一种基于原型学习的多示例卷积神经网络

何克磊[1)] 史颖欢[1)] 高 阳[1)] 霍 静[1)]
汪 栋[2)] 张 缨[2)]

[1)](南京大学计算机软件新技术国家重点实验室 南京 210093)
[2)](中国人民解放军第八一医院 南京 210002)

摘 要 卷积神经网络是一种全监督的深度学习模型,其要求样本类标完整.在样本类标缺失等弱监督的实际应用中,卷积神经网络的应用受到了极大的制约.为解决弱标记环境下的多示例学习问题,该文提出了一种新的多示例深度卷积网络模型.该模型引入了一种新的原型学习层.该层使用基于原型度量的算法,实现了示例特征至包特征的映射,从而使网络能够在包的层面给予类标信息,进而完成整个模型的学习过程.该文首先在肺癌病理图像细胞分类的问题中,验证了该网络的性能.实验表明,相较于传统基于手工图像特征的方法,该文所提出的方法在准确率方面约有 12%的提升.相较于卷积神经网络结合传统多示例学习的方法,所提出的方法在各项指标上同样取得了更好的效果.此外,在自然图像分类数据集 GRAZ-02 上,所提出的方法相较于目前最优的算法也取得了相当的效果.

关键词 深度学习;多示例学习;原型学习;卷积神经网络;图像分类;人工智能
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2017.01265

A Prototype Learning Based Multi-Instance Convolutional Neural Network

HE Ke-Lei[1)] SHI Ying-Huan[1)] GAO Yang[1)] HUO Jing[1)]
WANG Dong[2)] ZHANG Ying[2)]

[1)](*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093)
[2)](*Bayi Hospital, Nanjing* 210002)

Abstract Convolutional neural network is a fully supervised deep learning model. It requires that the labels of samples are fully provided. In weakly supervised applications where labels of samples are partly provided, the usage of convolutional neural networks is greatly limited. To solve the weakly supervised multi-instance learning problem, a new multiple instance convolutional neural network is proposed. The proposed model introduces a new prototype learning layer into the network. The prototype learning layer uses a prototype based metric method to transform instance features into bag features. The network therefore can use label information of bag and learning the whole model in a compact process. The network is firstly tested on a lung cancer cell pathology image classification dataset. Results show, compared with hand designed image feature based methods, the proposed method achieved an improvement of about 12% in accuracy. Compared with convolutional neural network and multi-instance learning combined methods, the

收稿日期:2015-09-21;在线出版日期:2016-05-09.本课题得到国家自然科学基金(61432008,61305068)、江苏省自然科学基金(BK20130581)资助.**何克磊**,男,1989 年生,博士研究生,主要研究方向为计算机视觉、模式识别、医学图像分析. E-mail: hekelei@gmail.com. **史颖欢**,男,1984 年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为机器视觉、医学图像分析. **高 阳**(通信作者),男,1972 年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为大数据分析、人工智能. E-mail: gaoy@nju.edu.cn. **霍 静**,女,1989 年生,博士研究生,主要研究方向为计算机视觉、人脸识别. **汪 栋**,男,1963 年生,硕士,主要研究方向为医学心胸外科. **张 缨**,女,1971 年生,硕士,主要研究方向为医学病理学诊断.

proposed method also achieved better results on all the evaluation criterion. Besides, the method is also tested on a natural image classification dataset (GRAZ-02). Comparable result is achieved by the proposed method compared with the state-of-the-art method.

Keywords deep learning; multi-instance learning; prototype learning; convolutional neural network; image classification; artificial intelligence

1 引 言

目前绝大多数行之有效的深度网络关注于全监督的情形[1-3]. 而在现实问题中,类标信息往往是弱监督的[4-6]. 在这样的数据集中使用以往全监督的深度网络会受到噪声的明显干扰,因其基于训练集中的所有示例对于网络的贡献都是均等而无差别的假设. 解决弱监督样本集的训练问题,一直以来都是机器学习领域的研究重点,并形成了系统的理论架构,其一是多示例学习. 多示例学习[7]思想由 Dietterich 等人首次提出,被用于预测药物分子活性. 多示例学习的假设是将样本集看作是一个包含了很多包的集合,每一个包中包含了若干数量的示例(其他机器学习任务中的样本特征),每个包中的示例数量是任意的. 当且仅当一个包中最少有一个示例为正时,这个包是正包,反之,是负包. 训练集中包的标记已知而示例的标记未知,目标则是在示例标记不可见的情形下对包的标记进行预测. 这样的假设建立了一个对于样本集的更高级的抽象,从关注对示例类标的预测转为关注包类标的预测,从而避免直接对每一个示例样本进行预测. 一些经典的多示例学习算法已经取得了显著的效果[4,8,9]. 但这些算法都是基于手工抽取的特征,例如 SIFT,LBP 等. 抽取特征的方法往往是通用的,对样本集不具有特异性. 抽取出的特征往往是浅层的,不能够有效地描述样本的结构信息.

近年来,通过深层卷积神经网络(Convolutional Neural Network,CNN)抽取得到的特征已被证实在图像分类任务中能够取得卓越的效果[10]. 卷积神经网络可以抽取出层次化结构化的目标的深层特征表示信息,其抽取特征的方法是基于数据集学习得到的,更加贴近输入数据的分布特点. 但是以往的深度学习模型关注于全监督的情形,所有的输入样本都与其类标一一对应,这样制约了其模型在使用上的灵活性. 因而如果将学习得到的深度特征与多示例学习算法结合可以解决弱监督条件下的图像分类问题,提高算法性能.

