视觉局部特征的表达学习

黄永祯 王 亮
中国科学院自动化研究所

关键词：局部特征 特征编码 表达学习 深度学习

引言

对图像中目标的认知是计算机视觉与模式识别领域中最基本的问题之一，涉及目标分类、检测、分割、检索以及场景理解和行为识别等研究内容。由于同一个目标受光照、尺度、旋转、形变等变化而形成不同的成像，用计算机对图像中的目标进行鲁棒的表达与识别依然是一个挑战性的问题。

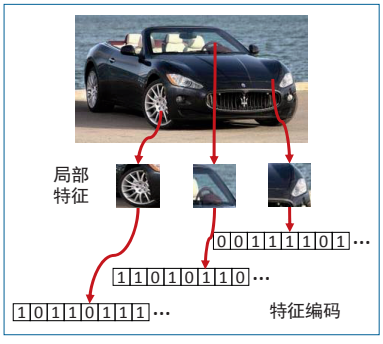

图1 目标局部特征表达示例

如何解决上述问题？我们认为从**目标局部特征来认知目标整体应该是一条可行的途径**。比如，识别出人脸五官的局部特征，并对它们之间的空间关系进行建模，就可以达到认知人脸的目的。这套研究方法通常称为“从局部到全局”，是学术界一个重要的方法论。图1展示了从一幅汽车图片中截取3个图像块组成局部特征，并用“0-1”向量对这些局部特征进行表达示例。

局部特征的表达学习是指用学习方法得到对局部特征的表达。局部特征的表达学习分为非监督学习和监督学习两大类。非监督的局部特征表达学习，强调从目标大量局部特征的分布找到能够鲁棒地表达局部特征的策略。通常这种策略与任务无关，局部特征通过非监督表达学习可用于不同的视觉任务。监督的局部特征表达学习则寻求建立从局部特征到其类别标签的学习模型，以便使局部特征表达有利于完成某种特定的视觉任务。

局部特征的非监督表达学习

局部特征的非监督表达学习广泛用于计算机视觉算法中，如图像分类和检索中常用的特征袋(Bag-of-Features, BoF)模型[1]的核心步骤之一就是局部特征的非监督表达学习。我们以BoF模型为算法平台介绍几种常见的局部特征非监督表达学习算法。BoF模型把每幅图像描述为其局部区域（局部特征）的无序集合，利用聚类算法将局部特征进行聚类，每个聚类中心被看做一个视觉单词，所有视觉单词组成一个视觉词典(visual vocabulary)。**在BoF中，局部特征的表达学习是指用视觉词典来表达每一个局部特征**。与之相应地，每一个局部特征都将在视觉词典的一个或多个视觉单词上产生表达。对视觉词典上的响应进行汇聚(pooling)操作就得到该图像的向量表达[1]。

[1] 常见的聚汇操作有取最大值(MAX pooling)和取平均值(average pooling)，分别是指对所有局部特征在同一个视觉单词上产生的响应取最大值或取平均值，这样所有局部特征在整个视觉词典上的响应就产生一个向量表达，该向量的维度和视觉单词的数量相同。

