山西大学学报(自然科学版)47(1):81－92,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023142

基于注意力机制的语义对比学习算法

陈俊芬,吕巧莉,谢博鋆*,孙劲松
(河北大学 数学与信息科学学院 河北省机器学习与计算智能重点实验室,河北 保定 071002)

摘　要:对比学习中不合适的数据增强会导致语义信息的失真,同一图像在不同类型的数据增强下语义信息有巨大的语义差距;此外,卷积神经网络(CNN)对纹理有强烈偏好,无法精准学习到下游任务所需的深层语义特征表示,针对以上问题,本文提出一种基于注意力的语义对比学习方法(Semantic attention contrastive learning method,SACL)。SACL首先利用卷积神经网络提取特征,然后注意力模块挖掘全局特征,获得更高级的语义特征,实现了对低级特征的补充和深层特征的语义融合。其次使用截然不同的数据增强方式构造正负样本对,将弱增强(几何增强)生成的正样本和强增强(纹理增强)生成的负样本进行对比,获得差异更为显著的图像输入。网格化增强视图增加了正样本的个数,加快网络收敛速度。在四个数据集上验证了所提出的语义对比学习算法的有效性,结果表明在ImageNet-100数据集上平均精度可以达到78.3%,可以有效提高模型的分类准确率。

关键词:对比学习;注意力机制;语义特征表示;数据增强;纹理;网格化
中图分类号:TP181　　**文献标志码**:A　　**文章编号**:0253-2395(2024)01-0081-12

Semantic Contrastive Learning Algorithm Based On Attention Mechanism

CHEN Junfen, LÜ Qiaoli , XIE Bojun *, SUN Jinsong
(*Hebei Key Laboratory of Machine Learning and Computational Intelligence,College of Mathematics and Information Science, Hebei University, Baoding* 071002*, China*)

Abstract: Inappropriate data augmentation in contrastive learning may lead to distortion of semantic information, and there is a huge semantic gap in semantic information about the same image under different types of data augmentation. In addition, the Convolutional Neural Network (CNN) has a strong preference for textures and cannot accurately learn the deep semantic feature representations required for downstream tasks. In response to the above issues, we propose a method—Semantic attention contrastive learning method (SACL). SACL first utilizes convolutional neural networks to extract features, and then the attention module mines global features to obtain higher-level semantic features, achieving the supplementation of low-level features and semantic fusion of deep features. Secondly, the positive and negative sample pairs are constructed using completely different data augmentation methods, and the positive samples generated by weak enhancement (geometric augmentation) and the negative samples generated by strong enhancement (texture augmentation) are compared to obtain the image input with more significant differences. Gridding augmented view increases the number of positive samples and accelerates network convergence speed. We verified the effectiveness of the proposed semantic contrastive learning algorithm on four datasets, and the results showed that the average accuracy of the ImageNet-100 dataset can reach 78.3%, which can effectively improve the classification accuracy of the model.

Key words: contrastive learning; attention mechanism; semantic feature representation; data augmentation; texture; gridding

收稿日期:2023-06-06;**接受日期**:2023-08-05
基金项目:河北省引进留学人员资助项目(C20200302);河北省教育教学改革研究与实践项目(2020GJJG007)
作者简介:陈俊芬(1976-),女,河北阜城人,博士,副教授,研究方向为机器学习、图像聚类。E-mail:chenjunfen2010@126.com
* **通信作者**:谢博鋆(XIE Bojun),E-mail: xiebojun@126.com
引文格式:陈俊芬,吕巧莉,谢博鋆,等.基于注意力机制的语义对比学习算法[J].山西大学学报(自然科学版),2024,47(1):81-92. DOI:10.13451/j.sxu.ns.2023142

0 引言

最近关于自监督学习的研究表明,在没有人工注释的情况下,视觉表征方面取得了巨大的成功。在预定义代理任务中使用大量未标记数据训练网络来学习特征表示,在完成代理任务后,将学习后的预训练模型进行迁移,在各种下游任务中微调后表现出卓越的性能,例如图像分类[1]、目标检测[2]、图像分割[3]、时间序列检测[4]等。

无监督学习和监督学习之间的差距通过对比学习[5-10]逐渐缩小。对比学习的黄金法则是在学习的实例[11]中,每个图像都被视为一个实例,训练网络使同一实例的不同增强视图的表示尽可能地相互接近,同时来自不同实例的不同视图的表示应该是彼此不同的。现有对比学习主要集中在精心设计的数据增强方面,包括几何变换(裁剪,旋转等)、颜色变换(噪声、模糊等)。Chen等[7]提出使用更强的数据增强方式——高斯模糊,使用更少的标记数据,就可以在分类任务中超越监督学习的性能。虽然这些积极的增强有助于提高模型的性能,但也给训练带来了严重的语义转移问题(图1)。沿用Bai等[12]的称呼,有限数据转换下不产生语义失真的数据增强(例如旋转、翻转)称为弱增强,将因更积极地转换造成图像扭曲语义失真的数据增强(例如纹理、高斯模糊)称为强增强。如何在弱增强以外探索更多强增强是需要考虑的。

视觉领域的局部冗余度是很大的,这种局部性常常会导致低效的计算。卷积操作通过将邻域聚合来减少局部冗余和不必要的计算,然而,有限的感受野使得卷积在学习全局依赖性方面存在困难,需要堆叠很多的卷积层。随着层数加深视野变大,特征图中的语义信息更加高级,学习到下游任务所需的深层语义特征表示更加精准。但是日益复杂的卷积网络结构导致规模和参数更大,带来巨大的计算开销。

并且当数据集的大小有限时,卷积神经网络(CNN)更有可能根据纹理特征对图像进行分类,而不是被人类青睐的形状特征,在Geirhos等[13]尝试从语义特征中获得CNN更多的可解释性,认为CNN更容易学习低级语义特征。特别地,在形状和纹理信息冲突的图像中,与人类表现不同,使用ImageNet训练的CNN,偏向于根据局部纹理而不是全局形状对目标进行分类。不仅如此,在特定条件下获取新信息的能力上CNN与人类认知也呈现相反的趋势。

显式地依赖全局进行建模可能是一种更强大的解决方案。注意力机制[14-16]克服了不同渠道之间缺乏交互的问题,有效地捕获和聚合了上下文信息。Dosovitskiy等[17]、Bao等[18]和Pan等[19]的方法就已经证明受益于捕获全局特征表示。然而注意力机制产生的海量操作导致其极高的复杂性和对GPU内存的高需求,导致训练和测试效率太低。

针对以上问题,本文提出了一种基于注意力机制的语义对比学习方法SACL(Semantic attention contrastive learning method),具体来说:

