计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210200090

FMNN:融合多神经网络的文本分类模型

邓维斌　朱　坤　李云波　胡　峰
重庆邮电大学计算智能重庆市重点实验室　重庆 400065
(dengwb@cqupt.edu.cn)

摘　要　文本分类是自然语言处理中一项基本且重要的任务。基于深度学习的文本分类方法大多只针对单一的模型结构进行深入研究,这种单一的结构缺乏同时捕获并利用全局语义特征与局部语义特征的能力,且网络的加深会损失更多的语义信息。对此,提出了一种融合多神经网络的文本分类模型 FMNN(A Text Classification Model Fused with Multiple Neural Network),FMNN 在最大限度减小网络深度的同时,融合了 BERT,RNN,CNN 和 Attention 等神经网络模型的特性。用 BERT 作为嵌入层获得文本的矩阵表示,用 BiLSTM 和 Attention 联合提取文本的全局语义特征,用 CNN 提取文本多个粒度下的局部语义特征,将全局语义特征和局部语义特征分别作用于 softmax 分类器,最后采用算术平均的方式对结果进行融合。在 3 个公开数据集和 1 个司法数据集上的实验结果表明,FMNN 模型实现了更高的文本分类准确率,其中在司法数据集上的准确率达到了 90.31%,证明了该模型具有较好的实用价值。

关键词:文本分类;深度学习;全局语义特征;局部语义特征;语义损失;融合
中图法分类号　TP391

FMNN:Text Classification Model Fused with Multiple Neural Networks

DENG Wei-bin,ZHU Kun,LI Yun-bo and HU Feng
Chongqing Key Laboratory of Computational Intelligence,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract　Text classification is a basic and important task in natural language processing. Most of the text classification methods based on deep learning only focus on a single model structure. The single structure lacks the ability to simultaneously capture and utilize both global and local semantic features. Besides,the deepening of the network will lose more semantic information. In order to overcome the above problems,a text classification model FMNN which is a text classification model fused with multiple neural network is proposed in this paper. The model combines the performances of BERT,RNN,CNN and Attention while minimizing the network depth. BERT is used as the embedding layer to obtain the matrix representation of the text. BiLSTM and Attention are used to jointly extract the global semantic features of the text. CNN is used to extract the local semantic features of the text at multiple granularities. The global semantic features and local semantic features are applied to the softmax classifier respectively. The results are finally fused by arithmetic average. The experimental results on three public data sets and one judicial data set show that the proposed FMNN model achieves higher accuracy rate,and the accuracy rate on the judicial data set reaches 90.31%,which proves that the model has good practical value.

Keywords　Text classification,Deep learning,Global semantic features,Local semantic features,Semantic loss,Fusion

1　引言

在信息呈指数级增长的大数据时代,文本是极其丰富的信息来源,如何从海量文本中获取目标信息是一个非常重要又具有挑战性的问题。文本分类作为自然语言处理中最基本和最重要的任务,是目前自然语言处理研究的一个热点问题。文本分类旨在为文本分配预先定义好的标签或标记[1],在情感分析、内容审核、问答匹配、新闻推荐等任务中具有广泛的应用。基于传统机器学习的文本分类方法常常依赖于人工设计的特征,并且文本表示存在着稀疏、高维度的问题。近年来,随着深度学习在自然语言处理领域的飞速发展,涌现出了大量基于深度学习的文本分类模型,它们通过端到端的方式学习特征表示并解决了文本表示稀疏、高维度的问题,在各种分类任务上的精度都超过了基于传统机器学习的方法。

目前,在基于深度学习的方法中,主流的文本分类模型包括:基于卷积神经网络(Convolutional Neural Network,CNN)

到稿日期:2021-02-09　返修日期:2021-05-03
基金项目:国家重点研发计划(2018YFC0832100,2018YFC0832102);国家自然科学重点基金项目(61936001)
This work was supported by the National Key Research and Development Program of China(2018YFC0832100,2018YFC0832102) and Key Program of National Natural Science Foundation of China(61936001).
通信作者:朱坤(1209562838@qq.com)

的模型、基于循环神经网络(Recurrent Neural Network, RNN)的模型、基于注意力机制(Attention Mechanism)的模型和基于预训练的语言模型(Pre-trained Language Model, PLM)。不同模型有着各自独特的结构,拥有各自的优势,同时也会存在一定的缺陷。CNN 用于学习识别跨空间的模式,它通过空间中的窗口能够很好地捕捉文本的局部语义特征,但是忽略了时间上的先后顺序;RNN 用于识别跨时间的模式,它通过递归计算捕捉全局语义特征,但是缺乏捕捉局部语义特征的能力;Attention 机制能将有限的注意力集中在重点信息上,但是它忽视了词之间的先后顺序关系;PLM 在大规模语料的预先训练下能够获得词的通用向量表示,考虑到从零开始训练一个语言模型会耗费大量的资源,相关研究人员已经公开了许多训练好的模型以供更多研究者使用。

