一种融合标签间强相关性的多标签图像分类方法

张辉宜[1],夏媛龙[1],周克武[2],包向华[2],陶 陶[1]
1. 安徽工业大学 计算机科学与技术学院,安徽 马鞍山 243000
2. 马鞍山市公安局,安徽 马鞍山 243000

摘 要:为了将标签间的语义相关性引入多标签图像分类模型中,传统的方法例如 ML-GCN 通过设置单阈值将标签条件概率矩阵二值化为标签共现矩阵,然而,仅设置单阈值很难归纳所有的标签语义关系情况。针对这一问题,提出一种融合标签间强相关性的多标签图像分类方法—MGAN(Multiple Graph Convolutional Attention Networks),通过设置多个阈值,将传统的标签条件概率矩阵按照不同的相关性程度分割为多个子图;同时,为了提升多标签分类性能,也引入图像区域空间相关性。另外,针对传统的"CNN+GCN"方法将标签与特征的融合张量视为预测分数缺乏可解释性问题,将标签与特征的融合张量视为注意力分数;在 MS-COCO 和 PASCAL VOC 数据集上与其他主流多标签图像分类方法进行了对比实验,平均准确率分别达到了 94.9%和 83.7%,相较于经典 ML-GCN 模型,分别获得了 0.9%和 0.8%准确率提升,且在"Binary"和"Re-weighted"邻接矩阵模式下,MGAN 都有较好的表现,验证了新的融合方法可以缓解图卷积神经网络过平滑问题对多标签图像分类的影响。

关键词:多标签图像分类;语义相关性;图卷积网络;注意力机制;区域空间相关性
中图分类号:O643 **文献标识码**:A doi:10.16055/j.issn.1672-058X.2023.0005.002

A Method of Multi-label Image Classification with Fusing Powerful Semantic Correlation

ZHANG Huiyi[1], XIA Yuanlong[1], ZHOU Kewu[2], BAO Xianghua[2], TAO Tao [1]
1. School of Computer Science and Technology, Anhui University of Technology, Anhui Maanshan 243000, China
2. Ma' anshan Public Security Bureau, Anhui Maanshan 243000, China

Abstract: In order to introduce semantic correlation between labels into multi-label image classification model, traditional methods, such as ML-GCN, transform label conditional probability matrix into label co-occurrence matrix by using single threshold value. However, it is difficult to sum up all semantic relationships of all labels by using single threshold value. To solve this problem, a method of multi-label image classification with fusing powerful semantic correlation, MGAN, was proposed. By setting multiple thresholds, the traditional conditional probability matrix of labels was divided into multiple subgraphs according to different degrees of correlation. Meanwhile, in order to improve the performance of multi-label classification, image region spatial correlation was also introduced. In addition, the traditional "CNN+GCN" method regards the fusion tensor of label and feature as the lack of interpretability of the predicted fraction. To solve this problem, MGAN regards the labels and feature' s fusion tensor as the attention score. Compared with other mainstream multi-label image classification methods on MS-COCO and PASCAL VOC datasets, the mAP were 94.9% and 83.7% respectively, which were 0.9% and 0.8% higher than traditional ML-GCN model. And MGAN performed well in both "Binary" and "Re-weighted" adjacency matrix mode, which verified that the new fusion method can alleviate the influence of graph

收稿日期:2022-03-09 **修回日期**:2022-04-18 **文章编号**:1672-058X(2023)05-0008-08
基金项目:安徽省重点研发计划项目(201904D07020020);安徽省自然科学基金项目(1908085MF212).
作者简介:张辉宜(1963—),四川自贡人,男,硕士,教授,从事深度学习研究.
通讯作者:夏媛龙(1994—),男,安徽合肥人,硕士研究生,从事深度学习研究. Email:1436746845@qq.com.
引用格式:张辉宜,夏媛龙,周克武,等. 一种融合标签间强相关性的多标签图像分类方法[J]. 重庆工商大学学报(自然科学版),2023,40(5):8—15.
ZHANG Huiyi, XIA Yuanlong, ZHOU Kewu, et al. A method of multi-label image classification with fusing powerful semantic correlation[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(5): 8—15.

convolutional neural network' s "over smoothing" problem on multi-label image classification.

Keywords: multi-label image classification; semantic correlation; graph convolutional network; attention mechanism; regional spatial correlation

1 引　言

多标签图像分类在很多如医学图像检测[1]、卫星地图检测[2]等机器视觉邻域得到应用。传统的卷积神经网络,如 Resnet[3],Densnet[4],VGG[5]等,在单标签图像识别邻域中取得了很高的成就,但在面对多标签分类问题时,简单地将多标签分类当作单标签分类问题处理,忽略标签之间的依赖关系和图像的空间信息,因而不能反映真实世界图像集复杂的数据关系。

因此,研究者尝试将视觉图像与和他关联的标签集合联系在一起,于是出现了基于概率图模型[6-7]、循环神经网络[8]、图卷积神经网络[9]等多标签图像的分类方法,以及基于概率图模型利用数学理论知识构建标签语义相关性。然而,基于概率图模型被认为高计算复杂度而导致扩展性很差,因此,神经网络被引入用来替代概率图建模。CNN-RNN 框架[8]利用循环神经网络建立标签语义相关性,将图像经过卷积后的特征和标签向量融合到一个嵌入空间中。然而,循环神经网络 RNN 是一个顺序的线性神经网络,能够表现出标签的高阶关系,却不能很好地表现出标签的结构关系。

