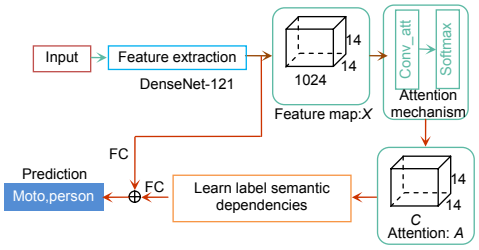
Opto-Electronic Engineering
光 电 工 程

Article

2019 年，第 46 卷，第 9 期

DOI: 10.12086/oee.2019.180468

融合注意力机制和语义关联性的多标签图像分类



薛丽霞，江　迪，汪荣贵，杨　娟*

合肥工业大学计算机与信息学院，安徽 合肥 230009

摘要：卷积神经网络在单标签图像分类中表现出了良好的性能，但是，如何将其更好地应用到多标签图像分类仍然是一项重要的挑战。本文提出一种基于卷积神经网络并融合注意力机制和语义关联性的多标签图像分类方法。首先，利用卷积神经网络来提取特征；其次，利用注意力机制将数据集中的每个标签类别和输出特征图中的每个通道进行对应；最后，利用监督学习的方式学习通道之间的关联性，也就是学习标签之间的关联性。实验结果表明，本文方法可以有效地学习标签之间语义关联性，并提升多标签图像分类效果。

关键词：多标签图像分类；卷积神经网络；注意力机制；语义关联性

中图分类号：TP391　　**文献标志码：**A

引用格式：薛丽霞，江迪，汪荣贵，等. 融合注意力机制和语义关联性的多标签图像分类[J]. 光电工程，2019，**46**(9): 180468

Multi-label classification based on attention mechanism and semantic dependencies

Xue Lixia, Jiang Di, Wang Ronggui, Yang Juan*

School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230009, China

Abstract: Multi-label image classification which is a generalization of the single-label image classification is aimed to assign multi-labels to the image to full express the specific visual concepts contained in the image. We propose a method based on convolutional neural networks, which combines attention mechanism and semantic relevance, to solve the multi label problem. Firstly, we use convolution neural network to extract features. Then, we apply the attention mechanism to obtain the correspondence between the label and channel of the feature map. Finally, we explore the channel-wise correlation which is essentially the semantic dependencies between labels by means of supervised learning. The experimental results show that the proposed method can exploit the dependencies between multiple tags to improve the performance of multi label image classification.

Keywords: multi-label classification; convolution neural network; attention mechanism; semantic dependencies

Citation: Xue L X, Jiang D, Wang R G, *et al.* Multi-label classification based on attention mechanism and semantic dependencies[J]. *Opto-Electronic Engineering*, 2019, **46**(9): 180468

收稿日期：2018-09-10；　收到修改稿日期：2018-12-25

作者简介：薛丽霞(1976-)，女，博士，副教授，硕士生导师，主要从事智能视频处理与分析、视频大数据与云计算、智能视频监控与公共安全、嵌入式多媒体技术等的研究。E-mail：xlxzzm@163.com

通信作者：杨娟(1983-)，女，博士，讲师，硕士生导师，主要从事视频信息处理、视频大数据处理技术、深度学习与二进神经网络理论与应用等的研究。E-mail：yangjuan6985@163.com

1 引 言

随着多媒体技术的发展和推广，各种多媒体数据(例如图像、视频等)迅速成为信息的主流，并对人们的生活和社会的发展产生重要的影响。图像分类作为计算机视觉领域的一个重要分支，通过给图像分配正确合适的标签，将图像的视觉信息转换为语义信息，便于人们更好地理解与分析图像，对多媒体相关技术(例如图像检索、语义分割等)的发展有重要的推动作用。

单标签图像分类作为图像视觉内容理解的传统课题，已有多年的研究，并取得了不错的进展。传统的单标签图像分类方法有词袋模型[1-2]、支持向量机[3-4]、随机森林[5]等。然而，在现实世界中，一幅图像往往包含丰富的语义信息，如多个目标，场景，行为等。因此，多标签图像分类是一个更为普遍和实际的问题，是单标签分类问题的推广，旨在为图像分配多个标签以充分表达图像中所包含的具体内容，其在图像检索、场景识别等计算机视觉领域具有更加广泛的应用。Harzallah 等[6]人提出使用线性支持向量机分类器进行预选，并使用非线性支持向量机进行评分来评估最终的分类结果。但是，传统方法的缺陷在于需要提取手工特征(例如 SIFT[7]、HOG[8]和 LBP[9])对图像进行预处理，并不能充分获取图像高层次的视觉信息。

