自监督学习综述

摘要:

为了在计算机视觉应用中从图像或视频中获得更好的视觉特征学习性能，通常需要大规模的标记数据来训练深度神经网络。为了避免大规模数据集收集和标注的大量开销.自监督学习作为无监督学习方法的一个子集，在不使用任何人类标注的标签的情况下，从大规模无标记数据中学习图像和视频的通用的特征。本文对基于深度学习的自监督通用视觉特征学习方法进行了广泛的综述。首先，描述了该领域的动机、通用流程(pipeline)和术语。在此基础上，总结了常用的用于自监督学习的深度神经网络体系结构。接下来，回顾了自监督学习方法的模式和评价指标，然后介绍了常用的图像数据集以及现有的自监督视觉特征学习方法。最后，总结和讨论了基于基准数据集的定量性能比较方法在图像中的应用。最后，对本文的研究进行了总结，并提出了一套具有发展前景的自监督视觉特征学习方法

关键词:自我监督学习,非监督学习,卷积神经网络,迁移学习,深度学习

1.介绍

1.1 动机

深度神经网络具有很强的学习不同层次通用视觉特征的能力，深度神经网络已被广泛应用于对象检测、[1]、[2]、[3]、语义分割、[4]、[5]、[6]、图像字幕[7]等计算机视觉应用中。从像ImageNet这样的大型图像数据集中训练出来的模型(这里指的是一般是训练的权重信息)被广泛地 用作预先训练的模型，并用于其他任务的微调(fine-tuning)，其主要原因有两个:(1)从大规模的参数不同的数据集提供了一个很好的起点,因此,网络训练其他任务可以收敛更快,(2)网络训练的大规模数据集已经学会了层次结构功能可以帮助减少在训练其他问题时候的过拟合问题,特别是当其他任务中的数据集的标签稀缺小的时候。

深度卷积神经网络(ConvNets)的性能在很大程度上取决于其训练能力和训练数据量。为了提高网络模型的训练能力，人们开发了各种不同的网络结构，并收集了越来越多的数据集。各种网络，包括AlexNet [8]， VGG [9]， GoogLeNet [10]， ResNet[11]，和DenseNet[12]和大规模的数据集如ImageNet[13]、OpenImage[14],这些数据集已经被用于训练非常深的卷积神经网络.随着复杂而精细的卷积神经网络和大规模的数据集的应用，ConvNets的表现不断打破许多计算机视觉任务的最新的水平.

然而，大规模数据集的收集和注释既耗时又昂贵。ImageNet[13]是最广泛用于深度2D卷积神经网络(2DConvNets)预处理的数据集之一，它包含约130万张标记图像，覆盖1000个类，而每个图像都是人工标注的。

为了避免耗时和昂贵的数据标注处理，人们提出了许多自监督方法(self-supervised)来学习大规模未标记图像或视频的视觉特征，不使用任何人工标注(无监督学习)。为了从未标记数据中学习视觉特征，一个流行的解决方案是提出各种借口任务(pretext task)让网络去解决，而网络可以通过学习借口任务(pretext task)的目标函数来训练，通过这个过程来学习特征。目前,已经提出了各种自监督学习的借口任务(pretext task)，包括给灰度图像着色[18]，图像修补[19]，图像拼图[20]等。借口任务(pretext task)有两个共同的属性:(1)卷积网络需要捕捉图像或视频的视觉特征来解决借口任务(pretext task)，(2)借口任务(pretext task)可以根据图像属性自动生成伪标签(pseudo labels).

图1: 自我监督学习的一般流程。视觉特征是通过训练卷积神经网络解决预先定义的借口任务(pretext task)所获得的。在完成自我监督学习的借口任务(pretext task)的训练后，将所学习的参数作为一个预先训练的模型，通过微调传递给其他下游计算机视觉任务(downstream tasks)。这些下游任务(downstream tasks)的表现用于评估在无监督学习中所学习特征的质量。在下游任务的知识转移过程中，只有前几层的特征被转移到下游任务.

自监督学习的一般流程如图1所示。在自监督训练阶段，设计了一个预先定义的借口任务用于卷积神经网络的训练，并根据数据的某些属性自动生成借口任务的伪标签。然后训练卷积神经网络学习借口任务的目标函数。在完成自我监督训练后，学习到的视觉特征可以进一步转移到下游任务(特别是当只有相对较小的数据可用时)作为预先训练的模型来提高性能和克服过拟合。通常，浅层捕获一般的低级特性，如边缘、角和纹理，而较深的层捕获与任务相关的高级特性。因此，在监督下的下游任务训练阶段，只传输前几层的视觉特征.

1.2 术语定义(Term Definition)

为了使这个综述易于阅读，我们首先定义在综述中将会使用到的一些术语。

- **人类标注的标签**:人工标注标签是指由人工标注的数据标签
- **伪标签(Pseudo label)**:伪标签是根据借口任务(pretext task)的数据属性自动生成的标签
- **借口任务(Pretext Task)**:借口任务是预先设计好的网络任务，通过学习借口任务的目标函数来学习视觉特征.
- **下游任务(Downstream Task)**:下游任务是用于评估通过自监督学习获得的特征的质量。当训练数据缺乏时，这些任务可以从预先训练的模型中得到很大的好处。在一般情况下，**需要使用人工标注的标签来解决下游的任务**.但是，在某些应用程序中，下游任务可以与借口任务相同，而不使用任何人工注释的标签。
- **非监督学习(Unsupervised Learning)**:无监督学习是指不使用任何人类标注的标签的学习方法
- **自监督学习(Self-supervised Learning)**:自监督学习是无监督学习方法的一个子集。自监督学习是指使用自动生成的标签(伪标签)对卷积神经网络进行训练的一种学习方法。本文只讨论了基于卷积神经网络的视觉特征学习的自监督学习方法，该方法将特征转移到多个不同的计算机视觉任务中来检测无监督学习方法的好坏.

由于在自监督训练过程中不需要人工标注来生成伪标签，因此可以使用非常大的数据集进行自监督训练。通过这些伪标签的训练，自监督方法取得良好的效果，与监督方法相比下游任务中的性能差距变小。本文综述了基于深度卷积神经网络的自监督视觉特征学习方法。本文的主要贡献如下:

- 这是一篇关于深度卷积自监督视觉特征学习的综合调查，对这一领域的研究有一定的帮助。
- 最近开发的自我监督学习方法和数据集的深入审查。
- 对现有方法进行了定量性能分析和比较。
- 指出了自监督学习未来可能的发展方向。

2 .不同学习模式的制定

基于训练标签，视觉特征学习方法可以分为四类:监督、半监督、弱监督和非监督。在这一节中，比较了四种类型的学习方法，并定义了关键术语.

2.1 监督学习公式

对于监督学习,给定一个数据集X,对于X中的每一个样本X_i都有一个与之唯一对于的Y_i,对于N个已经被标记的训练数据$D = \{X_i\}_{i=0}^{N}$.训练的损失函数为:

$$\text{loss}(D) = \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \text{loss}(X_i, Y_i) \tag{1}$$

用准确的人工标注的标签进行训练,监督学习方法在不同的计算机视觉任务上面获得了图突破性的结果[1],[4],[8],[16]。然而，数据收集和注释通常是花费巨大的，并且有时候可能还需要特殊的技能。为此，研究者们提出了半监督、弱监督和无监督学习方法来降低成本.

2.2 自监督学习

近年来，许多用于视觉特征学习的自监督学习方法被提出来，这些学习方法没有使用任何人类标注的标签[23]，[24]，[25]，[26]，[27]，[28]，[29，[30]，[31]，[32]，[33]，[33]，[34]，[35]。一些论文将这种学习方法称为无监督学习[36]，[37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48].与需要数据标注的监督学习方法相比,自监督学习使用数据X_i和pretext task任务生成的伪标签P_i来训练。可以使用图像或视频的属性,如图像的上下文[18]，[19，[20]，[36]来生成伪标签，也可以使用传统的手工设计方法[49]，[50]，[51]来生成伪标签P_i。给定一组N个训练数据$D = \{P_i\}_{i=0}^{N}$，将训练loss函数定义为:

$$\text{loss}(D) = \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \text{loss}(X_i, P_i) \tag{2}$$

只要在不涉及人工标注的情况下自动生成的伪标签P_i,就属于自监督学习.近年来,自监督学习方法取得了很大的进展.

本文研究的是基于视觉特征学习的自监督学习方法,这种学习方法主要是针对视觉特征的学习而设计的,而这些特征具有从有限的标记数据中学习并转移到多个视觉任务和执行新任务的能力.本文从网路结构,常用的借口任务,数据集,应用等方面综述了自监督特征学习方法.

