控制与决策
Control and Decision

基于自适应多尺度图卷积网络的多标签图像识别

王雪松, 荣小龙, 程玉虎, 陈正升

引用本文:
王雪松,荣小龙,程玉虎,陈正升. 基于自适应多尺度图卷积网络的多标签图像识别[J]. 控制与决策, 2022, 37(7): 1737−1744.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.0179

您可能感兴趣的其他文章

Articles you may be interested in

基于图卷积网络的行为识别方法综述

A survey of action recognition methods based on graph convolutional network

控制与决策. 2021, 36(7): 1537−1546 https://doi.org/10.13195/j.kzyjc.2020.0514

一种新的基于标签传播的复杂网络重叠社区识别算法

A novel algorithm for overlapping community detection based on label propagation in complex networks

控制与决策. 2020, 35(11): 2733−2742 https://doi.org/10.13195/j.kzyjc.2019.0176

基于改进DenseNet网络的人体姿态估计

Improved DenseNet network for human pose estimation

控制与决策. 2021, 36(5): 1206−1212 https://doi.org/10.13195/j.kzyjc.2019.1218

基于MobileNet的多目标跟踪深度学习算法

Deep learning algorithm based on MobileNet for multi−target tracking

控制与决策. 2021, 36(8): 1991−1996 https://doi.org/10.13195/j.kzyjc.2019.1424

复杂背景下全景视频运动小目标检测算法

Panoramic video motion small target detection algorithm in complex background

控制与决策. 2021, 36(1): 249−256 https://doi.org/10.13195/j.kzyjc.2019.0686

基于自适应多尺度图卷积网络的多标签图像识别

王雪松, 荣小龙, 程玉虎†, 陈正升

(1. 中国矿业大学 地下空间智能控制教育部工程研究中心，江苏 徐州 221116；
2. 中国矿业大学 信息与控制工程学院，江苏 徐州 221116)

摘 要: 利用一阶谱图卷积探索类别标签间关系是目前多标签图像识别常用的手段,但是,较多的图卷积层数易出现过度平滑现象,使得该方法存在局限性. 为此,提出一种基于自适应多尺度图卷积网络的多标签图像识别方法,主要思路为: 采用块Krylov子空间形式的谱图卷积来挖掘类别标签间的相关性,在每个图卷积层中拼接多尺度信息并扩展到深层结构,并在自适应标签关系图模块所构建的关系图上学习分类器,从而更加有效地进行多标签图像识别. 通过两个公开数据集PASCAL VOC 2007和MS-COCO 2014上的实验结果验证了所提出方法的有效性.

关键词: 自适应关系图；多尺度图卷积网络；多标签图像识别；块Krylov子空间

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0179

引用格式: 王雪松,荣小龙,程玉虎,等. 基于自适应多尺度图卷积网络的多标签图像识别[J]. 控制与决策, 2022, 37(7): 1737-1744.

Multi-label image recognition based on adaptive multi-scale graph convolutional network

WANG Xue-song, RONG Xiao-long, CHENG Yu-hu†, CHEN Zheng-sheng

(1. Engineering Research Center of Ministry of Education for Intelligent Control of Underground Space，China University of Mining and Technology，Xuzhou 221116，China；2. School of Information and Control Engineering，China University of Mining and Technology，Xuzhou 221116，China)

Abstract: Utilizing the first-order spectral graph convolution to explore the correlation between category labels is a common method for multi-label image recognition. However, more graph convolution layers are prone to over-smoothing, which causes some limitations for the method. With respect to the aforementioned problem, a multi-label image recognition method based on the adaptive multi-scale graph convolutional network is proposed. The main idea is as follows: the spectral graph convolution in the form of block Krylov subspace is employed to mine the correlation between category labels, and the multi-scale information eixsted in the convolutional layer is spliced and extended to the deep structure. At the same time, the classifier is learned on the relation graph constructed by the adaptive label relation graph module, accordingly the multi-label image recognition is performed more effectively. Experimental results on two public datasets including PASCAL VOC 2007 and MS-COCO 2014 verify the effectiveness of the proposed method.

Keywords: adaptive relation graph；multi-scale graph convolutional network；multi-label image recognition；block Krylov subspace

0 引 言

作为计算机视觉的基本任务之一,图像识别技术主要是解决如何准确预测给定图像或视频中的物体类别. 常规的图像分类任务所给定的图像中一般仅包含单个物体,该设定属于相对理想的情况,然而在类别繁多且复杂的世界中这是极不现实的,往往一个实际场景中会同时出现多个不同类别的物体. 因此,多标签图像识别问题引起了人们越来越多的关注,其研究任务是识别出给定图像中所存在的一系列物体. 近年来,得益于深度学习技术[1-3]的不断成熟,多标签图像识别技术的发展取得了长足的进步并在诸多领域有着广泛应用,如多目标分类[4]、时装属性识别[5]、人体属性识别[6]等.

目前,已有大量关于多标签图像识别的研究工

收稿日期: 2021-01-28；**录用日期:** 2021-04-21.
基金项目: 国家自然科学基金项目(61772532, 61976215).
责任编委: 张国山.
†通讯作者. E-mail: chengyuhu@163.com.

