软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(11):2879−2890 [doi: 10.13328/j.cnki.jos.005349]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

卷积神经网络特征重要性分析及增强特征选择模型*

卢泓宇[1,2], 张 敏[1,2], 刘奕群[1,2], 马少平[1,2]

[1](智能技术与系统国家重点实验室(清华大学),北京 100084)

[2](清华大学 计算机科学与技术系,北京 100084)

通讯作者: 张敏, E-mail: z-m@tsinghua.edu.cn

摘 要: 卷积神经网络等深度神经网络凭借着其强大的表达能力、突出的分类性能,已在不同领域内得到了广泛应用.当面对高维特征时,深度神经网络通常被认为具有较好的鲁棒性,能够隐含地对特征进行选择,但由于网络参数巨大,如果数据量达不到足够的规模,则会导致学习不充分,因而可能无法达到最优的特征选择.而神经网络的黑箱特性使得无法观测神经网络选择了哪些特征,也无法评估其特征选择的能力.为此,以卷积神经网络为例,首先研究如何显式地表达神经网络中的特征重要性,提出了基于感受野的特征贡献度分析方法;其次,将神经网络特征选择与传统特征评价方法进行对比分析发现,在非海量样本的情况下,传统特征评价方法对高重要性特征和噪声特征的识别能力反而能够超过神经网络.因此,进一步地提出了卷积神经网络增强特征选择模型,将传统特征评价方法对特征重要性的理解结合到神经网络的学习过程中,以辅助深度神经网络进行特征选择.在基于文本的社交媒体用户属性建模任务下进行了对比实验,结果验证了该模型的有效性.

关键词: 卷积神经网络;特征重要性分析;特征选择;文本分类

中图法分类号: TP181

中文引用格式: 卢泓宇,张敏,刘奕群,马少平.卷积神经网络特征重要性分析及增强特征选择模型.软件学报,2017,28(11):2879−2890. http://www.jos.org.cn/1000-9825/5349.htm

英文引用格式: Lu HY, Zhang M, Liu YQ, Ma SP. Convolution neural network feature importance analysis and feature selection enhanced model. Ruan Jian Xue Bao/Journal of Software, 2017,28(11):2879−2890 (in Chinese). http://www.jos.org.cn/1000-9825/5349.htm

Convolution Neural Network Feature Importance Analysis and Feature Selection Enhanced Model

LU Hong-Yu[1,2], ZHANG Min[1,2], LIU Yi-Qun[1,2], MA Shao-Ping[1,2]

[1](State Key Laboratory of Intelligent Technology and System (Tsinghua University), Beijing 100084, China)

[2](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Because of its strong expressive power and outstanding performance of classification, deep neural network (DNN), such as like convolution neural network (CNN), is widely used in various fields. When faced with high-dimensional features, DNNs are usually considered to have good robustness, for it can implicitly select relevant features. However, due to the huge number of parameters, if the data is not enough, the learning of neural network will be inadequate and the feature selection will not be desirable. DNN is a black box, which makes it difficult to observe what features are chosen and to evaluate its ability of feature selection. This paper proposes a feature contribution analysis method based on neuron receptive field. Using this method, the feature importance of a neural network, for example CNN, can be explicitly obtained. Further, the study finds that the neural network's ability in recognizing relevant and noise features is

* 基金项目: 国家自然科学基金(61622208, 61532011, 61672311); 国家重点基础研究发展计划(973)(2015CB358700)

Foundation item: National Natural Science Foundation of China (61622208, 61532011, 61672311); National Program on Key Basic Research Project of China (973) (2015CB358700)

本文由复杂环境下的机器学习研究专刊特约编辑胡清华教授、张道强教授、张长水教授推荐.

收稿时间: 2017-05-15; 修改时间: 2017-06-16; 采用时间: 2017-08-23

weaker than the tratitional evaluation methods. To enhance its feature selection ability, a feature selection enhanced CNN model is proposed to improve classification accuracy by applying traditional feature evaluation method to the learning process of neural network. In the task of the text-based user attribute modeling in social media, experimental results demonstrate the validity of the preoposed model.

Key words: convolution neural network; feature importance analysis; feature selection; text categorization

近年来,神经网络在多个领域均表现出突出的效果和潜力,在图像处理领域,以卷积神经网络(convolutional neural network,简称 CNN)为代表的深度学习模型已经突破了人类自身的识别能力[1,2].而在自然语言处理领域,卷积神经网络、递归神经网络(recurrent neural network,简称 RNN),包括门控递归神经网络(gated recurrent neural network,简称 GRU)、长短时记忆神经网络(long-short term memory,简称 LSTM)等模型,在文本分类[3−5]、语言模型[6]、机器翻译[7,8]、文本生成[9]等应用场景下取得了不错的效果.

特征在机器学习中占据重要作用,但特征数量的增加,往往不一定能给模型带来性能上的提升,这种现象被称为 Hughes 效应[10−12],通常是由于特征数量的增加会显著增大模型训练所需的样本规模,而充足的样本量往往很难获取.此时,其中的无关冗余特征反而会带来过拟合等风险.为了避免过多的特征带来的问题,需要对特征进行筛选,选出重要特征,消除无用、冗余特征,这被称为特征选择[13].

通常,神经网络在训练中能够自动学习并区分特征的重要性,例如,CNN 能够从图片中提取出颜色、纹理等不同粒度的特征,而忽略图片中的背景[14].但在实际应用过程中,神经网络的表现仍受到特征数量的影响,尤其是在样本量相对特征较少时[15,16].在自然语言处理任务下,文本描述多样,特征不仅高维,而且稀疏,这也对模型的特征选择能力提出了较高的要求.以往的研究更多地关注于神经网络最终的分类性能,而对其特征选择这一中间过程的分析相对较少.那么神经网络在文本特征选择上究竟有怎样的表现?这是本文想要研究的主要问题.