因此,本文设计了一种新型的基于原型学习的多示例深度卷积神经网络,可以将示例层面通过学习所得的深层特征表示映射至包的特征的层面,使用映射后的包的特征表示进行图像分类. 将多示例方法与深层神经网络学习的特征进行结合的想法是直接的. 虽然这样的想法还处于早期阶段,我们注意到已经有一些文献使用了这样的思想[11,12]. Xu 等人[11]的方法通过分别学习一个深层的神经网络获得特征并学习一个多示例的分类器得出最终预测,并通过实验证明了 CNN 在医学图像中的有效性. 但其并没有将多示例和深度学习方法结合成一个统一的整体,示例的特征抽取以及多示例分类是两个独立的阶段. Wu 等人[12]的方法则是通过使用一个深层神经网络对示例进行预测后将示例的类标进行综合得到包的类标,在自然图像分类实验中取得了令人满意的效果. 根据以往工作总结[13],不进行示例分类,而是采用示例之间关系的学习抽取包的特征,直接对包进行分类的多示例方法,通常效果更好.

此外,传统的多示例学习假设由于其较大的限制,不能随意应用到诸多其他领域. 因此,在图像分类任务中,一种扩展的多示例假设被提出来,其放松了对于示例与包之间关系的约束. 在图像分类中,将待分类的图像看作是包,而将图像中的子块看作是示例. 那么,正的示例是在一副图像中被标记为正有相关关系的一些图像子块,反之,负的示例则是其中没有这些相关关系的一些图像子块. 一些之前的图像分类工作广泛的应用了这样的假设[13]. 基于我们的观察,在图像分类问题中,假设以图像为包,图像中的目标(对应图像上的图像块)作为示例,如图 1 所示,图像上通常包含了很多的噪声示例,而这些噪声示例具有较大的形式上的差异,使得这些噪声示例不能够被准确预测出其具体类标,因而影响到基于示例类标预测的一些基于示例层面的多示例学习算法. 这些多示例学习算法常常综合预测得到的示

例类标，通过一些求和、平均或投票等算法预测包的类标. 明显地，其效果会因示例预测准确性而有所下降. 因而本文算法从包的特征层面进行分类并组建了一个完整的深度多示例学习模型. 其主要思路是，通过从示例中选取原型示例，将包中示例到原型示例之间的最小距离作为包的特征，因此构建出包层面的分类器，直接对包的特征进行分类. 该方法的优势是不依赖于示例的标记，对包的标记预测具有更高的准确率. 同时，本文所提方法基于深度神经网络实现，具有深度学习算法的一般优点，因此较适合于如今大样本量的数据集特点，但需要指出的是，由于本文使用了图像子块的集合而不是整张图像作为深度网络的输入，因此其单幅图像的计算时间会因使用图像子块的数目而加倍，但会使得该算法具有更高的鲁棒性.

本文的主要贡献有：(1) 提出了一种基于原型的多示例深层卷积神经网络，能够解决弱监督情况下的图像分类问题；(2) 在基于原型的多示例卷积神经网络中引入了一种基于原型的包特征表示方法，能够在包的层面学习包中示例之间的关系，而不仅仅是示例的特征；(3) 将模型应用到肺癌图像分类以及自然图像分类数据集中，取得了较好的效果.

本文第 2 节介绍相关工作；第 3 节对所提的网络进行整体的展示；第 4 节则详细介绍了样本特征的学习过程，尤其是原型层的学习方法；第 5 节说明整个网络的学习策略以及参数更新的计算方法；最后，在实验中验证了所提模型的效果并进行总结.

2　相关工作

根据度量对象的层次不同进行划分，多示例的学习方法可以被分为两大类：基于示例层度量的方法和基于包层度量的方法[13]. 其中，基于示例层度量的方法基于平均贡献假设，假设包中的所有示例对于包的标记的贡献都是相等的. 很多多示例的方法使用了这样的假设设计算法[4-6]. 这一类的方法独立地对每一个示例通过推理去判别他们的标签，因而对于分类器的训练是各自独立的，在判别一个示例标签时并没有使用同一个包中其他示例的信息. 而 Amores[13] 在他的文章中验证了使用包层面的特征判别的算法比使用示例层面特征判别的算法更能得到好的结果，同时节省了时间的开销. 因此，同样有一些算法基于包的层面的特征度量[14-16]. 这样的方法用机器学习的术语可以称为上层建模. 这些经典的多示例算法，以往主要使用的是手工设计的特征而非深度学习的特征，给算法性能带来了限制.

深度学习方法近年常被用来学习样本的深层的特征表示，常用的模型主要为两种：其中一种是擅长学习数据的结构的生成式深度玻尔兹曼机(DBM)模型；另一种为擅长处理图像等具有二维空间信息的图像等数据的判别式卷积神经网络(CNN)模型. 这些的方法被用在诸多的机器学习任务之中并且被证明是有效的，例如语音识别[1]、人脸识别和核实[2,3]、目标检测[17]等等. 尤其是在图像分类识别领域中，CNN 取得了比以往图像分类模型更好的效果，但是其针对的任务是全监督的，训练集中的每一个样本在网络的训练中都被假定有同样的价值. 但是在一些常见的场景中，需要处理弱监督的学习问题，数据中总是有大量的噪声数据，我们不能给出每一个样本一个确切的类标. 目前已经有一些学者关注到了这样的问题，Xu 等人[11] 提出了一种结合