3 常见的深度网络框架

无论哪一种自监督学习方法,他们在都使用相似的深度网络结构,本文回顾学习图像和视频的常见架构.

3.1 图像特征学习的架构

无论哪一种自监督学习方法,他们在都使用相似的深度网络结构,本文回顾学习图像和视频的常见架构.

针对图像特征学习设计了多种2D卷积神经网络。在这里，我们回顾了图像特征学习的5个里程碑架构，包括AlexNet[8]、VGG[9]、GoogLeNet[10]、ResNet[11]和DenseNet [12].

3.1.1 AlexNet

AlexNet在ImageNet数据集上的图像分类性能与以往最先进的方法[8]相比有了很大的提高。在强大的gpu支持下，拥有6240万参数的AlexNet在ImageNet上进行了130万张图像的训练。如图2所示，AlexNet架构共有8层，其中5层为卷积层，3层为全连接层。ReLU在每个卷积层之后应用。94%的网络参数来自于完全连接的层。有了这种比例的参数，网络很容易被过度拟合。因此，采用不同的技术来避免过拟合问题，包括数据增强、Dropout和数据归一化。

图2:AlexNet[8]的架构。数字表示每个feature map的通道数。

3.1.2 VGG

VGG由Simonyan和Zisserman提出，并在2013年ILSVRC竞赛[9]中获得第一名。Simonyan和Zisserman提出了各种深度的VGG网络，而16层VGG因其模型尺寸适中、性能优越而被广泛使用。VGG-16的结构如图3所示。它有16个层，属于5个卷积块。VGG和AlexNet之间的主要区别是,AlexNet的卷积步幅和卷积核尺寸比较大,VGG具有相同体积(33*)卷积和小的卷积步幅*11)。大卷积核导致太多的

参数和模型规模大,而大的卷积步长可能会导致网络较低的层的错过一些很好的特征。较小的卷积核尺寸使得在保留网络中细粒度信息的同时使得更深的卷积神经网络的训练成为可能。

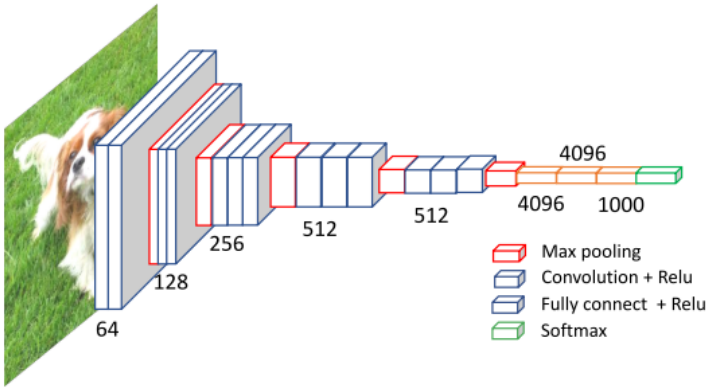


图3:VGG[9]的架构

3.1.3 ResNet

VGG证明了更深层次的网络可以获得更好的性能。然而，由于存在两个问题，较深的网络更加难以训练:梯度消失和梯度爆炸。ResNet是由He等人提出的，通过将前一个feature map发送到下一个卷积，利用卷积块中的skip连接块来克服梯度消失和梯度爆炸[11]。跳跃连接的细节如图4所示。通过跳跃连接，在gpu上进行深度神经网络的训练成为可能.

图4所示. ResNet块[11]的体系结构,能够有效地减少梯度消失和梯度爆炸，使更深层次网络的训练成为可能.

在ResNet[11]中，He等人也评估了不同深度的网络进行图像分类。由于ResNet模型体积小，性能优越，常被用作其他计算机视觉任务的基础网络。具有跳连接的卷积块也被广泛用作基本的构建块

3.1.4 GoogLeNet

GoogLeNet，一个22层的深层网络，是由Szegedy等人获得了ILSVRC-2014挑战赛的冠军，其测试精度最高的前5名为93.3%[10]。Szegedy等人为了构建更深层次的网络，与之前的工作相比，探索构建一个更广的网络，其中每一层都有多个并行的卷积层。GoogLeNet的基本块是inception块，它由4个不同核大小的并行卷积层组成，然后进行1×1的卷积以降维。GoogLeNet的inception模块的架构如图5所示。通过精心设计，他们增加了网络的深度和宽度，同时保持计算成本不变.

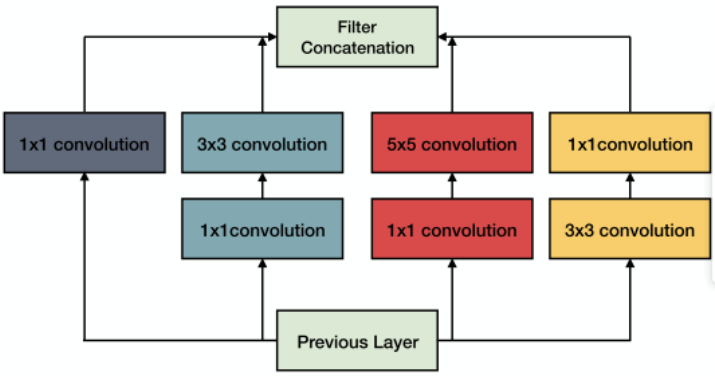


图5.inception块的结构

3.1.5 DenseNet

包括AlexNet、VGG和ResNet在内的大多数网络都遵循层次结构。图像被输入到网络中，由不同的层提取特征。浅层提取低级通用特性，而深层提取高级任务特定特性[52]。然而，当网络越深入，更深的层次可能会因为要记住网络完成任务所需的底层特征而受累。

为了缓解这个问题，Huang等人提出了稠密连接，将卷积块之前的所有特征作为输入发送到神经网络[12]中的下一个卷积块。如图6所示，之前所有卷积块的输出特征作为当前块的输入。这样，较浅的块集中于低级的通用特性，而较深的块可以集中于高级的特定于任务的特征。

3.2 卷积网络总结

深度卷积神经网络在各种计算机视觉任务中显示出了巨大的潜力。图像和视频特征的可视化表明，这些网络确实学习了响应核心任务[52].[78]、[79]、[80]所需要的有意义的特征。然而，一个常见的缺点是，当网络的参数过多并且数据量太少的时候,这些网络很容易过度适合.

为了在不需要昂贵的人工标注的情况下从大规模数据集中获得预先训练好的模型，提出了许多自监督学习方法来从预先设计的借口任务中学习图像和视频特征。下一节描述自监督图像和视频特征学习的一般流程.

4 常用基于变换的借口任务及下游任务

以"预测变换"作为自监督信号进行训练的模型

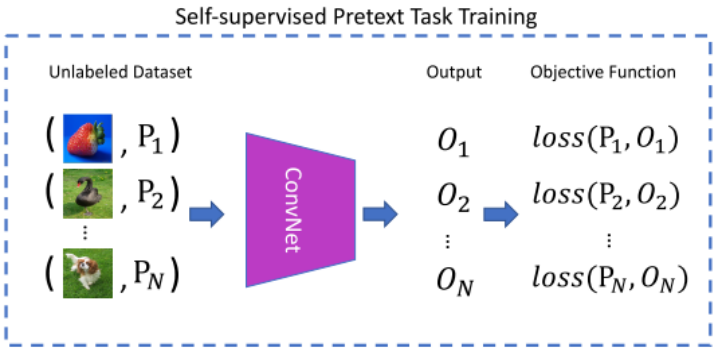


图6:自我监督的视觉特征学习图。通过最小化卷积神经网络伪标签P和预测O之间的误差来训练卷积神经网络。由于伪标签是自动生成的，所以在整个过程中不涉及人工注释的标签.

现有的大多数自监督学习方法都使用图6所示的模式。一般为卷积神经网络定义一个借口任务，通过完成这个借口任务的过程来学习视觉特征。可以自动生成用于借口任务的伪标签P，而无需人工注释。通过最小化ConvNet预测的O与伪标签p之间的损失函数，优化卷积网络。在完成借口任务的训练后，获得能够捕捉图像或视频视觉特征的卷积模型。

4.1 从借口任务中学习视觉特征

为了减轻大规模数据集注释的负担，通常设计一个用于网络优化的借口任务，根据数据属性自动生成借口任务的伪标签去优化卷积网络。目前,设计和应用了许多用于自监督学习的借口任务，如前景对象分割[81]、图像修复[19]、聚类[44]、图像着色[82]、时间顺序验证等等。有效的借口任务是通过完成借口任务的过程来学习语义特征的.

以将灰度图像上色为例，彩色化就是将灰度图像转化为彩色图像的过程。为了生成逼真的彩色图像，网络需要学习图像的结构和上下文信息。在这个借口任务中，数据X是对RGB图像进行线性变换生成的灰度图像，而伪标签P是RGB图像本身。可以实时生成训练对X_i和P_i，成本可以忽略不计。其他借口任务的自监督学习遵循类似的流程.