由于卷积神经网络强大的表征能力，其在单标签图像分类[10-12]任务上取得了突破性的进展，为多标签图像分类任务提供了一定的思路。然而，多标签图像中目标之间存在遮挡、背景复杂、目标不显著以及目标分布不集中等问题，基于卷积神经网络的单标签图像分类方法并不能直接应用到多标签图像分类任务中。处理多标签图像分类任务的一种简单方法是将其转换为多个单标签图像分类任务。Razavian 等[13]首先利用大规模单标签数据集 ImageNet[14]预训练网络模型，然后将该网络模型的参数迁移到多标签网络模型中，并使用网络模型输出的图像特征训练每个标签的支持向量机分类器，该方法取得了一定的效果。Wei 等[15]提出了 HCP 网络模型，该模型利用 BING[16]算法提取一系列图像候选块，并假设每个候选块包含单个目标，然后利用聚类算法从这些候选块中挑选出一定量的候选块作为网络输入。针对每个输入候选块，网络会输出一个分类结果，最后使用类别最大池化方式进行融合得到最终的多标签预测结果。但是，上述方法忽略了图像的多个标签之间存在的关联性。例如，“天空”和“云朵”、“船”和“海洋”常常会出现在同一幅图像中。Wang 等[17]提出 CNN-RNN 模型，使用卷积神经网络提取图像的特征，并利用循环神经网络[18]对标签依赖性进行了明确的建模，取得了很好的效果。Zhang[19]设计了 RLSD 模型，该模型依据区域建议网络(region proposal network，RPN)的思想设计了空间定位层，利用定位层提取图像局部区域特征，随后利用 LSTM(long-short term memory)网络获取该局部区域目标的标签，并整合所有局部区域标签从而获得一幅图像的预测标签。但是，RNN 网络模型在训练时普遍存在收敛速度较慢的问题。

本文提出了一种新的模型来学习标签之间的依赖关系。图 1 展示了本文方法的整体框架图。首先，使

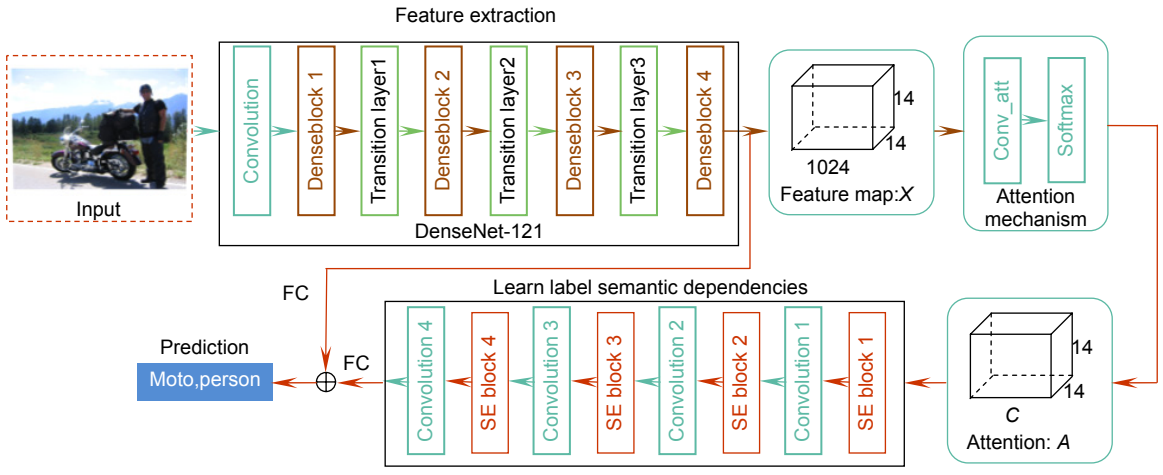


图 1 多标签图像分类整体框架图

Fig. 1 An illustration of the framework for multi-label classification

用 denseNet-121 模型作为特征提取模型，该模型初始参数是通过 ImageNet 数据集预训练得到的，在此基础之上，使用多标签的数据集对网络参数进行微调，以此提取多标签图像特征。其次，为探索标签之间的基本空间关系，利用注意力机制将数据集中的每个标签类别和特征图中的每个通道进行对应。其中，每个通道的响应可能是正响应或者负响应，本文使用一个 SE 模块来消除这些负响应；然后，根据注意力机制得到的每个特征通道，利用卷积操作去学习通道之间的关联性，也就是学习标签之间的关联性，并将其结果与特征提取网络的输出进行融合，得到最终的多标签图像分类结果。

2 融合注意力机制和语义关联性多标签图像分类

2.1 特征提取网络

本文使用 denseNet-121[11]网络模型提取图像特征。该模型的特性为其特征图采用堆叠的过程，即网络每一层的输入都是前面所有层输出的并集，而该层所学习的特征图也会被直接传给其后面所有层作为输入。整个网络是由一系列的密集块和过渡层组成，设第 l 层是接收前层的特征图，$x_0,x_1,\cdots,x_{l-1}$ 作为输入：

$$x_l = H_l([x_0,x_1,\cdots,x_{l-1}]) \quad , \tag{1}$$

其中：$[x_0,x_1,\cdots,x_{l-1}]$ 表示的是将第 0 层到 $l-1$ 层的输出特征图做一个堆叠。$H_l(\cdot)$ 是由两个三个连续操作的复合函数：批归一化(BN)[20]、线性整流函数(ReLU)[21]和卷积(Conv)组成。考虑到堆叠操作随着网络层数的加深使得通道的维度也在不断增长，同时参数的数量增长也呈二次方，因此，使用过渡层对网络进行降维。

本文将 H 表示输入图像，$y=[y^1,y^2,\cdots,y^C]^{\mathrm{T}}$ 表示真实标签，其中 y^l 是一个二进制指标，$y^l=1$ 表示图像中存在 l 标签，相反，$y^l=0$ 表示图像中不存在 l 标签，C 表示数据集的标签数目。网络特征提取方式可表示为

$$X = f_{\mathrm{cnn}}(H,\theta_{\mathrm{cnn}}), X \in R^{14\times14\times1024} \quad , \tag{2}$$

