基于深度学习中间层卷积特征的图像标注

于 宁[1]， 宋海玉[1]， 孙东洋[2]， 王鹏杰[1]， 姚金鑫[1]

(1. 大连民族大学计算机科学与工程学院，辽宁 大连 116600；
2. 安迅达盛医疗科技有限公司，北京 100020)

摘 要：针对基于深度特征的图像标注模型训练复杂、时空开销大的不足，提出一种由深度学习中间层特征表示图像视觉特征、由正例样本均值向量表示语义概念的图像标注方法。首先，通过预训练深度学习模型的中间层直接输出卷积结果作为低层视觉特征，并采用稀疏编码方式表示图像；然后，采用正例均值向量法为每个文本词汇构造视觉特征向量，从而构造出文本词汇的视觉特征向量库；最后，计算测试图像与所有文本词汇的视觉特征向量相似度，并取相似度最大的若干词汇作为标注词。多个数据集上的实验证明了所提出方法的有效性，就 *F*1 值而言，该方法在 IAPR TC-12 数据集上的标注性能比采用端到端深度特征的 2PKNN 和 JEC 分别提高 32%和 60%。

关 键 词：深度学习；图像标注；卷积；正例均值向量；特征向量
中图分类号：TP 391 **DOI：**10.11996/JG.j.2095-302X.2019050872
文献标识码：A **文 章 编 号：**2095-302X(2019)05-0872-06

Image Annotation Based on Middle-Layer Convolution Features of Deep Learning

YU Ning[1], SONG Hai-yu[1], SUN Dong-yang[2], WANG Peng-jie[1], YAO Jin-xin[1]

(1. College of Computer Science and Engineering, Dalian Nationalities University, Dalian Liaoning 116600, China;
2. Anxundasheng Medical Technology Company, Beijing 100020, China)

Abstract: Image annotation based on deep features always requires complex model training and huge space-time cost. To overcome these shortcomings, an efficient and effective approach was proposed, whose visual feature was described by middle-level features of deep learning and semantic concept was represented by mean vector of positive samples. Firstly, the convolution result is directly outputted as the low-level visual feature by the middle layer of the pre-training deep learning model, and the sparse coding method was used to represent image. Then, visual feature vector was constructed for each textual word by the mean vector method of positive samples, and the visual feature vector database of the text vocabulary was constructed. Finally, the similarities of visual feature vectors between test image and all textual words were computed, and some words with largest similarities were selected as annotation words. The experimental results on several datasets demonstrate the effectiveness of the proposed method. In terms of *F*1-measure, the experimental results on IAPR TC-12 dataset show that the performance of the proposed method was improved by 32% and 60% respectively, compared to 2PKNN and JEC with end-to-end deep features.

收稿日期：2019-07-31；**定稿日期：**2019-08-22
基金项目：国家自然科学基金项目(61300089)；辽宁省自然科学基金项目(201602199，2019-ZD-0182)；辽宁省高等学校创新人才支持计划项目(LR2016071)
第一作者：于 宁(1995–)，女，内蒙古呼伦贝尔人，硕士研究生。主要研究方向为图像理解、机器学习等。E-mail：877213412@qq.com
通信作者：宋海玉(1971–)，男，河南安阳人，副教授，博士，硕士生导师。主要研究方向为图像理解、计算机视觉等。E-mail：shy@dlnu.edu.cn

Keywords: deep learning; image annotation; convolution; mean vector of positive sample; feature vector

近 20 年来，自动图像标注一直是计算机视觉领域研究热点。图像标注的效果主要取决于标注模型和视觉特征向量，其中，视觉特征向量质量决定着图像标注质量的上限。近年来，随着图像标注模型越来越成熟，视觉特征向量已经成为图像标注效果的决定性因素。2012 年之前，计算机视觉领域的图像特征均为领域专家设计的人工特征，人工特征质量主要取决于专家的领域知识和经验，在复杂的计算机视觉应用中，人工特征很难提取出高质量的特征向量[1]。