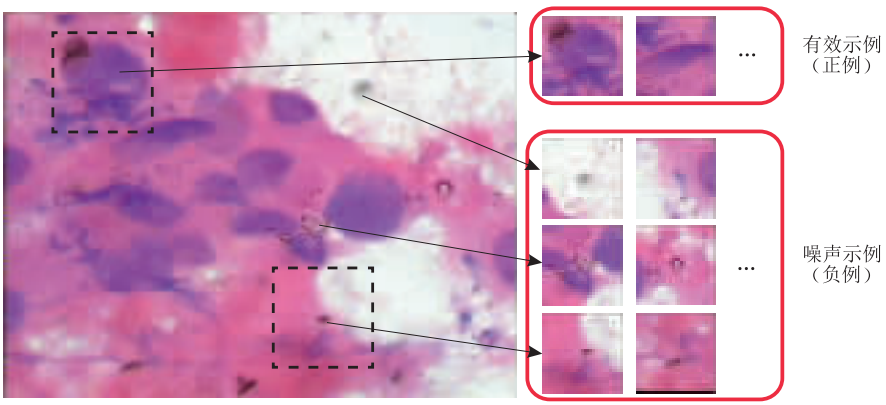

图 1　肺癌图像分类问题和多示例问题的对应关系

CNN 抽取特征和多示例分类器的方法去解决医学图像的分类问题. 但是他们将整个算法设计成两个阶段,训练 CNN 和多示例的分类器分开进行,使得残差并不能够通过回传消除,而我们的模型将多示例的分类器纳入到一个整合的深度学习网络中来. Wu 等人[12]使用一个整合的 CNN 学习示例图像的特征并得到该示例的预测标记,并综合包中所有示例的预测标记预测出包的预测标记信息,该方法虽然是一个整合的多示例深度网络,然而他们关注的是通过对示例分类后将示例的类标进行综合得到包的类标,即基于示例层度量的方法,并不能够学习示例之间的关系. 不同之处是,本文所提的方法是基于包层度量的方法.

3 基于原型学习的多示例深层卷积神经网络

基于原型的多示例深度学习网络的整体结构见图 2. 该网络主要由三部分构成:第 1 部分是抽取示例特征的深度卷积神经网络;第 2 部分为原型示例选取层,该层的输入为由一个预训练的深度卷积神经网络抽取的示例特征,输出为一系列原型示例的特征;第 3 部分将原型示例特征以及所有的示例特征,通过包特征变换函数,将示例特征映射为包的特征,最后对包的特征进行分类,判断一个包的类别. 总的来说,待分类的所有示例特征会被输入一个经过预训练的 CNN 网络(通常使用 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 12 数据集进行预训练)得到所有示例特征的集合,如图 2 中虚线框中所示,通过一些可能的原型选取方法,学习得到基于该数据集的原型特征向量集. 之后,在训练和测试阶段,一个待分类的图像(包)上的所有被选取出的图像子块(示例)通过一个深度的卷积神经网络,从而抽取出示例特征. 这些示例特征会被输入到一个原型学习层. 原型学习层构建了一个基于距离的特征表示器,根据之前的原型选取过程得到的原型示例的向量集,通过一个包特征转换函数,示例特征将被转换为包的特征. 最后,对于包的 Softmax 回归函数对包的特征进行分类,得到包的类别. 最终整个网络以包的分类损失为整体的网络损失进行反向微调. 网络的整体学习过程将在下文进行具体的介绍.

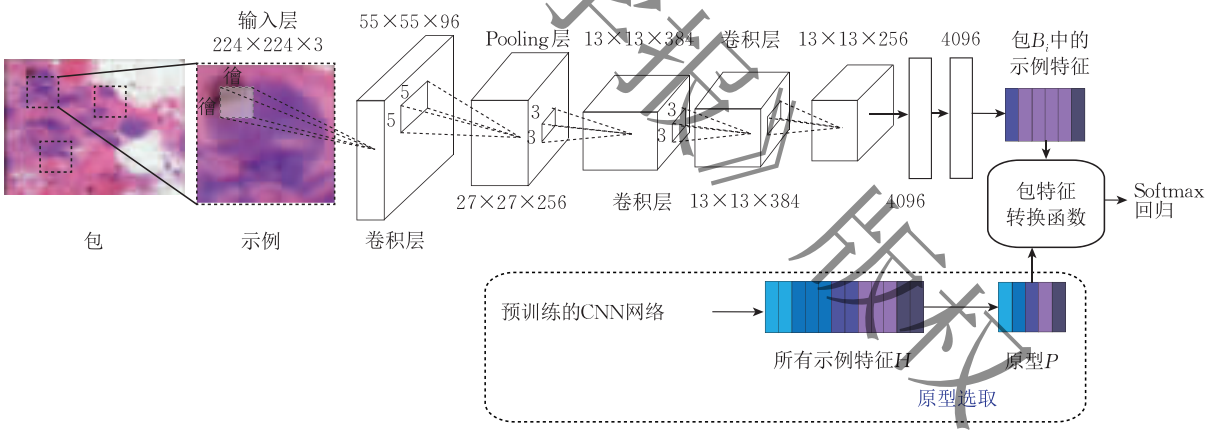

图 2 基于原型学习的多示例卷积神经网络的整体结构

4 网络结构

多示例学习问题中,训练样本通常由一系列的包构成,假设包的表示为$\mathcal{B}_i=\{x_{i1},x_{i2},\cdots,x_{in_i}\}$,$i=1,\cdots,m$,其中 m 是包的个数,$x_{ij}\in\mathbb{R}^d$ 为第 i 个包中的第 j 个示例,d 为示例的维度,n_i 为第 i 个包中示例的个数,在本文所研究的问题中,包为一幅图像,示例为图像上有目标的区域对应的一个图像小块. 假设包对应的标记为 $Y_i\in\{0,1\}^k$,表示训练样本中有 k 个类别,Y_i 中仅有一个元素取 1,取 1 元素的索引为包对应的类别. 输入数据首先被用来训练一个深度的卷积神经网络,深度的卷积神经网络由一系列成组的卷积(Convolution)以及池化(Pooling)层构成,通过对输入样本进行多次卷积以及池化操作后,输出输入样本的特征表示. 卷积神经网络中的层数是可以进行调整的,本文采用了同参考文献[10]中同样的设置,采用该设置的原因是,本文将所研究的模型主要用于图像分类任务,文献中的网络参数设置在图像分类任务中取得了良好的效果. 本文不

对 CNN 进行详细的回顾，我们将主要阐述本文的贡献，包括如何学习具有代表性的原型以及得到具有判别性的包的特征.

