计算机科学
COMPUTER SCIENCE

视觉表征学习综述

王帅炜, 雷杰, 冯尊磊, 梁荣华

引用本文

王帅炜, 雷杰, 冯尊磊, 梁荣华. 视觉表征学习综述[J]. 计算机科学, 2024, 51(11): 112-132.

WANG Shuaiwei, LEI Jie, FENG Zunlei, LIANG Ronghua. Review of Visual Representation Learning[J]. Computer Science, 2024, 51(11): 112-132.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100089

视觉表征学习综述

王帅炜[1]　雷　杰[1]　冯尊磊[2]　梁荣华[1]
1 浙江工业大学计算机科学与技术学院　杭州 310023
2 浙江大学计算机科学与技术学院　杭州 310027
(swwang@zjut.edu.cn)

摘　要　表征学习是人工智能算法中的重要一环,好的表征能够让后续的下游任务事半功倍。随着深度学习在计算机视觉领域的发展,视觉表征学习变得越来越重要,其目的是将复杂的视觉信息转换为更易于人工智能算法学习的表达。文中主要介绍了目前广泛使用的视觉表征学习的研究工作,根据数据依赖程度和类型的不同,将其划分为预训练视觉表征学习、生成式视觉表征学习、对比式视觉表征学习、解耦式视觉表征学习以及结合语言信息的视觉表征学习。具体而言,预训练视觉表征学习是基于有监督的预训练模型在视觉表征学习上的应用;生成式视觉表征学习利用生成模型学习视觉表征;对比式视觉表征学习主要介绍了利用对比学习思想来学习视觉表征的各类网络框架。此外,还介绍了利用变分自编码器和生成对抗网络在解耦式视觉表征学习中的应用,以及利用语言信息来增强视觉表征学习的各种方法。最后,总结了视觉表征学习的评价准则和未来展望。

关键词:视觉表征学习;人工智能算法;解耦式视觉表征学习;语言信息
中图分类号　TP391

Review of Visual Representation Learning

WANG Shuaiwei[1],LEI Jie[1],FENG Zunlei[2] and LIANG Ronghua[1]
1 College of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310023,China
2 College of Computer Science and Technology,Zhejiang University,Hangzhou 310027,China

Abstract　Representation learning is an important step of artificial intelligence algorithm,where well designed representation can boost downstream tasks. With the development of deep learning in computer vision,visual representation learning has become increasingly important,aiming at transforming complex visual information into representation that is easier for artificial intelligence algorithm to learn. In this paper,we focus on current research works widely used in visual representation learning,which are categorized as pre-trained visual representation learning,generative visual representation learning,contrastive visual representation learning,decoupled visual representation learning,and visual representation learning combined with language information according to the degrees and types of data dependency. Specifically,pre-trained visual representation learning is the application of supervised pre-training model in visual representation learning;generative visual representation learning uses generative model to learn visual representations;and contrastive visual representation learning focuses on the various network frameworks which using contrast learning to learn visual representations. Besides,the paper presents the applications of VAE and GAN in decoupled visual representation learning,as well as various approaches to improve visual representation learning with language information. Finally,evaluation metrics in visual representation learning and future perspectives are summarized.

Keywords　Visual representation learning,Artificial intelligence algorithm,Decoupled visual representation learning,Language information

1　引言

计算机视觉技术在人类的生活和生产过程中扮演着举足轻重的角色。在视觉任务中,输入数据(如图片、视频等)都是高维且冗余复杂的,传统的手动提取特征已变得不现实,如何从数据中学习更高质量的特征成为视觉领域重要的研究问题。表征学习是将复杂的原始数据化繁为简,消除原始数据的无效信息,提炼有效的信息,形成更容易应用于机器学习的特征的过程。因此,从视觉信息中学习更高质量的表征是解决视觉任务的关键一环。目前,大部分表征学习的综述[1-3]

到稿日期:2023-11-14　返修日期:2024-04-11
基金项目:国家自然科学基金(62106226,62036009);浙江省自然科学基金(LQ22F020013,LDT23F0202)
This work was supported by the National Natural Science Foundation of China(62106226,62036009) and Natural Science Foundation of Zhejiang Province,China(LQ22F020013,LDT23F0202).
通信作者:雷杰(jasonlei@zjut.edu.cn)

涉及多个领域的技术,如视觉领域、自然语言处理、网络表征等,或者总结了表征学习方法,如解耦表征学习综述[4-5]、多模态视觉语言表征学习研究综述[6-7]等等,但尚缺乏对视觉表征学习的系统性总结。随着技术的发展,视觉表征学习已经被广泛地应用于人工智能领域,良好的视觉表征学习技术能够更好地从数据中提取有用的特征,从而应用于下游任务。本文对视觉表征学习的研究现状以及应用进行了归纳总结,并展望了视觉表征学习的发展前景。

本文根据数据依赖程度和类型的不同,将视觉表征学习划分为预训练视觉表征学习、生成式视觉表征学习、对比式视觉表征学习、解耦式视觉表征学习,以及结合语言信息的视觉表征学习。前 4 种方法仅依赖视觉类型的数据,而结合语言信息的视觉表征学习则融合了文本类型的信息以学习表征。预训练视觉表征学习对数据的依赖程度较高,属于监督学习,而其他 3 种方法则属于无监督学习,其对数据的依赖程度较低。在无监督学习中,根据训练模型的方法不同,进一步将其分为生成式视觉表征学习、对比式视觉表征学习和解耦式视觉表征学习。

早期的视觉表征学习方法主要有主成分分析[8]、线性判别分析[9]、广义判别分析[10]和流形学习[11]。这些方法能够将高维数据映射为低维表示,为后续的发展奠定了基础。2006 年,Hinton 等[12]提出了神经网络算法,极大提升了神经网络学习视觉表征的能力。随着计算能力的提高和深度神经网络结构的不断发展,越来越多的研究者将深度神经网络应用于视觉表征学习中,并提出了许多预训练网络模型,如 VGG 和 ResNet 等,这使得有监督的视觉表征学习方法迅速发展。然而,这些训练网络需要大量的有标签的数据集,这限制了它们在实际场景中的应用。因此,生成式和对比式的无监督视觉表征学习方法相继出现,这些方法只需少量的标签数据,就可以在数据集上学习高质量的特征。

2012 年,Bengio 等[1]首次提出了基于解耦的表征学习,他们认为,在表征中每个维度对应于一个变化因子,当某个因子发生变化时,其他因子相对保持不变。简单来说,数据中某项特征的变化是由某个关键因素的变化引起的,与其他特征无关。解耦表征学习在模型可解释性和零样本学习等问题上具有很大优势,因此对解耦式视觉表征学习的研究也大量涌现,许多研究者通过改进网络模型,从视觉数据中分离出可解释性的潜在因子,生成更高质量的特征并将其用于下游任务。实际生活中,需要处理的信息不仅包括视觉信息,还有听觉信息、文本信息等,其中,结合语言信息的视觉表征学习是最有代表性的和最常见的形式。此外,很多语言信息中包含了对视觉任务有效的特征,如何提取语言信息中的有效特征并将其应用于视觉任务,对视觉表征学习的发展有很大帮助。在单模态表征学习的基础上,大量研究者对结合语言信息的视觉表征学习进行了研究,并取得了一定的进展。本文主要介绍视觉表征学习的方法和应用,以及解耦式视觉表征学习和结合语言信息的视觉表征学习。

本文第 2 章介绍了视觉表征学习中基于有监督预训练的方法;第 3—5 章分别介绍了无监督学习在视觉表征学习上的 3 种模式,即生成式视觉表征学习、对比式视觉表征学习和解耦式视觉表征学习;第 6 章介绍了结合语言信息的视觉表征学习方法;第 7 章和第 8 章介绍了视觉表征学习的评价准则和未来展望;最后,对本文进行了总结,概括了视觉表征学习的方法和应用。

2 预训练视觉表征学习

在视觉表征学习中,基于预训练模型的方法已经取代了早期的人工特征提取方法,成为了视觉表征学习的典型思路。其中,卷积神经网络(Convolutional Neural Network,CNN)是最为常用的预训练模型。1980 年,Fukushima[13]首次提出了一个包含卷积层和池化层的神经网络结构。基于此,Lecun 等[14]提出了 LeNet,该网络由输入层、卷积层、池化层、全连接层和输出层组成,有效学习到了用于手写数字识别的表征,为预训练模型在视觉表征学习上的应用提供了良好的开始。CNN 非常适合处理视觉表征学习,它可以通过卷积的方式从视觉信息中提取部分特定的表征信息,并且保留视觉信息的空间信息。2012 年,Alex 等[15]提出了 AlexNet。AlexNet 模型由 5 个卷积层、3 个池化层和 3 个全连接层构成,使用了更多的卷积层和更大的参数空间,能够更好地学习图片上的表征。AlexNet 在 ImageNet 大规模视觉识别挑战竞赛(ImageNet Large Scale Visual Recognition Challenge,LSVRC)中以较大的优势赢得了当年的冠军,使 CNN 在视觉表征学习和应用领域占据主导地位。

神经网络在下游任务中的性能取决于视觉表征学习的好坏,因此提升神经网络在下游任务中的性能,成为了提升视觉表征学习性能的关键。很多研究者通过改造网络模型的结构来提升预训练模型学习视觉表征的能力。2014 年,Simonyan 等[16]证明了增加网络深度会对网络最终性能产生一定的影响,并提出了 VGG 网络。VGG 网络采用连续的 3×3 的卷积核代替 AlexNet 中的较大卷积核,简化了网络结构。VGG 模型在 LSVRC14 竞赛中获得了图像分类“指定数据”组的第二名,这证明了深度在神经网络中的重要性,其提高了模型的性能。但是随着网络深度的加深,VGG 网络的参数也相应增加,这不利于网络训练。同年,Christian[17]提出了 GoogleNet 网络,该网络提出了 Inception 模块,可以轻松地添加和修改网络。Inception 模块采用并行结构,是一个超过 20 层的 CNN 结构,大大增加了 CNN 的深度。GoogleNet 采用了 3 种类型的卷积操作,大大减少了网络的参数。2015 年和 2016 年,GoogleNet 的发明团队在 GoogleNet 网络的基础上进行了修改,相继发布了新的版本 InceptionV3[18]和 InceptionV4[19],其不仅提高了网络的准确性,还减少了网络参数。GoogleNet 网络在 LSVRC14 竞赛中获得了图像分类“指定数据”组的第一名,大大提高了学习视觉表征的能力。

随着网络深度的加深,网络在反向传播时,梯度会越来越小,这会导致网络梯度消失。对此,He 等[20]提出了 ResNet 残差网络。残差网络引入了“快捷连接”技术,将输入跨层传递并与卷积的结果相加,使梯度传播可以跳过卷积层,即使网路层数达到了 1 000 层也可以继续训练,精度也随着网路深度的加深大大提高。ResNet 在 2015 年的 LSVRC 中获得了冠军。ResNet 网络的出现使预训练模型能够训练更深的

CNN 模型，从而实现更高的准确度，在提升视觉表征学习能力中起到了关键的作用。在 ResNet 和其他网络的影响下，Huang 等[21]从特征图入手，通过对特征图的极致利用提出了 DenseNet 网络。DenseNet 为了能够保证前馈的特性，每一层将之前所有层的输入进行拼接，之后将输出的特征图传递给之后的所有层，加强了特征的传递，更有效地利用了特征。另外，DenseNet 还应用了稠密连接模块，不再需要重新学习多余的特征图，因此它的参数比传统的卷积网络少。在参数和计算成本较低的情况下，DenseNet 的性能比 ResNet 更好，为视觉表征学习提供了更加有效的方法。

从 LeNet 到 DenseNet，视觉表征学习中的预训练模型都在向着提高精度的方向发展。除此之外，一些研究还专注于在不大幅降低模型精度的前提下，最大限度地提高运算速度，以加强网络学习视觉表征的能力。这种方法旨在减少模型训练和测试的计算量，缩小模型文件大小，更有利于模型的保存和传输。Landola 等[22]在 2016 年提出了 SqueezeNet。SqueezeNet 使用了 3 种模型压缩策略，使该网络能够在 ImageNet 数据集上达到与 AlexNet 近似的效果，但是参数量却只有 AlexNet 的 1/50。2017 年，Howard 等[23]提出了 MobileNet，该网络引入了由逐深度卷积和逐点卷积构成的深度可分离卷积，极大地减少了模型参数量，且在一定程度上仍保证了模型的性能。Zhang 等[24]利用组卷积和卷积随机提出了 ShuffleNet。ShuffleNet 通过将图像的特征通道进行分组，在组内进行卷积，并在某些层上，将不同通道的信息连接起来，从而减少网络的参数量。这些网络大大减小了网络模型的规模，更利于模型的部署和应用，促进了视觉表征学习在实际生活中的应用。

表 1 列出了各类网络结构在 ImageNet 数据集上 Top-1 的准确率和参数量。如今这些预训练的网络模型已成为视觉表征学习中的基础骨干网络，通过它们提取的原始特征能够在下游任务中取得更好的效果。

表 1 各网络在 ImageNet 数据集上 Top-1 的准确率及模型参数

Table 1 Top-1 accuracy and model parameters of each network on ImageNet dataset

网络	Top-1 准确率/%	参数量
AlexNet	62.50	60×10^6
GoogleNet	69.80	8.46×10^6
VGG-19	74.00	144×10^6
Inception V3	78.80	23.8×10^6
Inception V4	80.00	42.6×10^6
ResNet 152	78.60	60×10^6
DenseNet-201	77.42	20×10^6
SqueezeNet	57.50	4.8×10^6
MobileNet	74.70	6.9×10^6
ShuffleNet	75.40	7.4×10^6

3 生成式视觉表征学习

第 2 章介绍了基于监督学习方式的视觉表征学习。然而，在现实情况中，针对某些任务，无法获得这些带标注的数据集，这会导致监督学习所训练的网络效果较差。为了在没有标注的数据集中学习到有效的表征，许多研究工作致力于生成式视觉表征学习，其中自编码器（Autoencoder，AE）、生成对抗网络[25]（Generative Adversarial Network，GAN）和扩散模型为 3 种主流模型。自编码器是一种无监督的神经网络模型，能够学习输入数据的隐含特征，并且利用这些学到的新特征来重构原始输入数据。而生成对抗网络由生成器和判别器组成，生成器能够从噪声中生成类似于真实数据的合成数据，判别器则评估生成数据与真实数据之间的差异，以指导生成模型生成更真实的数据。扩散模型先通过正向过程将噪声逐渐加入到数据中，然后通过反向过程预测每一步加入的噪声，通过去掉噪声的方式还原得到无噪声图像。自编码器能够将高维的视觉信息转化为低维表示，从而学习视觉信息中的表征，并将其更好地应用在下游任务。生成对抗网络则通过生成与视觉信息类似的表征信息，来学习视觉信息中的表征。而扩散模型通过数据之间的相似性来推断或传播信息，学习视觉信息中的表征。这 3 种模型能够挖掘和学习视觉信息中的潜在特征，获得视觉信息中更紧凑的高层次特征表示，在表征学习方面具有显著的优势。

图 1 为经典自编码器的结构图。自编码器模型主要由两部分组成：一个由函数 $z=f(x)$ 表示的编码器和一个生成重构的解码器 $\tilde{x}=g(z)$。其主要目的是从输入 x 中提出隐藏编码 $z=f(x)$，然后将 z 重构成 $\tilde{x}=g(z)$，对比输入 x 和解码器产生的重构 $\tilde{x}$，使得它们两个尽可能相似，即在训练集上获得最小的重构损失 $L(x,\tilde{x})$，使编码器学习到与原视觉信息相似的表征，并将其用于下游任务。