作. Wang等[7]发现递归神经网络能以顺序方式捕获高阶标签相关性并提出了CNN-RNN,通过学习联合图像标签嵌入来表征语义标签依赖关系,但忽略了语义标签与图像区域间的联系. Wei等[8]将任意数量的目标假设区域作为卷积网络的输入,并将输出结果进行最大池化进而得到多标签预测,但该方法引入了大量冗余计算. Wang等[9]利用图像区域级别的空间注意力机制来直接从特征图中定位目标区域,并利用LSTM网络在捕获区域依赖性时进行语义标记评分,但该方法仅考虑了图像级别的局部区域特征,而忽略了类别间的关系. 为此,Zhu等[10]使用语义级别的注意力机制对类别标签的相关性进行建模,同时设计了正则化网络为所有标签生成注意力图,并利用可学习的卷积捕获标签的语义与空间关系,但该方法可能对非目标位置产生无效识别. 尽管上述工作考虑了图像中区域间的关系,即局部相关性,但忽略了超越图像本身经验知识中标签的全局相关性,因此,所能利用的局部区域辅助信息是有限的.

为解决以上方法存在的问题,Chen等[11]结合图卷积网络(graph convolutional network, GCN)[12]可使图的各节点间相互传递信息的优势,提出了更具可扩展性和灵活性的ML-GCN网络. 首先,根据非图像层次的先验知识获得类别标签的向量表示,并在构建类别标签关系图时考虑物体的共现性; 然后,利用GCN使关系图的节点间信息相互传递与融合; 最后,学习出彼此关联的类别分类器. 然而ML-GCN网络中所使用的标签相关图是人工设计的并且需要仔细修改调整,同时每一个标签关系图仅适用于一种特定场景,对其他新场景都需要重新制定关系图,因此该方法不具有普适性和实用性. 为此,Li等[13]提出了A-GCN网络,该网络设有对类别标签相关性进行自动建模的自适应标签关系图模块,因此具有较强的灵活性和实用性. 尽管上述方法探索了利用图卷积网络挖掘类别标签之间的相关性,但它们学习标签相互关联的分类器时采用的是传统的一阶谱图卷积,并且在邻接矩阵中赋予各节点特征信息相同的权重. 而一阶谱图卷积可视为拉普拉斯平滑的特殊形式,同时还造成了节点特征信息中自身成分占比过低的问题. 随着网络层数的增加,图中各节点的自身特征信息被稀释,各节点输出信息趋于相同,从而导致节点间可区分性降低,即产生过度平滑现象.

为解决以上问题,本文提出一种基于自适应多尺度图卷积网络(adaptive multi-scale graph convolutional network, AMS-GCN)的多标签图像识别方法,该方法主要思路为: 1)利用自适应关系图模块自动构建类别标签关系图,同时采用块Krylov子空间形式的多尺度图卷积网络对类别相关性进行建模,从而充分利用多尺度信息以使多层次特征互补; 2)为了使模型在对各节点信息进行融合时充分考虑节点自身信息的重要性,在邻接矩阵中赋予节点自连接以更大的权重,进而解决过度平滑问题并更好地挖掘标签相关性; 3)将与类别相关联的分类器应用于卷积神经网络提取的图像特征,最终实现多标签图像识别. 为验证本文所提出AMS-GCN的有效性,采用PASCAL VOC 2007数据集[14]和MS-COCO 2014数据集[15]开展实验研究,并分析块Krylov子空间形式的多尺度图卷积网络中层数与块数对识别效果的影响.

1 基于自适应多尺度图卷积网络的多标签图像识别

图1给出了AMS-GCN的结构框图,主要分为两个部分: 第1部分为用于图像特征提取的ResNet-101

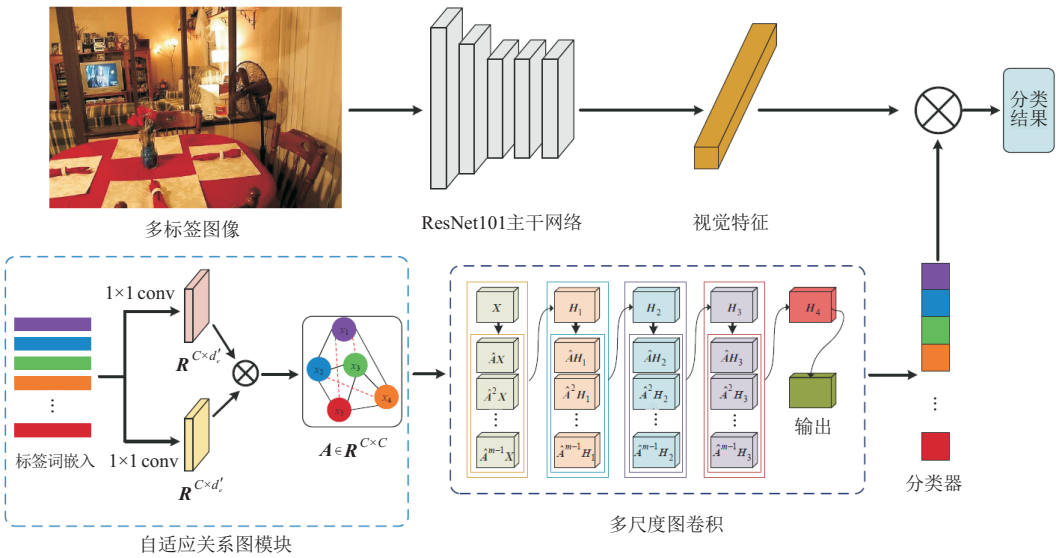


图1 AMS-GCN框图

主干网络; 第2部分为自适应模块所构建的关系图进行学习的多尺度图卷积网络.