4.2 常用的借口任务

根据用于设计借口任务的数据属性，如图10所示，我们将借口任务归纳为四类:基于生成、基于上下文、基于自由语义标签和基于跨模式。

基于生成的方法:这类方法通过解决涉及图像或视频生成的借口任务来学习视觉特征

- **图像生成**:通过图像生成任务的过程来学习视觉特征。这类方法包括图像着色[18]，图像超分辨率[15]，图像修复[19],用生成对抗网络(GANs)来生成图像.
- **视频生成**:通过视频生成任务的过程来学习视觉特征。这类方法包括GANs视频生成[85]、[86]和视频预测[87]

  基于上下文的借口任务:基于图像内容的借口任务的设计主要利用图像或视频的内容特征，如内容相似性、空间结构、时间结构等
- **基于内容的相似性**:根据图像块之间的上下文相似度设计虚拟任务。这类方法包括基于图像聚类的方法[34]、[44]和基于图约束的方法[43]。
- **基于context空间结构**:基于图像块之间的空间关系，提出了一种新的训练卷积神经网络的方法。这类方法包括图像拼图[20]，[87]，[88]，[89]，上下文预测[41]，几何变换记录[28]，[36]等.
- **基于context时间结构**:利用视频中的时间序列作为监控信号。训练卷积网络来验证输入帧序列或者识别帧序列是否以正确的顺序[40],[90].
- **基于自由语义标签的方法**:这种类型的借口任务用自动生成的语义标签训练网络。标签由传统的硬编码算法[50]、[51]或游戏引擎[30]生成。其借口任务包括运动目标分割[81]、[91]、轮廓检测[30]、[47]、相对深度预测[92]等.
- **基于辨别Instance的无监督训练**:以contrastive loss为代表的通过辨别不同instance来对实现自监督的训练。把每个样本看作是一个个独立的伪类别，用来无监督的训练网络。

4.3 常用下游任务评价

通过监督学习的方式去评估从图像或视频中学习到特征的质量,通过自监督学习到的模型作为下游任务的预训练的模型,比如说图像分类,语义分割、目标检测、行为识别等。通过迁移学习在这些高级视觉任务上的表现证明了所学习特征的泛化能力。如果自监督学习的卷积神经网络可以学习通用的特征，那么预训练模型可以作为其他视觉特征的任务的良好起点.

图像分类、语义分割和对象检测通常被用作评估无监督方式学习到的特征的泛化能力的任务.下面简要介绍了常用的用于视觉特征评估的高级任务.

4.3.1 语义分割

语义分割是为图像中的每个像素分配语义标签的任务，在自动驾驶、人机交互和机器人技术等许多应用中具有重要意义。目前，全卷积网络(FCN)[4]、DeepLab[5]、PSPNet[6]、PASCAL VOC[96]、CityScape[97]、ADE20K等数据集已经陆续推出，社区发展前景良好.

在所有这些方法中，FCN[4]是语义分割的一个里程碑，因为它开启了应用全卷积网络(FCN)来解决这一任务的时代。FCN的架构如图11所示。以AlexNet、VGG、ResNet等2DConvNet作为特征提取的基础网络，将全连接层替换为置换卷积层，得到稠密的预编码。网络通过像素级的注解进行端到端的训练.

当使用语义分割作为下游任务评价通过自我监督学习学习到的图像的质量,学习的特征用借口任务学习到的特征初始化FCN然后为语义分割数据集上微调,然后在将语义分割评估任务的表现与其他自监督学习方法相比较.

图7:提出了全卷积神经网络的语义分割框架。

4.3.2 物体检测

目标检测是一项定位目标在图像中的位置并识别目标的类别的任务，对于自动驾驶、机器人技术、场景文本检测等计算机视觉应用也非常重要。近年来，提出了许多用于目标检测的数据集，如MSCOCO[99]和OpenImage[14]，提出了许多基于卷积神经网络的模型，如[1]、[2]、[3]、[100]、[101]、[102]、[103]、[104]等，并取得了良好的性能

Fast-RCNN[2]是一个two-stage对象检测网络。Fast-RCNN的框架如图8所示。基于卷积神经网络生成的特征图生成目标建议，然后将这些建议反馈给几个全连接层，生成目标的边界框和目标的类别。

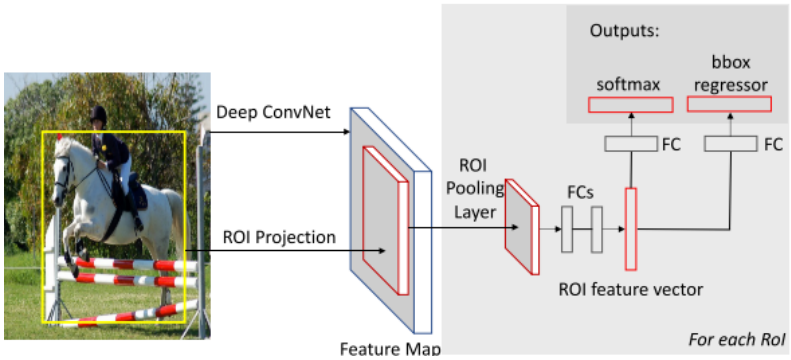


图8.Fast-RCNN用于物体检测的流程

当使用对象检测作为下游任务评估自监督学习训练图像产生的特征的质量时,将自监督中借口任务从大规模无标注的数据中训练出来的模型(参数)作为Fast-RCNN 预训练的模型(权重)[2],然后调整对象检测数据集,然后在目标检测这个任务上做实验观察其表现性能是为了评估证明自监督学习的泛化能力。

4.3.3 图像的分类

图像分类是一项识别图像中物体类别的任务。许多网络已经被设计用于这个任务，如AlexNet [8]，VGG [9]， ResNet [11]，GoogLeNet [10]， DenseNet[12]等。通常，每个图像只有一个类标签可用，尽管图像可能包含不同的对象类。

在选择图像分类作为下游任务评价从自监督学习获得的图像特征的质量,将自监督学习得到的卷积部分应用在每个图像提取特征,然后用于训练分类器,如支支持向量机(SVM) [105]。将测试数据的分类性能与其他自监督模型进行比较，以评估学习特征的质量.

4.3.4 定性评价

除了这些对学习特征的定量评价外，还有一些定性的可视化方法来评价自监督学习特征的质量。为此，通常使用三种方法:卷积核可视化、特征图可视化和图像检索可视化[28]、[36]、[41]、[44].

- **卷积核可视化** : 定性地可视化第一个从自监督学习中训练的卷积层的卷积核，并和监督学习训练出来的模型的卷积核相比较。通过比较监督模型和自监督模型学习的卷积核的相似性，说明了自监督方法[28]、[44]的有效性.
- **feature map可视化** : 特征图是可视化的，以显示网络的注意力。较大的激活表示神经网络对图像中相应区域的关注程度较高。特征图通常是定性可视化的，并与监督模型可视化[28]、[36]进行比较。
- **最近邻检索(Nearest Neighbor Retrieval)**: 一般情况下，外观相似的图像在特征空间中距离较近。采用最近邻法从自监督学习模型[40]、[41]、[43]学习的特征空间中找出K个最近邻的图像.

5 数据集(DATASETS)

本节总结了用于训练和评估自监督视觉特征学习方法的常用图像和视频数据集。自我监督学习方法可以通过抛弃人类标注的标签，用图像或视频进行训练.所有任何收集到的用于监督学习的数据可以用于不使用人工标注标签的自监督视觉特征学习。对学习特征质量的评估通常是通过对相对较小的数据集(通常带有精确的标签)的高级视觉任务进行微调来进行的，如视频动作识别、对象检测、语义分割等。值得注意的是，本文将使用这些合成数据集进行视觉特征学习的网络视为自监督学习，因为合成数据集的标签是由游戏引擎自动生成的，不涉及人工标注。表1总结了常用的图像和视频数据集.

