基于特征融合和代价敏感学习的图像标注方法

厍向阳，车子豪+，董立红

（西安科技大学 计算机科学与技术学院，陕西 西安 710054）

摘　要：针对图像标注数据集中存在的标注对象比例不一致和标签分布不平衡问题，提出基于特征融合和代价敏感学习的图像标注方法。在卷积神经网络中加入特征融合层，改进 VGG16 原有的网络结构，特征融合层结合注意力机制，对网络中不同卷积层提取的多尺度特征进行选择性融合，提升对不同尺度对象的标注精度；将代价敏感学习融入损失函数对网络模型进行训练，提升网络的泛化性能。实验结果表明，该方法能提升图像标注的准确率，增加对低频标签的召回率。

关键词：图像自动标注；深度学习；特征融合；卷积神经网络；代价敏感学习

中图法分类号：TP391　**文献标识号**：A　**文章编号**：1000-7024（2021）11-3114-07

doi：10.16208/j.issn1000-7024.2021.11.015

Image annotation method based on feature fusion and cost-sensitive learning

SHE Xiang-yang, CHE Zi-hao+, DONG Li-hong

(College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract: To solve the problems of object scale inconsistency and category imbalance in image datasets, an image annotation method based on feature fusion and cost-sensitive learning was proposed. The feature fusion layer was added to the convolutional neural network to improve the original network structure of VGG16, and the attention mechanism was combined to selectively fuse the multi-scale features extracted from different convolutional layers in the network to improve the performance of objects of different scales. Cost-sensitive learning was incorporated into the loss function to train the network model to improve the generalization performance of the network. Experimental results show that the proposed method can improve the accuracy of image annotation and increase the recall rate of low-frequency labels.

Key words: automatic image annotation; deep learning; feature fusion; convolutional neural network; cost-sensitive learning

0　引　言

网络技术的快速发展促进了数字图像的传播，使得用户可以通过互联网的检索工具搜索访问感兴趣的图像资源。但互联网的检索工具无法理解图像内容和语义，从而无法确定哪些图像满足查询要求，在这种情况下，图像标注是一个必需的过程[1]。然而互联网中图像的数量呈爆炸式增长，仅靠人工标注是无法满足需求的。因此，图像标注转向寻求机器学习算法来自动完成。目前图像自动标注方法可分为两类：基于生成模型的方法[2-4]计算已标注图像特征和标注词的联合概率分布，然后使用该模型计算每个标签匹配待标注图像的概率；基于判别模型的方法[5-9]将图像标注问题视为分类问题，使用图像的视觉特征训练分类器，通过训练的分类器将待标注图像划分到一个或多个标签类别中。近年来，基于卷积神经网络（convolutional neural network，CNN）的判别模型为图像标注提供了多种方法。文献［7］提出了 CNN-MSE 方法，通过改进均方误差函数来训练 CNN 网络。文献［8］在 CNN 模型中加入多标签平滑单元构成 CNN-MLSU 模型。深入分析现有工作，发现基于 CNN 的图像自动标注研究仍面临两个问题：①CNN 模型中，通过不断降采样过程使得深层的卷积层具有较大的感受野，如果感受野远大于物体的大小，那么很容易忽略小物体的特征，使图片中较小的物体不容易被标注和学习。②由于图像自动标注数据集中训练样本不足且标注类

收稿日期：2020-07-31；**修订日期**：2020-10-09

基金项目：陕西省自然科学基础研究基金项目（2019JLM-11）；陕西省自然科学基金项目（2017JM6105）

作者简介：厍向阳（1968-），男，陕西周至人，博士后，教授，研究方向为数据挖掘与智能信息处理、人工智能与模式识别；+通讯作者：车子豪（1996-），男，山东青岛人，硕士研究生，研究方向为机器学习、图像处理；董立红（1968-），女，河北唐山人，博士，教授，研究方向为智能信息处理技术、计算机监测与控制技术。E-mail：786058509@qq.com

别之间数量差异较大，使得训练出来的模型泛化性能较差。为解决以上问题，本文采用卷积神经网络构造端到端的图像标注模型，选择 VGG16 作为基网络，在其基础上引入特征融合机制融合不同卷积层提取的多尺度特征，最后在网络训练时使用代价敏感损失函数，来缓解标签分布不平衡引发的问题，进一步提升网络的性能。

