

面向多标签小样本学习的双流重构网络

方仲礼 王喆 迟子秋

华东理工大学信息科学与工程学院 上海 200237

(434383537@163.com)

摘要 多标签图像分类问题是计算机视觉领域的重要问题之一,它需要对图像中的所有标签进行预测。而一幅图像中待分类的标签个数往往不止一个,同时图像中对象的大小、位置和姿态的变化都会对模型的性能产生影响。因此,如何有效地提高图像特征的准确表达能力是一个亟需解决的难题。针对上述难题,文中提出了一个新颖的双流重构网络来对图像进行特征抽取。具体而言,该模型首先应用一个双流注意力网络来对图像进行基于通道信息和空间信息的特征提取,并经过特征拼接使得图像特征同时兼顾通道特征细节信息和空间特征细节信息。其次,该模型引入了重构损失函数,对双流网络进行特征约束,迫使上述两种分歧特征具有相同的特征表达能力,以此促使提取的双流特征共同向真值特征逼近。在基于 VOC 2007 和 MS COCO 多标签图像数据集上的实验结果表明,所提出的双流重构网络能够准确有效地提取出显著特征,并产生更好的分类精度。同时,鉴于重建损失对模型的解拟合作用,将该方法应用在小样本场景上,实验结果显示,所提模型对小样本数据同样具有较好的分类精度。

关键词: 多标签图像识别;特征重构;深度学习;小样本学习;图像注意力机制

中图分类号 TP183

Dual-stream Reconstruction Network for Multi-label and Few-shot Learning

FANG Zhong-li, WANG Zhe and CHI Zi-qiu

School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

Abstract The multi-label image classification problem is one of the most important problems in the field of computer vision, which needs to predict and output all the labels in an image. However, the number of labels to be classified in an image is often more than one, and the changeable size, posture, and position of objects in the image will increase the difficulty of classification. Therefore, how to effectively improve the accurate expression ability of image features is an urgent problem to be solved. In response to the above-mentioned problem, a novel dual-stream reconstruction network is proposed to extract features from images. Specifically, the model first proposes a dual-stream attention network to extract features based on channel information and spatial information, and uses feature stitching to make image features have both channel detail information and spatial detail information. Secondly, a reconstruction loss function is introduced to constrain the features of the dual-stream network, forcing the above two divergent features to have the same feature expression ability, thereby promoting the extracted dual-stream features to approach the ground-truth features. Experimental results on multi-label image datasets based on VOC 2007 and MS COCO show that the proposed dual-stream reconstruction network can accurately and effectively extract salient features and produce better classification accuracy. At the same time, in view of the sparse effect of reconstruction loss on model features, the proposed method is also applied to few-shot learning. The experimental results show that the proposed model also has good classification accuracy for few-shot learning.

Keywords Multi-label image recognition, Feature reconstruction, Deep learning, Few-shot learning, Image attention mechanism

1 引言

图像分类作为计算机视觉的一个基本问题,已经引起了人们的广泛关注。在过去的几年里,由于信息全球化的加速,大规模的图像^[1]已经变得广泛可用。这使得深度学习得以快

速发展,并显著提高了单标签图像的分类性能。但现实生活中的图像通常包含大量的信息,比如一幅常见的风景画中通常会包含白云、草地、河流、蓝天等标签属性,而单标签图像分类方法无法对其进行分类。由于这一现实问题的需要,人们逐渐将注意力转向多标签分类领域。

到稿日期:2020-11-23 返修日期:2021-03-27

基金项目:上海市科技计划项目(20511100600);国家自然科学基金(62076094)

This work was supported by the Shanghai Science and Technology Program(20511100600) and National Natural Science Foundation of China(62076094).

通信作者:王喆(wangzhe@ecust.edu.cn)

多标签分类是为了解决一幅图像与多个标签相关联的情况而提出的。鉴于深度学习^[2]在图像领域的巨大成功,该方法也被逐步应用于多标签图像分类^[3]。虽然基于深度学习的多标签图像的分类性能得到了较大提高,但仍然存在很多问题亟需解决:首先,多标签图像中的目标数量通常是不确定的。因此,我们需要独立预测图像中可能存在的每个标签。此外,图像中还有物体位置、角度甚至姿态的变化,这也是单标签图像中物体所不具备的特征。其次,由于一幅图中需要容纳多个对象,部分对象往往存在被遮盖或者对象所占比例过小的问题,而单标签图像的待识别对象通常位于图像中心且清晰可见。因此,多标签分类是一项具有挑战性的视觉任务。

鉴于深度神经网络在单标签图像识别任务中的巨大成功,人们将其引入到多标签领域中。例如,Wei等^[4]提出通过提取假设对象框的方法将多标签问题转化为单标签问题,解决了每张图像与多个标签关联的场景问题;Wang等^[5]提出了一个统一的CNN-RNN框架,利用RNNs模型学习标签之间的语义依赖关系,提高了预测精度。