其中 X 表示的是特征提取网络的输出特征图。

2.2 注意力机制

注意力机制模型[22-24]最近被广泛应用于计算机视觉任务，图像分类[22]也是注意力机制模型的重要应用之一。针对多标签图像分类，X 通常融合了多个目标的特征信息，然而，直接从这些特征信息中捕获标签之间的语义关联比较困难。为了更好地探索目标之间的关联性，本文首先使用注意力机制将特征图的通道和数据集中的类别相对应，这意味着不同的特征图通道能够去注意不同的类别。然后利用卷积学习标签之间的语义关联，该部分具体将在 2.3 讨论。

本文将网络提取的图像特征图 $X \in R^{14\times14\times1024}$ 作为注意力网络的输入，为了实现标签和特征图通道的对应，首先利用卷积层(conv_att)去初步学习标签和通道的转换关系，如图 1 中的“ attention mechanism ”部分所示。

$$Z = f_{\mathrm{conv\_att}}(X,\theta_{\mathrm{conv}}), Z \in R^{14\times14\times C} \quad , \tag{3}$$

其中：$f_{\mathrm{conv\_att}}$ 由三层卷积层实现，其维度分别是 $1\times1\times512$、$3\times3\times512$ 和 $1\times1\times C$，其中，C 表示的是多标签数据集标签类别。经过这三层后输出的特征图为 $Z \in R^{14\times14\times C}$，该特征图每个通道响应数据集中的某一个标签，因此通道数等于数据集标签数量。简单地利用卷积操作并不能像我们期待的那样学习到标签和特征图通道的对应关系，参考文献[23]，对 Z 的每个通道利用 softmax 函数归一化操作得到注意力特征图 A：

$$a_{i,j}^l = \frac{\exp(z_{i,j}^l)}{\sum_{i,j}\exp(z_{i,j}^l)}, A \in R^{14\times14\times C} \quad , \tag{4}$$

其中：$z_{i,j}^l$ 和 $a_{i,j}^l$ 分别表示为未经过和经过归一化的对应标签 l 的处于坐标 (i,j) 的响应值。softmax 操作会确保每个通道对每个类存在一个响应，然而给定的图像中只存在数据集中的部分标签，并非每个类都存在于图像中，因此这些通道响应有些是正响应(或者说该响应是图像中某个标签的响应)，其余的为负响应。因此，需要消除那些不存在与图像中的类的响应，才能更好地学习标签之间的语义关联性。

2.3 语义关联分类

注意力机制可以得到每个标签类别其对应学习得到的特征图的输出 A，$A(A \in R^{14\times14\times C})$的每一个通道对应于 $C(C$ 为数据集的标签类别)中的一个类。为了消除一些不存在与图像中的类的响应，本文使用 SE 模块[11]去除这些负响应。

如图 2 所示，SE 模块通过学习的方式获取到每个特征通道的重要程度，然后依照这个重要程度去提升有用的特征并抑制对当前任务用处不大的特征。SE 架构主要分为 Squeeze、Excitation 和 Reweight 操作。Squeeze 操作是对空间维度进行特征压缩，采用全局平均池化方式将每个二维的特征通道 a^l 变成一个实数 Z_l，Z_l 在某种程度上具有全局的感受野：

$$z_l = F_{\mathrm{Sq}}(a^l) = \frac{1}{14\times 14}\sum_{i=1}^{14}\sum_{j=1}^{14} a^l(i,j) \quad 。 \tag{5}$$

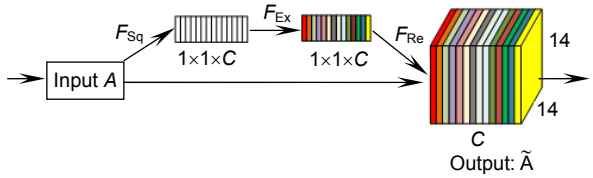


图 2 SE 块

Fig. 2 SE(Squeeze-and-excitation) block

Excitation 操作是为了显示建模特征通道间的相关性，考虑到通道之间复杂的相关性，Excitation 操作是采用两层全连接层和一个激活层的方式实现的。通过参数 W 来为每个特征通道生成权重，且 W_1 和 W_2 分别表示的第一层和第二层的可学习参数：

$$s = F_{\mathrm{ex}}(z,W) = \sigma(g(z,W)) = \sigma(W_2\sigma(W_1 z)), s\in R^{1\times 1\times C} 。 \tag{6}$$