虽然基于深度学习的文本分类方法已经取得不错的效果,但这些方法大多只针对单一模型进行深入研究。单一模型往往不能同时捕捉并利用多个方面的语义信息,并且由于模型的编码和嵌入是一个有损压缩的过程,因此网络深度的增加会进一步造成更多的语义损失。模型融合可以充分利用各模型的优势,选择更优的方式对文本进行表示,并提取多方面的语义特征来保留更多的语义信息,从而达到更佳的文本分类效果。本文提出的 FMNN 模型在尽可能减小网络深度的前提下融合了多个模型的优势,用 BERT 作为嵌入层,以获得更好的文本表示,用 BiLSTM 与 Attention 组合的方式提取全局语义特征,用 CNN 提取文本多个粒度下的局部语义特征,并将全局语义特征和局部语义特征分别用于类别预测,最后把二者的结果作进一步融合。

2 相关工作

2013 年以来,基于深度学习的文本分类方法逐渐取代了基于传统机器学习的方法,研究人员提出了大量用于文本分类的深度学习模型。与基于传统机器学习的文本分类方法相比,基于深度学习的模型通过端到端的方式学习特征表示并进行分类,有效解决了依赖于人工设计特征和文本表示的高维度、高稀疏的问题。其中一些主流的模型按照其体系结构可以分为:基于 CNN 的文本分类模型、基于 RNN 的文本分类模型、基于 Attention 机制的文本分类模型和基于 PLM 的文本分类模型。

2.1 基于 CNN 的文本分类模型

CNN 在图像领域取得的巨大成功促使研究人员将它用于文本分类任务。Kalchbrenner 等[2]提出的 DCNN 模型最早将 CNN 用于文本分类,DCNN 使用宽卷积和动态 k-max 池化交替的结构来生成句子的特征图,该特征图能很好地捕捉单词间的短距离关系和长距离关系,该模型在多个任务上都取得了不错的效果。Kim[3]提出了一种结构更为简单的 TextCNN 模型,该模型使用单层的一维卷积和最大池化来获得句子的特征表示,相比 DCNN 的宽卷积模式,一维卷积保证了每个单词的完整语义性。Johnson 等[4]提出了一种类似于金字塔结构的 DPCNN 模型,它通过增加网络深度来提升模型的性能。Conneau 等[5]提出了一种超深的 VDCNN 模型,它只使用很小的卷积和池化在字符级别的向量上操作,并通过实验证明,该模型的分类性能随着深度的增加而提高。虽然合理地增加网络深度会提升模型的性能,但同时也会增加时间复杂度,并且伴随着更多的语义损失。

2.2 基于 RNN 的文本分类模型

由于 RNN 递归的计算方式常用来捕捉长距离序列信息,而文本可以看作由一系列具有前后时间关系的单词序列组成,因此 RNN 常常用于文本分类。但是由于 RNN 的隐藏层变量梯度会出现消失和爆炸的问题,因此一般使用引入门结构的 RNN 变体长短期记忆网络(Long Short-Term Memory, LSTM)和门控循环单元网络(Gated Recurrent Unit, GRU)。Lai 等[6]提出了一种用于文本分类的 TextRCNN 模型,该模型利用双向 RNN 获得单词的上文向量表示和下文向量表示,并将单词的上文向量、词向量和下文向量进行拼接作为最终的词向量表示,然后使用卷积网络中的池化操作来筛选出有用的特征信息,最终将筛选后的特征作用于 softmax 分类器。Liu 等[7]提出的 TextRNN 模型不同于 TextRCNN,它使用更为常用的双向 LSTM 代替普通的 RNN 并舍弃了池化层,直接使用双向 LSTM 最后时刻的隐层状态向量作为序列表示,然后将该向量馈送至 softmax 层进行分类。基于 RNN 的模型能很好地捕捉序列化的全局语义信息,但是缺乏捕捉局部语义特征的能力。

2.3 基于 Attention 机制的文本分类模型

Attention 机制受人脑独特的信号处理机制的启发而产生,它将有限的注意力资源集中于重要的信息。在自然语言处理任务中,Attention 机制最早应用于机器翻译,随后研究人员将其广泛应用于其他任务。Yang 等[8]引入了 HAN 模型,该模型的显著特征在于,使用单词和句子两个层次的注意力机制,突出了不同层次结构的重要性,在 6 个文本分类任务上的结果显著优于以前的方法。Vaswani 等[9]提出的 Transformer 模型常用于文本分类,它完全基于自注意力机制来捕捉单词的权重分布以及单词间的长距离依赖关系,并通过多头的方式显著提升了模型的并行计算能力。Kim 等[10]提出的 DRCN 模型能够从最底层的单词嵌入层到最顶层的循环层保持原始的和共同关注的特征信息。Attention 机制能将有限的注意力集中在重点信息上,但是它忽略了单词之间的先后顺序关系。

2.4 基于 PLM 的文本分类模型

早期的 PLM 旨在将学习好的单词嵌入,通常不会用于下游任务,如 Word2vec[11]和 Glove[12]。自 2018 年以来,兴起了一系列基于 Transformer 的预训练语言模型,它们专注于学习上下文单词嵌入并用于各种下游任务。Radford 等[13]提出了 OpenGPT 模型,它通过无监督的方式从大规模语料中预先学习词的通用表示,并采用有监督的微调将这些表示用于包括文本分类在内的各种自然语言处理任务。但是它是一个单向模型,每个词的预测取决于以往的预测结果。Devlin 等[14]提出的 BERT 模型是自然语言处理领域发展的一个重要转折点,不同于 OpenGPT,BERT 通过其双向结构能学习到词的上下文表示,该模型提出时取得了包括文本分类在内的多项自然语言处理任务的最好成绩,并在电商[15]、医疗[16]和金融[17]等领域取得了不错的应用成果。随后,研究人员

