·Unsupervised Feature Learning via Non-Parametric Instance Discrimination

1. 摘要

我们是否可以学习一个好的特征表示，通过仅仅要求该特征对单个实例具有辨别力，来捕捉实例之间的明显相似性，而不是类别。我们将这种直觉表述为实例级的非参数化问题，并使用噪声对比估计来解决大量实例类所带来的计算挑战。

1. 问题重述

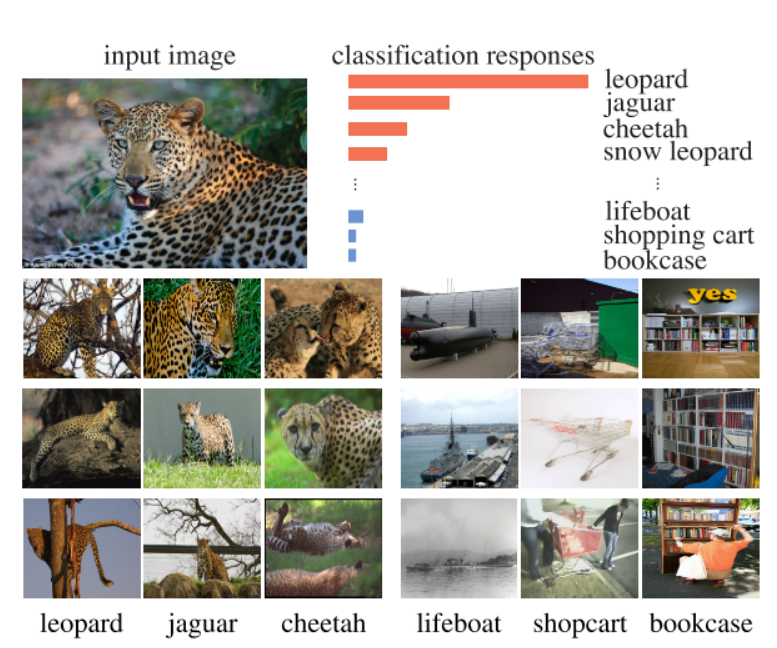


图 1 监督学习的结果激励了我们的非超视距方法。对于豹类的图像，从训练有素的神经网络分类器中得到最高响应的类别都是视觉上相关的，例如美洲虎和猎豹。不是语义标签，而是数据本身的明显相似性使一些类别比其他类别更接近。我们的无监督方法将类的监督发挥到了极致，并学习了一个能区分单个实例的特征表征。

图1显示，豹类的图像被美洲豹类而不是书架类评为高得多。这样的观察显示，典型的判别性学习方法可以自动发现语义类别之间的表观相似性，而不需要明确地指导这样做。换句话说，表面相似性不是从语义注释中学习的，而是从视觉数据本身学习的。

我们将 "类 "上的监督推向 "实例 "上的监督的极端，并问：我们能否通过纯粹的判别性学习来学习一个反映实例之间明显相似性的有意义的度量？一幅图像本身是有区别的，而且每幅图像都可能与同一语义类别中的其他图像有很大的不同[23]。如果我们学习辨别单个实例，而没有任何语义类别的概念，我们最终可能会得到一个捕捉实例间表面相似性的表征，就像类的监督学习仍然保留类间的表面相似性一样。这种将无监督学习表述为实例级的判别在技术上也很有吸引力，因为它可以从判别性监督学习的最新进展中获益，例如在新的网络结构上。

然而，我们也面临着一个重大的挑战，因为现在类 "的数量是整个训练集的大小。对于ImageNet来说，它将是120万而不是1000个类。简单地将softmax扩展到更多的类变得不可行了。我们通过使用[噪声对比估计（NCE）](https://baike.baidu.com/item/%E5%99%AA%E5%A3%B0%E5%AF%B9%E6%AF%94%E4%BC%B0%E8%AE%A1/22768348" \l ":~:text=%E5%99%AA%E9%9F%B3%E5%AF%B9%E6%AF%94%E4%BC%B0%E8%AE%A1%EF%BC%88NCE%2C%20Noise,Contrastive%20Estimation%EF%BC%89%E6%98%AF%E4%B8%80%E7%A7%8D%E6%96%B0%E7%9A%84%E7%BB%9F%E8%AE%A1%E6%A8%A1%E5%9E%8B%E4%BC%B0%E8%AE%A1%E6%96%B9%E6%B3%95%EF%BC%8C%E7%94%B1Gutmann%E5%92%8CHyv%C2%A8arinen%E6%8F%90%E5%87%BA%E6%9D%A5%EF%BC%8C%E8%83%BD%E5%A4%9F%E7%94%A8%E6%9D%A5%E8%A7%A3%E5%86%B3%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C%E7%9A%84%E5%A4%8D%E6%9D%82%E8%AE%A1%E7%AE%97%E9%97%AE%E9%A2%98%EF%BC%8C%E5%9B%A0%E6%AD%A4%E5%9C%A8%E5%9B%BE%E5%83%8F%E5%A4%84%E7%90%86%E5%92%8C%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86%E4%B8%AD%E5%BE%97%E5%88%B0%E5%B9%BF%E6%B3%9B%E5%BA%94%E7%94%A8%E3%80%82)[9]来逼近完整的softmax分布，并采用近似正则化方法[29]来稳定学习过程，从而应对这一挑战。