1　相关理论与技术

1.1　卷积神经网络

卷积神经网络由 4 个部分组成：输入层、特征提取层、全连接层和分类器。卷积神经网络结构如图 1 所示。

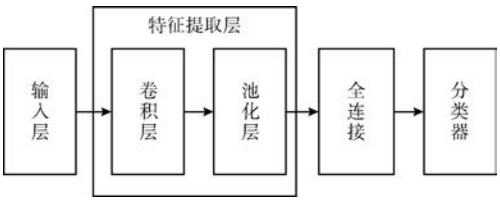

图 1　卷积神经网络结构

（1）输入层。接收预处理后的图像数据。

（2）卷积层。假设 X 为原始图像，X^0 为预处理后输入网络的图像。X^i 为第 i 层卷积特征图，由卷积核和偏置项计算出，卷积计算过程如下式

$$X^i = f(X^{i-1} \otimes w^i + b^i) \tag{1}$$

式中：w^i 表示第 i 层卷积中卷积核的权重矩阵；b^i 表示第 i 层卷积的偏置项；$\otimes$表示 2D 卷积运算操作；$f(\cdot)$ 表示激活函数，一般采用线性整流函数（ReLU），公式如下

$$f(x) = \begin{cases} x, & x > 0 \\ 0, & x \leqslant 0 \end{cases} \tag{2}$$

（3）池化层。池化层也称子采样层，通常使用平均池化（mean pooling）或最大池化（max pooling）。卷积操作后为了减少特征维数，降低数据复杂度，对特征进行池化操作，通过对下采样子区域取平均值或最大值来对特征图进行下采样。特征提取层通过对特征图重复执行卷积和池化操作，来递归提取高层特征。

（4）全连接层。在卷积神经网络的最后一般会连接全连接层来得到最后的分类结果，经过特征提取层对图像数据进行非线性特征提取后，输入到全连接层对特征进行聚合。将特征提取层看成自动提取图像特征的过程，提取完特征以后，仍需要通过全连接层来完成分类的任务。

（5）分类器。常用的分类器有 Softmax 分类器和 Sigmoid 分类器，分类器可以将最后一层全连接层的输出转换为当前样本属于每类标签的概率分布情况。

1.2　多核选择网络

Inception 结构仅是简单的对不同尺度的特征图进行融合，因此，文献［10］设计了 Selective kernel（SK）网络将多分支网络结构与软注意力机制相结合，有选择地融合不同尺度的特征信息，使网络更好获取不同感受野提取的信息。SK 网络结构如图 2 所示，主要包含 Split、Fuse 和 Select 这 3 个操作。

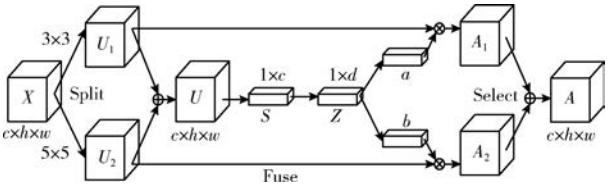

图 2　SK 网络结构

Split：使用不同尺寸的卷积核对输入特征图 X 进行特征提取，得到特征图 U_1 和 U_2。

Fuse：通过对特征图 U_1、U_2 进行逐元素相加，然后进行全局平均池化操作得到特征图 S。将特征图 S 输入全连接层进行线性变换，提取通道维度的信息，具体操作如下所示

$$Z = \delta(\beta(WS)) \tag{3}$$

式中：$\delta(\cdot)$ 为 ReLU 激活函数，$\beta(\cdot)$ 为批标准化操作。$W \in R^{d\times c}$ 表示全连接层的参数，d 为经过全连接层后输出的维度。

Select：特征 Z 输入全连接层，再使用 Softmax 函数来进行归一化得到 U_1，U_2 的通道权重 a 和 b。然后将通道权重乘以对应的 U_1，U_2 得到 A_1，A_2。最后，将 A_1，A_2 逐元素相加得到最终的融合特征 A。

1.3　损失函数

网络中的损失函数一般针对单标签分类问题，为了解决多标签标注问题需要将标签向量中对应元素的损失值求和。第 i 个样本包含第 j 个标签的真实概率 $\bar{t}_{ij}$ 定义为

$$\bar{t}_{ij} = t_{ij} / \| t_i \|_1 \tag{4}$$

$$t_{ij} = \begin{cases} 1 & (j = true\ lable) \\ 0 & (j \neq true\ lable) \end{cases} \tag{5}$$