通过进一步改进 BERT 提出了 BERT-wwm[18]，ALBERT[19]和 SpanBERT[20]等模型。训练一个预训练模型需要耗费大量资源，但有相关研究人员公开了许多训练好的模型供相关领域研究者直接使用。

3 FMNN 模型

本文提出的 FMNN 模型通过两个融合机制来分别获得文本的全局语义特征向量表示和局部语义特征向量表示，再通过一个结果融合机制得到最终分类结果。FMNN 模型的结构如图 1 所示，其由输入层、嵌入层、CNN 层、BiLSTM 层、Attention 层和融合输出层组成。

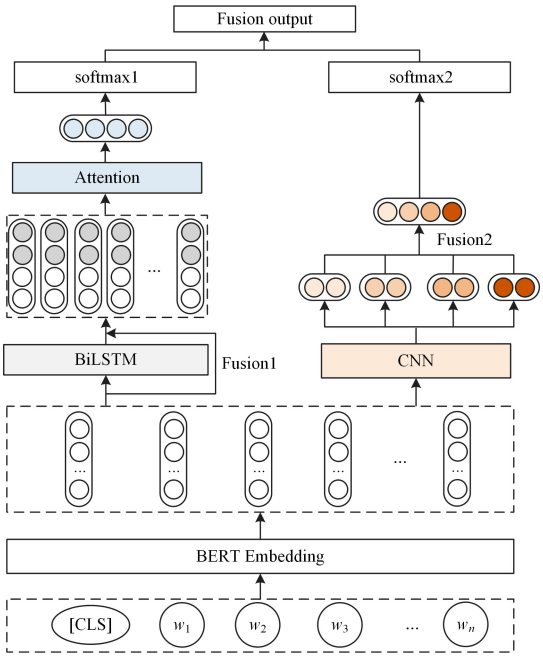

图 1 FMNN 模型

Fig. 1 FMNN model

3.1 嵌入层

预训练模型 BERT 使用上亿级别的未标注语料进行训练，旨在为每个单词生成高质量、通用的向量表示，其被广泛用于各种自然语言处理任务。如图 2 所示，传统的做法是把第一个位置的输出词向量并作为全局特征向量直接用于分类。不同于此，FMNN 模型使用 BERT 获得每个单词的输出向量，并将这些向量作为下游结构的输入。例如，向 BERT 输入一个文本 T，它由特殊符号[CLS]和单词序列 w_1，w_2，w_3，…，w_n 组成，输出得到文本的词向量矩阵 $\boldsymbol{M}\in\boldsymbol{R}^{(n+1)\times d}$，其中 n 代表文本的实际长度，d 代表词向量维度。

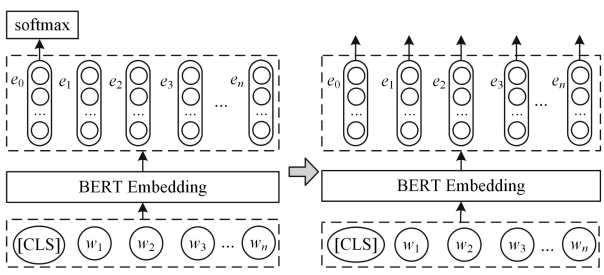

图 2 BERT 词嵌入

Fig. 2 BERT word embedding

3.2 CNN 层

卷积神经网络是一个很好的局部特征提取器，它能通过不同大小的卷积窗口来捕捉不同的 n-gram 特征。Kim[3]很早就证明，通过一个简单的 CNN 便能取得优异的分类效果，我们在此基础上做了一些改进，并用它来提取文本的局部特征。

一段文本除了具有整体的全局语义特征之外，其局部文本还包含了大量的语义特征，并且局部文本下的语义特征又是不同粒度大小的，因此需要考虑不同粒度下的特征提取。局部文本下最重要的是它的短语特征，构成一个短语的字符通常不超过 4 个，我们设计了多种大小的一维卷积核对它进行特征提取，并将不同粒度大小的特征融合在一起作为最终的局部特征表示。