为了评估无监督学习的有效性。过去的工作，如[2, 31]，依赖于线性分类器，如支持向量机（SVM），在测试时将学到的特征连接到分类中。然而，目前还不清楚为什么通过训练任务学到的特征对于未知的测试任务来说可以是线性分离的。

我们主张在训练和测试中采用非参数方法进行训练和测试。我们将实例级的辨别力表述为一个度量学习问题，其中实例之间的距离（相似性）是以非参数的方式直接从特征中计算出来的。也就是说，每个实例的特征被存储在一个离散的memory bank中，而不是网络中的权重。在测试时，我们根据所学到的方法，使用k-nearest neighbors（kNN）进行分类。因此，我们的训练和测试是一致的，因为我们模型的学习和评估都涉及到图像之间的相同度量空间。我们报告并比较了SVM和kNN准确率的实验结果。

1. 相关工作

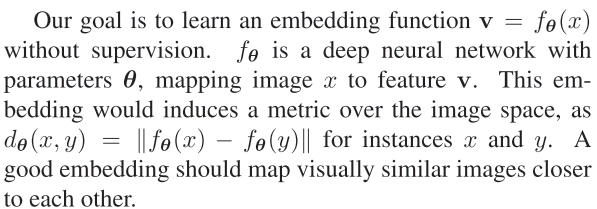
在没有人类提供的标签的情况下，人们对无监督学习的兴趣越来越大。以前的工作主要分为两类。1）生成模型和2）自我监督方法。

**生成模型**。生成模型的主要目标是尽可能忠实地重建数据的分布。经典的生成模型包括受限博尔兹曼机（RBMs）[12, 39, 21]和自动编码器[40, 20]。生成模型产生的潜在特征也可以帮助物体识别。最近的方法，如生成式对抗网络[8，4]和变异自动编码器[14]，既提高了生成的质量，也提高了特征学习的效果。

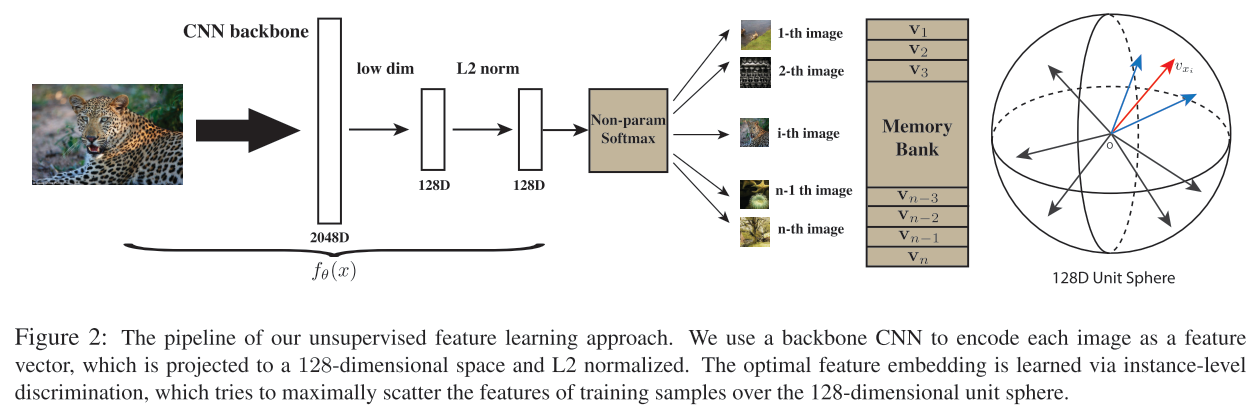
**自监督学习**。自监督学习利用数据的内部结构，制定预测任务来训练一个模型。具体来说，该模型需要预测一个实例中被遗漏的方面或组成部分，并给出其余部分。为了学习图像的表征，任务可以是：预测上下文[2]，计算物体[28]，填补图像的缺失部分[31]，从灰度图像中恢复颜色[47]，甚至解决一个拼图[27]。对于视频，自我监督的策略包括：通过跟踪[44, 45]来利用时间的连续性，预测未来[42]，或保留自我运动的等值性[13, 50, 30]。最近的工作[3]试图结合几个自我监督的任务来获得更好的视觉表征。虽然自我监督学习可以捕捉到一个实例的各部分或各方面之间的关系，但目前还不清楚为什么一个特定的自监督任务有助于语义识别，以及哪个任务会是最佳的。