1.1　一阶谱图卷积

图(graph)在表征结构化信息时表现出明显的优势,而且图卷积网络可使图中各节点间相互传递信息并增强节点自身的特征信息表示,因此,图卷积网络引起了人们广泛关注,相关学者已将其引入图像理解[16]、行为识别[17]和文本分类[18]等诸多领域.

对于一个给定的无向图$G=(\nu,\varepsilon,\boldsymbol{A})$,其中$\nu$、$\varepsilon$、$\boldsymbol{A}\in\boldsymbol{R}^{C\times C}$分别为图的节点集合、边集合以及对称邻接矩阵,利用GCN对输入特征$\boldsymbol{H}_l\in\boldsymbol{R}^{C\times d}$及邻接矩阵$\boldsymbol{A}\in\boldsymbol{R}^{C\times C}$进行学习,可得

$$\boldsymbol{H}_{l+1}=\delta(\hat{\boldsymbol{A}}\boldsymbol{H}_l\boldsymbol{W}_l), \tag{1}$$

$$\hat{\boldsymbol{A}}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}, \tag{2}$$

$$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}. \tag{3}$$

其中: $\boldsymbol{H}_{l+1}$为单层GCN的特征输出,$\hat{\boldsymbol{A}}$为标准化后的邻接矩阵,$\boldsymbol{W}_l\in\boldsymbol{R}^{d\times C}$为权值矩阵,$\tilde{\boldsymbol{A}}$为标准化前无向图具有自连接的邻接矩阵,$\boldsymbol{I}$为单位对角阵,$\tilde{\boldsymbol{D}}_{ii}=\sum_j\tilde{\boldsymbol{A}}_{ij}$为对角阵形式的度矩阵,$\delta(\cdot)$为非线性激活函数.

1.2　邻接矩阵

一般情况下,邻接矩阵在标准化前为$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$,表示计算图中各节点时添加了自连接. 在利用图卷积对图节点信息进行聚合时,对节点自身及其相邻近节点的特征信息赋予了相同的权重,但这造成了节点特征信息中自身成分占比过低的问题. 为此,文献[19]提出了相应的改进措施,即

$$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\alpha\boldsymbol{I}, \tag{4}$$

其中α为常数. 式(4)所示改进措施也是图信号处理过程中常用的处理方式,使得在信息聚合过程中,节点自连接时具有更大权重,从而使得节点自身信息更具有可辨识性,本文中α取值为2.

1.3　自适应标签关系图模块

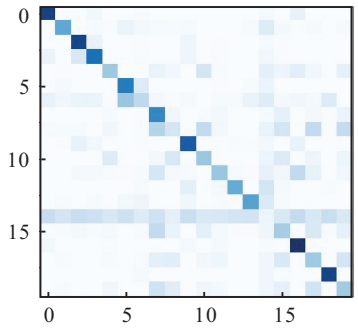
以各物体类别的标签作为图的各个节点,标签的词嵌入作为节点的特征信息,本文中词嵌入采用GloVe词向量,该模块利用具有d_e维度的C个类别标签的词嵌入自动构建标签相关矩阵. 具体过程为: 首先,利用两个卷积核为1×1的卷积层分别对词嵌入进行卷积操作; 然后,对二者进行点积运算,进而得出标签相关矩阵

$$\boldsymbol{A}=\frac{1}{C}(\boldsymbol{W}_\phi*\boldsymbol{E})^{\mathrm{T}}(\boldsymbol{W}_\theta*\boldsymbol{E}). \tag{5}$$

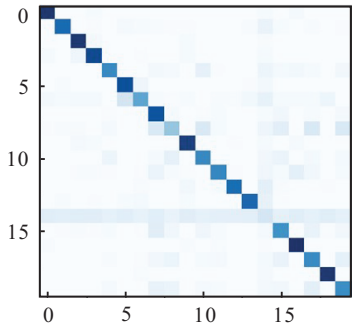
其中: $\boldsymbol{W}_\phi\in\boldsymbol{R}^{C\times C}$与$\boldsymbol{W}_\theta\in\boldsymbol{R}^{C\times C}$为卷积权值参数,$\boldsymbol{E}\in\boldsymbol{R}^{C\times d_e}$为标签词嵌入,$C$为类别数,$*$表示卷积运算. 根据式(2)和(4)对$\boldsymbol{A}$进行标准化,可得到最终所需的标签相关矩阵为

$$\hat{\boldsymbol{A}}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}(\boldsymbol{A}+2\boldsymbol{I})\tilde{\boldsymbol{D}}^{-\frac{1}{2}}. \tag{6}$$