Reweight 操作将 Excitation 操作的输出的权重通过乘法逐通道加权到先前的特征上，得到 $\tilde{A} = [\tilde{a}_1, \tilde{a}_2, \cdots, \tilde{a}_C], \tilde{A}\in R^{14\times 14\times C}$，其中：

$$\tilde{a}_l = F_{\mathrm{Re}}(a^l, s_l) = a^l \bullet s_l \quad 。 \tag{7}$$

如图 1 中“ Learn label semantic dependencies ”部分所示，在 A 的后面接上一个 SE 模块，通过网络训练，SE 模块能够抑制一些无用的特征，这些特征通常是一些不存在与图像中的类的特征(通道响应)。消除负响应后，特征图的通道和类别相对应，因此标签的关联性表现在通道之间的关联性，本文利用卷积操作去捕获通道之间的关联性，在实验中发现，不同的卷积层之间加入 SE 模块对网络的性能有一定的提升，原因在于 SE 模块可以很好地提升关联信息的同时抑制无关特征信息，本文使用 4 层卷积和 SE 模块交叉的方式是为了学习在不同特征维度下的标签之间的语义关联。其中，卷积层维度依次为 $1\times 1\times 256$ 、 $1\times 1\times 256$ 、 $1\times 1\times 512$ 、 $1\times 1\times 1024$ ，整个语义关联的模型输出 M 为

$$M = F_{\mathrm{cc}}(A,\theta_{\mathrm{cc}}), A\in R^{14\times 14\times C} \quad 。 \tag{8}$$

在此基础之上，将 M 经过平均池化层、全连接层得到一个 C 维的置信度 $\tilde{y}$ ，类似地，denseNet 特征提取网络的输出 $X\in R^{14\times 14\times 1024}$ 也会经过平均池化层和全连接层得到一个 C 维的置信度 y' ，通过加权操作获得最终的置信度 $\hat{y}$ ：

$$\hat{y} = \alpha y' + (1-\alpha)\tilde{y} \quad , \tag{9}$$

其中 $\alpha\in[0,1]$ 是可学习的参数，本文将其初始值设置为 0.48。

3 实验结果与分析

本文使用两种不同类型标签的数据集对模型进行评估：包含 20 个类别的 Pascal VOC2007 数据集[25]和包含 38 个类别的 MirFLickr25k 数据集。实验结果表明，相对于其它模型，本文方法在这两个数据集上都有更好的效果，在可以有效地捕捉标签之间的依赖关系的同时，也具有很好的泛化能力。

3.1 实现细节

本文实验是使用 pyTorch 框架实现，并使用单个 NVIDIA Geforce GTX Titan X GPU 对本文方法进行训练和测试。在训练阶段，首先，将图像大小调整为 512×512，然后随机裁剪图像为 448×448，并对剪裁后的图像进行随机水平翻转。模型训练的初始学习率设置为 0.001，且在每 30 个周期降低为之前的 1/10，动量为 0.9，权重衰减为 0.0005，使用随机梯度下降法进行优化。由于 GPU 内存的限制，批量大小设置为 24。在测试阶段，只需将图像大小调整为 448×448。

整个模型的训练分为 4 个阶段：第一阶段是使用多标签数据集微调特征提取网络的参数，该模型的初始参数是通过 ImageNet 单标签数据集预训练得到；第二阶段是固定特征提取网络的参数，训练注意力机制的参数；第三阶段是固定前两个阶段的参数，使用数据集训练语义关联分类部分的参数；最后一个阶段是使用数据集整体微调整个模型的参数。在网络模型的训练过程中，使用交叉熵损失函数(cross entropy)：

$$L_{\mathrm{loss}}(y,\hat{y}) = \sum_{j=1}^{T} y_j \log\sigma(\hat{y}_j) + (1-y_j)\log(1-\sigma(\hat{y}_j)) \quad , \tag{10}$$

其中：T 表示标签的种类数，y 表示真实值标签，$\hat{y}$ 表示标签的预测值，L_{loss} 作为训练多标签数据集训练的损失函数。

3.2 Pascal VOC2007 实验

Pascal VOC2007 数据集被广泛地应用于多标签分类的衡量标准，也可以应用于目标检测。数据集一共有 9963 张图像包括 20 个类别，其中训练数据有 5011 张，测试数据有 4952 张。该训练集除了会给出每张图像的标签外还会给出其相应的位置信息，因此，在网络的训练和测试过程中，需要对数据集的标签进行一定的预处理，然后再对网络进行训练与测试。该数据集的评价标准是平均精度(AP)和平均精度的平均值(mAP)，实验结果如表 1 所示，其中，红色表示方法的最佳效果，蓝色表示方法的次最佳效果。

表 1 Pascal VOC2007 数据集实验结果

Table 1 The experimental results on Pascal VOC2007 dataset

Labels	Plane	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table
CNN-SVM	88.5	81.0	83.5	82.0	42.0	72.5	85.3	81.6	59.9	58.5	66.5
CNN-RNN	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0
RLSD	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2
Very deep	98.9	95.0	96.8	95.4	69.7	90.4	93.5	96.0	74.2	86.6	87.8
Densenet	99.1	95.4	96.6	95.4	70.6	89.0	94.3	95.9	78.8	88.9	79.9
Proposed	99.3	95.8	97.2	95.4	73.2	88.5	94.3	95.5	77.3	91.8	81.4

Labels	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	mAP
CNN-SVM	77.8	81.8	78.8	90.2	54.8	71.1	62.6	87.2	71.8	73.9
CNN-RNN	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
RLSD	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
Very deep	96.0	96.3	93.1	97.2	70.0	92.1	80.3	98.1	87.0	89.7
Densenet	96.8	96.2	92.4	97.8	77.4	88.1	77.7	98.3	88.2	89.9
Proposed	97.1	96.3	91.7	98.0	78.3	92.2	75.7	98.4	88.8	90.4