[**度量学习**](https://zhuanlan.zhihu.com/p/348998459#:~:text=%E5%BA%A6%E9%87%8F%E5%AD%A6%E4%B9%A0%E6%98%AF%E4%BB%8E%E6%95%B0%E6%8D%AE%E4%B8%AD%E5%AD%A6%E4%B9%A0%E4%B8%80%E7%A7%8D%E5%BA%A6%E9%87%8F%E6%95%B0%E6%8D%AE%E5%AF%B9%E8%B1%A1%E9%97%B4%E8%B7%9D%E7%A6%BB%E7%9A%84%E6%96%B9%E6%B3%95%E3%80%82,%E5%A6%82%E5%9B%BE1c%E6%89%80%E7%A4%BA%EF%BC%8C%E5%85%B6%E7%9B%AE%E6%A0%87%E6%98%AF%E4%BD%BF%E5%BE%97%E5%9C%A8%E5%AD%A6%E5%BE%97%E7%9A%84%E8%B7%9D%E7%A6%BB%E5%BA%A6%E9%87%8F%E4%B8%8B%EF%BC%8C%E7%9B%B8%E4%BC%BC%E5%AF%B9%E8%B1%A1%E9%97%B4%E7%9A%84%E8%B7%9D%E7%A6%BB%E5%B0%8F%EF%BC%8C%E4%B8%8D%E7%9B%B8%E4%BC%BC%E5%AF%B9%E8%B1%A1%E9%97%B4%E7%9A%84%E8%B7%9D%E7%A6%BB%E5%A4%A7%E3%80%82)。每一个特征表示F都会在实例x和y之间诱导一个度量：dF(x, y) = ||F(x) - F(y) || 因此，特征学习也可以被看作是某种形式的度量学习。关于度量学习已经有了广泛的研究[15, 33]。成功应用度量学习往往能带来有竞争力的表现，例如在人脸识别[35]和人物重新识别[46]上。在这些任务中，测试时的类别与训练时的类别是不相干的。一旦一个网络被训练出来，人们只能从它的特征描述中推断，而不能从后续的线性分类器中推断。公制学习已经被证明对少数的学习是有效的[38, 41, 37]。关于人脸识别的度量学习的一个重要技术点是规范化[35, 22, 43]，我们在这项工作中也利用了这一点。请注意，这里提到的所有方法在某些方面都需要监督。我们的工作有很大的不同：它以无监督的方式学习特征，从而学习诱导的度量，没有任何人类注释。

1. 详细阐述



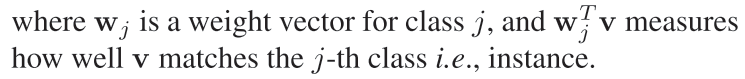
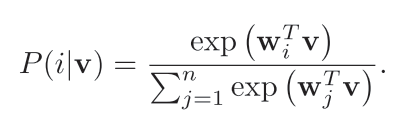
我们新颖的无监督特征学习方法是实例级辨别。我们将每个图像实例视为一个独立的类别，并训练一个分类器来区分各个实例类别（图2）。



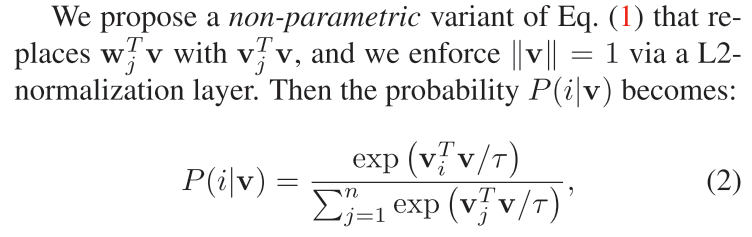
（极地距离会不会有效？）

* 1. Non-Parametric Softmax Classifier

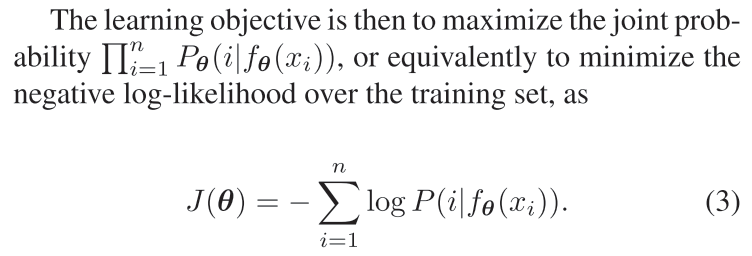
我们使用softmax准则来完成实例水平的目标分类。假定我们有N张图像x1-xn，自然分为n class，对应的特征是v1-vn。在传统的参数化softmax下，图像x被分类为第i个样本的概率是：



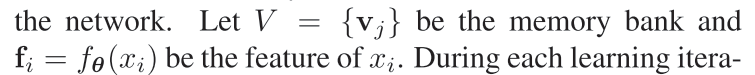
**非参数化的分类器**。公式（1）中的参数化softmax表述的问题是，权重向量w作为类别原型，阻止了实例之间的明确比较。



其中τ是一个温度参数，控制分布的集中程度。τ对监督特征学习很重要，对调整我们单位球体上的v的浓度也很必要。



**使用Memory Bank学习**。为了计算公式（2）中的概率P（i|v），需要所有图像的{vj}。我们不需要每次都详尽地计算这些表征，而是维持一个特征存储库V来存储它们[46]。在下文中，我们为存储库和从网络转发的特征引入单独的符号。



在每个学习迭代过程中，表征fi以及网络参数θ都通过随机梯度下降进行优化。然后，fi在相应的实例条目fi→vi处被更新到V。我们将存储库V中的所有表征初始化为单位随机向量。