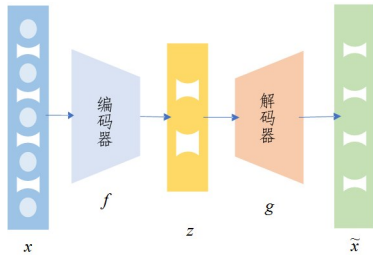


图 1 自编码器的结构

Fig. 1 Structure of autoencoder

图 2 为生成对抗网络的结构图。GAN 由两个网络生成器和判别器组成。生成器能将随机噪声生成相应的生成数据，再将生成数据输入到判别器中进行判别，输出是真实数据的概率。自编码器和 GAN 通过生成模型来学习视觉表征，在训练时只需原始视觉数据，就可进行视觉表征学习。这种无监督学习的方式使得自编码器和 GAN 的通用性得到大幅提升，为无监督的视觉表征学习提供了很好的方法。

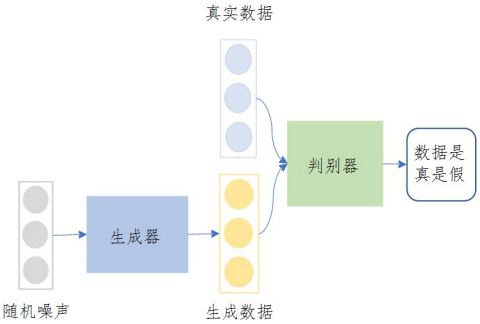


图 2 生成对抗网络的结构

Fig. 2 Structure of generative adversarial network

如图 3 所示，扩散模型主要可分为两个步骤：扩散过程和逆扩散过程。对于数据集中的任意图像的分布 $q(x_t)$，扩散过程 $q(x_t|x_{t-1})$ 指在前向的每一步对图像 x_{t-1} 添加高斯噪声得到 x_t，直到得到一个纯噪声的图像。通过扩散过程，未来任意 t 时刻的状态都可以利用 x_0 扩散得到。而逆扩散过程的目的是将随机噪声的分布逐渐去噪，直到得到真实的图像，即扩散过程的反转。下文将介绍自编码器、GAN 和扩散模型在视觉表征学习中的应用。

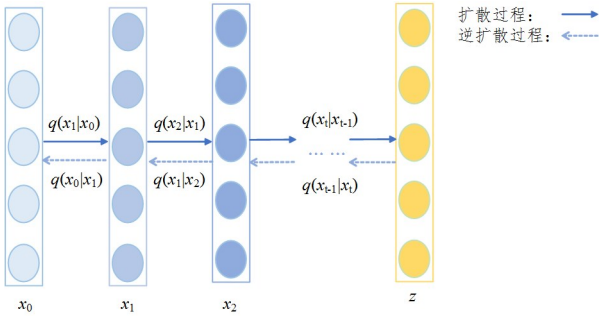


图 3 扩散模型结构

Fig. 3 Network structure of diffusion model

3.1 正则自编码器

从自编码器获得有用视觉特征的一种方法是限制隐藏编码 z 的维度小于输入 x 的维度。这种编码维度小于输入维度的自编码器被称为欠完备自编码器，它可以学习数据分布最显著的特征。但当隐藏编码的维数与输入相等，或隐藏编码的维数大于输入时，即处于过完备的情况时，自编码器将只执行复制任务，无法学习到任何有关数据分布的有用信息，从而无法提取数据中的视觉表征。正则自编码器通过损失函数，鼓励自编码器模型学习到其他视觉表征，而无需在浅层的编码器和解码器以及小的编码维数上来限制模型的容量。即使模型容量大到足以学习一个无意义的恒等函数，非线性且过完备的正则编码器仍然能够从中学习到有关数据分布的有用视觉表征。稀疏自编码器、去噪自编码器和收缩自编码器是 3 种最常用的正则自编码器。这些正则自编码器的设计旨在帮助自编码器学习到更加抽象和有用的数据表示，从而使其能够在视觉表征学习中发挥更为重要的作用。

3.1.1 稀疏自编码器

稀疏自编码器是由 Ng[26] 于 2011 年提出的。稀疏自编码在传统的自编码器的基础上对自编码器的隐层神经元增加了一些稀疏性约束。稀疏自编码器通过对隐层神经元输出的平均激活值进行约束，利用 Kullback-Leibler(KL)散度使其与一个给定的稀疏值相近，从而实现抑制效果。在训练时将编码层的稀疏惩罚与损失函数相结合，得到新的损失函数，即：

$$L=L(x,\tilde{x})+\beta\sum_{j=1}^{h}KL(\rho\|\tilde{\rho}_j) \tag{1}$$

其中，β 是用于控制稀疏惩罚的权重，取值范围为 0～1。

2016 年，Lin 等[27] 将稀疏自编码器和深度神经网络相结合来提取三维网格模型表面的三维关键点。他们利用稀疏自编码器提取多尺度空间中三维网格模型的局部信息和全局信息的内部结构，并为其制定高层特征，促进模型回归。另外，他们还使用 3 个稀疏自编码器来表示深度神经网络的隐藏层，然后训练逻辑回归层来处理第 3 个稀疏自编码器中产生的高级特征。实验表明，对于各种三维网格模型，该检测算法优于当前最先进的方法。稀疏自编码器在生成高维视觉数据的抽象特征方面取得了巨大的成功，但它没有考虑数据样本之间的关系，这会影响原始和新特征的实验结果。基于此，2018 年，Meng 等[28] 提出了一种同时考虑数据特征及其关系的关系自编码器模型，并将其扩展到了稀疏自编码器中。在基准数据集中，该模型都取得了较好的效果。与传统的自编码器相比，关系自编码器模型通过最小化数据和数据关系来重建损失，其定义如下：

$$L=(1-\alpha)\min_{\theta}L(x,\tilde{x})+\alpha\min_{\theta}L(R(x),R(\tilde{x}))+\beta\sum_{j=1}^{h}KL(\rho\|\tilde{\rho}_j) \tag{2}$$

其中，$R(x)$ 表示 x 中数据样本之间的关系，$R(\tilde{x})$ 表示 $\tilde{x}$ 中数据样本之间的关系，α 为比例参数，用于控制数据重建损失和关系重建损失的权重，θ 是自动编码器的神经网络参数。数据样本之间的关系通过相似性来表达，即 $R(x)$ 为 x 与 x^T 的乘积。在医学分类任务上，An 等[29] 于 2019 年利用稀疏自编码器提出了一种新的深度集成学习框架。他们运用稀疏自编码器融合多源数据，降低了原始特征之间的相关性，以此学习视觉信息中的表征，形成 3 个特征空间，并利用不同的学习算法和特征空间构建了广义分类器，在阿尔茨海默病分类的任务中取得了良好的效果。这显示了其在处理多源数据和改善特征表示方面的潜力。

3.1.2 去噪自编码器

当看到部分被遮挡或损坏的图像时，人类仍可以准确地识别图像中的物体。Vincent 等[30] 受这一现象的启发，提出了去噪自编码器(Denoising Autoender，DAE)。DAE 在传统自编码器的基础上，引入了一个损坏过程 $C(x'|x)$，这个条件分布代表给定数据样本 x 产生损坏样本 x' 的概率。在输入数据 x 中注入了噪声，使数据 x 变成了 x'，将 x' 作为编码器的输入并对其进行训练，来预测原始未被损坏的数据，如图 4 所示。通过编码器的输出来学习视觉信息中的表征。其中，去噪自编码器的损失函数 $L(x,\tilde{x})$ 中的 $\tilde{x}=g(f(x'))$。

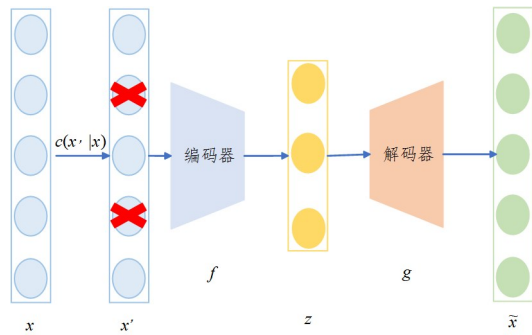


图 4 去噪自编码器的结构

Fig. 4 Structure of denoising autoender

2019 年，Gidaris 等[31] 提出使用去噪自编码器来解决少样本检测任务，使用了去噪自编码器网络来构建元模型。在训练期间，该网络以一组被高斯噪声破坏的分类权重作为输入，并学习重建目标-判别分类权重。在这种情况下，注入到

分类权重上的噪声起到了正则化权重生成元模型的作用，从而避免训练数据过拟合的危险，提升了模型学习视觉表征的能力。此外，Bo 等[32]通过将去噪自编码器以卷积的方式叠加，提出了堆叠卷积去噪自编码器。该网络通过分层训练方式来优化编码器，在每一层中，将下层的特征与去噪自编码器学习的卷积核进行卷积产生高维特征图，从而实现在没有任何标签信息的情况下学习视觉表征。2021 年，He 等[33]在去噪自编码器的基础上提出了掩蔽自编码器(Masked Autoencoders，MAE)。MAE 采用了非对称的编码器-解码器架构，编码器只应用于可见的补丁子集，即未被掩码的块上。再利用一个轻量级的解码器从潜在表示和掩码标记中重建原始图像。通过输入一个高掩蔽率(例如 75%)的图像，既能让编码器只处理小部分的补丁，从而减少内存消耗，提高训练速率，又能优化模型的精度。MAE 可以学习大容量模型并能更好地泛化，这些预先训练的表征可以更好地泛化到各种下游任务中。在 MAE 的基础上，Chen 等[34]提出了一种局部掩蔽重建模型，该模型只需一些局部视觉线索，就足以恢复丢失的信息，并将 MAE 中的重量级解码器替换为一个轻量级的多层感知器头，从而优化了局部掩蔽重建模型的性能。

3.1.3 收缩自编码器

收缩自编码器[35]使用了与稀疏自编码器类似的方式，在损失函数中添加一个惩罚项 $\Omega(z)$，但惩罚项的形式不同，如式(3)所示。

$$L=L(x,\tilde{x})+\Omega(z) \tag{3}$$

其中，$\Omega(z)=\lambda \left\| \frac{\partial f(x)}{\partial x} \right\|_F^2$；$\lambda$ 是用于控制惩罚项强度的超参数，可选择 0～1 之间的任意值。惩罚项 $\Omega(z)$为关于输入的隐层表达的 Jacobian 矩阵的 F 范数，这个惩罚项只对训练数据适用，它能够使特征空间在训练数据附近的映射达到收缩效果，从而让自编码器学习到可以反映训练数据分布信息的特征，但收缩自编码器只在局部收缩。收缩自编码器和去噪自编码器不同。前者是特征提取函数能抵抗极小的输入扰动，而后者能抵抗小且有限的输入扰动。在分类任务中，收缩自编码器在测试集的错误率上远远低于去噪自编码器的错误率。另外，Salah 等通过实验发现，使用堆叠的收缩自编码器的性能比单个收缩自编码器的性能更好。2022 年，Ganguli 等[36]利用一个卷积的收缩自编码器学习每个贴图可达性摘要的压缩表示，将其作为地理位置任务不可知的特征表示，进而解决下游的地理空间计算机视觉任务。

3.2 变分自编码器

变分自编码器(Variational Auto-Encoders，VAE)[37]是由 Kingma 等在 2014 年提出的基于变分贝叶斯推断的生成式网络结构。与传统的自编码器通过数值的方式描述潜在空间不同，它以概率的方式描述对潜在空间的观察，在数据生成方面表现出了巨大的应用价值。

自编码器将输入变量直接编码成隐藏层变量，再将其解码成输出变量。变分自编码器在“编码”的过程中，将输入变量“编码”成隐变量的分布，再从隐变量分布中采样，将隐变量分布“解码”成输出变量的分布，如图 5 所示。编码器负责根据输入变量建立隐藏变量的分布 $q_\phi(z|x)$，其中 ϕ 表示编码器中的所有参数。解码器负责根据从 $q_\phi(z|x)$中采样的数据，建立输出变量条件分布 $p_\theta(x|z)$，其中 θ 表示解码器的所有参数。

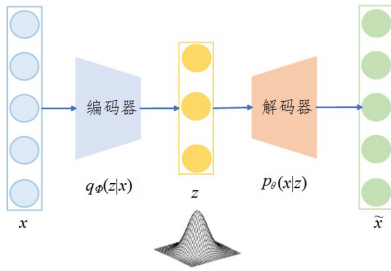


图 5 变分自编码器的结构

Fig. 5 Structure of variational auto-encoder

变分自编码器通过最大化与数据点 x 相关联的变分下界 L 来训练网络模型，如式(4)所示：

$$\begin{aligned} L &= \max_{\theta,\phi} \mathbb{E}_{z\sim q_\phi(z)}\left[\log \frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z)}\right] \\ &= \max_{\theta,\phi} \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\mathrm{KL}}(q_\phi(z|x)\,\|\,p_\theta(z)) \end{aligned} \tag{4}$$

其中，第一项表示在其他自编码器中出现的重构对数似然，第二项试图使近似后验分布 $q_\phi(y|x)$与模型先验 $p_\theta(y)$彼此接近。$p_\theta(y)$表示解码器中隐变量的先验分布；$q_\phi(y)$表示编码器中隐变量的先验分布。变分自编码器能将特征中各个属性的潜在因子分离开来，是很好的生成模型。

变分自编码器虽然能够很好地生成原来的图像，但仍无法对生成的数据类型进行控制，它只能生成与输入类似的输出数据，无法产生指定的数据。在 2015 年，Sohn 等[38]提出了条件变分自编码器。为了产生特定的数据，Sohn 等在变分自编码器的基础上引进了一个输出变量 c，得到了新的变分下界 L，如式(5)所示。通过最大化 L 来输出指定的数据类型。

$$\begin{aligned} L = \max_{\theta,\phi} &\mathbb{E}_{z\sim q_\phi(z|x,c)}[\log p_\theta(x|z,c)] - D_{\mathrm{KL}}(q_\phi(z|x,c)\,\| \\ &p_\theta(z|c)) \end{aligned} \tag{5}$$

变分自编码器在学习视觉表征中有很好的效果，还有很多研究者提出了多种另外的改进方法。文献[39-43]通过在损失函数上添加一个新的约束来改进变分自编码器。有研究者[44-45]通过将变分自编码器与自回归模型相结合，来优化变分自编码器，如卷积神经网络、循环神经网络等。还有通过修改变分自编码器的结构来学习有用的表征。比如 Razavi 等[46]通过将解码阶段分为两部分，来更好地优化编码器，学习更好的视觉表征，从而生成更加清晰的图像；Rueckert[47]通过对比目标来增强原始 VAE，以最大化相似视觉输入表征之间的互信息，从而解决了变分自编码器中出现的“文本坍塌”的问题。另外，变分自编码在解耦式视觉表征学习中也有非常广泛的应用，这一部分内容将在第 5 章介绍。

3.3 生成对抗网络

生成式对抗网络(Generative Adversarial Network，GAN)是视觉表征学习的另一类重要的模型。Goodfellow 等[25]在 2014 年提出了原始 GAN 结构，包含一个生成器 G 和一个判别器 D。生成器的任务是生成自然真实、与原始数据相似的实例；而判别器则被训练用于区分真实数据样本和由

生成器生成的样本，并给出了样本是真实训练样本而不是伪造样本的概率。GAN 模型的目标函数如下：

$$\min_G \max_D V(D,G)=\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)]+\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{6}$$

通过目标函数的训练，可以使判别器 D 正确地对样本进行分类，将其判别为真实数据或伪造数据，同时使生成器 G 生成让判别器误认为是真实数据的数据样本。在训练过程中，通过交替迭代的方式固定其中一个网络，更新另一个网络的参数，以此来训练模型。这种交替的训练方式使得编码器能够学习到视觉信息中更有意义的表征。