2012 年，Alex 和 Hinton 构建了基于卷积神经网络的深度学习模型 AlexNet[2]，在 ImageNet 图像分类比赛中以压倒性的优势夺取了冠军，自此开启了深度学习时代。此后研究者们在 AlexNet 的基础上提出了许多优秀的网络模型，如：VGG[3]，GoogleNet[4]，ResNet[5]等。深度学习特征是一个端到端的特征提取过程，不需要人工参与。对于图像特征提取而言，输入端输入原始图像，通过复杂的模型训练后，在输出端直接输出高质量的深度特征。基于高质量的端到端特征向量，深度学习在图像分类领域取得了突破性进展，并被尝试应用于很多领域[1]。

深度学习之所以可以取得突破性进展，很大程度上取决于其复杂的网络结构。为了取得较好的效果，深度学习网络结构设计越来越复杂，层数越来越深。然而数以百万计、千万计参数的模型训练不仅需要海量的训练样本支撑，而且需要巨大的时间开销以及较高的硬件配置，这些因素限制了深度学习的应用，例如 2012 年所提出的 AlexNet 网络模型共有 6 100 万参数[2]，此后提出的 VGG-16 模型参数有 1.38 亿[3]。在训练样本充足的情况下，模型训练充分，复杂的深度学习可以取得预期效果。但事实上，大多数应用很难提供充足的训练样本，往往会造成模型过拟合等，从而使得模型训练质量较差。针对以上缺点，相关学者提出了一些解决方法，比如基于预训练模型进行微调训练并应用于复杂的标注算法等等。尽管取得了较好的标注效果，但是依然没有提取适合于图像标注的高质量的深度特征。

1　相关工作

2014 年，Caffe 的设计者贾扬清团队率先将深度学习应用于图像标注[6]，此后，越来越多的学者基于深度学习技术开展图像标注的研究。表 1 为基于深度学习的图像标注模型与传统标注模型的效果对比，实验数据集均为 Corel5K。

表 1　基于深度学习的标注模型与传统标注模型的效果对比(Corel5k)

标注模型	特征向量	查准率 (%)	查全率 (%)	*F*1 值 (%)	N+
MBRM[7]	人工特征	24.00	25.00	24.88	122
JEC[8]	多种人工特征	27.00	32.00	29.28	139
文献[9]	深度学习特征	29.28	33.00	31.03	146
AlexNet	AlexNet 端到端特征	25.80	29.33	27.45	130
VGG-16	VGG-16 端到端特征	0.21	1.92	0.38	5

由表 1 可以看出，深度学习方法相对于传统的图像标注方法，效果虽然有所提升但是并不显著，特别是当网络模型较复杂时(例如 VGG-16 网络)，标注性能甚至会降低。主要原因是小规模图像数据无法满足复杂网络结构模型的训练。如果没有足够训练数据支撑，由于过拟合等原因，复杂的网络模型并不能取得理想的标注效果，甚至越是复杂的网络结构的深度学习模型其标注性能越差。同时该深度学习训练方法需要巨大的时间开销以及较高的硬件配置。针对以上不足，OQUAB 等[10]提出了迁移学习的方法。将迁移学习应用于图像标注的效果有了较好地提升，但是训练过程时空开销仍然较大并且需要较高的硬件配置。

尽管理论上很多深度学习模型有很好的理论基础和科学的网络结构，但如果没有足够训练数据支撑，复杂的网络模型并不能取得理想的标注效果，同时深度模型训练所需的巨大的时空开销和较高的硬件配置使得其应用受到限制。因此相关学者将研究重心转向更为复杂的标注算法或更优质的特征表示，例如将深度特征与其他特征融合进行图像标注。

目前，围绕深度学习开展图像标注的研究主要分 3 类：①设计新的网络结构或改进已有模型，通过修改损失函数、输出类别个数等方法，使用目标数据集进行训练，使得原有模型适合图像标注[11-12]。②基于预训练模型的微调型，仅修改在较大数据集上已训练好的网络模型的全连接层和输出层使其适应目标数据集，而其他网络结构和参数保持不变，即在现有网络权重的基础上对网