在 BoF 模型中，局部特征的非监督表达学习又称为局部特征编码。

局部特征编码方法

现有的大部分局部特征编码方法按照其原理分为 4 类，图 2 列举了几种代表性的局部特征编码方法。

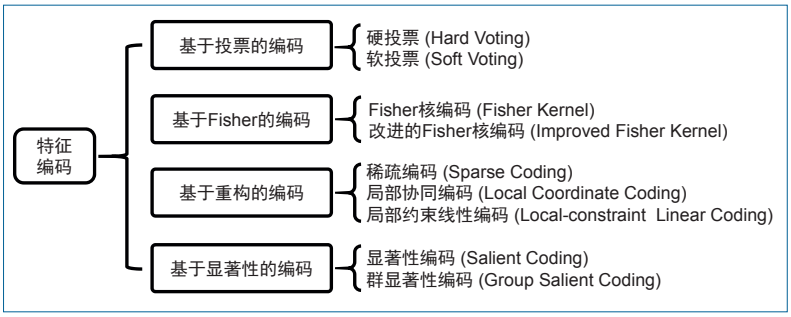

图2 常见的局部特征编码方法分类

基于投票的编码 该方法采用直方图来描述局部特征的分布，直方图中的一个值代表局部特征在一个视觉单词上出现的频率，这样的直方图通常是通过硬投票 (hard voting)[1] 或者软投票 (soft voting)[3] 来构建的。硬投票把每个局部特征分配给它最近的视觉单词，即每个局部特征都在其最近的视觉单词上产生“1”的响应，在其他视觉单词上产生“0”的响应。与硬投票相比，软投票对局部特征分布的描述更加准确，包括：(1) 使用“特征-单词”的距离核函数产生响应，从而进行局部特征编码；(2) 利用多个视觉单词而不是仅仅采用最近的视觉单词进行编码。基于投票的局部特征编码方法比较直观并且易于实现，但是用直方图对高维特征空间的概率密度分布进行近似操作是比较粗糙的。

Fisher 编码 受 Fisher 核的启发，从信号的概率密度函数中推导出梯度向量，用于描述该信号。Fisher 编码 [4] 是在特征空间中建立高斯混合模型，利用该模型来描述局部特征的分布，求解此模型概率密度对均值、方差的偏导，并以此作为局部特征编码表达。在描述局部特征分布的表达方面，与传统的利用直方图相比，由于高斯混合模型是高维的，因此具有更好的表达效果。

基于重构的编码 利用一小部分视觉单词对每一个局部特征进行重构。这个过程通过建立一个带约束的最小二乘优化问题来完成。不同的约束条件对应不同的编码方法，如稀疏编码 [5] 和局部约束线性编码 [6] 等。自从将稀疏编码应用于图像分类，基于重构的编码就成为局部特征非监督表达学习的研究热点。此外，学者们还相继提出了局部协同、拉普拉斯稀疏、混合稀疏、分层稀疏以及弱监督稀疏等编码，所有这些编码方式都是通过替换约束条件来解决不同的问题。

基于显著性的编码 [7] 通过显著度对每个局部特征进行编码，显著度是指局部特征到其最近几个视觉单词距离的差或比。

局部特征编码演变

各种局部特征编码方法之间的关系及演变过程主要体现在以下几个方面。

1. 基于投票的编码和 Fisher 编码都着眼于描述整个特征空间，它们的主要区别是描述局部特征概率密度分布的方式不同。在基于投票的编码方法中，直方图上的每一个值对应一个视觉单词参与局部特征编码的频率信息。由于视觉单词和局部特征也位于高维特征空间中，因此若一个视觉单词只用一个或少数几个值来表达，则很可能会忽略其他有用的信息。在 Fisher 编码方式中，混合高斯模型中每个高斯对应一个视觉单词（局部特征的聚类中心），包含更丰富的信息。需要指出的是，基于投票的编码可看做 Fisher 编码的一种简化形式，具体推导过程可以参见文献 [2]。

2. 基于重构的编码方式比基于投票的编码方式能够更加准确地描述一个局部特征，原因在于其目标被定义为用视觉单词来重构局部特征。基于重构的编码采用基于最小二乘的优化生成局部特征表达。由于用来重构的视觉单词数量一般少于局部特征的

