传递方案 Capsule[20].这种方案在分类任务上具有很好的性能表现与解释性.

2.4.1 注意力机制模型

注意力机制受启发于人类观察事物过程中的视觉注意力机制,人类观察图像总是注意在局部上而不是看到图像上的每个位置.与此类似,在阅读长文本时,人们通常不会关注全文,而是结合自身认识捕捉文本中重要的局部信息,以便快速分析内容.本文采用标准的注意力机制,从句子编码层的输出中提取对句子作用较大的特征,通过对特征自动加权的方法,可以有效地从长文本中捕捉到的重要的特征信息.对 $X\in\mathbb{R}^{L\times K}$,其中,$L$ 为最大语义词数,K 为特征维度.计算公式如下:

$$\left.\begin{aligned}\alpha &= softmax(\tanh(XW+b)V)\\ Q_1 &= \alpha^T X\end{aligned}\right\} \tag{3}$$

其中,$W\in\mathbb{R}^{K\times A}$,$b\in\mathbb{R}^{L\times A}$,$V\in\mathbb{R}^{A\times 1}$ 表示网络中需要训练的参数,A 为超参数.该层最终输出为 Q_1.

2.4.2 多粒度卷积核神经网络(MFCNN)模型

在文本分类任务中,传统 CNN 的隐藏层只使用单一粒度的卷积核.MFCNN 提出了在单粒度上扩展成多个粒度卷积核的方法,通过不同的卷积域,抽取句子中不同位置的 n-gram 特征.假设 $X\in\mathbb{R}^{L\times K}$ 表示输入的句子,L 为句长,K 为特征维度.$x_i\in\mathbb{R}^K$ 对应于句子中第 i 个词的 K 维特征.卷积操作卷积核为 $W^h\in\mathbb{R}^{h\times K}$,$h$ 为卷积核窗口大小,作用于句子上抽取新的特征.传统的 CNN 使用固定窗口值的多个卷积核,假设为 H,则 CNN 单个卷积核生成的特征 c_i 表示为

$$c_i^H = f(X_{i:i+H-1}\cdot W^H + b) \tag{4}$$

其中,$\cdot$为内积操作,f 为激活函数,i=1,…,L+1−H,$b\in\mathbb{R}$为偏差项.卷积核对句子$\{X_{1:H+1},X_{2:H+1},\ldots,X_{L-H+1:L}\}$生成特征如下:

$$c^H=[c_1,c_2,\ldots,c_{L-H+1}] \tag{5}$$

再使用 max-pooling 提取特征:

$$h^H=\max-pooiing(c^H) \tag{6}$$

为了减轻因为卷积层参数误差造成的估计均值偏移,我们在公式(6)的基础上额外加入 mean-pooling,即

$$h^H=\max-pooiing(c^H)\oplus mean\text{-}pooling(c^H) \tag{7}$$

其中,⊕为拼接操作.MFCNN 使用多个窗口卷积核,对窗口大小 a=1,…,B 的卷积核,抽取特征为

$$Q_3=h^1\oplus h^2\oplus\ldots\oplus h^B \tag{8}$$

最终得到输出 Q_3.这是单个卷积核的工作流程,在实验中,我们使用的卷积核个数为 m_1.MFCNN 的主要步骤是卷积与池化.我们通过不同宽度的卷积核在整个句子上滑动,每个卷积核都能得到 n 个激活值.CNN 更多的是关注关键词对应的特征,因此容易丢失结构化信息,不注意子结构之间的关系,难以发现长文本中的依存转折等复杂的关系.其优点是可以从不同的 n-gram 级别中抽取不同的特征,然后通过池化层提取出激活值中最重要的特征,为后级分类器提供分类依据.在短文本上效果卓越.