如图 3 所示，从嵌入层 BERT 的输出中获得文本的矩阵表示 $\boldsymbol{M}\in\boldsymbol{R}^{(n+1)\times d}$，并将其作为 CNN 层的输入，然后使用窗口大小分别为 1，2，3 和 4 的一维卷积核对 $\boldsymbol{M}$ 进行卷积，以获得的不同粒度大小的特征图，再对各个特征图使用最大池化操作进行特征筛选，最后采用拼接的方式将从不同粒度筛选出来的特征融合为一个一维向量 $\boldsymbol{F}_{\text{local}}$，并将向量 $\boldsymbol{F}_{\text{local}}$ 作为最终的局部特征表示。若用 h 表示窗口大小，则特征图可以表示为 $\boldsymbol{C}=[c_1,c_2,\cdots,c_{(n+1)-h+1}]$，特征图中的每个元素 c_i 都是通过式(1)计算得到的。

$$c_i=f(\boldsymbol{W}_c*\boldsymbol{W}_{i,i+h-1}+\boldsymbol{b}_c) \tag{1}$$

其中，f 表示激活函数，$\boldsymbol{W}_c\in\boldsymbol{R}^{h\times d}$ 表示卷积核，$*$ 表示元素方式的乘积运算，$\boldsymbol{b}_c$ 表示偏置项，$h\in\{1,2,3,4\}$ 表示窗口大小，$\boldsymbol{M}_{i,i+h-1}$ 表示从 $\boldsymbol{M}$ 的第 i 行到第 $i+h-1$ 行，i 的取值范围是 $[1,n-h+2]$。

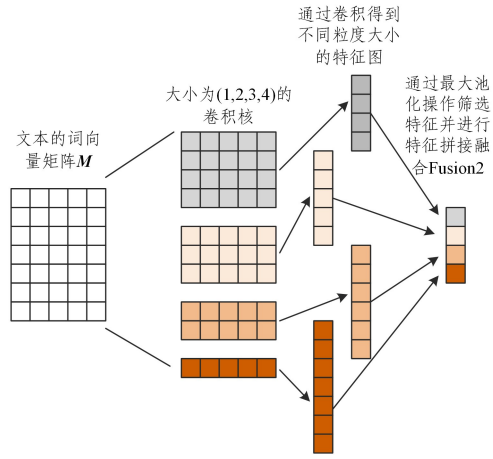

图 3 CNN 局部特征提取

Fig. 3 CNN local feature extraction

3.3 BiLSTM 层

RNN 将文本视为具有时间先后顺序的单词序列，旨在捕捉前后单词的相关性和文本结构，以便进行文本分类。然而，当文本序列过长时，传统的 RNN 容易出现梯度消失和爆炸问题，导致其效果常常不如前馈神经网络。长短期记忆网络(LSTM)是传统 RNN 的一个变体，它是目前主流的循环神经网络结构，旨在捕捉单词的长期依赖性，LSTM 通过引入一个存储单元来存储任意时间间隔内的值，并引入 3 个门(输入门 $\boldsymbol{i}_t$、输出门 $\boldsymbol{o}_t$、遗忘门 $\boldsymbol{f}_t$)来控制进出存储单元的信息流，从而

解决传统 RNN 面临的梯度消失和爆炸问题。

由于单向 LSTM 需要根据前一时刻的信息来预测当前时刻的输出，使当前时刻的输出只能包含该时刻之前的序列信息，而不包含该时刻之后的序列信息。因此，FMNN 模型使用了一个双向 LSTM 来捕捉每个单词的上下文语义信息，BiLSTM 通过前向和后向的 LSTM 来获得第 t 时刻前向隐藏层状态向量 $\boldsymbol{r}_t$ 和后向隐藏层状态向量 $\boldsymbol{l}_t$，并将 $\boldsymbol{r}_t$ 和 $\boldsymbol{l}_t$ 拼接起来作为最终隐藏层状态向量 $\boldsymbol{h}_t$，$\boldsymbol{h}_t$ 可以作为第 t 时刻对应单词的上下文向量表示。以前向 LSTM 为例，前向隐藏层状态向量 $\boldsymbol{r}_t$ 的计算过程如下：

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{r}_{t-1},\boldsymbol{e}_t]+\boldsymbol{b}_i) \tag{2}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{r}_{t-1},\boldsymbol{e}_t]+\boldsymbol{b}_f) \tag{3}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{r}_{t-1},\boldsymbol{e}_t]+\boldsymbol{b}_o) \tag{4}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t*\boldsymbol{c}_{t-1}+\boldsymbol{i}_t*\tanh(\boldsymbol{W}_c\cdot[\boldsymbol{r}_{t-1},\boldsymbol{e}_t]+\boldsymbol{b}_c) \tag{5}$$

$$\boldsymbol{r}_t=\boldsymbol{o}_t*\tanh(\boldsymbol{c}_t) \tag{6}$$

其中，$\boldsymbol{r}_{t-1}$ 表示第 $t-1$ 时刻的前向隐藏层状态向量，$\boldsymbol{e}_t$ 是嵌入层第 t 个位置的输出向量，σ 是 sigmoid 激活函数，$\boldsymbol{W}$ 是权重矩阵，$\boldsymbol{b}$ 是偏置项，tanh 是双曲正切函数，$*$ 是元素方式的乘积运算，$\boldsymbol{c}_t$ 是状态变量，$\boldsymbol{c}_t$ 与输出门共同决定最后输出。