多标签损失（multi label loss，ML Loss）函数定义为

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\bar{t}_{ij}\log(y_{ij}) \tag{6}$$

式中：N 表示样本数量；C 表示标签类别数；$y_{ij} \in [0,1]$ 表示网络预测第 i 个样本中包含第 j 个标签的概率，y_{ij} 由下式计算

$$y_{ij} = \frac{e^{x_j}}{\sum_{c=1}^{C} e^{x_c}} \tag{7}$$

式中：x_j 为网络模型最后一层第 j 个神经元的输出。

2　基于特征融合和代价敏感学习的图像标注算法

2.1　基本思想

融合不同卷积层的特征可以提高网络的学习能力，低层的卷积特征具有较多的细节特征，但是噪声多；高层卷

积特征语义信息丰富，但分辨率低，易忽略细小特征。直接将高低层特征连接在一起来融合特征会引入大量无用特征，增加网络的参数量和计算量，影响网络的性能。因此，本文借鉴 SK 网络的思想去融合不同层提取的多尺度特征，使得融合的特征能够更加全面的描述图像的内容，并改进损失函数引入代价敏感学习，使得不同类型标签的误分类代价具有较大差异。

首先将预处理后的样本输入到网络模型中，利用预训练的 VGG16 网络模型的卷积层进行特征提取生成特征图；其次将得到的特征图输入到采样层进行维度调整，再输入 L2 归一化层；然后特征融合层融合不同卷积层提取的多尺度特征；最后连接融合特征与全连接层的神经元，通过分类器得到每个标签标注样本的概率，提取前 K 个概率最大的标签作为标注结果。训练过程中使用代价敏感损失函数对网络参数进行训练，经过多次训练获得最终的图像标注模型。

2.2 代价敏感的多标签损失函数

本文对损失函数进行了修改加入权重敏感系数和错分敏感系数，设计代价敏感的多标签损失（cost sensitive multi label loss，CSML Loss）函数，计算公式如下

$$L=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\omega_j\times\left(1-\sin\left(\frac{\pi}{2}\times y_{ij}\right)\right)^{\gamma}t_{ij}\log(y_{ij}) \tag{8}$$

式中：N 表示样本数量；C 表示标签类别数；y_{ij} 表示网络预测第 i 个样本中包含第 j 个标签的概率，$y_{ij}\in[0,1]$；$\left(1-\sin\left(\frac{\pi}{2}\times y_{ij}\right)\right)^{\gamma}$ 表示错分敏感系数，控制标签在计算损失值时的权重，γ 表示调焦参数且 $\gamma>0$；ω_j 表示权重敏感系数，与标签出现频率相关，计算公式如下

$$\omega_j=1+\frac{S_{min}}{\beta\times S_j} \tag{9}$$

式中：S_{min} 表示数据集中出现频率最低的标签的数量；S_j 表示数据集中第 j 个标注词出现的频率；β 表示权重控制系数，通过调节 β 值可以控制不同标签在计算损失值时的权重。

式（8）中的错分敏感系数用来控制难易标签的权重，可以看出当预测值 y_{ij} 越接近真实值时，错分敏感系数值越小。通过降低简单标签在计算损失值时的权重，从而降低简单标签的损失值，使得网络把训练的重点放在难标注的标签上。权重敏感系数用来控制不同类别标签的权重，通过提高低频标签的权重，从而增加低频标签的损失值，使得损失函数把训练的重点放在低频标签上。因此，低频标签和难标注的标签在计算损失值时将被赋予较大的权重，而高频标签和易标注的标签将被赋予较小的权重。

2.3 融合多尺度特征的卷积神经网络

2.3.1 模型框架

为了更好地解决图像自动标注领域存在的问题，本文在 VGG16 模型基础上设计了新的网络结构，如图 3 所示。网络模型包含有 13 层卷积层、4 层最大池化层、3 层采样层、1 层特征融合层和 3 层全连接层。卷积层使用的是 VGG16 在 ImageNet 数据集上预训练的参数进行初始化。本文网络主要是在 VGG16 框架中添加特征融合层来融合高低卷积层提取的多尺度特征，从而提高网络的标注性能。