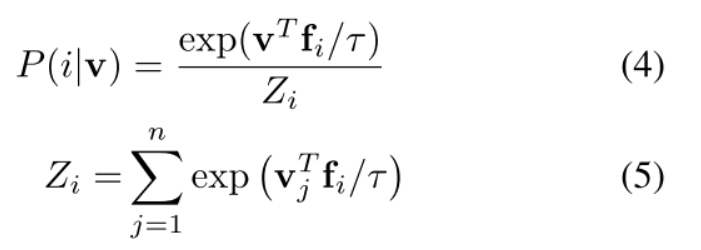
**讨论**。从类的权重向量wj直接到特征表示vj的概念性变化是很重要的。原始softmax公式中的权重向量{wj}只对训练类有效。因此，它们不能被推广到新的类别，或者在我们的环境中，新的状态。当我们去掉这些权重向量时，我们的学习目标就完全集中在特征表示和其诱导的度量上，它可以在空间的任何地方应用，并在测试时应用于任何新的实例。

在计算上，我们的非参数公式消除了计算和存储{wj}梯度的需要，使其在大数据应用中更具可扩展性。

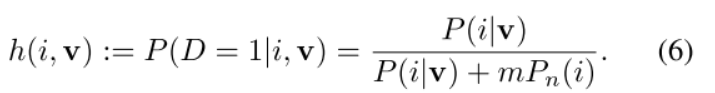
* 1. 噪声-对比估计

计算公式(2)中的非参数softmax的成本很高。当类的数量非常大时，例如在数百万的规模上，计算非参数softmax的成本是非常高的。类似的问题在学习单词嵌入的文献中得到了很好的解决[25, 24]，其中单词的数量也可以扩展到数百万。流行的减少计算的技术包括分层softmax[26]，噪声对比估计（NCE）[9]，和负采样[24]。我们使用NCE[9]来近似于完全的softmax。

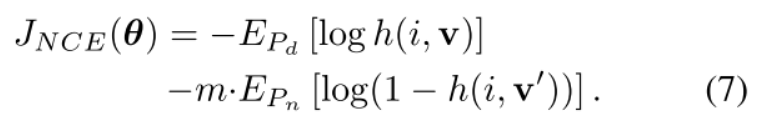
我们将NCE应用于我们的问题，以解决计算与训练集中所有实例的相似度的困难。其基本思想是将多分类问题变成一组二元分类问题，其中二元分类任务是区分数据样本和噪声样本。具体来说，在我们的模型下，存储库中的特征表示v对应于第i个例子的概率是



其中Zi是归一化常数。我们将噪声分布形式化为一个均匀分布。Pn=1/n。根据先前的工作，我们假设噪声样本比数据样本的频率高m倍。那么，具有特征v的样本i来自数据分布（用D=1表示）的后验概率为

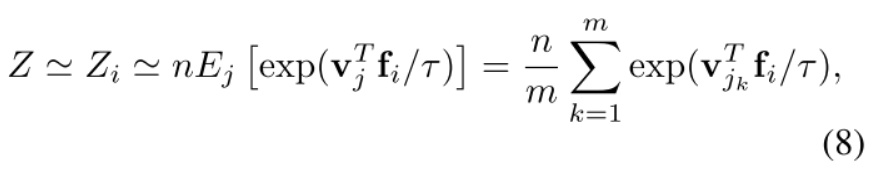


我们的近似训练目标是使数据和噪声样本的负对数后验分布最小化：



这里，Pd表示实际的数据分布。对于Pd，v是对应于xi的特征；而对于Pn，v’是来自另一幅图像的特征，根据噪声分布Pn随机取样。在我们的模型中，v和v’都是从非参数存储库V中取样的。

根据公式（4）计算归一化常数Zi是很昂贵的。我们遵循[25]，将其作为一个常数，并通过蒙特卡洛近似法估计其值。

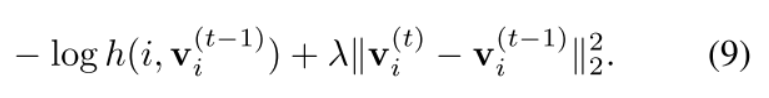


其中{jk}是一个随机的指数子集。从经验上看，我们发现从初始批次中得出的近似值在实践中足够好用了。

NCE将计算复杂性从O(n)降到每个样本的O(1)。在如此大幅度的减少下，我们的实验仍然产生了有竞争力的性能。

* 1. 近似正则化

与典型的分类设置不同，每个类别有许多实例，而我们每个类别只有一个实例。在每个训练周期中，每个类别只被访问一次。因此，学习过程会因随机抽样波动而出现大量振荡。我们采用近似优化方法[29]，并引入一个附加项来鼓励训练动态的平稳性。在当前迭代t，数据xi的特征表示是由网络v(t)计算出来的 vit = fθ(xi)。记忆库中的 V = {v(t-1)}，所有的表征都存储在上一次迭代中。来自Pd的正样本的损失函数是：