①使用加入注意力机制的卷积神经网络提取图像深度语义信息,利用注意力机制对全局依赖性进行建模,挖掘图像复杂的全局语义信息,加强全局特征的交互。

②充分利用数据增强策略提升特征表示的鲁棒性。根据CNN更容易识别纹理的偏好,选择纹理数据增强生成弱语义负样本,鼓励模型

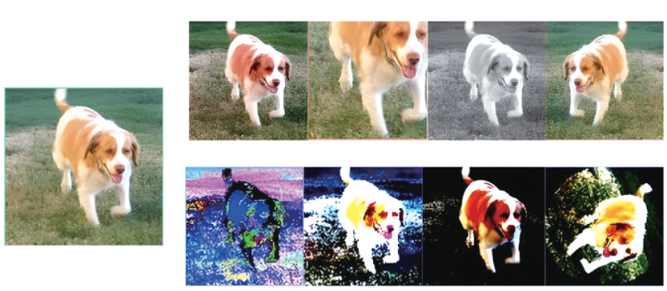


图1 同一图像在弱增强(上)方式和强增强方式(下)的可视化示意图

Fig. 1 The visualization of the weak augmentation (top) and strong augmentation (bottom) of an image

减少对不希望的特征的依赖,更加关注语义特征。将弱语义负样本与正样本进行比较,输入图像在正样本中仍然可以识别,但在负样本中难以理解。

③输入的网格化和组合编码操作产生更多的监督信号,缓解了加入注意力模块产生的复杂性,显著加速了模型运行。

本文在ImageNet-100、CIFAR-10(Canadian Institute For Advanced Research-10)、STL-10(Self-Taught Learning-10)等数据集上做了大量实验,结果显示模型在图像分类上具有较高的优越性。

1 相关工作

1.1 数据增强

对比学习的核心思想[11]是在每个实例判别的设置中,每个图像都被认为是一个单独的类,通过最大化同一图像的不同增强视图之间的相似度,并增加来自不同图像的不同增强视图之间的距离,实现了卓越的下游任务性能,如图2所示,图像领域对比学习经典框架可分为两种类型:基于网络结构形式的变换和基于损失函数的构造。

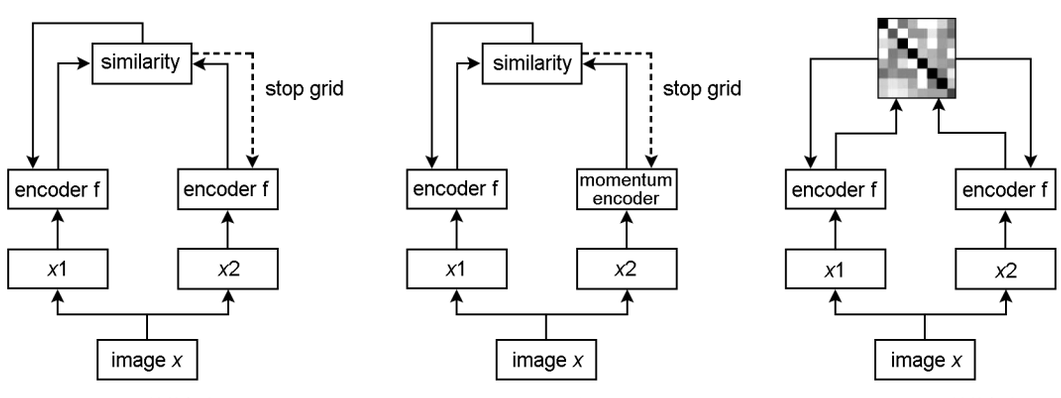
以上方法都采用了一种常见机制来为实例生成样本——积极地增强初始图像,数据增强在深度神经网络的训练中发挥了核心作用,在ImageNet中取得了出色的性能。一方面,它有助于学习到的表示更具有鲁棒性,这使得模型学习到不同变换下不变的表示。另一方面,增强技术也为训练引入了更丰富的数据。

Grill等[8]、Zheng等[20]和Geng等[21]已经讨论过对比学习方法对数据增强的敏感性,关于这个问题的结果得到的结论与本文是一致的——对比学习对数据增强的变化十分敏感。并有Huang等[22]指出足够复杂的数据增强组合对于对比方法更加有利,可转移性与数据增强组合的复杂性呈正相关。在缺少某些增强的情况下,组合缺少复杂性,从而导致表示更少的竞争性。

Chen等[5]表明,对比学习的不变性依赖于数据增强的特征,例如位置或颜色敏感,则可能对下游任务有害。带有随机扰动的数据增强可能会丢失有用的结构信息和语义信息,从而误导表示学习,这种对比方法的缺陷促使人们去寻找替代方案。

为了解决使用增强方式来解决语义信息丢失问题,Shen等[23]提出以真伪人脸信息差异特征的纹理增强作为输入,表示更深层次的语义特征,来很好应用到人脸活体检测。Bai等[12]提出强增强会扭曲图像的结构,导致严重的语义转移问题,同一图像的增强视图可能共享不同的语义,将语义不一致的少数对作为噪声对,通过平衡弱对和积极增强对的作用来抵消语义转移的影响。Abbasi等[24]和Chen等[25]通过使用弱增强来避免这个问题。然而,直接丢弃积极的增强可能会减少训练例子的多样性,导致表示能力有限。因此,在本文中,我们继续保留了积极的增强,并试图抵消随后的语义丢失问题。

针对数据增强的使用方式,一般情况下,正样本是通过数据增强的样本,负样本是一个批



(a) SimSiam 结构框架图 (b) MoCo 结构框架图 (c) Barlow Twins 结构框架图

图2 对比学习典型模型结构框架

Fig. 2 The structures of three typical contrastive models

次内剩下的样本,所以会导致同一个类别的数据,归为了负样本的情况。而Chen等[25]对负样本进行数据增强,提出改进编码器-解码器框架的新颖性检测方式,来解决GAN(Generative Adversarial Networks)[26]中的模式下降问题,Seyfi等[27]同样考虑到被忽略的负样本对直接的相互作用,通过引入一个交叉一致性正则化损失,将变换一致性扩展到不同的图像(负对)。受此启发,尝试在对比学习的正负样本中加入截然不同的数据增强方式,试图缓解语义信息的偏移问题。

1.2 基于CNN的语义偏差

高级场景在拥有数百万个参数训练的深度CNN的同时,也需要复杂的模式和大量的训练数据,这需要大量的计算资源。当数据集的大小有限时,网络往往无法学习场景中语义上重要的区域来进行对象分类,观察到CNN更容易学习低级特征,对识别纹理而不是形状的偏好非常明显。

