doi:10.3772/j.issn.1002-0470.2021.06.011

基于 transformer 的工单智能判责方法研究①

汪加婧②* 范 维**

(*中国电信股份有限公司湖北分公司 武汉 430022)

(**华中农业大学工学院 武汉 430070)

摘 要 在图像、文本、视频、语音以及社交类网络数据爆炸增长的时代,企业如何从海量非结构化数据中提取出有效信息并将之转化为生产效率的提升和流程自动化的实现,是目前迫切需要关注和解决的问题。本文以运营商集团电子工单自动判责场景为切入点,提出使用基于 transformer 架构的双向编码器表示(BERT)作为文本分类模型,自动收集各省份的反馈信息并进行各省份的工单责任智能判定。通过将 BERT 模型与 LightGBM 和 Bi-LSTM-Attention 模型进行实验对比,结果表明 BERT 模型对各类别工单的预测准确率均达到了 96% 以上,具有较好的实际应用效果。

关键词 工单智能判责; 文本分类; transformer; 双向编码器表示(BERT)

0 引 言

在互联网全球化的当今世界,许多半结构化数据比如聊天内容、热点时讯、交易信息等都在以每秒数亿兆的巨大数量迅速产生。面对这些复杂数据,如何利用人工智能技术提升生产力,是每个企业向智能化、自动化转型的重点与难点。随着通信网络的规模化发展,运营商也同样面临着海量的故障告警信息和故障处理反馈信息。可以通过使用自然语言处理技术迅速合理地定位故障主责任省份,以达到降低专业人员的工作量,提升工作效率和准确率的目的。本文使用基于 transformer 架构的 BERT 模型来做文本分类,通过加入一些优化策略,使其具有较好的实际应用效果。

1 相关研究

文本分类[1]最早使用的技术手段是专家经验,这种方法可以解决一些规则明显的文本分类问题,但是不具有普适性且人工耗费巨大。随着机器学习的兴起,一些传统文本分类方法,如朴素贝叶斯、K 近邻方法、支持向量机、集成算法中的轻量梯度提升算法(light gradient boosting machine, LightGBM)等开始在文本分类领域有所应用[2-4]。这些传统的机器学习方法在文本表示时没有考虑上下文关系,缺乏语义信息,且大多需要人工进行特征工程,较为费力。2006 年深度学习概念的提出为人工智能在复杂场景中的应用提供了更多的可能性。引入深度学习技术的神经网络模型可以自动提取文本中的有效特征进行学习,大大减少了人工的干预,表现更为智能。

卷积神经网络[5](convolutional neural network, CNN)架构中卷积层负责提取特征,采样层负责特征选择,全连接层负责分类。卷积层一般通过设置卷积核的高度、宽度和图片通道数来决定如何提取输入层的局部信息。采样层的操作即池化,目的是在尽量不丢失图像特征前提下,对特征图进行降维操

① 中国电信集团 AI 项目(ZDGG-2019-03)资助。

② 女,1990 年生,研究生,工程师;研究方向:大数据,人工智能;联系人,E-mail: 18907197886@189.cn

(收稿日期:2020-07-03)

作,形成最终的特征。一般最常用的是 max pooling,此外还有 sum pooling、crow pooling、ave pooling、mop pooling、rmac pooling 等其他池化操作。使用 CNN 做文本分类时,输入层中 n 个单词分别对应 n 个 k 维词向量,将其按照从上到下顺序排列形成的矩阵可以看作一副高度为 n、宽度为 k 的图像进行处理。

循环神经网络[6](recurrent neural network,RNN)相比普通神经网络在处理类似文本这种序列信息时更具有优势。因为普通神经网络会将一个句子里的每个词完全割裂开来,将其作为单独处理的输入,而 RNN 通过隐藏层的使用让网络拥有记忆力。RNN 通常采用基于时间的反向传播算法(back propagation through time,BPTT),但是由于反向传播算法(back propagation,BP)的特点和长距离依赖容易出现梯度消失或者梯度爆炸的问题,这导致训练时梯度不能在较长序列中一直传递下去,从而使 RNN 无法捕捉到长距离的影响。