由于神经网络特征选择的过程并不能显式地观测到,为了得到其对特征重要性的判断,本文首先提出了基于感受野的特征贡献度分析方法.该方法利用神经元的感受野及网络权重,观察输入样本中特征对神经网络决策的贡献程度,显式地挖掘出神经网络对特征的重要性评估和选择结果.进而通过与传统特征选择方法的对比,来更深刻地理解神经网络的特征选择能力.以卷积神经网络为例,通过遮挡与比较实验,本文发现在非海量样本问题上,神经网络的特征选择能力与传统特征评价方法相比有所欠缺,例如卡方检验等.

更进一步地,本文还提出了卷积神经网络的增强特征选择模型,将传统特征评价方法对特征的理解作为对特征空间的先验知识加入神经网络的训练过程中,来增强神经网络的特征选择能力.在基于文本的社交媒体用户属性建模任务下的对比实验结果,验证了所提出模型的有效性.

本文第 1 节介绍与本文研究相关的工作.第 2 节具体阐述何显式挖掘神经网络的特征重要性,介绍本文提出的基于感受野的特征贡献度分析方法,并通过遮挡实验验证了方法的有效性.第 3 节介绍神经网络特征选择能力和传统特征评价方法的对比实验及结果分析.第 4 节详细介绍增强特征选择的神经网络模型及验证实验结果.最后,第 5 节对全文内容进行总结.

1 相关工作

1.1 神经网络的样本特征分析

近年来,神经网络被广泛地应用,但其通常被认为是黑箱模型,可解释性较差.为了分析神经网络的优势和缺陷,很多研究者通过不同的方法来理解神经网络的分类过程,尤其是对于图像处理领域的 CNNs.例如,Krizhevsky 等人[17]直接可视化第 1 层卷积核的参数来观察神经元学习到的模式;Engelbrecht 等人[18]通过观察最大化激活神经元的图像样本,发现浅层卷积神经元更关注颜色等浅层特征,而深层神经元则更关注纹理等复杂特征;Alain 等人[19]通过在神经网络中添加线性分类器探针来理解深度网络层次之间信息量的变化;Zeiler 等人[14,20]集中于分析输入中的模式与神经网络类别判断的相关性;Samek 等人[21]利用网络的权重来传递计算输入样本不同特征与输出的相关性.

敏感性分析[18,22,23]是神经网络样本特征分析常用的方法,Simonyan 等人[24]应用输出类别概率对输入图像

像素的梯度来分析深层神经网络对图像中敏感模式的识别,揭示与特定类别相关的图像内的显著图(saliency map)模式.Denil 等人[25]使用损失梯度的大小从文本文档中提取关键句.Montavon 等人[26]提出了深度泰勒分解技术,其同样通过对图像像素的偏导数来计算敏感性.

神经网络的损失值通常是输入特征的非线性函数,为了衡量神经网络对于输入特征的敏感性,敏感性分析用一阶泰勒展开来近似神经网络的损失值 $L(\tilde{y}, f(x)) \approx w^T x + b$,则神经网络对输入特征的敏感性:

$$S_{x_{ij}} = \partial L(\tilde{y}, x) / \partial_{x_{ij}},$$

其中,x_{ij} 表示第 i 个特征的第 j 维.进而,Simonyan 等人[24]使用无穷范数,即 $S_{x_i} = \mathrm{Max}(|S_{x_{ij}}|)$ 来表示特征的整体敏感性,而 Li 等人[27]则使用其二范数,即 $S_{x_i} = \sum_j S_{x_{ij}}^2$.

但在严格意义下,敏感性分析并不能得到对模型分类的关键特征,敏感性较高的特征仅代表“当它出现时更易增强或减弱模型对样本类别的判断”,而不能代表“它对模型进行当前决策实际做出了多少贡献”.为了研究神经网络对特征的评价和选择,需要准确衡量特征对神经网络模型整体的重要性,这也是本文第 1 步的研究工作.

1.2 样本特征分析方法的评估

使用不同的样本特征分析方法,得到的特征重要性结果不同.因此,为了量化评估不同的样本特征分析方法的有效性,Samek 等人[28]提出了扰动曲线下的面积(area over the perturbation curve,简称 AOPC)的评估方法.该方法基于假设:当遮挡了样本中对模型较为重要的特征时,会显著影响模型对样本正确类别的判断概率.据此评估特征重要性分析方法的有效性.其实验步骤为:给定一个样本 x,首先,通过样本特征分析方法得到其中每个特征的重要性,依照重要性从高到低排序得到特征序列 $O=(r_1,r_2,\ldots,r_L)$;然后,根据该序列,用大值优先(most first 简称 MF)规则依次遮挡样本中的特征,并重新对遮挡后的样本进行分类,得到分类器对该样本正确类别的预测概率 $f(x_{MF}^{(k)})$.函数 g 代表对 r_k 特征的遮挡操作(例如,将 r_k 设置为全 0 向量),迭代规则为

$$x_{MF}^{(0)} = x; \forall 1 \le k \le L: x_{MF}^{(k)} = g(x_{MF}^{(k-1)}, r_k) \tag{1}$$

则分析方法的有效性被定义为遮挡前后 $f(x)$变化曲线上的面积 AOPC,计算方式为

$$AOPC = \frac{1}{L+1}\left\langle \sum_{k=0}^{L} f(x_{MF}^{(0)}) - f(x_{MF}^{(k)}) \right\rangle_x \tag{2}$$

理想的样本特征分析方法会将对模型分类最重要的特征排在遮挡序列的首端,随着特征词的遮挡,模型的分类性能曲线急剧下降,也随之产生较大的 AOPC 值.

本文也采用该评估方法对所提出的神经网络特征贡献度分析方法及传统的神经网络特征敏感性分析方法进行性能对比评价,以找到合适的方法来理解神经网络对特征的选择,并据此可以将传统特征选择方法与神经网络特征选择方法的效果进行对比分析.

