ISSN: 2661-3719(Print); 2661-3727(Online)

基于大语言模型 BERT 的文本分类

谢　宁

北京阿里巴巴云计算技术有限公司　北京市海淀区　100102

摘　要：随着科学技术的飞速发展，互联网、手机和计算机等已成为我们日常生活中不可缺少的工具，提升了文本信息的传播速度和传播效率。通过网络能够浏览各行各业的新闻资讯，大量信息的出现要求用户具有一定的筛选能力。因此，学术界和企业公司已将分类任务作为一个重点的研究方向。BERT 预训练语言模型利用 Transformer 的编码器原理进行构建，实现了上下文双向特征提取。本文基于 BERT 研究 BERT_RNN、BERT_CNN 模型在新闻分类数据集 Reuters-21578 和 THUCNews 上探究其性能表现。本文实验结果表明 BERT 仍然取得了最高的性能表现，在 Reuters-21578 和 THUCNews 数据集上分别达到了 0.93 和 0.95 的精度。BERT 模型首先进行自监督预训练，再进行监督学习（微调）。预训练 + 微调方式已成为一种流行训练方式。无标签数据占已有数据集的大部分，如何更好地利用无标签数据是各个领域亟待深入发掘和探索的问题。

关键词：预训练语言模型；BERT 模型；文本分类；深度学习

绪论

近年来 ,CPU、显卡等硬件设备不断升级迭代，计算机的运算速度和存储空间等都有极大的改善和提升。深度学习成为多个领域的热门交叉研究方向。深度学习神经网络已被广泛应用于文本分类任务，深层的网络结构有利于提取深层特征信息。深度神经网络中的卷积神经网络（CNN）、循环神经网络（RNN）、以及 RNN 的变体长短期记忆网络（LSTM）及等网络模型在文本任务上的应用日趋成熟。本文以 BERT 预训练语言模型为研究对象，基于 BERT 融合其他深度神经网络探究其在文本文类任务中的性能表现。以常见的 TextRNN、TextCNN、FastText 等模型作为基准，在 Reuters-21578 和 THUCNews 数据集上验证各模型的性能表现。

1 模型构建

1.1 模型构建

BERT 等基于编码器的预训练语言模型，在训练和微调时首先对文本进行分词处理，例如当输入为单词 pretraining 时，会根据词表被分成 pre，#train，#ing 标记，带有 # 表示并不是一个完整的单词，是单词的一部分。BERT 中的 wordpiecetokenizer 是根据给定的词表使用最长匹配优先算法执行分词操作。模型在读取两个句子后首先对输入文本进行分词，并在第一个句子的开头添加特殊标记 [CLS]，在第一个句子的末尾添加 [SEP]。如上所述，标记被送入嵌入层，标记嵌入、分段嵌入和位置嵌入对应相加，得到 BERT 的输入。模型的输入值送到多头注意力层，通过残差连接传入到叠加和归一化层，层归一化根据式（3.1）：

$$LN(x_i) = \alpha * \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta \quad (3.1)$$

xi 表示输入的第 i 个实例数据，μ 和 σ^2 分别表示每个实例数据特征值的均值和方差，表示一个小数。

随后特征值输送到前馈网络层进行特征值拼接继续前向传播，再通过残差连接输入到叠加和归一化层，至此完成一个编码器基本的特征提取流程，经过 N 个编码器模块逐一进行特征提取，得到语法、结构、语义的特征。

在将输入数据输入到 TextCNN 前，同样对文本先进行分词操作，再转化为标记对应的嵌入值，一段输入文本的词构成词嵌入矩阵。在本文中使用的是在搜狗新闻数据集上训练得到的词嵌入，并用 npz 格式进行存储。在卷积层，TextCNN 的卷积核不同于计算机视觉领域的卷积核沿图像矩阵的高和宽进行移动，而是只沿词向量矩阵的高进行移动。本文中，TextCNN 的卷积核的高度设置有 2，3，4 三种，宽度和词向量矩阵的宽度相同，保证对同一个词向量连续性，卷积核的数量是 256。卷积操作结束后，将卷积层中得到的特征值输入到最大池化层进行池化操作，减小特征空间。然后，进行 dropout 操作减少模型过拟合。最后，将特征值传

输到使用非线性函数的全连接层，实现分类。

TextRNN 使用的词嵌入来源与 TextCNN 相同。将单词转化为词向量后输入到双向 LSTM 层，LSTM 设置为两层，每层具有 128 个隐藏层。LSTM 后面连接两个全连接层，最终仅获取最后时刻的隐藏状态。