不同的工作都在试图解决CNN偏差的问题,Zhu等[28]将图像中的视觉提示添加到常规训练的CNN模型中,通过一系列的实验展示了深度CNN是通过学习背景纹理依赖于整个图像的视觉内容实现高性能。相比之下,人类的视觉感知往往关注图像的形状,结果表明,在背景下训练的网络优于人类,这意味着深度神经网络不能根据物体的形状和结构对图像进行分类。Geirhos等[13]使用深度CNN模型来演示纹理偏差,并使用ImageNet数据集的程式化版本进行实验,创建包含形状和纹理的物体的灰度版本,物体轮廓填充黑色的剪影图像,只有边缘的图像,最后是对比形状和纹理的图像,例如一只猫和大象的纹理。表明经典深度CNN图像分类不能识别形状,使用纹理更好。

这些结果证实了CNN在图像训练中,更多关注的是背景和纹理,而不是物体的形状。

1.3 CNN与人类的认知行为差异

除此之外,在获取新信息的能力上,CNN可充分利用的信息与被提供的图片趋势同样呈现与人类认知相反的趋势。具体来说:如图3所示,人类从得到少部分的图像块中和得到几乎所有的图像块的情况下只能提取很少的新信息,而在中等程度下能获取更多有意义的语义信息,而CNN呈现相反的编码交互方式,从而暴露出一种截然不同的分类方式。对于对比学习来说:基于CNN的对比学习更多的是学习表面规律,依赖非语义的底层特征甚至是虚假特征来作出决策,无意识中约束了对高级语义特征的学习,这使得模型在鲁棒性方面有所欠缺。而注意力的处理方式与人类保持高度一致,可以自然有效地找到复杂场景中的突出区域。借用加权求和的过程,增大有用的特征,弱化无用特征,从而起到特征筛选和增强的效果,挖掘图像复杂的全局语义信息,加强全局特征的交互。

1.4 注意力机制

人类可以自然而有效地在复杂的场景中找到突出的区域。基于这一观察结果,注意机制被引入到计算机视觉中,目的是模仿人类视觉系统的这一方面。Hu等[14]首先提出了通道注意的概念,并为此提出了SENet,简单压缩特征图就可以建立通道之间的相互依赖关系,通道注意力学习标量通道权值,提供一种相对计算效率高的重要通道选择,从而对全局平均池化的特征进行通道计算,结构简单且效果显著;CBAM(Convolutional Block Attention Module)[16]通过与大型核的卷积引入空间信息编码,进一步证明了这一想法;Non-local(Non-local Neural Networks)[15]为了减少对外部信息的依赖,尽可能地利用特征内部固有的信息进行注意力的交互,通过自注意力机制对全局的像素对进行建模,有效地捕获长距离的特征依赖,但同时

图3 CNN与人类对于新信息获取能力的认知

Fig. 3 The cognitive gap between CNN and human's ability to acquire new information

也受限于因全局像素点造成的过大计算量而难以很好地嵌入网络中应用。GC-Net(Geometry and Context)[29]受SENet[14]和SENet(Squeeze and Excitation Network)[15]思想的启发提出了一种更简化的空间自注意力模块。考虑到计算开销原因计算量受限和性能增益情况,SENet[12]是最好的选择。

表1 不同注意力机制的特性参数

Table 1 Characteristic parameters of different attention mechanisms

网络	参数量	浮点计算量	性能增益
ResNet[30]	25.56	4.12 G	低
SENet[12]	28.07	4.13 G	高
CBAM[14]	28.07	4.14 G	高
GC-Net[29]	28.11	4.13 G	高
Triplet Attention[31]	25.56	4.17 G	高
ABN(Attention Branch Network)[32]	43.59	7.18 G	低

2 方法

在线网络由一组权值η定义,由三部分组成:编码器$f\eta$、投影仪$g\eta$和预测器$q\eta$。目标网络由一组权值ξ定义,包括编码器$f\xi$和投影仪$g\xi$。在训练期间,在线网络对图像进行网格化处理和组合编码,用于学习图像的整体特征。目标编码器将完整的图像作为输入,通过与在线编码器的对比学习增强特征的可辨别性。

我们先将增强后的视图v进行$n\times n$网格化,再输入在线网络的编码器,为了避免n过大导致图像的数量过多、随机性变差和过小无法提高图像利用率,网格处理的n值选为2。增强视图v平均分成四网格$\{v1,v2,v3,v4\}$,经过编码器$f\eta$后得到

$$h_v^t=f\eta\left(X_v^t\right)\left[v=1,2,3,4\right]。\tag{1}$$

每一个网格图映射到一个潜在特征空间,得到特征编码,将他们拼接成组合特征$y\eta$,为了减少信息损失,两层的MLP(Multi-Layer Perception)投影头$g\eta$将特征映射到另一隐层空间z中,即$z\eta=g\eta(y\eta)$。经过预测头$q\eta$最终形成$q\eta(z\eta)$。

目标网络:将增强视图v'放入经过$f\xi$后提取特征空间h,即$h_v^t=f\xi\left(X_v^t\right)$形成组合嵌入$y\xi$,投影头$g\xi$将特征映射到隐层空间$z$中,即$z\xi=g\xi(y\xi)$。如图4所示,输入的毛色和嘴巴颜色相反的狗在经过纹理增强之后,局部信息保留并更加突出,能体现出更多高级语义,减少语义信息损失。

2.1 数据增强模块

给定N个图片$D=\{x_1,x_2,\cdots,x_N\}$,每个实例x通过预定义的数据增强变换t和t',转换后的视图分别为v和v',表示为$v=t(x)$和$v'=t'(x)$,其中t是一种弱增强(旋转)方式,t'为强增强(纹理)方式。我们使用传统的纹理合成工具——非参数采样纹理合成[23,33-35],将输入图像裁剪为图像块,通过分析图像块的特性来获得联合空间光谱表示,对联合空间-光谱空间中样本的概率分布进行了表征就可以自动生成纹理图像。目的是保持局部结构和细节的同时,增加其与周围环境之间的差异来提高负样本与正样本之间的语义差距。执行t和t'后,变成$2N$个数据样本$\left\{x_1^t,x_2^t\cdots x_N^t,x_1^{t'},x_2^{t'}\cdots x_N^{t'}\right\}$,组对的规律是$\left(x_i^t,x_i^{t'}\right)$为正样本对,剩下的$2N-2$对为负样本对。