1.3 传统特征选择方法

传统的特征选择方法大多采用单变量特征评价方法,例如皮尔逊相关系数[29]、卡方检验[30]等来衡量不同特征与目标类别的相关性,然后依照相关性的排序选取一定大小的特征子集用于模型的学习[13,31].

卡方检验分析方法针对原假设:特征变量与类别变量相互独立,进行假设检验.通过观察实际值与理论值的偏差来判断原假设是否成立,若实际值与理论值相差较大,则拒绝原假设,表明该特征与类别变量相关.卡方检验值的计算方法如下:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})}{E_{ij}} \tag{3}$$

其中,i 代表特征出现与否,k 为类别数量,A_{ij} 为统计计数,E_{ij} 为依据原假设得到的期望计数.

皮尔逊相关系数则直接计算特征变量 X 与类别变量 C 的相关系数,若相关系数呈现明显的正相关性或负相关性,则该特征对分类较为有用.皮尔逊相关系数的计算方法如下:

$$\rho_{X,C} = \frac{cov(X,C)}{\sigma_X \sigma_C} \tag{4}$$

本文实验中使用卡方检验及皮尔逊相关系数作为传统特征评价方法的代表,将神经网络的特征选择能力与传统特征评价方法进行对比分析,以更深刻地理解神经网络.

通常,机器学习方法,包括神经网络,大多将特征选择放于对数据的处理阶段[16].这样,所选择的特征数量等参数需要人为设定,带来了学习和泛化能力的损失.此外,目前还很少有在神经网络中加入传统特征选择过程的工作.本文基于对神经网络特征选择能力的评估和理解,尝试将传统特征选择方法结合到神经网络的训练过程中,使其能够自适应地进行特征选择,来增强其在高维特征问题上的表现.

2 神经网络的特征重要性分析

为了深入理解神经网络的特征选择,首先需要了解神经网络模型经过训练之后选择了哪些特征,即其认为哪些特征比较重要.传统的样本特征分析方法大多针对图像领域的神经网络,不能直接适用于自然语言处理领域,例如包含嵌入层的卷积神经网络;同时,以敏感性分析为代表的方法不能很好地衡量特征对神经网络决策的重要性.因此,本文提出了基于感受野的特征贡献度分析方法来观测单个输入样本中特征的重要性分布,进而通过对样本全集的统计得到模型整体对每个特征的重要性评价.

2.1 基于感受野的神经网络特征贡献度分析

神经元对应的输入区域称为感受野(receptive field,简称 RF),给定一个输入样本,每个神经元具有不同的感受野,例如,每个卷机神经元仅与输入的部分区域相连.通过观察激活神经元的感受野内特征的分布及其对网络输出的正确类别概率的贡献,可以探究不同特征对网络分类过程的重要性.因此,本文提出了基于神经元感受野的特征贡献度分析方法来评估神经网络在特征选择方面的表现.

本节以卷积神经网络为例,展示基于神经元感受野的特征贡献度分析方法的思想和流程,如图 1 所示.神经网络输出的类别概率被反向传播回卷积神经元的感受野内,分配给了其中的特征项.

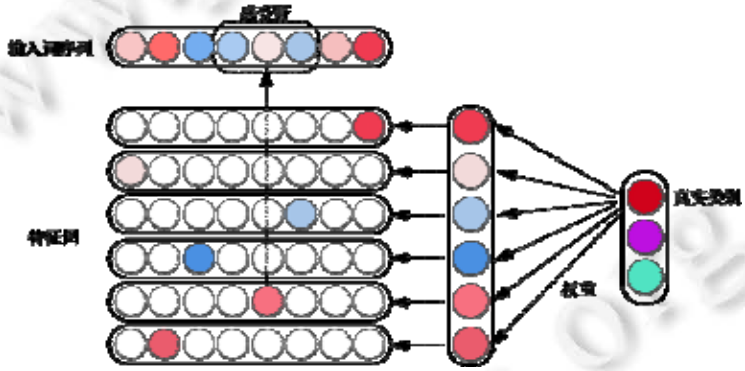

Fig.1 Sketch map of the feature contribution analysis based on receptive field

图 1 基于感受野的特征贡献度分析示意图

我们将反向传播的过程进行如下的形式化定义.

1. 输出层神经元 y_j 的贡献度被初始化为 $C_{y_j}=\delta_{jc}$,δ为克罗内克函数,c 为待观测的类别(例如样本的正确类别).
2. 输出层神经元 y_j 值由池化层神经元 p 经过一层全连接得到,因此 p_i 的贡献度 C_{p_i} 可以通过 C_{y_j} 和相应的全连接层权重 $w_{p_iy_j}$ 计算得到:

$$C_{p_i}=w_{p_iy_j}C_{y_j} \tag{5}$$

3. 最大池化层 p_j 仅保留对应的特征图 fm_i 中最大的一项,赢者通吃,池化神经元的贡献度 C_{p_j} 全部反向传播给特征图 fm_i 最大激活卷积神经元 $conv_{i,k_{\max}}$:

$$conv_{i,k} = I_{k=k_{\max}} C_{p_j} \tag{6}$$

4. 卷积神经元 $conv_{j,k}$ 的激活值由其感受野内特征 w_i 与卷积核参数进行卷积操作得来,因此,w_i 的贡献度 C_{w_i} 可以通过其词向量 x_{w_i} 与卷积核对应位置参数向量的点积得到:

$$C_{w_i} = \sum_j \sum_k I_{i \in RF(k)} conv\_kenel_{i-k+kenel\_size/2} x_{w_i} \times conv_{j,k} \tag{7}$$

已训练好的神经网络模型 M,给定一个样本 w,其可以被视为若干特征的序列(w_1,w_2,w_3,…),借助基于感受野的特征敏感性分析方法,可以得到其中每个特征对神经网络特定类别(例如该样本的真实类别 $\tilde{y}$)预测概率 $p_{\tilde{y}}$ 的贡献度(imp_{w_1},imp_{w_2},imp_{w_3},...).由于特征选择体现在特征的全局整体的重要性,进而本文提出神经网络的全局特征重要性计算方法.