本文方法相对于 CNN-SVM[13]、CNN-RNN[17]、RLSD[19]、Very deep[10]来说，最终平均识别精确率达到了最好的效果 90.4%，只使用多标签数据集微调后的 denseNet 网络模型的效果是 89.9%。并且，从表 1 中可以看出，在该数据集的 20 个标签的识别准确率中，13 个标签的识别精确率相较于以往方法有着更好的结果。在该数据集中，人(person)是出现频率最高的标签，针对表 1 的实验结果分析发现，一些与人存在语义相关的标签其识别精确率会有一定的提升，比如，人和瓶子(bottle)同时出现的频率较高，本文方法对瓶子的识别精确率为 73.2%，相对于上述 RLSD 模型的精确率 71.2%还有 2%的提高，相对于 denseNet 网络模型结果有了 2.6%的提高；人和自行车(bike)同时出现的频率也较高，自行车的识别精确率为 95.8%，相对于 Very deep 模型的精确率有了 0.8%的提升，相对于 denseNet 网络模型有 0.4%的提升，通过这些依赖关系，人的识别精确率相对于 RLSD 模型和 denseNet 模型分别有 0.1%和 0.2%的提升。除此之外，一些其他的标签之间也存在着一定的语义依赖关系，比如，桌子与盆栽(plant)、桌子(table)与电视(TV)。只使用 denseNet 模型时桌子、盆栽、电视的识别精确率分别为：79.9%、77.4%、88.2%，本文方法最终的识别准确率分别为 81.4%、78.3%、88.8%，分别有了 0.5%、0.9%、0.6%的提升。这说明，本文方法可以很好地学习标签之间的依赖关系并以此提升多标签分类结果。

3.3 MirFLickr25k 实验

MirFLickr25k 数据集是 MirFLickr 系列中的一个类型，包括 25000 张来自社交摄像网站 FLickr 上下载的图像。其中，已经标记了 24 个概念用于分类，其中，有 14 个概念当其在图像中显著体现时对其进行更严格的标注，这使得数据集一共有 38 个标签类别。在训练时，按照 3:2 的比例随机挑选训练图像和测试图像，最终训练图像和测试图像的数目分别为 15000 张和 10000 张。本文分别使用宽泛的 24 个标签集数据和对某些类别经过严格标注的 38 个标签数据集分别进行了实验。

24 个类别数据集的实验结果如图 3 所示，其中，绝大多数标签类别的识别准确率均比 MVAIACNN 模型要高。最终的平均识别精确率为 81.1%，相较于 MVAIACNN 模型(66.6%)的 mAP 提高了 14.5%。38 个类别数据集的实验结果如表 2 所示，从表中可以看出，大多数标签的识别精确率相较于 LDA、SVM、DBN[26]、AIACNN(MVAIACNN)[27]相比，都有了明显的提升；本文方法最终的平均识别精确率的结果为 78.2%，相比于 AIACNN(MVAIACNN)最好的效果 62.4%提高了 15.8%，相对于只使用主网络模型 denseNet 的效果 77.4%提升了 0.8%。由于该数据集图像来源于社交摄像网站，大多数图像所包含的内容十分丰富，每张图像的平均标签个数为 5~7 个，因此，该数据集存在着

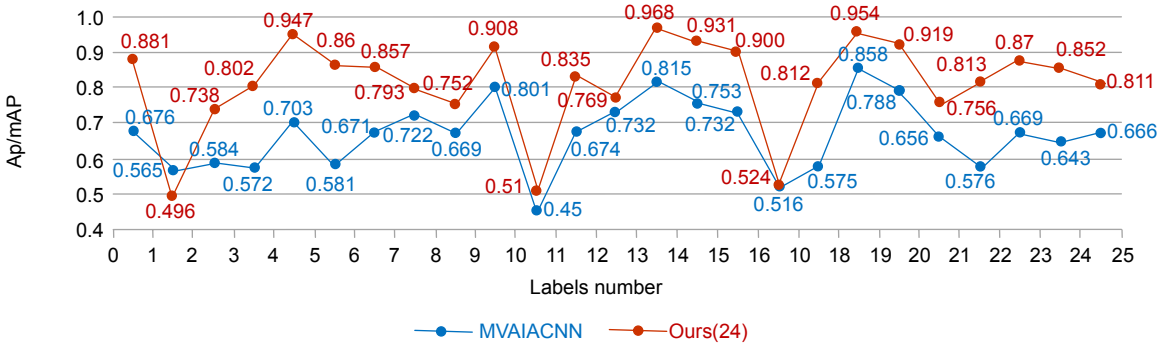


图 3 MirFLickr25k 数据集(24 个类)实验结果

Fig. 3 Experimental results of MirFLickr25k datasets (24 classes)

表 2 MirFLickr25k 数据集实验结果(38 个类)

Table 2 The experimental results on MirFLickr25k dataset (38 classes)