5.1 图像的数据

- **ImageNet** : ImageNet数据集[13]包含130万张图像，均匀地分布在1000个类中，并根据WordNet层次结构进行组织。每个图像只分配一个类标签。ImageNet是应用最广泛的自监督图像

特征学习数据集。

- **Places** : Places数据集[107]一般用于场景识别，包含超过250万图像，包括超过205个场景类别，每个类别超过5000张图像.
- **Places365** : Places365是place数据库的第二代，该数据库用于高级视觉理解任务，如场景上下文、对象识别、动作和事件预测以及心理理论推理[108]。有超过1000万张图片涉及400多个类别，每个类别有5000到30000张训练图片.
- **SUNCG** : SUNCG数据集是一个用于室内场景的大型合成3D场景存储库，包含超过45,000个不同的场景，并手动创建了逼真的房间和家具布局[109]。可以使用合成深度、对象级语义标签和容量ground truth.
- **MNIST** : MNIST是一个手写数字的数据集，它包含70,000张图像，60,000张图像属于训练集，其余10,000张图像用于测试[110]。所有数字都进行了大小标准化，并在固定大小的图像中居中。
- **SVHN** : SVHN是一个用于识别自然场景图像中数字和数字的数据集，它来自于谷歌街景图像中的房号[111]。数据集由60多万张图片组成，所有的数字都被调整为固定大小的32*32的图像.
- **CIFAR10** :CIFAR10数据集是用于图像分类的小图像[112]。有10个不同的类别。这10类包括6万张大小为32×32的图片，涵盖了汽车、鸟、猫、鹿、狗、青蛙、马、船和卡车。数据集是平衡的，每个类有6000张图像.
- **STL-10** : STL-10数据集是专门为发展无监督特征学习而设计的[113]。它由500张带标签的训练图像、800张测试图像和10万张未带标签的图像组成,一共10个类别,包括飞机、鸟、汽车、猫、鹿、狗、马、猴、船和卡车。
- **PASCAL Visual Object Classes (VOC)**: VOC 2012数据集[96]包含20个对象类别，包括汽车、家用、动物和飞机、自行车、船、汽车、摩托车、火车、瓶子、椅子、餐桌、盆栽、沙发、电视/显示器、鸟、猫、牛、狗、马、羊和人。该数据集中的每个图像都有像素级的标注、包围框注释和对象类注释。该数据集被广泛用作对象检测、语义分割和分类任务的基准。PASCAL VOC数据集分为三个子集:用于训练的1464张图片，用于valida- tion验证的1449张图片，以及用于私有测试的1449张图片[96]。所有的自监督图像表示学习方法在这个数据集上的评估都是通过这三个任务来进行的.
- **YFCC100M** : The Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M) (YFCC100M)是Flickr上的一个大型公共多媒体集合，包含1亿媒体数据，其中图像约9920万，视频约80万[114]。YFCC100M数据集的hashtags统计表明，数据分布严重不平衡.

6. 图像特征的学习

在这一节中，我们回顾了三组自我监督的图像特征学习方法，包括基于生成的方法、基于context的方法和基于自由语义标签的方法。图像特征自监督学习方法列表如表2所示。由于基于交叉模态的方法主要是从视频中学习特征，而这类方法大多可以同时用于图像和视频的特征学习，因此在视频特征学习一节中，我们将对基于交叉模态的方法进行综述。

Method	Category	Code	Contribution
GAN [83]	Generation	✓	Forerunner of GAN
DCGAN [120]	Generation	✓	Deep convolutional GAN for image generation
WGAN [121]	Generation	✓	Proposed WGAN which makes the training of GAN more stable
BiGAN [122]	Generation	✓	Bidirectional GAN to project data into latent space
SelfGAN [123]	Multiple	✗	Use rotation recognition and GAN for self-supervised learning
ColorfulColorization [18]	Generation	✓	Posing image colorization as a classification task
Colorization [82]	Generation	✓	Using image colorization as the pretext task
AutoColor [124]	Generation	✓	Training ConvNet to predict per-pixel color histograms
Split-Brain [42]	Generation	✓	Using split-brain auto-encoder as the pretext task
Context Encoder [19]	Generation	✓	Employing ConvNet to solve image inpainting
CompletNet [125]	Generation	✓	Employing two discriminators to guarantee local and global consistent
SRGAN [15]	Generation	✓	Employing GAN for single image super-resolution
SpotArtifacts [126]	Generation	✓	Learning by recognizing synthetic artifacts in images
ImproveContext [33]	Context	✗	Techniques to improve context based self-supervised learning methods
Context Prediction [41]	Context	✓	Learning by predicting the relative position of two patches from an image
Jigsaw [20]	Context	✓	Image patch Jigsaw puzzle as the pretext task for self-supervised learning
Damaged Jigsaw [89]	Multiple	✗	Learning by solving jigsaw puzzle, inpainting, and colorization together
Arbitrary Jigsaw [88]	Context	✗	Learning with jigsaw puzzles with arbitrary grid size and dimension
DeepPermNet [127]	Context	✓	A new method to solve image patch jigsaw puzzle
RotNet [36]	Context	✓	Learning by recognizing rotations of images
Boosting [34]	Multiple	✗	Using clustering to boost the self-supervised learning methods
JointCluster [128]	Context	✓	Jointly learning of deep representations and image clusters
DeepCluster [44]	Context	✓	Using clustering as the pretext
ClusterEmbegging [129]	Context	✓	Deep embedded clustering for self-supervised learning
GraphConstraint [43]	Context	✓	Learning with image pairs mined with Fisher Vector
Ranking [38]	Context	✓	Learning by ranking video frames with a triplet loss
PredictNoise [46]	Context	✓	Learning by mapping images to a uniform distribution over a manifold
MultiTask [32]	Multiple	✓	Using multiple pretext tasks for self-supervised feature learning
Learning2Count [130]	Context	✓	Learning by counting visual primitive
Watching Move [81]	Free Semantic Label	✓	Learning by grouping pixels of moving objects in videos
Edge Detection [81]	Free Semantic Label	✓	Learning by detecting edges
Cross Domain [81]	Free Semantic Label	✓	Utilizing synthetic data and its labels rendered by game engines

表二:基于借口任务类的自监督图像特征学习方法综述。多任务是指该方法显式或隐式地使用多个借口任务进行图像特征学习

6.1基于生成的图像特征学习

基于生成自监督的图像特征学习方法涉及生成图像的过程，包括使用GAN生成图像(生成假图像)、使用超分辨率生成高分辨率图像、图像修补(用于预测缺失的图像区域)和图像上色(用于将灰度图像着色为彩色图像)。对于这些任务，伪训练标签P通常是图像本身，训练过程中不需要人为标注的标签，因此这些方法属于自监督学习方法.

基于图像生成方法的先驱工作是自动编码器[131]，它学会了将图像压缩成一个低维向量，然后将其未压缩成与原始图像一致的堆层次的层。利用自动编码器，网络可以将图像降维为包含原始图像主要信息的低维向量。目前基于图像生成的方法都遵循了相似的思想，只是通过不同的流程(pipeline)来学习图像生成过程中的视觉特征.

6.1.1利用GAN生成图像

生成性对抗网络(GAN)是Goodfellow等[83]提出的一种深层生成性模型。GAN模型一般由两种网络构成:一种是由潜在向量生成图像的发生器，另一种是区分输入图像是否由发生器生成的鉴别器。在两方的对抗中，鉴别器强制生成器生成逼真的图像，生成器强制鉴别器提高鉴别能力。在训练中，这两个网络互相竞争，使对方更强大.

从潜在变量任务生成图像的一般架构如图13所示。该算法首先对潜在向量进行训练，然后将潜在向量映射到图像中，然后对真实数据分布和生成的数据分布进行判别。因此，需要识别器从图像中捕获语义特征来完成识别任务。鉴别器的参数可以作为其他计算机视觉任务的预训练模型。

图9所示。生成对抗网络的pipeline[83]。在两方对抗中，鉴别器强制生成器生成逼真的图像，生成器强制鉴别器提高鉴别能力.

数学上,生成器G是训练学习真实世界图像分布P_z去生成和真实数据一样的虚拟数据,对鉴别器D进行训练，区分真实数据P_{data}的分布和生成器G生成的数据分布P_z的分布。发生器G与鉴别器D之间的最小-最大博弈公式为:

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}\ (x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{3}$$

其中x是真实数据，G(z)是生成的数据

对鉴别器D进行训练，使其对真实数据x的概率最大(也就是:$\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)]$),并最小化生成数据G(z)的概率: $\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)]$,训练生成器生成接近真实数据x的数据，从而使鉴别器的输出最大化: $\mathbb{E}_{x\sim p_{data}(x)}[\log D(G(z))]$).

从随机变量生成图像的大多数方法不需要任何人为注释的标签。然而，这类任务的主要目的是生成真实的图像，而不是在下游应用程序上获得更好的性能。通常，生成图像的起初分数被用来评价生成图像的质量[132]，[133]。只有少数几种方法评估了高级任务描述器学习的特征的质量，并与其他方法进行了比较[120]、[122]、[123]。

对抗性训练可以帮助网络捕获真实数据的真实分布，生成真实的数据，在图像生成[134]、[135]、视频生成[85]、[86]、超分辨率[15]、图像平移[136].图像修补[19]，[125]等计算机视觉任务中得到了广泛的应用,。在不涉及人工标注的情况下，该方法属于自监督学习。

6.1.2 修补图像生成

图10所示。绘画任务中图像的定性说明。给定一幅缺失区域(a)的图像，人类艺术家在绘制它时不会遇到问 题(b)。在经过L2重构损失和对抗损失训练的[19]中提出的使用上下文编码器进行自动绘制的方法如(c)所示。

图像修补是一项基于图像其余部分预测任意缺失区域的任务。在图像修补任务中图像的定性说明如图10所示。图10(a)为缺失区域图像，图10(c)为网络预测。为了准确地预测缺失区域，网络需要学习共同的知识，包括共同物体的颜色和结构。只有了解了这些知识，网络才能根据图像的其余部分推断出缺失区域.