图结构数据被证明是更有效地建立标签相关性的方法。ML-GCN[9]用共现邻接矩阵建立标签相关性模型,利用图卷积神经网络学习标签节点表示,在模型结构最后,将学习到的标签节点表示与特征图相融合。ML-GCN[9]让多标签图像识别模型性能得到很大提升,并引发研究者利用图卷积建模标签相关性的潮流。但 ML-GCN[9]建立标签相关性时却忽略了标签内部的强相关性。例如,{"餐桌","茶杯","碗","人","椅子","雨伞","汽车"}有一定相关性,而{"茶杯","碗","餐桌"},{"人","椅子","雨伞","汽车"}有更强的相关性,且 ML-GCN 通过设置单阈值将标签关系条件概率矩阵二值化,而由于标签的语义关系较复杂,单阈值不能很好地划分所有的标签关系情况。因此,本文通过设置多个阈值将传统标签条件概率图按照不同的相关性强度分割为多个邻接子图。

先前的多标签分类方法存在类似的分割标签语义相关性工作,MGTN[10]模型将标签共现矩阵按照标签相关性强度分割为多个子图邻接矩阵,并用 multiple CNNs[3]结合社区检测算法学习不同子图标签。通过实验证明该方法对多标签图像分类有很强的提升,但是这样的方法需要消耗巨大计算量,由于 MGTN 中子图的个数与 CNN 基线网络数成正比,因此,随着子图数增加,整个网络计算量也近乎呈正比例增长。

本文利用多重图卷积神经网络学习不同子图标签,将不同图卷积神经网络学习到的标签表示划分为标签组。子图级别注意力池化 SLAP 利用特征图中的特征点,选择与之适配的标签组中的标签,特征级别注意力池化 FLAP 利用标签组中的标签,选择与之适配的特征点。不仅能学习到不同强度相关性的标签语义相关性,且大大降低计算复杂度。本文中,子图的个数与多重图卷积层[11]的重数成正比,而图卷神经网络需要消耗的计算量远小于卷积神经网络的骨干网络,因此可以很容易在子图个数上得到扩展。

同时,研究者尝试将图像的区域空间相关性应用于多标签图像分类中,基于 EdgeBoxes[12]或者 Selective Search[13]等目标提案[14-16](object proposal)是其中一类方法。例如 HCP[14],这些方法通常将提取出的目标提案送入共享卷积神经网络中,最后用最大池化计算提案分数。虽然目标提案方法能够很好地提取图像实体间的空间关系,但生成提案的方法计算复杂度高。基于注意力机制是另一类引入图像的区域空间相关性多标签图像分类方法,例如,SRN[17]引入注意力图像学习标签间的潜在空间关系。本文利用注意力机制将图像的区域空间相关性引入到卷积神经网络中,用于提到的多标签图像分类性能。

本文同时引入标签的语义相关性和图像的视觉区域空间相关性,并通过实验验证了标签语义相关性和图像区域相关性的引入对多标签分类方法有所提升。类似地,近年来也存在相关模型将标签的语义相关性和图像的空间信息引入到多标签图像分类中的现象:TSGCN[18]将标签共现关系视为标签的语义关系图,Mask R-CNN[19]提取图像的区域并设计物体间的空间关系图,TSGCN[18]用独立的模块分别学习标签的语义相关性和图像中各个实体之间的空间相关性,BMML[20]利用 MA-CNN[21]提取图像的区域空间特征并用循环神经网络 LSTM[22]学习标签语义关系。上述模型同时将标签的语义相关性和图像的空间信息引入到多标签图像分类中,且证明了语义关系和空间信息的引入都对多标签图像分类性能有所提升,然而这些模型结构相对庞大,消耗的计算量较多,本文的目的是提出一个简单且有效的方法学习标签的语义关系和图像区域空间关系。

另外,类似于 ML-GCN[9],传统"CNN+GCN"模式的网络结构还存在一个争议:将图像特征与图像特征毫无关系的标签表示相乘,且将没有其他突出图像特征重要性的方法直接送入分类器的设计是否合理尚待证明。

图像特征与标签表示在网络中不应当处于同等地

位,标签语义相关性应当是辅助卷积神经网络学习图像特征。因此,本文将标签与特征图的融合视为 softmax 函数的输入,用于计算标签与特征点之间的注意力分数。这样的设计在理论上更能表现标签的语义相关性对多标签图像识别的作用,且注意力分数由图像特征和标签两部分组成,相较于传统的方法学习图卷积层参数的难度在理论上会更小一些。以下为本文的主要工作:

(1) 提出一个融合标签间强相关性的多标签图像分类方法 MGAN(Multiple Sub-Graph Attention Network),学习标签间不同相关性强度的语义相关。方法中邻接子图个数与图卷积重数呈正比,相较于先前的融合标签间强相关性方法(MGTN[10]),分类模型在子图个数上扩展性更强。

(2) MGAN 同时引入标签的语义相关性和图像的区域空间相关性,SLAP 算法通过特征图中的特征点选择标签组中适配的标签,FLAP 通过标签选择适配的特征图上的特征点。通过实验证明了标签语义相关性和图像区域空间相关性的引入对多标签图像分类的作用。

(3) 改变传统的将标签与特征图融合直接视为预测标签并送入分类器的方法,将融合张量视为注意力分数辅助基线网络学习图像特征,在理论上更能表现标签的语义相关性对多标签图像分类的作用,且通过实验证明此设计能有效缓解图卷积神经网络过平滑问题对多标签图像分类的影响。

