计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(12):268−272 [doi: 10.15888/j.cnki.csa.008196]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Transformer 的改进短文本匹配模型①

蔡林杰, 刘　新, 刘　龙, 唐　朝

(湘潭大学 计算机学院·网络空间安全学院, 湘潭 411105)
通讯作者: 刘　新, E-mail: liuxin@xtu.edu.cn

摘　要: 短文本匹配是自然语言处理领域中的一个核心问题, 可应用于信息检索、问答系统、复述问题等任务. 过去的工作大多在提取文本特征时只考虑文本内部信息, 忽略了两个文本之间的交互信息, 或者仅进行单层次交互. 针对以上问题, 提出一种基于 Transformer 改进的短文本匹配模型 ISTM. ISTM 模型以 DSSM 为基本架构, 利用 BERT 模型对文本进行向量化表示, 解决 Word2Vec 一词多义的问题, 使用 Transformer 编码器对文本进行特征提取, 获取文本内部信息, 并考虑两个文本之间的多层次交互信息, 最后由拼接向量推理计算出两个文本之间的语义匹配度. 实验表明, 相比经典深度短文本匹配模型, 本文提出的 ISTM 模型在 LCQMC 中文数据集上表现出了更好的效果.

关键词: 短文本匹配; Transformer; DSSM; BERT; 多层次交互信息; LCQMC

引用格式: 蔡林杰,刘新,刘龙,唐朝.基于 Transformer 的改进短文本匹配模型.计算机系统应用,2021,30(12):268–272. http://www.c-s-a.org.cn/1003-3254/8196.html

Improved Short Text Matching Model Based on Transformer

CAI Lin-Jie, LIU Xin, LIU Long, TANG Chao

(School of Computer Science & School of Cyberspace Science, Xiangtan University, Xiangtan 411105, China)

Abstract: Short text matching is a core problem in the field of natural language processing, which can be applied to tasks such as information retrieval, question answering systems, and question paraphrase. Most of the past work only considered the internal information of the text when extracting text features, ignoring the interactive information between two texts, or only performed single-level interaction. Given the above problems, an Improved Short Text Matching model (ISTM) based on Transformer is constructed. The ISTM model takes DSSM as the basic architecture and uses the BERT model to vectorize the text to solve the ambiguity of Word2Vec. It relies on the Transformer encoder to extract features of the text and obtain its internal information. It considers the multi-level interactive information between the two texts and finally infers and computes the degree of semantic matching between two texts by the concatenated vector. Experiments show that compared with the classic deep short text matching model, the ISTM model proposed in this study shows better results on the LCQMC Chinese dataset.

Key words: short text matching; Transformer; DSSM; BERT; multi-level interactive information; LCQMC

短文本匹配[1] 是自然语言处理 (NLP) 领域中一项被广泛使用的核心技术, 它旨在分析和判断两个文本之间的语义关系, 广泛应用于信息检索[2]、问答系统[3]、复述识别[4] 和自然语言推理[5] 等任务. 在信息检索问题中, 用户想找到和给定查询相关的文档. 对于搜索引擎来说, 如何对给定的查询匹配到合适的文档是至关

① 基金项目: 智能化公共法律服务关键技术湖南省重点研发项目 (2022SK2106)
Foundation item: Key R&D Project of Hunan Province on Key Technologies of Intelligent Public Legal Services (2022SK2106)
收稿时间: 2021-02-24; 修改时间: 2021-03-19; 采用时间: 2021-03-26

重要的. 文本匹配还能被用来在问答系统中为问题匹配到合适的答案, 这对自动客服机器人非常有帮助, 可以大大降低人工成本. 复述识别用于识别两个自然问句是否语义一致, 而自然语言推理主要关注的问题是能否由前提文本推断出假设文本. 因此, 对短文本匹配的研究有重大意义.

传统的文本匹配算法主要解决词汇层面的匹配问题, 存在着词义局限、结构局限和知识局限等问题. 随着科学技术的飞速发展和深度学习的不断壮大, 深度神经网络模型在自然语言处理领域取得了巨大进展, 使用深度神经网络表示文本、学习文本之间的交互模式使得模型能够挖掘出文本之间复杂的语义关系, 关于短文本匹配问题的研究已经逐渐从传统的基于统计的方法转移到深度语义短文本匹配模型. 近年来提出的 Word2Vec[6]、GloVe[7]、ELMO[8]、BERT[9]、XLNet[10] 等预训练模型很好地解决了文本向量化表示的问题.