2.2 网格化处理模块

增加注意力模块导致原始ResNet[30]参数量增加10%,为了加快模型收敛,尝试对增强视图进行网格化处理。设图像被均分成$n\times n$个的局部不重叠的网格图$G=\left\{X^1,X^2,\cdots,X^{n^2}\right\}$,随机选取$i$个($1\leqslant i\leqslant n^2$)局部网格图放入深度学习进行编码,得到对应特征$g=\left\{g^1,g^2,\cdots,g^{n^2}\right\}$,为了得到高的计算效率,对所选特征进行求和平均处理,最终特征为e,公式为

$$e=\frac{\sum_{gi\in g}gi}{i}。\tag{2}$$

为了提高样本利用率,我们将所有可能的组合特征,记为集合$\left\{A\middle|a\in1,2,\cdots,C_{n^2}^i\right\}$,即$n\times n$个网格图对应的特征中随机选取$i$个进行组合。通过以上方式就可以产生大量的样本,而额外的开销可以忽略不计。这有利于快速收敛和稳定的模型训练。

2.3 基于注意力的特征提取模块

考虑到对深层语义信息的需求,本文尝试探索特征通道之间的关系,显式地建模特征通

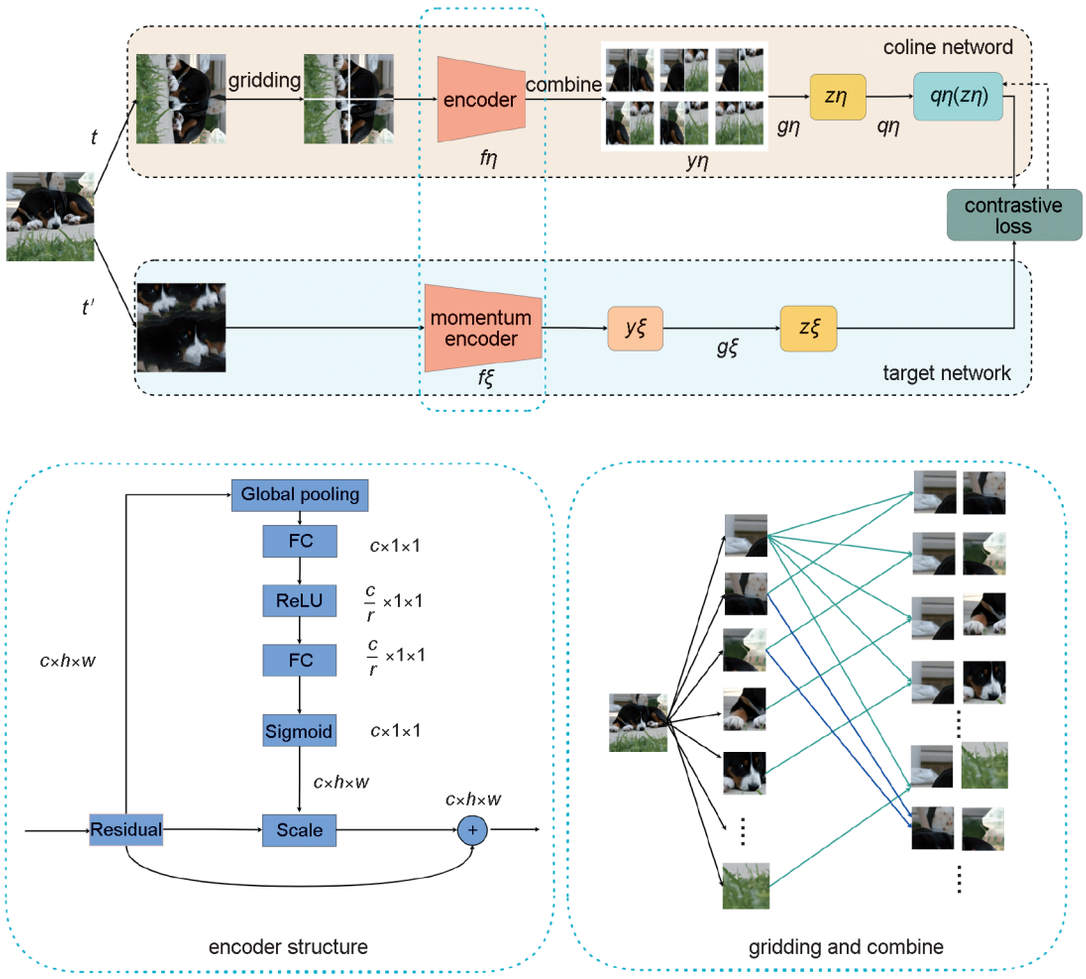


图4 SACL方法的框架结构

(每个实例图像通过预定义的数据增强变换t和t',转换后的视图分别为v和v',经过在线网络和目标网络两个分支处理,目标为被训练的特征$q\eta(z\eta)$和$z\xi$进行对比损失来最大化一致性)

Fig. 4 The framework of the proposed SACL method

(Each instance image is transformed by predefined data enhancement t and t', and the converted views are v and v' respectively, processed by two branches of the online network and the target network, and the target is the trained representations $q\eta(z\eta)$ and $z\xi$ to maximize the consistency to complete a contrastive loss)

道之间的相互依赖关系。遵循SENet[14]的设定,采用一种全新的"特征重标定"策略。通过学习的方式来自动获取到每个特征通道的重要程度,然后依照这个重要程度去提升有用的特征并抑制对当前任务用处不大的特征。将改进后的编码器结构记为SE-ResNet,SACL引入注意力编码器SE-ResNet来更好地将语义信息融合到嵌入空间中。通过考虑通道信息的交互来减少语义信息的损失和加强全局特征的交互。采用改进后的SE-ResNet作为特征提取的主干网络,将语义信息和嵌入空间进行更好的融合。

SENet分为压缩、激励、特征模块融合三个阶段,通道注意力块的输出U与输入u的关系表示如下:

$$S=F_{se}(u,r)=\sigma\left(W_2\delta\left(W_1\mathrm{GAP}(u)\right)\right), \quad (3)$$

$$U=Su, \quad (4)$$

其中r为缩放参数,σ为sigmoid函数,δ为Relu函数,W_1和W_2为网络权值。GAP为全局平均池化。

后续实验结果(见表4)表明,全连接轻量级的模块放入现有的网络结构,只需增加较少的计算量就可以获得性能改善。由于结合了CNN和注意机制的优点,注意力编码器提高特征提