络进行微调，如文献[10]提出的迁移学习方法等。③直接使用预训练模型完成目标数据集的特征提取，然后再使用其他复杂算法完成标注，或把深度特征与其他特征融合进行图像标注；文献[13]提出图像标注模型 CMRM 和 MBRM，将深度学习特征应用于 JEC、2PKNN 等复杂标注算法，并取得了较好的标注效果，主要原因是后续的复杂标注模型或多种特征融合。依然没有解决在数据集较小的情况下，如何提取适合于图像标注的高质量深度特征，以及如何设计高效的标注模型。为此，本文试图在资源受限、数据量不足的情况下，提取高质量的深度特征，并提出有针对性的图像标注模型。

2 改进的深度特征提取方法

目前，深度学习模型所提取的端到端特征可以视为是图像的全局特征，该方法在图像分类领域取得极大成功，但在图像标注领域并未取得显著成果。在基于深度学习的图像分类中，将仅使用模型末端输出层的全局特征，而忽略中间层的输出特征，称为端到端模型。然而深度学习模型在对图像特征进行层层抽象时，每层都有其自身价值，因感受野不同，提取的特征所描述范围不同，网络模型末端的全连接层刻画的是图像的全局视觉特征，而中间层刻画的是图像的区域或局部特征。深度学习中间层卷积核感受野小，但个数多，中间层卷积核更容易捕获局部或区域特征，因此，中间层特征更善于刻画多对象或复杂图像中的对象特征。而且，直接提取中间层特征可以避免深度学习全连接层较高的时空开销。本文提取了深度学习的中间卷积层特征，通过稀疏编码的方式生成图像的特征向量。特征生成过程如下：

(1) 提取预训练深度学习模型的中间层输出特征，即 $\boldsymbol{F}\in\boldsymbol{R}^{(K\times W\times H)}$，其中 K 为特征图的个数，W 和 H 分别为特征图的宽和高。然后对特征进行规格化并转换为二维特征矩阵，表示为 $\boldsymbol{F}_{(W\times H,\ K)}$。

(2) 对原始特征进行高斯规格化处理，并将数据应用主成分分析(principal component analysis, PCA)进行约减。此时卷积特征用 $\boldsymbol{F}_{\boldsymbol{new}(W\times H,\ n)}$表示，其中 n 代表约减后的维度。

(3) 将降维后的数据进行 K-means 聚类，构造 m 个视觉词汇。根据视觉词袋原理，每幅图像表示为 m 维的词袋向量。

(4) 利用获取到的聚类中心点将卷积特征进行(vector of locally aggregated descriptors, VLAD)编码[14]转换为图像的视觉特征向量 $\boldsymbol{Fv}$，即

$$\boldsymbol{F}_V=\left[\sum_{NN(\boldsymbol{f}_i)=\boldsymbol{C}_1}\boldsymbol{f}_i-\boldsymbol{C}_1,\cdots,\sum_{NN(\boldsymbol{f}_i)=\boldsymbol{C}_m}\boldsymbol{f}_i-\boldsymbol{C}_m\right] \quad (1)$$

其中，$\boldsymbol{f}_i$ 为图块特征；$[\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_m]$为聚类中心点；$NN(\boldsymbol{f}_i)$为离 $\boldsymbol{f}_i$ 最近的聚类中心。

3 图像标注方法

人工特征向量往往是底层视觉的统计量等，其视觉模式不显著，语义级别较低，因此，基于人工特征向量的图像分类/标注模型往往较抽象、算法较复杂、时空代价较大。与传统的人工特征相比，深度学习中间层特征视觉模式显著、语义级别较高，采用视觉词典等稀疏编码方式后，其视觉和语义刻画能力较强。如果能为每个文本词汇构造视觉特征向量，那么，传统的图像标注问题中计算词汇隶属于图像的置信度问题就转换为计算 2 个视觉特征向量(文本词汇视觉特征向量和图像视觉特征向量)的相似度问题了。基于此种文本词汇视觉特征向量标注方法的时空开销非常小，且与训练数据集规模无关。与传统方法相比，更能够胜任处理大规模数据集。