随着注意力机制的快速发展,其在图像字幕、视觉跟踪^[6]、图像问题回答^[7]、目标检测^[8]与语义分割^[9]等许多视觉任务中被证明是有益的。近年来,注意力机制已被用于多标签图像识别的研究,但大多数研究仍处于初级阶段。Zhu等^[10]利用注意力机制完成多标签图像识别。Wang等^[11]提出了一种通过动态注意力机制过滤特征图并使解码器模块聚焦于图像局部区域的方法。他们都采用注意力机制对特征图的空间信息进行过滤,使网络更多地关注重要信息而忽略无关信息。然而,这些方法大多采用单层注意力机制,没有考虑到通道信息对图像特征的重要性,同时也忽略了多视角下的特征信息对图像分类的影响。本文考虑到基于空间信息的注意力特征抽取和基于通道特征信息的注意力特征抽取对多标签图像识别的不同影响,提出了一种基于双流注意力机制的多标签图像识别网络。同时,为了约束上述不同视角下提取出的分歧特征,我们还基于该双流网络引入了一个新的重构损失,从而促使抽取的双流特征向真值特征逼近,如图1所示,以此最大限度地提高卷积网络对初始特征的卷积抽取能力,从而提高网络的分类性能。

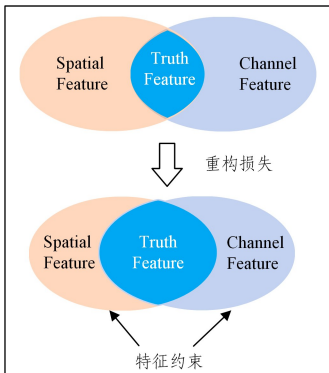


图1 重构损失对分歧特征的约束

Fig. 1 Constraints of reconstruction loss on divergent features

本文在VOC 2007和MS COCO两个主流的多标签数据集上开展了对所提方法的相关研究与讨论。实验结果表明,

本文提出的方法有效提高了网络对多标签图像的分类性能。此外,鉴于重建损失对网络解拟合能力的提高,我们还额外评估了本文模型在小样本场景下的泛化性能,实验结果表明,该方法在小样本学习下也拥有良好的分类性能。

2 相关工作

随着深度学习的快速发展和广泛应用,神经网络在多标签图像领域的应用变得更加多样化。在多标签图像发展的早期,由于缺乏图像处理方法,常用的方法是通过手工进行特征提取,并使用传统的分类方法进行标签预测。然而,随着ImageNet^[1],MS COCO^[12],PASCAL VOC^[13],ML-Images^[14]等大规模手工标注数据集的快速发展,基于深度网络的图像分类模型的性能得到了迅速提高。对于计算机视觉中的各种任务,如图像分类、图像标注、自动问答和目标检测等,许多研究者提出了性能优异的深度学习网络模型。下面我们将重点回顾这些多标签领域的研究。

由于注意机制所带来的优异性能,注意机制在多个领域得到了广泛应用。注意机制可以动态地为图像的不同区域分配不同的权值,并能更好地拟合不规则目标。因此,注意机制可以达到类似于人眼聚焦的效果。Zhu等^[10]提出了一种空间正则化网络,从仅含有图像级标签注释的注意图中学习语义标签与空间区域的关系。Guo等^[15]提出了一种分支网络,以保证在特殊情况下图像仍能保持视觉注意区域的一致性,从而提高多标签图像分类的精度。Luo等^[16]提出了一种新型双流神经网络,将传统的分类模型和显著性预测模型相结合,提高了多标签图像分类的性能。但是,上述方法都忽略了通信信息对特征的显著性影响。本文通过提出的双流网络分别对图像提取出基于空间注意力机制的特征和基于通道注意力机制的特征,并通过同特融合层充分结合了上述两个分歧特征的优点,从而进一步突出有效特征的显著性。

除了在模型结构上的改进,近年来还有许多研究显示,多标签学习算法在别的层面上同样能够对网络分类性能做出提高。这些多标签学习算法大多是为了优化现有的多标记学习损失而提出的。例如,Dembczyński等^[17]证明了估计单标签和多标签后验分布的方法分别是针对汉明损失和子集精度量身定制的,如二分类相关方法优化了汉明损失,Nam等^[18]则优化了子集精度。同时,Wang等^[19]优化了成对标签间关系,而Decubber等^[20]对F1值进行了函数优化。但是,上述所有的损失都是从优化特征预测结果或者标签排序出发,并没有考虑到损失函数对网络的特征提取的优化效果。本文创新性地引入了重构损失来优化网络的特征提取过程,并通过引入的重构损失,实现了特征之间的相互约束,进而迫使双流网络抽取得出的分歧特征共同向真值特征逼近,从而提高了网络的分类性能。

3 模型框架

3.1 问题定义

给定一组带标签的图像 $X = \{x_1, x_2, \dots, x_L\}$,其中 L 为图像集合的总数,多标签学习尝试学习每一幅图像 x 中包含的所有可能的正向标签。假设第 i 个图像 x_i 中对应的标签集为

$Y_i = [y_i^1, y_i^2, \dots, y_i^C]^T$, 其中 y_i^l 表示一个二元指标。令 I 表示具有真值标签的图像, 因此, 如果图像 I 被标签 l 标记为 1, 则 $y_i^l = 1$, 反之 $y_i^l = 0$ 。此外, C 表示数据集中所有标签的类别数。

