专业阅读与写作研究报告

朱启鹏
58119304
东南大学人工智能学院
中国，南京

方玉杰
58119105
东南大学人工智能学院
中国，南京

曾晗
58119303
东南大学人工智能学院
中国，南京

牟俊奇
58119207
东南大学人工智能学院
中国，南京

摘要—近些年来，随着大数据的发展，数据规模不断扩大，但人们也渐渐发现给数据打标签，将是一件耗费大量人力物力的事情。而对于使用传统神经网络进行分类任务，往往需要的数据是带有标签的，这就给研究工作者带来了一个不小的挑战。但自监督学习却巧妙的绕过了这个难题，自监督学习通过对无标签数据进行前置地打标签，采用聚类的方法进行分类，从而大大降低了训练的成本。但是，“天下没有免费的午餐”，自监督的模型并不总是带来十分令人满意的结果，所以自监督学习的研究在近几年仍具有很大的热度。

I. 概述

A. 学习范式

我们首先来回顾下机器学习中两种基本的学习范式，如图所示，一种是监督学习，一种是无监督学习。

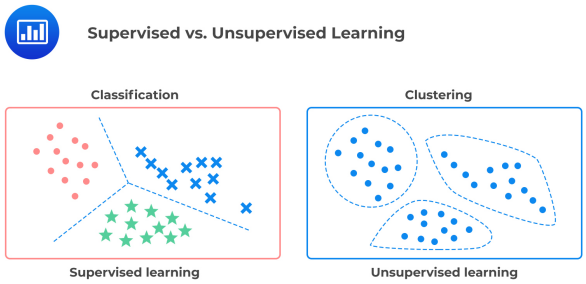

图 1: 监督学习 vs 无监督学习

监督学习利用大量的标注数据来训练模型，在计算模型的预测值与数据的真实值产生的损失后，进行反向传播（计算梯度、更新参数），经过不断的学习，最终可以获得识别新样本的能力。无监督学习则是不依赖任何标签值，而是通过对数据内在特征的挖掘，找到样本间的关系，比如聚类等相关任务。有监督和无监督最主要的区别在于模型在训练时是否需要人工标注的标签信息。

无监督学习中被广泛采用的方式是自动编码器（autoencoder）：

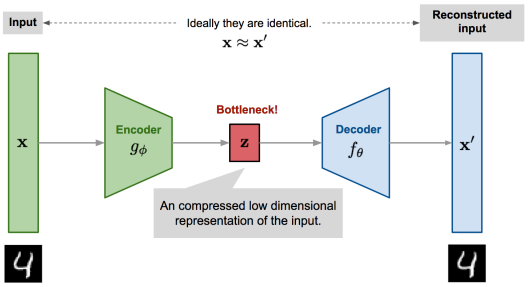

图 2: 自动编码器

编码器将输入的样本映射到隐层向量，解码器将这个隐层向量映射回样本空间。我们期待网络的输入和输出可以保持一致（理想情况，无损重构），同时隐层向量的维度远远小于输入样本的维度，以此达到了降维的目的，利用学习到的特征向量再进行聚类等任务，最终实验将更加的简单高效。对于如何学习特征向量的研究，可以称之为表征/表示学习（Representation Learning）。但这种简单的编码-解码结构仍然存在很多问题，基于像素的重构损失通常假设每个像素之间都是独立的，从而降低了它们对相关性或复杂结构进行建模的能力。尤其使用 L1 或 L2 损失来衡量输入和输出之间的差距其实是不存在语义信息的，而过分的关注像素级别的细节而忽略了更为重要的语义特征。对于自编码器，可能仅仅是做了维度的降低而已，我们希望学习的目的不仅仅是维度更低，还可以包含更多的语义特征，让模型懂的输入究竟是什么，从而帮助下游任务。而自监督学习最主要的目的就是学习到更丰富的语义表征，也是我们小组为何要研究自监督学习的根本所在。