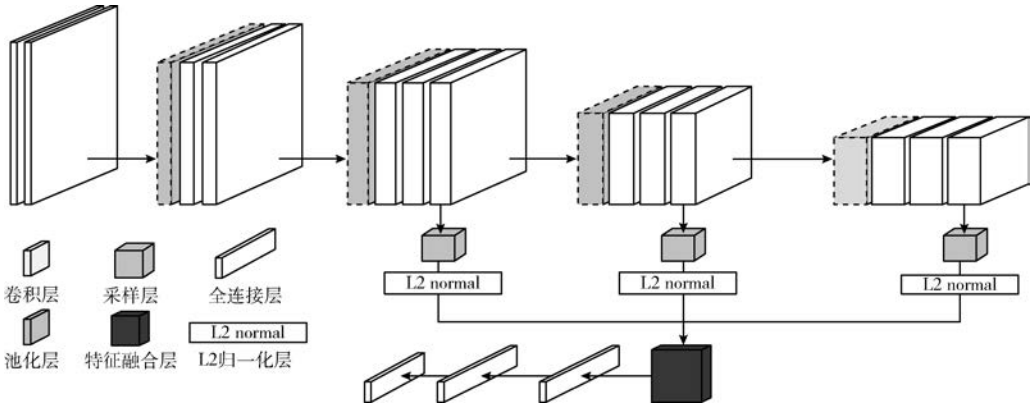

图 3 本文算法网络结构

为保证卷积特征在输入特征融合层时在通道维度上相匹配，采样层使用 1×1 的卷积在通道维度上进行降维或者升维操作。由于不同卷积层特征的激活值不同，直接对多尺度特征进行操作，会导致网络无法稳定训练。因此，在输入特征融合层前进行 L2 归一化操作，对卷积特征进行归一化。

2.3.2 特征融合层

特征融合层融合操作主要分为 3 个部分：①从卷积特征中提取多尺度特征；②改进 SK 网络融合特征；③融合层融合多尺度特征。

使用自适应最大池化（adaptive max pool，AdaMaxPool）操作提取多尺度特征，自适应最大池化中输入任意大小的特征图，都能产生指定大小的输出。因此，使用不同尺寸的自适应最大池化操作就可以提取到不同尺度的图像特征。自适应最大池化首先需要根据输出特征图的大

小计算滤波器的尺寸（Size）和步长（Stride），然后将得到的尺寸和步长输入最大池化中提取特征，Size 和 Stride 的计算公式如下

$$Stride = floor(inputSize \div outputSize) \tag{10}$$

$$Size = inputSize - (outputSize - 1) \times Stride \tag{11}$$

式中：$floor(\cdot)$ 为向下取整，$inputSize$ 为输入特征的尺寸，$outputSize$ 为输出特征的尺寸。

多尺度特征的提取过程如图 4 所示，使用自适应最大池化操作将图中 $a \times a \times N$ 的特征图转化为 $1 \times 1 \times N$、$2 \times 2 \times N$、$4 \times 4 \times N$ 的特征图。将 3 层卷积层的特征都经过自适应最大池化层进行多尺度特征提取，相同尺寸的特征输入到改进 SK 网络中进行融合。

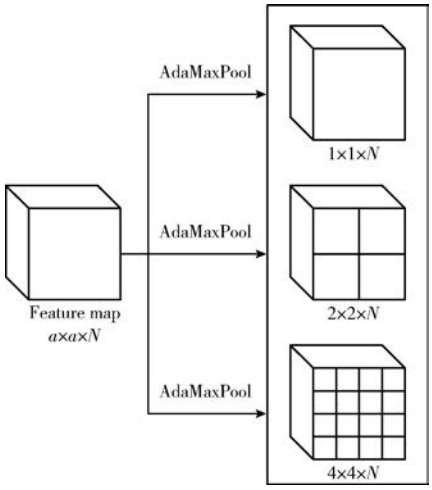

图 4　多尺度特征提取

改进 SK 网络融合特征。如图 5 所示，与原始的 SK 网络相比，本文去掉了 Split 操作，改为直接输入从不同卷积层中提取的多尺度特征。使用 SK 网络不仅能在通道维度上加强重要特征并压缩无用特征，还能根据不同层卷积特征的重要程度来融合特征，使得不同层提取出来的特征可以相互补充，并且该过程由网络自主学习。该操作包含以下步骤：