3.2 总体框架

本文方法的总体框架如图 2 所示。我们设计了一个具有双流注意力机制的识别网络, 并引入了一个新的重构损失, 以对特征进行额外的约束。具体而言, 我们设计的识别网络有

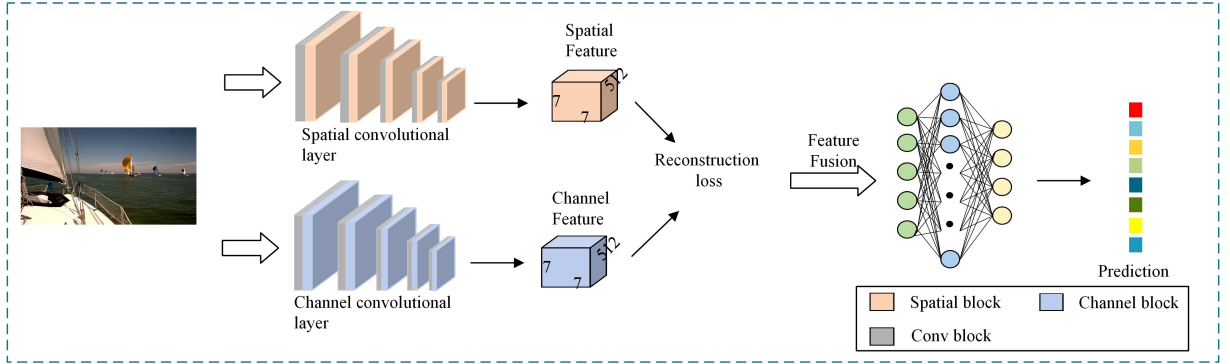


图 2 基于空间和通道注意力机制的双流重构网络的总体框架

Fig. 2 Overall framework of dual-stream reconstruction network based on spatial and channel attention mechanism

3.3 双流注意力网络

该双流网络选择 VGG-16 作为其卷积块, 参考 VGG 网络模型框架, 本文的双流注意力网络模型中每一个分支都有 5 个卷积块, 对应 VGG-16 中的 5 层网络模型。每当特征经过 1 层卷积块, 该模型都会将生成的特征送入相对应的特征优化模块, 如输入到空间分支网络的特征将通过空间卷积模块进行特征优化, 而处于通道卷积层的特征则输入至对应的通道卷积模块进行特征优化, 以此多角度地捕获特征的关键信息。具体来说, 令图像 x 中提取的特征 $F \in \mathbb{R}^{C \times H \times W}$, 一维通道特征注意力图 $C_a \in \mathbb{R}^{C \times 1 \times 1}$, 二维空间特征注意力图 $S_a \in \mathbb{R}^{1 \times H \times W}$, 其中 H 表示特征高度, W 表示特征宽度, C 表示图像特征的通道维度。因此, 基于空间注意力的分支特征为:

$$F_s = S_a(F') \otimes F' + F \quad (1)$$

基于通道注意力的分支特征为:

$$F_c = C_a(F') \otimes F' + F \quad (2)$$

其中, $F' \in \mathbb{R}^{C \times H \times W}$ 表示特征 $F \in \mathbb{R}^{C \times H \times W}$ 经过卷积核转化后的权重特征图。

随后通过特征融合手段将上述两个分支特征进行拼接, 并送入全连接层输出最后的预测结果。

3.4 空间特征卷积注意模块

首先, 该模型的第一个卷积块将输入的图像转化为图像特征, 通道特征由 3 通道卷积为 64 通道, 图像大小变更为 224×224 , 随后将特征 $F \in \mathbb{R}^{64 \times 112 \times 112}$ 输入对应的空间卷积层, 其网络细节如图 3 所示。

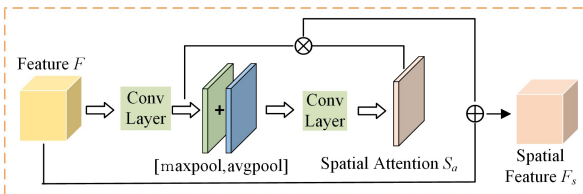


图 3 空间注意力网络

Fig. 3 Spatial attention network

两个分支, 第一个分支用来提取基于空间信息的特征, 第二个分支用来提取基于通道信息的特征。通过上述分支网络的设计, 可以充分利用特征的空间细节信息和通道细节信息。其次, 我们引入了一个新的重构损失, 通过对双流网络提取的分歧特征进行额外的特征约束, 迫使双流网络的分歧抽取不断向真值特征逼近, 从而提高有效特征的显著性, 进而提高网络的标签预测性能。