B. 自监督学习简介

自监督学习主要是利用辅助任务从大规模的无监督数据中挖掘自身的监督信息，通过这种构造的监督信息对网络进行训练，从而可以学习到对下游任务有价值的表征。也就是说自监督学习的监督信息不是人工标注的，而是算法在大规模无监督数据中自动构造监督信息，从而进行监督学习或训练。因此，大多数时候，我们称之为无监督预训练方法或无监督学习方法，但严格上讲，它应该叫自监督学习。

C. 自监督的 *Pretrain - Finetune* 流程

首先，在大量的无标签数据中，利用 pretext 来训练网络（自动在数据中构造监督信息），得到预训练的模型，然后对于新的下游任务，和监督学习一样，将迁移学习到的参数进行微调即可。所以自监督学习的能力主要由下游任务的性能来体现。

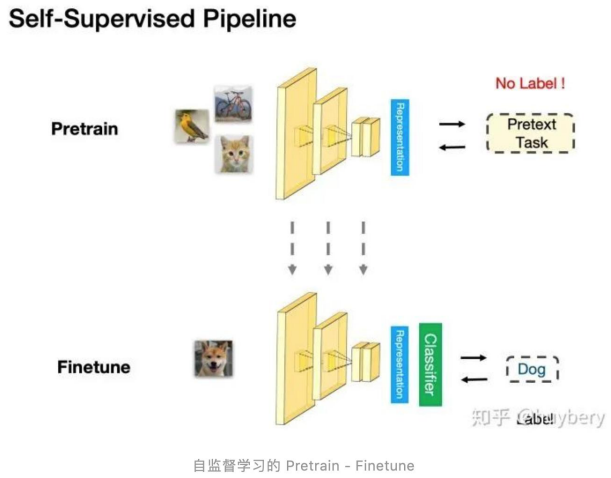

图 3: 自监督的 Pretrain - Finetune 流程

II. 研究过程

A. 论文来源

此次研究，由于时间有限，我们一共收集了六篇有关自监督学习有关的论文。具体如下：

Deep Clustering for Unsupervised Learning of Visual Features [1] 来自 ECCV (14) 2018

Unsupervised Pre-Training of Image Features on Non-Curated Data [2] 来自 ICCV 2019

Unsupervised Representation Learning by Predicting Image Rotations [3] 来自 ICLR (Poster) 2018

Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles [4] 来自 ECCV (6) 2016

Boosting Self-Supervised Learning via Knowledge Transfer [5] 来自 CVPR 2018

Local Aggregation for Unsupervised Learning of Visual Embeddings [6] 来自 ICCV 2019

查询论文时，我们首先查找到的时来自 ICCV 2019 的 Local Aggregation for Unsupervised Learning of Visual Embeddings [6] 以及来自 ICCV 2019 的 Unsupervised Pre-Training of Image Features on Non-Curated Data [2]，再通过扩展查找的方式查找到了余下四篇论文。

B. 论文来源会议简介

计算机视觉领域世界三大顶级会议分别为 CVPR、ICCV 和 ECCV。

1) ICCV: ICCV，英文全称 International Conference on Computer Vision，中文全称国际计算机视觉大会，这个会议也是由 IEEE 主办的全球最高级别学术会议，每两年在世界范围内召开一次，在业内具有极高的评价。ICCV 论文录用率非常低，是三大会议中公认级别最高的。与 CVPR 不同的是，CVPR 会议每年都在美国地区举办，而 ICCV 会议自 1987 年起至今每两年都会在全世界不同的国家举办会议，2005 年 ICCV 是在中国北京举办的会议。

2) CVPR: CVPR，英文全称 Conference on Computer Vision and Pattern Recognition，中文全称是国际计算机视觉与模式识别会议。这个会议是由 IEEE 主办的一年一度的全球学术性顶级会议，会议的主要内容是计算机视觉与模式识别技术，每年 CVPR 都会有一个固定的研讨主题。会议一般在每年六月举行，大部分情况下会议都在美国西部地区举办，也会在美国中部和东部地区之间循环举办。

3) ECCV: ECCV，英文全称 European Conference on Computer Vision，中文全称欧洲计算机视觉国际会议。ECCV 每年的论文接受率为 25-30% 左右，每次会议在全球范围会收录论文 300 篇左右，收录论文的主要来源是来自于美国、欧洲等顶级实验室及研究所，中国大陆的收录论文数量在 10-20 篇之间。