对于一个训练好的神经网络,将其训练集样本重新通过网络,记录网络中每一个神经元的激活情况来计算样本内每个特征对该样本所属真实类别的贡献度.进而,统计特征集合中每个特征出现时的平均贡献度,作为网络模型对该特征的重要性判断.

$$imp_{w_i} = \frac{1}{N} \sum_{j \in doc(w_i)} imp_{w_{ij}} \tag{8}$$

该方法具有通用性, $imp_{w_{ij}}$ 并不局限于特征贡献度.对于敏感性分析而言, $imp_{w_{ij}}$ 可以是特征 i 在样本 j 中的敏感性评价值.应用该方法对一个训练好的神经网络模型进行分析,可以得到该模型对特征集合中每个特征的重要性评价,进而可以理解其对特征的选择.

2.2 样本特征重要性分析方法的有效性对比实验

2.2.1 实验数据及模型

本文针对非海量数据情况下的文本分类问题,使用 SMP Challenge 2016 的公开微博数据集,基于用户的采样微博文本来判断其年龄属性(根据出生年份被划分为 3 个类别“–1979”“1980~1989”“1990+”).该数据集被划分为训练集、验证集和测试集,分别包含 3 200 个,920 个,1 240 个用户样本,每个样本平均包含 63 条微博文本,每条微博平均包含 15 个词,共 3 万余个不同的词语,样本量远少于词语特征的数量.

本文采用 Zhang 等人[32]提出的适用于文本分类任务的卷积神经网络作为实验模型进行分析,其中,嵌入层参数使用在未标注数据集上预训练的 Word2Vec[33]词向量,其结构如图 2 所示.

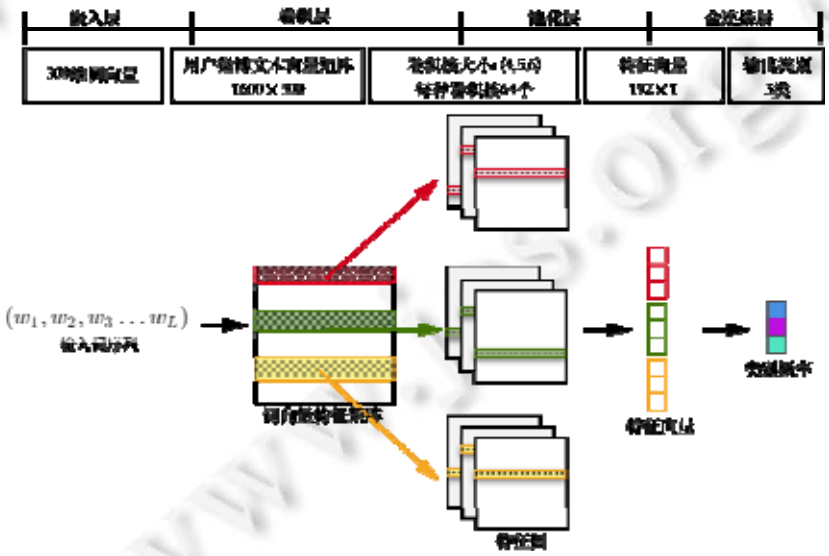

Fig.2 A convolution neural network model for text categorization tasks

图 2 文本分类任务下的卷积神经网络模型

基于实验数据和模型,本文对基于感受野的特征贡献度分析方法(RF-contribution)与传统的基于梯度的特征敏感性分析方法(sensitivity,包括无穷范数 $Sensitivity_\infty$,二范数 $Sensitivity_2$)和随机排序方法(random)进行了对比评估.

2.2.2 有效性实验及结果分析

利用前文提出的样本特征分析方法,可以将神经网络判断的类别相关性反向映射到输入文本上,得到文本中每个词语对神经网络模型分类的重要性,并可视化展示出来.图 3 展示了出生日期位于“1990+”类别中某用户的示例微博特征词的重要性分布.其中,颜色偏蓝为负相关,偏红为正相关.

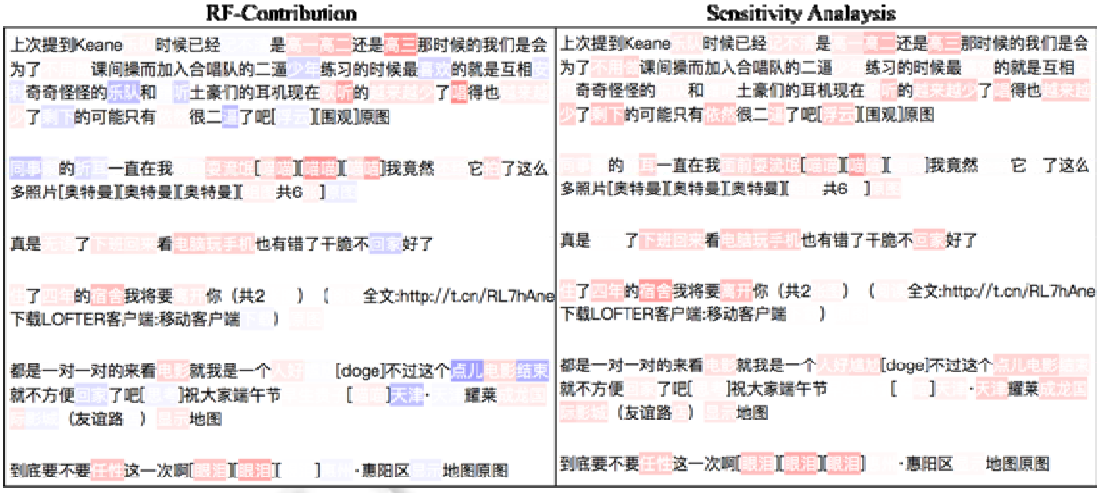

Fig.3 Visual display of feature contribution and feature sensitivity analysis

图 3 特征贡献度和特征敏感性分析可视化展示

从图中可以看出,文本中仅有部分词语对神经网络的决策做出了贡献;而一些常见词语,例如“一直”“都是”等,则对最终的决策没有起到直接的作用.此外,特征贡献度分析方法可以得到特征对分类正、负面的作用,例如“任性”“高三”等词语对神经网络判断该用户为“1990+”类别起到了正面作用;而“同事”“下班”等词语则起到了负面作用.