特征 F 通过两个 3×3 卷积块对初始特征进行特征重排, 生成特征 F' , 其公式如下:

$$F' = conv(F) \quad (3)$$

其中, $conv(\cdot)$ 表示两个 3×3 卷积层模块。

然后分别对上述特征做基于通道的最大池化和平均池化, 生成特征 $F_{spatial}^{max} \in \mathbb{R}^{1 \times H \times W}$ 和特征 $F_{spatial}^{avg} \in \mathbb{R}^{1 \times H \times W}$, 其公式如下:

$$F_{spatial}^{max} = max\ pool(F') \quad (4)$$

$$F_{spatial}^{avg} = avg\ pool(F') \quad (5)$$

然后将它们连接起来并进行卷积, 通过 Sigmoid 函数生成 2D 空间注意力图 S_a , 其公式如下:

$$S_a = \sigma(f_{7 \times 7}[F_{spatial}^{avg}; F_{spatial}^{max}]) \quad (6)$$

最后, 将特征 F' 与 2D 空间注意力图 S_a 进行元素级点乘, 并与原特征相加生成最后的空间特征 F_s :

$$F_s = S_a(F') \otimes F' + F \quad (7)$$

其中, $\sigma(\cdot)$ 表示 Sigmoid 函数, $f_{7 \times 7}$ 表示带有 7×7 卷积核的卷积操作, $S_a(F')$ 表示空间注意权重图 S_a 基于 F' 产生。

3.5 通道特征卷积注意模块

本文的通道特征处理方法与空间注意力图类似, 首先将输入图像转化为图像特征, 图像大小变更为 224×224 , 随后将特征 $F \in \mathbb{R}^{64 \times 112 \times 112}$ 输入对应的通道卷积层, 其网络细节如图 4 所示。特征 F 通过两个 3×3 卷积块对初始特征进行特征重排, 生成特征 F' , 其公式如下:

$$F' = conv(F) \quad (8)$$

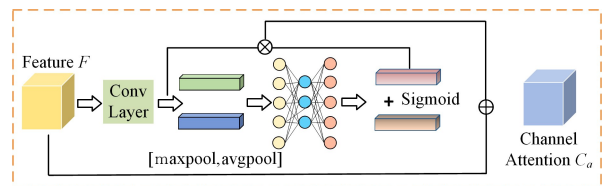


图 4 通道注意力网络

Fig. 4 Channel attention network

分别对上述特征做基于 W 和 H 的最大池化和平均池化,生成特征 $F_{channel}^{max} \in \mathbb{R}^{C \times 1 \times 1}$ 和特征 $F_{channel}^{avg} \in \mathbb{R}^{C \times 1 \times 1}$,其公式如下:

$$F_{channel}^{max} = \max pool(F') \quad (9)$$

$$F_{channel}^{avg} = \text{avg pool}(F') \quad (10)$$

然后,将其输入全连接层进行权重的转化,随后将它们相加,并通过 Sigmoid 函数生成 1D 通道特征注意力图 C_a :

$$C_a = \sigma(MLP(F_{channel}^{avg}) + MLP(F_{channel}^{max})) \quad (11)$$

最后,将特征 F' 与 1D 通道特征注意力图 C_a 进行元素级点乘,并与原特征相加生成最后的通道特征 F_c :

$$F_c = C_a(F') \otimes F' + F \quad (12)$$

其中, $MLP(\cdot)$ 表示全连接层, $C_a(F')$ 表示通道注意力权重图 C_a 基于 F' 产生。

3.6 现有损失函数

本节将讨论现有的损失函数对多标签模型的意义,并简要讨论上述损失函数与本文损失函数的区别。

汉明损失^[21]的公式如下:

$$HammingLoss = \frac{1}{N} \sum_{i=1}^N \frac{x \text{ or } (Y_{i,j}, P_{i,j})}{L} \quad (13)$$

其中, N 是样本的数量, L 是标签的个数, $Y_{i,j}$ 是第 i 个预测结果中第 j 个分量的真实值, $P_{i,j}$ 是第 i 个预测结果中第 j 个分量的预测值, XOR 是抑或函数。

排序损失^[22]的公式如下:

$$RankingLoss = \frac{1}{N} \sum_{i=1}^N \frac{1}{|Y_i| | \bar{Y}_i |} | \{ (y', y'') \mid f(x_i, y') \leq f(x_i, y''), (y', y'') \in Y_i \times \bar{Y}_i \} | \quad (14)$$

其中, Y_i 为对应标签真值为 1 的标签集, \bar{Y}_i 为标签真值为 0 的标签集, $f(\cdot)$ 表示模型的预测结果。

One-error^[23]的公式如下:

$$One-error = \frac{1}{N} \sum_{i=1}^N [[\arg \max_{y \in Y} f(x_i, y)] \notin Y_i] \quad (15)$$

其中, $\arg \max_{y \in Y} f(x_i, y)$ 表示排名最高且不在相关标签集中的标签的分数。

以往的图像损失函数大多关注模型预测结果的标签内部的排序或者是标签真值与模型预测标签之间的差异,然而这些函数都是基于模型的标签输出与真值标签之间的对比而设计的损失函数。目前,鲜有研究能够考虑为特征设置一个专门的损失函数,用来提高图像特征的显著性,从而提高模型预测的准确性。