针对 RNN 的缺陷,Hochreiter 和 Schmidhuber[7] 在网络结构上做了相应改进,提出了长短期记忆网络(long short-term memory,LSTM)。LSTM 使用了“输入门”、“输出门”和“遗忘门”3 个门结构来将短期记忆与长期记忆结合起来,并控制不同时刻的状态和输出。与 RNN 不同的是,LSTM 的记忆和输入是相加的,因此只要“遗忘门”是开启的,一般就不会出现梯度消失的问题。在文本涉及领域,模型有时需要通过获取上下文来消除歧义,而 LSTM 对文本序列的记忆理解是单向的即无法获取从后到前的信息。双向长短期记忆网络 Bi-LSTM[8] 有一种在时间上前向学习过程和向后学习过程,相比 LSTM 可以更好地捕捉双向的语义依赖。Bi-LSTM 是由两个 LSTM 上下叠加组成,每个时刻的输出通过前后两个 LSTM 的计算结果共同决定,因此 Bi-LSTM 能更好地捕获句子中上下文的信息,并能在实际中获得更好的效果。

2 Transformer 模型

2014 年 Google Brain 团队[9] 提出了序列到序列(sequence to sequence,Seq2Seq)概念。Seq2Seq 简单来说就是利用 LSTM 或门控循环单元[10](gated recurrent unit,GRU)将一个输入序列映射为一个输出序列,即 encoder-decoder 结构思想。在 Seq2Seq 模型中,encoder 负责将序列转换成一个固定长度的语义向量,decoder 则将该向量转换成最终序列进行输出。当输入的文本序列长度较长时,这样的做法会造成一些问题:一是语义向量无法包含长序列所有信息,限制了模型的理解能力;二是由于 LSTM 或 GRU 是序列性循环结构,计算无法进行并行化处理,导致模型训练效率低下。解决的方法就是在模型计算时引入 attention 机制,让模型学会关注重要信息。

文本中,attention 可以理解为输入序列中每个单词和输出序列中某个单词的对应模型,输入序列中每个单词可当成一个(key-value)元素,输出序列中某个单词看为一个(query)元素。Attention 机制的具体计算过程可抽象为两个过程:(1)对 query 和每个 key 分别计算两者的相似性或者相关性来获得权重系数。(2)根据权重系数对 value 进行加权求和。权重系数的大小代表了输入序列中每个单词所包含信息的重要程度,这样就能从长文本序列中有选择地筛选出重要信息,忽略其他不重要的信息。2014 年 Bahdanau 等人[11] 提出了 attention 模型并将其运用于机器翻译。2015 年 Xu 等人[12] 关于图像描述生成场景部分提出了两种 attention 模式,即 hard attention 和 soft attention。同年 Luong 等人[13] 以机器翻译为场景提出了改进版的 global attention 和 local attention。2017 年 Google 翻译团队[14] 提出了解决 Seq2Seq 问题的 transformer 模型。Transformer 是第一个彻底抛弃 RNN、完全依赖 self-attention 机制的 Seq2Seq 模型。

2018 年 Google 提出一个基于双向 transformer 的 encoder 架构的语言模型,即 transformer 的双向编码器表示(bidirectional encoder representations from transformers,BERT)。在完成 BERT 模型的预训练后,对于不同自然语言处理(natural language processing,NLP)的下游任务,只需要用一小部分带标签的数据集,就可以使用 fine-tuning 方法完成相应任务。fine-tuning 方法通过结合下游任务的数据集

进行微调从而调整预训练模型参数，使模型能够更好捕捉到下游任务的数据特征。到目前为止，BERT 是在句子和词级别任务上达到最佳表现的 transformer 模型。因此本文可以通过使用 BERT 来解决文本分类的问题，以实现工单的智能判责。

3　基于 transformer 工单智能判责系统

为了提高生产环境的稳定性与自动化运维能力，工单智能判责系统部署在 Kubernetes[15] 集群上。Kubernetes 属于主从分布式架构，Master 是管理节点，负责集群的控制和调度，Node 是工作节点，执行具体的业务容器。Kubernetes 架构图如图 1 所示。一个 Node 上面可以运行多个 Pod，同时每个 Pod 可以包含多个容器。