4.1　原型学习层

从示例中选取具有代表性的示例(原型示例)并将示例特征转换为包的特征进行分类的多示例学习方法已经有若干代表性的工作[16,18]. 本文将示例特征的学习、原型示例的学习以及判别性包的特征的分类 3 个部分的内容整合到一个完整的深度学习网络中. 而在以往的方法中，这 3 个部分是分开进行学习和优化的. 接下来首先介绍如何在网络中进行原型示例的学习.

原型学习层的主要目的是从示例数据的分布中学习得到具有代表性的，显著的示例，并且构建一个基于距离度量的包的层面的特征表示，进而完成对于包的类标的预测. 该算法的核心思想是，将每个原型向量看作某个非确定类别(因其是基于学习得到的类别而非传统人工指定的类别)的图像分布的中心. 因而同属于原型向量那一类的示例图像将和原型向量具有很小的距离，不同类别的示例图像与原型之间的距离相对较大，从而在包特征的层面，不同类别的包的特征会在不同的维度表现出不一样的形式. 整体来说，包的特征在属于自己那一类的原型的那几个维度上，特征值将较小. 包特征值的抽取，实际上我们是对示例到原型的相对距离进行了一次池化操作. 相对于卷积神经网络中采用的最大池化，均值池化而言，卷积神经网络中的池化操作是在某一个图像的局部进行计算. 而我们的池化操作是对图像上全局的所有的示例来说，选取与原型最相近的示例的距离，由于仅用到了最有效的示例，该操作中可以消除原图上的噪声图像块对分类的影响. 以下详细介绍该算法的实现过程.

假设以矩阵 $\boldsymbol{H}=[h_1,h_2,\cdots,h_n]\in\mathbb{R}^{d_h\times n}$ 代表通过 CNN 抽取的所有示例的特征对应的特征矩阵，d_h 为 CNN 输出的示例特征的维度，n 为示例的数量. 我们的目标是从整个数据集中学习得到 p 个原型向量，得到这样的原型向量的原型选取方法可以有两种：第 1 种方法是随机选取，得到符合要求的原型向量即可；第 2 种方法是通过聚类方法得到，可以观察到，原型向量可以被看作向量空间中的一些锚点，通过度量这些锚点和示例之间的距离，可以得到示例对于这些原型向量的表示. 因此，可以假设原型向量为空间中的一些中心点，那么这些中心点就能较好地刻画示例在整个空间中的相互关系. 假设需要将数据集分成 k 类，那么可以在空间中找到 k 个这样的中心，而每个中心选取距离最近的 p/k 个样本作为原型向量. 这两种方法的对比，本文将在之后的实验部分给出. 这里主要介绍第 2 种方法所构成的网络，使用第 1 种方法与此类似.

4.2　包特征转换函数

得到原型之后，学习从示例特征映射到包的特征有一系列的转换函数可以实现，本文采用了通过将包中的示例特征与原型示例特征进行距离计算后，选取包中与原型向量距离最近的距离作为一个特征，假设包 $\mathcal{B}_i$ 的特征为 B_i，则具体如下：

$$B_i=\varnothing(\mathcal{B}_i,P)=\begin{bmatrix}\min\limits_{j=1,\cdots,n_i}||\boldsymbol{h}_{ij}-P_1||_2^2\\ \cdots\\ \min\limits_{j=1,\cdots,n_i}||\boldsymbol{h}_{ij}-P_p||_2^2\end{bmatrix}\tag{1}$$

其中，$\boldsymbol{h}_{ij}$ 表示第 i 个包中第 j 个示例的特征向量，因此，第 i 个包的第 k 维特征为包中的所有示例与原型 P_k 计算欧氏距离的最小值. 值得注意的是，其他的距离度量方法也可以用在上式的计算中替换欧氏距离度量.

4.3　联合的包特征以及示例特征学习

基于原型的多示例卷积神经网络的整体目标是学习一个可以正确分类包的神经网络，这样的目标函数依赖于以下的两个主要因素，包的特征具有判别性以及一个好的分类函数. 本文采用了 Softmax 回归函数作为包的分类函数，Softmax 函数可以直接对包的特征进行多分类. 而包的特征具有判别性取决于具有判别性的示例特征以及原型向量. 因而，多示例深度神经网络的目标函数可定义为

$$L=-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k}1\{Y_j=1\}\log\frac{\mathrm{e}^{-\boldsymbol{\theta}_j^{\mathrm{T}}\frac{B_i}{\beta}}}{\sum\limits_{l=1}^{k}\mathrm{e}^{-\boldsymbol{\theta}_l^{\mathrm{T}}\frac{B_i}{\beta}}}\right]\tag{2}$$

其中，$1\{\cdot\}$ 为判断函数，若大括号中的陈述为真，则其值为 1，反之为 0. θ 为 Softmax 回归函数的参数. $\mathrm{e}^{\boldsymbol{\theta}_j^{\mathrm{T}}B_i}$ 为包 B_i 的类标为 j 的概率，将其做归一化后的误差，使用负的 log 损失进行累积. 容易看出，由于该损失函数在 $Y=1$ 时会最大化第 i 个包属于第 j 类的概率 $\mathrm{e}^{\boldsymbol{\theta}_j^{\mathrm{T}}B_i}$，而如前文所述算法的目标需最小化原型与其最相似示例间的距离，即最小化特征 B_i. 因此，式(2)中，需对包的特征取反. 特别地，由于包的特征 B_i 基于距离的度量，在数量级上会比以信号

传导为基础的神经网络大很多，因此我们用一个基于经验的参数 β 对其进行规范化. 根据经验，参数 β 与示例特征 $\boldsymbol{h}_{ij}$ 的维度有关，在本文的实验中，通常取 $\beta=1000$. 整个函数对包的分类损失定义在包特征层面，将对包的分类损失整体回传，可用于微调整个 CNN 网络，因此该网络可以同时进行示例特征以及包特征的学习.