本模块自动构建各个节点间的连接关系,即节点邻接矩阵. 图2(a)和图3(a)表示构建出的邻接矩阵的

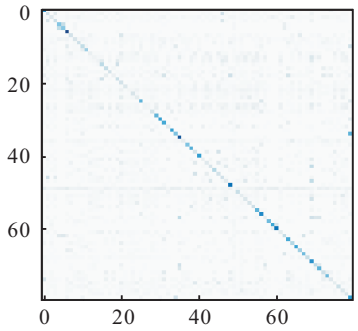


(a)　初始邻接矩阵

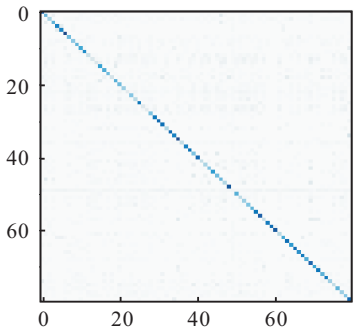


(b)　最终邻接矩阵

图 2　PASCAL VOC 2007数据集上的邻接矩阵可视化图



(a)　初始邻接矩阵



(b)　最终邻接矩阵

图 3　MS-COCO 2014数据集上的邻接矩阵可视化图

可视化图,其中对角线元素表示节点自身,颜色越深表示连接关系越显著,即描述该节点的权重越大. 由图2(b)和图3(b)可以看出,使用增强节点自身信息的邻接矩阵后,节点自身信息更为突出.

1.4 多尺度图卷积网络

目前,针对GCN的研究主要集中于一阶谱图卷积,然而随着层数增加时,该方法呈现出显著的“过度平滑”和“过拟合”等现象. 针对该问题,Luan等[20]开展了激活函数及网络结构形式对谱图卷积的表达能力限制性的研究,并将GCN推广到块Krylov子空间形式的图卷积中,同时将多尺度信息扩展到层数更多、结构更丰富的表达形式中. 因多尺度信息可以取得多层次信息互补的效果[21-22],故能够减轻每一层输出中图节点自身信息的稀释性,进而增强节点信息的可区分性.

由文献[20]对Krylov子空间的相关定义,将矩阵$\boldsymbol{P}\in\boldsymbol{R}^{N\times N}$及$\boldsymbol{Q}\in\boldsymbol{R}^{N\times F}$的$m$阶块Krylov子空间$\phi$表示为

$$\kappa_m^{\phi}(\boldsymbol{P},\boldsymbol{Q}):=\mathrm{span}^{\phi}\left\{\boldsymbol{P},\boldsymbol{PQ},\dots,\boldsymbol{P}^{m-1}\boldsymbol{Q}\right\},\quad(7)$$

同时,相应的块Krylov矩阵为

$$K_m(\boldsymbol{P},\boldsymbol{Q}):=[\boldsymbol{P},\boldsymbol{PQ},\dots,\boldsymbol{P}^{m-1}\boldsymbol{Q}].\quad(8)$$

对于由谱图滤波器$g(\boldsymbol{A})$所滤波的图信号$\boldsymbol{X}$,有如下关系:

$$g(\boldsymbol{A})\boldsymbol{X}=\sum_{n=0}^{\infty}\frac{g^{(n)}(0)}{n!}\boldsymbol{A}^n\boldsymbol{X}=[\boldsymbol{X},\boldsymbol{AX},\boldsymbol{A}^2\boldsymbol{X},\dots]\cdot\left[\frac{g^{(0)}(0)}{0!}\boldsymbol{I}_F,\frac{g^{(1)}(0)}{1!}\boldsymbol{I}_F,\frac{g^{(2)}(0)}{2!}\boldsymbol{I}_F,\dots\right]^{\mathrm{T}}=\boldsymbol{P}'\boldsymbol{Q}'.\quad(9)$$

其中:$\boldsymbol{P}'\in\boldsymbol{R}^{N\times\infty},\boldsymbol{Q}'\in\boldsymbol{R}^{\infty\times F}$,且$\boldsymbol{I}_F$为单位矩阵. 对于式(9),存在一个最小的$m$,使得

$$\mathrm{span}^{\phi}\left\{\boldsymbol{X},\boldsymbol{AX},\boldsymbol{A}^2\boldsymbol{X},\dots\right\}\approx\mathrm{span}^{\phi}\left\{\boldsymbol{X},\boldsymbol{AX},\dots,\boldsymbol{A}^{m-1}\boldsymbol{X}\right\}.\quad(10)$$

根据式(9)和(10),可得

$$g(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}'=[\boldsymbol{X},\boldsymbol{AX},\boldsymbol{A}^2\boldsymbol{X},\dots,\boldsymbol{A}^{m-1}\boldsymbol{X}]\boldsymbol{W}^{\tau},\quad(11)$$

其中$\boldsymbol{W}^{\tau}\in\boldsymbol{R}^{mF\times O}$为需要学习的输出权重.

结合式(6),基于块Krylov子空间的单层多尺度图卷积网络输出特征的表达形式为

$$\boldsymbol{H}_{l+1}=f([\boldsymbol{H}_l,\hat{\boldsymbol{A}}\boldsymbol{H}_l,\dots,\hat{\boldsymbol{A}}^{m-1}\boldsymbol{H}_l]\boldsymbol{W}_l),\quad l=0,1,2,\dots,n-1.\quad(12)$$

其中:$\boldsymbol{H}_0=\boldsymbol{X}$为输入,$\boldsymbol{H}_l$为第$l$层的输出特征,$\boldsymbol{W}_l\in\boldsymbol{R}^{(mF_l)\times F_{l+1}}$为学习的权重,$m$为块的数量,$f(\cdot)$为非线性激活函数.