为了量化评估特征分析方法的有效性,本文使用通用特征分析方法评价指标 AOPC 分别对两种分析方法,在验证集上进行了遮挡实验.实验结果如图 4 所示.

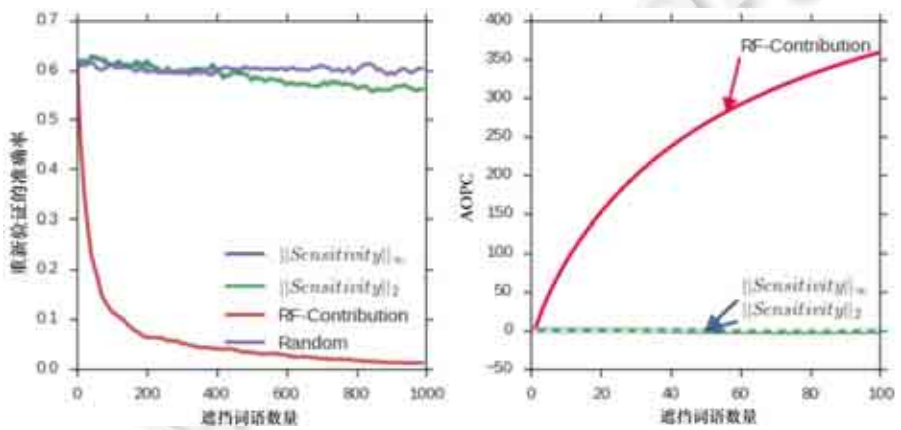

Fig.4 Effective experiments of feature analysis method

图 4 特征分析方法有效性实验

实验结果中,遮挡特征贡献度分析方法、特征敏感性分析方法所认为的关键词均会降低验证集的分类准确率,这表明两种方法可以衡量特征对神经网络模型分类的重要性判断.而贡献度分析方法的 AOPC 值显著地优于敏感性分析,可能的原因是:严格意义下,敏感性分析并不能得到对模型分类重要的特征,敏感性较高的特征词仅代表“当它出现时更易增强或减弱模型对样本类别的判断”,因此遮挡后未必降低模型的分类性能.综上所述,本文所提出的特征贡献度分析方法更能代表卷积神经网络对特征选择的理解.

2.3 神经网络的特征选择结果

前文介绍了如何得到神经网络模型对特征集合的重要性分布,表 1 展示了分析神经网络得到的基于贡献度的特征重要性(RF-Contribution)、基于敏感性的特征重要性(Sensitivity)得到的对于年龄分类评价值 Top10 的特征词列表.

Table 1 Top10 keywords of different feature importance evaluation methods

表 1 不同特征重要性评价方法 Top10 特征词

No.	RF-Contribution	Sensitivity
0	摇一摇(247)	高三(32)
1	–冷笑(45)	原图(3177)
2	酷网(341)	自习(17)
3	新增(1211)	有感而发(42)
4	眼泪(305)	甜言蜜语(15)
5	学院(234)	Cry(247)
6	有感而发(42)	愧疚(16)
7	三好学生(88)	丫头(20)
8	Cry(247)	说(2407)
9	个人主页(112)	用处(15)

从表中可以看到:不同特征重要性评价方法得到的结果不尽相同,其中,基于贡献度的特征重要性分析方法较为关注中频词,而敏感性分析则较关注低频词和部分高频词,但两者对于部分关键词的判断具有一致性.

神经网络通过优化训练样本的误差,不断修改内部参数,隐式地对输入特征进行评价和选择.而传统特征选择方法基于特征变量和类别变量之间的相关性,给予每个特征显式的评价用于选择.那么神经网络对特征的评价是否优于传统的特征选择方法呢?

3 神经网络特征选择能力与传统特征选择方法的对比分析

3.1 特征选择能力的评估

对于分类任务,特征集合的质量一般体现为其分类性能,即基于该特征集合所训练的分类器的表现.为了量化评估神经网络模型和传统特征选择算法所选择的特征集合 S 的分类性能,本文选取了两种不同类型的分类算法作为评估模型:基本的线性分类算法逻辑回归(logistic regression)以及本文所采用的卷积神经网络.对比了以皮尔逊相关系数、卡方检验为代表的传统特征选择方法和卷积神经网络(通过 RF-Contribution 方法计算得到)的特征选择能力.

特征选择的目的是识别有用(高重要性)特征,消除噪声(低重要性)特征,实现对特征空间的缩减.那么特征选择的能力也体现在两个方面:对高重要性特征的识别能力、对噪声特征的识别能力.因此,本文从正向选择和反向遮挡来开展研究,对两者分别进行评估.

3.2 高重要性特征的识别能力的实验性对比研究(正向选择)

正向选择实验,是针对特征评价方法所认为的最相关、最重要的特征子集进行评估.已知特征评价方法对特征集合所有特征的重要性评价值(imp_1,imp_2,imp_3,…),按照重要性从高到低依次对特征进行排序可以得到特征序列(fea_1,fea_2,fea_3,…).依次选取 Top K 的特征集合 S,基于 S 对样本数据重构(遮挡未在 S 中的特征)和学习,并对训练好的模型在测试集进行评估,得到其分类精度随 K 的变化情况.