取能力,并提供内在的归纳偏差,从而提高了基于注意力的架构在对比学习环境下的性能。

观察表1得出,在保持性能的前提下,SACL选择参数少的注意力模块SENet[14],但是理论上计算增加SE模块后产生的复杂度得出:与ResNet50相比,理论上,计算损失增加不到1%,GPU推理时间增加10%;CPU推理时间增加不到2%;SE-ResNet50的参数量增加2%~10%。观察图4框架中的编码器模块,得出加入SE注意力机制后,增加的参数主要来自两个全连接层,在特征通道数为C、缩放参数为r的两个全连接层的参数量为$\frac{2C^2}{r}$,若ResNet包含S个stage,每个Stage包含N_S个重复的残差块,输出的通道数量为C_s^2,那么添加了SE块后的增加的参数量为

$$\frac{2}{r}\sum_{s=1}^{S}N_s\cdot C_s^2。\tag{5}$$

模型的收敛速度因为参数量的增加是有所减慢的,为此在模型框架中设计了网格化处理的操作以产生更多的监督信号和显著加速模型运行。

2.4 动量更新

使用梯度回传对在线网络的参数进行更新,从对比损失中产生的梯度被反向传播到在线网络中,目标网络编码器使用在线网络编码器权重的指数移动平均进行更新,在保证动量编码器得到的特征尽可能一致性的前提下使得编码器动量更新,动量更新迫使动量编码器比编码器网络变化得更慢,以此提高模型的泛化能力和稳定训练。

形式上,目标网络编码器$f\xi$被更新为在线网络编码器$f\eta$的指数移动平均,$\theta^{f\eta}$和$\theta^{g\eta}$为在线网络中编码器和投影仪的参数,$\theta^{f\xi}$和$\theta^{g\xi}$是目标网络中编码器和投影仪的参数,m为动量系数,其中$m\in[0,1)$,本文取$m=0.999$。

$$\theta^{f\xi}\leftarrow m\theta^{f\xi}+(1-m)\theta^{f\eta},\tag{6}$$

$$\theta^{g\xi}\leftarrow m\theta^{g\xi}+(1-m)\theta^{g\eta}。\tag{7}$$

2.5 损失函数

本文采用InfoNCE(Info Noise Contrastive Estimation loss)[33]作为对比损失函数,由于训练所用的样本对来自于同一幅图像的两次不同增强,优化目标为正样本对之间的相似度尽可能大,负样本之间的相似度尽可能小。

如上所述,样本x在数据增强处理下得到增强视图v和v',(v,v')为一对正样本对,之后在两个网络分支的处理下提取特征$q\eta(z\eta)$和$z\xi$,进行L2正则化:

$$q_v=\frac{q\eta(z\eta)}{\left\|q\eta(z\eta)\right\|_2}q_{v'}=\frac{z\xi}{\left\|z\xi\right\|_2},\tag{8}$$

sim为相似度度量函数,通过余弦距离测量样本对的相似性,即公式(8),样本对的损失函数为公式(9)。τ为温度参数,$1[k\neq v]$为指示函数,当$[k\neq v]$时候为1,否则为0。在计算损失后对目标网络编码器进行更新。

$$\mathrm{sim}(q_v,q_{v'})=\frac{(q_v)^T(q_{v'})}{\left\|q_v\right\|_2\left\|q_{v'}\right\|_2},\tag{9}$$

$$L_{(v,v')}=-\log\frac{\exp\left(\mathrm{sim}\left(qv,qv'\right)/\tau\right)}{\sum_{k=1}^{2N}1[k\neq v]\exp\left(\mathrm{sim}\left(qv,k\right)/\tau\right)}。\tag{10}$$

样本x的损失为:

$$L_x=\frac{L_{(v,v')}+L_{(v',v)}}{2},$$

每个批次内所有对的损失之和取平均为

$$L_{batch}=\frac{1}{2N}\sum_{k=1}^{N}\left[l(2k-1,2k)+l(2k,2k-1)\right]。$$

算法的伪代码在算法1中描述。

单个卷积层的时间复杂度为:

$$Time\sim O(F^2\cdot K^2\cdot C_{in}\cdot C_{out}),$$

整体卷积神经网络的时间复杂度为:

$$Time\sim O\left(\sum_{d=1}^{D}{F_d}^2\cdot{K_d}^2\cdot C_{d-1}\cdot C_d\right),$$

空间复杂度:

$$Space\sim O\left(\sum_{d=1}^{D}K_d^2\cdot C_{d-1}\cdot C_d+\sum_{d=1}^{D}K^2\cdot C_d\right),$$

其中F为每个卷积核输出特征图的边长;K为每个卷积核的边长;C_{in}为输入通道数;C_{out}为输出通道数;D为卷积层数;d为第d个卷积层;C_d为第d个卷积层的输出通道数。

3 实验

在本节中,我们将在多个基准数据集上进行广泛的实验,以证明算法的有效性和通用性。首先3.1节介绍了数据集。3.2节描述具体

算法1 基于注意力机制的语义对比学习算法

```
# f_η:在线网络[编码器,投影仪,预测器];f_ξ:目标网络[编码器,投影仪]
# m: 动量系数;τ: 温度超参数
f_ξ.params = f_η.params  # 参数初始化
for x in loader:#加载一个 mini batch 的样本
x1,x2=aug(x),aug(x)  #进行数据增强
x1_s,x2_s=split(x1),split(x2)  #网格化:N×C×H×W→4N×C×(H/2)×(W/2)
v1,v2=f_η[0](x1_s),f_η(x2_s)  #经过编码器得到向量表示:
e1,e2=combine(v1),combine(v2)  #组合
z1_e,z2_e=f_η[1:](e1),f_η[1:](e2)  #在线网络投影仪和预测头处理
z1,z2=f_t(x1),f_t(x2)  #目标网络编码器和投影仪处理
loss=(ctr(z1_e,z2)+ctr(z2_e,z1))/2  #对称化
loss.backward( )
updata(f_η.params)  # 在线编码器参数的更新
f_ξ.params=m*f_ξ.params+(1−m)*f_η[:2].params  # 动量编码器参数的更新
# 组合样本的对比损失
loss=0
for_z in z_c:
        logits=mm(_z,z,ξ( ))#mm:矩阵乘法
    loss+=CrossEntropyLoss(logits/τ,labels)
return loss /=len(z_e)
```

实施细节,3.3和3.4节分别为数据集和注意力模块的对比实验结果,消融实验结果见3.5节。

3.1 数据集

实验中使用了四个具有挑战性的图像数据集,分别是CIFAR-10、CIFAR-100、STL-10、和ImageNet-100。

表2 数据集的信息描述

Table 2 Detailed description of the datasets

Dataset	Class	Size	Training Set	Testing Set
CIFAR-10	10	32×32	50 000	10 000
CIFAR-100	100	32×32	50 000	10 000
ImageNet-100	100	224×224	100 000	30 000
STL-10	10	96×96	5 000	8 000