2 MGAN 的构建

如图 1 所示,本文提出的多标签图像分类方法包含两个部分。上面部分为多标签图像分类的基线网络(例如 Resnet-101)提取图像特征,并将得到的特征图转化为特征点形式;下面部分为将传统的条件概率矩阵转化为多个邻接子图,并将子图邻接张量和标签嵌入矩阵送入多重图卷积神经网络中。之后将学习到的标签表示张量转化为标签组形式。在网络的最后,将特征点形式特征图与标签组融合,通过 SLAP 和 FLAP 算法各自求得预测标签,再相加得到最终的预测分数。

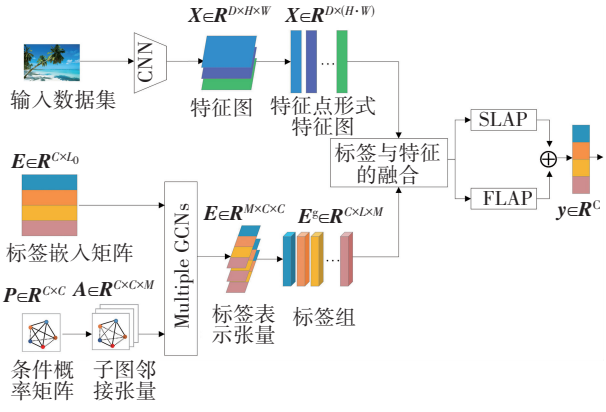

图 1 MGAN 网络模型全局结构示意图

Fig. 1 Model structure of MGAN network

2.1 多重图卷积层与标签组

相较于传统构造标签共现关系方法,本文参照 MGTN[10] 中方法将标签设置不同阈值 $T=[t_1,t_2,\cdots,t_M]$ 并将标签共现条件概率矩阵 $\boldsymbol{P}$ 分割为多个邻接子图,其中,$t_i\in[0,1]$ 且 $t_i<t_j\ \forall\, i<j$。

$$\boldsymbol{A}_{kij}=\begin{cases}1 & \text{if } P_{ij}\in[t_{k-1},t_k)\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

多重邻接子图 $\boldsymbol{A}\in\boldsymbol{R}^{C\times C\times M}$ 由邻接矩阵 $\boldsymbol{A}_k\in\boldsymbol{R}^{C\times C}$,$k=\{1,\cdots,M\}$,其中 M 为子图个数。图卷积神经网络(GCN)[11] 开始时被用来处理半监督分类任务,其本质是在节点之间传播信息来更新节点表示。本文为了让标签嵌入矩阵在多个邻接子图中传播,更新节点表示信息,特引入多重图卷积层学习不同子图中的节点表示。

$$\boldsymbol{E}_k^{l+1}=h(\hat{\boldsymbol{A}}_k\boldsymbol{E}_k^l\boldsymbol{W}_k^l) \tag{2}$$

多重图卷积层第一层输入为标签嵌入矩阵 $\boldsymbol{E}\in\boldsymbol{R}^{C\times L_0}$ 和归一化后的多重邻接子图 $\hat{\boldsymbol{A}}_k\in\boldsymbol{R}^{C\times C}$,子图被送入到不同的图卷积层中,经过多重图卷积层后,得到标签表示张量 $\boldsymbol{E}\in\boldsymbol{R}^{M\times C\times L_l}$,其中,$L_l$ 为第 l 层图卷积输入的标签表示张量的节点维度,$l\geqslant 2$,$l\in\boldsymbol{R}^N$;$h(\cdot)$ 为非线性激活函数。式(2)为多重图卷积层在大于两层时的表达式,其中的 $\boldsymbol{W}_k^l$ 为图卷积层中的参数矩阵,类似于式(2),第一层图卷积层表达式为 $\boldsymbol{E}_k^2=h(\hat{\boldsymbol{A}}_k\boldsymbol{E}\boldsymbol{W}_k^l)$。

在标签经过多重图卷积层后,将最终得到的标签表示张量 $\boldsymbol{E}\in\boldsymbol{R}^{M\times C\times L}$ 重新组合、归纳为新的标签组。具体方法如下:得到的标签表示张量 $\boldsymbol{E}$ 的维度为 $M\times C\times L$,此时代表 $\boldsymbol{E}$ 可以表示为 $[\boldsymbol{E}_1,\boldsymbol{E}_2,\cdots,\boldsymbol{E}_M]$ 的组合,此时标签表示张量由标签表示矩阵 $\boldsymbol{E}_i\in\boldsymbol{R}^{C\times L}$ 组成。将标签表示张量维度变换为 $C\times L\times M$,此时代表 $\boldsymbol{E}$ 可以表示为 $[\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_C]$ 的组合,此时标签表示张量由标签组 $\boldsymbol{G}_i\in\boldsymbol{R}^{L\times M}$ 组成。将经过不同子图和不同图卷积层更新的标签节点表示归纳为一组,方便之后与卷积神经网络特征图计算注意力分数。