生成对抗网络是一种在不需要大量标注训练视觉数据的情况下，能够很好地学习视觉表征的模型。Denton 等[48]将 CNN 与 GAN 相结合，提出了拉普拉斯金字塔形对抗网络(LAPGAN)来提取更有效的视觉表征。LAPGAN 使用多尺度分解生成过程的形式，将真实图片分解成为拉普拉斯金字塔，并训练一个条件性卷积 CNN 来生成每一层图像，这种方法使得 LAPGAN 能够学习高质量的视觉表征，从而生成更高质量的自然图像样本。此外，Radford 等[49]于 2015 年提出了深度卷积生成对抗网络(DCGAN)。DCGAN 采用带步长的卷积和小步长卷积，在训练过程中学习空间下采样和上采样算子，通过这些算子来处理采样率和位置的变化，提升网络学习视觉表征的能力。DCGAN 网络结构在之后的各种改进的 GAN 中有广泛的应用，可将其视为各类改进 GAN 的前身。在 DCGAN 的基础上，很多研究者通过修改其训练策略[50-51]、改进模型[52-55]，提出了更多的网络模型，从而学习到更好的视觉表征，并生成更高质量的图像。

为了生成更多形式的数据和学习特定的表征，文献[56-60]将特定的条件信息 C 作为网络的输入，提出了条件 GAN，使网络能够生成符合条件信息的不同形式的数据。另外一些研究人员通过修改 GAN 的损失函数得到高训练稳定性的 GAN。例如，Arjovsky 等[61]使用 Wasserstein 距离来衡量两种分布之间的距离，将其加入到损失函数中，解决了两种分布之间没有重叠的情况，提高了训练的稳定性。此外，部分研究人员将自编码器和生成对抗网络相结合，提出了对抗自编码器[62-63]，用于学习更有效的视觉表征。该方法在解耦式视觉表征学习中应用广泛，将在第 5 章中详细介绍。

3.4 扩散模型

扩散模型(Diffusion Model)是一类概率生成模型，以序列化的方式生成图像。这些模型使用递归过程生成图像，每一步都通过引入一些噪音来改变先前的图像。这个过程重复多次，逐步生成完整的图像。Sohl-Dickstein 等[64]于 2015 年首次提出了扩散模型，其主要目的是消除训练图像在连续应用过程中产生的高斯噪声，可将其视为一系列去噪自编码器的组合，通过训练编码器将图像转换为低维潜在空间，学习图像中的表征信息。

2020 年，Ho 等[65]将扩散模型应用于图像生成任务中，提出了去噪扩散概率模型(Denoising Diffusion Probabilistic Models，DDPM)。DDPM 在逆扩散过程中通过拟合噪声的均值和方差来预测添加的噪声，将在 t 时刻的状态减去添加的噪声来得到 $t-1$ 时刻的形态，其采用了一个 U-Net 的网络结构，并且在逆扩散过程中参数是共享的。Nichol 等[66]对 DDPM 网络模型进行了改进，将用常值表示的拟合噪声方差改为用模型学习得到，并将添加噪声的时间表从线性改为余弦的方式，使其能够更好地学习低分辨率图片的特征。受 GAN 的实验的影响，Dhariwal 等[67]在原始的 U-Net 网络上添加了一个单头全局注意力模块，从模型的宽度、深度、注意头的数量及注意力的分辨率等方面出发，对扩散模型进行了大量的消融实验，找到了更深、更宽、架构更好的模型。

虽然扩散模型能够很好地生成高质量的样本，但其潜在变量通常缺乏语义，在表征学习中的表现仍不是很好。对此，Wang 等[68]提出了 InfoDiffusion 模型，该模型利用观测变量和隐藏变量之间的相互信息、正则化的学习目标来提高潜在空间的质量，防止解码器忽略潜在空间。另外，该模型还使用一种能够捕捉数据中高级变化因素的低微潜在变量来增强扩散模型算法，促进模型学习更高质量的表征。而 Yang 等[69]通过提出一种新的知识提取方法来增强模型学习视觉表征的能力，其利用预先训练的扩散概率模型来增强识别任务，并建立扩散概率模型和去噪自编码器之间的关系来提高模型的学习能力，在多个任务中证明了该方法的有效性。扩散模型相比自编码器和生成对抗网络，网络模型的速度更快，并且学习到的视觉表征质量更高。

3.5 性能表现

各生成模型在不同下游任务中展现的性能如表 2 所列。正则自编码器和变分自编码器主要应用在图像检测、分类和分割等任务中，而生成式对抗网络和扩散模型则主要使用在图像生成领域。随着对生成式网络框架的深入研究，其在各个下游任务中的性能逐步提升，这表明了生成式网络框架在学习视觉数据中高质量表征信息方面具有卓越的潜力，能够更好地适用于表征学习。表 2 中，对数似然、负对数似然和 FID 是衡量图像生成任务好坏的重要指标。对数似然和负对数似然能够很好地反映网络的重建能力，而 FID 评价指标则量化了真实图片与生成图片在特征空间之间的距离，较低的 FID 值代表着更高水平的图片质量和多样性。

表 2 各生成模型在不同下游任务中的性能统计

Table 2 Performance statistics of each generative model in different downstream tasks

模型	网络模型	下游任务	数据集	评价指标	评价指数
正则自编码器-稀疏自编码器	DNN[27]	3D 关键点检测	3DInterestPoint[70]	IOU	0.275
正则自编码器-稀疏自编码器	RAE[28]	图像分类	CIFAR-10[71]	错误率	13.40%
正则自编码器-稀疏自编码器	DELearning[29]	医学图像分类	NACC 数据集[72]	准确率	87.00%
正则自编码器-去噪自编码器	wDAE-GNN[31]	少样本图像分类	ImageNet-FS[73]	准确率	48.00%
正则自编码器-去噪自编码器	MAE[33]	图像分类	ImageNet-1K[74]	准确率	83.60%
正则自编码器-去噪自编码器	LoMaR[34]	图像分类	ImageNet-1K	准确率	84.10%
变分自编码器	CVAE[38]	图像分割	Caltech-UCSD Birds(CUB)[75]	IOU	98.52

（续表）

模型	网络模型	下游任务	数据集	评价指标	
变分自编码器	VQ-VAE[45]	图像分类	ImageNet	准确率	54.83
变分自编码器	CR-VAE[47]	图像生成	CIFAR-10	负对数似然	62.34
生成对抗网络	LAPGAN[48]	图像生成	CIFAR-10	对数似然	−1799
生成对抗网络	DCGAN[49]	图像分类	CIFAR-10	准确率	82.80%
生成对抗网络	GANU[54]	少样本图像分类	CIFAR-10	准确率	77.00%
扩散模型	DDPM[65]	图像生成	CIFAR-10	FID	3.17
扩散模型	Improved Diffusion[66]	图像生成	ImageNet	FID	2.92
扩散模型	ADM[67]	图像生成	ImageNet	FID	2.07
扩散模型	InfoDiffusion[68]	图像生成	CIFAR-10	FID	31.50

4 对比式视觉表征学习

无监督视觉表征学习的另一种方法是利用对比学习思想。对比式学习着重于学习同类实例间的共性特征，同时区分非同类实例间的差异。其目标是训练一个编码器，使得对于正样本数据，其编码相似度较高，而对于负样本数据，其编码差异尽可能大。相比生成式学习，对比式学习不需要过多关注样本的细节，而是在抽象的语义级特征空间上学习数据的区分性特征。因此，对比式学习的模型和优化相对简单，并且在泛化能力上更强。在所有的对比式视觉表征学习的框架中，编码器是下游任务的核心部分，其他组件和损失函数则辅助训练编码器，以产生更优质的视觉表征。本章从训练过程中是否使用负样本的角度出发，将对比式视觉表征学习分为基于负例的对比式视觉表征学习和基于正例的对比式视觉表征学习。

4.1 基于负例的对比式视觉表征学习

文献[76]指出一个好的对比学习系统应该具有两个属性：对齐性和均匀性。对齐性表示在特征空间中，正样本之间距离应该较近；均匀性表示特征向量应该保留尽可能多的信息。这等价于特征向量尽可能均匀地分布在特征空间汇总。而当所有的数据都被映射到特征空间中的同一个点时，就会出现模型坍塌的问题，此时所有数据的信息都被丢掉，数据极度不均匀地分布在了特征空间中，违背了均匀性。因此，解决模型坍塌问题是对比学习的关键。在基于负例的对比式视觉表征学习中，主要通过选取尽可能多的负样本和合理的损失函数来解决模型坍塌问题。如何选取正负样本，并在每一次训练中尽可能多地使用负样本，成为了基于负例的对比式视觉表征学习的一个关键。

在基于监督学习范式的 ImageNet 分类结果中可以发现，top-5 的错误率总是低于 top-1 的错误率，并且第二大响应的类别与实际类别有更多的视觉相似性。这表明视觉相似性是通过数据本身学习得到的，而不是语义标签。基于这一发现，Wu 等[77]引入了在个体级别的无参分类器，提出了 InstDisc 网络模型。它将每一张图片当作一个类别，学习一个实例级别的分类器，使该分类器能够捕捉到实例间相似性的特征表达。通过无语义类别标签的无监督学习，获得区分单个个体相似度的特征表示，在一个度量空间中拉近与正样本之间的距离，推远负样本之间的距离。

如图 6 所示，Wu 等[77]使用一个主干网络将每个图像编码为特征向量，将其投影到 128 维空间并进行 L2 归一化，将获得的特征使用无参数的 Softmax 函数对其进行分类，并创建了一个存储体来存储所有图片的 128 维特征。在训练时，随机从存储体中抽取 4 096 个负样本进行对比学习，通过 NCE 损失更新主干网络和存储体，使训练样本的特征最大限度地分散在 128 维单位球上。通过实验发现，训练过程中负样本的数量越多，模型学习到的视觉表征越有效，下游任务的效果就越好。

InstDisc 网络虽然能很好地学习到视觉表征，但需要额外的数据结构来存储视觉特征，因此效率较低。Ye 等[78]提出了一种在低维的度量空间中学习一个有效的样本相似度测量的方法。他们不使用额外的数据结构，采用端到端的方式在一个训练批量中进行对比学习。对于含有 N 张图片的训练批量，先将所有的数据进行数据增强，得到 $2N$ 张图片，再选取一张图片和数据增强后的这张图片当作正样本对，将剩下的 $2N-2$ 张图片作为负样本，将所有样本放入一个孪生网络进行对比学习，使得在一个度量空间中正样本之间的距离更近，负样本之间的距离更远，优化编码器学习高质量的视觉特征。

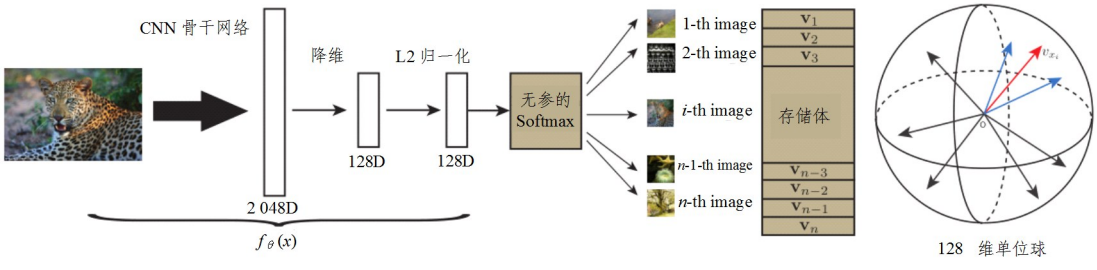


图 6 InstDisc 的网络结构[77]

Fig. 6 Network structure of InstDisc[77]

InstDisc 网络模型使用了存储体来保存负样本的特征，但其中存储图像的特征都是过去的编码器编码的特征，会导致采样的特征具有不一致性。鉴于此，He 等[79]从将对比学习作为字典查找的角度，构建了一个带有队列和移动平均编码器的动态字典，即构建一个既大又能保持特征一致性的字典来代替存储体，提出了动量对比法（MoCO）用于无监督视觉表征学习。

图 7 为 MoCo 的框架图。该模型定义了一个查询值 q 和

一个队列 k。在队列中，一般包含一个查询 q 的正样本和多个负样本。再通过对比损失函数来学习特征表示。MoCo 使用了一个队列来存储和采样负样本，队列中存储了多个近期用于训练批量的特征向量。当新的训练批量进入队列后，队尾的训练批量出队列，以此来更新队列。在训练时，将基准点样本记为 x^{quary}，经过编码器网络 f_q 对其进行编码得到 q。随后从队列中采样 $K+1$ 个样本 $\{k^0,\cdots,k^K\}$ 作为值，这些值通过不同的队列编码器网络 f_k 进行编码得到。编码器 f_k 采用基于动量更新的方式进行更新，这使得 f_k 的变化变得非常缓慢，也保证了队列中的值即使是由不同的编码器编码而成，数据特征仍具有一致性。MoCo 在多个下游任务中进行了实验，都取得了与有监督学习相近甚至优于有监督学习的结果。

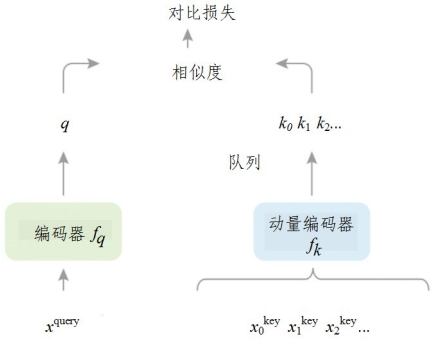


图 7 MoCo 的网络结构[79]

Fig. 7 Network structure of MoCo[79]

在文献[78]的基础上，Chen 等[80]于 2020 年提出了一种新的对比学习视觉表征框架 SimCLR。SimCLR 不需要专门的体系结构和内存库存储负样本，就能实现对比式视觉表征学习，如图 8 所示。

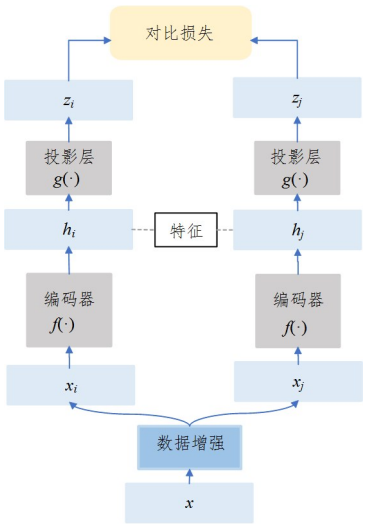


图 8 SimCLR 的网络结构

Fig. 8 Network structure of SimCLR

对于每一个实例，随机使用两种数据增强方法产生两个对应的视图 x_i 和 x_j，将这一对视图作为一对正样本，训练时将训练批量内的其他任意图像作为负样本。Chen 等通过大量的实验发现，随机裁剪和颜色失真是两种效果最好的数据增强方式。使用基编码器 $f(\cdot)$ 从增强后的数据中提取表征向量，得到 h_i 和 h_j。再使用一个非线性变换结构投影层 $g(\cdot)$，将表示映射到使用对比损失的空间中，$g(\cdot)$ 由单层的隐藏层 MLP 构成。为了使空间中的正样本之间的距离更近，负样本之间的距离更远，SimCLR 使用了 InfoNCE[81]损失函数来训练网络模型，使编码器学习更好的视觉特征。