本文对卷积神经网络、卡方检验以及皮尔逊相关系数通过对训练集的学习得到的特征重要性分布进行评估.图 5 展示了不同特征选择方法选取的 Top K 特征集合的分类精度随 K 值的变化曲线.左图基于逻辑回归模型,右图基于卷积神经网络.

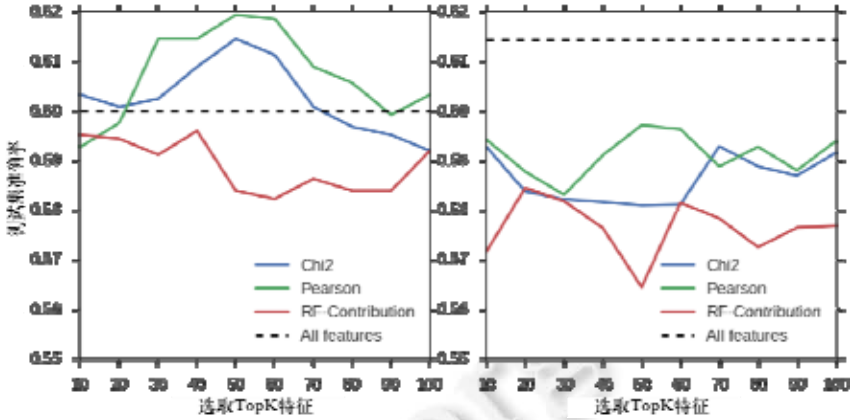

Fig.5 Experimental result of positive selection

图 5 正向选择实验结果

从实验结果中可以看到,卡方检验、皮尔逊相关系数所选取的 Top 特征集合的分类性能在 LR 和 CNN 模型上均优于神经网络所选取的特征集合.因此,这也说明传统特征选择方法对最相关特征的判断优于神经网络,尤其是在非海量样本的情况下.

3.3 噪声特征的识别能力的实验性对比研究(反向遮挡)

反向遮挡实验是针对特征评价方法所认为的最不相关、噪声冗余的特征进行评估分析.不相关、噪声的特征通常对分类器的性能有着负面作用,因此当去除这些特征时,会使得模型的分类性能有所改善.例如,在图像识别任务下对目标主体无关的背景特征进行过滤,或在文本分类任务下对与分类目标无关的特征词进行过滤,都会降低模型的学习负担,可能会提高模型的分类性能.本文先使用全部特征训练 CNN 模型,进而基于不同特征评价方法得到评价值最低的 K 个特征词集合 S',将测试样本内处于 S'中的特征词进行遮挡,然后重新对测试集进行分类,分析其准确率的变化情况.实验结果如图 6 所示.

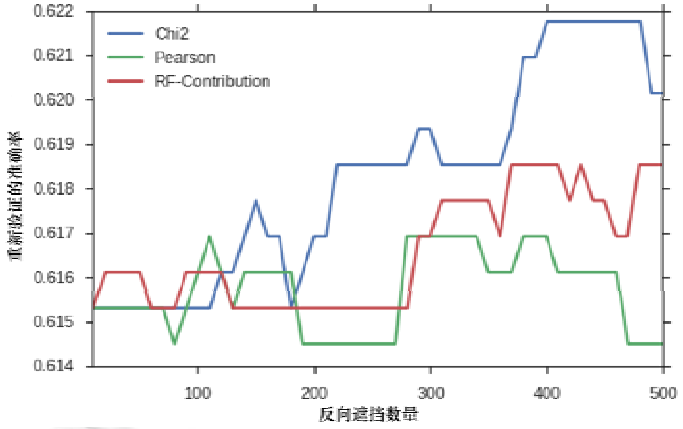

Fig.6 Experimental result of reverse occlusion

图 6 反向遮挡实验结果

从实验结果中可以看到:当遮挡了卡方检验所认为的无用特征时,模型的分类精度有较大的提升,超过了遮挡神经网络所认为的无关特征.这表明传统特征选择方法在识别无关冗余特征方面的能力同样优于神经网络.

综合正、反两个方面评估实验的结果,我们发现:传统的特征选择方法,例如卡方检验、皮尔逊相关系数,优于神经网络模型对特征重要性的判断.其原因可能是当特征数量较多时,神经网络模型需要优化的参数量过多,其对特征的分析和选择效果受限于样本量等因素.

4 卷积神经网络的增强特征选择模型

借助于前文提出的分析方法,我们发现:在非海量样本情况下,传统特征选择方法对特征重要性的理解优于神经网络.那么是否可以将传统特征选择方法与神经网络相结合呢?通常,特征选择被用来对输入进行预处理,与训练过程隔离,但选用的特征数量需人为设定,降低了模型的抗过拟合和泛化能力.那么是否可以将传统特征选择方法对特征的评价结合入神经网络的训练过程中,来辅助神经网络进行特征选择,使其更快、更好地关注富有信息量的特征,避免被无关冗余特征所影响呢?为此,本文提出了增强特征选择的神经网络模型.

4.1 特征选择层

本文在传统的神经网络模型中引入额外一层特征选择层,其中,权重 $W_{m\times 1}$ 与相应输入特征 $X_{n\times m}$ 相对应进行元素相乘,起到对输入特征的放缩效果.使用 ReLU 激活函数,对特征集合进行截断,偏置 b 为特征选择的阈值.神经网络在优化 w 和 b 的同时,起到了对输入特征自适应选择的效果.W 被初始化为传统特征选择的评价值.工作过程如图 7 所示.

$$x'=ReLU(x\odot w+b) \tag{9}$$

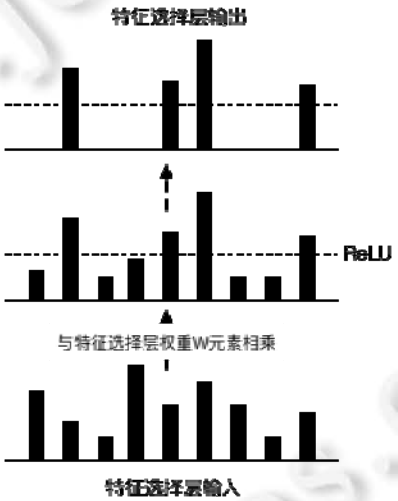

Fig.7 Sketch map of feature selection layer

图 7 特征选择层示意图

特征选择层基于传统特征选择方法对特征项的评价值,对输入特征进行放缩,从而增强或者减弱某些特征对网络训练的影响,这使得网络在学习过程中更关注于有用信息.该方法等效于给神经网络添加关于特征的先验知识.