1.5 损失函数

将多尺度图卷积网络学习得到的类别权重$\hat{\boldsymbol{W}}\in\boldsymbol{R}^{C\times D}$应用于第$i$个图像特征$\boldsymbol{x}_i\in\boldsymbol{R}^D$,可得到预测概率为

$$\hat{\boldsymbol{y}}=\hat{\boldsymbol{W}}\boldsymbol{x}_i.\quad(13)$$

对于真实标签$\boldsymbol{y}\in\boldsymbol{R}^C$,其中$y_i\in\{0,1\}$,多标签图像识别损失函数为

$$L_{\mathrm{cls}}=-\frac{1}{C}\sum_{j=1}^{C}\boldsymbol{y}_i^j\log(\sigma(\hat{\boldsymbol{y}}_i^j))+(1-\boldsymbol{y}_i^j)\log(1-\sigma(\hat{\boldsymbol{y}}_i^j)).\quad(14)$$

实际上,GCN可视为拉普拉斯平滑的特殊形式,它从每个节点自身和相邻所连接节点的特征中聚合信息. 因此随着层数增加,节点特征所表示的信息将会被稀释,进而极大降低了距离较大节点之间的可区分性,即“过度平滑”现象. 为此,对相关矩阵的稀疏性进行如下约束:

$$L_A=|\hat{\boldsymbol{A}}-\boldsymbol{I}|.\quad(15)$$

根据上述分析,损失函数可表示为

$$L_{\mathrm{total}}=L_{\mathrm{cls}}+L_A.\quad(16)$$

2 实 验

2.1 数据集及评估标准

PASCAL Visual object classes challenge (VOC 2007)数据集是多标签图像识别任务中广泛使用的一个数据集,它也是目标检测和语义分割等任务中常用的数据集之一. 该数据集包含trainval集(5 011张图片)和test集(4 952张图片),总共有9 963张图片并涵盖20个物体类别. Microsoft COCO 2014数据集是另一个多标签图像识别数据集,同时它也是目标检测和语义分割等任务中常用的数据集之一,其中每一张图片平均含有3个以上物体标签. 该数据集共含有80个物体类别,公众可获得部分数据集,包括含有共82 081张图片的train集以及含有共40 504张图片的val集. 实验过程中,主要采用的评估标准为每个类别的平均精度(AP)及所有类别平均精度的均值(mAP),在MS-COCO数据集实验上还增加了总体精度(OP)、总体召回率(OR)、总体F1值(OF1)、每类精度(CP)、每类召回率(CR)以及每类F1值(CF1),计算方式如下:

$$\begin{cases} \mathrm{OP}=\dfrac{\sum\limits_i N_i^c}{\sum\limits_i N_i^p},\ \mathrm{OR}=\dfrac{\sum\limits_i N_i^c}{\sum\limits_i N_i^g}, \\ \mathrm{CP}=\dfrac{1}{C}\sum\limits_i \dfrac{N_i^c}{N_i^p},\ \mathrm{CR}=\dfrac{1}{C}\sum\limits_i \dfrac{N_i^c}{N_i^g}, \\ \mathrm{OF1}=\dfrac{2\times \mathrm{OP}\times \mathrm{OR}}{\mathrm{OP}+\mathrm{OR}}, \\ \mathrm{CF1}=\dfrac{2\times \mathrm{CP}\times \mathrm{CR}}{\mathrm{CP}+\mathrm{CR}}. \end{cases} \tag{17}$$

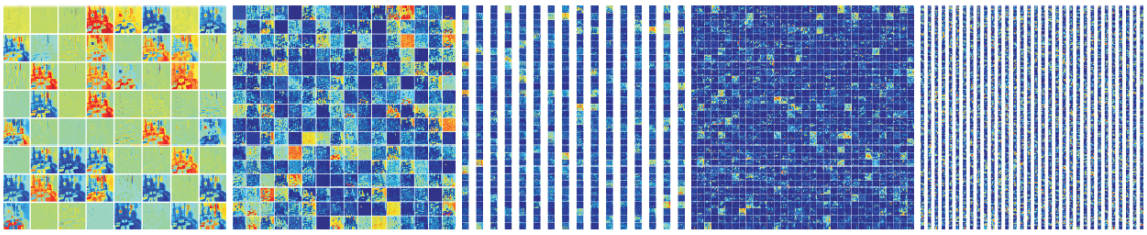
其中: N_i^c 为对第 i 个标签正确预测的图片数量, N_i^p 为对第 i 个标签进行预测的图片数量, N_i^g 为第 i 个标签的真实图片数量.

2.2　实验设置

在本文实验过程中, 硬件环境为 Intel Core i9-9900K CPU 处理器、3.60 GHz 16 GB 内存以及单个 NVIDIA GeForce RTX 2080Ti 图形处理单元, 操作系统采用 Ubuntu16.04 软件系统, 并在 PyTorch 1.40 框架上进行网络训练和测试. 在训练过程中, 将图片尺寸缩放为 512×512, 测试时图片尺寸为 448×448. 特征提取网络选择在 ImageNet 数据集上经过预训练的 ResNet-101 主干网络, 其参数见表 1; ResNet-101 网络每个卷积阶段输出的特征可视化如图 4 所示. 优化算法采用动量为 0.9 及权重衰减系数为 0.000 1 的 SGD 方法, 批量大小为 6, 初始学习率为 0.001, 并且每 30 个 epochs 以 0.1 的学习衰减率更新学习率, 直到完成 80 个 epochs 后结束训练. 各类标签表示采用具有 300 维的 GloVe 词向量, 其中包含多个词汇标签的词向量为所有单个词汇词向量的平均值. 方法中, 图卷积网络为 2 层, 其中每一层的输出维度分别为 1 024 和 2 048, 多尺度拼接过程中, 块 (blocks) 的数量初始设置为 2, 并且在每一层图卷积网络之间采用非线性函数 Tanh 作为激活函数, dropout 设置为 0.2.