2.2 卷积神经网络与特征点形式特征图

类似于传统单标签图像识别,首先将图片剪裁成统一规格送入卷积神经基线网络 φ 中提取特征。在这里用 Resnet-101[3] 作为例子,将图片剪裁为 224×224 规格,则送入 Resnet-101 后能得到维度为 $\boldsymbol{X}\in\boldsymbol{R}^{2\,048\times 7\times 7}$ 特征图。如果将特征图分解为 $[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{49}]$,则,$\boldsymbol{x}_i\in\boldsymbol{R}^{2\,048}$ 为特征图上的某一个特征点。

$$\boldsymbol{x}=\varphi(I;\theta)$$

其中,θ 是 CNN 主干网络中的参数。将特征图 $\boldsymbol{X}\in\boldsymbol{R}^{D\times H\times W}$ 表示为 $\boldsymbol{X}^f=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{(H\cdot W)}]\in\boldsymbol{R}^{D\times(H\cdot W)}$ 特征

点形式,其中 D 为特征图的通道,H 和 W 为特征图的高和宽,$\boldsymbol{x}_i \in \boldsymbol{R}^{(H \cdot W)}$ 为特征图上的某一个特征点。通过改变最后多重图卷积参数矩阵 $\boldsymbol{W}_k^l$ 维度,使得 $D=L$。同样的,可以得到标签表示张量的标签组形式 $\boldsymbol{E}^g=[\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_C]\in\boldsymbol{R}^{C\times L\times M}$。

因此,通过变换维度来转换特征图和标签表示张量的表现形式:

(a) $\boldsymbol{X}\rightarrow\boldsymbol{X}^f:\boldsymbol{R}^{D\times H\times W}\rightarrow\boldsymbol{R}^{D\times(H\cdot W)}$,此为特征图的特征点形式转换。

(b) $\boldsymbol{E}\rightarrow\boldsymbol{E}^g:\boldsymbol{R}^{M\times C\times L}\rightarrow\boldsymbol{R}^{C\times L\times M}$,此为标签表示张量的标签组形式转换。

2.3　特征点与标签的融合

将 $\boldsymbol{X}^f$ 的转置与每一个标签组 $\boldsymbol{G}_i$ 做矩阵乘法,可以保证每个特征点向量与标签组内的每个标签表示向量相乘。再用同样的方法将特征图与每个标签组相乘,最终堆叠起来,如下:

$$\boldsymbol{Z}=\|_{i=1}^{C}(\boldsymbol{X}^f)^{\mathrm{T}}\boldsymbol{G}_i$$

其中,$(\boldsymbol{X}^f)^{\mathrm{T}}\in\boldsymbol{R}^{(H\cdot W)\times D}$,$\boldsymbol{G}_i\in\boldsymbol{R}^{L\times M}$,$L=D$,$\|(\cdot)$ 表示矩阵的堆叠,将所有特征图与标签组乘积得到的矩阵堆叠为一张类属张量。由上,可以得到一张由特征点与标签组中的标签节点融合而成的融合特征图 $\boldsymbol{Z}\in\boldsymbol{R}^{C\times(H\cdot W)\times M}$,其中,$\boldsymbol{Z}=[\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_C]$,$\boldsymbol{Z}_i\in\boldsymbol{R}^{(H\cdot W)\times M}$ 为特征图与每个标签组融合得到的矩阵。

分析融合特征图 $\boldsymbol{Z}\in\boldsymbol{R}^{C\times(H\cdot W)\times M}$,其由特征图 $(\boldsymbol{X}^f)^{\mathrm{T}}$ 和标签组 $\boldsymbol{G}_i$ 相乘得来,因此,里面的每一个元素都是特征点与标签的乘积 $\boldsymbol{x}_i^{\mathrm{T}}\cdot\boldsymbol{g}_j$,其中,$\boldsymbol{x}_i\in\boldsymbol{R}^D$,$\boldsymbol{g}_j\in\boldsymbol{R}^L$,$L=D$,$i$ 为 1 到 $H\times W$ 之间正整数,j 为 1 到 $C\times M$ 之间正整数,且融合特征图 $\boldsymbol{Z}$ 的维度为 $C\times(H\cdot W)\times M$。如果在 M 维度上使用 softmax 函数归一化,可以计算出标签组中的标签相对于特征点的注意力分数,如果在 $H\cdot W$ 维度上使用 softmax 函数归一化,可以计算出特征点相对于标签的注意力分数。因此,子图级别注意力池化(SLAP)算法可以让特征点选择与之适配的标签组中的标签,特征级别注意力池化(FLAP) 算法可以让标签选择与之适配的特征点。

2.4　子图级别注意力池化(SLAP,Sub-Graph Level Attention Pooling)

图 2 为子图级别注意力池化方法流程。其中,“R”,“T”,“×”,“‖”,“S_1”,“+”分别表示“维度变换”、“转置”、“矩阵乘法”、“堆叠”、“计算子图级别注意力分数”、“矩阵加法”操作,“1×1 conv”为 1×1 卷积层。

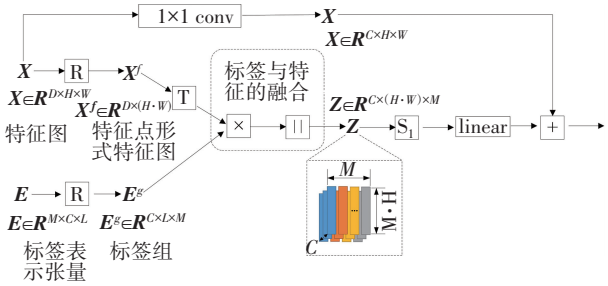

图 2　子图级别注意力池化流程(SLAP)

Fig. 2　Process of sub-graph level attention pooling (SLAP)