SimCLR 的实验结果表明，对实例进行数据增强，并使用两次非线性映射的策略对无监督的视觉表征学习非常有效。于是，Chen 等[82]在 MoCo 的基础上添加了数据增强和 MLP 层，并使用了余弦衰减学习率，提出了 MoCo V2，大大提高了 MoCo 提取视觉表征的能力。另外，Chen 等[83]在 SimCLR 的模型框架中加入了 MoCo 的动量编码器，并使用 152 层的 ResNet 网络和 SK 模块，提高了编码器的网络规模和 MLP 层的深度，大大优化了 SimCLR 在对比式视觉表征学习中的效果。很多研究在选取正负样本的方式上提出了新的想法。文献[84-85]将上下文中提取到的表征信息和未来时刻样本的表征信息作为正负样本进行对比学习，获得了最能预测未来的关键表征信息，提出了对比预测编码(CPC)模型[84]。还有研究者将同一场景的多个视角图作为正样本，将不同场景的视图作为负样本，提出了多视角对比学习框架(CMC)[86-88]。这些框架在选取正负样本时都做了修改，并取得了很好的效果。但上述方式在选取负样本时，都是在整个数据集或者一个训练批量中选择负样本，这会导致一些问题。例如，这可能会重复抽取到同一数据，既把它当作正样本，也把它当成了负样本；也可能抽取的数据不具有整个数据集的代表性。对此有研究者提出在模型训练过程中引入聚类[89-91]来提高模型学习视觉表征的能力。图 9 给出了 Caron 等[89]提出的 SwAV 模型结构。其中，图像增强、编码器结构以及投影层结构与 SimCLR 基本保持一致。不同的是，SwAV 模型加入了一个聚类模块，使编码器获得的表征与聚类中心进行对比学习，并使用换位预测的方法，对 SwAV 模型进行训练，大大提高了模型学习视觉表征的能力。

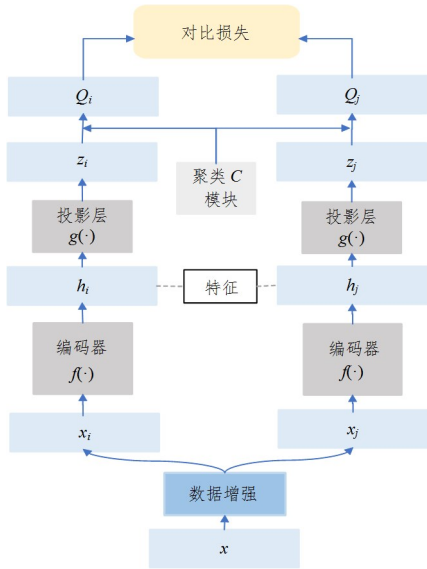


图 9 SwAV 的网络结构

Fig. 9 Network structure of SwAV

随着视觉转换器(ViT)在视觉领域的不断发展，用 ViT 主干来代替 CNN 主干的方法也成为了一种趋势。但不稳定

的 VIT 训练可能会导致更差的实验结果。于是，He 等[92]改进了 MoCo，提出了 MoCo V3。通过对梯度变化的经验观察，MoCo V3 在训练时冻结了 ViT 中的补丁投影层，使用固定的随机补丁投影解决了训练的不稳定性问题。Caron 等[93]则提出了一种新的自监督学习方式，没有标签的自蒸馏框架。该框架由两个不同的编码器（学生编码器和教师编码器）组成：对于教师网络，采用动量更新的方式更新，并在编码器的后面加了一层中心层，以一个批量为单位来计算教师网络输出的平均值，再对输出进行归一化，从而避免了模型坍塌，优化了模型学习视觉表征的能力。

4.2 基于正例的对比式视觉表征学习

在 4.1 节介绍的对比视觉表征学习方法中，负样本起着至关重要的作用。这些方法通过缩小与正样本之间的距离，并增加与负样本之间的距离来进行训练，以避免视觉表征学习中出现的坍塌问题。然而，这些方法需要投入大量的成本来处理负样本，如需要大规模数据、专门的数据结构等。因此，只利用正样本来学习视觉表征的模式成为了新的研究热点。从形式上看，如果只使用正例而不使用负例来训练对比学习模型，模型会推动正例在表示空间内相互靠近。然而，如果只有这一优化目标，模型很快就会收敛到常数，将所有数据映射到表示空间中的同一点，从而很容易出现模型坍塌的问题。尽管存在这些挑战，仍有研究者在此基础之上提出了新的模型框架，一定程度上解决了这个问题。

Grill 等[94]在只使用正样本的情况下，提出了 BYOL（Bootstrap Your Own Latent）模型，如图 10 所示。该模型由两个分支组成：在线分支和目标分支。在线分支的编码器和投影层与其他对比学习模型一样，但在投影层之后，新增了一个非线性变换模块 Predictor，该模块与投影层模块类似，是一个 MLP 映射网络，由 Linear->BN->ReLu->Linear 构成。目标分支由编码器和投影层组成，类似于 MoCo V2 中的动量更新结构，即与在线分支不共用参数，且模型的参数也不参与梯度更新。对于任意一张图片，使用数据增强的方法产生两组增强图片的视图 x_1 和 x_2，将这两组视图作为正样本，并分别放入两个分支中，以提取特征向量。BYOL 的优化目标是使在线分支部分的正例在表示空间中向目标分支侧对应的正例靠近。首先计算两个特征向量之间的距离，得到损失函数 $L1$；再将 x_1 和 x_2 交换位置，重复上诉操作，得到新的损失函数 $L2$；最后将 $L1+L2$ 作为最终的损失函数来训练网络优化编码器学习视觉表征的能力。Grill 等认为 BYOL 在训练时没有发生模型坍塌，主要是由这两个分支的结构不对称造成的。也有其他研究[95-96]认为，BYOL 没有发生坍塌是由于在非线性变换模块中使用了批量归一化，批量归一化中采用的训练批量内统计量起到了类似负例的作用。但是很快，Grill 等进行了反驳[97]，把非线性变换模块中的批量归一化替换成组归一化和权重标准，这样使得非线性变换模块无法知晓训练批量内的信息，但是同样可以达到与采用批量归一化类似的效果，这说明并非批量归一化在起作用。BYOL 只用正例进行对比学习不会发生期望中的模型坍塌，其原因目前仍还未有定论，但是可以推导出主要是由存在非线性变换模块结构造成的。

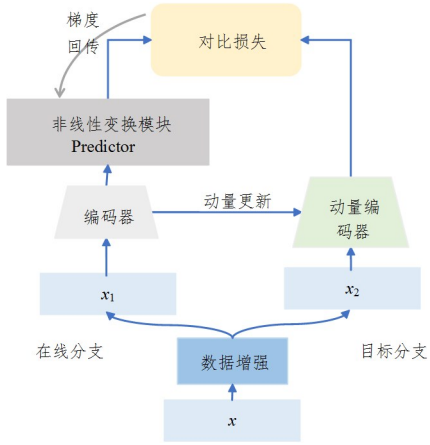


图 10 BYOL 网络模型

Fig. 10 Network structure of BYOL

如图 11 所示，基于各项研究，Chen 等[98]使用了一种新的解决模型坍塌问题的方法，提出了 SimSiam 网络框架。SimSiam 去除了 BYOL 的动量更新机制，并且使用一个孪生网络作为编码器来学习视觉特征，上下分支使用参数共享的编码器。在不使用负样本、动量编码器的情况下，SimSiam 使用停止梯度回传的机制，解决了模型坍塌的问题。另外一种解决模型坍塌问题的方法是使用新的损失函数。Zbontar 等[99]在只使用正例的情况下，提出了新的 Barlow Twins 网络模型。与 BYOL 不一致的是，该网络模型采用了上下分支对称结构，且两者参数共享。为了防止模型发生坍塌，Barlow Twins 使用了“冗余消除损失函数”来解决这个问题。另外，Barlow Twins 在增强图片经过编码器之后，对两例正例分别做了批量归一化的正则，并对两个正例的表示矩阵做矩阵乘法，求出两者的互相关性矩阵，通过这两个矩阵来计算损失函数。该方案也说明了可以通过修改损失函数来解决模型坍塌的问题。

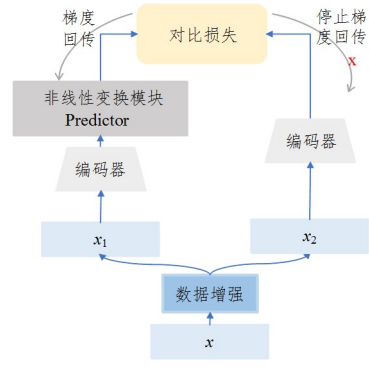


图 11 SimSiam 网络框架

Fig. 11 Network structure of SimSiam

利用海量的无标注图像数据，根据对比学习指导原则，学习出好的编码器模型，使其能够学习好的视觉表征，是目前所有对比式视觉表征学习的一个关键。对于输入图像，一个好的编码器能学会并抽取出关键视觉特征。在解决下游具体任务时，它能用学到的参数初始化编码器中的主干网络，用下游任务标注数据来微调模型参数，利用预训练阶段学到的知识对下游任务进行迁移。在基于负例的对比式视觉表征学习框架中，主要研究的问题是如何有效地选择正负样本来提高

模型学习视觉表征的能力。这个问题主要通过大的训练批量、特殊的数据结构和动量更新的方式来解决。而在基于正例的对比式学习表征学习框架中,在不使用负例的情况下,如何训练模型从而使模型不发生坍塌成为了关键。本文主要介绍不对称结构、停止梯度回传和修改损失函数这 3 种方法。表 3 列出了各个对比式视觉表征学习框架在 ImageNet 数据集上的 Top-1 准确率和使用的主干网络。可以发现,随着 MoCo 和 SimSLR 的提出,模型的准确率得到大大提高,模型学习视觉表征的能力也有所提高。很多网络模型也沿用了这两个框架所提出的方法。另外,随着 Transformer 在视觉中的应用,也进一步提高了对比式视觉表征学习框架学习表征的质量。目前,已经有很多网络模型在无监督学习的情况下达到与监督学习相同的效果,甚至优于监督学习的效果。

表 3 各个对比式视觉表征学习框架在 ImageNet 数据集上的 Top-1 准确率

Table 3 Top-1 accuracy of each contrastive visual representation learning framework on ImageNet dataset

框架	主干网络	是否使用负例	Top-1 准确率/%
监督网络	Resnet 50	是	79.3
InstDisc	Resnet 50	是	54.0
CPCv2	Resnet 50	是	63.8
MoCov1	Resnet 50	是	60.6
MoCoV2	Resnet 50	是	71.1
SimCLRv1	Resnet 50	是	70.0
SimCLRv2	Resnet 50	是	77.5
SwAV	Resnet 50	是	75.3
MoCoV3	ViT-S	是	72.7
DINO	ViT-S	是	77.0
BYOL	Resnet 50	否	74.3
SimSiam	Resnet 50	否	71.3
Barlow Twins	Resnet 50	否	73.2

5 解耦式视觉表征学习

解耦表征学习最早是由 Bengio[1] 提出的,其思想是数据集中的样本是由不同的因素生成的,每个隐单元对应一个生成因子,而其他的生成因子保持不变。解耦表征学习对影响数据形态的关键因素进行建模,通过改变某一关键因素来改变数据在某项特征上的变化,即这一因素的变化不会引起其他特征的变化。本章主要介绍了在视觉领域的解耦表征学习。基于第 3 章的介绍,变分自编码器和生成对抗网络被认为是解决解耦式视觉表征学习的关键技术,因此本章将着重介绍变分自编码器、生成对抗网络以及两项技术的结合在解耦式视觉表征学习中的应用。

5.1 基于 VAE 的解耦表征学习

变分自编码器在第 3 章已经进行简单的介绍,VAE 模型通过极大对数似然思想来优化网络中的参数 θ,ϕ。如式(4)所示,第一项 $\log p_\theta(x|z)$ 表示数据的条件对数似然,反映了潜在变量 z 对于真实数据 x 的表征能力,第二项为 KL 项,反映了变分后验分布 $q_\phi(z|x)$ 与先验分布 $p_\theta(z)$ 间的相似性。$p_\theta(z)$ 是人为选择的先验分布,通常满足独立特性,如高斯正态分布等。因此,KL 项相当于对网络施加了一定程度的独立性约束,导致训练出的网络模型具有一定的解耦性。但在实际应用过程中,VAE 的解耦能力还远不能实现对视觉数据的有效解耦。基于此问题,大量的研究通过在原始 VAE 的基础上增添各类无监督正则项归纳偏好,促使网络学习数据内部各个可解释生成因子的有效解耦表征。

2017 年,Higgins 等[100] 在 VAE 的基础之上引入了一个可调超参数 β,提出了一种新的无监督框架 β-*VAE*,在 VAE 的变分下界 L 中在 KL 项上加入了一个超参数,构成了新的变分下界 L,如式(7)所示:

$$L_{\beta\text{-VAE}}=\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]-\beta D_{\mathrm{KL}}(q_\phi(z|x)\parallel p_\theta(z)) \tag{7}$$

其中,$\beta\geqslant 1$。超参数 β 限制了潜在瓶颈的有效编码能力,能够促进潜在表征的更多因式分解。当 $\beta>1$ 时,数据中至少包含一些独立的潜在变化因素,会推动模型学习视觉数据中更有效的潜在表示,且可以将其解开。并且 β-VAE 在更加复杂的数据集中都能取得不错的效果。但当 β 的值过大时,β 值的额外压力会导致潜在变量中的有效信息在经过解码器时高频细节丢失,出现解耦效果好但重构效果差,或解耦效果差但重构效果好的现象,无法达到数据表征和解耦表征之间的最佳权衡。基于此,Burgess 等[101] 用信息瓶颈理论分析了这个问题产生的原因,认为在式(7)中第二项近似后验分布的约束项为第一项重构项的信息瓶颈。因此,在训练过程中,Burgess 等采用渐进策略逐渐增加潜在变量的信息容量,来更好地权衡强表征能力与强解耦能力,使潜在变量的表征空间更大,如式(8)所示,$L_{\beta\text{-}VAE}$ 为新的变分下界 L。

$$L_{\beta\text{-VAE}}=\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]-\gamma|D_{\mathrm{KL}}(q_\phi(z|x)\parallel p_\theta(z)-C)| \tag{8}$$

其中,γ 为随着网络训练不断线性增大的超参数。

2018 年,Kim 等深度分析了原始 VAE 的 KL 散度项,如式(9)所示:

$$\mathbb{E}_{p_{\mathrm{data}}(x)}[KL(q(z|x)\parallel p(z))]=I(x;z)+KL(q(z)\parallel p(z)) \tag{9}$$

其中,$I(x;z)$ 为联合分布 $p_{\mathrm{data}}(x)q(z|x)$ 下 x 与 z 的互信息。$KL(q(z)\parallel p(z))$ 表示使 $q(z)$ 与 $p(z)$ 更接近,使 z 的各维之间更独立。他们发现,β-VAE 使用超参数 β 来乘以 KL 散度,这同时惩罚了散度项 $KL(q(z)\parallel p(z))$ 和互信息项 $I(x;z)$,导致强表征能力和强解耦能力无法达到权衡。于是,Kim 等使用了一个新的能够直接鼓励后验累积分布 $q(z)$ 服从因式阶乘分布的惩罚项,提出了 Factor-VAE[102]。Factor-VAE 直接将这惩罚项添加到原始 VAE 的优化函数中,得到了新的优化函数,如式(10)所示,进一步改善了模型在强表征能力与强解耦能力之间的权衡关系。

$$L_{\text{Factor-VAE}}=\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]-D_{\mathrm{KL}}(q_\phi(z|x)\parallel p_\theta(z))-\gamma KL(q(z)\parallel\prod_{i=1}^{d}q(z_i)) \tag{10}$$

其中,$KL(q(z)\parallel\prod_{i=1}^{d}q(z_i))$ 称为全相关惩罚系数 TC(Total Correlation Penalty),能促进表征 z 的每一维度之间尽可能独立,提高了解耦效果。而 Tian 等将式(9)中的第二项 $KL(q(z)\parallel p(z))$ 进行了进一步分解,通过给各个正则项赋予不同的权重得到新的优化函数,提出了新的模型 β-TC-VAE[103]。这两种模型都采用对抗方式来求解 TC 项的相似性度量。虽然这些模型很好地解决了强表征能力与强解耦能力的权衡问题,但在潜在变量子集中仍存在一些无关噪声的干扰,无法区分有意义的潜在变量和讨厌的潜在变量,这可能导致解耦性能的下降。Kim 等在 Factor-VAE 的基础上引入