3.7 基于特征重构的损失函数

本文针对图像特征的特点,为多标签网络引入了一个新的损失函数,基于该损失函数的特性,我们称之为基于特征重构的损失函数。

特征是图像的一种抽象表示,因此,特征可以称为抽象的图像。因为模型总是倾向于从图像中提取出最能表达图像关键信息的显著特征并从中识别出正确的标签结果 $P_{i,j} = Y_{i,j}$ 。因此在理想状态下,我们可以认为特征 $F \in \mathbb{R}^{C \times H \times W}$ 中包含着真值特征 $F_T \in \mathbb{R}^{C \times H \times W}$ 。而在双流注意力网络中,通过不同的特征抽取方式,获得的分歧特征 F_c 与 F_s 都能够识别

出相同的结果 $P_{i,j}$,因此我们可以认为分歧特征 F_c 与 F_s 具有相同的真值特征 $F_T \in \mathbb{R}^{C \times H \times W}$,其概念如图 5 所示。因此,我们通过引入新的损失函数来对特征进行约束,从而促使分歧特征向真值特征逼近。其损失函数如下:

$$L_{Reconstruction} = \frac{1}{P} \sum_{p=1}^P \left(\frac{1}{M} \sum_{c=1}^M |F_{s,c}^1 - F_{s,c}^2| \right) \quad (16)$$

其中, $F^1 \in \mathbb{R}^{C \times H \times W}$ 和 $F^2 \in \mathbb{R}^{C \times H \times W}$ 为相同维度的分歧特征, M 表示 F^1 特征的最大通道数, P 表示特征的空间像素点。

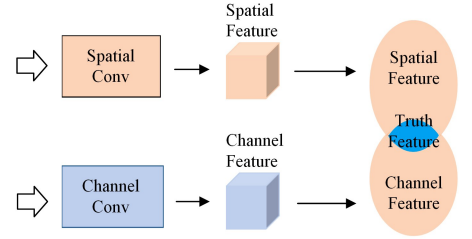


图 5 分歧特征之间的真值特征

Fig. 5 Truth feature of different features

基于上述特征损失函数的构建,本文模型会不断要求空间特征 F_s 以及通道特征 F_c 互相逼近,以此要求两者基于不同方法所抽取的特征图逐步向真值特征靠拢。本文双流网络同时兼顾了特征的时空信息和通道信息,从而优化了网络参数,达到了更高的分类预测效果。

4 数据集介绍

4.1 VOC 2007 数据集

PASCAL VOC 2007 (PASCAL Visual Object Classes Challenge)数据集被广泛用作多标签分类的基准数据集。它有 9963 张图像,标签数据共有 20 个类,每张图像都与一个或多个标签相关联。该数据集分为训练集、验证集和测试集。为了与其他算法保持一致,我们使用训练集(5 011 张 VOC 2007 图片)进行训练,使用测试集(4 952 张 VOC 2007 图片)进行性能测试。

4.2 MS COCO 数据集

MS COCO (Microsoft Common Objects in COtext)^[12]是一个非常流行的大型数据集,其中包含 82 783 张用于训练的图像和 40 504 张用于测试的图像。它经常被用作分类、对象检测、分割和看图说话的基准数据集。MSCOCO 总共有 80 个类,每幅图像与一个或多个标签相关联。本文选择 MS-COCO 来评估本文方法是否可以用于更大的数据集和更多的类别,以证明本文模型的泛化效果。

5 实验设置

5.1 实验设置

我们选择在 Pytorch 框架中对提出的深度模型进行训练和测试。具体地,我们采用基于 ImageNet 2012 分类挑战数据集进行预训练的 VGG-16^[24]作为所提双流网络的基模型。VGG-16^[24]首先在 ImageNet^[1]上进行预训练,然后在 VOC 2007^[13]和 MS COCO^[12]目标数据集上进行二次训练来微调所提模型。此外,为了体现出本文重构损失对网络带来的提

升,我们首先使用普通的交叉熵损失函数对模型进行微调,然后在此基础上再加入本文所引入的重构损失函数进行再次训练。具体来说,首先对所提双流网络的基模型进行参数固定,然后通过新增的重构损失函数对空间注意层和通道注意层进行微调,最后通过交叉熵损失对整个网络进行联合调整。所有的训练图像均被调整为 256×256 大小,并随机裁剪为 224×224 大小,同时通过水平翻转来增加训练样本。

5.2 评价指标

我们选择 (macro/micro) precision, recall, F1-measure 和 mean Average Precision (mAP) 评价指标来讨论本文方法的有效性。我们将每个类的平均精度缩写为 C-P, 将 micro precision 记为 O-P, 它表示对所有类的所有图像的真实预测进行计算的总体度量。同样,我们也可以评估 macro/micro Recall (C-R/O-R) 和 macro/micro F1 (C-F1/O-F1)。如果这些标签的信任度大于 0.5, 则这些标签被预测为正值。为了公平地比较本文算法与其他算法,我们将在相同的实验条件下评估各个模型的性能。