（1）输入相同尺寸的特征图 F_1、F_2、F_3 进行对应位置元素相加得到融合特征 F。再对融合特征 $F = [f_1, f_2, \cdots, f_c]$ 在通道维度上进行全局平均池化操作，得到代表每个通道上全局信息的特征 $S \in R^{1\times c}$，计算公式如下所示

$$S = \frac{1}{h \times w} \sum_{m=1}^{h} \sum_{n=1}^{w} f_c(m,n) \tag{12}$$

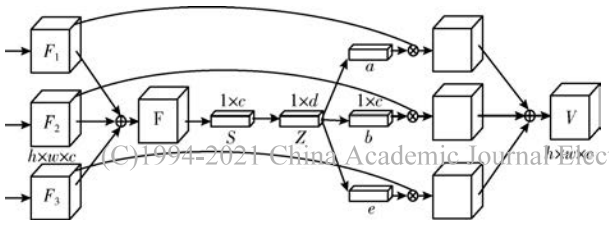

图 5　改进 SK 网络结构

（2）特征 S 经过两层全连接层，第一层对特征 S 进行降维得到特征图 $Z \in R^{1\times d}$；第二层对特征图 Z 进行升维，然后使用 Softmax 函数激活，生成各层卷积特征的注意力权重 $a, b, e \in R^{1\times c}$。具体计算公式如下所示

$$Z = \mathrm{ReLU}(W_1 S) \tag{13}$$

$$a = \frac{e^{W_a Z}}{e^{W_a Z} + e^{W_b Z} + e^{W_e Z}}, b = \frac{e^{W_b Z}}{e^{W_a Z} + e^{W_b Z} + e^{W_e Z}},$$

$$e = \frac{e^{W_e Z}}{e^{W_a Z} + e^{W_b Z} + e^{W_e Z}} \tag{14}$$

式中：$W_1 \in R^{d\times c}$ 表示第一层全连接层的参数，W_a、W_b、$W_e \in R^{c\times d}$ 表示第二层全连接层的参数。

（3）根据计算的注意力权重对特征图 F_1、F_2、F_3 加权更新并融合，得到融合后的特征 $V = [v_1, v_2, \cdots, v_c]$，如式（15）所示

$$V = a \cdot F_1 + b \cdot F_2 + e \cdot F_3 \tag{15}$$

最终得到融合后尺寸为 $1 \times 1 \times N$、$2 \times 2 \times N$、$4 \times 4 \times N$ 的特征图，并将 3 种不同尺度的特征输入到融合层中。

融合多尺度特征。融合层结构如图 6 所示，先通过 flatten 操作将特征图展开，scale 操作对展开后的特征使用不同的权重系数来进行缩放，最后通过 concat 操作将多尺度特征连接起来输入到全连接层。scale 操作中的权重系数可以看作去除偏置项的神经元，重要的特征设置较大的权重系数，辅助特征设置较小的权重系数，并且设置的权重系数可以在网络学习过程中自适应调节，自动更新不同融合特征的权重值。

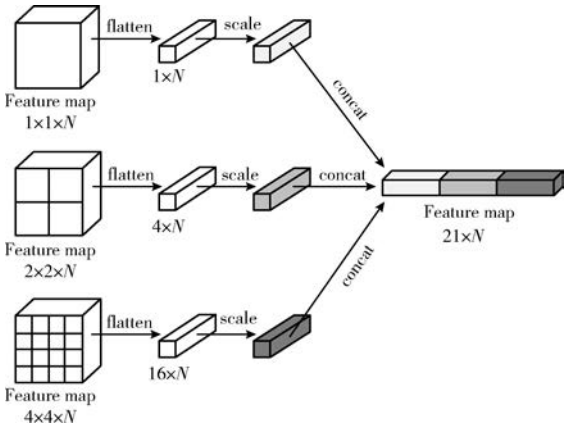

图 6　融合层结构

3　算法验证与分析

3.1　数据集与评价指标

3.1.1　数据集

IAPR TC-12 数据集包括 19 627 张图片和 291 个标注

词，其中 17 665 张图片用于训练，1962 张图像用于测试。数据集涵盖了运动、城市、风景、动物、建筑物和植物。训练集中平均每张图片包含 5.7 个标注，平均每个标签标注 347.7 张图片，最少标注词的训练样本量只有 44 张，最多标注词的训练样本量有 4999 张。

ESP game 数据集包括 20 770 张图片和 268 个标注词，其中 18 689 张图片用于训练，2081 张图像用于测试。数据集涵盖了徽标、绘画、风景和个人肖像。训练集中平均每张图片包含 4.7 个标注，平均每个标签标注 326.7 张图片，最少标注词的训练样本量只有 18 张，最多标注词的训练样本量有 4553 张。

