# 专业阅读与写作第三次作业

58119304 朱启鹏

*A. 第一篇论文 [1]Deep Clustering for Unsupervised Learning of Visual Features*

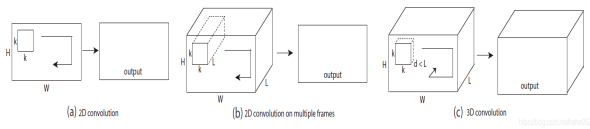
*1) Abstract:* 在无监督学习中，聚类是一个十分常用的算法，而且在计算机视觉领域应用十分广泛，但由于是无标签的学习过程，很难保证十分优秀的正确率，而此时，为了能极大利用样本中的数据，许多人提出了一种从样本中让模型自身学习出标签的方法——自监督（self-supervised）。而在这篇论文中作者采用了一种新式的聚类方法，而他们的方式可以在图上呈现，这种新颖的思想和常规的自监督方法也不尽相同。

*B. 第二篇论文 [2]Learning Spatiotemporal Features with 3D Convolutional Networks*

*1) Abstract:* 本文提出了一种简单而有效的时空特征学习方法，使用在大规模有监督视频数据集上训练的 3D 卷积网络。本文的发现有三个方面：1）与 2D ConvNets 相比，3D ConvNets 更适合于时空特征学习；2）一个在所有层都有 3 × 3 × 3 卷积核的同质结构是 3D ConvNets 的最佳性能结构；3）使用 3D 卷积学习到的特征，使用简单的线性分类器在 4 个不同的基准上均优于最新的方法。此外，它的特点是基于 3D 卷积的快速推理，计算效率非常高。而且在概念上非常简单，易于训练和使用。

*2) 理解:*

*a)* 分析比较：*：* 首先论文介绍了 3D 卷积与 2D 卷积的区别，如图所示：




1）对图像应用二维卷积可生成图像。

2）在视频序列上应用二维卷积（多帧作为多个通道）也会产生图像。

3）在一个视频序列上应用 3D 卷积会产生另一个序列，从而保留输入信号的时间信息。

三维卷积网络非常适合时空特征学习。与 2D-ConvNet 相比，3D-ConvNet 具有更好的时间信息建模能力，这得益于 3D 卷积和 3D 池化操作。在 3D ConvNets 中，卷积和池化操作是在时空上执行的，而在 2D ConvNets 中，卷积和池化操作只是在空间上执行的（如上图）。而 2D ConvNets 在每次卷积运算后都会丢失输入信号的时间信息。只有 3D 卷积才能保留产生输出时间信息。同样的 3D 池化操作也是如此。

根据 2D ConvNets 的研究结果，3*3 卷积核的小感受野和较深的结构产生了最好的结果。因此本文将空间感受野固定为 3*3，并且仅改变 3D 卷积核的时间深度。

*b)* 网络结构*:* 8 个卷积层，5 个池化层，2 个全连接层，1 个 softmax 输出层。所有卷积核均为 3×3×3。第一个 pooling 层 1×2×2，Stride=1×2×2，之后都是 2×2×2，stride=2×2×2。两个全连接层都是 4096。(注：为简单起见，假设视频序列大小为 c × l × h × w，其中 c 是频道数，l 是帧的长度，h 和 w 分别是帧的高度和宽度。三维卷积和池化的核大小为 d × k × k，其中 d 是核的时间深度，k 是核的空间大小。)

*c)* 探索时间核长度*:* 本文主要关注如何通过深度网络聚合时间信息。** 为了寻找一个好的 3D ConvNet 架构，作者只改变卷积层的内核时间深度 di，同时保持所有其他公共设置不变。

作者注意到，所有这些网络在最后一个池化层具有相同大小的输出信号，因此它们对于全连接层具有相同数量的参数。由于核的时间深度不同，卷积层的参数个数也不同。与全连接层中的数百万个参数相比，这些差异非常微小。

*C. 第三篇论文 [3]Local Aggregation for Unsupervised Learning of Visual Embeddings*

Abstract: 神经网络中的无监督学习方法对于推进人工智能具有重要意义，因为它们可以在不需要大量昂贵打标的情况下训练神经网络。而且它们将是由人类部署的更好通用学习类型模型。然而，无监督网络长期以来一直落后于有监督网络的性能，尤其是在大规模视觉识别领域。就在最近，在训练无监督学习嵌入深度卷积神经网络，以实现最大化非参数对象分离和聚类目标方面的发展上，已经显示出缩小这一差距的希望。在这里，我们描述了一种训练嵌入函数以最大化局部聚合度量的方法，从而实现相似的数据对象在所嵌入的空间中一起移动，同时允许不同的对象分离。这种聚合指标是动态的，并且允许出现不同规模的软集群。我们在几个大规模视觉识别数据集上进行评估我们的程序，其中，在 ImageNet 中的对象识别、Places 205 中的场景识别和 PASCAL VOC 中的对象检测方面都实现了最先进的无监督转移学习性能

## References

[1] Mathilde Caron et al. “Deep clustering for unsupervised learning of visual features”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 132–149.

[2] Du Tran et al. “Learning spatiotemporal features with 3d convolutional networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4489–4497.

[3] 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*. 2019, pp. 6002–6012.