而同一小批次中的其他嵌入向量则作为负样本。以 $\boldsymbol{u}_1^{m_1}$ 为例，其正样本为 $\boldsymbol{u}_1^{m_2}$，负样本为 $\boldsymbol{u}_{2-N}^{m_2}$，其大小为 $N-1$。成对样本的相似度通过内积来衡量，即：

$$s(\boldsymbol{u}_i^{k_1}, \boldsymbol{u}_j^{k_2}) = (\boldsymbol{u}_i^{k_1})(\boldsymbol{u}_j^{k_2})^{\mathrm{T}} \tag{12}$$

其中，$k_1, k_2 \in \{m_1, m_2\}$，且 $i, j \in [1, N]$。为不失一般性，基于相似度函数，给定嵌入向量 $\boldsymbol{u}_i^{m_1}$（对应于样本 $x_i^{m_1}$）的损失定义为：

$$L_i^{m_1} = -\log \frac{\exp(s(\boldsymbol{u}_i^{m_1}, \boldsymbol{u}_i^{m_2})/\tau)}{\sum_{j=1}^{N} [\exp(s(\boldsymbol{u}_i^{m_1}, \boldsymbol{u}_j^{m_2})/\tau)]} \tag{13}$$

其中，τ 为温度参数。确切地说，是一个超参数，用于控制模型区分负样本的能力。因此，两种模态之间的最终对比对齐损失形式如下：

$$L_{\mathrm{Con}}^{m_1, m_2} = \frac{1}{2N} \sum_{i=1}^{N} (L_i^{m_1} + L_i^{m_2}) \tag{14}$$

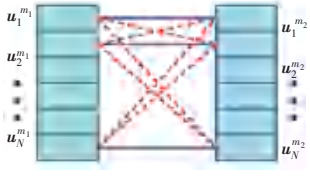


图 2 对比策略示意图

Fig. 2 Illustration of the contrastive strategy

1.2.3 最终损失函数

为了让模型同时兼顾单模态特征学习与跨模态特征对齐。本文将 CNF-MVAE 的损失函数设计为单模态损失以及跨模态对比对齐损失的加权组合。对于具有两种模态的模型，最终损失函数的形式为：

$$L_{\mathrm{CNF-MVAE}} = L_{\mathrm{VAE-NF}}^{m_1} + L_{\mathrm{VAE-NF}}^{m_2} + \lambda L_{\mathrm{Con}}^{m_1, m_2} \tag{15}$$

其中，λ 为正则化权重。

可将上面的方法扩展至多模态场景。将 CNF-MVAE 推广到包含两种以上模态的多模态数据集是十分直观的。设 D^M 为包含 M 种模态的多模态数据集。对于采用标准化流的单模态变分自编码器，其损失函数的扩展非常简单，只需将各模态的损失相加即可，需要注意的是对比损失的构建方式。具体而言，依次将每种模态与除自身以外的所有其他模态进行配对，从而得到共计 $(M-1)$ 种组合。每一对组合均使用公式(14)计算损失，随后将所有组合的损失求和。通过这种方式，CNF-MVAE 的通用损失函数如下式：

$$L_{\mathrm{CNF-MVAE}} = \sum_{i=1}^{M} L_{\mathrm{VAE-NF}}^{i} + \sum_{i=1}^{M} \sum_{j=1; j \neq i}^{M} \lambda_{ij} L_{\mathrm{Con}}^{m_i, m_j} \tag{16}$$

其中，λ_{ij} 是模态对 m_i 和 m_j 的正则化权重。

2 实验

为验证 CNF-MVAE 的有效性，在基准数据集的多个任务上将其与不同类型的先进模型进行了对比。任务涉及语义连贯性、生成质量以及潜在表征；数据集为 MNIST、SVHN。用于对比的模型包括 CADA-VAE、MMVAE 和 MoPoE-VAE。

训练过程中使用了 Adam 优化器，初始学习率为 0.000 1，并采用每 40 步降低 10%的学习率衰减策略。同时为确保所有实验中所用模型的一致性，均采用了高斯形式的先验和后验分布。此外，根据数据集的维度和类型，使用了不同的网络架构，但同一数据集下对比模型的主要骨干网络保持一致。例如，MNIST 采用多层感知器（MLP），SVHN 则采用卷积神经网络（CNN）。

2.1 实验设置

根据 MMVAE 中的设定，首先构建了一个图像-图像多模态数据集，该数据集由 MNIST[19] 和 SVHN[20] 的成对图像组成，以探索与感知复杂性截然不同的概念复杂性。每一对图像均描绘相同的数字类别。采用多对多配对方式：具体而言，一个数据集中的每个样本随机与另一个数据集中相同数字类别的 10 个样本进行配对，反之亦然。

尽管 MNIST 和 SVHN 本身都不是复杂的数据集，且均已得到充分研究，但要有效捕捉不同风格图像对之间直接的类别匹配信息仍颇具挑战性。如前所述，MNIST 的所有模型均采用多层感知机，SVHN 的所有模型均采用卷积神经网络。训练使用 100 个 epoch，mini-batch 大小为 256。标准化流的层数为 5，潜在空间的维度为 40。在 CNF-MVAE 中，对比正则化权重 λ 设置为 10。由于 MNIST-SVHN 数据