4) ICLR: 国际学习表示会议（ICLR）是每年春季举行的机器学习会议。会议包括受邀演讲以及参考论文

的口头和海报展示。自 2013 年成立以来，ICLR 一直采用开放的同行评审流程来裁判论文提交（基于 Yann LeCun 提出的模型）。2019 年共提交论文 1591 篇，其中 500 篇通过海报展示（31%）和 24 篇通过口头报告（1.5%）。2021 年论文投稿 2997 篇，其中 860 篇被接受（29%）。ICLR 与 ICML 和 NeurIPS 一起是三大机器学习和人工智能会议之一，在三者中影响最大。

C. 论文关系

我们研究的六篇论文的关系如图所示。

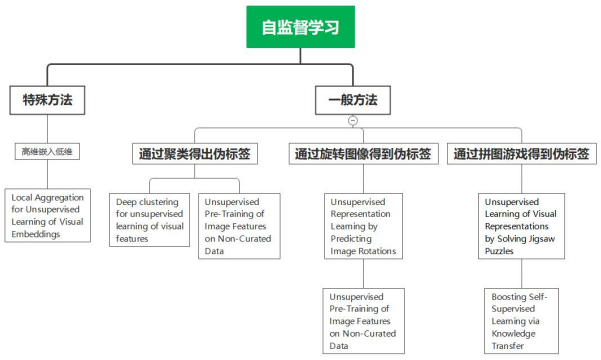

图 4: 论文关系

III. 研究内容

A. 论文内容概述

1) Deep Clustering for Unsupervised Learning of Visual Features: 在无监督学习中，聚类是一种十分常用的算法，而且在计算机视觉领域应用十分广泛，但由于它是无标签的学习过程，很难保证十分优秀的正确率。所以，为了能极大利用样本中的数据，许多人提出了一种让模型自身从样本中学习出标签的方法——自监督（self-supervised）。而在这篇论文中，作者采用了一种新式的聚类方法，并且可以在图上呈现聚类的方式，这种新颖的思想和常规的自监督方法也不尽相同。

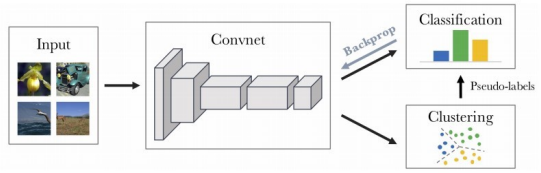

图 5: DeepCluster

此过程首先是选用数据集中的一些图片，在一个传统的预训练的 CNN 模型（比如 AlexNet，或者 VGG 模型）上进行特征提取。但由于选取的图片不尽相同，即使一个训练好的模型也不一定能提取出十分好的特征。因此，文章中作者提出了使用伪标签的做法，即对这些模型提取出来的特征向量进行聚类算法。聚类算法有很多种，作者在这里采用了比较常规的一种做法 KMeans。但他利用了 KMeans 的性质，产生的聚类之后再人为设置伪标签，并且把其作为分类的结果训练 CNN 中的参数，从而使得新训练的 CNN 拥有很好的产生特征向量的能力，进而完成更好的聚类。

同时，作者也在文章中提出了两种新颖的方法来解决他们实验问题。第一个问题就是空聚类。我们知道在 KMeans 中如果 k 设置的较大，可能会出现有一类全空的状态，作者此时就采用随机选取另一个聚类的轴并且对其做轻微扰动的做法，使这里旁边的点聚成两个类别，而不是依照传统把他们都聚成一个类。第二个问题就是关于参数训练的问题。我们知道如果许多的图片聚成很少的类，参数会专门提供给他们具体的类。在最戏剧性的场景中，一个群集以外的所有群集都是单例的，从而最大程度地减少了导致平凡的参数化，卷积网络将不管输入如何，都会输出相同的预测。当每个类别的图像数量高度不平衡时，在监督分类中也会出现此问题。而作者提出的解决方法——基于伪标签的抽样，能很好的规避这个问题。