表 1 各种方法在 PASCAL VOC 2007 数据集上的分类结果

Table 1 Comparisons of classification results of each method on PASCAL VOC 2007 dataset

| | plane | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | mAP |
|------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| INRIA ^[25] | 77.2 | 69.3 | 56.2 | 66.6 | 45.5 | 68.1 | 83.4 | 53.6 | 58.3 | 51.1 | 62.2 | 45.2 | 78.4 | 69.7 | 86.1 | 52.4 | 54.4 | 54.3 | 75.8 | 62.1 | 63.5 |
| AGS ^[26] | 82.2 | 83.0 | 58.4 | 76.1 | 56.4 | 77.5 | 88.8 | 69.1 | 62.2 | 61.8 | 64.2 | 51.3 | 85.4 | 80.2 | 91.1 | 48.1 | 61.7 | 67.7 | 86.3 | 70.9 | 71.1 |
| AMM ^[27] | 84.5 | 81.5 | 65.0 | 71.4 | 52.2 | 76.2 | 87.2 | 68.5 | 63.8 | 55.8 | 65.8 | 55.6 | 84.8 | 77.0 | 91.1 | 55.2 | 60.0 | 69.7 | 83.6 | 77.0 | 71.3 |
| Multi-VGG | 95.4 | 92.4 | 94.3 | 91.8 | 58.8 | 82.2 | 93.1 | 91.7 | 66.8 | 79.5 | 78.5 | 90.3 | 94.7 | 88.6 | 96.4 | 71.2 | 83.3 | 70.9 | 96.6 | 83.5 | 85.0 |
| CNN-RNN ^[4] | 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 | 84.0 |
| HCP-1000C ^[3] | 95.1 | 90.1 | 92.8 | 89.9 | 51.5 | 80.0 | 91.7 | 91.6 | 57.7 | 77.8 | 70.9 | 89.3 | 89.3 | 85.2 | 93.0 | 64.0 | 85.7 | 62.7 | 94.4 | 78.3 | 81.5 |
| HCP-2000C ^[3] | 96.0 | 92.1 | 93.7 | 93.4 | 58.7 | 84.0 | 93.4 | 92.0 | 62.8 | 89.1 | 76.3 | 91.4 | 95.0 | 87.8 | 93.1 | 69.9 | 90.3 | 68.0 | 96.8 | 80.6 | 85.2 |
| Attend and Imagine ^[28] | 97.0 | 92.5 | 93.8 | 93.3 | 59.3 | 82.6 | 90.6 | 92.0 | 73.4 | 82.4 | 76.6 | 92.4 | 94.2 | 91.4 | 95.3 | 67.9 | 88.6 | 70.1 | 96.8 | 81.5 | 85.6 |
| DSN (ours) | 95.9 | 93.4 | 94.9 | 92.3 | 59.5 | 83.7 | 93.2 | 92.9 | 69.4 | 83.9 | 80.5 | 91.7 | 95.3 | 90.4 | 97.0 | 71.9 | 86.3 | 75.8 | 96.9 | 83.7 | 86.4 |
| DSRN (ours) | 96.5 | 93.8 | 95.0 | 93.3 | 61.2 | 86.5 | 93.9 | 93.8 | 69.9 | 84.9 | 81.4 | 93.0 | 95.7 | 91.7 | 97.3 | 72.9 | 88.5 | 78.5 | 97.5 | 84.1 | 87.5 |

(单位: %)

由表 1 可知,所提双流重构网络——DSRN(Dual-Stream Reconstruction Network) 取得了最优性能,同时为了体现本文重构损失所带来的提升,我们还做了消融实验,表 1 中的 DSN(Dual-Stream Network) 是指模型仅存在双流注意力且没有重构损失约束的网络。实验结果表明,通过引入重构损失函数,本文方法的 mAP 提高了大约 0.8%, 且 DSRN 中的每一类相比 DSN 都有较大的提高。

同时,通过与其他方法的对比发现,本文模型较当前的主流方法有所提高,并在大部分单类精度上处于最优性能。其中, CNN-RNN^[4] 在 cat 类和 sheep 类中取得最优精度,这可能得益于该网络设计的语义关系模型能够较好地对待识别对象进行标签语义上的关联,从而增强关联标签的识别率,但是 DSRN 在其他类别上都有较大的性能优势。这说明了通过引入基于特征的重构损失,本文模型确实利用特征约束方式迫使双流网络的分歧特征不断向真值特征靠拢,从而使网络在整体分类性能上都有较大的提高。

6.2 MS COCO 数据集上的实验结果

本文在 MS-COCO^[12] 数据集上对各方法的分类性能进行对比实验,结果如表 2 所列。所选对比算法有 CNN-RNN^[4], RLSD^[29] 和 WARP^[30], 其中 RLSD^[29] 可以通过提取与标签具

6 实验结果分析

6.1 PASCAL VOC 2007 数据集上的实验结果

首先,我们在 PASCAL VOC 2007^[13] 上计算这些方法的总精度和每个类的精度。除了基模型 VGG-16^[24] 和自身的消融实验外,我们还将本文方法与其他常用方法进行了比较,包括 INRIA^[25], AGS^[26], AMM^[27], HCP^[3], CNN-RNN^[4] 和 Attend and Imagine^[28]。INRIA^[25] 是一种对象定位和分类的组合方法。AGS^[26] 引入了基于子类感知的对象分类框来提高对象的分类性能。AMM^[27] 通过相互输入其他任务的结果来提高模型性能。HCP-1000C/HCP-2000C^[3] 倾向于在图像中提取假设区域,然后对每个假设区域进行分类,最后利用 cross-hypothesis max-pooling 融合所有假设区域的分类结果,得到整个图像的标签。与上述方法不同的是, CNN-RNN^[4] 倾向于学习词嵌入之间的语言关联来学习图像标签之间的关联性。表 1 列出了本文方法与其他最新方法在 PASCAL VOC 2007 数据集上的比较结果。