2.4.3 Capsule 模型

Capsule 将标量输入与标量输出特征替换成向量输入与向量输出特征,并用动态路由算法代替反向传播算法.在自然语言处理中,可以用以表征如单词的长度、本地顺序或者语义等特征,改善 CNN 在表征上的局限性.我们假设 capsules 数量为 m,d 为 capsule 维度.我们在 capsules 的第 1 层中用宽度为 c 的卷积核 $W^c\in\mathbb{R}^{c\times K}$ 对上一层的表示 $X\in\mathbb{R}^{L\times K}$ 进行卷积操作.其中,L 为句长,K 为上一层句子表示的向量维度,卷积核个数为|$m\times d$|,所有卷积核结果 $\hat{u}$ 为

$$\left.\begin{aligned}\hat{u} &= f(X_{p:p+c-1}W^c+b)\\ \hat{u} &= [u_1,u_2,\ldots,u_{m\times d}]\end{aligned}\right\} \tag{9}$$

其中,f 为激活函数,p=1,…,L+1−c,$b\in\mathbb{R}$为偏差项.然后,通过 reshape 操作分发给各个 capsule:

$$\hat{u}=[u_{(i-1)xd+1},\ldots,u_{i\times d}] \tag{10}$$

其中,i=1,…,m,j=1,…,d.其他层 capsules 的所有输入 s_j 为 $\hat{u}_{j|i}$ 的加权和:

$$s_j=\sum_i c_{ij}\hat{u}_{j|i} \tag{11}$$

我们的向量输出与 hinton 的不同在于,我们只对模长进行了归一化处理:

$$v_j=\frac{s_j}{\|s_j\|} \tag{12}$$

其中,耦合系数 c_{ij} 由算法 1 给出的动态路由算法决定.

算法 1. 动态路由算法.

1. procedure ROUTING($\hat{u}_{j|i}$,r,l)
2. for all capsule i in layer l and capsule j in layer l+1:
3. b_{ij}=0
4. for r iterations do:
5. $c_i \leftarrow softmax(b_i)$
6. $s_j \leftarrow \sum_i c_{ij}\hat{u}_{j|i}$
7. $v_j \leftarrow squash(s_j)$
8. $b_{ij} \leftarrow \hat{u}_{j|i}v_j$
9. return v_j

最后,我们对 Capsule 得到的 $v_j \in \mathbb{R}^{L\times m\times d}$ 做一次压缩操作,得到 $Q_2 \in \mathbb{R}^{L\times|m\times d|}$作为 capsule 网络层的输出.

Softmax 预测层通过混合层我们得到文本最终的特征表示 $X_f=Q_1\oplus Q_2\oplus Q_3\in\mathbb{R}^{1\times K_f}$,通过 $W_s\in\mathbb{R}^{K_f\times N}$ 映射得到总的类别得分:

$$S=X_fW_s+b_s \tag{13}$$

其中,L 为句长,K_f 为最终特征维度,N 为类别数.各个类别的概率为

$$z_i=\frac{e^{s_i}}{\sum_k e^{s_k}} \tag{14}$$

其中,s_i 表示 S 中第 i 类得分.我们使用交叉熵损失函数,训练的目标是最小化损失函数:

$$C=-\sum_i y_i \ln z_i \tag{15}$$

通过小批量随机梯度下降反向传播算法更新权重,其中,y_i 表示正确分类结果.

3 实 验

本节先介绍实验数据,然后介绍实验设置与评价方法,最后介绍实验结果与分析.实验中比较的模型包括 Yang 等人提出的 HAN[4]、Lai 等人提出的 RCNN[5]、Johnson 等人提出的 DPCNN[25]、李超等人提出的 LSTM-MFCNN[6]以及本文提出的混合模型.