5 学习方法

精确的对上述的网络进行求解是不可行的，因而我们采用近似求解的方式对上述网络中的参数进行学习. 整体来说，对上述的网络参数进行求解可分为以下的几个步骤，首先，采用传统 CNN 网络的学习方式，学习得到对输入示例输出最优示例特征表示的网络，在传统 CNN 网络的学习过程中，假设示例的类标与包的类标一样来进行训练，这样的设置实际上，示例类标中是存在噪声的，但是对于初始化来说，这样的假设已经足够学习到一个较优的网络参数. 然后，通过学习后的示例特征选取出原型向量. 最后，对整个网络总的参数进行全局的反向微调.

在第 3 步中，反向微调依赖于反向求解各个网络层的参数的导数. 而根据链式法则的原理，示例特征的误差由原型层传递，首先给出原型学习层输入的导数形式：

$$\frac{\partial L}{\partial h_{it}}=\sum_{j\in C_{it}}\frac{\partial L}{\partial B_{ij}}\frac{\partial B_{ij}}{\partial h_{it}}=\sum_{j\in C_{it}}\frac{\partial L}{\partial B_{ij}}(-2(h_{it}-p_j)) \tag{3}$$

式(3)中，C_{it} 为包 $\mathcal{B}_i$ 中使用 h_{it} 计算包特征的维度索引，通过链式求导法则，$\frac{\partial L}{\partial h_{it}}$ 的公式如式(3)，如果 h_{it} 和所有原型的距离都很远，包特征的计算中该示例没有参与，则 C_{it} 为空，$\frac{\partial L}{\partial h_{it}}$ 为 0，对应该示例特征不会被更新，只有对包特征计算有贡献的示例才参与后续更新，该方法为次梯度算法. 之后的误差传递过程与传统的 CNN 相同，通过全局的反向传播，对整个网络进行微调，最终学习得到多示例的深度学习网络，最终测试的过程为，使用学习得到的网络，输入一个包中的所有示例，输出为该包的类标.

6 实 验

这一章本文首先介绍数据集的设置以及对图像进行预处理的方法. 之后本文验证了所提模型在数据集上的验证效果，一些参数的设置以及和现有的方法进行对比.

6.1 数据集设置及预处理

实验采用了两个真实图像数据集：

(1) 肺癌图像数据集：由中国人民解放军八一医院提供的肺癌数据集中共有 4 类肺癌细胞图像(依次标记为核异性癌(NA)，共 21 张，鳞癌(SC)，共 342 张，腺癌(AC)，共 329 张，小细胞癌(SCC)，共 370 张)和 1 类正常肺细胞图像(标记为正常(NC)，共 120 张)，共计 1182 张. 分辨率为 768×576 像素. 由于癌症细胞之间会有重叠和粘连，并且细胞只占到图像的一部分，所以传统的机器学习方法并不能很好地解决这样的 5 分类问题.

对图像的预处理的主要目的是提取图像上包含肺癌细胞的图像区块，以整幅图像作为一个包，一个图像区块作为一个示例. 肺癌图像的预处理方法主要如下，首先将输入图像转到 HSV 空间，提取该图像在 H 通道上的图像，将像素值小于一定阈值的像素值截断为白色，将上述处理得到的图像二值化，然后进行最大连通区域计算，计算连通区域的中心，按该中心，提取图像上一个 227×227 大小的区块，作为该图像上的示例.

(2) GRAZ-02 数据集[19,20]：是一个十分常用的自然图像数据集，包含 Cars，Persons，Bikes 和 None 四类图像，每类的图像数量在 300 到 500 张之间，图像的分辨率为 640×480. 数据集中的图像有很高的空间复杂度，类内差距很大，具体表现在图像中的待识别目标具有多种形状、角度和光照条件，是一个难度较高的图像和目标识别数据集.

该数据的预处理首先采用 BING 算法[21]生成图像上一系列的目标框，然后每幅图像上保留一定数量的目标框中的图像进行缩放后，作为该图像上的示例.

6.2 评价指标

实验中采用的评价指标包括：准确率(Accuracy)、精度(Precision)、召回率(Recall)、F1 值(F1-Score)以及真阴率(True Negative Rate，TNR). 准确率即正确分类的样本数量除以总的样本数量. 精度、召回率、F1 值以及真阴率在实验中均是对 5 个类别分别计算，另外对 5 个类别的数值进行了求平均得到.

6.3 实验结果

本节的实验结果主要验证以下几个方面，所提

方法中原型示例数量对实验结果的影响，所提方法在两个数据集中的效果，所提方法与传统方法（包括采用手工特征的方法、CNN 与传统多示例方法结合的方法）的效果对比.

6.3.1　肺癌图像数据集

在实验中，整个 CNN 网络的参数设置与参考文献[17]相同，在原型学习层，主要实验不同的原型数量对分类准确率效果的影响.

（1）调整原型数量对结果的影响

首先验证本文所提方法中的原型示例数量对分类效果的影响，以肺癌数据集的实验结果为例，表 1 列出了原型数量对最终图像五分类的分类准确率影响，测试的原型数量包括 80，100，150，200，由实验显示，100 的效果最优，因而，在后续实验中，采用的原型数量为 100.