集的标签信息完整，通过不同任务从多个角度测试了模型的性能，包括生成样本之间的联合与跨语义一致性、图像样本的生成质量，以及潜在空间的表征能力。

2.2　语义一致性测试

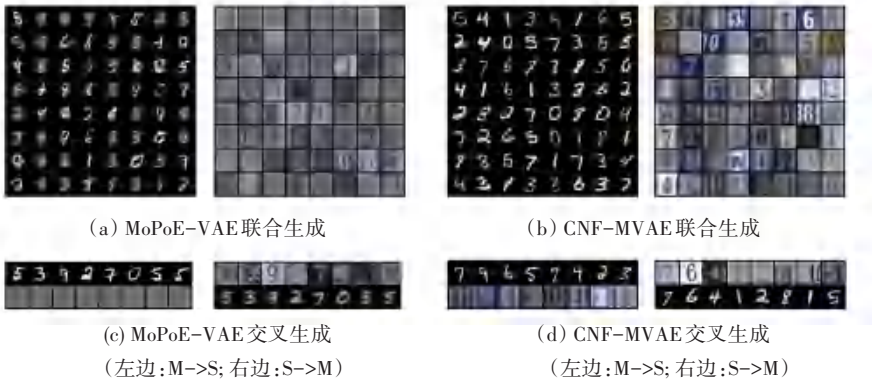
首先，遵循 MMVAE 模型评估准则，评估 CNF-MVAE 的语义一致性，这是多模态生成模型最关键的指标之一。具体而言，语义一致性测试采用两种方法，包括联合一致性和交叉一致性。在联合一致性评估中，模型需要在生成的成对数据中保留共同特性。具体操作流程包括：从同一个先验样本 z 生成一个数据点的所有模态的样本，使用训练好的分类器判断生成的模态是否属于同一类别。交叉一致性评估则着重验证模型在给定另一可观测模态条件下恢复缺失模态的能力。其实现过程为：通过交换潜在变量 z 实现交叉生成，使用训练好的分类器判断交叉生成的模态类别是否与可观测模态的类别一致。这种双重一致性验证机制确保模型能够同时满足模态间特征保留和条件生成的质量要求。

结果汇总见表 1。总体而言，该模型优于所有其他基线模型，表明由 CNF-MVAE 生成的样本能够更好地保留不同模态之间的共性（包括联合模态和跨模态）。尤其在 M→S 的设置中，CNF-MVAE 表现出显著优势，与当前比较先进模型 MoPoE 相比，提升幅度仍达到 $\Delta = 53.23$。定性结果如图 3 所示。图 3（a）与 3（b）左右两图分别表示 MoPoE-VAE 与 CNF-MVAE 的联合生成效果，即由同一隐空间采样点同时生成两种模态数据。图 3（c）与 3（d）上下两图分别表示 MoPoE-VAE 与 CNF-MVAE 的交叉生成效果，即由一个模态（上面）映射到隐空间后，生成另一模态的结果。可以看出，在联合一致性和交叉一致性生成任务中，该模型生成的样本质量均优于 MoPoE，进一步验证了 CNF-MVAE 的有效性。

表 1　MNIST-SVHN（S：SVHN；M：MNIST）中联合与跨语义一致性的比较结果

Table 1　Comparison results of joint and cross - semantic consistency in MNIST - SVHN （S：SVHN；M：MNIST）

模型	联合	交叉（S->M）	交叉（M->S）
CADA-VAE	28.29	75.36	26.87
MMVAE	28.93	74.93	26.43
MoPoE	31.42	64.45	29.99
CNF-MVAE	**33.45**	**75.67**	**83.22**



（a）MoPoE-VAE联合生成　　（b）CNF-MVAE联合生成

（c）MoPoE-VAE交叉生成　　（d）CNF-MVAE交叉生成

（左边：M->S；右边：S->M）　　（左边：M->S；右边：S->M）

图 3　MoPoE-VAE 与 Conf-MVAE 的生成效果对比

Fig. 3　Comparison of generation effects between MoPoE-VAE and Conf-MVAE

2.3　生成质量

在现有研究中，多模态生成模型通常在潜在空间中引入语义对齐约束以增强模态间的语义连续性，但这往往以牺牲单模态生成质量为代价。需要强调的是，生成质量是评估生成模型（尤其是多模态模型）的核心指标。为验证所提出模型的有效性，对 CNF-MVAE 生成样本的质量进行了量化评估。具体采用弗雷歇 inception 距离（FID）[21] 作为评估指标，该指标通过计算生成图像与真实标注图像分布之间的差异来衡量生成质量（FID 值越低表明生成图像与真实图像的相似度越高）。针对 MNIST-SVHN 数据集，对所有基线模型随机采样 20 000 对图像，并分别计算各模态的 FID 值，结果见表 2。可以发现，CNF-MVAE 在各模态下的生成质量均优于大多数基线模型。