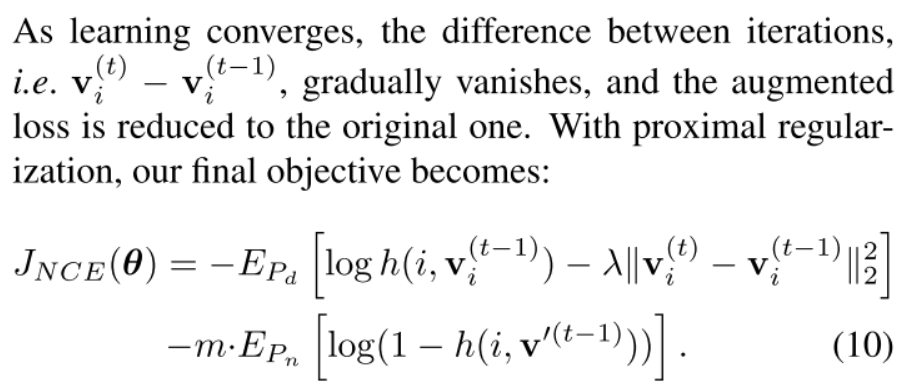
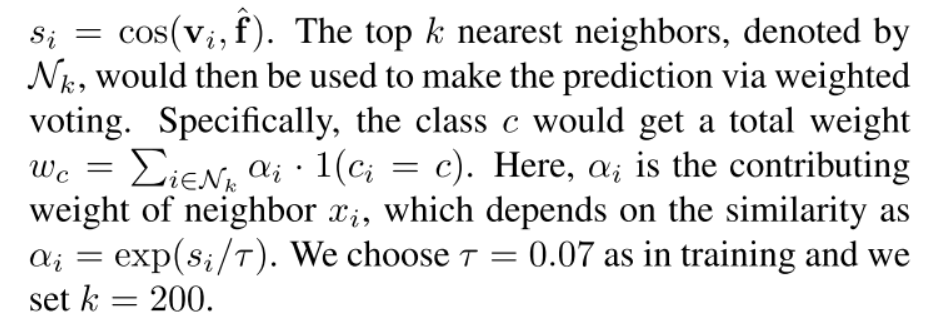


图3显示，根据经验，近似正则化有助于稳定训练，加快收敛速度，并改善学习到的表征，其额外成本可以忽略不计。

* 1. 加权的K-近邻分类器

为了对测试图像ˆx进行分类，我们首先计算其特征ˆf =fθ(ˆx)，然后用余弦相似度si = cos(vi, ˆf)将其与存储库中所有图像的嵌入进行比较。



1. 实验

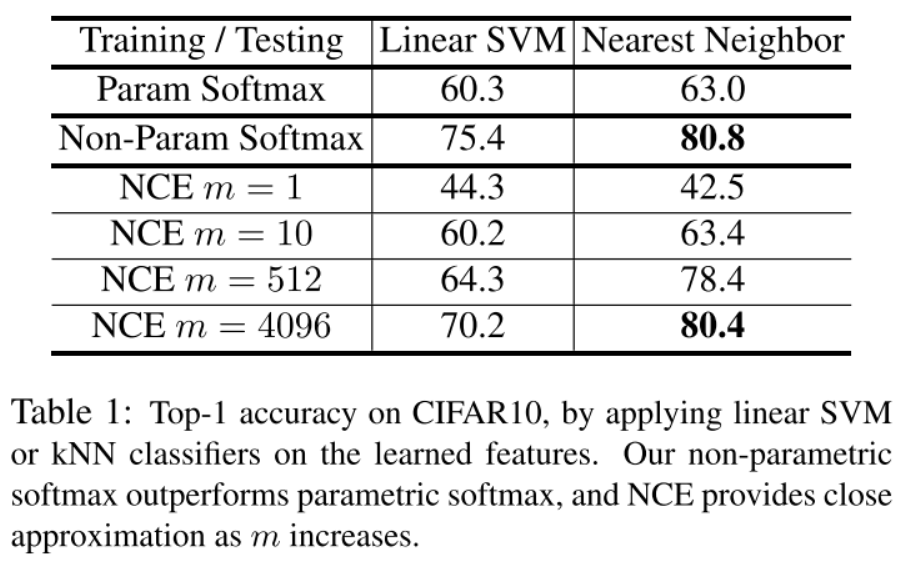
我们进行了4组实验来评估我们的方法。第一组是在CIFAR-10上，将我们的非参数化softmax与参数化softmax进行比较。第二组是在ImageNet上，将我们的方法与其他无监督的学习方法进行比较。最后两组实验研究了两个不同的任务，即半监督学习和物体检测，以显示我们学习的特征表示的泛化能力。

* 1. Parametric vs. Non-parametric Softmax

我们的方法的一个关键创新点是非参数化的softmax函数。与传统的参数化softmax相比，我们的softmax允许非参数化指标转移到监督任务中。我们在CIFAR-10[17]上比较了参数化和非参数化的公式，该数据集有10个类别的50,000个训练实例。这种规模使我们能够计算公式（2）中的非参数softmax而不需要任何近似值。我们使用ResNet18作为骨干网络，其输出特征被映射为128维向量。我们根据学习到的特征表示评估分类效果。

学到的特征表示。一个常见的做法[48, 2, 31]是在训练集上对所学特征进行SVM训练，然后根据从训练网络中提取的特征对测试实例进行分类。此外，我们还使用近邻分类器来评估学到的特征。后者直接依赖于特征度量，可能会更好地反映表征的质量。