表 1　原型数量与准确率之间的关系（单位：%）

原型数量	Accuracy	Precision	Recall	F1	TNR
50	98.5	98.5	98.5	98.5	99.6
100	99.0	99.0	99.0	99.0	99.8
150	98.5	98.6	98.5	98.5	99.6
200	98.5	98.5	98.5	98.5	99.6
250	98.0	98.0	98.0	98.0	99.5

（2）所提算法在 5 个肺癌图像类别上的分类效果

本节主要分析所提方法在 5 类肺癌图像上的分类效果，图 3 展示了所提算法在 5 类肺癌图像分类问题上的混淆矩阵（confusion matrix）. 从图中可以看出，5 个类别的图像均可以被很好的分开，其中 SC 与 AC 有部分图像不能被很好地分开，其原因主要是 SC 与 AC 图像在外观是比较相近的.

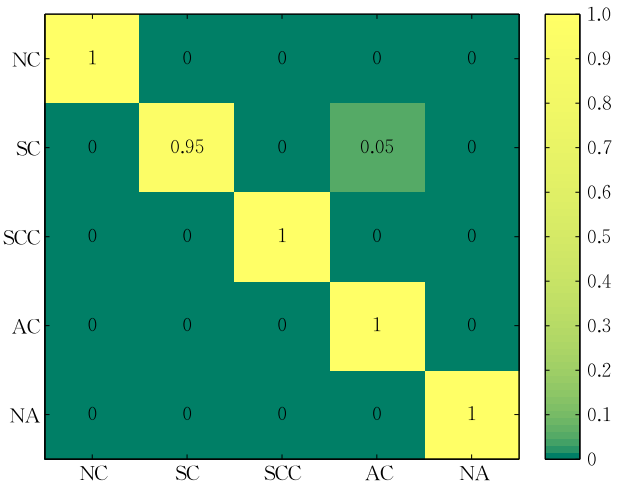

图 3　所提算法在 5 类肺癌图像分类问题上的混淆矩阵，从图中可以看出 5 个类别的图像均可以被很好地分开，其中 SC 与 AC 有部分图像不能被很好地分开

在 5 个类别上的分类效果与 CNN 结合传统分类方法的效果对比，从图 4 中可以看出，在 Precision 指标上，NC 与 NA 的分类效果是最好的，几乎都为 1，所提方法在 SC 的分类效果略差，而在其它 3 类上均优于传统方法. 在 Recall 指标上，同 Precision 指标，NA 与 NC 的分类效果趋近于 1，所提方法在其他几类上均优于以往的算法. 在 F1-Score 的指标上，所提方法也取得较以往算法更好的效果. 在 TNR 的指标上，所提方法仅在 AC 的指标上存在略低于其它方法.

（3）所提算法与其他方法的效果对比

我们将所提算法与相关的方法在上述的 5 个指标上进行了比较，表 2 中前 6 行为传统通过手工方式抽取特征的方法，对比的方法包括 LapRLS[18]，MCMI-AB[22]，mcSVM[23]，ESRC[24]，KSRC[25] 以及 mSRC[26]，可以看到的是传统方法在准确率指标上的最好效果是 86.7%，未达到 90%. 而采用 CNN 结合传统多示例学习方法的效果，均达到了 95%以上. 其中 CNN+SVM 的方法未采用多示例学习，其效果略差于采用了多示例方法 mi-SVM 以及 MI-SVM 的效果. 表 2 中，所提方法-随机以及所提方法-KMeans 分别为原型学习采用随机选取以及采用 KMeans 距离，所提方法-KMeans 在 5 个指标上均取得了最优的效果，验证了本文算法的有效性，然而深度学习算法在该数据集上的分类效果趋近于饱和，因而本文的算法相对于 CNN 结合传统分类方法的效果的优势并未完全体现.

此外，本实验采用经 ILSVRC 12 数据集预训练的 CNN 抽取图像的深层特征，此数据集中并未包含本实验所需验证的肺癌数据等. 但实验结果的对比说明此预训练网络依然具有较高的效能. 因此可以看出，不同种类图像，例如自然图像和医学图像在底层特征上具有某些共性，因此通过共享经自然图像数据集训练的 CNN 依然可以在医学图像中得到较高的效果.

6.3.2　GRAZ-02 数据集

为验证所提算法在普通自然图像分类问题中的适用性，我们在 GRAZ-02 图像数据集上验证了我们的算法. 在实验中，我们将数据集中每一类的 80%作为训练样本，余下的 20%作为测试样本. 在每张图像（包）中，选取 20 个图像子块（示例）.

表 3 将我们所提方法在 ROC-Equal Error Rate 指标上与该数据集上作为基准的算法进行了比较.

表 2 与相关方法在肺癌分类数据集上的对比(每类方法中最优者用粗体表示)

		Accuracy	Precision	Recall	F1	TNR
非深度学习方法	LapRLS[18]	0.625	0.533	0.538	0.657	0.907
	MCMI-AB[22]	0.608	0.585	0.564	0.563	0.899
	mcSVM[23]	0.674	0.598	0.577	0.576	0.921
	ESRC[24]	0.800	0.730	0.884	0.777	0.940
	KSRC[25]	0.830	0.782	0.843	0.804	0.953
	mSRC[26]	**0.867**	**0.834**	**0.913**	**0.862**	**0.962**
深度学习方法	CNN+SVM(baseline)	0.968	0.968	0.968	0.968	0.992
	CNN+mi-SVM	**0.987**	**0.988**	**0.987**	**0.987**	**0.997**
	CNN+MI-SVM	0.981	0.981	0.981	0.981	0.995
	本文所提方法-随机	0.975	0.975	0.975	0.975	0.994
	本文所提方法-KMeans	**0.990**	**0.990**	**0.990**	**0.990**	**0.998**