表 1　ResNet101 主干网络相关参数

ResNet101 主干	conv1	conv2_x	conv3_x	conv4_x	conv5_x
参数	7×7, 64, stride 2	3×3 max pool, stride 2, $\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1\,024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2\,048 \end{bmatrix}\times3$
输出尺寸	112×112	56×56	28×28	14×14	7×7



(a) conv 1　(b) conv 2_x　(c) conv 3_x　(d) conv 4_x　(e) conv 5_x

图 4　ResNet101 主干网络各卷积层的输出特征

2.3　参数分析

在 AMS-GCN 中, 图卷积网络的层数和多尺度拼接过程中的块数是两个主要的参数, 本节将主要分析这两个参数对图像识别效果的影响.

在进行图卷积网络层数对图像识别效果的影响实验时, 仅改变图卷积的层数, 保持其他部分不变, 其中第 1 层图卷积层的输出维度为 1 024, 随后各层输出维度为 2 048, 即 1 024→2 048→ … →2 048. 本文方法 AMS-GCN 在 PASCAL VOC 2007 数据集和 MS-COCO 2014 数据集上识别结果分别如图 5 和图 6 所示. 可见: 当图卷积网络的层数为 2 层时, 最终的识别

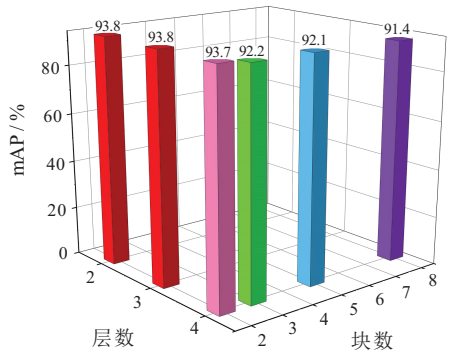


图 5　AMS-GCN 的不同层数及块数在 PASCAL VOC 2007 数据集上的性能对比

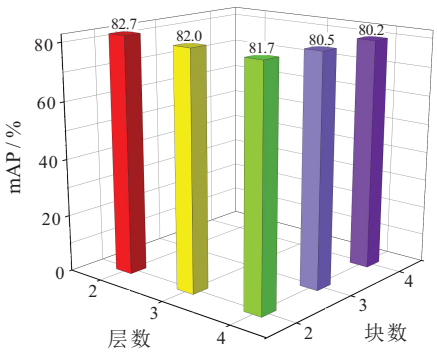


图 6 AMS-GCN 的不同层数及块数在 MS-COCO 2014 数据集上的性能对比

效果最佳;而当增加层数为 3 层或 4 层时,最终的识别效果反而有所变差. 经分析可知,实验中物体虽然具有 20 或 80 个类别,但是,由此所构建的标签关系图仍然属于规模较小的图数据,因此,虽然本文方法对多尺度信息有了更好的利用,但是,所输出的局部信息仍可能随着图卷积网络层数的增加而部分丢失.

在开展多尺度拼接过程中块的数量对图像识别效果的影响实验时,仅改变截断块 Krylov 网络中块的数量,保持其他部分不变,观察块数量的改变对识别效果的影响,此时图卷积网络的层数保持为 4 层,其中每 1 层的输出维度分别为 1 024→2 048→2 048→2 048. 由图 5 和图 6 可以看出:在 PASCAL VOC 2007 数据集与 MS-COCO 2014 数据集上,当块的数量为 2 时,AMS-GCN 的识别效果表现最佳;而当增加块的数量时,最终的性能反而会有所降低. 这是因为与大规模图数据相比,本文实验中数据集的标签所构建的关系图仍然属于小规模图数据,并且在获得块 Krylov 矩阵时采取了近似操作,所以随着小规模图数据上块数目的增加,图卷积层间输出误差变大,进而使得最终效果变差.

2.4 实验结果

由于 PASCAL VOC 2007 数据集含有的图片数量较少并且为了便于与其他方法进行对比,实验过程中采用 trainval 数据进行网络训练,利用 test 数据进行最终的效果测试,其损失收敛过程如图 7 所示. AMS-GCN 与其他多个多标签图像识别算法之间的对比实验结果见表 2,主要对比了各算法之间每一类的平均精度 (AP) 和总体类别的平均精度均值 (mAP). 由表 2 可以看出,AMS-GCN 在 PASCAL VOC 2007 数据集上所表现的整体性能比 ML-GCN 提升了 0.4 %.