虽然双向 LSTM 能很好地捕捉到每个词的上下文语义信息，但是由于 LSTM 结构中遗忘门的存在，使传播过程中必然伴随着语义的损失。为了尽可能保留更多的语义信息，如图 4 所示，FMNN 模型将 BERT 输出的词向量与 BiLSTM 的输出进行了拼接融合，得到每个单词最终的上下文表示 $\boldsymbol{h}_t^*$。

$$\boldsymbol{h}_t=[\boldsymbol{r}_t\oplus\boldsymbol{l}_t] \tag{7}$$

$$\boldsymbol{h}_t^*=[\boldsymbol{h}_t\oplus\boldsymbol{e}_t] \tag{8}$$

其中，$\oplus$ 表示向量的拼接操作。

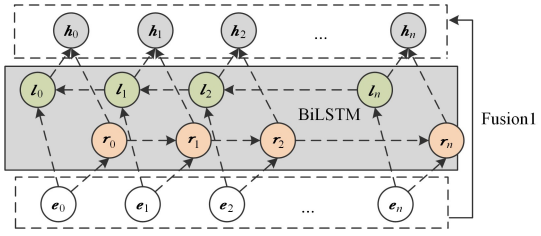

图 4 双向 LSTM 层

Fig. 4 Bidirectional LSTM layer

3.4 Attention 层

从 BiLSTM 层获得的各个时刻的单词表示 $\boldsymbol{h}_t^*$ 包含大量的上下文信息，因此许多研究人员将 BiLSTM 最后一个时刻的隐藏层状态向量 $\boldsymbol{h}_n$ 作为文本的全局语义特征表示，并将其用于分类。考虑到每个单词对文本的整体语义贡献程度是不一样的，我们并没有直接将 $\boldsymbol{h}_n^*$ 用于分类，而是在 BiLSTM 层后使用了注意力机制，通过注意力机制为每个单词的上下文表示向量分配一个权重，以体现不同单词对文本全局语义特征的重要程度，最后通过加权求和的方式得到文本的全局语义特征表示 $\boldsymbol{F}_{\text{global}}$，$\boldsymbol{F}_{\text{global}}$ 通过下式计算得到：

$$\boldsymbol{u}_t=\tanh(\boldsymbol{W}_w\boldsymbol{h}_t^*+\boldsymbol{b}_w) \tag{9}$$

$$\boldsymbol{a}_t=\frac{\exp(\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_w)}{\sum_{t=0}^{n}\exp(\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_w)} \tag{10}$$

$$\boldsymbol{F}_{\text{global}}=\sum_{t=0}^{n}\boldsymbol{a}_t\boldsymbol{h}_t^* \tag{11}$$

其中，$\boldsymbol{W}_w$ 是可训练参数，$\boldsymbol{b}_w$ 为偏置项，$\boldsymbol{u}_t^{\mathrm{T}}$ 是 $\boldsymbol{u}_t$ 的转置，$\boldsymbol{u}_w$ 是一个随机初始化的、可学习的词级上下文向量，$\boldsymbol{a}_t$ 是第 t 时刻单词的归一化权重。

3.5 融合输出层

通过 Attention 层获得文本的全局语义特征表示 $\boldsymbol{F}_{\text{global}}$ 和通过 CNN 层获得文本的局部语义特征表示 $\boldsymbol{F}_{\text{local}}$ 后，我们将 $\boldsymbol{F}_{\text{global}}$ 和 $\boldsymbol{F}_{\text{local}}$ 分别作用于 softmax1 分类器和 softmax2 分类器，从而得到两个类别预测概率 $p1$ 和 $p2$，然后对这两个概率采用算数平均的方式进行融合，得到最终的类别预测概率 p。

$$p1=softmax1(\boldsymbol{W}_g\boldsymbol{F}_{\text{global}}+\boldsymbol{b}_g) \tag{12}$$

$$p2=softmax2(\boldsymbol{W}_l\boldsymbol{F}_{\text{local}}+\boldsymbol{b}_l) \tag{13}$$

$$p=\frac{1}{2}(p1+p2) \tag{14}$$

其中，$\boldsymbol{W}_g$ 和 $\boldsymbol{W}_l$ 为可训练权重，$\boldsymbol{b}_g$ 和 $\boldsymbol{b}_l$ 为偏置项。

最后，将最小化交叉熵损失作为模型训练的目标函数。

$$H(p,q)=-\sum_{i\in N}\sum_{j\in C}q_i^j\log p_i^j \tag{15}$$

其中，N 为训练样本数，C 为类别数，q 为样本的真实标签，使用的是独热编码。

4 仿真实验

4.1 实验数据集

为验证 FMNN 模型的有效性，在 3 个公开数据集和 1 个非公开的司法领域数据集上进行了实验，4 个数据集的信息统计如表 1 所列。其中，*class* 表示类别数量，*length* 表示数据平均长度，*train* 表示训练集样本数量，*dev* 表示验证集样本数量，*test* 表示测试集样本数量。

表 1 数据集信息统计

Table 1 Information statistics of data sets

Data sets	*class*	*length*	*train*	*dev*	*test*
SST-2	2	19	6920	872	1821
AG_News	4	52	120000	—	7600
IMDB	2	294	25000	—	25000
Judicial	2	1080	15308	1563	3827

(1)SST-2[21]：SST-2 是由斯坦福大学公开的一个情感分析数据集，主要针对电影评论来作情感分类，情感类别分为积极和消极两个粗粒度。

(2)AG_News[22]：AG_News 是一个新闻主题分类数据集，它由原始语料中出现最为频繁的 4 种类别构建，类别标签名称分别为 Worlds，Sports，Business 和 Sci/Tech，每个类别包括 30000 个训练样本和 1900 个测试样本，每个样本由所属标签名称、标题和内容描述构成。本文将每个样本中的标题和内容描述进行拼接，并将拼接后的文本作为模型的输入。