No.	Labels	LDA	SVM	DBN	AIACNN	Denseset	Proposed(38)
1	Animals	0.537	0.531	0.498	0.612	0.854	0.862
2	Baby(r1)	0.285(0.308)	0.200(0.165)	0.129(0.134)	0.487(0.462)	0.459(0.591)	0.467(0.587)
3	Bird(r1)	0.426(0.500)	0.443(0.520)	0.184(0.255)	0.529(0.534)	0.704(0.902)	0.725(0.918)
4	Car(r1)	0.297(0.389)	0.339(0.434)	0.309(0.354)	0.502(0.521)	0.758(0.818)	0.777(0.827)
5	Cloud(r1)	0.651(0.528)	0.695(0.434)	0.759(0.691)	0.667(0.682)	0.937(0.833)	0.945(0.840)
6	Dog(r1)	0.621(0.663)	0.607(0.641)	0.342(0.376)	0.555(0.523)	0.832(0.902)	0.864(0.914)
7	Female(r1)	0.494(0.454)	0.465(0.451)	0.540(0.478)	0.623(0.630)	0.840(0.842)	0.844(0.855)
8	Flower(r1)	0.560(0.623)	0.480(0.717)	0.593(0.679)	0.612(0.630)	0.774(0.886)	0.775(0.881)
9	Food	0.439	0.308	0.447	0.645	0.714	0.720
10	Indoor	0.663	0.683	0.750	0.793	0.902	0.903
11	Lake	0.258	0.207	0.262	0.369	0.458	0.463
12	Male(r1)	0.434(0.354)	0.413(0.335)	0.503(0.406)	0.623(0.625)	0.821(0.801)	0.830(0.814)
13	Night(r1)	0.615(0.420)	0.588(0.450)	0.655(0.483)	0.712(0.709)	0.760(0.640)	0.761(0.681)
14	People(r1)	0.731(0.664)	0.748(0.565)	0.800(0.730)	0.789(0.787)	0.967(0.969)	0.967(0.973)
15	Plant_life	0.703	0.691	0.791	0.721	0.921	0.927
16	Portrait(r1)	0.543(0.541)	0.480(0.558)	0.642(0.635)	0.692(0.698)	0.911(0.909)	0.903(0.901)
17	River(r1)	0.317(0.134)	0.158(0.109)	0.263(0.110)	0.488(0.246)	0.462(0.184)	0.492(0.146)
18	Sea(r1)	0.477(0.197)	0.529(0.201)	0.586(0.259)	0.526(0.301)	0.788(0.498)	0.792(0.499)
19	Sky	0.800	0.823	0.873	0.833	0.950	0.949
20	Structures	0.709	0.695	0.787	0.756	0.913	0.916
21	Sunset	0.528	0.613	0.648	0.649	0.745	0.747
22	Transport	0.411	0.369	0.406	0.516	0.785	0.787
23	Tree(r1)	0.515(0.342)	0.559(0.321)	0.660(0.483)	0.639(0.388)	0.852(0.706)	0.860(0.713)
24	Water	0.525	0.527	0.629	0.629	0.841	0.836
25	mAP	0.492	0.475	0.503	0.624	0.774	0.782

较多的标签依赖关系，例如：1) 女性(female)与室内(indoor)、食物(food)与室内、动物(animals)与室内都存在共生关系；2) 天空(sky)与建筑(structures)、太阳(sunset)与建筑、交通工具(transport)与建筑都存在共生关系；3) 河流(river)与天空、云(clouds)与天空都存在共生关系。由于标签之间的语义依赖关系十分多样，从表中可以看出，相对于只使用 denseNet 模型的识别准确率，本文方法单个标签的识别准确率绝大多数都有所提升，最终的平均识别准确率也相应的提升。这说明融合注意力机制与语义关联性的方法可以通过学习标签之间的依赖关系来提升最终的多标签分类效果。

3.4 模型分析与讨论

为证明本文方法的有效性，在两个数据集上同时做了如表 3 所示的各模块的对比实验。由表可知，只使用 denseNet 模型进行分类时，最终的 mAP 结果在两个数据集上分别为 89.9%和 77.4%。在此基础之上，结合注意力机制，并利用一个 SE 模块来消除其产生的负响应，然后使用一个全局平均池化层和全连接层并与 denseNet 的输出融合，最终的 mAP 结果为 90.0%和 77.6%，这说明，只结合注意力机制对整体的分类效果影响不大，本文旨在利用注意力机制使得特征图通道和标签类别一一对应，并没有学习标签之间的关联性。不使用注意力机制，只采用 denseNet 和卷积与 SE 的结合，分别得到 89.7%和 77.0%的结果，相较于只使用 denseNet 的结果分别下降了 0.2%和 0.4%，这说明只增加网络层并不能促进分类效果的提升。最终的实验结果说明，利用注意力机制获取特征图通道和标签之间的对应关系，在此基础之上学习标签之间的依赖关系可以提升多标签图像分类结果。

为验证语义关联模块的有效性，本文分别在两个数据集上做了如表 4 所示的对比实验。结合图 1，注意力机制产生的负响应利用 SEblock1 进行消除，然后，直接使用 4 个卷积层是可以学习标签之间的语义关联性的，在两个数据集上的平均识别精确率分别为 90.1%和 77.8%。在此基础之上，在两个卷积层之间插入一个 SEblock3，其分类效果会有一定的提升，原因在于 SE 模块可以抑制无关特征信息，从而有利于捕获标签之间的关联性。此外，通过实验发现，过多堆叠卷积层和 SE 模块，并不能获得更好的效果，反而会导致最终的识别效果一定程度的下降，原因可能是过多层数会导致网络收敛效果不好，增加训练难度。如表中采用 5 个 SE 模块和 5 个卷积层的组合最终的实验效果并没有超过本文方法。