表 2　MNIST-SVHN 中各模态随机采样图像的 FID 分数对比

Table 2　Comparison of FID scores for randomly sampled images from each modality on MNIST-SVHN

模型	随机 MNIST	随机 SVHN
CADA-VAE	80.22	93.65
MMVAE	88.27	126.11
MoPoE	129.53	120.74
CNF-MVAE	**50.76**	**68.11**

2.4 隐空间表征学习

为验证 CNF-MVAE 在潜在空间中共享信息(数字类别分辨)的能力,在 2 000 个随机训练样本的潜在变量上训练数字分类器,并在完整测试集上评估分类准确率。结果见表 3,该方法优于所有基线模型,表明 CNF-MVAE 能够学习更具结构化和语义意义的潜在空间表示。

表 3 MNIST-SVHN 中潜在表示分类准确率的比较结果

Table 3 Comparison of classification accuracy of latent representations on MNIST-SVHN

模型	MNIST	SVHN
CADA-VAE	82.34	75.52
MMVAE	81.23	76.12
MoPoE	73.24	69.45
CNF-MVAE	**96.25**	**77.53**

3 结束语

在本文中,提出了一种新的多模态深度生成模型 CNF-MVAE,该模型基于对比标准化流,旨在提升多模态变分自编码器的语义连贯性和生成质量。通过利用标准化流的可逆性,首先引入“元”潜在空间(一个与特定模态潜在空间分离的高层次语义空间),并在该空间中采用对比约束实现跨模态语义对齐。该模型在保证生成样本语义连贯性的同时,最小化了对单模态生成质量的影响。在多个任务和基准多模态数据集上的实验验证了其有效性。

参考文献

[1] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423-443.

[2] KARPATHY A, FEI-FEI L. Deep visual-semantic alignments for generating image descriptions [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3128-3137.

[3] RAHATE A, WALAMBE R, RAMANNA S, et al. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions[J]. Information Fusion, 2022, 81: 203-239.

[4] KINGMA D P, WELLING M. Auto-encoding variational bayes [J]. arXiv preprint arXiv, 1312.6114, 2013.

[5] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic backpropagation and approximate inference in deep generative models[C]// Proceedings of International Conference on Machine Learning. IMLS, 2014: 1278-1286.

[6] GREGOR K, BESSE F, REZENDE D J, et al. Towards conceptual compression[J]. arXiv preprint arXiv, 1604.08772, 2016.

[7] CHEN R T, LI X, GROSSE R, et al. Isolating sources of disentanglement in variational autoencoders[J]. arXiv preprint arXiv, 1802.04942, 2018.

[8] AI Q, HE L, LIU S, et al. ByPE-VAE: Bayesian Pseudocoresets Exemplar VAE [C] //Proceedings of Advances in Neural Information Processing Systems. NeurIPS, 2021: 5910-5920.

[9] SCHONFELD E, EBRAHIMI S, SINHA S, et al. Generalized zero-and few-shot learning via aligned variational autoencoders [C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8247-8255.

[10] WU M, GOODMAN N. Multimodal generative models for scalable weakly-supervised learning [C] //Proceedings of Advances in Neural Information Processing Systems. NeurIPS, 2018: 5575-5585.

[11] SHI Y, PAIGE B, TORR P, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models[C] //Proceedings of Advances in Neural Information Processing Systems. NeurIPS, 2019: 4769-4779.

[12] SUTTER T M, DAUNHAWER I, VOGT J E. Generalized multimodal elbo[J]. arXiv preprint arXiv, 2105.02470, 2021.

[13] DAUNHAWER I, SUTTER T M, CHIN-CHEONG K, et al. On the limitations of multimodal vaes[J]. arXiv preprint arXiv, 2110.04121, 2021.

[14] JO D U, LEE B, CHOI J, et al. Associative variational auto-encoder with distributed latent spaces and associators [C] // Proceedings of AAAI Conference on Artificial Intelligence. AAAI, 2020: 11197-11204.

[15] DINH L, KRUEGER D, BENGIO Y. Nice: Non-linear independent components estimation[J]. arXiv preprint arXiv, 1410.8516, 2014.

[16] KINGMA D P, DHARIWAL P. Glow: Generative flow with invertible 1×1 convolutions[C] //Proceedings of Advances in Neural Information Processing Systems. NeurIPS, 2018: 10215-10224.

[17] SU J, WU G. f-VAEs: Improve VAEs with conditional flows [J]. arXiv preprint arXiv, 1809.05861, 2018.

[18] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C] //Proceedings of International Conference on Machine Learning. IMLS, 2021: 8748-8763.

[19] LECUN Y. The mnist database of handwritten digits[DB/OL]. (1998-11-30)[2025-09-06]. http://yann.lecun.com/exdb/mnist/, 1998.

[20] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning [C] // Proceedings of Advances in Neural Information Processing Systems. NeurIPS, 2011: 1097-11041.

[21] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C] //Proceedings of Advances in Neural Information Processing Systems. NeurIPS, 2017: 6626-6637.