对于使用了嵌入层的神经网络而言,输入序列需要经过嵌入层映射为词向量矩阵,其输入序列不同位置代表的特征词不固定,无法通过直接对输入向量按维度进行放缩.此时,本文基于特征评价值,对嵌入层的词向量参数矩阵进行放缩.相应的处理结构如图 8 所示.

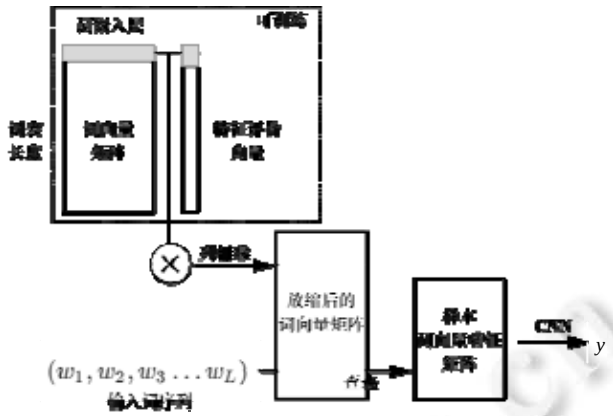

Fig.8 Feature selection enhanced model applied to the convolutional neural network with embedded layer

图 8 增强特征选择模型应用于包含嵌入层的卷积神经网络

该方法不局限于使用嵌入层的卷积神经网络,对于输入为等长特征向量的神经网络(例如输入为数值、图像等),可以直接在输入向量后附加一层特征选择层,如图 9 所示.

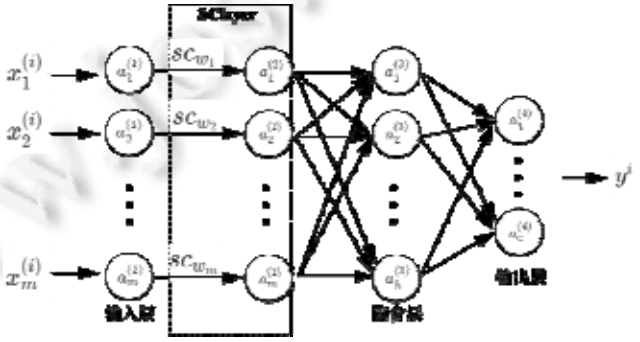

Fig.9 Feature selection enhanced model applied to the neural networks with fixed length features

图 9 增强特征选择模型应用于定长特征的神经网络

4.2 模型有效性验证

本文选取皮尔逊相关系数以及卡方检验两种特征评价方法,将其对词特征的评价值附加入神经网络之中,为了避免评价值的长尾分布导致过多的特征被赋予较小的权重,本文对评价值进行了两种不同的归一化方法:离差归一化(MaxMinScale)、Sigmoid 归一化(SigmoidScale).

本文使用包含嵌入层的卷积神经网络进行文本分类任务,特征评价值被用于对嵌入层相应词向量进行放缩.本文对不同的特征评价方法和归一化方法的组合进行了 30 次重复实验.实验中,模型的迭代次数作为待调整的超参数,选取在验证集上的最优值,计算此迭代次数下模型在测试集上的准确率.实验结果见表 2.

Table 2 Experimental results of feature selection enhanced convolution neural network

表 2 增强特征选择的卷积神经网络模型实验结果

Scale	CNN	CNN-Pearson	CNN-Chi2
No scale	0.614 355	0.604 677	0.592 446
MaxMinScale	–	0.604 489	0.608 198
SigmoidScale	–	**0.618 037***	**0.626 102****

注:上标*代表 *p-value*<0.05,**代表 *p-value*<0.01

从结果中可以看到:在神经网络中附加传统特征选择方法,可以优化神经网络在文本分类任务上的效果,其中,附加经过 Sigmoid 归一化的 Pearson 和 Chi2 评价值的模型效果显著优于无附加神经网络的效果.这表明了本文提出的增强特征选择的神经网络模型的有效性.

5 结论与展望

本文从理解神经网络特征选择的角度出发,提出了基于感受野的特征贡献度分析方法来显式地挖掘神经网络模型对不同特征的重要性判断和选择.进而通过比较传统特征选择方法和神经网络对特征的选择能力,我们发现:以卡方检验、皮尔逊相关系数为代表的传统特征选择方法在提取高重要性特征、过滤噪声特征方面具有优势,尤其当样本量并非海量时.因此,可以在神经网络中结合传统特征选择方法来提高神经网络的特征选择能力和分类效果.据此,本文设计了增强特征选择的神经网络.与通常将特征选择放于模型训练之前的做法不同,该模型将特征选择直接加入网络训练过程中,使得网络可以更有效地进行特征选择.在社交媒体用户建模数据集上的实验结果验证了模型的有效性.本文提出的分析方法和改进模型为理解神经网络特征选择能力、改进神经网络在高维特征问题上的效果提供了一种思路.

神经网络的黑箱特性使我们对其工作原理知之甚少,本文通过对网络权重的分析,尝试打开神经网络的黑箱,从特征选择的角度理解和评价神经网络的特点和不足.未来我们会进一步从理论的角度对本文的分析方法和结论进行验证.近年来,许多研究者尝试将传统机器学习的成果加入神经网络中.我们初步尝试了将传统特征评价方法作为先验知识来辅助神经网络的学习.未来我们会进一步从实验和理论的角度探究不同的结合方法来优化神经网络的效果.