1.2 数据集

THUCNews 和 Reuters-21578 是两个在文本分类领域广泛使用的数据集，常被用于测试模型性能表现。Reuters-21578 数据集包含 11228 条新闻文档，涵盖金融、经济、政治等多个领域的新闻报道，具有 90 个类别。THUCNews 数据集是清华大学自然语言处理实验室整理，重新整合划分为“财经”、“房产”、“娱乐”“游戏”，“体育”等 14 个类别。本文对使用的两个数据集的样本分布情况进行了分析。可以看出 Reuters-21578 数据集中样本分布很不均衡，类别 earn 具有 3964 个样本，而 castor-oil、copra-cake、cotton-oil、dfl 等类别仅具有两个或三个样本。因为样本过少可能导致模型训练不充分对某些类别分类不准确降低模型精度。因此在用于训练模型之前将所含样本过少的种类剔除。为保证数据均衡每个类别有足够的数据量，从 THUCNews 数据集中选取 10 个类别，每个类别抽取 2 万个样本构成在本文中使用的数据集，并按照 0.9，0.05，0.05 的比例划分训练集、测试集和验证集。

2 实验设计和结果分析

为探索文本分类新方法进一步提升分类准确率，本文提出 BERT 与神经网络相融合的分类模型，并进行性能评估。在 Reuters-21578 和 THUCNews 数据集上进行模型测试，选择了几种常用的性能指标，与基线模型的性能表现进行对比评估。

2.1 结果评估

为比较不同分类模型的性能，本文使用不同的评价指标进行结果评估。分类任务中常用的评价指标有准确率、召回率、F1 分数、宏平均、加权平均计算公式如下。

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{m} TP_i \quad (4.1)$$

$$\text{Recall} = \frac{1}{N}\frac{TP_i}{TP_i+FN_i} \quad (4.2)$$

$$F_1 = 2\cdot\frac{1}{N}\frac{\text{precision}_i\cdot\text{recall}_i}{\text{precison}_i+\text{recall}_i} \quad (4.3)$$

$$P_{macro} = \frac{1}{k}\sum_{i=1}^{k} P_i \quad (4.4)$$

$$\text{weighted avg} = \frac{\text{sum}(\text{score}_i*\text{weight}_i)}{\text{sum}(\text{weight}_i)} \quad (4.5)$$

N 代表实例的总数，m 代表实例的总类别数，TPi、FNi 分别代表第 i 个类别中被正确分类的实例个数和错误分类的正实例的个数。

本文的全部实验在驱动云平台（https://platform.virtaicloud.com/）使用云服务器完成。本文采用 Python 语言作为开发语言，使用 VS code 进行代码编写，软件利用深度学习框架 Pytorch 完成模型的搭建，配合其他第三方库如 Numpy、Sklearn 等。服务器操作系统为 Ubuntu，搭载 24GB 显存的 GPU 和 8 核 16GB 内存的 CPU，在 windows 10 PC 端完成程序开发，运行内存为 16GB。

2.2 基线模型

本文选取了几种常见且性能较好的深度学习模型作为基线模型，即 BERT、TextRNN、TextCNN、FastText、Transformer 和 DPCNN。

BERT：本文中使用的 BERT 是谷歌 AI 团队在 2018 年发布的 BASE 版本，使用官方发布的预训练参数对模型初始化。

TextRNN：TextRNN 是利用双向 LSTM 实现对文本从左到右和从右到左中两个方向提取特征值，最后将特征进行拼接。在全连接层维度转换，在输出层利用 softmax 激活函数计算分类概率。

TextCNN：2014 年，Kim 等人首次提出了 TextCNN 文本分类模型，将 CNN 神经网络用于文本处理。CNN 的核心操作是通过卷积层捕捉局部特征，卷积核的权重共享能够减小参数量，池化层可以减小数据规模，加快模型训练速度。

FastText：FastText 由 Facebook 人工智能研究团队开发。FastText 是一种基于神经网络的文本分类，该模性体积小但运算速度快，使用词向量和 n-gram 信息捕捉输入数据中的特征信息。与其他神经网络相比 FastText 在分类精度等指标与 TextCNN、TextRNN 等神经网络相近的情况下将训练和推理速度降低了几个数量级。