3.2 实施细节

网络架构 使用改进的SE-ResNet-50作为所有数据集的编码器,投影仪$g\eta$和预测器$q\eta$被实现为MLP,投影MLP有3层,全连接层维度是2 048维。每一全连接层都包含BN,输出全连接层中不包含ReLU。预测MLP有两层,隐藏层中有BN操作,输出层中没有BN和ReLU。预测器的输入和输出维度都是2 048,隐藏层维度为512。预训练使用了权重衰减为1×10^{-4}的SGD优化器,动量设置为0.99,温度τ设置为0.2,使用2×2网格化处理方式;初始化学习率0.1,采用余弦学习率调度器。

数据增强 弱增强方式在常见数据增强组成的搜索空间内随机选取,强增强方式依赖于Efros等[33]、Wei等[34]、Han等[35]的经典纹理合成工具,根据从输入图像中提取的两个图像块生成真实的纹理图像。

线性评估 根据Chen等[5]、Grill等[8]、Chen等[7]、Zbontar等[36]、Oord等[37],我们采用了一种通用的线性评估协议来评估SACL在所有数据集上由不同变量下学习到的特征。在主干网络后加上一个线性分类器,冻结主干权重,仅训练线性分类器来验证SACL方法。该分类器由LARS(Layer-wise Adaptive Rate Scaling)优化器[38]进行微调,学习速率设置为0.9。在四个数据集上与多个对比模型进行比较,中等数据集:在ImageNet100数据集,设置批处理大小为128,训练轮数为100。小数据集:对于CIFAR-10、CIFAR-100、STL-10数据集,设置批处理大小为128,训练轮数为200。

3.3 不同方法分类性能对比

本节实验探究不同数据集对模型分类准确率的影响。实验结果如表3、图5所示,可以看出,基于注意力的语义对比学习方法在所有数据集上都能一致地提高框架的性能,验证了SACL的注意力层可充分利用高级语义信息学习聚合权重。

表3 几个对比模型在数据集上的线性分类精度(%)

Table 3 Linear classification accuracy of several contrastive models on the datset

	CIFAR-10	CIFAR-100	STL-10
PIRL[39]	55.78	31.55	50.26
SimCLR[5]	52.58	21.26	44.50
BYOL(Bootstrap Your Own Latent)[8]	80.14	58.28	84.88
Simsiam[40]	75.58	49.21	71.78
MoCov2[7]	82.35	53.44	81.25
SACL	83.74	58.90	84.90

3.4 注意力对比实验

如图6所示,Hu等[14]根据压缩激励模块与残差单元的位置关系,给出标准形式的同时设

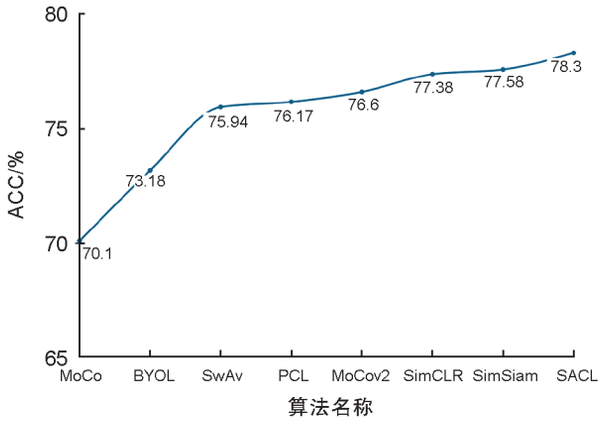


图5 在ImageNet-100上线性分类精度比较

Fig. 5 Comparison of the linear classification accuracy on the ImageNet-100

计了三种变体:SE-PRE(the SE block is moved before the residual unit);模块、SE-POST(the SE unit is moved after the summation with the identity branch);模块、SE-Identity(the SE unit is placed on the identity connection in parallel to the residual unit)模块。从表4中得出结论,多种与残差网络的结合方式中,标准SE与其他三种相比有性能略高的优势,因此继续选择标准的SE模块。

表4 残差模块和SE模块在不同结合方式下的正确率

Table 4 Accuracy of the residual modules and SE modules using different combinations

Combination	SE	SE-PRE	SE-POST	SE-Identity
Top-1 Acc/%	78.0	77.77	77.25	77.84
Top-5 Acc/%	94.0	93.57	93.65	93.85

3.5 消融实验

为了更加深入地研究基于注意力的语义对比学习方法,本节将介绍一系列在CIFAR-10、CIFAR-100、STL-10等小数据集上的消融实验的结果。

分析温度τ的影响:尝试在CIFAR-10和CIFAR-100对SACL模型进行线性评估,从图7中观察得到:为了有效缓解均匀性-容忍性困境,选择合适的温度系数才可以在对比学习中得到有用的表示。比较合适的小温度可以得到更均匀的表示,而如果在温度取值过大时,过分强迫与困难样本分开会破坏学习的潜在语义结构,性能明显下降。

分析批处理大小batchsize的影响:SACL模型训练200个epoch,批处理大小为128;并且随着批处理大小的增大,精度在不断的提升。如表5所示,SACL也可以从更长训练批次的训练中显著受益,我们推测,通过扩大实验规模,可以进一步缩小自监督和监督的差距。

分析动量m的影响:尝试使用不同大小的动量在ImageNet-100对模型进行线性评估,m通常是一个比较大的值,如表6所示,可以看出在$m=0.99$时,精度取到最大为78.0%,m与精度整体呈现u形状,当$m\leqslant 0.5$时模型崩溃。表明保持在线网络编码器和目标网络的编码器一致性是重要的,保持目标网络编码器训练的一致性同样是重要的。

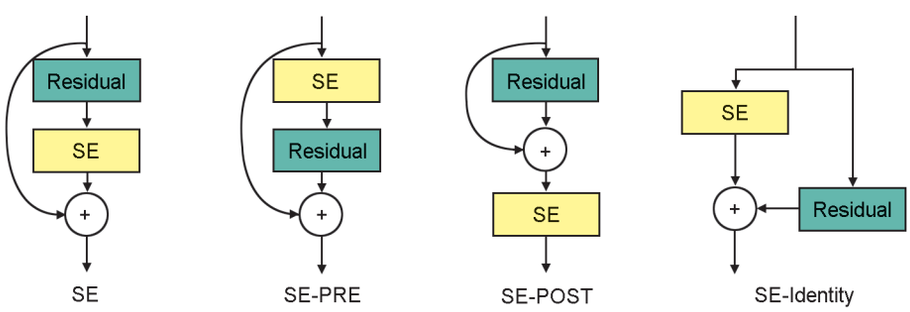


图6 残差模块和SE模块的结合方式

Fig. 6 Multiple combination modes of the residual module and the SE modules block