子图级别注意力池化的目的是利用特征点选择与之适配的标签组标签。首先,计算经过标签组的标签与特征图上的特征点在子图级别的注意力分数 $\boldsymbol{S}_k^g\in\boldsymbol{R}^{C\times(H\cdot W)\times M}$。

$$\boldsymbol{S}_k^g=\frac{\exp(Z_{ijk})}{\sum_{k=1}^{M}\exp(Z_{ijk})}$$

因此,子图级别注意力池化如下:

$$P_{ij}^g=\sum_{k=1}^{(H\cdot W)}S_{ijk}^gZ_{ijk}$$

为了保留或者突出图像特征,本文参照 Resnet[3] 中的“残差链接”或者 Transformer[23] 中的“Add”层,将原始特征图加到子图级别池化特征图上。首先,原始特征图通过“1×1 卷积”维度变换与子图级别池化特征图 $\boldsymbol{P}^g\in\boldsymbol{R}^{C\times(H\cdot W)}$ 保持一致($\boldsymbol{X}:\boldsymbol{R}^{D\times H\times W}\rightarrow\boldsymbol{R}^{C\times H\times W}\rightarrow\boldsymbol{R}^{C\times(H\cdot W)}$),并与子图级别池化特征图相加在一起

$$\boldsymbol{P}^g=\boldsymbol{P}^g+\boldsymbol{R}^{C\times(H\cdot W)}$$

其中,$c(\cdot)$ 表示使用 1×1 卷积层,$\boldsymbol{R}^{C\times(H\cdot W)}$ 是经过维度变换将特征图 $C(x)\in R^{C\times H\times W}$ 从三维转换为二维后的矩阵。最后使用平均池化方法求出预测标签:

$$P_i^g=\frac{1}{M}\sum_{j=1}^{M}P_{ij}^g$$

2.5　特征级别注意力池化(FLAP, Feature Level Attention Pooling)

图 3 为特征级别注意力池化方法流程。图 3 中,“R”,“T”,“×”,“‖”,“S_2”,“+”分别表示“维度变换”、“转置”、“矩阵乘法”、“堆叠”、“计算特征级别注意力分数”、“矩阵加法”操作,“1×1 conv”为 1×1 卷积层,“linear”为线性变换。

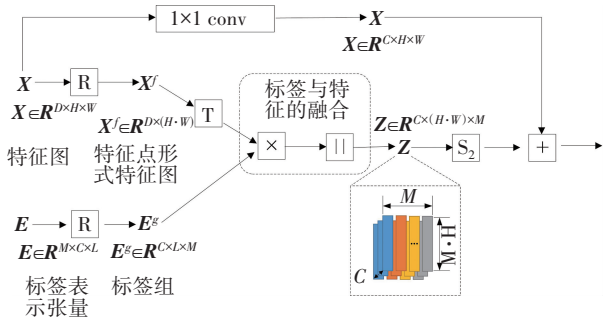

图 3　特征级别注意力池化流程(FLAP)

Fig. 3　Process of feature level attention pooling (FLAP)

特征注意力池化与子图级别注意力池化类似,其目的是利用标签组的标签选择与之适配的特征点:$\boldsymbol{R}^{C\times M\times(H\cdot W)}$为对融合张量$\boldsymbol{Z}$作维度转换后的矩阵($\boldsymbol{Z}:\boldsymbol{R}^{C\times(H\cdot W)\times M}\rightarrow\boldsymbol{R}^{C\times M\times(H\cdot W)}$),再送入softmax函数中,计算特征点相对于标签的注意力分数,计算的特征级别注意力分数为$\boldsymbol{S}^f\in\boldsymbol{R}^{C\times M\times(H\cdot W)}$。

$$S_k^f=\frac{\exp(\boldsymbol{R}^{C\times M\times(H\cdot W)}(Z_{ijk}))}{\sum_{k=1}^{M}\exp(\boldsymbol{R}^{C\times M\times(H\cdot W)}(Z_{ijk}))}$$

再将注意力分数乘到维度转换后的融合张量Z,在最后一个维度求和,得到在特征级别池化特征图$\boldsymbol{P}^f\in\boldsymbol{R}^{C\times M}$。

$$P_{ij}^f=\sum_{k=1}^{M}S_{ijk}^f\boldsymbol{R}^{C\times M\times(H\cdot W)}(Z_{ijk})$$

同样地,将原始特征图加到特征级别类属特征图上。类似Transformer[23]网络中的FFN层,本文在特征级别池化特征图上做线性变化,而不是在原始特征图上做线性变换。利用linear(·)函数对特征级别池化特征图$\boldsymbol{P}^f$做线性转换:$\boldsymbol{P}^f:\boldsymbol{R}^{C\times M}\rightarrow\boldsymbol{R}^{C\times(H\cdot W)}$,再与经过1×1卷积和维度变换($\boldsymbol{X}:\boldsymbol{R}^{D\times H\times W}\rightarrow\boldsymbol{R}^{C\times H\times W}\rightarrow\boldsymbol{R}^{C\times(H\cdot W)}$)后的原始特征图相加,得到最终的特征级别注意力特征图$\boldsymbol{P}^f\in\boldsymbol{R}^{C\times(H\cdot W)}$。

$$\boldsymbol{P}^f=\text{linear}(\mathrm{P}^f)+\boldsymbol{R}^{C\times(H\cdot W)}(C(x));\ P_i^f=\frac{1}{(H\cdot W)}\sum_{j=1}^{H\cdot W}P_{ij}^f$$