表1显示了CIFAR10的前1名分类精度。在用参数化softmax学习的特征上，我们用线性SVM和kNN分类器获得的准确率分别为60.3%和63.0%。在用非参数化softmax学习的特征上，线性分类器和近邻分类器的准确率分别上升到75.4%和80.8%，后者有18%的显著提升。我们还研究了NCE接近非参数化softmax的质量（第3.2节）。近似是由m控制的，即每个实例的否定数。当m=1时，kNN的准确性明显下降到42.5%。随着m的增加，性能稳步提高。当m=4,096时，准确率接近m=49,999时的准确率--没有任何近似的全形式评估。这个结果保证了NCE是一个有效的近似。



* 1. 图像分类

我们在ImageNet上学习一个特征表示ILSVRC[34]，并将我们的方法与代表性的无监督学习方法进行比较。

**实验设置**。我们通过经验性的验证来选择设计参数。特别是，我们设定温度τ=0.07，并使用NCE的m=4，096来平衡性能和计算成本。该模型使用SGD的动量训练了200个epochs。批量大小为256。学习率初始化为0.03，在第一个120个历时之后，每40个历时的系数为0.1。

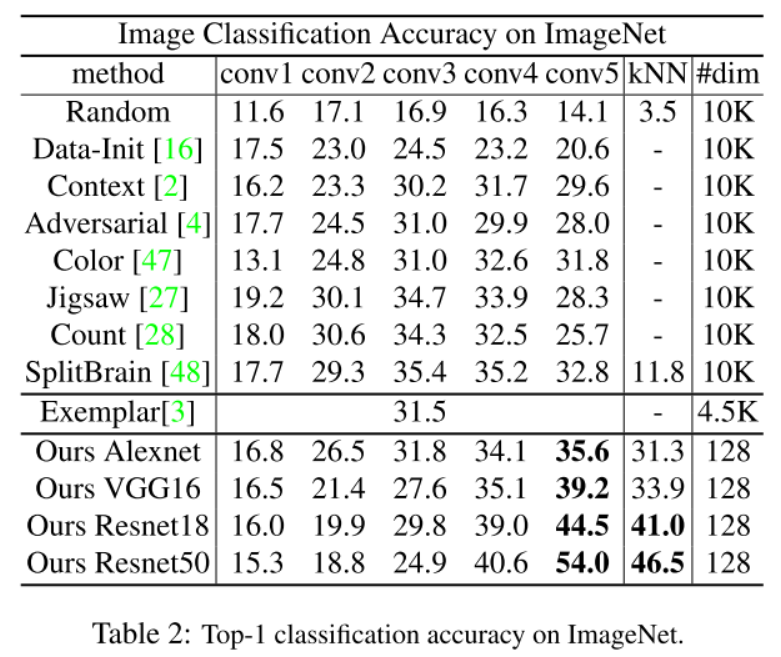
**比较**。我们将我们的方法与随机初始化网络（作为下限）和各种无监督学习方法进行比较，包括自监督学习[2, 47, 27, 48]、对抗学习[4]和Exemplar CNN[3]。分裂脑自动编码器[48]是一个强有力的基线，代表了目前的技术水平。这些方法的结果在其原始论文中都是用AlexNet架构[18]报告的，除了典范CNN[5]，其结果是用ResNet-101[3]报告。由于网络架构对性能有很大的影响，我们考虑几个典型的架构。AlexNet [18], VGG16 [36], ResNet-18, 和ResNet-50 [10]。我们用两种不同的协议来评估性能：（1）对从conv1到conv5的中间特征执行线性SVM。注意，在VGG16和ResNet[36, 10]中也有相应的层。(2) 在输出特征上执行kNN。表2显示：

1. 通过AlexNet和中间特征的线性分类，我们的方法达到了35.6%的准确率，超过了所有的基线，包括最先进的。我们的方法可以轻易地扩展到更深的网络。当我们从AlexNet转移到ResNet-50时，我们的准确率提高到了54.0%，而使用典范CNN[3]的准确率只有31.5%，即使使用ResNet-101也是如此。

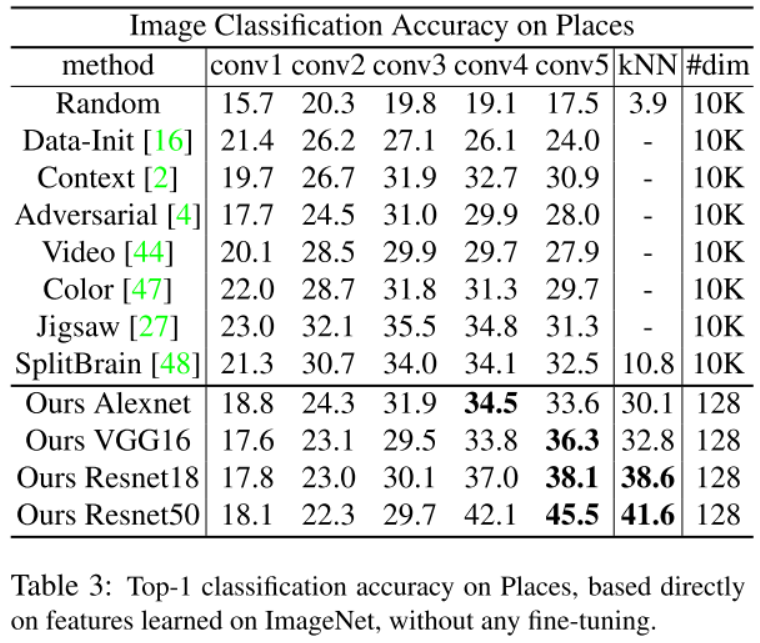
2. 在最终的128个维度特征上使用近邻分类，我们的方法在AlexNet、VGG16、ResNet-18和ResNet-50上取得了31.3%、33.9%、41.0%和46.5%的准确率，不比线性分类结果低多少，表明我们学到的特征能引起一个合理的好指标。作为比较，对于Split-brain，用conv3特征的近邻分类，准确率下降到8.9%，而将特征投影到128维后，准确率下降到11.8%。