目前大多数短文本匹配模型在提取文本特征时只考虑文本内部信息, 忽略了两个文本之间的交互信息, 或者仅进行单层次交互, 丧失了文本间丰富的多层次交互信息. 为解决上述问题, 本文提出一种基于 Transformer[11] 改进的短文本匹配模型 ISTM. ISTM 模型以 DSSM[12] 为基本架构, 利用 BERT 模型对文本进行向量化表示, 解决 Word2Vec 一词多义的问题, 使用 Transformer 编码器对文本进行特征提取, 获取文本内部信息, 并考虑两个文本之间的多层次交互信息, 分别得到两个文本最终的语义表示, 经过最大池化后进行拼接操作再通过全连接神经网络分类输出两个文本之间的语义匹配度. 实验表明, 相比经典深度短文本匹配模型, 本文提出的 ISTM 模型在 LCQMC[13] 中文数据集上表现出了更好的效果, 证明了该模型的有效性和可行性.

1 背景技术

DSSM 是一个非常出名的文本匹配模型, 它首先被应用于 Web 搜索应用中匹配查询 (query) 和相关文档 (documents). DSSM 使用神经网络将查询和文档表示为向量, 两向量之间的距离被视为它们的匹配得分.

Transformer 模型于 2017 年由 Google 提出, 主要用于自然语言处理领域, 有一个完整的 Encoder-Decoder 框架, 其主要由注意力 (attention)[14–16] 机制构成. 其中, 每个编码器均由自注意力机制和前馈神经网络两个主要子层组成. Transformer 旨在处理顺序数据 (例如自然语言), 但不需要按顺序处理顺序数据, 与循环神经网络 (RNN)[17] 相比, Transformer 允许更多的并行化, 因此减少了训练时间, 可以对更大的数据集进行训练. 自问世以来, Transformer 已成为解决 NLP 中许多问题的首选模型, 取代了旧的循环神经网络模型.

2018 年 Google 提出了 BERT 模型, BERT 模型主要利用了 Transformer 的双向编码器结构, 采用的是最原始的 Transformer. 与最近的其他语言表示模型不同, BERT 旨在通过联合调节所有层中的上下文来预先训练深度双向表示, 能够解决一词多义的问题.

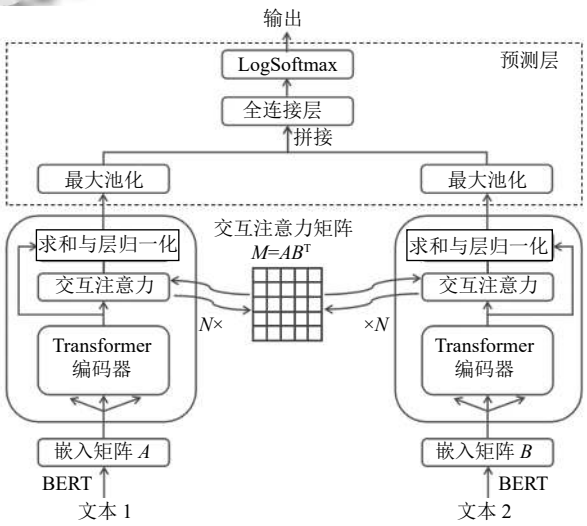

图 1 ISTM 模型架构图

2 改进的短文本匹配模型 ISTM

现有的短文本匹配方法大多只考虑文本自身的内部信息, 忽略了两个文本之间的交互信息, 或者仅在提取文本特征之后才进行一次交互, 只能获取单层次交互信息, 而丧失了多层次交互信息. 为此, 本文提出了基于 Transformer 改进的短文本匹配模型 ISTM, 该模型以 DSSM 为基本架构, 利用 BERT 模型实现文本向量化表示, 将 Transformer 编码器作为一个特征提取器, 同时借助 Transformer 的多层编码器使得两个文本进行多层次交互, 进而获取多层次交互信息, 最终分别得到两个文本的语义表示, 其中 Transformer 编码器个数即为文本交互次数. 两个文本的语义表示进入预测层, 分别进行最大池化, 而后得到的拼接向量进入全连接

层, 最终通过分类器 LogSoftmax 输出匹配结果. ISTM 模型架构如图 1 所示, 其中 N 为编码器数量.

2.1 Transformer 编码器

Transformer 编码器有两个子层, 分别是多头自注意力层和前馈神经网络层, 同时每个子层的周围都有一个求和与层归一化[18] 步骤, 其结构如图 2 所示, 其中 N 为编码器数量.