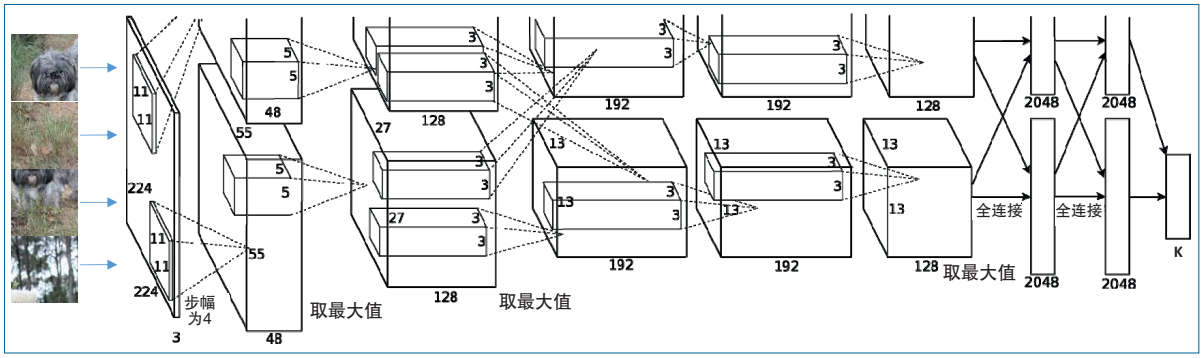

图3 基于卷积神经网络的目标局部特征表达学习

维度，基于最小二乘的重构通常是一个欠定问题。因此，在重构时基于最小二乘的优化就会不可避免地产生重构误差。尽管如此，在一些数据库中，该类编码方法仍然能够取得较好的性能，依靠的是在与取最大值汇聚操作相结合时所表现出的显著性表达。以局部约束线性编码[6]为例，在重构局部特征时，每一个视觉单词都会被多次使用，因此会产生多个响应。但是在取最大值操作中，只会保留最大值响应。当一个视觉单词获得一个非常强的响应时，则这个视觉单词就可以独立地描述这一局部特征（显著性表达）。当所有响应在描述局部特征方面都表现很弱时（非显著性表达），若单个视觉单词上的响应不能用来独立地表达局部特征，就要使用所有相关的视觉单词来描述该局部特征。在这种情况下，响应是不稳定的，因为在后续的取最大值操作中，弱响应很可能会被抑制掉。因此显著性表达是一种稳定的描述，具体可参见文献[7]。

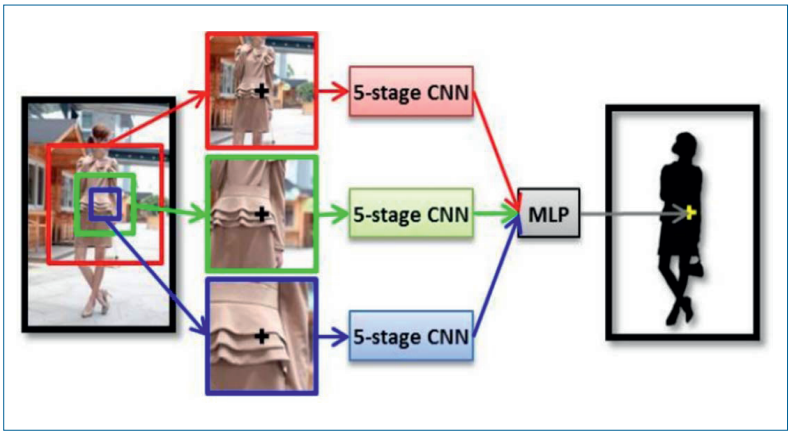

图4 基于深度学习的图像分割[10]

3．基于显著性的编码方式是在理解显著性表达的重要性以及基于最小二乘的重构方法存在的问题之后提出的。这种编码方式是根据最邻近的视觉单词与其他视觉单词之间的差异来获取显著性表达的，差异越大，则表达越显著。与基于重构的编码方式相比，基于显著性的编码方式具有如下优势：(1) 可直接从显著性定义中推导得到，不存在基于最小二乘重构中的欠定问题；(2) 易于执行，无须迭代优化，速度很快。

限于篇幅，有关各种局部特征编码方法的研究背景和数学形式的介绍可参阅文献[2]。

局部特征的监督表达学习

近年来，监督的表达学习特别是监督的深度学习在图像分类、目标检测等多种视觉任务中展现出优越的性能。

在面临同样复杂任务的情况下，深层模型比浅层模型使用更少的参数（以指数形式减少），这意味着深层描述可以更有效地对数据进行类内不变性和类间判别性局部特征的学习与表达。由

于这一良好的性质，机器学习领域出现了对数据局部特征进行多层次抽象化的学习和表达的热潮，这些研究被称为“深度学习”。在计算机视觉方面，深度学习最具代表性的模型是卷积神经网络(Convolutional Neural Networks, CNN) [8, 9]，是近年来一系列国际竞赛中取得最好成绩的方法。卷积神经网络与其他传统监督学习方法的重要区别是不再需要人工设计局部特征，而是在层叠的卷积和下使采样操作自动提取具有平移不变性的视觉局部特征。此外，为了克服深度神经网络易于过拟合的不足，研究人员相继提出了 Drop-out[8] 等策略，用于增强学习能力。

鉴于卷积神经网络在视觉计算中的突出表现，我们将其应用于对目标局部特征进行监督的表达学习。如图 3 所示，每一个局部特征都可以输入到一个卷积神经网络中，并在该网络的顶层用一个向量作为局部特征的监督反馈信息。该向量的长度 K 表示局部特征有 K 个类别，每一个局部特征可以被标注为其中的一个或多个类别。