3. 3.用我们的方法，当我们检查从早期到后期的学习特征表示时，性能逐渐增加，这通常是可取的。而其他所有的方法，在超过conv3或conv4之后，性能就会下降。

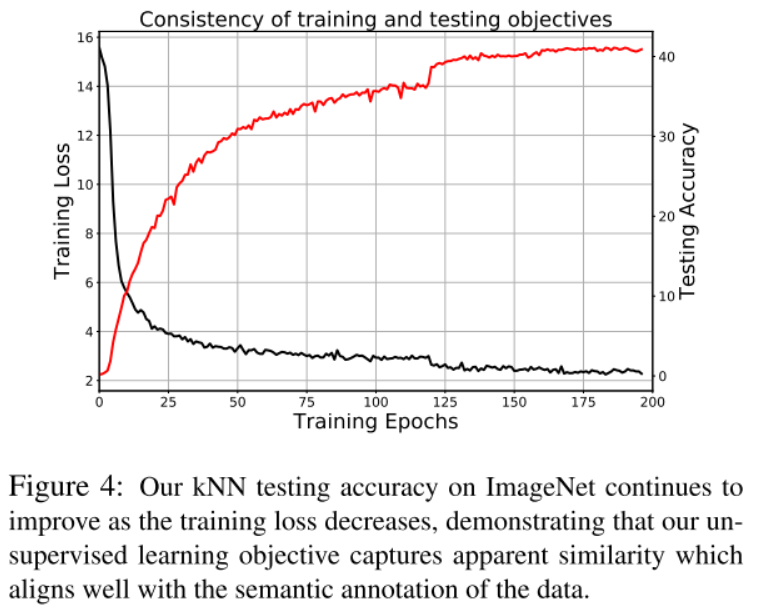
4. 值得注意的是，来自中间卷积层的特征可能超过10,000维。因此，对于其他方法来说，使用表现最好的层的特征会产生巨大的存储和计算成本。我们的方法在最后一层产生一个128维的表示，这是非常有效的工作。ImageNet中所有128万张图像的编码特征只需要大约600MB的存储空间。在Titan X GPU上对这个数据集进行详尽的近邻搜索，每张图片只需要20毫秒。



**特征泛化**。我们还研究了所学的特征表征如何能推广到其他数据集。在相同的设置下，我们在Places[49]上进行了另一个大规模的实验，这是一个用于场景分类的大型数据集，其中包含205个类别的245万张训练图像。在这个实验中，我们直接使用了在ImageNet上训练的特征提取网络，没有进行微调。表3比较了不同方法和不同评价策略下的结果。同样，在使用conv5特征的线性分类器的情况下，我们的方法取得了具有竞争力的性能，AlexNet的最高准确率为34.5%，ResNet-50为45.5%。在最后一层的最近邻居比中间层小得多的情况下，我们与ResNet-50的准确率达到了41.6%。这些结果表明，用我们的方法学到的表征具有显著的泛化能力。



**训练和测试目标的一致性**。无监督的特征学习是困难的，因为训练目标与测试目标无关。一个好的训练目标应该反映在测试性能的持续改善上。我们研究了训练损失和测试精度之间的关系。图4显示，随着训练的进行，我们的测试准确率不断提高，没有过拟合的迹象。这也表明，对训练目标进行更好的优化可能会进一步提高我们的测试精度。



嵌入特征大小。我们研究当我们将嵌入大小从32到256变化时，性能如何变化。表4显示，性能从32开始增加，在128时趋于平稳，到256时似乎达到了饱和。

训练集大小。为了研究我们的方法是如何随数据大小而变化的，我们用不同比例的ImageNet数据训练不同的表征，并使用最近的邻居来评估全部标记集的分类性能。表5显示，我们的特征学习方法受益于更大的训练集，并且测试精度随着训练集的增加而提高。这一特性对于成功的无监督学习至关重要，因为在野外并不缺乏未标记的数据。

定性案例研究。为了说明学到的特征，图5显示了使用学到的特征进行图像检索的结果。上面四行显示了最好的情况，所有前10个结果都与查询的类别相同。下面四行显示了最坏的情况，即前10名中没有一个是相同类别的。然而，即使在失败的情况下，检索到的图像仍然与查询结果非常相似，这证明了我们无监督学习目标的力量。