(3)IMDB[23]：IMDB 数据集由电影评论文本构成，是一个情感二分类数据集，其包含相同数量的正负情感样本。

(4)司法数据集(Judicial)：该数据由吉林大学相关专业人员提供，每条数据由原告诉称文本和被告辩称文本构成。针对被告行政行为是否合法这一争议焦点，通过司法领域专家人工和撰写规则两种方式进行数据标注，存在争议焦点的标签为 1，反之为 0，其中的一个具体样例如表 2 所列。总数

据集中被告行政行为违法和被告行政行为没有违法的数据比例为1:2,保持该数据比例将数据划分为训练集、验证集和测试集。

表2 司法数据样例

Table 2 Example of Judicial data

原告诉称	被告辩称	被告行政行为是否合法
原告海鑫公司诉称:姜利锋与海鑫公司签订的承包合同于2008年5月19日已经到期,不存在事实上的劳动关系。市人社局将姜利锋按海鑫公司员工对待,根据《工伤保险条例》第十四条第(一)项的规定认定工伤,系适用法律错误,请求依法撤销市人社局作出的决定书。	被告市人社局辩称:姜利锋在工作时间、工作地点、因工作原因受到事故伤害,根据《工伤保险条例》第十四条第(一)项的规定,应当认定为工伤,请求维持市人社局作出的决定书。	0

4.2 实验设置

(1)数据集划分:SST-2和司法数据集已经预先分为训练集、验证集和测试集,而AG_News和IMDB中没有给定验证集,在给它们分配验证集时,保持了不同类别的数据在训练集中所占的比例,从训练集中随机抽取10%作为验证集。例如,IMDB的训练集中,情感态度为积极的样本数量和情感态度为消极的样本数量的比例为1:1,则随机抽取的验证集中情感态度为积极的样本数量和情感态度为消极的样本数量的比例也为1:1,且样本总数占训练集的10%。

(2)权重初始化:模型中的权重通过从[-1,1]的均匀分布中随机采样来初始化。

(3)训练超参数:对于英文数据集,FMNN使用BERT-base[14]作为词嵌入,对于中文数据集,FMNN使用BERT-wwm[24]作为字嵌入,获得的词向量和字向量维度都为768,且嵌入层的参数会被更新。CNN层中使用4种不同大小的一维卷积核,卷积窗口分别为1,2,3和4,每种卷积核的数量为200,BiLSTM层中的隐藏单元个数为300,CNN层和BiLSTM层中均使用dropout机制,dropout参数大小设置为0.2。模型训练的最小批次大小batch_size为64,整个模型通过Adam优化方法进行训练,学习率大小为0.0005。FMNN模型的具体参数设置如表3所列。

表3 FMNN模型的参数设置

Table 3 FMNN model parameter settings

参数描述	参数值
嵌入维度	768
卷积核大小	(1,2,3,4)
每种卷积核的数量	200
隐藏层单元个数	300
dropout随机失活率	0.2
优化器	Adam
学习率	0.0005
训练批次大小	64

4.3 实验结果与分析

4.3.1 对比实验

本文将FMNN模型与以下被广泛使用的文本分类方法进行了对比。

TextCNN[3]:Kim在CNN网络中使用一维卷积核来提取文本的n-gram特征,然后通过最大池化保留最重要的特征,并将该特征作用于softmax分类器。

TextRNN[7]:Liu等提出的一种通用于文本分类的循环神经网络结构,它将单向LSTM最后一个时间步的隐藏层状态向量作为文本的全局语义特征表示,然后将该向量馈送到softmax分类器。

TextRCNN[6]:Lai等引入循环卷积神经网络用于文本分类,与TextRNN不同的是,该网络通过循环神经网络学习到文本的上下文表示后,又引入了一个最大池化层来捕捉文本中的关键信息。

DPCNN[4]:Johnson等提出的一种类似于金字塔结构的分类模型,通过增加网络深度来提升模型的性能。

FastText[25]:Joulin等提出的一种快速的文本分类方法,FastText的性能与基于CNN和RNN模型的性能相当,但是其训练速度要快得多。

Att-BLSTM[26]:Zhou等通过结合双向长短期记忆网络和注意力机制来捕捉文本中重要的语义信息,以便为重要单词赋予更大的权重。

Transformer[9]:Vaswani等在机器翻译任务中提出该模型,在文本分类任务中我们只使用了Transformer的Encoder部分,它能够强有力地提取文本的长距离依赖特征。

BERT-base[14]:Devlin等提出的BERT模型是自然语言处理的一个重要转折点,它在文本分类任务中被广泛使用。它设计了特殊的起始符[CLS],将[CLS]的输出向量作为文本的上下文语义特征,并将其作用于softmax分类器。本文改变了这种传统的分类方式,将BERT所有的词向量输出作为FMNN模型的词嵌入。