表5 不同批处理下精度

Table 5 Classification accuracy of the different batches

Dataset	CIFAR-10			CIFAR-100			STL-10		
Batchsize	32	64	128	32	64	128	32	64	128
Acc/%	55.78	78.25	83.74	21.55	41.26	58.90	71.87	79.68	84.90

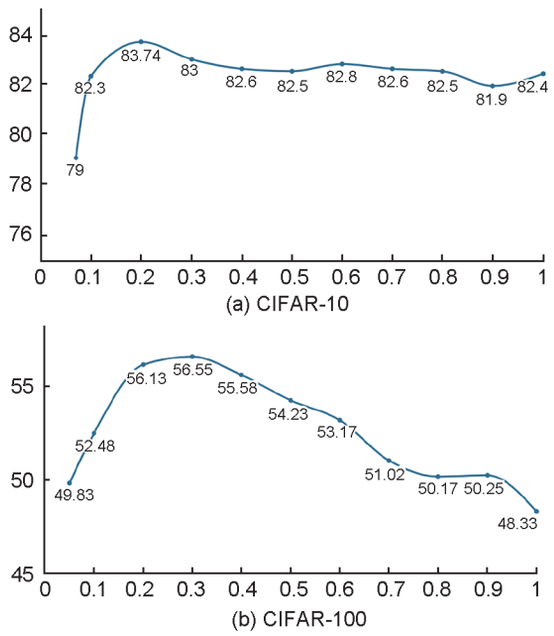


图7 在(a) CIFAR-10和(b) CIFAR-100上不同温度下的精度

Fig. 7 Classification accuracy of the different temperatures on (a) CIFAR-10 and (b) CIFAR-100 datasets

表6 不同动量***m***的分类精度

Table 6 Classification accuracy of the different momentums *m*

m	⩽0.5	0.6	0.7	0.8	0.9	0.99	0.999	1
Acc/%	0	40.36	50.37	57.34	66.17	78.00	73.13	51.21

表7 更长训练轮次的分类精度

Table 7 Classification accuracy of the longer epoches

Model	MoCoV2			SACL		
Epoch	200	400	800	200	400	800
Acc/%	53.44	59.62	63.74	58.90	62.26	65.83

分析训练时间的影响:尝试使用更长的训练时间在CIFAR-100对模型进行线性评估。如表7所示,观察到更多的迭代轮次下,SACL性能仍然优于MoCoV2(Momentum Contrast V2)。

4 结束语

本文提出一种基于注意力机制的语义对比学习方法,尝试从编码器的角度探索,使用加入注意力机制的卷积神经网络作为主干网络,卷积神经网络提取图像深度语义信息,注意力机制对全局依赖性进行建模,挖掘图像复杂的全局语义信息,加强全局特征的交互;同时充分利用数据增强策略提升特征表示的鲁棒性,选择卷积神经网络偏好的数据增强方式作为强增强方式使负样本和正样本呈现语义差异,输入图像在正样本中仍然可以识别,负样本难以理解。并且为了缓解加入注意力模块引起的复杂性,加入网格化处理和组合编码的操作来加速模型运行。综合以上处理,显著提升下游任务的性能,减少语义信息的损失,在ImageNet-100等数据集上与MoCov2[7]等经典网络模型进行对比实验,取得了进步。由于篇幅和时间的问题,本文工作只在分类下游任务中检验改进算法的性能,未来工作可以在多个下游任务中实现改进,如目标检测、语义分割等,让模型更好的应用到实际任务中。

参考文献:

[1] YANG Z Y, WANG J, ZHU Y Y. Few-shot classification with contrastive learning[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 293-309. DOI: 10.1007/978-3-031-20044-1_17.

[2] ZHANG Y X, WANG J D, CHEN Y Q, *et al*. Adaptive Memory Networks with Self-supervised Learning for Unsupervised Anomaly Detection[J]. *IEEE Trans Knowl Data Eng*, 2022(99): 1. DOI: 10.1109/TKDE.2021.3139916.

[3] SAUTIER C, PUY G, GIDARIS S, *et al*. Image-to-lidar Self-supervised Distillation for Autonomous Driving Data[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 9881-9891. DOI: 10.1109/CVPR52688.2022.00966.

[4] WANG X, YIN K, OUYANG Q, *et al*. Identifying Erroneous Software Changes Through Self-Supervised Contrastive Learning on Time Series Data[C]///Proceedings of the 33rd International Symposium on Software Reliability Engineering (ISSRE), 2022: 366-377. DOI: 10.1109/ISSRE55969.2022.00043.

[5] CHEN T, KORNBLITH S, NOROUZI M, *et al*. A Simple Framework for Contrastive Learning of Visual Representations[C]//Proceedings of the 37th International Conference on Machine Learning. 2020: 1597-1607. DOI: 10.5555/3524938.3525087.

[6] HE K M, FAN H Q, WU Y X, *et al*. Momentum Contrast for Unsupervised Visual Representation Learning[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 9726-9735. DOI: 10.1109/CVPR42600.2020.00975.

[7] CHEN X L, FAN H Q, GIRSHICK R, *et al*. Improved Baselines with Momentum Contrastive Learning[J]. arXiv Preprint: 2003.04297, 2020.

[8] GRILL J B, STRUB F, ALTCHÉ F, *et al*. Bootstrap your Own Latent a New Approach to Self-supervised Learning [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 21271−21284. DOI: 10.5555/3495724.3497510.

[9] CARON M, MISRA I, MAIRAL J, *et al*. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 9912−9924. DOI: 10.5555/3495724.3496555.

[10] HJELM R D, FEDOROV A, LAVOIE-MARCHILDON S, *et al*. Learning Deep Representations by Mutual Information Estimation and Maximization[J]. arXiv Preprint:1808.06670, 2018.

[11] WU Z R, XIONG Y J, YU S X, *et al*. Unsupervised Feature Learning *via* Non-parametric Instance Discrimination[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 3733−3742. DOI: 10.1109/CVPR.2018.00393.

[12] BAI Y B, YANG E, WANG Z Q, *et al*. MSR: Making Self−supervised learning Robust to Aggressive Augmentations[J]. arXiv Preprint:2206.01999, 2022.

[13] GEIRHOS R, RUBISCH P, MICHAELIS C, *et al*. ImageNet-trained CNNs are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness[J]. arXiv Preprint:1811.12231, 2018.

[14] HU J, SHEN L, SUN G. Squeeze-and-excitation Networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 7132−7141. DOI: 10.1109/CVPR.2018.00745.