本文提出的图像标注方法的系统结构如图 1 所示。在训练阶段，提取所有训练图像的深度特征并形成 VLAD 向量，从而构成图像视觉特征图库，采用正例样本均值向量法为每个文本词汇构造能代表其最本质视觉信息的视觉特征向量，从而构成包含所有词汇对应特征的正例均值向量词库。在标注(测试)阶段，在线提取该测试图像的特征向量并生成其 VLAD 向量，测试图像的 VLAD 特征向量逐一与正例均值向量词库中各个词汇的正例均值向量计算视觉相似度，最终，排序选择相似度最大特征向量所对应的文本语义词汇作为该测试图像的标注词汇。

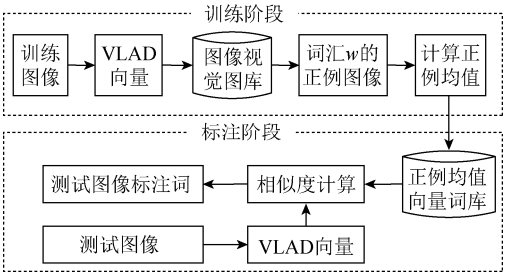

图 1 图像标注的系统结构图

在传统的视觉词典表示方法中，若有 M 个视觉词汇，相当于在视觉词典中构造一个 M 维的视觉特征空间，每个图像都是该空间中的一个对象，因此可以由 M 个基底特征向量线性表示。从语义角度看，每幅图像可以视为若干个文本概念的组合，如果每个文本概念 w 均可以表示为视觉特征空间中的特征向量 $\boldsymbol{w}$，那么任一图像 I 的视觉特征向量 $\boldsymbol{I}$ 可以视为若干个文本语义词汇对应视觉特征向量的线性和，即

$$\boldsymbol{I}=a_1\boldsymbol{w}_1+\cdots+a_j\boldsymbol{w}_j+\cdots+a_M\boldsymbol{w}_M \tag{2}$$

其中，系数 a_j 为布尔型，若图像中有对应词汇则为 1，否则为 0；$\boldsymbol{w}_j$ 为词汇 w_j 的视觉特征向量。

当已知图像特征向量以及其所包含的词汇信息时，可以根据矩阵知识求出每个文本词汇的视觉向量。但该方程组求解存在如下困难：①理想情况下，所有语义对象的特征向量都是线性无关的，其可以作为该语义空间的基底向量，但事实上，不同概念之间会有相关视觉模式，因此，这一假设很难严格成立；②大多数图像数据集词汇分布不均衡，一些低频词汇对应的图像个数远低于向量维数；③当特征向量维数较高时，求解的时空复杂度过高。因此，很难采用传统的矩阵方法或机器学习方法求解。

针对深度学习中间层特征的区域或局部调整描述能力强、区分度大，且具有一定语义刻画能力，本文提出一种基于正例样本均值向量的快速标注方法。尽管无法直接对方程式求解，但针对深度学习中间层特征的特点，任一文本词汇的特征向量可由包含该词汇的所有图像特征向量的均值近似表示。以词汇 w_j 为例，若有 N 幅图像包含这个词汇，则 N 幅图像均由语义概念特征向量表示，即方程为

$$\begin{cases}\boldsymbol{I}_1=a_{1,1}\boldsymbol{w}_1+\cdots+a_{1,j}\boldsymbol{w}_j+\cdots+a_{1,M}\boldsymbol{w}_M\\ \boldsymbol{I}_2=a_{2,1}\boldsymbol{w}_1+\cdots+a_{2,j}\boldsymbol{w}_j+\cdots+a_{2,M}\boldsymbol{w}_M\\ \vdots\\ \boldsymbol{I}_N=a_{N,1}\boldsymbol{w}_1+\cdots+a_{N,j}\boldsymbol{w}_j+\cdots+a_{N,M}\boldsymbol{w}_M\end{cases} \tag{3}$$

本文提出词汇 w_j 的视觉特征向量 $\boldsymbol{w}_j$ 可由包含该词汇正例样本均值向量 $\boldsymbol{v}_j$ 近似表示，即

$$\boldsymbol{w}_j\approx\boldsymbol{v}_j=\frac{1}{N}\sum_{I\in\boldsymbol{S}_j}\boldsymbol{I} \tag{4}$$