与自动编码器类似，Pathak等人首先训练一个卷积神经网络，根据图像[19]的其余部分生成任意图像区域的内容。他们的贡献有两方面:一是使用卷积网络来解决图像修补中的图像问题，二是使用对抗损失来帮助网络生成一个现实的假设。最近的大多数方法都遵循类似的流程(pipeline)[125]。通常有两种网络:生成器网络是用像素级重建损失来生成缺失区域;鉴别器网络是用对抗性损失来区分输入图像是否真实。在对抗性损失的情况下，该网络能够对缺失的图像区域产生更清晰、更真实的假设。这两种网络都能从图像中学习语义特征，并能转移到其他计算机视觉任务中。然而，只有Pathak等人研究了从图像修补任务中对生成器的学习参数进行转移学习的性能。

生成器网络是由全卷积网络构成,它有两部分构成:编码器和解码器。编码器的输入是需要绘制的图像，上下文编码器学习图像的语义特征。上下文解码器就是根据这个特征来预测缺失的区域。为了生成一个合理的假设，需要生成器理解图像的内容,对鉴别器进行训练，以区分输入图像是否是生成器的输出。为了完成图像嵌入任务，两个网络都需要学习图像的语义特征。

6.1.3具有超分辨率的图像生成

图像超分辨率(SR)是提高图像分辨率的一项重要任务。在全卷积网络的帮助下，低分辨率图像可以生成更精细、更真实的高分辨率图像。SRGAN是Ledig等人提出的用于单幅图像超分辨率的广义对抗网络。这种方法的核心是利用感性损失，感性损失包括对抗性损失和内容损失。有了感知器的损失，SRGAN能够从大量向下采样的图像中恢复真实感纹理，并显示出感知质量的显著提高.

有两种网络:一种是增强输入低分辨率图像分辨率的生成器，另一种是区分输入图像是否是生成器的输出的鉴别器。生成器的损失函数是像素方向的L2损失加上内容损失，内容损失是预测的高分辨率图像和高分辨率原始图像特征的相似性，而鉴别器的损失是双元分类损失。网络相比,只有最小化均方误差(MSE),通常会导致较高的峰值信噪比率但缺乏高频细节,SRGAN能够恢复高分辨率图像的细节,因为对抗的损失将输出到自然图像的鉴别器网络.

图像超分辨任务网络能够学习图像的语义特征。与其他GAN网络相似，鉴别器网络的参数可以传递给其他下游任务。然而，还没有人测试转移学习在其他任务上的表现。增强图像的质量主要是用来评价网络的性能.

6.1.4彩色图像生成

图像着色是一项给定一个灰度图像预测一个可信的彩色版本的照片的任务。图像着色任务的定性说明如图15所示。为了正确地给每个像素着色，网络需要识别对象并将相同部分的像素分组在一起。因此，在完成这个任务的过程中可以学习到视觉特征.

图11所示。在[18]中提出了图像着色的结构

近年来，许多基于深度学习的着色方法被提出，[137]，[138]。一个直接的想法是使用一个完整的卷积神经网络，它由一个用于特征提取的编码器和一个用于颜色幻化的解码器组成，可以用L2损失对网络进行优化.Zhang等人提出通过发布分类任务来处理不确定性，并使用类再平衡来增加预测颜色[18]的多样性。Zhang等人提出的图像着色框架如图11所示。经过大规模图像采集的训练，该方法在彩色化测试中取得了良好的效果，并在32%的测试中欺骗了人类.

有些工作专门以图像着色任务为借口，进行自监督图像表示学习[18]，[42]，[82]，[124]。在完成图像着色训练后，通过着色过程学习到的特征将在其他带有转移学习的下游高级任务中进行具体评估.

6.2 基于context的图像特征学习

基于context的借口任务主要利用图像的一些特征来作为监督信号,比如context相似性、空间结构、时间结构等.

6.2.1 context相似的学习

聚类是在相同的集群中对相似数据集进行分组的一种方法。由于其强大的利用数据属性对数据进行分组的能力，在机器学习、图像处理、计算机图形学等领域得到了广泛的应用。许多经典的聚类算法被提出用于各种应用[139].

在自监督场景中，聚类方法主要用于图像数据的聚类。基于HOG[140]、SIFT[141]、Fisher Vector[49]等手工设计的特征对图像数据进行聚类，是一种较为简单的方法。聚类后得到多个簇，其中在同一个簇中的图像在特征空间中的距离较小，不同聚类中的图像在特征空间中的距离较大。特征空间中的距离越小，RGB空间中的图像外观越相似。然后训练一个卷积神经网络来分类.

图12所示。深度聚类[44]的体系结构。通过迭代聚类的方法对网络的特征进行聚类，并将聚类赋值作为伪标签来学习网络的参数。

现有的使用聚类作为借口任务的方法遵循这些原则[34], [43], [44],[128],[129]。首先将图像聚类为不同的聚类，来自同一聚类的图像距离较小，来自不同聚类的图像距离较大。然后训练一个卷积神经网络来识别分配给它的簇[34]、[44]或识别两个图像是否来自同一个簇[43]。基于聚类的方法DeepCluster的pipeline如图12所示。DeepCluster使用Kmeans迭代地对图像进行集群，并使用子序列分配作为监督来更新网络的权重。这是目前最先进的自我监督技术.

6.2.2 空间语境结构的学习

图像包含丰富的空间背景信息，如图像中不同斑块之间的相对位置，可以用来设计自我监督学习的借口任务。其借口任务可以是从相同的图像[41]预测两个patch的相对位置，也可以是从相同的图像[20]，[88]，[89]识别打乱的patch序列的顺序。整个图像的背景也可以作为一个监督信号来设计借口任务，比如识别整个图像[36]的旋转角度。为了完成这些借口任务，ConvNets需要学习空间环境信息，如物体的形状和物体不同部分的相对位置。

图13。七巧板图像拼图[20]的可视化。(a)为9个采样图像块的图像，(b)为打乱图像块的例子，(c)为9个采样图像块的正确顺序

Doersch等人提出的方法是利用空间上下文线索进行自监督视觉特征学习的开创性工作之一。从每个图像中提取随机对的图像patch，然后训练一个卷积神经网络来识别两个图像patch的相对位置。为了解决这个难题，ConvNets需要识别图像中的对象，并学习对象不同部分之间的关系。为了避免网络学习琐碎的解决方案，例如简单地使用补丁中的边来完成任务，在训练阶段应用了大量的数据增强.

基于这一思想，提出了更多的方法来学习图像特征，解决更困难的空间难题[20]，[27]，[87]，[88]，[89]。如图17所示，Noroozi等人提出的一个典型工作是尝试使用ConvNet[20]解决一个图像拼图。图13(a)为9个采样图像块的图像，图13(b)为打乱图像块的例子，图13(c)为9个采样图像块的正确顺序。将变换后的图像块输入到网络中，网络通过学习空间环境图像的结构如物体颜色、结构和高级语义信息来识别输入块的正确空间位置.

6.3基于语义标签的图像特征学习

自由语义标签是指在不涉及任何人类干扰的情况下获得的具有语义意义的标签。一般来说，自由的语义标签，如分段掩码、深度图像、视神经流和表面法线图像，可以由游戏引擎渲染或由硬编码方法生成。由于这些语义标签是自动生成的，因此使用合成数据集或与未标记的大型图像或视频数据集结合使用的方法被认为是自监督学习方法。

6.3.1使用游戏引擎生成的标签进行学习

给定模型的各种对象和环境布局，游戏引擎能够渲染现实的图像，并提供准确的像素级标签。由于游戏引擎能够以极低的成本生成大规模的数据集，因此不同的游戏引擎如Airsim[142]和Carla[143]被用于生成具有高级语义标签的大规模合成数据集，这些语义标签包括深度、轮廓、表面法线、分割掩模和用于训练深度网络的光流。图18所示为生成准确标签的RGB图像示例.