[15] WANG X L, GIRSHICK R, GUPTA A, *et al*. Non-local Neural Networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 7794−7803. DOI: 10.1109/CVPR.2018.00813.

[16] WOO S, PARK J, LEE J Y, *et al*. CBAM: Convolutional Block Attention Module[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3−19. DOI: 10.1007/978-3-030-01234-2_1.

[17] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is Worth 16×16 Words: Transformers for Image Recognition at Scale[J]. arXiv Preprint: 2010.11929, 2020.

[18] BAO H, DONG L, PIAO S, *et al*. Beit: Bert Pre-Training of Image Transformers[J]. arXiv Preprint: 2106.08254, 2021.

[19] 盘展鸿, 朱鉴, 迟小羽, 等. 基于特征融合和注意力机制的图像超分辨率模型[J]. 计算机应用研究, 2022, **39**(3): 884−888. DOI: 10.19734/j.issn.1001-3695.2021.07.0288.
PAN Z H, ZHU J, CHI X Y, *et al*. Image Super-resolution Model Based on Feature Fusion and Attention Mechanism[J]. *Appl Res Comput*, 2022, **39**(3): 884−888. DOI: 10.19734/j.issn.1001-3695.2021.07.0288.

[20] ZHENG M K, YOU S, WANG F, *et al*. Ressl: Relational Self-supervised Learning with Weak Augmentation[J]. *Adv Neural Inf Process Syst*, 2021, **34**: 2543−2555.

[21] 耿传兴, 谭正豪, 陈松灿. 自监督增广的监督分类学习增强[J]. 软件学报, 2023, **34**(4): 1870−1878. DOI: 10.13328/j.cnki.jos.006433.
GENG C X, TAN Z H, CHEN S C. Self-supervisedly Augmented Supervised Classification Enhancement[J]. *J Softw*, 2023, **34**(4): 1870−1878. DOI: 10.13328/j.cnki.jos.006433.

[22] HUANG J Q, KONG X W, ZHANG X Y. Revisiting the Critical Factors of Augmentation-invariant Representation Learning[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 42−58. DOI: 10.1007/978-3-031-19821-2_3.

[23] 沈超, 何希平. 基于纹理特征增强和轻量级网络的人脸防伪算法[J]. 计算机科学, 2022, **49**(6A): 390−396. DOI: 10.11896/jsjkx.210600217.
SHEN C, HE X P. Face Anti-spoofing Algorithm Based on Texture Feature Enhancement and Light Neural Network[J]. *Comput Sci*, 2022, **49**(6A): 390−396. DOI: 10.11896/jsjkx.210600217.

[24] KOOHPAYEGANI S A, TEJANKAR A, PIRSIAVASH H. Mean Shift for Self-supervised Learning[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2022: 10306−10315. DOI: 10.1109/ICCV48922.2021.01016.

[25] CHEN C W, XIE Y, LIN S H, *et al*. Novelty Detection via Contrastive Learning with Negative Data Augmentation[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2021: 606−614. DOI: 10.24963/ijcai.2021/84.

[26] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, *et al*. Generative Adversarial Networks[J]. *Commun ACM*, 2020, **63**(11): 139−144. DOI: 10.1145/3422622.

[27] SEYFI M, BANITALEBI-DEHKORDI A, ZHANG Y. Extending Momentum Contrast with Cross Similarity Consistency Regularization[J]. *IEEE Trans Circuits Syst Video Technol*, 2022, **32**(10): 6714−6727. DOI: 10.1109/TCSVT.2022.3169145.

[28] ZHU Z, XIE L X, YUILLE A. Object Recognition with and without Objects[C]//Proceedings of the Twenty−Sixth International Joint Conference on Artificial Intelli-

gence. California: International Joint Conferences on Artificial Intelligence Organization, 2017: 3609−3615. DOI: 10.24963/ijcai.2017/505.

[29] LIN X , GUO Y, WANG J . Global Correlation Network: End-To-End Joint Multi-Object Detection and Tracking[J]. arXiv Preprint:2103.12511, 2021.

[30] HE K M, ZHANG X Y, REN S Q, *et al*. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770−778. DOI: 10.1109/CVPR.2016.90.

[31] MISRA D, NALAMADA T, ARASANIPALAI A U, *et al*. Rotate to Attend: Convolutional Triplet Attention Module[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021: 3138−3147. DOI: 10.1109/WACV48630.2021.00318.

[32] KWAK J G, HAN D K, KO H. CAFE-GAN: Arbitrary Face Attribute Editing with Complementary Attention Feature[C]//Computer Vision − ECCV 2020: 16th European Conference, 2020: 524−540. DOI: 10.1007/978-3-030-58568-6_31.

[33] EFROS A A, LEUNG T K. Texture Synthesis by Non-parametric Sampling[C]//Proceedings of the Seventh IEEE International Conference on Computer Vision. 2002: 1033−1038. DOI: 10.1109/ICCV.1999.790383.

[34] WEI L Y, LEVOY M. Fast Texture Synthesis Using Tree-structured Vector Quantization[C]//Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM, 2000: 479−488. DOI: 10.1145/344779.345009.

[35] 韩建伟, 王青, 周昆, 等. 基于Wang Tiles的几何纹理合成[J]. 软件学报, 2009, **20**(12): 3254−3264. DOI: 10.3969/j.issn.1006-7043.2011.11.018.
HAN J W, WANG Q, ZHOU K, *et al*. Wang Tile Based Geometric Texture Synthesis[J]. *J Softw*, 2009, **20**(12): 3254−3264. DOI: 10.3969/j.issn.1006-7043.2011.11.018.

[36] ZBONTAR J, JING L, MISRA I, *et al*. Barlow Twins: Self-Supervised Learning *via* Redundancy Reduction [C]//Proceedings of the 38th Int Conference on Machine Learning, 2021: 12310−12320. DOI: 10.21437/Interspeech.2022-11301.

[37] OORD A, LI Y, VINYALS O. Representation Learning with Contrastive Predictive Coding[J]. arXiv Preprint: 1807.03748, 2018.

[38] YOU Y, GITMAN I, GINSBURG B. Large Batch Training of Convolutional Networks[J]. arXiv Preprint: 1708.03888, 2017.

[39] MISRA I, VAN DER MAATEN L. Self-supervised Learning of Pretext-invariant Representations[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 6706−6716. DOI: 10.1109/CVPR42600.2020.00674.

[40] CHEN X L, HE K M. Exploring Simple Siamese Representation Learning[C]//Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 15750−15758. DOI: 10.1109/CVPR46437.2021.01549.