3.1 数据集

本文实验数据来自于 CCL 2018-Task1:中国移动客服领域用户意图分类评测赛事(http://www.cips-cl.org/static/CCL2018/call-evaluation.html)初复赛数据语料,属于客服领域对话文本,我们可以将其视为含有若干对话句的段落.该数据共有两个省份的真实数据集,由于不同省份的标注规范与质量的差别,官方将其分为数据集 *A* 与数据集 *B*.每个数据集为 2 万条真实客服对话标注数据,我们将其随机打乱,通过 8:1:1 比例划分,分别分为训练、开发和测试集.除此之外,还有 5 万条真实客服对话未标注数据,我们将其与相应训练集合并,用以训练 Word2Vec 和 ELMo 词向量.对于类别标签,我们将业务类型与用户意图合并,合并后数据集共有 35 种类别.表 2 给出了业务类型与用户意图的种类.

Table 2 Type of business and user intent

表 2 业务类型种类与用户意图种类

业务类型	用户意图
咨询(含查询)	业务订购信息查询
	业务规定
	业务订购信息查询
	业务资费
	产品/业务功能
	使用方式
	办理方式
	号码状态
	宽带覆盖范围
	工单处理结果
	服务渠道信息
	用户资料
	电商货品信息
	营销活动信息
	账户信息
投诉(含抱怨)	不知情定制问题
	业务使用问题
	业务办理问题
	业务规定不满
	信息安全问题
	服务问题
	网络问题
	营销问题
	费用问题
办理	下载/设置
	停复机
	取消
	变更
	开通
	打印/邮寄
	移机/装机/拆机
	缴费
	补换卡
	重置/修改/补发
	销户/重开

由于文本内容长度对分类效果有一定的影响,因此我们对极少数较长文本进行截断,截断后文本最大长度为所有文本长度占比为 95%的数值.考虑到 RNN 隐藏层初始化为 0 的冷启动问题,我们对于长度不足的文本,在句首补齐占位符.表 3 和表 4 分别给出了数据集 *A* 和数据集 *B* 的相关统计信息.

Table 3 Statistics of dataset *A*

表 3 数据集 *A* 统计数据

	训练集	验证集	测试集
原实例数	16 000	2 000	2 000
段落最大长度	3 728	1 865	2 612
段落占比 95%长度	601	573	591
单句最大长度	582	204	260
单句占比 95%长度	37	37	36
最大句子数量	247	120	136
句子占比 95%长度	41	40	42

Table 4 Statistics of dataset *B*

表 4 数据集 *B* 统计数据

	训练集	验证集	测试集
原实例数	16 000	2 000	2 000
段落最大长度	2 043	1 414	2 562
段落占比 95%长度	512	520	517
单句最大长度	372	190	572
单句占比 95%长度	39	39	39
句子数量	111	89	106
句子占比 95%长度	35	35	36

3.2 实验设置

实验中,每训练迭代 $\frac{4000}{B}$ 次,测实验证集的得分值,每次达到极大值时存储模型,其中,$|B|$为批次(batch size),我们设置为 128.为了防止过拟合,如果连续 8 次(即约为两个 epoch 的数量)没有新的极大值出现,我们则认为其达到收敛,提前结束训练,最大 epoch 值为 20.Dropout 都设为 0.5,优化器采用 Adam,学习率为 0.001,β_1 为 0.9,β_2 为 0.999.对于 HAN 模型,我们设置最大句子数量为 50,最大句长为 50,超过的截取,不足的补齐.

词编码层以词作为语义单元,以各自的训练集和 5 万条未标注数据作为预训练语料.设置最大段长为 600,Word2Vec 维度为 300,窗口为 5,最小词频为 5.ELMo 采用原论文默认参数[8],数据集 *A* 上迭代 80 100 个 batch,混乱度(preplexity)为 7.991;数据集 *B* 上迭代 82 000 个 batch,混乱度为 8.423.句子编码层 LSTM 与 GRU 隐藏层 h_1,h_2 为 128 与 100.多粒度卷积核神经网络层卷积核窗口大小为 1、2、3、4,卷积核个数 m_1 为 64.胶囊网络层设置 m 为 10,d 为 16.注意力层设置 A 为 300.