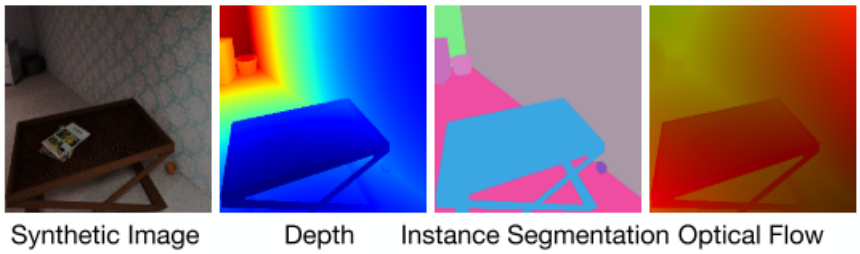


图14所示。一个由游戏引擎生成的室内场景的例子[115]。对于每个合成图像，引擎可以自动生成相应的深度、实例分割和光流。

游戏引擎可以生成逼真的图像与准确的像素级标签并且成本非常低。然而，由于合成图像与真实世界图像的领域差距，单纯训练合成图像的ConvNet不能直接应用于真实世界图像。为了利用合成数据集进行自监督特征学习，需要显式地填补领域的空白。这样，经过语义标签训练的卷积神经网络可以有效地应用于真实世界的图像.

为了克服这一问题，Ren和Lee提出了一种基于对抗性学习[30]的无监督特征空间域自适应方法。如图14所示，该网络对合成图像的表面法线、深度和实例轮廓进行预测，并使用鉴别器网络D来最小化真实数据和合成数据之间的特征空间域差异。借助对抗性训练和精确的合成图像语义标签，该网络能够为真实图像捕获视觉特征。

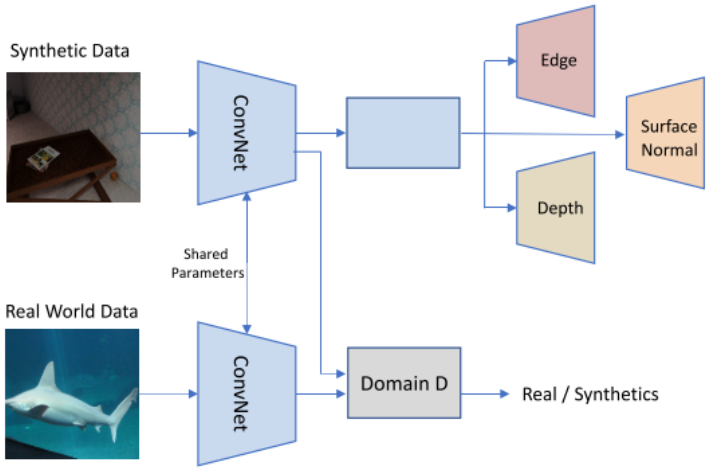


图19所示。利用合成图像和真实图像进行自监督特征学习的体系结构

与其他通过虚拟任务隐式强制学习语义特征的借口任务相比，这种方法通过精确的语义标签训练，显式强制学习与图像中物体高度相关的特征.

6.3.2使用硬编码程序生成的标签进行学习

应用硬编码程序是自动生成语义标签的另一种方式，例如图像和视频的显著性、前景遮罩、轮廓、深度等。通过这些方法，可以使用带有语义标签的大规模数据集进行自监督特征学习。这类方法一般有两个步骤:(1)通过在图像或视频上使用硬编码程序生成标签来获得标签，(2)使用生成的标签训练卷积神经网络。

各种各样的硬编码程序已经被应用于自监督学习方法的生成标签，包括前景对象分割方法[81]、边缘判定[47]方法和相对深度预测方法[92]。Pathak等人提出通过训练卷积神经网络在视频的每一帧中分割前景对象来学习特征，而标签是视频中移动对象的掩码[81]。Li等人提出通过训练一个卷积神经网络进行边缘预测来学习特征，而标签则是从视频中获得的运动边缘[47]。Jing等人提出在光流产生标签的同时，通过训练卷积神经网络来预测相对场景深度来学习特征[92]。

无论用于训练卷积神经网络的标签是什么，这种方法的一般思想是从硬编码检测器中提取出已知边。硬码检测器可以是边缘检测器、显著性检测器、相对检测器等。只要检测器的设计不涉及到人的注释，那么检测器就可以用来生成自我监督训练的标签。

与其他的自监督学习方法相比，这些借口任务中的监督信号是语义信号，可以直接驱动卷积神经网络学习语义特征。然而，一个缺点是由硬代码检测器生成的语义标签通常是非常嘈杂的，需要专门处理。

6.4 基于实例的无监督学习方法

6.4.1 基于NCE的无监督学习方法

通过训练基于实例（将每一个样本视为单独的类别）的分类器代替基于类别的分类器，得到可以捕捉视觉相似性的特征表达。我们将其总结为**非参数化实例级判别**，并且通过**噪声对比估计（noise-contrastive estimation）**解决大量实例类别引起的计算困难。

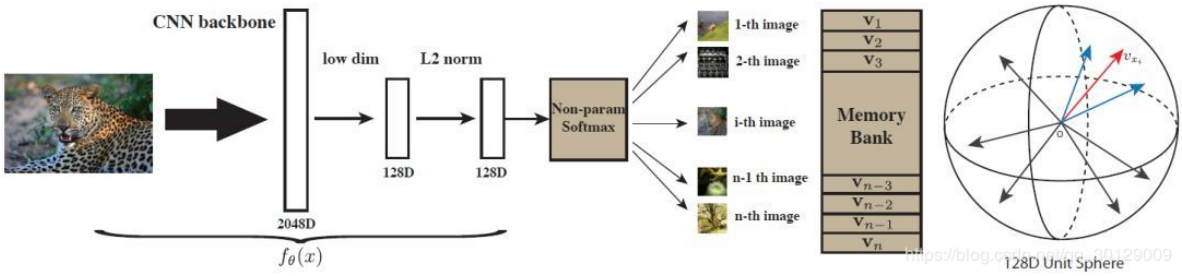


图 20：提出的无监督特征学习方法的工作流图。研究者使用骨干 CNN 将每个图像编码为 128 维空间并进行 L2 归一化的特征向量。最佳特征嵌入过程是通过实例级判别器学习的，该判别器尝试将训练样本的特征最大程度地散布在 128 维的单位球上。

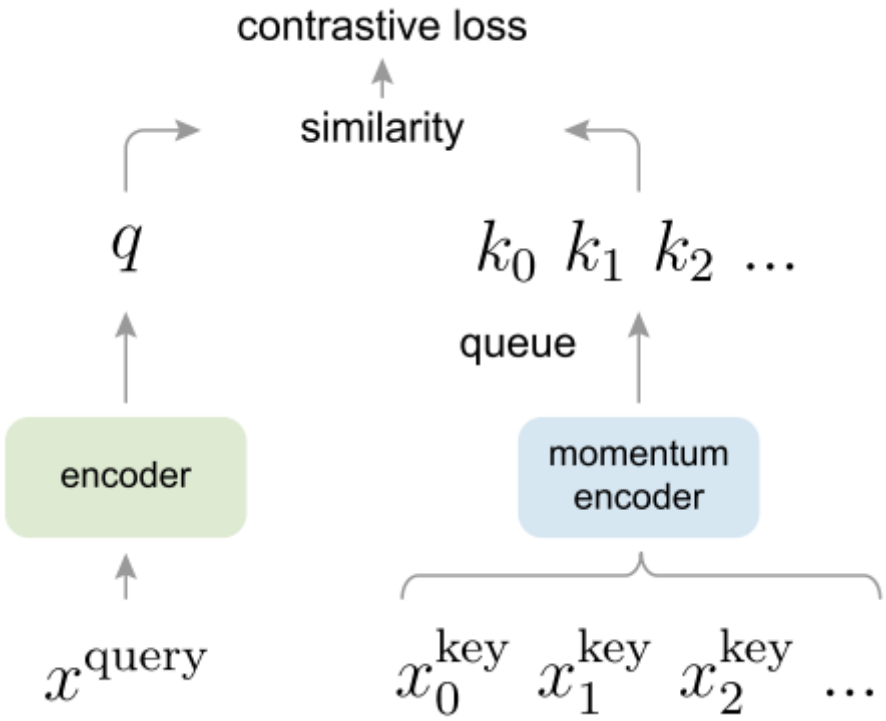
我们的目标是无需监督信息学习一个特征映射：v = f_θ(x), fθ是以θ为参数的卷积神经网络，将图片x映射成特征v。映射同样包含了图像空间的度量d_θ(x,y)=||f_θ(x)−f_θ(y)||对于实例x和y.一个好的映射应该能够将视觉相似的图片投影得相近。我们的无监督特征学习是**实例级别的判别式学习**，我们将**每张图片都当作一个独特的类别对待并训练一个分类器将这些类别分开.**

研究者们提出了非参数的公式:用$\mathbf{v}_j^T\mathbf{v}$取代$\mathbf{w}_j^T\mathbf{v}$,并且通过L2正则化使得$\|\mathbf{v}\| = 1$,然后概率公式$P(i|\mathbf{v})$为$P(i|\mathbf{v}) = \frac{\exp(\mathbf{v}_i^T\mathbf{v}/\tau)}{\sum_{j=1}^n \exp\left(\mathbf{v}_j^T\mathbf{v}/\tau\right)}$,这里比较的是$\mathbf{v}'$和V之间的匹配程序.τ是个温度参数控制分布的集中程度。学习的目标是最大化联合概率密度$\prod_{i=1}^n P_\theta\left(i|f_\theta\left(x_i\right)\right)$,等价于最小化:$J(\theta) = -\sum_{i=1}^n \log P\left(i|f_\theta\left(x_i\right)\right)$.