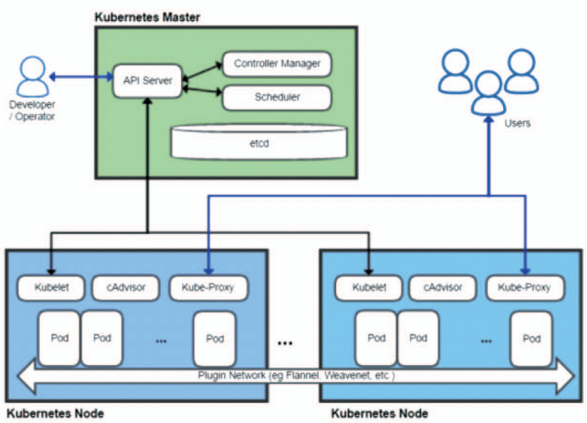

图 1　Kubernetes 架构图

在部署过程中，使用 Dockerfile 制作镜像文件并指定端口信息和启动命令，创建持久化卷声明（peritet volume claim，PVC）用于请求特定的资源存储空间，同时利用 Deployment 控制器自动监控 Pod 资源实现容器的自动扩缩和自动迁移。整个系统功能封装在一个 Pod 内，并通过使用 kube proxy 的代理服务来实现高可用的应用负载均衡。整个工单自动判责系统设计如图 2 所示。

首先电子工单系统自动采集告警处理回单信息和告警处理反馈信息，将数据加密后推送到数据接收端口。数据接收端口将数据解密后按订单号和省份进行分类，形成单独的文本信息。预处理模块通过构建的专家知识经验库对文本信息数据进行深度清洗，提升文本内容的质量。预处理过程分为两方面：一是增加关联信息，比如天气、时间、地理位置；二是删除冗余信息，比如无效自动反馈内容、故障专业词汇等。将预处理后的结果输入到文本分类器 BERT 中完成文本分类后，再将最终结果通过数据输出端口回传给电子工单进行页面展示。

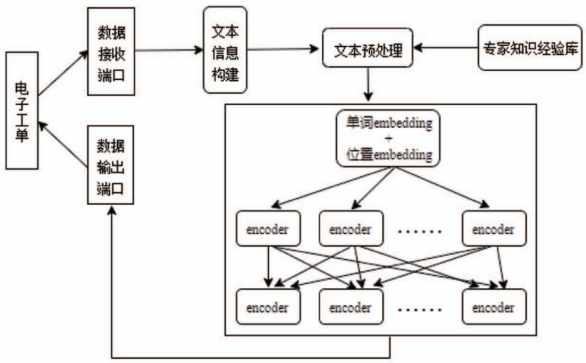

图 2　工单智能判责系统图

在工单智能判责系统中，使用基于双向 transformer 架构的 BERT 作为分类器来实现其核心功能。实现文本分类的基础是文本序列内容的向量化，即对每个字做向量编码。因为 self-attention 是与位置无关的，那么无论句子的顺序是怎样的，通过 self-attention 计算的 hidden embedding 都是一样的，这显然是不符合逻辑的。因此 Encoder 的输入是包含内容和位置两个信息，即文本序列中的每个单词 embedding 和位置 embedding 拼接后得到的带有位置信息的单词表示向量。对于位置编码有 2 种方式，第 1 种是训练出一个位置编码；第 2 种是使用三角函数编码的方法。具体公式如下：

$$PE(pos,\ 2i) = \sin\left(\frac{pos}{10\,000^{\frac{2i}{d}}}\right)$$

$$PE(pos,\ 2i+1) = \cos\left(\frac{pos}{10\,000^{\frac{2i}{d}}}\right)$$

其中，pos 表示单词的位置，i 表示 embedding 的维度。三角函数的性质使得位置编码信息可以既考虑到绝对位置又可以考虑到相对位置。

BERT 是一个以 transformer 架构为基础的深度双向网络模型，该网络能有效地从任意标记的左右或上下文捕获信息，目前有两种变体 BERT Base 和

BERT Large。本文采用 BERT Base,即一共包含 12 层 encoder layer, 以及 1.1 亿个参数。BERT 将工单文本序列中的每个字转换为一维向量作为输入,输出则是对应的融合全文语义信息后的向量表示,再经过 softmax 函数得到最终的分类结果。