BAC_BiLSLM:为了进一步验证本文所提FMNN模型优于传统的直接融合的方法,参照了Liu等[27]提出的AC_BiLSTM模型,它直接融合了CNN,BiLSTM和Attention这3个模型,我们在AC_BiLSTM的基础上增加了BERT词嵌入层,以直接融合BERT,CNN,BiLSTM和Attention这4个模型,并将该融合模型称作BAC_BiLSLM。

4.3.2 整体表现

我们使用准确率(accuracy)作为实验的评价指标,实验结果如表4所列。

表4 不同模型的效果对比

Table 4 Effects comparison of different models

(单位:%)

Model	SST-2	IMDB	AG_News	Judicial
TextCNN	90.35	88.44	91.99	88.48
TextRNN	89.66	85.03	91.28	86.31
TextRCNN	89.56	89.1	91.79	88.35
DPCNN	89.61	88.76	91.43	88.22
FastText	89.06	88.82	91.33	86.62
Att-BLSTM	91.49	89.55	91.58	87.82
Transformer	88.47	86.81	90.97	84.47
BERT-base	93.67	93.06	94.24	88.61
BAC_BiLSTM	94.16	93.29	94.12	88.89
FMNN	**94.86**	**93.57**	**94.55**	**90.31**

FMNN模型在SST-2、IMDB、AG_News和司法数据这4个数据集上分别获得了94.86%,93.57%,94.55%和

90.31%的准确率。在前8个基准实验中，实验结果最好的是BERT-base模型，与BERT-base模型相比，本文提出的FMNN模型在4个实验数据集上分别获得了1.19%，0.51%，0.31%，1.7%的精度提升，从而证明了FMNN模型通过分别提取并利用文本的全局语义特征和多个粒度下的局部语义特征能有效提升模型性能。相比采用传统直接融合方法的BAC_BiLSTM模型，FMNN在4个数据集上分别获得了0.70%，0.28%，0.43%和1.42%的精度提升，进一步说明了所提模型优于传统直接融合的方法。

4.3.3 消融实验

为了验证FMNN模型中的不同组件对模型的有效增益，我们设计了相应的消融实验。

FMNN-CNN为原模型中去掉CNN层的模型，仅将全局语义特征用于分类，输出由原来的融合输出改为直接输出；FMNN-BiLSTM-Att为原模型中去掉BiLSTM和注意力机制的模型，只在嵌入层后使用CNN层来捕捉多个粒度下的局部语义特征并将其用于分类，输出改为直接输出；FMNN-Att为原模型中去掉BiLSTM层后面的注意力机制的模型，将BiLSTM最后一个时间步的隐藏状态向量作为全局语义特征表示；FMNN-BiLSTM为原模型中去掉BiLSTM层的模型，直接对嵌入层获得的词向量矩阵使用注意力机制。消融的实验结果如表5所列。

表5 消融实验结果

Table 5 Ablation experiment results

（单位：%）

Model	SST-2	IMDB	AG_News	Judicial
FMNN-CNN	94.16	93.15	94.25	88.79
FMNN-BiLSTM-Att	94.56	93.18	94.13	89.73
FMNN-Att	94.61	93.48	94.38	89.56
FMNN-BiLSTM	94.62	93.35	94.34	89.79
FMNN	**94.86**	**93.57**	**94.55**	**90.31**

首先将FMNN-CNN的实验结果同基准模型BERT-base进行对比，可以看到，FMNN的准确率均高于BERT-base，说明在BERT模型后面融合BiLSTM和注意力机制能有效提取文本的全局语义特征，提升模型效果；其次将FMNN-BiLSTM-Att的实验结果同基准模型BERT-base进行对比，FMNN-BiLSTM-Att除了在AG_News数据集上的准确率比BERT-base低0.11%以外，在其他数据集上的效果均优于BERT-base，说明在BERT模型后面融合CNN能有效提取文本的局部语义特征，为模型带来增益；然后把FMNN-CNN和FMNN-BiLSTM-Att的实验结果同FMNN进行对比，可以看到，仅将全局语义特征用于分类的FMNN-CNN模型效果和仅将局部语义特征用于分类的FMNN-BiLSTM-Att模型效果均显著低于FMNN，说明FMNN采用的融合输出方式能有效地将全局语义特征和局部语义特征同时作用于分类结果，提升模型效果；最后FMNN-Att和FMNN-BiLSTM在保持FMNN融合结构的前提下，分别只去除了FMNN中的注意力机制和BiLSTM，实验结果表明，FMNN-Att在IMDB和AG_News两个数据集上的效果均优于FMNN-BiLSTM，而FMNN-Att在SST-2和司法数据两个数据集上的效果均低于FMNN-BiLSTM，但FMNN-Att和FMNN-BiLSTM的总体效果基本上高于FMNN-CNN和FMNN-BiLSTM-Att但低于FMNN，从而证明了在FMNN融合模型中通过BiLSTM和注意力机制这种组合方式提取全局语义特征的有效性，它能为模型带来稳定的提升效果。

4.3.4 特征融合讨论

FMNN模型通过CNN层提取到了文本多个粒度下的局部语义特征，通过BiLSTM层的注意力机制的组合方式提取到了文本的全局语义特征，我们尝试将两种特征通过相加和拼接的方式进行融合，然后将融合后的特征向量用于分类，并将这两种方式和FMNN中的结果融合机制进行对比，得到的结果如表6所列。