如果能够对以每一个像素为中心的图像块（局部特征）进行较好的监督学习，就可以获得目标分割的预期效果，其基本原理如图 4 所示。首先，对一幅图像上的每一个像素点分别提取多个尺度窗口；其次，将每个尺度下窗口分别输入到一个多层卷积神经网络，通过学习得到各尺度下的局部特征表达；第三，采用一个三层感知器来预测以该像素为中心的局部特征的标签（前景还是背景）。在三个尺度窗口中，

图5　基于局部特征深度学习的目标分割算法在百度人形分割图像测试集上的结果（每幅原始彩色图像右侧的两幅图分别为真实分割结果和我们算法的分割结果）[10]

最小尺度窗口用于描述局部细节，最大尺度窗口用于获取目标与场景的上下文关系。

我们通过深度学习挖掘了目标局部特征与所属目标的关系[10]，在只有约5000幅图片的训练数据库上完成了大规模卷积神经网络的模型训练，达到了接近87%的分割精度。在2013年中国云·移动互联网创新大奖赛中，我们的模型在分割效果上不仅超过国际竞赛PASCAL VOC[2] 2012人形分割的最好成绩，也远领先于其他参赛选手，最终获得了大赛冠军[3]。我们的分割结果（见图5）表明，基于局部特征深度表达学习的目标分割算法可以克服复杂背景、尺度变化、姿态变化等视觉任务经常遇到的挑战，而在传统算法中解决这些任务是比较困难的。

总结与展望

综上所述，我们认为卷积神经网络中的卷积操作是一种带参数的非监督局部特征编码操作，且参数是从数据中学习到的。卷积操作反映了滤波器与图像中每一个局部区域之间的相似性，因此滤波器和局部区域可分别被看做视觉单词和局部特征。不难发现卷积操作与软投票是非常相似的。对局部特征编码的深入研究有可能启发我们设计出新的操作来代替卷积，开发出性能更好的深度模型。反之，深度学习取得的巨大成功也有可能为我们进一步研究BoF模型带来启发，使之在局部特征表达学习中获得更大的感受野(receptive field)，还能够保持有效的区分力，从而开发出更好的局部特征编码算法。■

黄永祯

CCF会员。中国科学院自动化研究所副研究员。主要研究方向为计算机视觉与模式识别。yzhuang@nlpr.ia.ac.cn

王　亮

CCF高级会员。中国科学院自动化研究所研究员。主要研究方向为计算机视觉、模式识别和数据挖掘。wangliang@nlpr.ia.ac.cn

参考文献

[1] G. Csurka, C. Bray, C. Dance, and L. Fan, Visual categorization with bags of keypoints, ECCV, 2004.

[2] Y. Huang, Z. Wu, L. Wang, T. Tan, Feature coding in image classification: a comprehensive study, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 36(3), 493~506.

[3] J. Gemert, J. Geusebroek, C. Veenman, and A. Smeulders, Kernel codebooks for scene categorization, ECCV, 2008.

[4] F. Perronnin and C. Dance, Fisher kernels on visual vocabularies for image categorization, CVPR, 2007.

[5] J. Yang, K. Yu, Y. Gong, and T. Huang, Linear spatial pyramid matching using sparse coding for image classification, CVPR, 2009.

[6] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, Locality-constrained linear coding for image classification, CVPR, 2010.

[7] Y. Huang, K. Huang, Y. Yu, and T. Tan, Salient coding for image classification, CVPR, 2011.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton, ImageNet classification with deep convolutional neural networks, in NIPS, 2012.

[9] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989.

[10]Z. Wu, Y. Huang, L. Wang and T. Tan, Early hierarchical contexts learned by convolutional networks for image segmentation, International Conference on Pattern Recognition (ICPR), 2014.

[2] 国际计算机视觉算法竞赛，Pattern Analysis, Statistical Modelling and Computational Learning Visual Object Classes。

[3] 2013年中国云·移动互联网创新大奖赛设置了一项特别大奖：如果采用深度学习技术实现的人形图像分割超过PASCAL VOC 2012年目标分割竞赛的最好成绩，则被授予百度公司赞助的20万元奖金。