4 工单智能判责应用实例

4.1 数据来源

本文实验数据来源于运营商集团电子工单系统中 2019 年传输专业的工单数据,数据集各标签的样本数量如表 1 所示。

表 1 数据集各标签的样本数量

数据集	标签 0	标签 1	标签 2	标签 3
实验集	780	83	3773	1194
测试集	91	46	232	198
总样本	871	129	4005	1392

数据集一共有 6397 个样本,数据集标签分为 0、1、2、3 四类,其中标签 1 的样本数量只有 129,存在样本不均衡的问题。数据集的文本长度最小为 36,最大为 6689,平均长度为 270,属于长文本数据。

4.2 实现环境

本实验采用 Windows 10 操作系统,GPU 内存为 16 GB 的运行环境。编程语言为 Python 3.7,在 Jupyter notebook 平台上配合 Tensorflow 1.13.1 版本深度学习框架进行开发。

4.3 模型评价

为验证工单自动判责的准确性,在使用相同实验集的基础上,将本文所采用基于 transformer 的 BERT 模型与 LightGBM 和 Bi-LSTM-Attention 模型进行实验对比,并使用统一评价指标来对实验结果进行评价。本文分类效果评价指标采用准确率(precision)、召回率(recall)和 F1-score 值作为综合评价指标。各个评价指标的计算公式如下:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = 2 \times \frac{TP}{2 \times TP + FP + FN}$$

其中,*TP* 表示正确的正例,*FP* 表示错误的正例,*FN* 表示错误的负例,*TN* 表示正确的负例。

实验时需要对所有样本统一进行优化处理以提升文本可读性和关键内容归纳能力,并大幅度降低专业词汇出现频率。对于样本不均衡问题采用焦点损失函数以提升小样本分类效果。实验过程采用三折交叉验证方法,将实验集分成 3 份,每次选择 2 份作为训练集,剩余的 1 份作为验证集。LightGBM、Bi-LSTM-Attention、BERT3 个模型训练 3 次后,在不同验证集上的准确率如表 2 所示。

表 2 验证集分类效果

验证集	LightGBM	Bi-LSTM-Attention	BERT
1	0.73	0.86	0.95
2	0.69	0.89	0.97
3	0.72	0.88	0.99

BERT 模型在 3 个验证集上准确率可分别达到 0.95、0.97 和 0.99。由于使用了双向 transformer 架构和 fine-tuning 方法,BERT 分类准确率明显高于仅引入 attention 的 Bi-LSTM 模型。同时可以看出,基于树模型的 LightGBM 分类效果远没有深度神经网络模型的效果好。为了进一步验证 3 个模型的分类效果,使用准确率、召回率和 F1-score 值来具体评价 3 个模型在测试集上的表现,结果如表 3 所示。

表 3 模型测试集效果

分类模型	准确率	召回率	F1-score
LightGBM	0.71	0.73	0.71
Bi-LSTM-Attention	0.88	0.89	0.88
BERT	0.98	0.98	0.98

从表 3 的模型测试集结果可以看出,BERT 的准确率、召回率和 F1-score 值均为 98%,效果较佳。由于数据集存在样本不均衡的问题,所以还需要关注 BERT 在测试集上对每个分类预测的效果如何,其结果如表 4 所示。

从表 4 可以看出,BERT 对每个类别的分类效

果也较好,每个分类的准确率均达到 96% 以上。目前该模型已经成功在生产环境中上线使用,经过 2 个月的人工观察评估,模型预测准确率稳定在 97% 左右。

表 4 测试集分类效果

标签	准确率	召回率	F1-score
0	0.96	0.99	0.97
1	1.00	0.98	0.99
2	0.97	0.99	0.98
3	0.99	0.96	0.97