3.1.2 评价指标

(1) 平均准确率 P。计算数据集中每个标签正确预测占实际预测的比例，并根据该数据集中的标签类别数量进行求和平均，计算公式如下

$$P=\frac{1}{N}\sum_{i=1}^{N}\frac{\text{Precision}(y_i)}{\text{Prediction}(y_i)} \tag{16}$$

式中：N 表示标签类别数；$\text{Precision}(y_i)$ 表示在数据集中正确预测标签 y_i 的总数；$\text{Prediction}(y_i)$ 表示在数据集中预测标签 y_i 的总数。

(2) 平均召回率 R。计算数据集中每个标签正确预测占真实标注的比例，并根据该数据集中的标签类别数量进行求和平均，计算公式如下

$$R=\frac{1}{N}\sum_{i=1}^{N}\frac{\text{Precision}(y_i)}{\text{Ground}(y_i)} \tag{17}$$

式中：N 表示标签类别数；$\text{Precision}(y_i)$ 表示在数据集中正确预测标签 y_i 的总数；$\text{Ground}(y_i)$ 表示在数据集中真实标注标签 y_i 的总数。

(3) 综合性能 F_1。由于平均召回率和平均准确率都是重要的评价指标，只有当平均召回率和平均准确率都高时，模型才有良好的性能。因此，需要计算 F_1 值，以反映模型的综合性能，计算公式如下

$$F_1=\frac{2\times P\times R}{P+R} \tag{18}$$

(4) N^+ 指数。统计至少正确预测过 1 次的标签个数，表示模型在数据集所有标签上的覆盖性能，计算公式如下

$$N^+=\sum_{i=1}^{N}\text{Sgn}(y_i) \tag{19}$$

式中：N 表示总的样本数；$\text{Sgn}(\cdot)$ 表示符号函数计算公式如下

$$\text{Sgn}(x)=\begin{cases}1, & x>0\\ 0, & x=0\end{cases} \tag{20}$$

3.2 实验环境与参数设置

实验基于 Tensorflow 深度学习框架，使用 NVIDIA TITANXp GPU 进行计算，操作系统为 Ubuntu16.04，编程语言为 Python。训练中参数设置见表 1。

表 1 参数设置

参数名称	参数
batch_size	50
学习率	0.0005
激活函数	ReLU
优化方法	Adam
权重控制系数 β	1
调焦参数 γ	2

3.3 实验方案及结果分析

3.3.1 实验方案

方案 1：探究 SK 网络中降维全连接层节点数 d 对网络性能的影响。将降维全连接层节点数分别设置为 32，64，128 进行对比实验。

方案 2：损失函数比较。使用多标签损失（ML Loss）函数和代价敏感的多标签损失（CSML Loss）函数训练本文设计的网络与原始 VGG16 进行对比实验。

方案 3：本文方法与其它图像标注方法进行对比。与近些年提出的先进方法进行对比，涉及方法包括：KCCA、2PKNN_ML、SEM、SNDF、ADA、CNN-Regression、CNN-MSE 和 CNN-MLSU。

3.3.2 结果分析

方案 1：降维全连接层节点数对网络性能影响的实验结果见表 2。

表 2 不同融合方案性能对比

方案	IAPR TC-12				ESP game			
	P	R	F_1	N^+	P	R	F_1	N^+
d=32	0.449	0.424	0.437	279	0.412	0.397	0.404	255
d=64	0.434	0.422	0.428	279	0.418	0.399	0.408	255
d=128	0.420	0.355	0.385	252	0.406	0.348	0.375	240

从表 2 可以看出，当降维全连接层节点数为 32 和 64 时，网络都可以取得较优的性能；当降维全连接层节点数为 128 时，网络性能较差。节点数在取 32 时，不仅能够保证网络性能，而且还可以减少网络的参数量。因此，本文网络将降维全连接层节点数 d 设置为 32。

方案 2：损失函数比较方案的实验结果见表 3、表 4。

表 3　损失函数性能对比（IAPR TC-12）

方案		IAPR TC-12			
		P	R	F_1	N^+
VGG16	ML Loss	0.419	0.356	0.385	264
	CSML Loss	0.415	0.379	0.396	269
本文网络	ML Loss	0.410	0.393	0.401	260
	CSML Loss	0.449	0.424	0.436	279