用平均池化消除最后一个维度求得预测标签。将子图级别池化和特征级别池化得到的预测分数相加,得到最终的预测标签,$\hat{y}\triangleq(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_C)=(P_1,P_2,\cdots,P_C)$,并送往分器。

$$\boldsymbol{P}=\boldsymbol{P}^g+\boldsymbol{P}^f$$

最后,选择二分类交叉熵损失函数(BCE)计算预测标签和真实标签之间的损失,并用SGD优化方式最小化损失函数。

3 实验结果与分析

本文选择被广泛应用于多标签图像分类的MS-COCO[24]和PASCAL VOC 2007[25]两个数据集并与其他典型的多标签分类模型进行对比评估实验,通过消融实验分析了各组成部分、图卷积层的重数与层数对实验结果的影响。

3.1 实验设置

评价指标:采用被广泛应用的多标签图像分类平均准确率(mAP)作为主要评价指标,同时,参照先前的多标签图像相关研究[9],采用全局准确率(OP),全局召回率(OR),全局综合评价指标(OF1),每个类别准确率(CP),召回率(CR),综合评价指标(CF1)作为性能评估。

实验细节:将多重图卷积神经网络的层数和重数设置为2,阈值设置为$T=[0.2,0.4,1]$。两层图卷积层输出维度分别为1 024和2 048,并将经过维基百科数据集预训练的Glove模型[26]对每个标签生成300维词向量作为图卷积神经网络的输入。类似于ML-GCN[9],本文采用LeakyReLU作为图卷积中间层的非线性激活函数,负斜率设置为0.2。对于数据增强,参照先前的相关工作[9],对数据集做随机水平翻转(random horizontal flip)和随机大小裁剪处理(random resized crop)。Resnet-101[3]作为对比实验的基线网络,且Resnet-101[3]分别在ImageNet数据集[27]和经过cutmix[28]数据增强的ImageNet数据集[31]提前预训练。无论是在MS-COCO数据集[24]还是PASCAL VOC 2007数据集[25]上的实验,都将批量大小设置为16,且将图片裁剪为448×448大小输入基线网络中。对于基线网络部分,学习率设置为0.01,对于非基线网络的其他部分,学习率全设置为0.1。

3.2 实验结果

如表1、表2所示,在PASCAL VOC 2007数据集上与其他主流多标签图像分类模型进行对比实验,并以类别平均准确率(AP)、平均精准率(mAP)作为评价指标。其中,PASCAL VOC 2007数据集包含的9 963张图片分别被划分为训练集、验证集和测试集,且包含20个目标类别。其中,"MGAN-FLAP"表示MGAN网络只保留特征级别注意力池化部分,计算模型为式(9)—式(12)。"MGAN-cut"表示MGAN网络中的基线网络Resnet-101在在cutmix处理的ImageNet数据集中预训练。

表1 PASCAL VOC 2007数据集上的对比实验结果1

Table 1 Comparative experimental results on PASCAL VOC 2007 dataset 1

方 法	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table
CNN-RNN[8]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0
ResNet-101[3]	99.5	97.7	97.8	96.4	65.7	91.8	96.1	97.6	74.2	80.9	85.0
HCP[14]	98.6	97.1	98.0	95.6	75.3	94.7	95.8	97.3	73.1	90.2	80.0
ML-GCN(Binary)[9]	99.6	98.3	97.9	97.6	78.2	92.3	97.4	97.4	79.2	94.4	86.5
ML-GCN(Re-weighted)[9]	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4
MGAN(Binary)	99.8	98.2	97.6	98.4	81.9	93.3	97.1	97.8	83.2	94.9	89.1
MGAN(Re-weighted)	99.7	98.1	97.7	98.5	81.9	94.1	97.2	98.0	82.4	93.8	89.1
MGAN-FLAP(Re-weighted)	99.8	98.4	98.0	98.2	81.1	94.8	97.2	97.9	82.5	95.0	89.9
MGAN-cut(Binary)	99.6	98.6	97.9	98.7	80.9	94.4	97.8	97.7	82.9	96.0	92.2
MGAN-cut(Re-weighted)	99.6	97.6	97.9	98.2	83.1	95.6	97.9	98.7	81.8	95.4	90.9

表 2　PASCAL VOC 2007 数据集上的对比实验结果 2

Table 2　Comparative experimental results on PASCAL VOC 2007 dataset 2

方　法	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN[8]	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
ML-GCN(Binary)[9]	86.5	97.4	97.9	97.1	98.7	84.6	95.3	80	98.6	90.4	93.1
ML-GCN(Re-weighted)[9]	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
MGAN(Binary)	89.1	97.5	98.0	96.1	99.0	83.5	95.5	86.4	99.0	93.1	94.0
MGAN(Re-weighted)	89.1	97.8	97.8	96.2	99.1	93.1	95.3	87.8	99.1	92.7	94
MGAN-FLAP(Re-weighted)	89.9	98.0	98.0	96.7	99.0	84.6	96.9	86.6	99.1	93.4	94.2
MGAN-cut(Binary)	92.2	98.5	98.5	97.8	99.1	86.2	95.9	88.2	99.2	95.4	94.8
MGAN-cut(Re-weighted)	90.9	98.5	97.7	97.5	99.0	87.6	97.2	88.8	99.1	95.8	94.9