不过，本文方法也存在一定的不足：

首先，在 2.2 节中利用注意力机制获得数据集标签类别与其学习得到的特征图通道的对应关系，但是仍存在一些值得讨论的内容：1) 特征图通道是否能充分表征其对应标签类别特征信息。使用若干个特征图通道作为一组来表征数据集中的一个标签类别，通过对比不同的分组方式的实验效果挑选出最合适的分组方式；2) 标签类别与特征图通道的对应关系能否进一步的明确。可以对通道进行可视化，观察对应关系是否存在一定的特点，再做进一步的分析。

其次，受 Li 等[28]在构建向量空间模型时利用了大规模文本数据集进行训练的启发，在利用卷积学习标签之间的依赖关系时，考虑利用文本集中的语义关联模型对数据集的标签关系进行一定的语义度量，结合语义关联模块对数据集的标签权重进行一定的调整，提升分类效果。

表 3 注意力机制与语义关联分类模块对图像分类效果的影响

Table 3 The influence of attention mechanism and semantic association classification model on image classification

Methods	Pascal VOC2007	MirFLickr25k(38)
DenseNet	89.9%	77.4%
DenseNet+att	90.0%	77.6%
DenseNet+SE	89.7%	77.0%
Proposed	90.4%	78.2%

表 4 SE 模块和卷积的结合方式对语义关联分类效果的影响

Table 4 The influence of the combination of SE block and convolution on semantic association classification

Methods	Pascal VOC2007	MirFLickr25k(38)
SEblock(1) +4Convs	90.1%	77.8%
SEblock(1,3) +4Convs	90.2%	77.9%
5SEblock +5Convs	90.4%	78.1%
Proposed	90.4%	78.2%

4 结 论

多标签图像分类研究在计算机视觉领域中具有重要意义，本质上是单标签图像分类的一种推广。本文结合注意力机制和语义关联性提出了一种有效的多标签图像分类方法。利用卷积神经网络来提取图像的视觉特征，然后使用注意力机制使得数据集的标签类别与特征图中的通道一一对应，并利用基于监督学习的方式学习特征图通道之间的依赖关系，也就是学习标签之间存在的语义依赖关联性。实验结果表明，我们的方法可以很好地学习标签之间的依赖关系并提升图像多标签分类结果。

同时也看到本文改进之处，使用分组更好的表征图像的特征信息，并更好地探索数据集标签类别与每组特征图通道之间的对应关系(例如对通道进行可视化)；除此之外，引入文本集中语义度量方式对数据集标签集进行一定的度量调整最终的分类效果也是值得研究的方向。

参考文献

[1] Sivic J, Zisserman A. Video Google: a text retrieval approach to object matching in videos[C]//*Proceedings 9th IEEE International Conference on Computer Vision*, 2003: 1470–1477.

[2] Wang R G, Ding K, Yang J, *et al*. Image classification based on bag of visual words model with triangle constraint[J]. *Journal of Software*, 2017, **28**(7): 1847−1861.
汪荣贵, 丁凯, 杨娟, 等. 三角形约束下的词袋模型图像分类方法[J]. 软件学报, 2017, **28**(7): 1847–1861.

[3] Huang Q H, Liu Z. Multiple-hyperplane SVMs algorithm in image semantic classification[J]. *Opto-Electronic Engineering*, 2007, **34**(8): 99–104.
黄启宏, 刘钊. 基于多超平面支持向量机的图像语义分类算法(英文)[J]. 光电工程, 2007, **34**(8): 99–104.

[4] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. *ACM Transactions on Intelligent Systems and Technology*, 2011, **2**(3): 27.

[5] Breiman L. Random forests[J]. *Machine Learning*, 2001, **45**(1): 5–32.

[6] Harzallah H, Jurie F, Schmid C. Combining efficient object localization and image classification[C]//*Proceedings of the 12th International Conference on Computer Vision*, 2009: 237–244.

[7] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004, **60**(2): 91–110.

[8] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//*Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005: 886–893.

[9] Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions[J]. *Pattern Recognition*, 1996, **29**(1): 51–59.

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556[cs.CV], 2015.

[11] Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks[C]//*Proceedings of 2017 IEEE Computer Vision and Pattern Recognition*, 2017: 2261–2269.

[12] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778.

[13] Razavian A S, Azizpour H, Sullivan J, *et al*. CNN features off-the-shelf: an astounding baseline for recognition[C]// *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 512–519.

[14] Deng J, Dong W, Socher R, *et al*. ImageNet: a large-scale hierarchical image database[C]//*Proceedings of 2009 IEEE Computer Vision and Pattern Recognition*, 2009: 248–255.

[15] 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.

[16] Cheng M M, Zhang Z M, Lin W Y, *et al*. BING: binarized normed gradients for objectness estimation at 300fps[C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 3286–3293.

[17] Wang J, Yang Y, Mao J H, *et al*. CNN-RNN: a unified framework for multi-label image classification[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2285–2294.

[18] Hochreiter S, Schmidhuber J. Long short-term memory[J]. *Neural Computation*, 1997, **9**(8): 1735–1780.