2) Unsupervised Pre-Training of Image Features on Non-Curated Data: 这篇文章正式的引入了自监督的做法，不同于上一篇论文，这里采用的模型，是经过一个 pre-task 的，也就是在预训练好的模型的基础上再次对参数进行修改，以更好的完成后期的 DeepCluster 的工作。这篇论文的中心思想可以用一张图概括：

这里展示了论文的中心思想，即交替使用 DeepCluster 中的聚类方法，首先通过两次聚类得到 subclass，然后通过对图片旋转的方式来对 CNN 进行训练。这里的旋转可以认为是一个 pre-task，就是可以先通过旋转然后放入 CNN 中让他能得到相应的标签，之后再和 cluster 聚类出来的 subclass 中比较来对此训练 CNN 中的参数，最终达到更加精准的目的。

3) Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles: 本文采用拼图游戏作为自监督的 pretext task，来获取到图像中的某一部分，

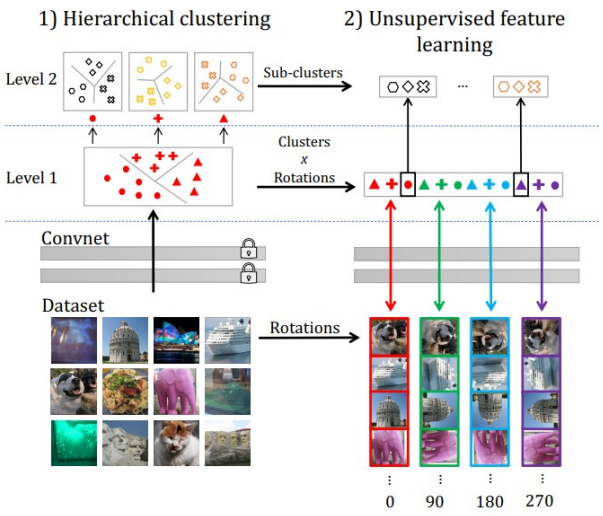

图 6

简单的说，就是将一个图像中的某一部分分成 9 份，按照一定的策略打乱，然后依次输入到接受九个输入的 context-free network 中，预测不同的排序的类别，通过一定的数据处理，以学到较为良好的语义特征的目的。

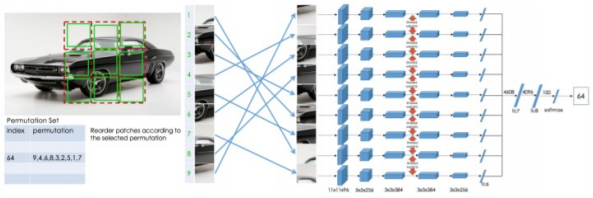

图 7: 网络结构

作者在这里将拼图问题转换为了一种分类问题，按照九个图块的排列数序的不同，可以产生 9! 种排列，也就是 9! 种类别。但是这些类别中大多数是很相似的，有可能两两之间只有两块的位置有所差别。因此，为了保证不同的排列之间具有着足够的差别，确保网络学习到的不是它们的相对位置，在确定排列的时候，计算不同的排列之间的 Hamming 距离，选择具有足够大 Hamming 距离的排列。

除此之外，为了防止网络学习一些底层的无关特征，如图块的纹理特征，边缘信息等等（这些信息会影响到自监督学习的能力），作者采取三种策略来保证网络学习到良好的语义特征：

1. 不同的图块是取自同一张图片的，具有相同的均值和方差。作者将不同的图块单独计算均值和方差进行归一化，从而尽量让其差距变大，增加网络区分的难度

2. 不同的图块间很有可能存在边界上的连续性，取样的时候在不同的图块之间增加了一个 21 像素的间隔，来消除这种连续性。

3. 不同的图块在颜色上是相似的，因此对于颜色做了不同的变换（随机裁剪图像；随机将彩色图像转化为灰度图；随机变化图像的颜色）。总的来说，用拼图游戏来训练自监督模型，从实验上来看，这种方法可以在分类检测和分割任务上取得不错的效果。

4) Boosting Self-Supervised Learning via Knowledge Transfer: 在 Self-Supervised Learning 的设置中，如果我们想要在下游任务中使用一个小型的模型，那么我们将会被迫在 pretext task 中使用相同的模型。这可能会限制模型的表示能力，以及 pretext task 的复杂程度。为了解决这样的一个问题，作者提出了一种从 self supervision 中使用更大的模型来迁移知识到下游任务的方法。