MGAN 方法与 ML-GCN 方法在平均精准率 mAP 上持平(基线网络 Resnet-101 在 ImageNet 数据集中预训练),但当将 MGAN 中的特征级别注意力池化单独实验时(MGAN-FLAP),发现有更好的性能提升。这主要归因于图像的空间相关性对多标签图像分类的作用,也说明了标签辅助选择特征点能有效提升模型的分类性能。

“Re-weighted”是 ML-GCN 模型为了防止图卷积神经网络过平滑,而为二值化的共现邻接矩阵设计的“重新加权模式”。通过实验证明,“Binary”模式的邻接矩阵和“Re-weighted”模式的邻接矩阵对 MGAN 模型影响很小。其中的原因是将标签与特征的融合视为注意力分数,弱化了图卷积神经网络对 MGAN 模型性能的影响,因此,即使采用原始的“0-1”共现矩阵作为图卷积神经网络的输入,也有很好的性能表现。

“MGAN-cut”表示 MGAN 中基线网络 Resnet-101 在经过 cutmix[28] 数据增强后的 imageNet[27] 数据集中的提前预训练。相较于 ML-GCN 模型[9],MGAN 模型在“Binary”邻接矩阵模式下,平均精准率(mAP)从 93.1%提升到 94.8%,获得了 1.7%准确率提升;在“Re-weighted”邻接矩阵模式下,平均精准率(mAP)从 94.0%提升到 94.9%,也得到了 0.9%准确率提升。由此,可以证明 MGAN 方法的有效性。

表 3 为在 MS-COCO 数据集上的实验结果。MS-COCO 是最为流行的多标签图像数据集之一,其训练集中包含超过 80 000 张图片而验证集中包含超过 40 000 张图片。其中的目标实体被分为 80 个标签类别且每张图片大概包含 2.9 个标签类别。MGAN 的基线网络 Resnet-101[3] 在 cutmix[28] 数据增强处理后的 ImageNet[27] 数据集中预训练,并以 mAP、CP、CR、CF1 和 top-3 标签的 CP、CR、CF1 作为对比实验评价指标。从表 3 可以看出:MGAN 方法在大部分指标上优于其他主流方法,特别是平均精准率(mAP)较好,同时,没有达到最优的部分指标的数值也非常接近于最优指标。与 ML-GCN 相比,在“Binary”邻接矩阵模式下,平均精准率从 80.3%提升到 83.4%,得到 3.7%准确率提升;在“Re-weighted”邻接矩阵模式下,平均精准率从 83.0%提升到 83.7%,也得到 0.8%准确率提升。与 PASCAL VOC 2007 数据集上实验相似,在 MS-COCO 数据集上,是否对共现矩阵“重新加权”对实验结果的影响很小。这同时说明了将标签与特征的融合视为注意力分数缓解了图卷积神经网络的过平滑问题,对多标签图像分类的影响在 MS-COCO 数据集上也得到了验证。

表 3　MS-COCO 数据集上的对比实验结果

Table 3　Comparative experimental results on MS-COCO dataset

方　法	All							Top-3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
CNN-RNN[8]	61.2	—	—	—	—	—	—	66.0	55.6	60.4	69.2	66.4	67.8
SRN[17]	77.1	81.6	65.4	71.2	82.7	69.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
Resnet-101[3]	77.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6
Resnet-101-cut[32]	82.1	86.2	68.7	76.4	88.9	73.1	80.3	88.7	61.3	72.5	92.1	65.2	76.3
ML-GCN(Binary)[9]	80.3	81.1	70.1	75.2	83.8	74.2	78.7	84.9	61.3	71.2	88.8	65.2	75.2
ML-GCN(Re-weighted)[9]	83.0	85.1	72.0	78.0	85.8	75.4	80.3	89.2	64.1	74.6	90.5	66.5	76.7
MGAN-cut(Binary)	83.4	84.9	71.7	77.8	88.1	75.2	81.1	88.0	63.9	74.0	91.9	66.4	77.1
MGAN-cut(Re-weighted)	83.7	84.8	72.6	78.2	87.9	76.1	81.5	88.0	64.4	74.4	91.7	67.0	77.4

3.3 消融实验

表4为在PASCAL VOC 2007和MS-COCO数据集上验证MGAN不同组成部分的有效性消融实验结果。其中,“MGAN-SLAP”表示在基线网络上加入子图级别池化,计算模型为式(5)—式(8);“MGAN-FLAP”表示在基线网络上加入特征级别池化,计算模型为式(9)—式(12);“MGAN-Both”表示在基线网络上加入子图级别池化和特征级别池化,计算模型为式(5)—式(12)。

表4 MGAN网络中不同组成部分的有效性

Table 4 The effectiveness of different components on MGAN networks

方 法	voc 07		MS-COCO		
	Resnet-101	Resnet-cut	Resnet-cut		
	mAP	mAP	mAP	CF1	OF1
MGAN-SLAP	94.1	94.9	83.2	77.3	80.7
MGAN-FLAP	94.2	95.0	82.5	77.6	81.1
MGAN-Both	94.0	94.9	83.7	78.2	81.5

MGAN方法中的SLAP强调的是标签的语义相关性对多标签图像分类的作用,而FLAP强调的是空间区域相关性对多标签图像分类的作用。如表4所示,在PASCAL VOC 2007数据集上,加入空间区域相关性的MGAN模型性能要强于加入标签语义相关性的MGAN模型性能,并且当把两者都加入网络时,模型的平均精准率甚至有所下降。其中的原因是VOC数据集较为简单,类别数较少,导致标签的语义相关性作用没有充分展现。