其中，$\boldsymbol{I}$ 为图像 I 的特征向量；$\boldsymbol{S}_j$ 为词汇 w_j 的正例图像集合，正例图像集合是指包含该词汇的所有图像的集合；N 为正例图像个数。将式(3)代入式(4)，得到正例样本均值向量 $\boldsymbol{v}_j$，即

$$\boldsymbol{v}_j=\frac{\sum_{i=1}^{N}a_{i,1}}{N}\boldsymbol{w}_1+\cdots+\frac{\sum_{i=1}^{N}a_{i,j}}{N}\boldsymbol{w}_j+\cdots+\frac{\sum_{i=1}^{N}a_{i,M}}{N}\boldsymbol{w}_M \tag{5}$$

数据集规模越大，则包含词汇的图像子集规模也越大，式(5)计算出的 $\boldsymbol{v}_j$ 越接近于词汇 w_j 的向量 $\boldsymbol{w}_j$。因为当图像规模增大时，第 j 个向量 $\boldsymbol{w}_j$ 的系数 $\frac{1}{N}\sum_{i=1}^{N}a_{i,j}$ 越接近于 1，而其他向量的系数 $\frac{1}{N}\sum_{i=1,x\neq j}^{N}a_{i,x}$ 越来越接近于 0。因此，数据集规模越大，正例均值向量法所构造的词汇视觉特征向量 $\boldsymbol{v}_j$ 越接近于词汇 w_j 的真实特征向量 $\boldsymbol{w}_j$。根据式(4)生成所有关键词的正例均值向量，实现文本语义概念向视觉词汇向量表示的转换，可以构造出文本词汇的视觉特征向量库。

基于正例均值向量的标注过程是，词库中所有词汇的视觉特征向量均与待标注图像的视觉特征向量计算相似度，取视觉相似度最大的若干词汇作为图像的标注词。相似度距离采用 L2 距离，即

$$sim(I,w_j)=Dist_{L2}(\boldsymbol{I},\boldsymbol{w}_j) \tag{6}$$

4　实验结果与分析

为了准确、客观地比较并评价本文方法的各项性能，实验数据集采用经典数据集 Corel5k 和 IAPR TC-12，这 2 个数据集是图像标注领域最常用的实验数据集，已经成为事实上的标准数据集[1]。Corel5k 数据集规模较小，包括 4 500 幅训练图像、500 幅测试图像，共包含 260 个语义概念；IAPR TC-12 数据集规模较大，共有 19 623 幅图像，其中训练图像 17 663 幅，测试图像 1 960 幅，数据集内共包含 291 个语义概念。实验平台为 64 位 Windows7 操作系统，硬件配置为 3.60 GHz 的 i7-3790 CPU、NVIDA GeForce GTX 1080 的显卡、28 GB 内存，软件环境为 Matlab2016a。

性能评价指标采用图像标注领域最广泛的查准率(precision)、查全率(recall)、$F1$ 值($F1$-score)和 N+[1]。给定词汇 w_j 的查准率、查全率和 $F1$ 值的计算式为

$$precision=\frac{r}{r+w} \tag{7}$$

$$recall=\frac{r}{n} \tag{8}$$

$$F1=\frac{2\times precision\times recall}{precision+recall} \tag{9}$$

其中，r 为正确标注词汇 w_j 的图像个数；w 为错误标注词汇 w_j 的图像个数；n 为数据集中含有标注词 w_j 的图像个数。对数据集中所有词汇分别计算查准率、查全率和 $F1$，最后计算所有词汇查准率和查全率的平均值作为系统整体评价。N+为标注结果中所能标注出的词汇个数，即标注词出现次数大于零的词汇个数，作为正确标注词汇多样性的评价指标。各项指标数值越高标注性能越好，理论上，查准率、查全率和 $F1$ 性能上限可以接近于 100%，N+可以接近于数据集中词汇个数。

为了客观评价本文所提取的深度学习中间卷积层特征的性能，将实验的结果与使用传统人工特征经典图像标注模型(MBRM 模型、JEC 模型)以及将深度学习特征应用于 JEC、2PKNN 等复杂标注算法[13]的结果进行了比较。与文献[13]中深度学习网络模型相同，本文采用 VGG-16 网络，根据网络结构及卷积核等信息，选用 Conv5-2 层数据作为图像的局部特征信息。预训练数据集为 ILSVRC-2012[15]。在较小规模数据集 Corel5k 和较大规模数据集 IAPR TC-12 上完成的实验结果分别见表 2 和表 3。