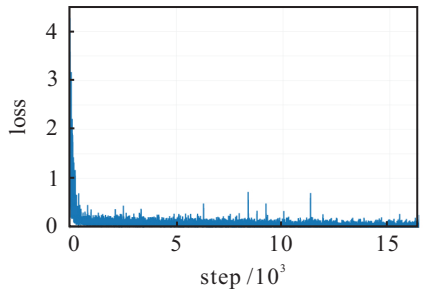


图 7 PASCAL VOC 2007 上的损失收敛曲线

表 2 AMS-GCN 与各识别方法在 PASCAL VOC 2007 数据集上的性能对比 %

算法	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv
CNN-RNN[7]	84.0	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	**99.7**	78.6
ResNet-101[1]	89.9	99.5	**97.7**	97.8	96.4	65.7	91.8	96.1	97.6	74.2	80.9	85.0	**98.4**	96.5	95.9	98.4	70.1	88.3	80.2	98.9	89.2
HCP[8]	90.9	98.6	97.1	**98.0**	95.6	75.3	94.7	95.8	97.3	73.1	90.2	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5
RNN-Atten[9]	91.9	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	**87.7**	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3
VGG[2]	91.1	98.3	97.1	96.1	96.7	75.0	91.4	95.8	95.4	76.7	92.1	85.1	96.7	96.0	95.3	97.8	77.4	93.1	79.7	97.9	89.3
ML-GCN[11]	93.4	99.5	97.5	**98.0**	**98.2**	78.8	94.9	96.7	97.3	80.9	95.1	85.3	97.7	98.2	95.9	98.6	84.6	**97.6**	82.5	98.7	92.5
A-GCN[13]	89.5	97.6	95.5	95.6	93.0	72.2	89.4	95.2	95.8	74.8	89.2	77.0	93.9	96.4	89.6	98.0	77.8	90.6	75.9	96.8	89.1
AMS-GCN	**93.8**	**99.7**	97.6	97.6	98.1	**79.6**	**95.4**	**97.3**	**97.9**	**81.1**	**96.0**	85.4	97.8	**98.5**	**96.7**	**99.0**	**84.8**	96.4	**83.7**	98.6	**94.7**

无论是物体类别数量还是图片数量,MS-COCO 2014 数据集均比 PASCAL VOC 2007 数据集大得多. 因此,在 MS-COCO 2014 数据集上进行实验更加符合实际场景设置,同时也更具有挑战性. 实验过程中采用与其他方法相同的 train 数据集进行网络训练,完成训练后利用 test 数据集进行最终的识别效果验证,其损失收敛过程如图 8 所示.

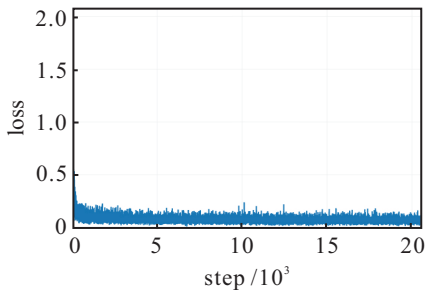


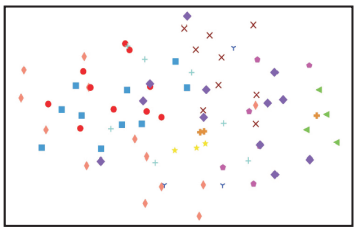
图 8 MS-COCO 2014 上的损失收敛曲线

实验包括了AMS-GCN在MS-COCO 2014数据集上运行的结果,以及与ResNet-101、ML-GCN、A-GCN等方法的对比结果,见表3. 显然,AMS-GCN在MS-COCO 2014数据集上的实验结果同样具有显著优势,其整体性能相对于ML-GCN提升了0.9%,这是由于本文方法更加充分地利用了类别标签信息,学习出了性能更强的分类器,进而表明了本文方法的有效性.

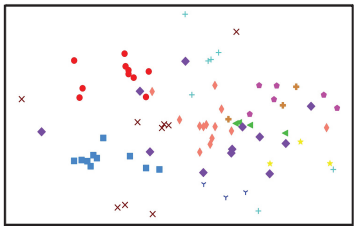
表3　AMS-GCN与各识别方法在MS-COCO 2014数据集上的性能对比　%

算法	mAP	CP	CR	CF1	OP	OR	OF1
CNN-RNN[7]	61.2	66.0	55.6	60.4	69.2	66.4	67.8
SRN[10]	77.1	81.6	65.4	71.2	82.7	69.9	75.8
ResNet-101[1]	77.3	80.2	66.7	72.8	**83.9**	70.8	76.8
ML-GCN[11]	81.8	**83.7**	71.0	76.8	82.1	74.2	77.9
A-GCN[13]	82.0	81.2	71.3	75.9	81.5	73.9	77.6
AMS-GCN	**82.7**	**83.7**	**71.5**	**77.1**	82.2	**74.5**	**78.2**

本文最终所使用的多尺度图卷积网络为两层,其输出特征的维度分别为1 024和2 048,可利用t-SNE降维方法对特征进行观察. t-SNE是一种将高维数据降维到二维或三维数据的降维方法,利用该方法可对高维数据降维进而观察其分布情况. 当高维数据降维到二维或三维并进行可视化时,类别越多其可区分性越强. 由于PASCAL VOC 2007数据集只有20类,而MS-COCO 2014数据集有80类,本文将对MS-COCO 2014数据集的标签关系图进行t-SNE降维及可视化,结果如图9所示. 图9中每个点表示每一个类别,同一种颜色表示具有相关联系的类别,可以看出,多尺度图卷积网络可使具有关联的类别明显靠近.