首先对于所有没有标签的图像，作者用一个已经预训练好的模型来计算出他们的特征；其次通过使用 K-means 方法来聚类图像，使用每一个图像所属于的类名来作为伪标签，基于在良好的视觉表示中，语义相似的数据点理应彼此相互靠近，因此在特征空间中，使用简单的聚类方法来获得每个图片的群集分配作为伪标签。最后，用伪标签预训练一个小一些的网络（如 AlexNet），用它来作为下游任务进行 fine tuning 网络的开始。

这个方法可以分为四个步骤：1. 用 Jigsaw++ 自监督的方法来预训练

2. 在参数空间中进行聚类

3. 为每个图片提取伪标签

4. 群集分类：为了将我们在自监督中学习到的 knowledge 迁移，我们用伪标签训练了一个小一些的分类网络

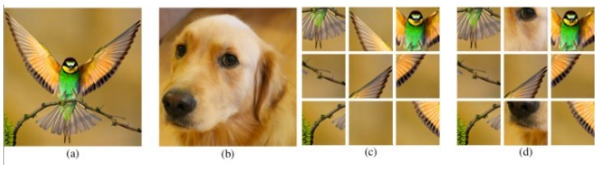

图 8: Jigsaw++ 任务

在原始的前置任务中，目标是找到从图像裁剪的 3*3 的网格中找到图块的重新排序。

在 Jigsaw ++ 的前置任务中，作者从另一个随机图像选取图块中替换网格中的随机数量的图块（图块的位置也是随机选取）。这些遮蔽的图片使得任务变得更加复杂。首先，模型需要检测到这些遮蔽的图块，其次，它需要仅用剩余的图块来解决拼图问题。为了确保没有生成歧义，类似的排列都被删除，使得任意两个排列之间的 Hamming 距离不小于 3。这样对于任意数量的遮蔽图块，拼图任务都会有一个唯一的解决方案。

5) Unsupervised Representation Learning by Predicting Image Rotations: 近年来卷积神经网络（Convolutional Neural Networks, CNN）因其在高维图像上提取语义信息特征的出色表现而被大量研究使用，但作为一种有监督的学习，其不可避免地要使用大量有标签的数据，而在实际应用中是难以实现的。RotNet 采用经典的自监督学习方法，通过以特定角度旋转图片来产生旋转后图片，旋转角度的数据和伪标签，再通过这些“有标签”的数据为前置任务来训练 CNN，来达到提取图片特征向量的目的。而这些特征向量可以被用来进行下游任务。具体实现如下图所示：

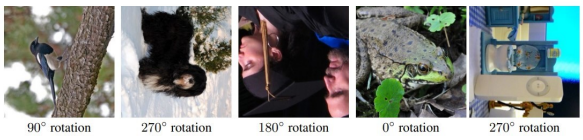

图 9: 旋转后的图像

该算法首先将数据集图片随机旋转 0°，90°，180°，270°，获得四倍的数据（如图 9）；再将图片输入 CNN，以四个旋转角度为“伪标签”训练 CNN 预测图片旋转角度（如图 10），从而让 CNN 真正习得图片的语义信息。

在上述 CNN 的训练过程中，神经网络会学习丢弃图片中与主体旋转无关的背景部分，不予考虑；专注于旋转后差异较大的主题部分，从而识得图片的语义特征。例如：在识别以猫为主体的图片时，CNN 会将更高的权重赋予猫的头部，眼睛，鼻子等“高级部分”，而不去太多考虑与主题无关联的草地等“低级部分”。

6) Local Aggregation for Unsupervised Learning of Visual Embeddings: 本文提出了另外一种不同的自监督方式：通过直接对高维图片在低维嵌入空间（embedding space）中对应的特征点进行移动来训练 CNN，使相似点间距离较近，而不相似点间距离较远，直接实现在嵌入空间中的聚类。这种想法相当于将训练 CNN 与下游聚类任务合二为一，正与我们的目的相符。