为了计算Eq.(2)中的概率P(i|v)，需要对所有的图像使用{$\mathbf{v_j}$}。我们保留了一个用于存储[46]的特征{v_j}的memory bank V，而不是每次都对这些特征进行详尽的计算。在接下来的文章中，我们将介绍memory bank的独立设置和特征在网络中的前向传播。让$V = \{\mathbf{v}_j\}$是memory bank和$\mathbf{f}_i = f_{\boldsymbol{\theta}}\left(x_i\right)$是Xi的特征。在每个迭代通过随机梯度下降优化学习特征**fi**以及网络参数θ。然后$\boldsymbol{f}_i$在其相对应的实例下去更新V,我们初始化memory bank V中的特征为随机的单元向量.

6.4.2 基于Momentum Contrast的无监督学习方法

提出动量对比度（MoCo）用于无监督的视觉表示学习。 从作为字典查找的对比学习的角度来看，构建了一个带有队列和移动平均编码器的动态字典。 这样就可以实时构建大型且一致的词典，从而促进对比性的无监督学习。



假设有一个编码的查询q以及一组编码的样本{k_0,k_1,k_2,...}（字典中的键值），假设q和一个字典里单独的键值k_+

匹配，Contrastive loss 即对比损失，它的取值低表示q和键值k_+（positive key）相似而和字典中其它所有键值（negative keys）不相似。

定义一种对比损失函数InfoNCE，形式为：

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)} \tag{5}$$

τ用来控制concentration level of distribution。

对比损失作为无监督的目标函数用来训练编码器网络来表示查询（queries）和键值（keys），

查询表示为$q = f_q\left(x^q\right)$,f_q为一个编码器网络,x_q为查询样本.同样有$q = f_k\left(x^k\right)$,输入的具体形式由特定的任务决定。

方法的核心是将词典保持为数据样本队列。这样可以重新利用当前mini-batch中已编码的键值。同时队列能够将字典大小和mini-batch大小进行解耦，字典大小可以远远大于mini-batch的大小，可被当作超参数。由于mini-batch遵循先进先出的准则，字典总是表示一个所有数据的子集。

使用队列可以扩充字典的大小，但是对键值编码器key encoder进行反向传播变得更难（梯度会在队列中的所有数据进行传播）。而简单地将query encoder f_q直接复制给key encoder f_k，这样快速地改变key encoder会破坏键值表示的一致性。于是作者提出动量更新方法：

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, m \in [0,1)$$

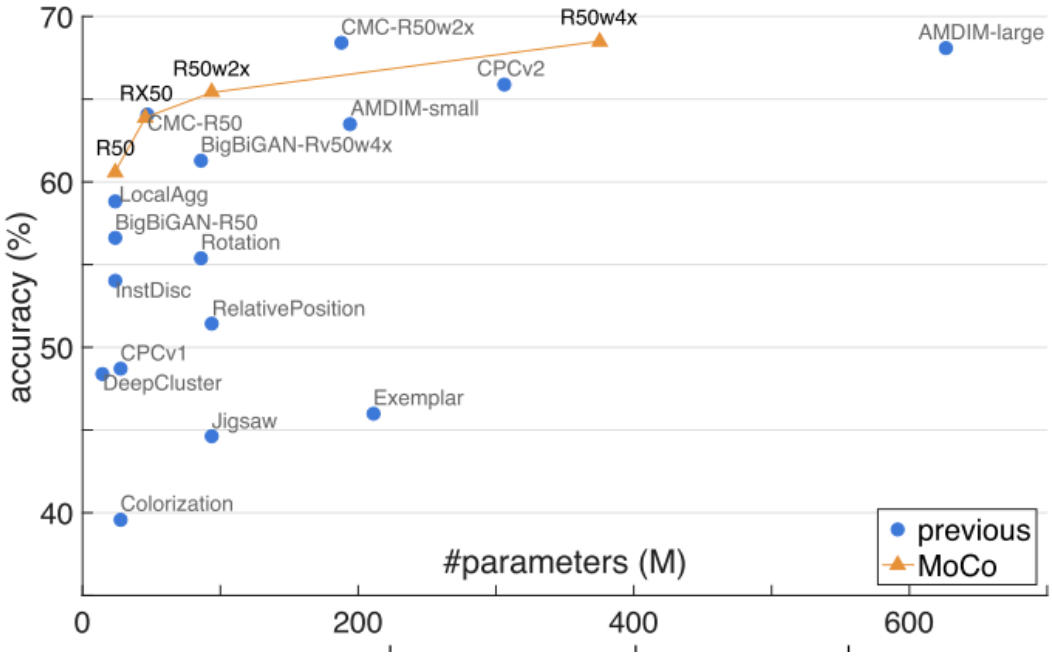
只有θ_q 通过反向传播更新，θ_k 的变换更加平滑，这样一来，尽管队列中的键值被不同的编码器进行编码，但是这些编码器的差别很小，在实验中，大的动量（例如0.999）往往效果好于小的动量（例如0.9），意味着缓慢变化的key encoder是利用好队列的关键所在。

7 性能比较

本节比较图像和视频特征自监督学习方法在公共数据集上的性能。在图像特征自监督学习中，比较了图像分类、语义分割和目标检测等下游任务的性能。对于视频特征自监督学习，本文报道了视频中人类动作识别这一下游任务的性能。

7.1图像特征学习的表现

如4.3节所述，通过对自监督学习模型的下游任务(如语义分割、目标检测和图像分类)进行微调，来评估自监督学习模型学习特征的质量。本节对现有的图像特征自监督学习方法的性能进行了总结。



method	architecture	#params (M)	accuracy (%)
Exemplar [15]	R50w3×	211	46.0 [36]
RelativePosition [11]	R50w2×	94	51.4 [36]
Jigsaw [43]	R50w2×	94	44.6 [36]
Rotation [17]	Rv50w4×	86	55.4 [36]
Colorization [62]	R101*	28	39.6 [12]
DeepCluster [3]	VGG [51]	15	48.4 [4]
BigBiGAN [14]	R50	24	56.6
	Rv50w4×	86	61.3
methods based on contrastive learning follow:			
InstDisc [59]	R50	24	54.0
LocalAgg [64]	R50	24	58.8
CPC v1 [44]	R101*	28	48.7
CPC v2 [33]	$R170^{*}_{wider}$	303	65.9
CMC [54]	$R50_{L+ab}$	47	64.1†
	$R50w2\times_{L+ab}$	188	68.4†
AMDIM [2]	$AMDIM_{small}$	194	63.5†
	$AMDIM_{large}$	626	68.1†
MoCo	R50	24	60.6
	RX50	46	63.9
	R50w2×	94	65.4
	R50w4×	375	**68.6**

表4:在ImageNet上，MoCo和其它方法在线性分类评价下的对比结果。

		ImageNet					Places				
Method	Pretext Tasks	conv1	conv2	conv3	conv4	conv5	conv1	conv2	conv3	conv4	conv5
Places labels [8]	—	—	—	—	—	—	22.1	35.1	40.2	43.3	44.6
ImageNet labels [8]	—	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random(Scratch) [8]	—	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
ColorfulColorization [18]	Generation	12.5	24.5	30.4	31.5	30.3	16.0	25.7	29.6	30.3	29.7
BiGAN [122]	Generation	17.7	24.5	31.0	29.9	28.0	21.4	26.2	27.1	26.1	24.0
SplitBrain [42]	Generation	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34.0	34.1	32.5
ContextEncoder [19]	Context	14.1	20.7	21.0	19.8	15.5	18.2	23.2	23.4	21.9	18.4
ContextPrediction [41]	Context	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Jigsaw [20]	Context	**18.2**	28.8	34.0	33.9	27.1	23.0	32.1	35.5	34.8	31.3
Learning2Count [130]	Context	18.0	30.6	34.3	32.5	25.7	**23.3**	**33.9**	36.3	34.7	29.6
DeepClustering [44]	**Context**	13.4	**32.3**	**41.0**	**39.6**	**38.2**	19.6	33.2	**39.2**	**39.8**	**34.7**

表5:在ImageNet和Places上的线性分类，并使用来自AlexNet的卷积层的激活作为特征。“Convn”表示基于AlexNet的第n个卷积层对线性分类器进行训练。“Places Labels”和“ImageNet Labels”表示使用经过人工标注标签训练的监督模型作为预训练模型。

表5列出了ImageNet[13]和Places[107]数据集上的图像分类性能。在自我监督的借口任务训练中，大多数方法都是在ImageNet数据集上，以AlexNet为基础网络进行训练，不使用类别标签。在完成文本前任务的自监督训练后，在ImageNet和Places数据集的训练分割上，在卷积网络不同的冻结卷积层上训练一个线性分类器。在两个数据集上的分类性能被用来表示学习特征的质量.