了相关性指标 r，提出了 RF-VAE[104]，使得 Factor-VAE 中的惩罚项只作用在对数据有用的相关潜在变量中，进一步提高了模型学习解耦视觉表征的能力。

Chen 等[105]通过实验发现，当条件分布充分表达时，潜在变量往往会被忽略，模型只使用了单个条件分布组件对数据进行建模，VAE 的混合建模能力未被充分利用。因此，Zhao 等引入了新的训练目标，来权衡正确推断和拟合数据分布之间的偏好，并指定模型对潜在变量的依赖程度，提出新的优化模型 InfoVAE[106]，如式(11)所示。为了促进解耦学习，他们重新考虑了式(11)中最后一项的权重值，并提出可使用 Jensen-Shannon 散度来替代最后一项进行训练。另外，Kumar 等[107]还建议在优化训练目标时，将潜在变量后验累积分布 $q_\phi(y)$ 与先验分布 $p(y)$ 假设为高斯分布，利用了矩估计思想设计两种矩匹配项来对后验分布的协方差矩阵进行约束，从而解决了在对抗训练中出现的鞍点问题。另外，Han 等[108]基于可逆列网络提出了 RecColV2 模型，首次在掩码图像建模(Masked Image Moding，MIM)视觉预训练中实现了特征解耦学习，统一了上下游任务的网络结构，在多个任务中都展现了较好的结果。

$$L_{\text{InfoVAE}}=\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]+\lambda_1 D_{\text{KL}}(q_\phi(z|x)\parallel p_\theta(z))+\lambda_2 KL(q_\phi(z)\parallel p(z)) \tag{11}$$

上述这些模型能够实现较好的解耦式视觉表征学习，也比较灵活，但难以添加条件独立性等结构性限制。基于此，Lopez 等利用希尔伯特-施密特独立性准则[109](Hilbert-Schmidt Independence Criterion，HSIC)来加强潜在表示和任意干扰因素之间的独立性，提出了 HSIC-VAE[110]。与上述方法不同，该模型除了控制潜在变量的依赖关系外，还可以通过添加的正则化器从潜在表示中移除敏感信息，从而提高解耦能力。此外，Esmaeili 等在 2019 年从多级隐变量的角度，提出了基于 VAE 的两级分层 HFVAE 模型[111]，HFVAE 将 KL 项进一步分解为组间潜在变量和组内潜在变量，通过对不同正则化项进行重新加权，来控制组内隐变量和组间隐变量之间不同程度的解耦程度，从而学习更高质量的解耦视觉表征。

5.2 基于 GAN 的解耦表征学习

传统的生成对抗网络能够有效地学习到语义特征，但在 GAN 中，没有规则来引导生成模型生成数据，很难学习到可解释性的有价值的特征，使其无法被应用到解耦式视觉表征学习中。为此，在 2016 年 Chen 等[112]提出了一种信息理论与 GAN 相结合的生成式对抗网络 InfoGAN。InfoGAN 将隐含规则变量 c 和随机变量 z 作为生成模型 G 的输入，提出最大化隐含规则变量 c 与生成数据 $G(z,c)$ 间的互信息 $I(c;G(z,c))$，使隐含规则变量 c 在数据生成过程中发挥作用。另外，Chen 等引入了一种能有效优化互信息的目标下界 $\mathscr{L}_I(G,Q)$，其定义如式(12)所示。将其加入 GAN 的优化目标函数中，如式(13)所示，GAN 网络模型就能学习可解释性的有意义的视觉表征。

$$\begin{aligned}\mathscr{L}_I(G,Q)&=E_{c\sim P(c),x\sim G(z,c)}[\ln Q(c|x)]+H(c)\\&=E_{x\sim G(z,c)}[E_{c'\sim P(c|x)}[\ln Q(c'|x)]]+H(c)\\&\leqslant I(c;G(z,c))\end{aligned} \tag{12}$$

$$\min_{G,Q}\max_{D}V_{\text{InfoGAN}}(D,G,Q)=V(D,G)-\lambda\mathscr{L}_I(G,Q) \tag{13}$$

与 InfoGAN 网络类似，Singh 等[113]以 GAN 为基础采用了信息论的相同原则，提出了 FineGAN 模型。但不同的是，InfoGAN 中对象的所有细节都是一起生成的，而 FineGAN 提出了明确的分离，分别对背景、形状和外观的生成进行了控制，将 4 个随机采样的潜码作为输入，将生成图片过程分离成 3 个阶段：背景阶段、父阶段和子阶段。背景阶段通过控制背景图像生成的潜在变量进行背景建模；父阶段利用控制对象形状和姿势的潜在变量生成前景对象，并将其融合进背景图像中；子阶段利用控制对象外观纹理的潜在变量进行外观建模。3 个阶段采用 3 个独立的网络，3 个网络首尾相连将每个网络生成的图像拼接在一起，形成最终的图像。将生成图像的过程进行分离，使得网络能够达到很好的解耦表征效果。尽管 FineGAN 可以分解多个因子并且重新生成逼真的图像，但它是无条件的，且在采样过程中使用的是随机编码而不是图像。基于此，在 FineGAN 的基础上，Li 等[114]于 2020 年提出了条件生成模型 MixNMatch。MixNMatch 单独训练 4 个编码器，从真实图像中学习图像的姿势、背景、形状和纹理，并将其作为生成器的条件输入到生成器中。MixNMatch 能从图片信息中学习到 4 类不同的视觉表征，取得了不错的解耦效果。

2019 年，Li 等[115]提出了一种完全无监督的生成式对抗网络 SCGAN，用于学习可解释性的视觉表示。SCGAN 从平滑度假设和对图像内容和表征的假设出发，通过在条件和合成图像之间添加有效的相似度约束来解耦出可解释性的表征。与 InfoGAN 相比，SCGAN 直接在传统的 GAN 中添加了一个正则化项，即相似约束，取得了与 InfoGAN 一样的效果。上诉这些解耦式视觉表征模型在单目标的数据集中能够达到很好的解耦效果，但在多目标的数据集中效果却不尽如人意。对此，Ojha 等[116]在 FineGAN 的基础上，提出了一种能够在多个领域上解开对象形状和外观特征的生成模型。Ojha 等认为，不能实现跨域解耦的原因在于在单域数据集下未考虑源域属性信息，导致其耦合在形状、外观等表征中。于是，他们利用视觉特征的可微直方图表示物体外观，并优化生成器，使两幅具有相同潜在外观因子但潜在形状因子不同的图像产生相似的直方图，从而实现跨领域的解耦式视觉表征学习。

5.3 基于 VAE-GAE 的解耦表征学习

VAE 和 GAN 在解耦式视觉表征学习中都有着较好的效果，但都有各自的不足。如在 VAE 中，解码产生的图片往往都是比较模糊的，而 GAN 中生成器的输入都是随机的噪声，导致网络的模型训练比较困难。基于此，Larsen 等[117]提出了将 VAE 和 GAN 结合的 VAE/GAN 模型，如图 12 所示。他们认为 GAN 的鉴别器网络能够学习到丰富的图像相似性度量，能将真实图像和生成图像区分出来，而 VAE 中编码器输出的编码符合某些空间分布，有利于模型的训练。于是，Larsen 等将真实图片 x 作为输入，通过 VAE 的编码器和解码器生成 $\tilde{x}$，将 VAE 的解码器当作 GAN 的生成器，把提取到的编码特征作为输入传递给 GAN 代替传统噪声，再将 x 与 $\tilde{x}$ 输入到判别器中判别图片的真假。在训练时，VAE/GAN 将在变分目标下界增加对抗性损失来训练网络模型，使编码器能够学习到可解释性的视觉表征。VAE/GAN 模型结合了

VAE 和 GAN 网络各自的优点，是一种新的生成模型。与 VAE/GAN 模型的研究类似，Rosca 等[118]提出了 α-GAN。α-GAN 运用对抗训练使得编码隐变量分布和先验分布更接近，同时用生成器来增强隐变量到样本的映射拟合能力。α-GAN 的编码器不直接限制输出的隐变量的分布类型，而是引入一个二分类器 C 来辨别先验分布和编码后分布，从而用分类器输出比值代替传统 VAE 中后验估计与先验的 KL 散度，将 VAE 的解码器当作 GAN 的生成器，损失函数由样本编码解码后的重构损失和 GAN 的生成器损失组成，训练时使用对抗学习的方式对生成器进行训练。通过在不同的数据集上与不同的模型进行比较，α-GAN 在解耦式视觉表征学习中有很好的效果。随着改进特征归因在医学图像分类任务中的影响，Bass 等[119]使用 VAE-GAN 的架构提出了 ICAM 框架。ICAM 在 GAN 的网络框架中引入了 VAE 的思想，使用编码器编码的特征作为重构输入而不是随机噪声，在生成虚假图像后再次进行一次图像生成，用二次生成的图像与真实图像之间的差异作为模型的损失。ICAM 通过在更容易解释的潜在空间可视化类别和类别内部之间的差异，挖掘出决定相似图像类别差异的特征，实现类别与无关特征之间的解耦，从而实现更有效的解耦式视觉表征学习。

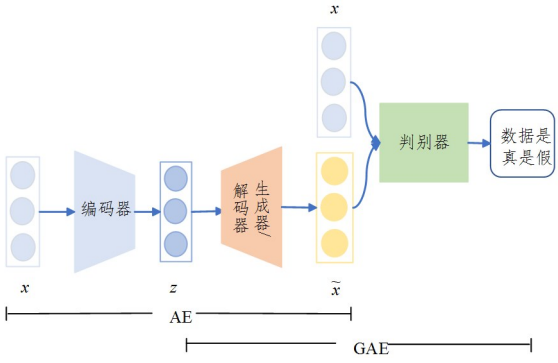


图 12 VAE/GAN 的网络结构[117]

Fig. 12 Network structure of VAE/GAN[117]

5.4 性能表现

图 13 为 β-VAE，VAE，InfoGAN 算法在 celebA 数据集[120]上的解耦性能展示图。

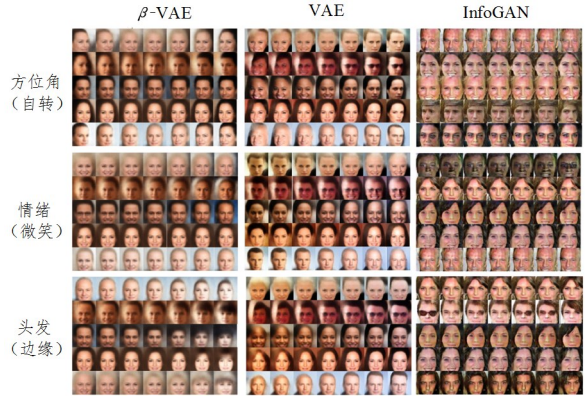


图 13 β-VAE，VAE，InfoGAN 算法在 celebA 数据集上的解耦性能展示图[100]

Fig. 13 Disentangled performance of β-VAE，VAE and InfoGAN on celebA dataset[100]

图中，只有 β-*VAE* 和 InfoGAN 学会了解耦方位角、情绪和发型等因素，而 VAE 则将其中两个因素纠缠在了一起，无法解耦。通过上述的分析和观察可视化结果图可知，解耦表征学习的重点是挖掘数据生成背后复杂耦合的物理机理，从而学习更灵活、更高质量的视觉表征，并将其应用到下游任务中。

6 结合语言信息的视觉表征学习

前文主要介绍了在视觉领域学习表征的各种方法，针对的问题仅包含单一模态信息，属于单模态表征学习。然而，在实际问题中，研究的对象通常包含图像、视频、语音、文本等各种模态信息。在存在多个模态信息的情况下，需要同时将多个模态数据中所蕴含的语义信息转化为实值向量，并且还要考虑多个模态信息之间的一致性和互补性。从模态形式的角度，多模态表征学习主要分为图像加音频、视频加音频、图像加文本、视频加文本等情形。其中，结合语言信息的视觉表征学习是最典型和最常见的情形。前述的单模态视觉表征学习已经取得重大进展。在自然语言处理（NLP）领域中，word2vec[121]等隐式表征学习的使用，也充分挖掘了文本信息中的潜在表征。另外，Transformer[122]和 BERT[123]在 NLP 领域的成功使用，也证明了对数据表征进行充分学习的重要性。这些单模态表征学习方法的提出也促进了多模态表征学习从早期的简单连接单一模态方式发展到了现在的多模态统一表征。本章主要介绍结合语言信息的视觉表征学习的方法和应用。

结合语言信息的视觉表征学习是多模态表征中的一个重要研究方向，在多个领域都有广泛的应用。该模式旨在同时学习视觉表征和语言表征，并将它们映射到一个共同的空间中。为了学习到更好的特征表示，需要充分利用视觉表征和语言表征之间的互补性，并消除模态之间的冗余性。2019 年，Baltrušaitis[124]等基于输出的表征是否在一个统一的表征空间上，将多模态表征分为联合表征和协调表征。联合表征将多个单模态表征投影到同一个统一的表征空间中，使多模态特征能够融合；而协调表征单独处理单模态表征，但在不同模态之间施加一定的相似性约束，使多模态表征能够相似协调。这两种方式在不同的场景中都有广泛的应用。随着 Transformer 等神经网络框架的涌现，研究者发现这些框架的编码器能够同时学习到不同模态的信息，于是基于 Transformer 等神经网络预训练框架来编码融合统一各种模态的表征的方法开始流行。另外一种方法是利用视觉语言相似性的方法来优化编码，以便让不同模态的信息在同一表征空间内相互协调。本章主要从视觉语言表征融合统一和视觉语言表征相似协调两个方面来介绍结合语言信息的视觉表征学习。

6.1 视觉语言表征融合统一

受自然语言处理领域和计算机视觉预训练的启发，利用预训练架构来实现视觉语言表征融合统一的研究大量涌现。特别是受自然语言处理领域的 Transformer 和 BERT 的广泛应用影响，许多研究者将这些技术作为基础，提出了基于预训练架构和 Transformer 特征抽取的多模态模型。在结合语言信息的视觉表征学习中，关键问题是将语言信息和视觉信息进行融合统一，以生成图像和文本的上下文联合表征。目前的研究主要采用两种方式来融合输入的视觉特征向量和文本

特征向量，分别为单流建模和双流建模，如图 14 所示。单流建模将视觉特征和文本特征进行拼接，输入到 Transformer 中进行融合；双流建模先将视觉特征和文本特征映射到相同的语义空间中，然后通过交叉注意力机制对不同模态信息进行融合。当然还有研究者提出了其他的融合方式，表 4 列出了多模态视觉语言模型的建模方式及其网络模型。

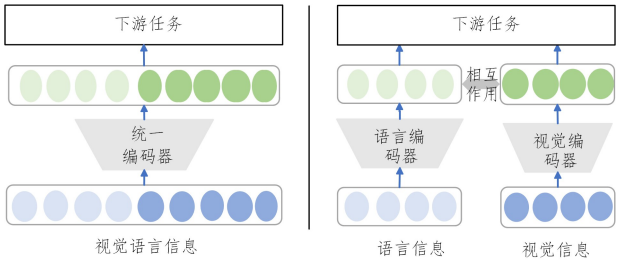


图 14 单流建模和双流建模

Fig. 14 Single stream modeling and dual stream modeling

表 4 多模态视觉语言模型的建模方式及其网络模型

Table 4 Modeling methods of multimodal visual language models and their network models

建模方式	网络模型	适用领域
单流建模	VideoBERT	视频文本
	VisualBERT	图文
	UNITER	图文
	VL-BERT	图文
	Unicoder-VL	图文
	B2T2	图文
	Pixel-BERT	图文
	SOHO	图文
	ViLT	图文
	Oscar	图文
	VinVL	图文
	VIVO	图文
双流建模	ViLBERT	图文
	LXMERT	图文
	12-in-1	图文
	Ernie-ViL	图文
其他建模方式	SemVLP	图文