有高度关联性的区域来进行标签的分类,而 WARP^[30] 则通过结合卷积结构和传统的标签排序方法来优化结果。

表 2 各种方法在 MS COCO 数据集上的分类结果

Table 2 Comparisons of classification results of each method on MS

COCO dataset

| Method | C-P | C-R | C-F1 | O-P | O-R | O-F1 | mAP |
|------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| CNN-RNN ^[4] | 66.0 | 55.6 | 60.4 | 69.2 | 66.4 | 67.8 | 61.2 |
| WARP ^[30] | 59.3 | 52.5 | 55.7 | 59.8 | 61.4 | 60.7 | — |
| Multi-CNN ^[29] | 54.8 | 51.4 | 53.1 | 56.7 | 58.6 | 57.6 | 60.4 |
| Attend and Imagine ^[28] | — | — | — | 59.1 | 71.9 | 64.9 | 64.7 |
| RLSD ^[29] | 67.7 | 56.4 | 61.5 | 70.5 | 59.9 | 64.8 | 67.9 |
| DSN (ours) | 78.6 | 45.8 | 57.8 | 84.8 | 51.1 | 63.7 | 66.1 |
| DSRN(ours) | 80.4 | 48.5 | 60.5 | 86.7 | 53.7 | 66.3 | 68.9 |

(单位: %)

表 2 中, C-P, C-R, C-F1 等评价指标被用来作为模型性能的度量指标。在 MS COCO^[12] 数据集上, DSRN 的分类性能相比 DSN 模型仍有较大的提升,并达到了前沿水平。从表 2 可以看出, DSRN 的每一个评估指标的数值都要优于 DSN, 这与在 PASCAL VOC 2007^[13] 数据集上的实验结果相同,说明重构损失对模型的提升是全方位的,它优化了特征网络中的大部分参数,从而提高了模型的各方面性能。

同时,通过与其他方法的对比发现,本文模型已经达到了当前最主流的分类性能,并在 O-P 和 C-P 评价指标上远高于其他模型。其中本文模型的 C-P 达到了 80.4%,而其他算法最高为 67.7%;O-P 则达到了 86.7%,其他算法最高为 70.5%。但是本文模型的召回率有一定的下降,可能是因为本文的双流注意网络在一定程度下忽略了小物体的识别,因此召回率较之基模型有一定的降低,但总体 F1 值仍有提高。在此基础上,本文提出的 DSRN 又在一定程度上提高召回率和识别精度,达到了最前沿算法的分类性能。

6.3 面向小样本场景的实验结果

鉴于重建损失对网络解拟合能力的提高,我们还额外评估了本文模型在小样本场景下的泛化性能,其在 PASCAL VOC 2007^[13]数据集上的实验结果如表 3 所列。

表 3 本文方法在小样本场景下的分类结果

Table 3 Classification results of proposed method in few-shot

| | | scenario | | |
|-------|--------|-------------|-------------|-------------|
| Ratio | Method | C-F1 | O-F1 | mAP |
| 1.0 | VGG | 77.1 | 79.4 | 85.0 |
| | DSRN | 79.3 | 81.9 | 87.5 |
| 0.9 | VGG | 76.7 | 79.0 | 85.2 |
| | DSRN | 79.1 | 81.8 | 87.1 |
| 0.8 | VGG | 75.3 | 77.2 | 83.6 |
| | DSRN | 78.1 | 80.6 | 86.3 |
| 0.7 | VGG | 74.5 | 78.6 | 82.5 |
| | DSRN | 77.6 | 81.0 | 85.9 |
| 0.6 | VGG | 75.1 | 77.9 | 82.7 |
| | DSRN | 77.0 | 79.9 | 84.9 |
| 0.5 | VGG | 75.5 | 78.3 | 81.6 |
| | DSRN | 77.2 | 80.3 | 84.8 |
| 0.4 | VGG | 73.9 | 75.8 | 81.6 |
| | DSRN | 77.3 | 80.4 | 84.6 |
| 0.3 | VGG | 72.9 | 75.3 | 80.5 |
| | DSRN | 75.2 | 78.8 | 82.9 |
| 0.2 | VGG | 72.0 | 75.5 | 78.7 |
| | DSRN | 73.8 | 76.9 | 81.3 |
| 0.1 | VGG | 68.1 | 71.9 | 74.9 |
| | DSRN | 70.6 | 75.2 | 77.0 |

表 3 中,Ratio 表示对数据集的采样比例,1.0 表示使用所有的训练集进行训练,而 0.9 表示使用 90% 的训练集进行模型的训练。为了确保实验的有效性,我们在每次欠采样时,都会事先打乱训练集的图像顺序,并采用 C-F1,O-F1 和 mAP 3 种评价指标来全面地检测模型的性能。