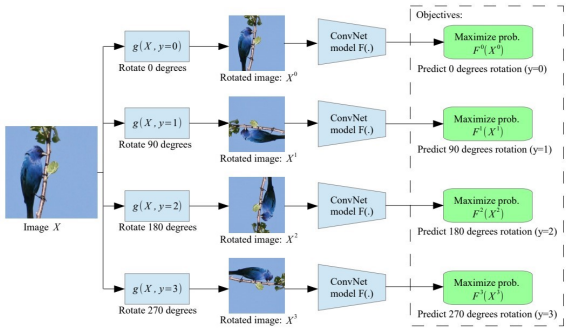

图 10: 自监督学习示意图

本文具体的算法核心为：对于某一张图片，通过 CNN 将其映射到嵌入空间中的上的某一点 v。所有图片对应的点集为 V。定义该点的 background neighbours（集合 B）和 close neighbours（集合 C）：B 集合即为距离 v 最近的 k 个点；集合 C 为在对嵌入空间作多次 k-means 聚类后包含 v 在内的多个聚类之并集。C 集合是在优化时应当与点 v 接近的点集，B 集合是用来衡量 C 集合与 v 的接近程度的参数。并定义与之相关的局部聚集（local aggregation）程度函数 $L(C_i, B_i|\theta, x_i)$。该算法优化的目标即是最大化局部聚集程度，每轮迭代都会将某点 v 移动到使局部聚集程度更高的位置。

此外，作者还使用了记忆库（memory bank），即将 v 集合均初始化为单位向量，在每轮迭代中通过超参数混合系数 t 混合当前点与真实 v 来替代真实特征点进行计算。这样替换后就不必再每轮定义 B 和 C 集前重新计算 V，是一种以空间换时间的操作。

通过图 11 我们可以形象地理解该算法

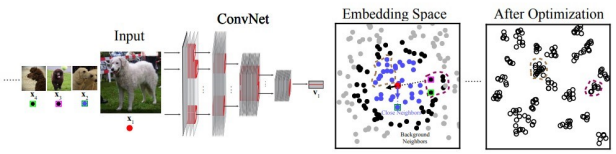

图 11: Local Aggregation 说明

图中 Embedding Space 中的每个点代表一幅图片。对于红色的点，蓝色的点为其 close neighbours，黑色的点为其 background neighbours。训练使得红点与蓝

点更近，且与黑点更远。经过多轮迭代后的情况如图 After Optimization，其中特征点已被聚类。

B. 重点论文介绍

在本部分，我们重点介绍其中 Unsupervised Pre-Training of Image Features on Non-Curated Data 此篇论文。

1) 提出问题:

预训练好的特征提取器，在迁移学习方面发挥着重要作用，例如：

- 当训练数据比较少时，先用预训练好的 feature 提取器初始化网络，然后在用少量的训练数据微调网络（fine tuning），可以提高网络的泛化能力。
- 当训练数据比较多时，也可以先用预训练好的 feature 提取器初始化网络，这样能够加快收敛速度。

现有的比较好的特征提取器（如 VGG、ResNet），是在 ImageNet 这样的大规模标注数据集上预训练好的。还有一些特征提取器，是基于元数据（raw metadata）训练出来的，但元数据并不是总能获取到。这些方法都属于监督学习算法，构建数据集的成本很大，所以大家逐渐开始关注如何用非监督方法来提取高质量的特征。

目前已经有了一些非监督提取特征的方法，其中性能最好的方法是在 ImageNet 数据集上进训练出来的。虽然在训练过程中没有使用 ImageNet 自带的标签信息，但是 ImageNet 排除标签后，还是有很多“人为的监督因素”隐含在里面，例如：ImageNet 里面的所有图片都经过人工挑选、标注，并且在挑选的时候，还考虑到了图片类别的多样性、平衡性。之前的一些实验结果发现：这种“隐藏的监督因素”对特征质量也是有影响的，换一个其它的完全未处理的数据集，特征质量会下降。

如何在完全未处理的数据（uncurate data）上训练出好的 feature 提取器，就是这篇论文所解决的问题。

2) 解决方法:

文章提出了一个基于大规模未处理数据的非监督特征学习方法：DeeperCluster。该方法受到以下思路的启发：

1) Self-supervised learning。这类方法通过设计一个“辅助任务”来实现。“辅助任务”会在输入数据上加上一个伪标签，然后用一个网络来预测这个伪标签。例如：把输入的图片旋转一个角度，然后预测旋转的角度。（也可以是对图片做一个变换，然后预测出变换）
2) Clustering。通过 k-means 在图片的特征空间上进行聚类，能给每个图片设置一个类标签，然后可以用一个网络来预测这个类标签。

这两个方法，相当于用两种不同的方式，”创造“出两种不同类型的类的集合。假设用 Self-supervised learning” 创造“出来的类的集合是 A，Clustering 创造出来的类的集合是 B，A 和 B 的笛卡尔乘积为 C，用 C 来表示所有图片的类型的集合（每张图片从属的两个不同类型的类，组合成的新类，肯定在 C 里面）。然后用一个网络来预测图片所属的新类。通过这种方式，只需要用一个网络，就可以预测两个类型，把 Self-supervised learning 和 Clustering 结合到了一起。

表 I: DeeperCluster, RotNet 和 DeepCluster 的比较

Method	Data	ImageNet	Places	VOC2007
Supervised	ImageNet	70.2	45.9	84.8
RotNet	ImageNet	32.7	32.6	60.9
DeepCluster	ImageNet	48.4	37.9	71.9
RotCluster	YFCC100M	45.6	42.1	73.0

3) 具体实现:

我们假设输入 $x_1 \dots x_n$ 为被旋转的图像，每个图像与一个编码其旋转角度的目标标签 y_n 和一个赋值为 z_n 的簇相关. 在训练过程中，随着视觉表示 (特征) 的变化，聚类产生的标签也发生了变化. 我们用 Y 表示可能的旋转角度集合，用 Z 表示可能的簇分配集合。将自监督与深度聚类相结合的一种方法是将定义的损失相加。然而，把这些损失加在一起假定旋转和聚类关系是两个独立的任务，这可能会限制可以捕获的信号. 相反，我们使用笛卡尔积空间 $Y \times Z$，它可以捕获两个任务之间更丰富的交互。我们得到如下优化问题:

$$\min_{\theta,W} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(y_n \otimes z_n, W f_\theta(x_n)) \tag{1}$$

y_n 是聚类生成的伪标签,$f_\theta(x_n)$ 是卷积层提取出来的特征，经过分类器 V 之后得到预测的标签，将聚类

的伪标签 yn 作为真实标签，去优化 θ 和 V. 在优化过程中，伪标签是固定的，学习过程的质量完全取决于它们的相关性.Self-supervised learning 采用了判断图片旋转类型的方法（参考《Unsupervised representation learning by predicting image rotations》）。Clustering 采用了分级的方式：先把所有图片聚类成 m 个大类；再通过旋转（总共有 4 中旋转），把一个大类变成 4 个大类；最后每个大类内的图片再聚类成 n 的小类。

按照这一流程，我们将目标标签划分为一个 2 级层次结构。每一个都预测相对应的超类 s 中的子类成员。线性分类器的参数 $(V, W_1, ..., W_S)$ 和 θ 是通过最小化损失函数共同学习的，损失函数如下：

$$\frac{1}{N}\sum_{n=1}^{N}[\mathcal{L}(V f_\theta(x_n), y_n) + \sum_{s=1}^{S} y_{ns}\mathcal{L}(W_s f_\theta(x_n), z_n^s)] \quad (2)$$

用 k-means 在图片的特征空间上进行聚类时，由于在训练过程中，模型参数在不断的更新，相当于特征提取器在不断更新，所以图片的特征空间也在不断的更新。算法并不是每次更新模型参数时，都重新用 k-means 为所有图片重新聚类，而是每经过 T 个 epoch 重新聚类一次。