3.3 评价方法

官方评价指标为

$$P=\frac{0.4\times|G_1|+0.6\times|G_1\cap G_2|}{|A|} \tag{16}$$

其中,$|A|$代表预测集总数;$|G_1|$代表一级类别正确识别总数;$|G_1\cap G_2|$表示一级类别正确情况下,二级类别正确识别总数.因为在不同任务中,层级类别权重的不确定性会导致官方评价指标有一定的局限性,因此,我们选择更有泛化意义的准确率评价指标:

$$P=\frac{|G\cap A|}{|A|} \tag{17}$$

其中,$|A|$代表预测集总数,$|G\cap A|$代表预测集与合并标签完全匹配的结果总数.

3.4 实验结果与分析

在本节中,我们在数据集 *A* 和 *B* 上分别进行实验.我们共有 Baseline、Single、Hybrid 和 Hybrid ELMo 这 4 种实验组,实验中如没有特殊说明,则表示采用 Word2Vec 词向量.其中,+Capsule 表示在 Baseline 中加入 Capsule 层,+Attention 在 Baseline 中加入 Attention 层,+MFCNN 在 Baseline 中加入 MFCNN 层,+ELMo 替换 Word2vec 词向量为 ELMo 词向量.表 5 给出了我们的所有模型的实验结果.从实验结果可以看出:

- 通过 Single 组看出,在意图任务上,MFCNN 逊色于 capsule,证明了向量输出优于标量输出方法.
- 对比 Hybrid 与 Single 组实验结果,在 word2vec 词向量上,任意混合优异网络层的网络,在两个数据集上都能达到很好的效果,说明了混合模型的有效性.
- 对比 Hybrid ELMo 与 Hybrid 组实验结果,我们验证了语言模型词向量在意图任务上的有效性.在同样的模型上,语言模型在数据集 *A* 和数据集 *B* 上能够分别取得 2.0%和 2.2%的性能提升,并能结合混合模型取得目前最佳的效果.

Table 5 Result of experiment

表 5 实验结果

实验组	模型	数据集 *A*	数据集 *B*
Baseline	BiLSTM-BiGRU	66.45	54.50
Single	+Attention	66.85	55.45
	+MFCNN	66.70	55.50
	+Capsule	67.35	55.75
Hybrid	+Capsule +MFCNN	67.70	56.10
	+Capsule +MFCNN +Attention	67.45	56.25
Hybrid ELMo	+Capsule +ELMo	68.70	58.25
	+MFCNN +Capsule +ELMo	69.35	58.30
	+MFCNN +Attention +Capsule +ELMo	**69.45**	**58.45**

3.5 长中短文本意图分类对比分析

为了进一步验证混合模型的鲁棒性和有效性,我们将官方数据集*A*和官方数据集*B*的测试集按照长度从短到长排序,按比例为 3:4:3 划分成短、中、长文本.使用不同的模型依次对其进行对比实验,实验结果见表 6,图 3 是对应的直方图.

Table 6 Intention classification score of long, medium and short text

表 6 长、中、短文本意图分类得分结果

模型	短文本		中文本		长文本	
	数据集 *A*	数据集 *B*	数据集 *A*	数据集 *B*	数据集 *A*	数据集 *B*
+MFCNN	73.56	64.90	67.55	54.37	58.64	47.61
+Capsule	74.38	65.39	67.42	54.37	60.13	47.94
+Attention	72.08	63.60	67.30	54.05	60.96	49.00
Hybrid Elmo (Ours)	76.52	67.72	69.20	58.17	61.96	49.42

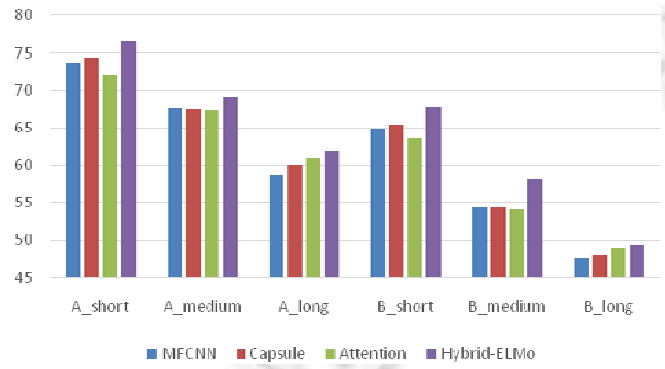


Fig.3 Intention classification score of long, medium and short text

图 3 长、中、短文本意图分类得分结果

从实验结果可以看出:

- 3 种网络的性能都满足短文本>中文本>长文本;
- 在短文本上,性能满足 Capsule>MFCNN>Attention;
- 在中文本上,3 种模型效果持平;

- 在长文本上,性能满足 Attention>Capsule>MFCNN.

MFCNN 主要是在单窗口大小卷积核 CNN 模型的基础上增加不同卷积核窗口,类似于对不同 *n*-gram 的特征提取.而每个卷积核都是为了抽取出文本中和卷积核窗口大小相等的最重要的特征,因此,MFCNN 几乎不受非重要词特征的影响,所以在文本较短的语料上效果优异.但是其缺点在于难以考虑到词之间的关联和不同词特征的重要程度,从而在文本较长的语料上效果基本不如 Attention 网络.Capsule 在自然语言处理中可以用以表征如单词长度、本地顺序或者语义等特征,改善 CNN 在局部表征上的局限性,在中等长度文本上有优异表现.三者相互合作具有一定的 bagging[27]效果.避免混合模型在短文本和长文本的极端情况下可能出现的性能剧烈波动情况,从而增强混合模型的鲁棒性和性能.

3.6 与其他工作的对比

我们将本文提出的方法与其他先进的方法进行了对比,结果见表 7.其中,Hybrid 对应表 5 中 Hybrid 组最佳模型,Hybrid ELMo 对应表 5 中 Hybrid ELMo 组最佳模型.通过对比看出,我们的 Hybrid 模型能取得了相对于其他模型更好的效果.相对于最优的 HAN 模型,在数据集 *A* 和数据集 *B* 上分别有 0.95%和 1.65%的性能提升,证明了混合模型的有效性.同时,结合语言模型词向量在两个数据集上相对于 HAN 模型取得了 2.95%和 3.85%的性能提升.我们最终将所有表现优异的模型进行融合[28],获得了表 8 中官方评分 Rank 1 的成绩.

Table 7 Comparison with other methods

表 7 与其他方法的对比

模型	数据集 *A*	数据集 *B*
RCNN[5]	64.00	50.25
BiLSTM-CNN[16]	64.90	51.00
DPCNN[25]	65.20	52.30
BiLSTM-MFCNN[6]	66.40	53.00
HAN[4]	66.50	54.60
Hybrid (Ours)	67.45	56.25
Hybrid ELMo (Ours)	**69.45**	**58.45**

Table 8 Official ranking

表 8 官方排名

	得分
Rank 1 (Ours)	**70.20**
Rank 2	70.06
Rank 3	68.38
Rank 4	68.07
Rank 5	67.82

4 结论与未来的工作

本文提出一种混合神经网络层的模型,结合 MFCNN 和 Capsule 在短文本特征处理和 Attention 在长文本特征处理上的优势,混合使用 Capsule、Attention 与 MFCNN 层.在此基础上,结合语言模型词向量 ELMo,将语言模型拥有的语义挖掘能力应用到混合网络中.实验结果表明,本文提出的新模型对客服领域的意图分类有较好的性能表现,并在 CCL 2018 中国移动客服领域用户意图分类评测任务中取得第 1 名.

本文使用的神经网络层和特征编码词向量还有一定的局限性,我们准备从下述几个方面进行改进.

(1) 探究更多、更好的短文本和长文本处理优异模型的结合方式.

(2) 在特征编码词向量上,在语义单元上可以考虑字级别特征以及字级别与词级别的结合.