从表 3 中可以发现,DSRN 的 mAP 均优于基模型 VGG,并且在 C-F1 和 O-F1 评价指标上的评估性能也高于 VGG。此外,在训练集的采样比例为 70%,50%,40% 时,DSRN 的 mAP 较之 VGG 有大约 3% 的提高,同时在采样比例偏小时也有较大的提高。因此,DSRN 具有较强的鲁棒性,能较好地适应小样本场景下的模型学习。

结束语 多标签图像分类是视觉领域的一个基本问题。本文提出了一种面向多标签小样本学习的双流重构网络。该方法探索了基于特征的损失函数的设计对最终结果的影响,通过所设计的特征损失函数,迫使分歧特征共同向真值特征逼近,从而提取出更准确的真值特征。实验结果表明,该方法具有优异的性能。

将来我们会继续探索更多新式的损失函数在神经网络中

的贡献,并在更多的数据集上进行实验来验证本文方法的有效性。此外,我们还将采用更多特征提取策略,如基于对象大小敏感的双流网络特征提取等,来丰富网络的分歧特征提取手段,从而进一步探索基于特征重建的损失函数的应用场景。

参考文献

- [1] 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. IEEE, 2009: 248-255.
- [2] CHEN T R,LING J. Differential Privacy Protection Machine Learning Method Based on Features Mapping[J]. Computer Science,2021,48(7):33-39.
- [3] WANG Q,JIA N,BRECKON T P. A baseline for multi-label image classification using an ensemble deep CNN[J]. IEEE International Conference on Image Processing(ICIP),2019.
- [4] 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.
- [5] 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.
- [6] YANG T,CHAN A B. Learning dynamic memory networks for object tracking[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018:152-167.
- [7] YANG Z,HE X,GAO J,et al. Stacked attention networks for image question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 21-29.
- [8] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:779-788.
- [9] WANG P,CHEN P,YUAN Y,et al. Understanding convolution for semantic segmentation[C]// 2018 IEEE Winter Conference on Applications of Computer vision (WACV). IEEE, 2018: 1451-1460.
- [10] 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.
- [11] WANG Z,CHEN T,LI G,et al. Multi-label image recognition by recurrently discovering attentional regions[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017:464-472.
- [12] 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.
- [13] 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.

- [14] WU B, CHEN W, FAN Y, et al. Tencent ml-images: A large-scale multi-label image database for visual representation learning[J]. *IEEE Access*, 2019, 7: 172683-172693.
- [15] GUO H, ZHENG K, FAN X, et al. Visual attention consistency under image transforms for multi-label image classification [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019: 729-739.
- [16] LUO Y, JIANG M, ZHAO Q. Visual attention in multi-label image classification[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019.
- [17] DEMBCZYŃSKI K, WAEGEMAN W, CHENG W, et al. On label dependence and loss minimization in multi-label classification [J]. *Machine Learning*, 2012, 88(1/2): 5-45.
- [18] NAM J, MENCÍA E L, KIM H J, et al. Maximizing subset accuracy with recurrent neural networks in multi-label classification [J]. *Advances in Neural Information Processing Systems*, 2017, 30: 5413-5423.
- [19] WANG Y, WANG S, TANG J, et al. Ppp: Joint pointwise and pairwise image label prediction [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016: 6005-6013.
- [20] DECUBBER S, MORTIER T, DEMBCZYŃSKI K, et al. Deep f-measure maximization in multi-label classification: A comparative study [C]// *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Cham: Springer, 2018: 290-305.
- [21] WU X Z, ZHOU Z H. A unified view of multi-label performance measures [C]// *International Conference on Machine Learning*. PMLR, 2017: 3780-3788.
- [22] LI Y, SONG Y, LUO J. Improving pairwise ranking for multi-label image classification [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017: 3617-3625.
- [23] ELISSEEFF A, WESTON J. A kernel method for multi-labelled classification [C]// *Advances in Neural Information Processing Systems*. 2002: 681-687.
- [24] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. *arXiv: 1409. 1556*, 2014.
- [25] HARZALLAH H, JURIE F, SCHMID C. Combining efficient object localization and image classification [C]// *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009: 237-244.
- [26] DONG J, XIA W, CHEN Q, et al. Subcategory-aware object classification [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013: 827-834.
- [27] SONG Z, CHEN Q, HUANG Z, et al. Contextualizing object detection and classification [C]// *CVPR 2011*. IEEE, 2011: 1585-1592.
- [28] LYU F, WU Q, HU F, et al. Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks [J]. *IEEE Transactions on Multimedia*, 2019, 21(8): 1971-1981.
- [29] ZHANG J, WU Q, SHEN C, et al. Multi-label image classification with regional latent semantic dependencies [J]. *IEEE Transactions on Multimedia*, 2018, 20(10): 2801-2813.
- [30] GONG Y, JIA Y, LEUNG T, et al. Deep convolutional ranking for multilabel image annotation [J]. *arXiv: 1312. 4894*, 2013.



FANG Zhong-li, born in 1996, postgraduate, is a member of China Computer Federation. His main research interests include multi-label learning and deep learning.



WANG Zhe, born in 1981, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include pattern recognition and image processing.

(责任编辑:柯颖)