算法步骤如下：

1、最终目的是要训练出来一个好的特征提取网络。所以，先要初始化一个未训练的特征提取网络，用来提取图片特征。

2、用 k-means 对所有图片（未旋转）的特征进行聚类（clustering），分成 m 个大类。（每张未旋转的图片有了一个大类的标签）

3、对 m 个大类进行扩展：用“旋转角度的类型”（4 种）和“m 个大类“计算笛卡尔乘积，得到 4m 个新的大类。每张经过旋转的图片，只属于 4m 个大类里面的一个类型。（每张经过旋转的图片有了一个大类的标签）

4、4m 个大类的每个类，再分别使用 k-means 对属于本类中的图片的特征进行聚类，分成 k 个子类。（每张经过旋转的图片，在一个大类下面，又有了一个子类的标签）

5、构建 1 个大类的分类器（4m 个大类），每个大类下面再构造 1 个子类分类器（4 个子类），总共 1 + 4m 个分类器。所有分类器分成两部分：特征提取网络，分类网络。不同分类器的特征提取网络是共享的，分类网络各不相同。然后用 Deeper Clustering loss 来训练，训练 T 个 epochs。

6、T 个 epochs 后，重新回到第 2 步。

4) 实验结果:

作者给出的实验结果表明：DeeperCluster 在大规模未处理数据集上（YFCC100M）训练出的 feature 质量比其它非监督学习方法训练出来的 feature 质量好一些，用 Places205 数据集测试时，性能逼近监督学习算法。

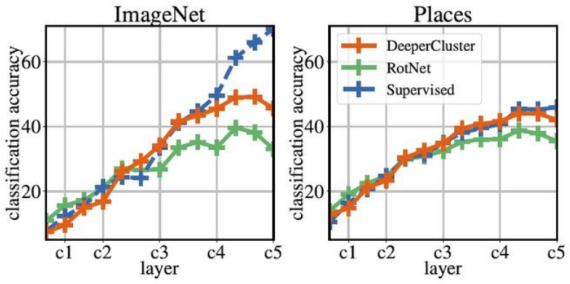

图 12: 线性分类器在 ImageNet 的准确率

5) 总结:

文章提出了一个基于大规模未处理数据的非监督 feature 学习方法：DeeperCluster。该方法受 self-supervised learning 和 clustering 两种方法的启发，用一个网络，预测了两个类型，把 Self-supervised learning 和 Clustering 结合到了一起，并用实验说明了方法的有效性。

IV. 意义与展望

A. 意义

通过引入自监督标签来为下游任务学得有效的特征向量表示，自监督学习确实显著地提高了下游任务的学习性能。但是现阶段如何设计前置任务，或如何进一步提高自监督学习方法的性能，仍是一个很大的问题。据我们所知，当前仍缺乏相关理论对其设计进行指导。

事实上，从多视图角度看，自监督学习中引入的自监督信号实质上是对原始数据进行了各种变换 (如旋转、着色和拼图等) 从而产生多个变换数据 (可视为多个视图数据)，这恰好落入早期提出的单视图的多视图学习框架。换句话说，自监督学习的本质就是对原数据

进行多视角的数据增广，这不同于传统的数据增广，因为它考虑到了所附的自监督信号。从该视角来看，我们相信在理论上能借鉴已有的多视图学习理论，弥补自监督学习理论的缺乏，并对其进一步拓展，从而实现计算机视觉等领域的进一步突破。

B. 未来展望

我们的世界是在严格的物理学、化学、生物学规则下运行的，视觉信号是这些内在规则的外在反映，而深度学习，正好非常擅长处理高维的视觉信号。所以，自监督学习的存在和发展是必然的，因为世界本身就是有序的、低熵的，这使得数据本身就已经包含了丰富的信息。自监督学习看似神奇，但理解了其本质之后，也就会觉得是情理之中了。当然，目前学术界对自监督学习的理解程度，可能也只是九牛一毛而已。未来会走向什么方向，谁也说不准。目前是基于数据之间的结构的 instance discrimination 处于 state-of-the-art，未来，基于 priors 的方法更胜一筹也是有可能的。所以，千万不要受限于一类方法，不要让自监督学习变成了调参游戏，自监督领域的想象空间其实非常大。

参考文献

[1] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[2] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2959–2968, 2019.

[3] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[4] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[5] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9359–9367, 2018.

[6] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019.