References:

[1] Szegedy C, Liu W, Jia YQ, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich, A. Going deeper with convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014. 1−9. [doi: 10.1109/CVPR.2015.7298594]

[2] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770−778. [doi: 10.1109/CVPR.2016.90]

[3] Graves A. Long short-term memory. In: Proc. of the Supervised Sequence Labelling with Recurrent Neural Networks. Berlin, Heidelberg: Springer-Verlag, 2012. 1735−1780. [doi: 10.1007/978-3-642-24797-2_4]

[4] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, 2014. [doi: 10.3115/v1/P14-1062]

[5] Santos CND, Gattit M. Deep convolutional neural networks for sentiment analysis of short texts. In: Proc. of the Int'l Conf. on Computational Linguistics. 2014. 69−78.

[6] Mikolov T, Karafiát M, Burget L, Cernocký J, Khudanpur S. Recurrent neural network based language model. In: Proc. of the 11th Annual Conf. of the Int'l Speech Communication Association (INTERSPEECH 2010). 2010.

[7] Liu SJ, Yang N, Li M, Zhou M. A recursive recurrent neural network for statistical machine translation. In: Proc. of the Meeting of the Association for Computational Linguistics. 2014. 1491−1500. [doi: 10.3115/v1/P14-1140]

[8] Cho K, Merrienboer BV, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[9] Sutskever I, Martens J, Hinton GE. Generating text with recurrent neural networks. In: Proc. of the Int'l Conf. on Machine Learning (ICML 2011). 2011. 1017−1024.

[10] Shahshahani BM, Landgrebe DA. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. IEEE Trans. on Geoscience and Remote Sensing, 1994,32(5):1087−1095. [doi: 10.1109/36.312897]

[11] Tadjudin S, Landgrebe DA. Covariance estimation with limited training samples. IEEE Trans. on Geoscience and Remote Sensing, 1999,37(4):2113−2118. [doi: 10.1109/36.774728]

[12] Lu S, Oki K, Shimizu Y, Omasa K. Comparison between several feature extraction/classification methods for mapping complicated agricultural land use patches using airborne hyperspectral data. Int'l Journal of Remote Sensing, 2007,28(5):963−984. [doi: 10.1080/01431160600771561]

[13] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research, 2003,3:1157−1182.

[14] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2014. 818−833. [doi: 10.1007/978-3-319-10590- 1_53]

[15] Mares MA, Wang S, Guo Y. Combining multiple feature selection methods and deep learning for high-dimensional data. Trans. on Machine Learning and Data Mining, 2016,9:27−45.

[16] Poria S, Cambria E, Gelbukh AF. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proc. of the EMNLP. 2015. 2539−2544. [doi: 10.18653/v1/D15-1303]

[17] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2012. 1097−1105.

[18] Engelbrecht AP, Cloete I, Zurada JM. Determining the significance of input parameters using sensitivity analysis. In: Proc. of the Int'l Workshop on Artificial Neural Networks. Berlin, Heidelberg: Springer-Verlag, 1995. 382−388. [doi: 10.1007/3-540-59497-3_199]

[19] Alain G, Bengio Y. Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644, 2016.

[20] Zeiler MD, Taylor GW, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning. In: Proc. of the Int'l Conf. on Computer Vision. IEEE, 2011. 2018−2025. [doi: 10.1109/ICCV.2011.6126474]

[21] Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR. Evaluating the visualization of what a deep neural network has learned. IEEE Trans. on Neural Networks and Learning Systems, 2016. [doi: 10.1109/TNNLS.2016.2599820]

[22] Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, MÃžller KR. How to explain individual classification decisions. Journal of Machine Learning Research, 2010,11:1803−1831.

[23] Dimopoulos Y, Bourret P, Lek S. Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Processing Letters, 1995,2(6):1−4. [doi: 10.1007/BF02309007]

[24] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv Preprint arXiv:1312.6034, 2013.

[25] Denil M, Demiraj A, de Freitas N. Extraction of salient sentences from labelled documents. arXiv preprint arXiv:1412.6815, 2014.

[26] Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 2017,65:211−222. [doi: 10.1016/j.patcog.2016.11.008]

[27] Li JW, Chen XL, Hovy E, Jurafsky D. Visualizing and understanding neural models in NLP. arXiv preprint arXiv:1506.01066, 2015.

[28] Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR. Evaluating the visualization of what a deep neural network has learned. IEEE Trans. on Neural Networks and Learning Systems, 2016. [doi: 10.1109/TNNLS.2016.2599820]

[29] Seiler MC, Seiler F. Numerical recipes in C: The art of scientific computing. Risk Analysis,1989,9(3):415−416.

[30] Liu H, Setiono R. Chi2: Feature selection and discretization of numeric attributes. In: Proc. of the 7th IEEE Int'l Conf. on Tools with Artificial Intelligence. 1995. 388−391.

[31] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: Proc. of the 14th Int'l Conf. on Machine Learning. Morgan Kaufmann Publishers Inc., 1998. 412−420.

[32] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820, 2015.

[33] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

卢泓宇(1994－),男,河南南阳人,硕士生,CCF 学生会员,主要研究领域为推荐系统,机器学习应用.

刘奕群(1981－),男,博士,副教授,博士生导师,CCF 高级会员,主要研究领域为网络搜索技术,信息检索,用户行为分析.

张敏(1977－),女,博士,副教授,博士生导师,CCF 高级会员,主要研究领域为信息检索与推荐,用户行为分析与建模,机器学习应用,数据挖掘.

马少平(1961－),男,博士,教授,博士生导师,主要研究领域为智能信息处理,信息检索,用户行为分析.