[19] Zhang J J, Wu Q, Shen C H, *et al*. Multilabel image classification with regional latent semantic dependencies[J]. *IEEE Transactions on Multimedia*, 2018, **20**(10): 2801–2813.

[20] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//*Proceedings of the 32nd International Conference on Machine Learning*, 2015: 448–456.

[21] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//*Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011: 315–323.

[22] Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention[J]. arXiv:1412.7755[cs.LG], 2015.

[23] Xu K, Ba J, Kiros R, *et al*. Show, attend and tell: neural image caption generation with visual attention[J]. arXiv:1502.03044 [cs.LG], 2015.

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

[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] Srivastava N, Salakhutdinov R. Learning representations for multimodal data with deep belief nets[C]//*Proceedings of 2012 ICML Representation Learning Workshop*, 2012: 79.

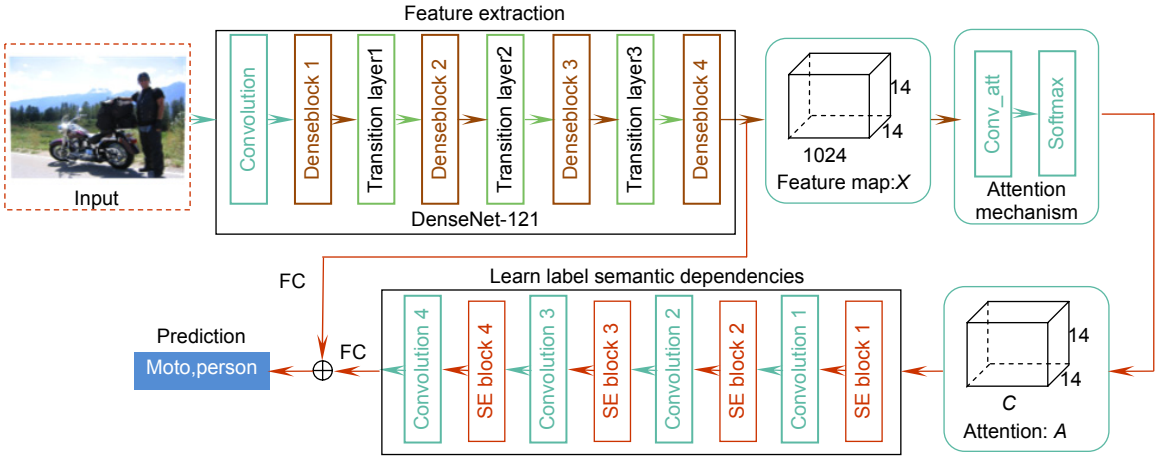
[27] Wang R G, Xie Y F, Yang J, *et al*. Large scale automatic image annotation based on convolutional neural network[J]. *Journal of Visual Communication and Image Representation*, 2017, **49**: 213–224.

[28] Li Y N, Yeh M C. Learning image conditioned label space for multilabel classification[J]. arXiv:1802.07460[cs.CV], 2018.

Multi-label classification based on attention mechanism and semantic dependencies

Xue Lixia, Jiang Di, Wang Ronggui, Yang Juan*

School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230009, China



Overview of the proposed method

Overview: As a fundamental task of image classification problems, single-label image classification has been researched for decades and has made good progress. However, multi-label image classification task is not only a general and practical problem, but also a challenging task, because most real-world images often contain rich semantic information, such as multiple objects, scenes, attributes, and actions. In this paper, combines attention mechanism and semantic relevance, a method based on convolutional neural networks is proposed to solve the multi label problem. Firstly, we use the recent most popular convolutional neural network denseNet-121 to extract image features. Traditional methods usually to pre-process the images by extracting hand-craft features and train a classifier. However, these hand-craft are designed for different visual tasks. In contrast, the method based on convolutional neural network can extract more discriminative features from images by powerful feature learning ability. Secondly, the attention mechanism which can explore the basic spatial relation has recently been applied to many computer vision tasks. For multi-label images classification, most of the images have different semantic information and we tag them with several labels. We hope that we will use the attention mechanism to focus on the areas of interest where we need to identify and the channels of the feature map can correspond to the categories of the dataset so as to better explore the dependencies between labels. Consequently, we use the image feature map extracted from the network as the input of the attention mechanism and utilize convolution operation to preliminarily learn the conversion relationship between the label and the channel. Then, we employ the softmax function to ensure that each group channel of the feature map has a tag response. The softmax operation may cause visual feature redundancies, because the network also learns some negative feature information, that is, the corresponding labels that are not existed in the images. So, we exploit the SE module to eliminate the negative feature information. The Squeeze-and-Excitation (SE) block which is a structural unit is able to definitely model inter-dependencies between channels. And this unit focuses on channels through adaptively adjusting channel-wise feature information. Finally, we explore the channel-wise correlation which is essentially the semantic dependencies between labels by means of supervised learning. This special approach using the SE block and the convolution operation alternately is able to more accurately learn the dependencies between channels. The experimental results show that the proposed method can exploit the dependencies between multiple tags to improve the performance of multi label image classification.

Citation: Xue L X, Jiang D, Wang R G, *et al*. Multi-label classification based on attention mechanism and semantic dependencies[J]. *Opto-Electronic Engineering*, 2019, **46**(9): 180468

* E-mail: yangjuan6985@163.com