5 结 论

本文使用基于 transformer 的 BERT 模型来实现运营商工单智能判责功能。通过与 LightGBM 和 Bi-LSTM-Attention 模型进行实验对比,实验结果表明 BERT 模型对各类别工单的预测准确率均达到了 97% 以上,并在实际应用中准确率稳定在 96%,效果较佳。Transformer 架构模型虽然实际应用效果较好,但是依然存在两个问题,一是由于自注意力机制每次都要计算所有词之间的注意力,其计算复杂度为文本长度的平方;二是对文本长度有限制,直接切割可能造成语义上的碎片化,损失部分有效信息。下一步可以尝试研究 Star-Transformer[16] 的星状结构,Star-Transformer 在注意力机制的计算上进行了优化,构建了一个星状的结构,所有序列中直接相邻的元素可以直接相互作用,而非直接相邻的元素则通过中心元素实现间接的信息传递。面对文本长度限制问题可以考虑学习 Transformer-XL[17],引入循环机制和相对位置表示。Transformer-XL 会把每个序列计算后的隐状态带入到下一个序列的计算当中,但是这样跨层的方式也使得其能够学习到的语义长度受限于网络的深度。

参考文献

[1] 王强, 王晓龙, 关毅,等. K-NN 与 SVM 相融合的文本分类技术研究[J]. 高技术通讯, 2005, 15(5): 19-24

[2] 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3): 18-24

[3] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 17(9): 1848-1859

[4] 周青松, 范兴容. 基于 Stacking 融合深度学习模型和传统机器学习模型的短文本情感分类研究[J]. 无线互联科技, 2018, 15(24): 63-65

[5] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks[J]. *Communications of the ACM*, 2017, DOI: 10.1145/3065386

[6] Mikolov T, Martin K, Burget L, et al. Recurrent neural network based language model[C]//Interspeech, Conference of the International Speech Communication Association, Chiba, Japan, 2015: 1045-1048

[7] Hochreiter S, Schmidhuber J. Long short-term memory [J]. *Neural Computation*, 1997, 9(8): 1735-1780

[8] 朱嘉莹, 王荣波, 黄孝喜, 等. 基于 Bi-LSTM 的多层面隐喻识别方法[J]. 大连理工大学学报, 2020, 60(2):102-108

[9] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//28th Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 1-9

[10] 陈虎, 高波涌, 陈莲娜, 等. 结合注意力机制与双向切片 GRU 的情感分类模型[J]. 小型微型计算机系统, 2020, 41(9):1793-1799

[11] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translation[J]. *arXiv*:1409.0473, 2014

[12] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention[J]. *arXiv*:1502.03044, 2015

[13] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation[J]. *Computer Science*, 2015, DOI: 10.18653/v1/D15-1166

[14] Guo M, Zhang Y, Liu T, et al. Gaussian transformer: a lightweight approach for natural language inference[C]// National Conference on Artificial Intelligence, Hawaii, USA, 2019:6489-6496

[15] Bernstein D. Containers and cloud: from LXC to Docker to Kubernetes[J]. *Cloud Computing*, *IEEE*, 2014, 1(3):81-84

[16] Guo Q, Qiu X, Liu P, et al. Star-transformer[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies, Minneapolis, USA, 2019, DOI: 10.18653/v1/N19-1133

[17] Dai Z H, Yang Z L, Yang Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context [C]// 57th Annual Meeting of the Association-for-Computational-Linguistics, Florence, Italy, 2019: 2978-2988

The realization of intelligent judgments of the work order responsibilities based on transformer

Wang Jiajing*, Fan Wei**

(*Hubei Branch of China Telecom, Wuhan 430022)

(**College of Engineering, Huazhong Agricultral University, Wuhan 430070)

Abstract

In the era of explosive growth of images, texts, videos, voice and social network data, how enterprises extract and transform effective information from massive unstructured data to improve the production efficiency and realize the process automation has been an urgent concern that needs to be solved. Taking the electronic work order automatic responsibility judgment scenario of the Telecom Operator Group as the starting point, the bidectional encoder representations from transformer (BERT) based on the transformer architecture is used as the text categorization model to automatically collect feedback information from provinces and make intelligent judgments of the work order responsibilities of each province. The comparison and analysis of the BERT model, the LightGBM model and the Bi-LSTM-Attention model indicate that the prediction accuracy of the BERT model on all types of work orders is over 96%, showing excellent practical effects.

Key words: intelligent judgments of the work order responsibility, text categorization, transformer, bidirectional encoder representations from transformer(BERT)