(3) 我们也可以使用迁移学习,比如通过百度百科或者维基百科语料来预训练语言模型,再通过相应的训练语料进行微调(finetune).

(4) 最近提出的语言模型 BERT[14],其效果在多个任务上优于 ELMo,在后续的研究中,我们也会主要对其进行尝试.

References:

[1] Morbini F, De Vault D, Sagae K, Gerten J, Nazarian A, Traum D. FLoReS: A forward looking, reward seeking, dialogue manager. In: Natural Interaction with Robots, Knowbots and Smartphones. New York: Springer-Verlag, 2012. 313−325.

[2] Tur G, Celikyilmaz A, Hakkani-Tür D. Latent semantic modeling for slot filling in conversational understanding. In: Proc. of the 2013 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing. IEEE Computer Society Press, 2013. 8307−8311.

[3] Eyben F, Wöllmer M, Graves A, Schuller B, Douglas-Cowie E, Cowie R. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. Journal on Multimodal User Interfaces, 2010,3(1-2):7−12.

[4] Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2016. 1480−1489.

[5] Lai S, Xu L, Liu K, Zhao J. Recurrent convolutional neural networks for text classification. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. AAAI, 2015. 2267−2273.

[6] Li C, Chai YM, Nan XF, Gao ML. Research on problem classification method based on deep learning. Computer Science, 2016, 43(12):115−119 (in Chinese with English abstract).

[7] Szegedy C, Liu W, Jia Y, 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. 2015. 1−9.

[8] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.

[9] Mathew J, Radhakrishnan D. An FIR digital filter using onehot coded residue representation. In: Proc. of the 10th European Signal Processing Conf. IEEE Computer Society Press, 2000. 1−4.

[10] Irsoy O, Cardie C. Opinion mining with deep recurrent neural networks. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2014. 720−728.

[11] Goldberg Y, Levy O. Word2vec explained: Deriving Mikolov *et al.*'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722, 2014.

[12] 12Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2014. 1532−1543.

[13] Howard J, Ruder S. Universal language model fine-tuning for text classification. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, Vol.1. Association for Computational Linguistics, 2018. 328−339.

[14] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[15] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: Proc. of the European Conf. on Machine Learning. Berlin, Heidelberg: Springer-Verlag, 1998.

[16] Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.

[17] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. on Neural Networks, 1994,5(2):157−166.

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

[19] Chung J, Gulcehre C, Cho KH, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

[20] Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: Advances in Neural Information Processing Systems. MIT Press, 2017. 3856−3866.

[21] Mnih V, Heess N, Graves A. Recurrent models of visual attention. In: Advances in Neural Information Processing Systems. MIT Press, 2014. 2204−2212.

[22] Chen H, Sun M, Tu C, Lin Y, Liu Z. Neural sentiment classification with user and product attention. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2016. 1650−1659.

[23] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proc. of the 3rd Int'l Conf. on Learning Representations. arXiv preprint arXiv:1409.0473, 2014.

[24] Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P. Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems. MIT Press, 2015. 1693−1701.

[25] Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, Vol.1. Association for Computational Linguistics, 2017. 562−570.

[26] Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv: 1207.0580, 2012.

[27] Breiman L. Bagging predictors. Machine Learning, 1996,24(2):123−140.

[28] 2018. https://mlwave.com/kaggle-ensembling-guide/

附中文参考文献:

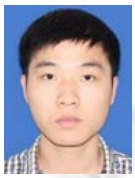
[6] 李超,柴玉梅,南晓斐,高明磊.基于深度学习的问题分类方法研究.计算机科学,2016,43(12):115−119.

周俊佐(1995－),男,四川安岳人,硕士,CCF 学生会员,主要研究领域为自然语言处理.

陈文亮(1977－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为自然语言处理.

朱宗奎(1994－),男,硕士,CCF 学生会员,主要研究领域为自然语言处理.

张民(1970－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,机器翻译,人工智能.



何正球(1993－),男,硕士,CCF 学生会员,主要研究领域为自然语言处理.