基于预训练架构的多模态视觉语言模型大多采用单流建模的方式来融合不同模态之间的信息。Sun 等[125]于 2019 年提出的 VideoBERT 是结合语言信息的视觉表征学习的一项开创性工作。VideoBERT 通过视频来学习跨模态表示。通过自动语音识别技术将视频语音转换为文本，获得文本信息，通过卷积网络从视频的片段中提取视觉信息，对视频中提出的特征向量采用聚类的方法进行离散化，再将文本信息和视觉信息输入到构建在 BERT 上的单流模型中，学习视觉表征。VideoBERT 在视频说明、零样本预测等多个下游任务中取得了很好的效果。同年 Li 等[126]提出了一种更为简单且灵活的框架 VisualBERT。VisualBERT 采用了 Transformer 层的堆叠，将学习到的视觉和文本信息直接组合输入到 Transformer 中，隐式地将相关的文本信息和视觉信息进行融合。VisualBERT 在不降低模型性能的情况下，简化了模型结构，更加灵活。基于 BERT 的单流模型的基本结构都与 VisualBERT 类似，采用 Transformer 中的自注意力机制隐式地将相关的文本信息和视觉信息融合统一，复用 BERT 的加掩码操作的编码方式，最后采用预训练加下游任务微调的方式来训练模型。一些并行的研究，如 UNITER[127]，VL-BERT[128]，Unicoder-VL[129]等也采用了单流架构。这些模型的主要不同就是在预训练方式和数据上存在差异。另外，Alberti 等在单流模型的文本信息中融合了检测到的视觉目标特征，提出了 B2T2[130]。通过实验发现，正确的视觉信息和语言信息结合能够提高视觉问答的效果，并证明了早期融合对象和文本标记有利于模型学习更好的特征。上诉多模态的预训练模型主要使用基于 BERT 的语言特征和基于区域的视觉特征作为输入来进行学习。一些研究对视觉表征的形式进行了改变。Pixel-BERT[131]通过从像素特征中学习视觉表征，直接建立了图像像素与语义信息的连接，无需标注边界框。并在训练时使用随机像素采样机制来降低计算成本和提高模型的鲁棒性。SOHO[132]直接将整张图片作为视觉特征的输入，不需要目标检测器来找到目标区域。SOHO 是一个端到端的预训练框架，可以直接利用图像文本对学习跨模态的表征。为了更好地对齐图像和文本表征，SOHO 还提出了一个动态更新的视觉字典来提取视觉表征。另外 Kim 等[133]将图像块作为输入来提取特征，提出了 ViLT 模型，ViLT 在不使用卷积操作的情况下，也能达到很好的效果，并且还减少了模型的参数，提高了训练的效率。另外还有一些研究对视觉语言表征对齐工作进行了改进，比如，Li 等[134]提出的 Oscar 模型利用图像中检测到的相同语义下的对象标签作为锚点，来简化语言特征和视觉特征空间的对齐。这一想法主要是因为视觉中检测到的对象标签会在对应的文本中出现，并且在视觉中很有可能很难区分语义特征，而在文本信息中却很好区分。VinVL[135]则采用 ResNeXt-152 作为新的目标检测模型，生成新的视觉特征输入到 Oscar 网络模型中，并在更大的数据集中进行训练，在多个下游任务中都取得了最好的成绩。与 Oscar 网络模型类似，VIVO[136]也通过改进视觉语言的对齐工作，来优化模型以学习更好的特征。

双流建模采用两个编码器分别对视觉信息和文本信息进行编码，然后使用交叉编码的方式学习每种模态的特征，使双流建模的特征学习更加充分，如图 14(b)所示。Lu 等[137]于 2019 年提出了 ViLBERT 网络模型，将单流的 DERT 架构扩展到多模态的双流模型。Lu 等认为对语言的理解比对图像的理解复杂，因此两者所需要的编码深度应该是不一样的，ViLBERT 使用两个并行流分别处理视觉信息和文本信息，然后采用共同注意力转换模块来融合不同模态的表征，使视觉特征和文本特征都包含其他另一种信息特征的先验条件。这种双流结构能够允许每个模态的特征深度变化，并通过共同注意力转换模块实现稀疏交互。TAN 等[138]提出了另一种双流模型 LXMERT。LXMERT 由对象关系编码器、语言编码器、跨模态编码器 3 个编码器组成。对象关系编码器和语言编码器分别对图视觉信息和语言信息进行处理，得到单一模态特征；跨模态编码器以两种单一模态作为输入，负责将文本特征和视觉特征进行融合，产生联合表征。为了获得更好的融合表征，LXMERT 在 5 种预训练任务中进行了大量的训练，以提高模型学习模态内和模态间关系的能力。Lu 等[139]在 ViLBERT 的基础上进行了拓展，提出了一种多任务训练模型 12-in-1。

在12个数据集上进行了预训练,实验结果表明,多任务预训练能够提高下游任务的性能,并产生更轻量级的模型。另外,受ERNIE[140]的启发,Yu等[141]首次将结构化知识引入到双流模型的多模态视觉语言模型中,提出了ERNIE-ViL。ERNIE-VIL将场景图引入到了多模态预训练任务中,利用场景图预测任务对图中的对象、属性和关系进行建模,学习对象级和属性感知表示。通过实验发现,ERNIE-VIL能够在较少的训练资源和训练数据的情况下取得较好的结果。

从融合特征的时间上看,双流建模先提取每种模态的高级特征之后,再进行特征融合统一,而单流建模会更早地将底层特征空间中的视觉特征和语言特征连接起来,直接进行融合统一。Li等观察到一些图像文本对很容易就能将不同模态的特征进行融合统一,而其他图片需要在更高级的抽象特征中才能实现特征融合统一。对此,提出了一种结合两种建模架构的网络模型SemVLP[142]。SemVLP由共享Transformer编码器和跨模态注意模块组成,通过迭代训练的方式来组合两种建模结构。在单流模式中,直接将连接的图像特征和文本特征输入到共享的Transformer编码器中,进行预训练;在双流模式中,使用两个共享Transformer编码器分别对图片信息和文本信息编码,再将输出的特征输入到跨模态注意模块进行特征融合对齐,得到联合表征。实验结果表明,SemVLP在跨模态表示与不同语义粒度对齐方面具有很好的效果,并且共享的Transformer编码器也减少了模型的参数,优化了模型,更利于视觉语言表征的融合统一。

6.2 视觉语言表征相似协调

视觉语言表征相似协调是在表征空间中最小化视觉信息和语言信息之间的距离,利用对比学习的思想,学习两个模态信息之间的表征。与单模态对比式视觉表征不同,在结合语言信息的视觉表征学习中,一般将正例的图像文本对作为正样本,而将负例的文本图像对和负例的图像文本对作为负样本。2018年,Lee等提出的SCAN[143]网络模型将正负样本作为网络的输入,采用三元损失函数(见式(14)),来计算总的损失来训练网络。

$$I_{\text{hard}}(I,T)=[\alpha-S(I,T)+S(I,\widetilde{T}_h)]_+ +[\alpha-S(I,T)+S(\widetilde{I}_h,T)]_+ \tag{14}$$

其中,α为一个常量,S为一个相似度函数方程。正样本为(I,T)正例的图像文本对,负样本为$(I,\widetilde{T}_h)$负例的图像文本对和$(\widetilde{I}_h,T)$负例的图像文本对。通过该损失函数训练网络模型,使得在特征空间中正样本之间的距离远大于负样本之间的距离,从而优化编码器,学习更好的特征。Faghri等[144]通过加大样本与疑难负例之间的距离,提出了新的损失函数,如式(15)所示:

$$\ell_{MH}(i,c)=\max_{c'}[\alpha+s(i,c')-s(i,c)]_+ +\max_{i'}[\alpha+s(i',c)-s(i,c)]_+ \tag{15}$$

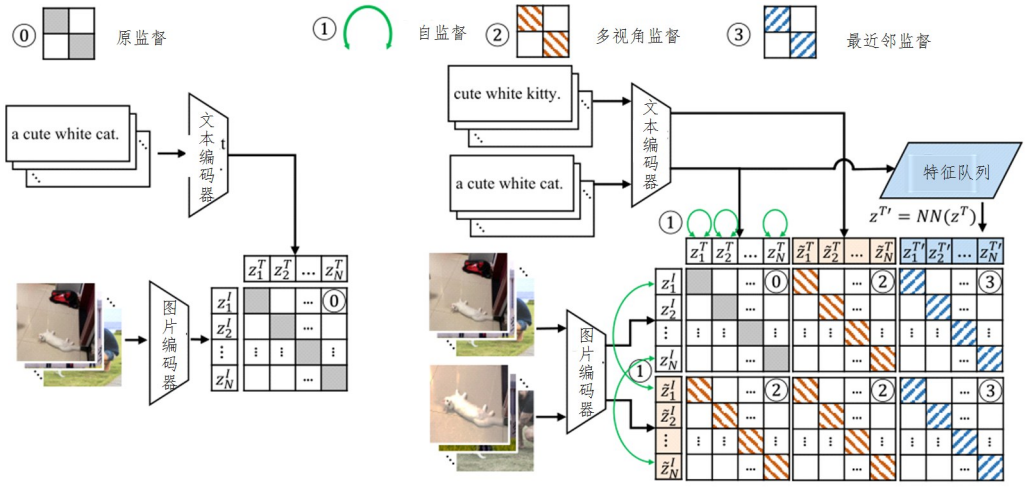
其中,(i,c)表示正样本对,c'和i'表示负例,s表示距离函数。在负例的选择上,$i'=\arg\max_{j\neq i}s(j,c)$,$c'=\arg\max_{d\neq c}s(i,d)$,即在训练时选择距离正样本较远的负样本进行训练,从而学习更好的表征。

在NLP领域中,直接预训练的方法取得了巨大成功,这些预训练不需要人工注释的原始文本训练也能取得良好的效果,而视觉领域预训练模型必须在有很多标签的数据集上训练才能取得较好的效果。因此,有研究提出使用纯文本来引导视觉领域模型的学习。基于对这个问题的思考,Radford等[145]提出了CLIP网络,如图15(a)所示。他们创建了包含4亿个图像文本对的数据集WIT(Web ImageText)来进行训练。CLIP使用两个编码器对图像信息和文本信息进行编码。其中,图片编码器使用ResNet或VIT,文本编码器使用Transformer模型。将生成的特征向量映射到多模态的特征空间中,使用对比学习的思想,拉近匹配的图片和文本的距离,推远不匹配的图片和文本距离。图15(a)的矩阵中,对角线上都是配对的正样本对,其他都是负样本对。对于训练批量大小为N的训练样本,正样本对为N对,负样本对为N^2-N对。使用损失函数InfoNCE来训练网络。CLIP预训练的模型在多个视觉任务上都取得了较好的结果。2021年Jia等[146]提出了ALIGN网络模型,如图15(a)所示。与CLIP网络模型一样,ALIGN也使用两个编码器来分别学习图片信息和文本信息中的特征,使用对比学习损失训练两个编码器,以便编码器学习对齐图像和文本对的视觉和语言表示。正负样本的选取方式也与CLIP一致。但ALIGN训练时使用一个没有进行数据预处理的噪声数据集,该数据包含超过10亿个图像文本对,且这些图像文本对都没有进行数据过滤和后处理。ALIGN模型学习到的表征在多个下游任务中获得了不错的结果。Jia等证明了数据规模的提升可以弥补数据内部存在噪声的不足,使用简单的对比学习方式,模型也能学习到好的特征。

与CLIP和ALIGN网络模型一样,大多数视觉语言模型[148]协调和对齐文本和图像之间的信息都是通过一组跨模态Transformer来实现的,这些方法需要专用的跨模态转换层或现成的对象检测器来实现,严重阻碍了它们的可扩展性。虽然CLIP和ALIGN网络训练的数据集不需要太多的人工标注信息,但它们仍需要大量的训练数据进行训练,占用了大量的资源。为了提高训练效率,Li等[147]提出了DeCLIP,这是一种可以在较少训练数据下仍能取得不错效果的多模态预训练模型。DeCLIP模型使用了双塔框架,仅在顶部进行多模态的交互,如图15(b)所示。DeCLIP直接从互联网上丰富的图像-文本对中学习,不使用单一的图像和文本对比监督,而是充分利用图像文本对的广泛监督,通过模式内的自监督、跨模式的多视角监督和来自其他类似配对的最近邻监督来学习多模态的表征。DeCLIP首先在每种单个模态内使用各自的自监督学习框架.在视觉领域,采用了简单而有效的SimSiam网络框架,最大化两个增强图像特征之间的相似性;对于文本自监督,采用了掩码语言模型(MLM)[123]框架。然后使用跨模式的多视角监督对图片和文本进行随机数据增强,计算所有2×2对的图像文本对比损失来增加更多的监督。最后使用最近邻监督,更好地利用数据集中的相似文本描述。DeCLIP大大减少了数据集的容量,提高了训练效率。另外,其充分利用了各个模态的自监督学习,提高了模型的可扩展性。

在以上的研究中,大多针对的是图像和文本的检索任务,

而对于图像和文本的其他任务，这些研究缺乏图像和文本之间更复杂交互的能力。于是，Li 等[149]提出了 ALBEF 模型，将图像-文本对比学习(ITC)、掩蔽语言建模(MLM)和图像-文本匹配(ITM)作为预训练任务，并将 3 者的损失函数当作整个模型的损失函数来训练模型。首先用两个编码器独立对图像和文本进行编码，然后利用多模态编码器，将图像特征和文本特征进行相似协调，并提出了动量蒸馏来对抗数据中的噪声，学习更好的表征，ALBEF 在多个下游视觉语言任务上达到了最佳的效果。与 ALBEF 网络类似的还有 Wang 等[150]提出的 VLMO 模型，其引入了一个模态混合专家的 Transformer，能够根据不同的输入数据类型选择不同的专家，并且 VLMO 采用了分阶段的预训练方式，得到了更泛化的表示，可以在更多的下游任务中取得更好的效果，进一步提高了视觉语言表征的相似协调。



(a)CLIP 和 ALIGN 结构图 (b)DeCLIP 结构图

图 15 CLIP，ALIGN 和 DeCLIP 网络结构图[147]

Fig. 15 Network structure of CLIP, ALIGN and DeCLIP[147]

7 评价准则

如上文所述，不同网络框架在学习视觉表征的过程中展现出了不同的能力。基于预训练的网络模型在有监督条件下表现出卓越的性能，能够有效学习高质量的视觉表征，从而显著提升下游任务的性能。相较之下，无监督学习的 3 种方式的性能略逊于预训练的网络模型，但它们为数据收集工作带来了便利，并在不断的研究中持续提升对视觉表征的学习能力，从而更好地被应用到下游任务中。解耦式网络模型能够对视觉表征进行解耦，通过改变某个生成因子来实现对不同的视觉表征的学习，进而在多个视觉任务中灵活应用。另外，结合语言信息的视觉表征学习将语言信息和视觉信息结合，通过有效地利用语言信息，能够充分地学习到视觉信息中的高质量视觉表征，从而提高网络学习视觉表征的能力。在语言信息的辅助下，融合语言信息的视觉表征网络模型在各类下游任务中呈现出了显著的性能优势。

视觉表征学习通常与分类、检测或其他下游任务一起考虑，因此良好的表征对于机器学习模型的性能至关重要。为了使下游任务获得良好的表征，在构建模型时必须正确考虑不同下游任务共有的先验知识。将这些先验知识形式化为一种评价准则对视觉表征学习的性能至关重要。目前，主要可以通过下游任务和建立分离式度量两种方式来评价视觉表征学习。