在MS-COCO数据集上做了类似的消融实验,并采用mAP、CF1、OF1作为评价指标。如表4所示,在MS-COCO数据集上加入空间相关性的MGAN网络优于加入语义相关性的MGAN网络,并且将二者同时加入网络效果最好。MS-COCO数据集的复杂性和较多的标签种类数导致语义相关性的作用更加显著。

表5为在PASCAL VOC 2007数据集上验证图卷积层的不同重数和层数对实验结果的影响,基线网络Resnet-101[3]分别在ImageNet[27]和经过cutmix[28]数据处理后的ImageNet[27]数据集上提前预训练。实验结果显示:图卷积层的重数对MGAN模型的影响较小,图卷积层层数在增加到两层后有下降趋势,图卷积层层数为两层,重数为两重时为最好的实验设置。

表5 图卷积层重数与层数对实验结果的影响

Table 5 The influence of graph convolution layer multiplicity and layer number on experimental results

网 络	重数	层数		
		1层	2层	3层
MGAN	2重	92.3	94.0	93.9
(Reweighted)	3重	92.6	94.0	94.0
MGAN-cut	2重	94.4	94.9	94.5
(Reweighted)	3重	94.3	94.7	94.6

4 结束语

本文提出一种融合标签间强相关性的多标签图像分类方法MGAN。将不同相关性强度的标签语义相关性融入卷积神经网络中,用于提高多标签图像分类性能;同时,基于注意力机制将图像区域空间相关性也引入网络结构中。通过实验证明:加入空间相关性和语义相关性的MGAN网络模型优于其他主流多标签图像分类网络性能,且二者优于不加空间相关性和语义相关性的基线网络性能。此外,本文改进了传统的将标签与特征融合视为预测标签的方法,采用将标签与特征融合视为注意力分数的方法,使得图卷积层对网络模型的影响有所降低。通过实验证明该方法使得图卷积层的过平滑问题对MGAN方法的影响较小,对多标签图像分类方法的研究具有一定参考价值。

参考文献(References):

[1] 朱润笋. 面向近视性眼底疾病的图像分类算法研究[D]. 重庆: 重庆邮电大学, 2021.
ZHU Run-sun. Research on image classification algorithm for myopic funfus diseases[D]. Chongqing: Chongqing University of Posts and Telecommunications, 2021.

[2] 杨敏航, 陈龙, 刘慧, 等. 基于图卷积网络的多标签遥感图像分类[J]. 计算机应用研究, 2021, 38(11): 3439—3445.
YANG Min-hang, CHEN Long, LIU Hui, et al. Multi-label remote sensing image classification based on graph convolutional network[J]. Application Research of Computers, 2021, 38(11): 3439—3445.

[3] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770—778.

[4] HUANG G, LIU Z, VAN D M L, et al. Densely connected convolutional networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700—4708.

[5] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for largescale image recognition[C]//In International Conference on Learning Representations(ICLR), 2015: 1—14.

[6] ANTONUCCI A, CORANI G, MAUA D D, et al. An ensemble of Bayesian networks for multi label classification[C]// Twenty-Third International Joint Conference on Artificial Intelligence. 2013.

[7] LI Q, QIAO M, BIAN W, et al. Conditional graphical lasso for multi-label image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2977—2986.

[8] WANG J, YANG Y, MAO J, et al. CNN-RNN: a unified framework for multi-label image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2285—2294.

[9] CHEN Z M, WEI X S, WANG P, et al. Multi-label image recognition with graph convolutional networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5177—5186.

[10] NGUYEN H D, VU X S, LE D T. Modular graph transformer networks for multi-label image classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(10): 9092—9100.

[11] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//In International Conference on Learning Representations(ICLR), 2017.

[12] ZITNICK C L, DOLLAR P. Edge boxes: locating object proposals from edges[C]// European Conference on Computer Vision. Cham: Springer, 2014: 391—405.

[13] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154—171.

[14] WEI Y, 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, 2015, 38(9): 1901—1907.

[15] LI Y, HUANG C, LOY C C, et al. Human attribute recognition by deep hierarchical contexts [C]//European Conference on Computer Vision. Cham: Springer, 2016: 684—700.

[16] WANG M, LUO C, HOMG R, et al. Beyond object proposals: random crop pooling for multi-label image recognition[J]. IEEE Transactions on Image Processing, 2016, 25(12): 5678—5688.

[17] ZHU F, LI H, OUYANG W, et al. Learning spatial regularization with image-level supervisions for multi-label image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5513—5522.

[18] XU J, TIAN H, WANG Z, et al. Joint input and output space learning for multi-label image classification[J]. IEEE Transactions on Multimedia, 2020, 23: 1696—1707.

[19] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961—2969.

[20] LI P, CHEN P, XIE Y, et al. Bi-modal learning with channel-wise attention for multi-label image classification[J]. IEEE Access, 2020(8): 9965—9977.

[21] ZHENG H, FU J, MEI T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition [C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5209—5217.

[22] HOCHREUTER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735—1780.

[23] VANSWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998—6008.

[24] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context[C]// European Conference on Computer Vision. Cham: Springer, 2014: 740—755.

[25] EVERINGHAM M, VAN 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.

[26] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532—1543.

[27] DENG J, DONG W, SOCHER R, et al. Imagenet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. LEEE, 2009: 248—255.

[28] YUN S, HAN D, OH S J, et al. Cutmix: regularization strategy to train strong classifiers with localizable features [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6023—6032.

责任编辑:李翠薇