表6 特征融合实验结果

Table 6 Feature fusion experiment results

（单位：%）

融合方式	SST-2	IMDB	AG_News	Judicial
相加	94.41	92.34	94.22	88.34
拼接	94.46	92.75	94.14	89.02
结果融合	**94.86**	**93.57**	**94.55**	**90.31**

从表6中可以看到，将文本的全局语义特征和多个粒度下的局部语义特征通过相加或拼接的方式进行融合得到的结果低于通过结果融合机制得到的结果，其在IMDB和AG_News数据集上的准确率甚至低于BERT-base模型。常用的信息整合方式将语义特征向量进行简单的相加或者拼接，容易造成语义混乱的问题以及引入更多的噪声，从而影响模型性能。

结束语 本文提出的FMNN模型融合了BERT、CNN、RNN和注意力机制的优势，该模型先通过两个特征融合机制分别提取了全局语义特征和多个粒度下的局部语义特征，然后将这两种特征分别作用于softmax分类器，最终通过结果融合机制得到最终结果。在3个公开数据集和1个司法数据集上进行了实验，结果表明，本文所提FMNN模型在4个数据集上均达到了最优的效果。近年来，研究人员还将图神经网络用于捕捉文本的结构信息，未来工作将结合现有模型和图神经网络进行研究。

参 考 文 献

[1] TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory network[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. ACL, 2015: 1556-1566.

[2] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[C]// Proceedings of 52th Annual Meeting of the Association for Computational Linguistics. ACL, 2014: 1-11.

[3] KIM Y. Convolutional neural networks for sentence classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1746-1751.

[4] JOHNSON R,ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:526-570.

[5] CONNEAU A,SCHWENK H,BARRAULT L,et al. Very Deep Convolutional Networks for Text Classification[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017:1107-1116.

[6] LAI S,XU L,LIU K,et al. Recurrent convolutional neural networks for text classification[C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015:2267-2273.

[7] LIU P,QIU X,HUANG X. Recurrent neural network for text classification with multi-task learning[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016:2873-2879.

[8] YANG Z,YANG D,DYER C,et al. Hierarchical attention networks for document classification[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. 2016:1480-1489.

[9] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017:5998-6008.

[10] KIM S,KANG I,KWAK N. Semantic sentence matching with densely-connected recurrent and co-attentive information[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:6586-6593.

[11] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. arXiv:1301.3781,2013.

[12] PENNINGTON J,SOCHER R,MANNINGC D. Glove:Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014:1532-1543.

[13] RADFORD A,WU J,CHILD R,et al. Language models are unsupervised multitask learners[J]. OpenAI Blog,2019,1(8):9.

[14] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. 2019:4171-4186.

[15] LI K Y,CHEN Y,NIU S Z. Social E-commerce Text Classification Algorithm Based on BERT[J]. Computer Science,2021,48(2):87-92.

[16] RASMY L,XIANG Y,XIE Z Q,et al. Med-BERT:pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction[J]. NPJ Digital Medicine,2021,4(1):1-13.

[17] WENG X F,ZHAO J H,JIANG C X,et al. Research on sentiment classification of futures predictive texts based on BERT[J/OL]. Computing,2021. https://doi.org/10.1007/s00607-021-00989-9.

[18] CUI Y,CHE W,LIU T,et al. Pre-training with whole word masking for chinese bert[J]. arXiv:1906.08101,2019.

[19] LAN Z,CHEN M,GOODMAN S,et al. Albert:A lite bert for self-supervised learning of language representations[C]// Proceedings of the 8th International Conference on Learning Representations. ICLR,2020:1-17.

[20] JOSHI M,CHEN D,LIU Y,et al. Spanbert:Improving pre-training by representing and predicting spans[J]. Transactions of the Association for Computational Linguistics,2020,8:64-77.

[21] SOCHER R,PERELYGIN A,WU J,et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013:1631-1642.

[22] ZHANG X,ZHAO J,LE C Y. Character-level convolutional networks for text classification[J]. Advances in Neural Information Processing Systems,2015,28:649-657.

[23] DIAO Q,QIU M,WU C Y,et al. Jointly modeling aspects,ratings and sentiments for movie recommendation (JMARS)[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014:193-202.

[24] CUI Y,CHE W,LIU T,et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing[J]. arXiv:2004.13922,2020.

[25] JOULIN A,GRAVE É,BOJANOWSKI P,et al. Bag of Tricks for Efficient Text Classification[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,Short Papers. 2017:427-431.

[26] ZHOU P,SHI W,TIAN J,et al. Attention-based bidirectional long short-term memory networks for relation classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Short Papers). 2016:207-212.

[27] LIU G,GUO J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification[J]. Neurocomputing,2019,337:325-338.

DENG Wei-bin, born in 1978, Ph.D, professor. His main research interests include intelligent information processing, natural language processing and uncertainty decision-making.

ZHU Kun, born in 1997, postgraduate. His main research interests include natural language processing and intelligent information processing.

(责任编辑:李亚辉)