表 2　本文方法与其他方法实验结果性能对比(Corel5k)

标注模型	特征向量	查准率 (%)	查全率 (%)	$F1$ 值 (%)	N+
MBRM[7]	人工特征	24.00	25.00	24.88	122
JEC[8]	多种人工特征	27.00	32.00	29.28	139
JEC[13]	VGG-16 端到端特征	31.00	32.00	32.00	141
2PKNN[13]	VGG-16 端到端特征	33.00	30.00	32.00	160
本文方法	VGG-16 中间层特征	26.93	41.43	32.64	161

表 3　本文方法与其他方法实验结果性能对比(IAPR TC-12)

标注模型	特征向量	查准率 (%)	查全率 (%)	$F1$ 值 (%)	N+
MBRM[7]	人工特征	24.00	23.00	23.48	223
JEC[8]	多种人工特征	28.00	29.00	28.49	250
JEC[13]	VGG-16 端到端特征	28.00	21.00	24.00	237
2PKNN[13]	VGG-16 端到端特征	38.00	23.00	29.00	261
本文方法	VGG-16 中间层特征	46.15	32.80	38.35	258

表 2 和表 3 实验结果表明，无论是在较小规模数据集 Corel5k，还是在较大数据集 IAPR TC-12 中，本文所提出方法主要性能指标不仅优于采用人工特征的标注模型 MBRM 和 JEC，也优于使用深度学习端到端特征的标注模型 JEC 和 2PKNN。表 2 实验结果表明本文所提出方法标注效果略优于其他方法，主要是因为 Corel5k 数据集中语义词汇出现频次很不均衡，最高频次和最低频次的词汇出现次数分别是 1 004 次和 1 次。而表 3 表明在 IAPR TC-12 中本文所提出方法除 N+略低于 2PKNN 外，其他指标明显优于其他方法，综合评价指标 $F1$ 值分别比 MBRM、JEC (人工特征)、JEC (深度特征)、2PKNN (深度特征)提升了 63%、35%、60%、32%。这印证了，式(4)和式(5)中的推理，图像数据集规模越大，所构造的 $\boldsymbol{v}_j$ 越接近于目标词汇 w_j 的特征向量 $\boldsymbol{w}_j$。因此，采用该方法所完成的图像标注效果越好。

由于 2PKNN 和 JEC 涉及的多种人工特征模型复杂度较高，现有相关资料中均未给出这些特征的时间开销，且模型较复杂，实验环节未能在同一环境下完成对比实验，因此表 2 和表 3 仅对比了这些算法的标注效果，而没有列出时间开销。但理论上，此类算法的时间复杂度、空间复杂度均大于本文所提出的方法。本文所提出方法无需训练深度学习模型，Corel5k 测试图像数据集特征提取时间为 55 s，而传统的端到端的深度学习微调方法模型训练时间是 8 h，测试图像数据特征提取时间为 70 s。在较大规模数据集 IAPR TC-12 中，本文方法测试图像特征提取时间为 330 s，而传统的端到端的深度学习微调方法模型训练时间是 10 h，测试图像数据特征提取时间为 360 s。若测试图像个数为 T、训练图像个数为 N、数据集中所包含词汇个数为 W，JEC 和 2PKNN 时间复杂度为 O(TN)，而本文方法时间复杂度为 O(TW)，由于数据集中训练图像个数 N 远大于词汇个数 W，因此，标注阶段本文所提出方法的时间开销也明显低于 JEC 和 2PKNN，远低于 MBRM 等标注模型。

本文所采用的 VGG-16 模型预训练所用的 ImageNet 数据集，以及图像标注领域最常用的图像数据集 Corel5k 和 IAPR TC12 均为自然场景领域图像，为了验证所提出方法对领域迁移的适应性，在 ESP Game 数据集上完成了图像标注实验。ESP Game 数据集是双人游戏图像数据，与自然场景数据集 ImageNet 属于完全不同的领域。该数据集共有 20 770 幅图像，其中训练图像 18 689 幅，测试图像 2 981 幅，数据集内共包含 268 个语义概念。相同实验方法下，在 ESP Game 数据集上完成的实验结果见表 4。实验结果表明，在其他领域的图像数据集上，本文方法的标注性能也优于其他方法，说明本文方法对领域迁移有较强的