表 4　损失函数性能对比（ESP game）

方案		ESP game			
		P	P	F_1	N^+
VGG16	ML Loss	0.351	0.340	0.345	243
	CSML Loss	0.385	0.376	0.380	252
本文网络	ML Loss	0.355	0.356	0.355	253
	CSML Loss	0.412	0.397	0.404	255

对表 3、表 4 分析可以得出，代价敏感多标签损失（CSML Loss）函数相比于多标签损失（ML Loss）函数，在 IAPR TC-12 数据集和 ESP game 数据集上均有较好表现，尤其在平均召回率和 N^+ 指数上有明显提升。N^+ 指数和平均召回率可以表明本文提出的损失函数能够缓解训练中标注类别不平衡对网络的影响，提升对低频词的标注性能。表中数据还可以分析出本文改进的网络相比于原始 VGG16 取得了更好的效果，并且本文网络的参数仅为 0.77 亿个，远小于 VGG16 中 1.35 亿个参数。

方案 3：对比不同图像标注方法在 IAPR TC-12 和 ESP game 数据集上的平均准确率 P、平均召回率 R 和综合性能 F_1。表 5 给出了本文方法与其它图像标注算法在 IAPR TC-12 数据集上的实验结果对比，表 6 给出了本文方法与其它图像标注算法在 ESP game 数据集上的实验结果对比。

表 5　本文算法与其它图像标注方法实验结果性能对比（IAPR TC-12）

方法	P	R	F_1
KCCA[4]	0.44	0.34	0.38
SEM[5]	0.41	0.39	0.40
2PKNN_ML[6]	0.53	0.32	0.40
CNN-MLSU[7]	0.44	0.38	0.41
CNN-MSE[8]	0.35	0.40	0.37
CNN-Regression[11]	0.49	0.31	0.38
SNDF[12]	0.48	0.30	0.37
ADA[13]	0.42	0.30	0.35
本文方法	0.45	0.42	0.44

表 6　本文算法与其它图像标注方法实验结果性能对比（ESP game）

方法	P	R	F_1
KCCA[4]	0.34	0.38	0.36
SEM[5]	0.38	0.42	0.40
2PKNN_ML[6]	0.43	0.26	0.32
CNN-Regression[11]	0.44	0.28	0.34
SNDF[12]	0.50	0.29	0.37
ADA[13]	0.35	0.21	0.26
本文方法	0.41	0.40	0.40

通过表 5 可以看出，本文方法相比于最近提出的 SEM 方法，在平均召回率与平均准确率上高出 3 个和 4 个百分点；与同样使用卷积神经网络的 CNN-MLSU 方法相比，本文方法在平均召回率和平均准确率上高出 4 个和 1 个百分点。综合来看本文方法在 IAPR TC-12 数据集上与其它的方法相比，虽然平均准确率低于 2PKNN_ML 方法，但本文模型的平均召回率和综合性能 F_1 优于其它方法。通过表 6 可以看出，在 ESP game 数据集上，本文提出的方法较其它方法在各项评价指标上都有较好表现，与较先进的 SEM 方法相比，虽然在平均召回率上存在差距，但在平均准确率上优于 SEM，在综合评价指标 F_1 值上也不相上下。从整体来看本文提出的方法较其它方法在平均准确率和平均召回率上都取得一个较好的结果，从而使得综合性能 F_1 与其它方法相比具有明显的提升。

表 7 列出了本文方法在 IAPR TC-12 测试集上有代表性的标注结果，每幅测试图像根据本文方法给出的结果选择概率最大的前 5 个标签作为图像的标注结果。其中表 7 的第一和第二个示例，场景简单且图像中物体特征明显，本文方法得出的标注结果与真实标签匹配度高。在第二和第三个示例的标注中“people”，“man”以及“house”，“building”是具有相近语义的标注词，本文方法虽然未能准确预测出真实的标签，但预测出的标签同样也符合图像的语义。表中第三个示例，真实标注显然遗漏了标签“sky”，该标签在图像中占据了很大的区域；在第四个示例中，“camera”和“hat”被识别标注，但由于其在图像中占据区域较小而被真实标注忽略，事实上“camera”和“hat”也符合图像本身的语义。表中第三、第四个示例中预测的新标签是对图像中真实标签的扩充，能够更加精确地描述图像的语义信息。

