Deep Clustering for Unsupervised Learning of Visual Features

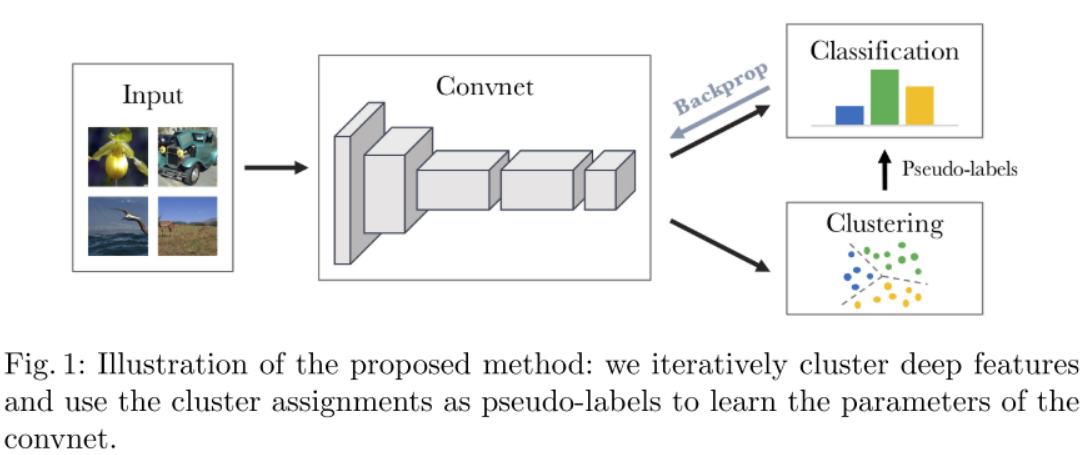
1. 摘要

聚类是一类无监督的学习方法，已经在计算机视觉中得到了广泛的应用和研究。在使其适应大规模数据集上的视觉特征的端到端训练方面所做的工作很少。在这项工作中，我们提出了DeepCluster，一种聚类方法，联合学习神经网络参数和所产生特征的聚类分配。DeepCluster用标准聚类算法k-means对特征进行迭代分组，并使用后续的分配作为监督来更新网络的权重。我们将DeepCluster应用于大型数据集如ImageNet和YFCC100M上的卷积神经网络的无监督训练。

1. 问题重述

预先训练的卷积神经网络，或称卷积网，已经成为大多数计算机视觉应用的构建模块[1,2,3,4]。它们产生了优秀的通用特征，可以用来提高在有限的数据量上学习的模式的通用性[5]。ImageNet[6]的存在，一个大型的完全监督的数据集，一直在推动着ConvNet预训练的进展。然而，Stock和Cisse[7]最近提出了经验性的证据，表明最先进的分类器在ImageNet上的性能在很大程度上被低估了，而且很少有错误没有被解决。这在一定程度上解释了为什么尽管近年来提出了许多新的架构[2,8,9]，但其性能却一直处于饱和状态。事实上，以今天的标准来看，ImageNet相对较小；它 "只 "包含了一百万张涵盖物体分类这一特定领域的图像。前进的一个自然途径是建立一个更大、更多样化的数据集，可能包括数十亿张图像。这又需要大量的人工注释，尽管社区多年来积累了众包方面的专家知识[10]。用原始元数据代替标签会导致视觉表达的偏差，产生不可预知的后果[11]。这就需要有能够在没有监督的情况下对互联网规模的数据集进行训练的方法。例如，"特征袋 "模型在手工制作的局部脚本上使用聚类来产生良好的图像级特征[16]。它们成功的一个关键原因是，它们可以应用于任何特定的领域或数据集，如卫星或医学图像，或用新的模式拍摄的图像，如深度，其中注释并不总是可用的数量。一些工作表明，有可能将基于密度估计或降维的无监督方法适应于深度模型[17,18]，从而产生有前途的多用途视觉特征[19,20]。尽管聚类方法在图像分类中取得了初步的成功，但很少有工作[21,22]被提出来，以使它们适应ConvNet的端到端训练，而且从来没有规模。一个问题是，聚类方法主要是为固定特征之上的线性模型而设计的，如果要同时学习特征，它们几乎不起作用。例如，用k-means学习convnet会导致一个简单的解决方案，即特征被归零，而聚类被溃缩成一个单一类别。

在这项工作中，我们提出了一种新的聚类方法，用于大规模的端到端信念网的训练。我们表明，通过聚类框架有可能获得有用的通用视觉特征。



我们的方法总结在图1中，包括在图像描述符的聚类和通过预测聚类任务更新convnet的权重之间交替进行。为了简单起见，我们把研究重点放在k-means上，但也可以使用其他的聚类方法，比如Power Iteration Clustering（PIC）[23]。整个管道足够接近于标准的convnet的监督训练，可以重复使用许多常见的技巧[24]。与自监督方法[25,26,27]不同，聚类的优点是不需要什么领域知识，也不需要输入的特定信号[28,29]。尽管我们的方法很简单，但在ImageNet分类和转移任务上，我们的方法取得的性能明显高于以前发表的无监督方法。