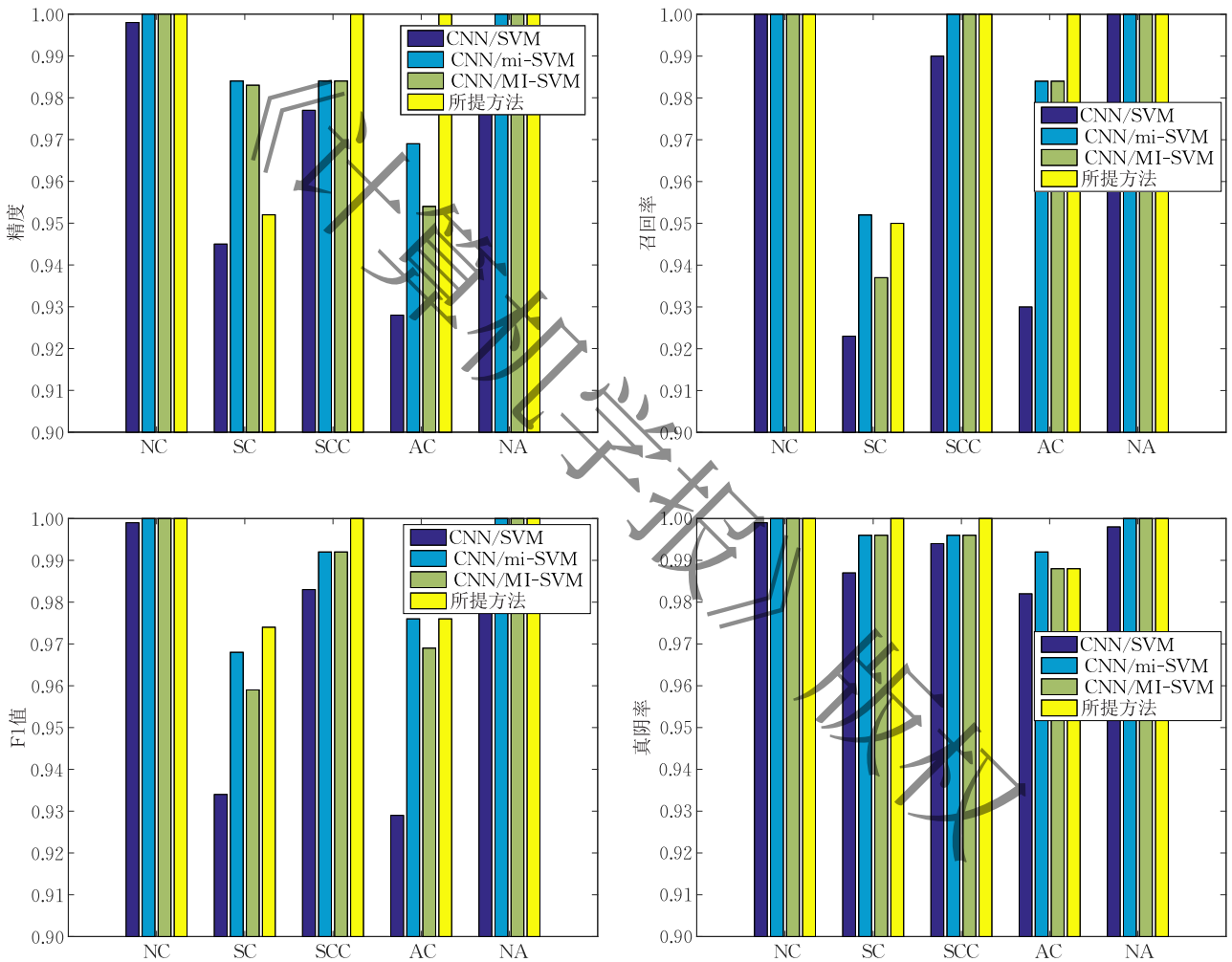

图 4 所提算法与 CNN+传统分类方法在 5 个类标的图像上的分类效果对比

可以看到,我们的方法相较于表 3 中的前 3 个算法取得更好的效果,此外相较于表 3 中的第 4 个算法,我们的算法取得了与之相当的效果.需要说明的是,表 3 中的前 4 个算法均是采用特定的目标检测加识别的算法实现的,训练样本为全监督形式的.而我们的算法采用了 BING 进行目标检测,保留的图像块上有大量的噪声,实验设置为弱监督环境下的.在这样的情形下,所提算法仍然取得了和目前算法相当的效果,从而验证了所提算法在抗示例噪声干扰上的有效性.

表 3 不同方法的 ROC-Equal Error Rate 指标比较

方法	Bikes	Cars	Persons
Opelt 等[19]	77.8	70.5	81.2
Hegazy 等[27]	74.7	81.3	81.3
Zhang 等[28]	88.9	85.2	88.1
Gupta 等[29]	96.0	90.7	89.3
本文所提方法	86.7	85.6	88.5

7 总　结

基于多示例的深度学习方法在图像分类领域目前取得了广泛的关注，本文提出了一种基于原型向量提取的多示例深度学习方法，可以同时学习示例以及包的特征，从而更好地进行包层的分类. 该方法在肺癌图像分类数据集上进行了效果验证，实验表明，相对于传统手工特征抽取的图像分类方法以及CNN结合经典多示例学习的方法，所提方法效果均优于传统方法. 而在GRAZ-02数据集上，所提算法在更难的实验设置下，取得了与以往算法相当的效果，从而验证了方法的有效性.