下游任务效果的好坏是评价视觉表征学习中所学表征是否高效的直接依据，是评价视觉表征学习的直观方法。Bengio 等[1]首次提出一个良好的表征应该具有平滑性、时间和空间相干性、稀疏性和自然聚类等属性。对于每一个目标函数 f，目标函数的平滑性要求数据尽可能覆盖目标函数的空间。对于编码器学习到的表征，在多个下游任务中都有效，不需要针对每一个下游任务进行单独训练。例如，使用对比学习方法的动量对比法 MoCo 能在检测和分割等不同的任务中都达到很好性能，并且它对于不同的数据集也是有效的。为了使学习到的表征能在下游任务中更好地被利用，在维数上，表征向量的维数应该要远低于原始空间的维度，这样才能更好地被机器学习理解，下游任务才能取得更好的效果。另外，为了更好地在表征空间中存储表征，学习到的表征应该尽可能的稀疏，并且在观测数据上的微小扰动不会影响其他大部分表征。在表征空间中，不同类的数据尽可能分布在分散的流形上，并且不同类的数据的线性插值应该处在低密度区域。越抽象的表征应该在越高层，包含的信息也应该更丰富。对于多模态的表征，需要找到对多模态信息的统一表示，来自同一个体的不同模态信息的表征间应具备更高的相似程度，在表征空间的相似性应能够反映出表征所对应的概念的相似性，学习到的统一表征能在多个下游任务中起到好的效果。即使缺失某些模态数据，多模态表征仍能够轻松地获得，并且能够在给出被观察到其他模态数据后，填补出缺失的模态数据并且能够在只包含部分模态数据的情况下，填补出缺失的模态数据。

另外一种评价视觉表征学习的方式是建立稳健的分离式度量。分离式度量是为了量化视觉表征学习与影响因子之间的统计关系，主要体现在解耦式视觉表征学习中。在评价准则中，量化由不同模型实现的解耦程度非常重要，但在实际的

设计过程中并不容易。最早有研究者提出使用可预测性的概念来量化解耦,从潜在编码中预测真实因子的值。Yang等[151]在线性ICA环境中学习从表征到因子的线性映射,并量化此映射与排列矩阵的接近程度。Eastwood等[152]则通过Lasso回归器将潜在编码映射到真实因子中,并使用训练后的权值对解耦进行量化。与这些方法不同,2017年,Higgins等[100]提出了β-VAE度量,通过固定一个数据生成因子的值,随机采样其他所有数据,生成大量的图像进行推断。如果推理表征具有独立性和可解释性,则推断潜在因素与固定性因素的差异就会更小。再使用一个低容量的线性分类器来识别这个因素,并将精度值作为最终的解缠度量值。而Kim等认为这种度量方式也存在一些问题,β-VAE度量对线性分类器优化的超参数太敏感,且线性分类器也不直观,最重要的是,这个度量有一个失效模式,当K个因素中只有$K-1$个被解耦,它也能给出100%的准确度。基于此,Kim等[102]提出了新的解耦度量方法,即FactorVAE度量,FactorVAE度量选择一个因子,生成带有该固定因子、随机变化其他因子的数据,获得它们的表征,将整个数据或者足够大的随机子集的经验标准差的每个维度进行归一化,得到这些归一表征的每个维度的经验方差,有着最低方差维度的索引和固定因子的目标索引为分类器提供了一个训练输入和输出示例。如果表征能完美地解耦,与固定因子对应维数的经验方差将为0。将表征进行标准化,使参数最小值对每个维度的表征的缩放是不变的,当输入和输出分布在一个离散空间中时,最终分类器为多数投票分类器,度量为分类器的错误率。FactorVAE度量比β-VAE度量更简单,更自然。FactorVAE度量中的分类器需要看到给定因子的潜在维数的最小方差才能正确分类,避免了β-VAE中的失效模式。Do等[153]在2020年从信息论的概念出发,借鉴了信息论、可分离性和可解释性3个维度,通过形式化地描述解耦表示,为这些属性设计了稳健的量化度量。

下游任务和建立分离式度量是评价视觉表征学习的两种主要方式,在很多模型中,都通过这两种方式来衡量模型学习到的表征的好坏。但目前,这两种方式只能在一些模型上进行评价判断,仍没有一种通用的评价准则,无法对各个网络进行标准的评价,这个问题仍需要更深入的研究。

8 未来展望

无监督的视觉表征学习是目前视觉表征学习的一个主要趋势,相对于有监督的表征学习,无监督表征学习无需大量的标注数据集即可学习到视觉信息的表征。然而,生成高质量的表征仍需要依赖大量的预训练数据集。尽管有许多无监督学习方法在下游任务中取得了不错的效果,但其效果仍不及有监督学习。因此,如何使无监督的视觉表征模型学习到更高质量的表征,仍需进一步研究。综合不同表征学习框架的优缺点,未来有以下几点值得深入研究。

1)提升模型训练及推理的性能。目前无监督学习的视觉表征学习框架在检测和分割等下游任务中性能有了较大的提升,但在实际推理任务中,其速度仍然较慢,主要的原因有以下几个方面。(1)大部分网络框架复杂且参数量大,另外还有一些网络为了提高模型的效果,采用多阶段的方式提取特征。比如,在多模态的视觉语言任务中,很多模型都采用了Faster-RCNN两阶段目标检测模型,导致训练和推理速度降低。可以通过采用效率更高的单阶段检测框架来优化网络模型。(2)随着Transformer在视觉领域的应用,越来越多的网络使用Transformer架构作为预训练模型。而Transformer网络架构的计算量大,参数较多,因此可以采用蒸馏、压缩等方式来进行提升。例如,Jiao等[154]通过两阶段蒸馏的方式提出了TinyBERT模型,TinyBERT同时对预训练任务和下游任务进行蒸馏,大大减小了模型的规模,提升了网络训练及推理的速度。

2)利用数据的可解释性提高表征的质量。Bengio等[1]指出,数据中的样本是由不同的因子生成的,完美的表征不应该扭曲生成数据的变量的潜在因子。因此需要充分利用数据的可解释性来提高表征的质量,以更好地用于下游任务。

3)提出通用的表征框架。目前,大多数视觉表征框架仅适用于特定的下游任务和某些类似的数据集,对于不同数据集和不同下游任务,学习的模型并不适合。例如,在多模态的视觉语言表征学习中,大多数框架如VideoBERT,VisualBERT,UNITER等都是偏向于内容理解的任务,但这些模型在内容生成等任务中,展现的效果并不好,无法实现跨任务学习。另外,很多视觉表征学习框架在不同的数据集上展现的效果也不一样,表征学习框架不能普遍适用于多个数据集,数据集的变化可能会导致截然不同的结论。因此提出通用的表征框架是未来值得研究的问题。

4)制定评价指标。在解耦式视觉表征学习中,缺乏有效的评估度量来实现模型之间的公平比较。虽然有学者提出了各种指标,例如β-TCVAE模型中提出的互信息度量[103],FactorVAE中的FactorVAE度量[102]、β-VAE度量[100]以及2020年Do等[153]提出的用于度量信息性、可分离性和可解释性的评价指标等,但这些度量仍无法在定量衡量解耦性能的最佳表征上形成一致。目前,对视觉表征评价指标的研究还处于初级阶段,在未来需要进一步研究和探索。

结束语 本文总结了视觉表征学习的相关研究工作,根据模型对数据的依赖程度和类型的不同,将其分为了5种不同方式。首先介绍了有监督的基于预训练网络的视觉表征学习,包括近几年来比较流行的CNN网络框架。接着重点介绍了无监督学习的视觉表征学习,将无监督的方法分为了3类:生成式视觉表征学习、对比式视觉表征学习和解耦式视觉表征学习。生成式视觉表征学习介绍了不同的自编码器、生成对抗网络和扩散模型在视觉表征学习中的应用;对比式视觉表征学习主要介绍了应用对比学习的流行网络模型框架;解耦式视觉表征学习主要介绍了基于变分自编码器和生成对抗网络在解耦表征学习中的应用。然后又介绍了结合语言信息的视觉表征学习,包括视觉语言表征融合统一和视觉语言表征相似协调两种不同的方法,它们将视觉信息和语言信息结合起来用于视觉任务,以学习更好的视觉表征。最后介绍了视觉表征学习的评价指标和未来的发展趋势。

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] ZHANG D, YIN J, ZHU X, et al. Network Representation Learning: A Survey[J]. IEEE Transactions on Big Data, 2020, 6(1): 3-28.

[3] CHEN F X, WANG Y C, WANG B, et al. Graph representation learning: a survey[J]. Transactions on Signal and Information Processing, 2020, 9: e15.

[4] CHENG K Y, MENG C Y, WANG W S, et al. Research advances in disentangled representation learning[J]. Journal of Computer Applications, 2021, 41(12): 10.

[5] WEN Z D, WANG J R, WANG X X, , et al. A Review of Disentangled Representation Learning[J]. Acta Automatica Sinica, 2022, 48(2): 351-374.

[6] DU P F, LI X Y, GAO Y L. Survey of Multimodal Visual Language Representation Learning[J]. Journal of Software, 2021, 32(2): 22.

[7] YIN J, ZHANG Z D, GAO Y H, , et al. A survey on visual language pre-training[J]. Journal of Software, 2023, 34(5): 2000-2023.

[8] PEARSON K. On lines and planes of closest fit to systems of points in space[J]. London, Edinburgh & Dublin Philosophical Magazine & Journal of Science, 1901, 2(11): 559-572.

[9] FISHER R A. The Use of Multiple Measurements in Taxonomic Problems[J]. Annals of Human Genetics, 2012, 7(7): 179-188.

[10] BAUDAT G, ANOUAR F. Generalized Discriminant Analysis Using a Kernel Approach[J]. Neural Computation, 2000, 12(10): 2385-2404.

[11] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290: 2323-2326.

[12] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[13] FUKUSHIMA K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics, 1980, 36(4): 193-202.

[14] 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.

[15] ALEX K, ILYA S, GEOFFREY E. ImageNet classification with deep convolutional neural networks[C/OL]// ACM, 2017: 84-90. https://doi.org/10.1145/3065386.

[16] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv: 1409.1556, 2014.

[17] SZEGEDY C. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 1-9.

[18] SZEGEDY C, VANHOUCKEV, IOFFE S, et al. Rethinking the Inception Architecture for Computer Vision[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2818-2826.

[19] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[C]// AAAI. 2017.

[20] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2016: 770-778.

[21] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely Connected Convolutional Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2261-2269.

[22] IANDOLA F N, MOSKEWICZ M W, ASHRAF K, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size[J]. arXiv: abs/1602.07360.

[23] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [J]. ArXiv, abs/1704.04861.

[24] ZHANG X, ZHOU X, LIN Met al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.

[25] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Networks. June 2014[J/OL]. http://arxiv.org/abs/1406.2661, 2014.

[26] NG A. Sparese Autoencoder[J]. CS294A Lecture Notes, 2011, 72(2011): 1-19.

[27] LIN X, ZHU C, ZHANG Q, et al. 3D Keypoint Detection Based on Deep Neural Network with Sparse Autoencoder[J/OL]. 2016. https://www.semanticscholar.org/paper/3D-Keypoint-Detection-Based-on-Deep-Neural-Network-Lin-Zhu/f0226fd05ff951ca63d318ec71cca02925a887b9.

[28] MENG Q, CATCHPOOLE D, SKILLICOM D, et al. Relational autoencoder for feature extraction[C]// 2017 International Joint Conference on Neural Networks(IJCNN). 2017: 364-371.

[29] AN N, DING H, YANG J, et al. Deep ensemble learning for Alzheimers disease classification[J/OL]. Journal of Biomedical Informatics, 2019. https://www.sciencedirect.com/science/article/pii/S1532046420300393.

[30] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[R]. Universite de Montreal, 2008.

[31] GIDARIS S, KOMODAKIS N. Generating classification weights with gnn denoising autoencoders for few-shot learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 21-30.

[32] BO D, WEI X, JIA W, et al. Stacked convolutional denoising auto-encoders for feature representation[J]. IEEE Transactions on Cybernetics, 2016, 47(4): 1017-1027.

[33] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16000-16009.

[34] CHEN J, HU M, LI B, et al. Efficient self-supervised vision pretraining with local masked reconstruction[J]. arXiv: 2206.00790, 2022.

[35] SALAH R,VINCENT P,MULLER X. Contractive auto-encoders:Explicit invariance during feature extraction[C]// Proceedings of the 28th International Conference on Machine Learning. 2011:833-840.

[36] GANGULI S,IYER C V K,PANDEY V. Reachability Embeddings: Scalable Self-Supervised Representation Learning from Mobility Trajectories for Multimodal Geospatial Computer Vision[C]// 2022 23rd IEEE International Conference on Mobile Data Management(MDM). IEEE,2022:44-53.

[37] KINGMA D P,WELLING M. Auto-Encoding Variational Bayes [EB/OL]. https://www. ee. bgu. ac. il/~rrtammy/DNN/StudentPresentations/2018/AUTOEN~2. PDF.

[38] SOHN K, YAN X, LEE H, et al. Learning Structured Output Representation using Deep Conditional Generative Models[C]// International Conference on Neural Information Processing Systems. MIT Press,2015.

[39] LOUIZOS C,SWERSKY K,LI Y,et al. The Variational Fair Autoencoder[J/OL]. Computer Science, 2015. https://www. semanticscholar. org/paper/The-Variational-Fair-Autoencoder-Louizos-Swersky/cbef7a84a53e19e019e5a05d232eb3c487c0e0c6?p2df.

[40] ZHAO S,SONG J,ERMON S. Infovae:Information maximizing variational autoencoders[J]. arXiv:1706. 02262,2017.

[41] RAMACHANDRA G. Least Square Variational Bayesian Autoencoder with Regularization[J]. arXiv:1707. 03134,2017.

[42] CHEN X,KINGMA D P,SALIMANS T,et al. Variational lossy autoencoder[J]. arXiv:1611. 02731,2016.

[43] SHANG W,SOHN K,AKATA Z,et al. Channel- recurrent variational autoencoders[J]. arXiv:1706. 03729,2017.

[44] CAI L,GAO H,JI S. Multi-Stage Variational Auto-Encoders for Coarse-to-Fine Image Generation. CoRR abs/1705. 07202 (2017)[J]. arXiv:1705. 07202,2017.

[45] VAN DEN OORD A,VINYALS O. Neural discrete representation learning[J/OL]. Advances in Neural Information Processing Systems, 2017, 30. https://proceedings. neurips. cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract. html.

[46] RAZAVI A,VAN DEN OORD A,VINYALS O. Generating diverse high-fidelity images with vq-vae-2[J/OL]. Advances in Neural Information Processing Systems, 2019, 32. https://proceedings. neurips. cc/paper/2019/hash/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Abstract. html.

[47] RUECKERT F L. CR-VAE:Contrastive Regularization on Variational Autoencoders for Preventing Posterior Collapse[J]. arXiv:2309. 02968,2023.

[48] DENTON E L,CHINTALA S,FERGUS R,et al. Deep generative image models using a laplacian pyramid of adversarial networks[C]// Advances in Neural Information Processing Systems. 2015:1486-1494.

[49] RADFORD A,METZ L,CHINTALA S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[J/OL]. Computer Science,2015. http://arxiv.org/pdf/1511. 06434.

[50] SALIMANS T,GOODFELLOW I,ZAREMBA W,et al. Improved techniques for training gans[J/OL]. Advances in Neural Information Processing Systems,2016,29. https://proceedings. neurips. cc/paper_files/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract. html.

[51] LU S,DONG Z,CAI D,et al. MIM-GAN-based Anomaly Detection for Multivariate Time Series Data[C]// 2023 IEEE 98th Vehicular Technology Conference(VTC2023-Fall). IEEE,2023:1-7.

[52] WU J,ZHANG C,XUE T,et al. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling[J/OL]. Advances in Neural Information Processing Systems, 2016, 29. https://proceedings. neurips. cc/paper/2016/hash/44f683a84163b3523afe57c2e008bc8c-Abstract. html.

[53] YANG W X, YAN Y, CHEN S, et al. Multi-scale Generative Adversarial Network for Person Re-identification under Occlusion[J]. Journal of Software,2020,31(7):1943-195.