如表5所示，自监督模型的整体性能低于使用ImageNet标签或使用place标签训练的模型。在所有的自监督方法中，Deepcluster[44]在两个数据集上的性能最好。从表中我们可以得出三个结论:(1)不同层次的特征总是受益于自我监督的借口任务训练。自监督学习方法的性能总是优于从零开始训练的模型。(2)所有的自监督方法对来自conv3和conv4层的特征表现良好，而对来自conv1、conv2和conv5层的特征表现较差。这可能是因为浅层捕获一般的低层特性，而深层捕获借口任务相关的特性。(3)当用于借口任务训练的数据集与用于下游任务的数据集之间存在域差距时，自监督学习方法能够达到与使用ImageNet标签训练的模型相当的性能.

除了图像分类外，后续任务还包括目标检测和语义分割，以评估自监督学习学习特征的质量。通常，ImageNet通过丢弃类别标签来进行自我监督借口任务的预训练，而AlexNet作为基础网络，对这三个任务进行微调。表6列出了在PASCAL VOC数据集上的图像分类、对象检测和语义分段任务的性能。通过对模型的检验，获得了分类和检测的性能.

Method	Pretext Tasks	Classification	Detection	Segmentation
ImageNet Labels [8]	—	79.9	56.8	48.0
Random(Scratch) [8]	—	57.0	44.5	30.1
ContextEncoder [19]	Generation	56.5	44.5	29.7
BiGAN [122]	Generation	60.1	46.9	35.2
ColorfulColorization [18]	Generation	65.9	46.9	35.6
SplitBrain [42]	Generation	67.1	46.7	36.0
RankVideo [38]	Context	63.1	47.2	35.4†
PredictNoise [46]	Context	65.3	49.4	37.1†
JigsawPuzzle [20]	Context	67.6	53.2	37.6
ContextPrediction [41]	Context	65.3	51.1	—
Learning2Count [130]	Context	67.7	51.4	36.6
DeepClustering [44]	**Context**	**73.7**	**55.4**	**45.1**
WatchingVideo [81]	Free Semantic Label	61.0	52.2	—
CrossDomain [30]	Free Semantic Label	68.0	52.6	—
AmbientSound [154]	Cross Modal	61.3	—	—
TiedToEgoMotion [95]	Cross Modal	—	41.7	—
EgoMotion [94]	Cross Modal	54.2	43.9	—

表6:基于PASCAL VOC数据集的自监督图像特征学习方法在分类、检测和分割方面的比较。“ImageNet标签”表示使用经过人工标注标签训练的监督模型作为预训练模型。

如表5所示，自监督模型在分割和检测数据集上的性能与在训练前使用ImageNet标签进行训练的监督方法非常接近。在目标检测和语义分割任务上的性能差异小于3%，说明自监督学习学习的特征具有良好的泛化能力。在所有的自监督学习方法中，深度聚类的[44]在所有的任务中都获得了最好的性能.

7.2 总结

根据这些结果，可以得出关于自我监督学习方法的性能和复现性的结论。

性能:对于图像特征自监督学习，由于其借口任务设计良好，自监督方法在一些下游任务上的性能可与监督方法相媲美，特别是在目标检测和语义分割任务上。在目标检测和语义分割任务上的性能差异小于3%，说明自监督学习学习的特征具有良好的泛化能力,然而,视频特征自监督学习模型在下游任务上的性能仍远低于监督模型。基于3DConvNet的方法在UCF101数据集上的最佳性能降低了18%以上,而非监督模型[70],3DCovnNet自监督学习方法的性能较差，可能是由于3DConvNets通常有较多的参数，由于视频的时间维度，容易导致过拟合和视频特征学习的复杂性.

可复现性:我们可以观察到，对于图像特征自监督学习方法，大多数网络使用AlexNet作为基础网络，在ImageNet数据集上进行预训练，然后对相同的下游任务进行评估，以进行质量评估。而且，大多数方法的代码都会被发布，这对重现结果有很大的帮助。然而，在视频自监督学习中，各种数据集和网络被用于自监督前训练，因此直接比较不同的方法是不公平的。此外，一些方法使用UCF101作为自监督的训练前数据集，这是一个相对较小的视频数据集。由于数据集的大小，更强大的模型(如3DCovnNet)的能力可能无法被完全发现，并可能受到服务器过度拟合的影响。因此，应该使用较大的数据集进行视频特征的自监督预训练。

评估指标:另一个事实是，需要更多的评估指标来评估不同层次的学习特性的质量。当前的解决方案是使用下游任务的性能来指示特性的质量。然而，这一评估指标并没有给出网络通过自我监督的预训练所学习到的东西。应该使用更多的评价指标，如网络剖分[78]来分析自监督学习特征的可解释性.

8 未来的发展方向

在一些计算机视觉任务中，自监督学习方法已经取得了巨大的成功，并取得了接近于超级可视模型的良好性能。在这里，讨论了一些自监督学习未来的方向。

从合成数据中学习特征:自我监督学习的一个上升趋势是使用同步数据来训练网络，这些数据可以很容易地由游戏引擎来呈现，而人类的参与非常有限。在游戏引擎的帮助下，可以轻松生成数百万具有精确像素级注释的合成图像和视频。通过精确而详细的注释，可以设计各种借口任务来从合成数据中学习特性。需要解决的一个问题是如何弥合合成数据和真实数据之间的领域差距。只有少数研究探索了利用GAN桥接域隙[30]从合成数据中进行自监督学习[166]。随着大规模数据的增多，将会有更多的自监督学习方法被提出。

从web数据中学习:另一个上升趋势是使用基于现有关联标记的web收集数据[22][167]和[168]来训练网络。有了这个搜索引擎，人们可以从Flickr和YouTube等网站上下载数百万张图片和视频，而成本可以忽略不计。除了原始数据之外，标题、关键字和评论也可以作为数据的一部分，作为培训网络的额外信息。通过精心策划的查询，可靠的搜索引擎检索到的web数据可以相对干净。使用大规模的web数据及其相关的元数据，可以提高自我监督方法的性能。从web数据学习的一个开放问题是如何处理web数据及其相关元数据中的噪声.

从视频中学习时空特征:自监督图像特征学习已经得到了很好的研究，在语义分割和目标检测等下游任务中，自监督模型和自监督模型之间的性能差距很小。然而，使用3DConvNet进行自监督视频时空特征学习的问题尚未得到很好的解决。需要更有效的借口任务，专门从视频中学习时空特征.

利用来自不同传感器的数据学习:大多数存在的自我监督的视觉特征学习方法只专注于图像或视频。然而，如果来自不同传感器的其他类型的数据是可用的，那么不同类型数据之间的约束可以用作训练网络学习特性的额外来源[155]。自动驾驶汽车通常配备各种传感器，包括RGB摄像头、灰度摄像头、3D激光扫描仪、高精度GPS测量和IMU加速度。很容易获得非常大规模的数据集,并且不同设备捕获的数据对应可以作为自监督特征学习的监督信号.

多借口任务学习:现有的自监督视觉特征学习方法大多是通过训练卷积网络来学习特征，解决一个借口任务。不同的借口任务提供不同的监督信号，帮助网络学习更多有代表性的特征。只有少数研究探索了自监督特征学习[30]、[32]的多借口任务学习。通过对多借口任务自监督特征学习的研究，可以做更多的工作。

9 结论

基于深度卷积神经网络的自监督图像特征学习取得了很大的成功，在一些下游任务中，自监督方法与监督方法的性能之间的差距变得很小。本文从常用的网络结构、虚拟任务、算法、数据集、性能比较、讨论和未来发展方向等方面，对近年来基于深度卷积神经网络的自监督图像和视频特征学习方法进行了广泛的综述。以表格形式对方法、数据集和性能的比较总结清楚地说明了它们的特性，这将有利于计算机视觉领域的研究人员的研究.