最后，我们通过修改实验协议，特别是训练集和convnet结构来探测我们框架的稳健性。实验结果扩展了由Doersch等人[25]发起的关于这些选择对无监督方法性能的影响的讨论。我们证明了我们的方法对架构的改变是稳健的。用VGG[30]代替AlexNet，可以显著提高特征的质量和随后的转移性能。更重要的是，我们讨论了使用ImageNet作为无监督模型的训练集。虽然它有助于理解标签对网络性能的影响，但ImageNet有一个特殊的图像分布，这是从它用于细粒度的图像分类挑战中继承下来的：它是由均衡的类组成的，例如包含各种各样的狗的品种。我们考虑用Thomee等人[31]的YFCC100M数据集中的随机Flickr图像作为替代。我们表明，我们的方法在对这种未固化的数据分布进行训练时保持了最先进的性能。最后，目前的基准集中在无监督网络捕捉类信息的能力上。我们建议在图像检索的基准上对其进行评估，以衡量其捕获实例级信息的能力。

在本文中，我们做出了以下贡献：(i)一种新颖的无监督的Convnets端到端学习方法，该方法可与任何标准聚类算法（如K-means）一起使用，并且需要最少的额外步骤；(ii)在无监督学习中使用的许多标准转移任务上具有最先进的性能；(iii)在未固化的图像分布上训练时，性能高于以前的技术水平；(iv)对目前无监督特征学习的评价协议进行了讨论。

1. 相关工作

**无监督特征学习**。与我们的工作相关的几种方法在没有监督的情况下学习深度模型。Coates和Ng[32]也使用k-means来预训练convnets，但他们是以自下而上的方式依次学习每一层，而我们则是以端到端的方式进行。其他的聚类损失[21,22,33,34]也被考虑用来联合学习convnet特征和图像聚类，但它们从未被测试过，无法对现代convnet架构进行彻底研究。特别值得注意的是，Yang等人[21]用一个递归框架反复学习convnet特征和集群。他们的模型在小数据集上提供了很好的性能，但要扩展到convnets所需的图像数量上，可能是一个挑战。与我们的工作更接近的是，Bojanowski和Joulin[19]在一个大型数据集上学习视觉特征，其损失试图保留流经网络的信息[35]。他们的方法以类似于examplar SVM[36]的方式对图像进行判别，而我们只是对其进行聚类。

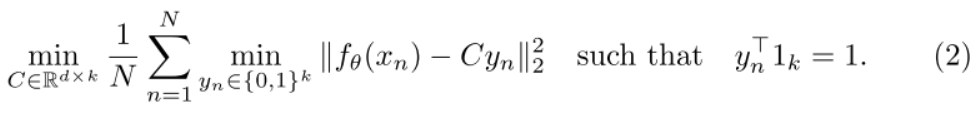
**自监督学习**。一种流行的无监督学习形式，称为 "自我监督学习"[37]，使用借口任务，用直接从原始输入数据中计算出来的 "伪标签 "取代人类标注的标签。例如，Doersch等人[25]使用预测图像中斑块的相对位置作为借口任务，而Noroozi和Favaro[26]训练一个网络来重新安排打乱的图像块。另一个使用空间线索的例子是Pathak等人[38]的工作，即根据其周围的情况猜测缺失的像素。Paulin等人[39]利用图像检索设置学习补丁级卷积核网络[40]。其他一些人通过预测连续帧之间的相机转换[41]，利用跟踪斑块的时间一致性[29]或基于运动的视频分割[27]来利用视频中可用的时间信号。除了空间和时间上的一致性，许多其他信号也得到了探索：图像着色[28,42]、跨频道预测[43]、声音[44]或实例计数[45]。最近，人们提出了几种结合多种线索的策略[46,47]。与我们的工作相反，这些方法是依赖于领域的，需要专家知识来精心设计一个可能导致可转移特征的借口任务。

**生成模型**。最近，无监督学习在图像生成方面取得了很大的进展。通常，在预定义的随机噪声和图像之间学习一个参数化的映射，使用自动编码器[18,48,49,50,51]、生成对抗网络（GAN）[17]或更直接地使用重建损失[52]。特别值得注意的是，GAN的判别器可以产生视觉特征，但其性能相对令人失望[20]。Donahue等人[20]和Dumoulin等人[53]已经表明，在GAN中加入一个编码器可以产生更具竞争力的视觉特征。

1. 详细阐述
   1. 通过聚类进行无监督学习

当θ从高斯分布中取样时，在没有任何学习的情况下，fθ不会产生好的特征。然而，这种随机特征在标准转移任务上的表现，远远高于偶然水平。例如，在随机AlexNet的最后一个卷积层之上的多层感知器在ImageNet上取得了12%的准确率，而随机网络只有0.1%[26]。随机卷积网的良好性能与它们的卷积结构密切相关，它给了输入信号一个强大的先验。这项工作的想法是利用这种弱信号来引导卷积网的判别能力。我们对卷积网的输出进行聚类，并将后续的聚类分配作为 "伪标签 "来优化公式（1）。这种深度聚类（DeepCluster）方法迭代地学习特征并将它们分组。