适应性。

表 4　本文方法与其他方法实验结果性能对比(ESP Game)

标注模型	特征向量	查准率 (%)	查全率 (%)	*F*1 值 (%)	N+
MBRM[7]	人工特征	18.00	19.00	18.50	209
JEC[8]	多种人工特征	22.00	25.00	23.40	224
JEC[13]	VGG-16 端到端特征	26.00	22.00	23.83	234
2PKNN[13]	VGG-16 端到端特征	40.00	23.00	29.20	250
本文方法	VGG-16 中间层特征	43.74	33.08	37.67	258

5　结 束 语

深度学习是近年来的研究热点，但模型训练所要求的数据门槛和系统配置都比较高，制约了深度学习的应用。本文根据深度学习模型中间层视觉特征模式的通用性，采用提取深度学习中间层卷积特征的方法，并在此基础之上提出了基于正例均值的图像标注方法。与传统的依赖于大规模数据模型训练的端到端深度特征相比，本文所采用的基于深度学习中间层卷积特征提取方法，无需大规模数据集训练模型，降低了深度特征的数据和硬件门槛、扩大了深度学习应用范围；所提出的标注方法时空开销较小，更适合于大规模数据集的处理和在线标注。此外，由于测试图像的最终标注词汇主要取决于文本词汇的视觉特征向量，而不是训练图像的特征向量，所以，本文所提出的方法也有助于缓解训练数据类别不均衡的难题。

参 考 文 献

[1] CHENG Q M, ZHANG Q, FU P, et al. A survey and analysis on automatic image annotation [J]. Pattern Recognition, 2018, 79: 242-259.

[2] 张顺, 龚怡宏, 王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用[J]. 计算机学报, 2019, 42(3): 453-482.

[3] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10). [2018-09-11]. https://arxiv.org/pdf/1409.1556.pdf.

[4] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C]//Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2015: 1-9.

[5] HE K, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2016: 770-778.

[6] GONG Y C, JIA Y Q, LEUNG T K, et al. Deep convolutional ranking for multilabel image annotation [EB/OL]. (2014-04-14). [2018-11-14]. https://arxiv. org/pdf/1312.4894v2.pdf.

[7] FENG S L, MANMATHA R, LAVRENKO V. Multiple Bernoulli relevance models for image and video annotation [C]//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2004: 1002-1009.

[8] MAKADIA A, PAVLOVIC V, KUMAR S. A new baseline for image annotation [C]//Proceedings of the 10th European Conference on Computer Vision. Heidelberg: Springer, 2008: 316-329.

[9] 罗世操. 基于深度学习的图像语义提取与图像检索技术研究[D]. 上海: 东华大学, 2016: 55.

[10] OQUAB M, BOTTOU L, LAPTEV I, et al. Learning and transferring mid-level image representations using convolutional neural networks [C]//Proceedings of the 2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2014: 1717-1724.

[11] JOHNSON J, BALLAN L, LI F F. Love thy neighbors: Image annotation by exploiting image metadata [C]// Proceedings of the 2015 IEEE Computer Society International Conference on Computer Vision (ICCV). New York: IEEE Press, 2015: 4624-4632.

[12] 黎健成, 袁春, 宋友. 基于卷积神经网络的多标签图像自动标注[J]. 计算机科学, 2016, 43(7): 41-45.

[13] MURTHY V N, MAJI S, MANMATHA R. Automatic image annotation using deep learning representations [C]// Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. New York: ACM Press, 2015: 603-606.

[14] NG J Y H, YANG F, DAVIS L S. Exploiting local features from deep networks for image retrieval [C]// Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVRPW). New York: IEEE Press, 2015: 53-61.

[15] 田萱, 王亮, 丁琪. 基于深度学习的图像语义分割方法综述[J]. 软件学报, 2019, 30(2): 440-468.