[54] SUN H,ZHU T,CHANG W,et al. Generative Adversarial Networks Unlearning[J]. arXiv:2308. 09881,2023.

[55] ATHREYA S,RADHACHANDRAN A,IVEZI? V,et al. Ultrasound Image Enhancement using CycleGAN and Perceptual Loss[J]. arXiv:2312. 11748,2023.

[56] MIRZA M,OSINDERO S. Conditional generative adversarial nets[J]. arXiv:1411. 1784,2014.

[57] REED S,AKATA Z,MOHAN S,et al. Learning What and Where to Draw[J/OL]. New Republic, 2016. https://proceedings. neurips. cc/paper/2016/hash/a8f15eda80c50adb0e71943adc8015cf-Abstract. html.

[58] ZHANG H,XU T,LI H,et al. Stackgan:Text to photo-realistic image synthesis with stacked generative adversarial networks [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:5907-5915.

[59] BOUROU A,BOYER T,DAUPIN K,et al. PhenDiff:Revealing Invisible Phenotypes with Conditional Diffusion Models[J]. arXiv:2312. 08290,2023.

[60] LI J,GUO Y M,YU T Y,et al. Multi-target Category Adversarial Example Generating Algorithm Based on GAN[J]. Computer Science,2022,49(2):83-91.

[61] ARJOVSKY M,CHINTALA S,BOTTOU L. Wasserstein GAN [J]. arXiv:1701. 07875,2017.

[62] MESCHEDER L,NOWOZIN S,GEIGER A. Adversarial variational bayes:Unifying variational autoencoders and generative adversarial networks[C]// International Conference on Machine Learning. PMLR,2017:2391-2400.

[63] LI J,GUO Y M,YU T Y,et al . Multi-target Category Adversarial Example Generating Algorithm Based on GAN[J]. Computer Science,2022,49(2):83-91.

[64] SOHL-DICKSTEIN J,WEISS E,MAHESWARANATHAN N,et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]// International Conference on Machine Learning. PMLR,2015:2256-2265.

[65] HO J,JAIN A,ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems,2020,33:6840-6851.

[66] NICHOL A Q,DHARIWAL P. Improved denoising diffusion probabilistic models[C]// International Conference on Machine

Learning. PMLR, 2021: 8162-8171.
[67] DHARIWAL P, NICHOL A. Diffusion models beat gans on image synthesis[J]. Advances in Neural Information Processing Systems, 2021, 34: 8780-8794.
[68] WANG Y H, YAIR S, AARON G, et al. InfoDiffusion: Representation Learning Using Information Maximizing Diffusion Models[C]// ICML. 2023.
[69] YANG X, WANG X. Diffusion Model as Representation Learner [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 18938-18949.
[70] SONG R, LIU Y, MARTIN R R, et al. 3d point of interest detection via spectral irregularity diffusion[J]. The Visual Computer, 2013, 29(6): 695-705.
[71] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[J/OL]. Handbook of Systemic Autoimmune Diseases, 2009, 1(4). https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf.
[72] BEEKLY D L, RAMOS E M, LEE W W, et al. The National Alzheimer's Coordinating Center (NACC) database: The uniform data set[J]. Alzheimer Disease & Associated Disorders, 2007, 21: 249-258.
[73] HARIHARAN B, GIRSHICK R. Low-shot visual recognition by shrinking and hallucinating features[J]. arXiv: 1606.02819, 2016.
[74] JIA D, WEI D, RICHARD S, et al. ImageNet: A large-scale hierarchical image database[C]// CVPR. 2009.
[75] WELINDER P, BRANSON S, MITA T, et al. Caltech-UCSD Birds 200[R]. California Institute of Technology, 2010.
[76] WANG T, ISOLA P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere [C]// International Conference on Machine Learning. PMLR, 2020: 9929-9939.
[77] WU Z, XIONG Y, YU S X, et al. Unsupervised Feature Learning via Non-parametric Instance Discrimination [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.
[78] YE M, ZHANG X, YUEN P C, et al. Unsupervised embedding learning via invariant and spreading instance feature[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6210-6219.
[79] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9729-9738.
[80] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]// International Conference on Machine Learning. PMLR, 2020: 1597-1607.
[81] AARON V D O, LI Y Z, ORIOL V. Representation learning with contrastive predictive coding[J]. arXiv: 1807.03748, 2018.
[82] CHEN X, FAN H, GIRSHICK R, et al. Improved Baselines with Momentum Contrastive Learning[J]. arXiv: 2003.04297, 2020.
[83] CHEN T, KORNBLITH S, SWERSKY K, et al. Big self-supervised models are strong semi-supervised learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 22243-22255.
[84] OORD A, LI Y, VINYALS O. Representation Learning with Contrastive Predictive Coding[J]. arXiv: 1807.03748v1, 2018.
[85] HENAFF O. Data-efficient image recognition with contrastive predictive coding[C]// International Conference on Machine Learning. PMLR, 2020: 4182-4192.
[86] TIAN Y, KRISHNAN D, ISOLA P. Contrastive multiview coding[C]// European Conference on Computer Vision. Cham: Springer, 2020: 776-794.
[87] HASSANI K, KHASAHMADI A H. Contrastive multi-view representation learning on graphs[C]// International Conference on Machine Learning. PMLR, 2020: 4116-4126.
[88] TIAN Y, SUN C, POOLE B, et al. What makes for good views for contrastive learning? [J]. Advances in Neural Information Processing Systems, 2020, 33: 6827-6839.
[89] CARON M, MISRA I, MAIRAL J, et al. Unsupervised learning of visual features by contrasting cluster assignments[J]. Advances in Neural Information Processing Systems, 2020, 33: 9912-9924.
[90] LI Y, HU P, LIU Z, et al. Contrastive clustering [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 8547-8555.
[91] VAN GANSBEKE W, VANDENHENDE S, GEORGOULIS S, et al. Scan: Learning to classify images without labels[C]// European Conference on Computer Vision. Cham: Springer, 2020: 268-285.
[92] CHEN X, XIE S, HE K. An empirical study of training self-supervised vision transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 9640-9649.
[93] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 9650-9660.
[94] GRILL J B, STRUB F, ALTCH? F, et al. Bootstrap your own latent-a new approach to self-supervised learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 21271-21284.
[95] ABE F, JOSH A. Understanding self-supervised and contrastive learning with bootstrap your own latent (BYOL)[OL]. https://untitled-ai.github.io/understanding-self-supervised-contrastive-learning.html, 2020.
[96] TIAN Y D, YU L T, CHEN X L, et al. Understanding self-supervisedlearning with dual deep networks[J/OL]. 2020. http://arxiv.org/abs/2010.00578v2.
[97] RICHEMOND P H, GRILL J B, ALTCHÉ F, et al. BYOL works even without batch statistics[J]. arXiv: 2010.10241, 2020.
[98] CHEN X, HE K. Exploring simple siamese representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15750-15758.
[99] ZBONTAR J, JING L, MISRA I, et al. Barlow twins: Self-supervised learning via redundancy reduction[C]// International Con-

ference on Machine Learning. PMLR,2021:12310-12320.

[100]HIGGINS I,MATTHEY L,PAL A,et al. beta-vae:Learning basic visual concepts with a constrained variational framework[J/OL]. 2016. https://www. semanticscholar. org/paper/beta-VAE%3A-Learning-Basic-Visual-Concepts-with-a-Higgins-Matthey/a90226c41b79f8b06007609f39f82757073641e2.

[101]BURGESS C P,HIGGINS I,PAL A,et al. Understanding disentangling in β-VAE[J]. arXiv:1804.03599,2018.

[102]KIM H,MNIH A. Disentangling by factorising[C]// International Conference on Machine Learning. PMLR, 2018: 2649-2658.

[103]CHEN R T Q,LI X,GROSSE R B,et al. Isolating sources of disentanglement in variational autoencoders[J/OL]. Advances in Neural Information Processing Systems,2018,31. https://proceedings. neurips. cc/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract. html.

[104]KIM M,WANG Y,SAHU P,et al. Relevance factor VAE: Learning and identifying disentangled factors[J]. arXiv: 1902.01568,2019.

[105]CHEN X,KINGMA D P,SALIMANS T,et al. Variational lossy autoencoder[J]. arXiv:1611.02731,2016.

[106]ZHAO S,SONG J,ERMON S. Infovae:Balancing learning and inference in variational autoencoders[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:5885-5892.

[107]KUMAR A,SATTIGERI P,BALAKRISHNAN A. Variational inference of disentangled latent concepts from unlabeled observations[J]. arXiv:1711.00848,2017.

[108]HAN Q,CAI Y,ZHANG X. RevColV2:Exploring Disentangled Representations in Masked Image Modeling[J]. arXiv: 2309.01005,2023.

[109]ARTHUR G,OLIVIER B,ALEX S,et al. Measuring statistical dependence with Hilbert-Schmidt norms[C]// Algorithmic Learning Theory. 2005:63-77 .

[110]LOPEZ R,REGIER J,JORDAN M I,et al. Information constraints on auto-encoding variational bayes[J/OL]. Advances in Neural Information Processing Systems,2018,31. https://proceedings. neurips. cc/paper/2018/hash/9a96a2c73c0d477ff2a6da3bf538f4f4-Abstract. html.

[111]ESMAEILI B,WU H,JAIN S,et al. Structured disentangled representations[C]// The 22nd International Conference on Artificial Intelligence and Statistics. PMLR,2019:2525-2534.

[112]CHEN X,DUAN Y,HOUTHOOFT R,et al. Infogan:Interpretable representation learning by information maximizing generative adversarial nets[J/OL]. Advances in Neural Information Processing Systems,2016,29. https://proceedings. neurips. cc/paper_files/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract. html.

[113]SINGH K K,OJHA U,LEE Y J. Finegan:Unsupervised hierarchical disentanglement for fine-grained object generation and discovery[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:6490-6499.

[114]LI Y,SINGH K K,OJHA U,et al. Mixnmatch:Multifactor disentanglement and encoding for conditional image generation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:8039-8048.

[115]LI X,CHEN L,WANG L,et al. SCGAN:Disentangled Representation Learning by Adding Similarity Constraint on Generative Adversarial Nets[J/OL]. IEEE Access, 2018: 147928-147938. https://ieeexplore. ieee. org/document/8476290/.

[116]OJHA U,SINGH K K,LEE Y J. Generating furry cars:Disentangling object shape & Appearance across Multiple Domains[J]. arXiv:2104.02052,2021.

[117]LARSEN A B L,SØNDERBY S K,LAROCHELLE H,et al. Autoencoding beyond pixels using a learned similarity metric[C]// International Conference on Machine Learning. PMLR, 2016:1558-1566.

[118]ROSCA M,LAKSHMINARAYANAN B,WARDE-FARLEY D,et al. Variational approaches for auto-encoding generative adversarial networks[J]. arXiv:1706.04987,2017.

[119]BASS C,DA SILVA M,SUDRE C,et al. Icam:Interpretable classification via disentangled representations and feature attribution mapping[J]. Advances in Neural Information Processing Systems,2020,33:7697-7709.

[120]LIU Z,LUO P ,WANG X. ,et al. Deep learning face attributes in the wild[C]// ICCV. 2015.

[121]MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781,2013.

[122]VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017:5998-6008.

[123]DEVLIN J,CHANG M W,LEE K,et al. Bert:Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805,2018.

[124]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.

[125]SUN C,MYERS A,VONDRICK C,et al. Videobert:A joint model for video and language representation learning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:7464-7473.

[126]LI L H,YATSKAR M,YIN D,et al. VisualBERT:A Simple and Performant Baseline for Vision and Language[J/OL]. 2019. https://zhuanlan. zhihu. com/p/535357931.

[127]CHEN Y C,LI L,YU L,et al. UNITER:UNiversal Image-TExt Representation Learning[C]// European Conference on Computer Vision. Cham:Springer,2020.

[128]SU W,ZHU X,CAO Y,et al. Vl-bert:Pre-training of generic visual-linguistic representations[J]. arXiv:1908.08530,2019.

[129]LI G,DUAN N,FANG Y,et al. Unicoder-VL:A Universal Encoder for Vision and Language by Cross-Modal Pre-Training[J]. Proceedings of the AAAI Conference on Artificial Intelligence,2020,34(7):11336-11344.

[130]ALBERTI C,LING J,COLLINS M,et al. Fusion of Detected Objects in Text for Visual Question Answering[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference

on Natural Language Processing (EMNLP-IJCNLP). 2019.
[131]HUANG Z, ZENG Z, LIU B, et al. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers[J]. arXiv: 2004.00849, 2020.
[132]HUANG Z, ZENG Z, HUANG Y, et al. Seeing out of the box: End-to-end pre-training for vision-language representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12976-12985.
[133]KIM W, SON B, KIM I. Vilt: Vision-and-language transformer without convolution or region supervision[C]// International Conference on Machine Learning. PMLR, 2021: 5583-5594.
[134]LI X, YIN X, LI C, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks[C]// European Conference on Computer Vision. Cham: Springer, 2020: 121-137.
[135]ZHANG P, LI X, HU X, et al. VinVL: Making Visual Representations Matter in Vision-Language Models[J/OL]. 2021. https://ieeexplore.ieee.org/document/9577951.
[136]HU X, YIN X, LIN K, et al. VIVO: Surpassing Human Performance in Novel Object Captioning with Visual Vocabulary Pre-Training[J/OL]. 2020. http://arxiv.org/abs/2009.13682.
[137]LU J, BATRA D, PARIKH D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J/OL]. Advances in Neural Information Processing Systems, 2019, 32. https://proceedings.neurips.cc/paper_files/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.
[138]TAN H, BANSAL M. Lxmert: Learning cross-modality encoder representations from transformers[J]. arXiv: 1908.07490, 2019.
[139]LU J, GOSWAMI V, ROHRBACH M, et al. 12-in-1: Multi-Task Vision and Language Representation Learning[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
[140]SUN Y, WANG S, LI Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv: 1904.09223, 2019.
[141]YU F, TANG J, YIN W, et al. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 3208-3216.
[142]LI C, YAN M, XU H, et al. Semvlp: Vision-language pre-training by aligning semantics at multiple levels[J]. arXiv: 2103.07829, 2021.
[143]LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 201-216.
[144]FAGHRI F, FLEET D J, KIROS J R, et al. Vse++: Improving visual-semantic embeddings with hard negatives[J]. arXiv: 1707.05612, 2017.
[145]RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PMLR, 2021: 8748-8763.
[146]JIA C, YANG Y, XIA Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]// International Conference on Machine Learning. PMLR, 2021: 4904-4916.
[147]LI Y, LIANG F, ZHAO L, et al. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm[J]. arXiv: 2110.05208, 2021.
[148]CHEN Y C, LI L, YU L, et al. Uniter: Universal image-text representation learning[C]// European Conference on Computer Vision. Cham: Springer, 2020: 104-120.
[149]LI J, SELVARAJU R, GOTMARE A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in Neural Information Processing Systems, 2021, 34: 9694-9705.
[150]WANG W, BAO H, DONG L, et al. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts[J]. arXiv: 2111.02358, 2021.
[151]YANG H H, AMARI S I. Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information[J]. Neural Computation, 1997, 9(7): 1457-1482.
[152]EASTWOOD C, WILLIAMS C K I. A framework for the quantitative evaluation of disentangled representations[C]// International Conference on Learning Representations. 2018.
[153]DO K, TRAN T. Theory and Evaluation Metrics for Learning Disentangled Representations[J]. arXiv: 1908.09961, 2019.
[154]JIAO X, YIN Y, SHANG L, et al. Tinybert: Distilling bert for natural language understanding[J]. arXiv: 1909.10351, 2019.

WANG Shuaiwei, born in 1999, postgraduate. His main research interests include few-shot video segmentation and visual representation learning.

LEI Jie, born in 1991, assistant professor. His main research interests include deep network optimization and visual representation learning。

（责任编辑：喻藜）