(a)　初始标签图t-SNE可视化



(b)　最终标签图t-SNE可视化

图9　MS-COCO 2014数据集上的t-SNE可视化结果

3　结　论

在进行多标签图像识别时,利用类别标签间的潜在关系辅助识别可以有效提升识别性能. 本文提出的基于自适应多尺度图卷积网络的多标签图像识别方法通过在标签词嵌入上自动构建标签关系图,同时在每一个图卷积层中结合多尺度信息,并利用多尺度图卷积网络在关系图上学习出比以往研究工作更有效的分类器,进而实现了多标签识别. 在数据集PASCAL VOC 2007和MS-COCO 2014上的实验结果表明了AMS-GCN的有效性. 此外,鉴于现有工作仅研究了学习分类器过程中的类别相关性,后续工作将在AMS-GCN基础上对特征提取过程中的类别特征相关性开展进一步探索.

参考文献(References)

[1] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. 2014, arXiv: 1409.1556.

[3] 吕恩辉, 王雪松, 程玉虎. 基于反卷积特征提取的深度卷积神经网络学习[J]. 控制与决策, 2018, 33(3): 447-454.
(Lv E H, Wang X S, Cheng Y H. Deep convolution neural network learning based on deconvolution feature extraction[J]. Control and Decision, 2018, 33(3): 447-454.)

[4] 赵静, 王弦, 王奔, 等. 基于神经网络的多类别目标识别[J]. 控制与决策, 2020, 35(8): 2037-2041.
(Zhao J, Wang X, Wang B, et al. Multi-category target recognition based on neural network[J]. Control and Decision, 2020, 35(8): 2037-2041.)

[5] Inoue N, Simo-Serra E, Yamasaki T, et al. Multi-label fashion image classification with minimal human supervision[C]. IEEE International Conference on Computer Vision Workshops. Venice, 2017: 2261-2267.

[6] Li Y N, Huang C, Chen C L, et al. Human attribute

recognition by deep hierarchical contexts[C]. Proceedings of European Conference on Computer Vision(ECCV). Berlin: Springer Verlag, 2016: 684-700.

[7] Wang J, Yang Y, Mao J H, et al. CNN-RNN: A unified framework for multi-label image classification[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 2285-2294.

[8] Wei Y C, Xia W, Lin M, et al. HCP: A flexible CNN framework for multi-label image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(9): 1901-1907.

[9] Wang Z X, Chen T S, Li G B, et al. Multi-label image recognition by recurrently discovering attentional regions[C]. IEEE International Conference on Computer Vision. Venice, 2017: 464-472.

[10] Zhu F, Li H S, Ouyang W L, et al. Learning spatial regularization with image-level supervisions for multi-label image classification[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 2027-2036.

[11] Chen Z M, Wei X S, Wang P, et al. Multi-label image recognition with graph convolutional networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 5172-5181.

[12] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J/OL]. 2016, arXiv: 1609.02907.

[13] Li Q, Peng X J, Qiao Y, et al. Learning label correlations for multi-label image recognition with graph networks[J]. Pattern Recognition Letters, 2020, 138: 378-384.

[14] Everingham M, Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[15] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]. Proceedings of European Conference on Computer Vision (ECCV). Berlin: Springer Verlag, 2014: 740-755.

[16] 莫宏伟, 田朋. 一种基于多层语义特征的图像理解方法[J]. 控制与决策, 2021, 36(12): 2881-2890.
(Mo H W, Tian P. An image understanding method based on multi-level semantic features[J]. Control and Decision, 2021, 36(12): 2881-2890.)

[17] 孔玮, 刘云, 李辉, 等. 基于图卷积网络的行为识别方法综述[J]. 控制与决策, 2021, 36(7): 1537-1546.
(Kong W, Liu Y, Li H, et al. A survey of action recognition methods based on graph convolutional network[J]. Control and Decision, 2021, 36(7): 1537-1546.)

[18] Chen G B, Ye D H, Xing Z C, et al. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization[C]. International Joint Conference on Neural Networks (IJCNN). Anchorage, 2017: 2377-2383.

[19] Gao H Y, Ji S W. Graph U-Nets[C]. Proceedings of International Conference on Machine Learning (ICML). New York: ACMPress, 2019: 2083-2092.

[20] Luan S T, Zhao M D, Chang X W, et al. Break the ceiling: Stronger multi-scale deep graph convolutional networks[J/OL]. 2019, arXiv: 1906.02174.

[21] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 2261-2269.

[22] 卢健, 王航英, 陈旭, 等. 基于多尺度特征表示的行人再识别[J]. 控制与决策, 2021, 36(12): 3015-3022.
(Lu J, Wang H Y, Chen X, et al. Multi-scale feature representation for person re-identification[J]. Control and Decision, 2021, 36(12): 3015-3022.)

作者简介

王雪松 (1974–), 女, 教授, 博士生导师, 从事机器学习及模式识别、人工智能等研究, E-mail: wangxuesongcumt@163.com;

荣小龙 (1994–), 男, 硕士生, 从事图像识别的研究, E-mail: xiaolong_rong@cumt.edu.cn;

程玉虎 (1973–), 男, 教授, 博士生导师, 从事机器学习、模式识别与智能系统等研究, E-mail: chengyuhu@163.com;

陈正升 (1984–), 男, 讲师, 博士, 从事机器人动力学建模与控制等研究, E-mail: chenzhengsheng@cumt.edu.cn.

(责任编辑：李君玲)