聚类已被广泛研究。在缺乏比较点的情况下，我们把重点放在一个标准的聚类算法上，即k-means。k-means将一组向量作为输入，在我们的例子中是由Convnet产生的特征fθ(xn)，并根据一个几何标准将它们聚成k个不同的组。更确切地说，它通过解决以下问题来联合学习一个d×k的中心点矩阵C和每个图像n的聚类任务yn。



解决这个问题提供了一组最佳分配(y∗ n)n≤N和一个中心矩阵C∗。然后，这些分配被用作伪标签；我们不使用中心点矩阵。

总的来说，DeepCluster在对特征进行聚类以产生 使用公式(2)产生伪标签，并通过使用公式(1)预测这些伪标签来更新convnet的参数。这种类型的交替过程容易产生琐碎的解决方案；我们在下一节描述如何避免这种退化的解决方案。

* 1. Avoiding trivial solutions

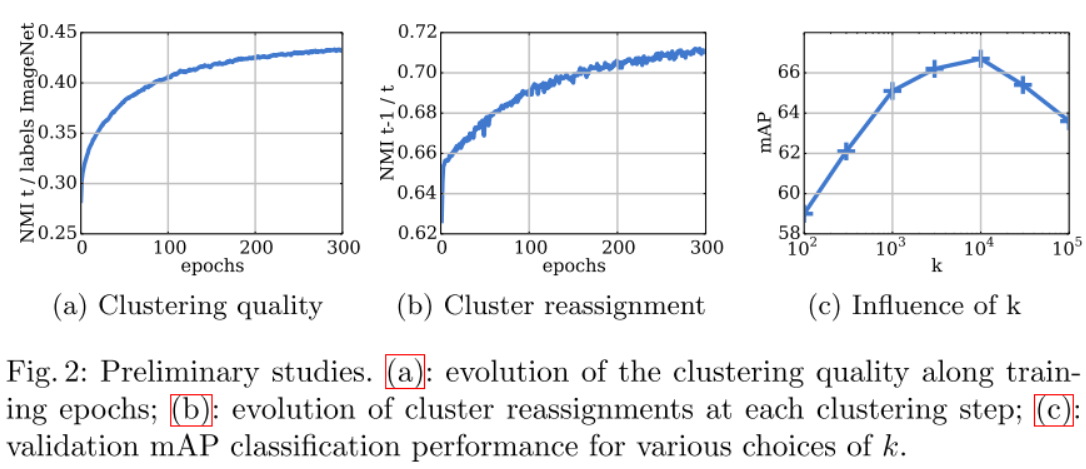
琐碎解决方案的存在并不是神经网络的无监督训练所特有的，而是存在于任何联合学习判别性分类器和标签的方法中。判别性分类器甚至在应用于线性模型时也会受到这个问题的影响[57]。解决方案通常是基于约束或惩罚每个聚类的最小点数[58,59]。这些条款是在整个数据集上计算的，这不适用于大规模数据集上的convnets的训练。在本节中，我们简要介绍了这些琐碎解的产生原因，并给出了简单和可扩展的解决方法。

**空的群组**。鉴别性模型学习类之间的决策边界。一个最佳的决策边界是将所有的输入都分配到一个单一的集群中[57]。这个问题是由缺乏防止空集群的机制引起的，在线性模型中和在Convnets中都会出现。在特征量化[60]中使用的一个通用技巧是在k-means优化过程中自动重新分配空集群。更确切地说，当一个簇成为空的时候，我们随机选择一个非空的簇，用它的中心点作为空簇的新中心点，并进行小的随机扰动。然后，我们将属于非空簇的点重新分配到所产生的两个簇中。

**琐碎的参数化**。如果绝大部分图像被分配到几个簇中，参数θ将专门区分它们。在最夸张的情况下，除了一个聚类外，所有的聚类都是单子，最小化公式（1）会导致一个琐碎的参数化，无论输入是什么，convnet都会预测相同的输出。当每个类别的图像数量非常不平衡时，这个问题也会出现在监督分类中。例如，元数据，如标签，表现为Zipf分布，少数标签在整个分布中占主导地位[61]。规避这个问题的策略是根据类的均匀分布或伪标签对图像进行采样。这相当于将输入对公式（1）中损失函数的贡献按其分配的群组大小的倒数加权。

1. 实验

聚类和标签之间的关系。图2(a)显示了训练期间群组分配和ImageNet标签之间NMI的演变。它衡量的是模型预测类级信息的能力。请注意，我们只在这个分析中使用这个指标，而不是在任何模型选择过程中使用。聚类和标签之间的相关性随着时间的推移而增加，表明我们的特征逐渐捕捉到了与物体类别相关的信息。



1. 评析