该方法在解决图像分类的问题上，尚有进一步研究的空间，包括目前的方法中，包特征转换函数的研究，原型学习方法的研究，此外，在图像分类问题中，多示例多标签问题也是值得研究的方向之一，本文的后续工作包括，进一步研究结合深度网络的原型学习方法，包特征转换函数方法以及与多示例多标签结合的深度学习方法等.

参　考　文　献

[1] Hinton G E, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Transactions on Signal Processing Magazine, 2012, 29(6): 82-97

[2] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 1988-1996

[3] Taigman Y, Yang M, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1701-1708

[4] Li W, Duan L, Xu D, et al. Text-based image retrieval using progressive multi-instance learning//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 2049-2055

[5] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2002: 561-568

[6] Han Y, Tao Q, Wang J. Avoiding false positive in multi-instance learning//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2010: 811-819

[7] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 1997, 89(1): 31-71

[8] Zha Z J, Hua X S, Mei T, et al. Joint multi-label multi-instance learning for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8

[9] Zhou Z H, Zhang M L, Chen K J. A novel bag generator for image database retrieval with multi-instance learning techniques//Proceedings of the IEEE International Conference on Tools with Artificial Intelligence. Sacramento, USA, 2003: 565-569

[10] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097-1105

[11] Xu Y, Mo T, Feng Q, et al. Deep learning of feature representation with multiple instance learning for medical image analysis//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy, 2014: 1626-1630

[12] Wu J, Yu Y, Huang C, et al. Deep multiple instance learning for image classification and auto-annotation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3460-3469

[13] Amores J. Multiple instance classification: Review, taxonomy and comparative study. Artificial Intelligence, 2013, 201: 81-105

[14] Babenko B, Verma N, Dollár P, et al. Multiple instance learning with manifold bags//Proceedings of the 28th International Conference on Machine Learning. Bellevue, USA, 2011: 81-88

[15] Babenko B, Yang M H, Belongie S. Visual tracking with online multiple instance learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 983-990

[16] Chen Y, Bi J, Wang J Z. MILES: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12): 1931-1947

[17] Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations//Proceedings of the 26th Annual International Conference on Machine Learning. San Francisco, USA, 2009: 609-616

[18] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. The Journal of Machine Learning Research, 2006, 7(—): 2399-2434

[19] Opelt A, Pinz A, Fussenegger M, et al. Generic object recognition with boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(3): 416-431

[20] Marszatek M, Schmid C. Accurate object localization with shape masks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[21] Cheng M M, Zhang Z, Lin W Y, et al. BING: Binarized normed gradients for objectness estimation at 300fps//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 3286-3293

[22] Zhu L, Zhao B, Gao Y. Multi-class multi-instance learning for lung cancer image classification based on bag feature selection//Proceedings of the International Conference on Fuzzy Systems and Knowledge Discovery. Shandong, China, 2008, 2: 487-492

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

[24] Liu M, Zhang D, Shen D, et al. Ensemble sparse classification of Alzheimer's disease. NeuroImage, 2012, 60(2): 1106-1116

[25] Zhang L, Zhou W D, Chang P C, et al. Kernel sparse representation-based classifier. IEEE Transactions on Signal Processing, 2012, 60(4): 1684-1695

[26] Shi Y, Gao Y, Yang Y, et al. Multimodal sparse representation-based classification for lung needle biopsy images. IEEE Transactions on Biomedical Engineering, 2013, 60(10): 2675-2685

[27] Hegazy D, Denzler J. Generic object recognition using boosted combined features//Proceedings of the 2nd International Workshop on Robot Vision. Auckland, New Zealand, 2008: 355-366

[28] Zhang Z, Li Z N, Drew M S. Learning image similarities via probabilistic feature matching//Proceedings of the IEEE International Conference on Image Processing. Hong Kong, China, 2010: 1857-1860

[29] Gupta N, Das S, Chakraborti S. Hierarchy of visual features for object recognition//Proceedings of the IEEE International Conference on Image Processing. Paris, France, 2014: 5901-5905

HE Ke-Lei, born in 1989, Ph. D. candidate. His current research interests include computer vision, pattern recognition and medical image analysis.

SHI Ying-Huan, born in 1984, Ph. D., lecturer. His current research interests include computer vision and medical image analysis.

GAO Yang, born in 1972, Ph. D., professor, Ph. D. supervisor. His current research interests include big data analytics and artificial intelligence.

HUO Jing, born in 1989, Ph. D. candidate. Her current research interests include computer vision and face recognition.

WANG Dong, born in 1963, M. S., chief physician. His current research interest is cardiothoracic surgery.

ZHANG Ying, born in 1971, M. S., associate chief physician. Her current research interest is medical pathology.

Background

Recently, deep learning has attracted much interest in machine learning society. Deep learning methods have been widely applied in many machine learning tasks with their promising results, e. g., computer vision, speech recognition and neural language processing. However, there are still many unsolved problems for deep neural networks, specifically, for convolutional neural networks (CNN), as CNN is a fully-supervised network, solving weakly supervised tasks in image recognition by CNN is required to be settled.

Previous works on addressing weakly supervised problem in multiple instance learning (MIL) can be classified into two categories. The first is about building MIL classifiers based on hand-crafted features and the second is to learn multiple instance features and classifiers separately. Such two schemes may cause performance degradation, as feature representation and classifier learning are not optimized together. In this paper, we proposed an integrated CNN with MIL via bag feature representation and prediction. The problem is solved in a more reasonable way. Our results show a marked improvement compared with previous works.

Our work is partly supported by the National Natural Science Foundation of China (Nos. 61432008, 61305068), the Natural Science Foundation of Jiangsu Province (No. BK20130581). The projects aim at learning and inferring from big data. Some works of the research fields have been published in the international journals and international conferences, such as AAAI, TNNLS, etc.