Transformer：Transformer 由编码器和解码器型组成，通过自注意力机制计算句子中每个单词与其他单词相对的关联程度，根据关联性重新计算特征值。

DPCNN：2017 年腾讯 AI 实验室发布 DPCNN（Deep Pyramid Convolutional Neural Networks），在 TextCNN 的基础上进行网络改造，该模型解决了难以捕捉长距离特征的问题。

ISSN: 2661-3719(Print); 2661-3727(Online)

2.3 BERT 结合神经网络

BERT_CNN 由 BERT 与 CNN 融合而成。BERT 模型对文本输入提取特征，对 BERT 的特征值调整维度作为后续卷积操作的准备。BERT 的输出层与 CNN 相连，用于进一步提取特征。

BERT_RNN 是 BERT 预训练语言模型和 RNN 结合而成。BERT 对输入文本进行特征提取，将结果输入到双向 LSTM 层进一步序列建模。

2.4 实验参数设置

下面是部分网络模型的参数设置。训练 BERT 模型时，epoch 设置为 3，batch 设置为 128，学习率设置为 5e-5，隐藏单元个数为 768。BERT 具有 12 层隐藏层，由 12 层 Transformer Encoder 堆叠而成。TextCNN 的 epoch 设置为 20，batch 设置为 128，学习率设置为 1e-3，卷积核大小设置为 2，3，4 三种。FastText 训练时 epoch 设置为 3，batch 设置为 128，学习率设置为 1e-3，隐藏单元个数为 256。为缩短训练时间和节省计算资源，所有模型训练时均设置若超过 1000 batch 效果还没提升，则提前结束训练。

2.5 实验结果与分析

本文在 Reuters-21578 数据集上，对两种模型进行训练测试。BERT 具有最高的准确率 0.83，而 TextCNN 的准确率仅为 0.58。TextCNN 的精度、召回率和 F1-score 均低于 BERT。经过分析认为，BERT 具有更好的鲁棒性，在数据分布不均衡的情况下也能达到较好的预测效果。TextCNN 的准确率低可能是对具有较少数量样本的类别训练不够导致的。数据的分布不均衡影响了模型的性能表现。

从 THUCNews 数据集中获取的数据分布均衡能够充分显示模型的性能，因此在该数据集上对 8 种模型进行了性能测试。BERT、BERT_RNN 和 BERT_CNN 分别取得了精度 0.95、0.94、0.95。BERT 融合其他深度模型并没有显著提升分类任务的精度，反而 BERT_RNN 的精度相较于 BERT 有所下降。各模型的宏平均（先计算每个类别的性能指标，再计算所有类别的算术平均值。）指标同样是 BERT、BERT_RNN 和 BERT_CNN 取得最高值。BERT 在 THUCNews 数据集上训练后对各类别的 Precision、recall 和 F1-score。每个类别的精度均达到 0.92 及以上，体现出 BERT 较强的分类能力。

3 总结与展望

随着数字化和信息化建设不断推进，网络为信息传播提供了巨大便利上。互联网是当下信息的主要传播媒介。如何实现文本分类快速获取相关资讯成为重要研究方向。本研究基于 BERT 和深度神经网络展开文本分类任务研究，主要完成工作有介绍国内外文本分类技术的研究现状、BERT 及深度学习网络的基本原理。通过将 BERT、TextRNN、TextCNN 和 BERT 模型与深度神经网络结合构建的 BERT_RNN 和 BERT_CNN 模型，在 Reuters-21578 和 THUCNews 数据集上进行性能测评，测试 BERT 和其他深度神经网络结合能否提升分类性能。

参考文献：

[1] 胡少云 , 翁清雄 . 基于词向量融合的建筑文本分类方法研究 [J]. 微型电脑应用 ,2024,40(02):18-20+25.

[2] 慎金花 , 陈红艺 , 张更平等 . 基于层次分类器的专利文本分类模型研究 [J]. 情报杂志 ,2023,42(08):157-163+68.

[3] 谢莉萍 . 基于卷积神经网络的中文文本分类研究 [J]. 信息与电脑 (理论版),2023,35(20):94-96.

[4] 金罡 . 基于词嵌入分布式表示特征的卷积循环神经网络长文本自动分类研究 [J]. 电子技术 ,2022,51(06):52-54.

[5] 王道康 , 张吴波 . 基于 MacBERT-BiLSTM 和注意力机制的短文本分类研究 [J]. 现代电子技术 ,2023,46(21):123-128.