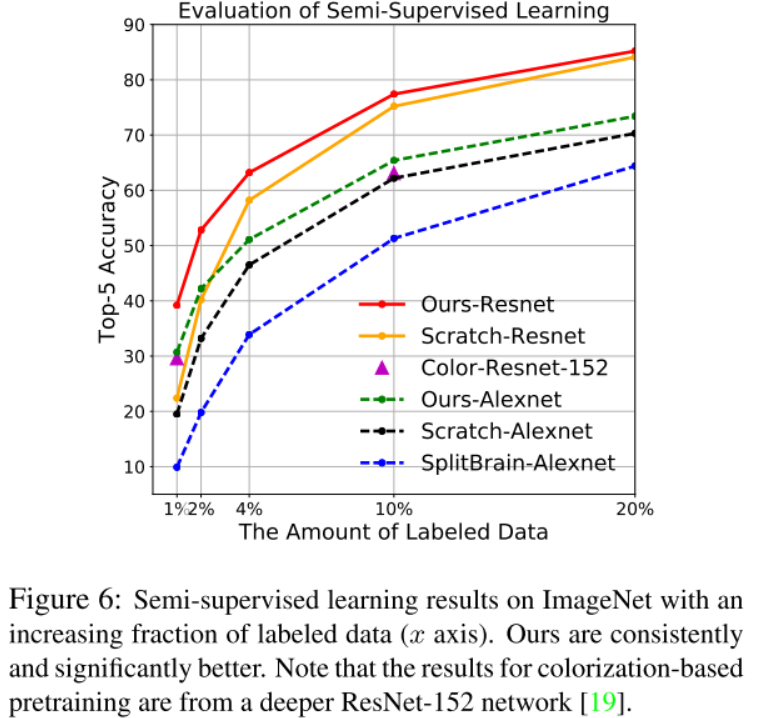
* 1. 半监督式学习

我们现在要研究的是，学到的特征提取网络如何受益于其他任务，以及它是否可以为其他任务的迁移学习提供良好的基础。一个可以从无监督学习中获益的常见情况是，我们有大量的数据，但其中只有一小部分是有标签的。一个自然的半监督学习方法是首先从大量的无标签数据中学习，然后在少量的有标签的数据中微调模型。我们随机选择ImageNet中的一个子集作为标记，并将其他子集视为未标记的。我们进行上述的半监督学习，并在验证集上测量分类精度。为了与[19]进行比较，我们在此报告前五名的准确率。我们将我们的方法与三个基线进行比较：

1. Scratch，即在小的标记子集上进行完全监督训练
2. 用于预训练的Split-brain[48]
3. 用于预训练的Col- orization[19]。

标签子集上的微调需要70个历时，初始学习率为0.01，每30个历时的衰减率为10。我们将标签子集的比例从整个数据集的1%到20%不等。

图6显示，我们的方法明显优于其他所有的方法，而且我们的方法是唯一一个优于从有限的标记数据中进行监督学习的方法。当只有1%的数据被标记时，我们的表现要比其他方法高出10%，这表明我们从未标记的数据中学习的特征对任务适应是有效的。



* 1. 物体检测

为了进一步评估所学特征的泛化能力 为了进一步评估所学特征的泛化能力，我们将所学网络转移到PASCAL VOC 2007[6]的物体检测新任务中。从头开始训练物体检测模型通常很困难，一个普遍的做法是在ImageNet上预训练基础CNN，并针对检测任务对其进行微调。我们用Fast R-CNN[7]与AlexNet和VGG16架构进行实验。VGG16架构，以及使用ResNet-50的Faster R-CNN[32]。当微调Fast R-CNN时，学习率被初始化为0.001，并在每50K次迭代后缩减10倍。当对AlexNet和VGG16进行微调时，我们遵循标准做法，固定conv1模型的权重。当微调Faster R-CNN时，我们固定第三类残差块以下的模型权重，只更新上面的层，并冻结所有批处理正常化层。我们按照标准的管道进行微调，不使用[2]中提出的重新缩放方法。我们使用VOC 2007中的标准trainval集进行训练和测试。我们比较了三种设置。1）直接从头开始训练（下限），2）以监督的方式在ImageNet上进行预训练（上限），3）使用各种无监督的方法在Ima-geNet或其他数据上进行预训练。表6列出了以平均精度（mAP）表示的检测性能。

平均精度（mAP）。使用AlexNet和VGG16，我们的方法实现了48.1%和60.5%的平均精度，与最先进的无监督方法相当。使用Resnet-50，我们的方法实现了65.4%的mAP，超过了所有以前的无监督学习方法。这也表明，随着网络的深入，我们的方法具有良好的扩展性。在有监督的预训练中，仍有11%的差距需要缩小，才能达到76.2%的mAP。

1. 评析
   1. 总结

我们提出了一种无监督的特征学习方法，通过一种新的非参数化的softmax公式使实例之间的区别最大化。它的动机是有监督的学习会导致明显的图像相似性。我们的实验结果表明，我们的方法在ImageNet和Places上的图像分类方面优于最先进的方法，其紧凑的128维描述在更多的数据和更深的网络工作中具有良好的扩展性。它还在半监督学习和物体检测任务上提供了有竞争力的泛化结果。