4　结束语

本文在 VGG16 的基础上加入特征融合机制，融合多尺度特征提高对图像中不同尺度对象的标注能力。同时，引入代价敏感损失函数，在一定程度上提升了低频标签的召回

表 7 IAPR TC-12 数据集上的预测效果

图像	真实标签	预测标签
	bed,blanket lamp,room wall,wood	bed,blanket lamp,wall wood
	cap,fence front,man sky,tree	cap,fence **people**,sky tree
	car,hill house,lamp road,street tree,village	car,**building** road,**sky** tree
	desert,middle rock,tourist	**camera**,desert **hat**,**man** **sky**

率，有效解决了训练过程中标签类别不平衡引发的问题。实验结果表明，本文提出的方法在标注性能上有所提升，优于其它经典方法和近年来所提出的先进方法。但本文方法未探究标注词之间的关系，无法通过标注词之间的关系来改善标注结果。如何挖掘标注词之间错综复杂的关系，是未来研究的关键问题。

参考文献：

[1] Marin-Castro HM, Hernandez-Resendiz JD, Escalante-Balderas HJ, et al. Chained ensemble classifier for image annotation [J]. Multimedia Tools and Applications, 2019, 78 (18): 26263-26285.

[2] LIU Kai, ZHANG Limin, SUN Yongwei, et al. An automatic image annotation algorithm using deep boltzmann machine and canonical correlation analysis [J]. Journal of Xi'an Jiaotong University, 2015, 49 (6): 33-38 (in Chinese). [刘凯，张立民，孙永威，等. 利用深度玻尔兹曼机与典型相关分析的自动图像标注算法 [J]. 西安交通大学学报，2015，49 (6)：33-38.]

[3] ZHANG Lei, CAI Ming. Image annotation based on topic fusion and frequent patterns mining [J]. Computer Science, 2019, 46 (7): 246-251 (in Chinese). [张蕾，蔡明. 基于主题融合和关联规则挖掘的图像标注 [J]. 计算机科学，2019，46 (7)：246-251.]

[4] Uricchio T, Ballan L, Seidenari L, et al. Automatic image annotation via label transfer in the semantic space [J]. Pattern Recognition, 2017, 71: 144-157.

[5] Ma Y, Liu Y, Xie Q, et al. CNN-feature based automatic image annotation method [J]. Multimedia Tools and Applications, 2019, 78 (2): 3767-3780.

[6] Verma Y, Jawahar CV. Image annotation by propagating labels from semantic neighbourhoods [J]. International Journal of Computer Vision, 2017, 121 (1): 126-148.

[7] GAO Yaodong, HOU Lingyan, YANG Dali. Automatic image annotation method using multi-label learning convolutional neural network [J]. Journal of Computer Applications, 2017, 37 (1): 228-232 (in Chinese). [高耀东，侯凌燕，杨大利. 基于多标签学习的卷积神经网络的图像标注方法 [J]. 计算机应用，2017，37 (1)：228-232.]

[8] WANG Peng, ZHANG Aofan, WANG Liqin, et al. Image automatic annotation based on transfer learning and multi-label smoothing strategy [J]. Journal of Computer Applica-tions, 2018, 38 (11): 3199-3203 (in Chinese). [汪鹏，张奥帆，王利琴，等. 基于迁移学习与多标签平滑策略的图像自动标注 [J]. 计算机应用，2018，38 (11)：3199-3203.]

[9] KE Xiao, ZHOU Mingke, NIU Yuzhen. End-to-end automatic image annotation based on deep cnn and multi-label data augmentation [J]. IEEE Transactions on Multimedia, 2019, 21 (8): 2093-2106.

[10] Li X, Wang W, Hu X, et al. Selective kernel networks [C] //IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 510-519.

[11] Murthy VN, Maji S, Manmatha R. Automatic image annotation using deep learning representations [C] //Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015: 603-606.

[12] KE Xiao, ZHOU Mingke, NIU Yuzhen. Automatic image annotation combining semantic neighbors and deep features [J]. Pattern Recognition and Artificial Intelligence, 2017, 30 (3): 193-203 (in Chinese). [柯逍，周铭柯，牛玉贞. 融合深度特征和语义邻域的自动图像标注 [J]. 模式识别与人工智能，2017，30 (3)：193-203.]

[13] ZHOU Mingke, KE Xiao, DU Mingzhi. Enhanced deep automatic image annotation based on data equalization [J]. Journal of Software, 2017, 28 (7): 1862-1880 (in Chinese). [周铭柯，柯逍，杜明智. 基于数据均衡的增进式深度自动图像标注 [J]. 软件学报，2017，28 (7)：1862-1880.]