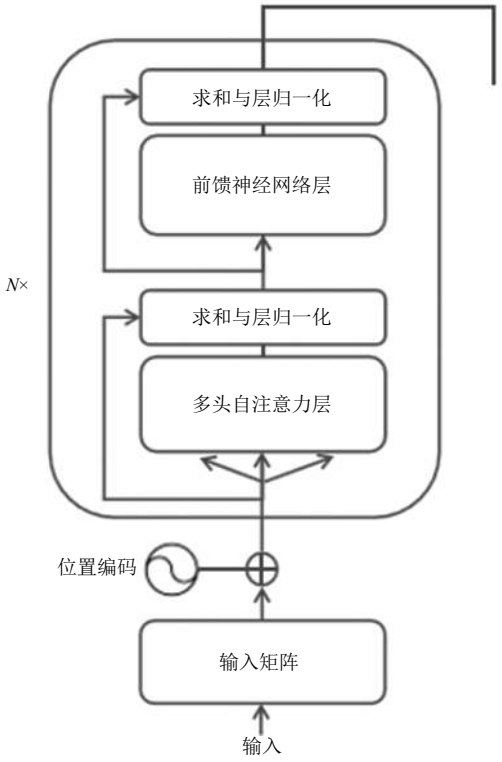

图 2 Transformer 编码器结构图

输入矩阵 X 的维度为 $S\times E$, 其中 S 为最大序列长度, E 为嵌入向量的维数, 本文中 S 为 25, E 为 768. Transformer 编码器的计算过程如下:

(1) 自注意力机制通过输入矩阵 X 和权重矩阵 W^Q, W^K, W^V 分别计算查询矩阵 Q, 键矩阵 K 和值矩阵 V.

$$Q = XW^Q \tag{1}$$

$$K = XW^K \tag{2}$$

$$V = XW^V \tag{3}$$

(2) 计算自注意力层的输出矩阵 Z, 其中 d_k 为键向量的维数.

$$Z = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{4}$$

(3) 计算多头自注意力层的输出矩阵 Z_{mul}, 其中 H 为注意力头数, Z_i 表示第 (i+1) 个注意力头, *Concat* 函数表示将所有的 H 个注意力头拼接起来, W^O 为附加的权重矩阵, Z_{mul} 的维度与 X 相同.

$$Z_{\mathrm{mul}} = Concat(Z_i)W^O,\ i = 0,1,\cdots,H-1 \tag{5}$$

(4) 进行求和与层归一化, 其中 LN 函数表示层归一化.

$$Z_{\mathrm{mul}} = LN(Z_{\mathrm{mul}} + X) \tag{6}$$

(5) 将 Z_{mul} 传递到前馈神经网络, 之后再次进行求和与层归一化.

2.2 文本向量化表示

文本向量化表示是自然语言处理任务中的一个重要过程. Word2Vec 是自然语言处理中最早的预训练模型之一, 过去的工作大多采用 Word2Vec 实现文本的向量化表示. Word2Vec 的优点是简单、速度快、通用性强, 但它受限于语料库, 产生的词表示是静态的, 无法解决一词多义的问题, 它的建模较为简单, 不能体现词的多层特性, 包括语法、语义等.

本文采用 BERT 模型来处理文本的向量化表示, BERT 模型相比以 Word2Vec 为代表的词嵌入方法, 一个比较突出的进步就是更加动态, 能够解决一词多义的问题. 此外, BERT 基于 Transformer, 利用了 Transformer 的双向编码器结构, 能够体现词的多层特性. 本文使用 BERT-BASE 中文模型, 该模型有 12 个 Transformer Encoder, 每个 Transformer Encoder 有 12 个注意力头, 隐藏层维数为 768. BERT 模型实现文本向量化的过程如图 3 所示.

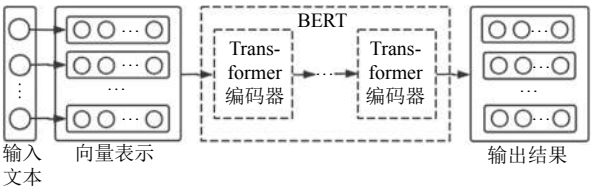

图 3 BERT 文本向量化

2.3 交互注意力层

假设文本 1 的矩阵表示为 $A_{S\times E}$, 文本 2 的矩阵表示为 $B_{S\times E}$, MP 函数表示最大池化操作, 则两个文本的交互注意力矩阵 $M_{S\times S}$ 的计算如下:

$$M = AB^{\mathrm{T}} \tag{7}$$

对 M 的每一行进行最大池化操作, 得到向量 r_S, 表

示 B 对 A 中每个字符的注意力权重, r_S 中的每个元素乘以 A 中的每一行即可得到交互后的 A:

$$A = MP_{\text{row}}(M) \times A \tag{8}$$

对 M 的每一列进行最大池化操作, 得到向量 c_S, 表示 A 对 B 中每个字符的注意力权重, c_S 中的每个元素乘以 B 中的每一行即可得到交互后的 B:

$$B = MP_{\text{col}}(M) \times B \tag{9}$$

如此, 两个文本进行了一次信息交互. 两个文本依据编码器数量可进行多层次信息交互, 获得丰富的上下文信息和交互信息.

2.4 预测层

假设经过编码组件后两个文本的矩阵表示分别为 $A_{S\times E}$ 和 $B_{S\times E}$, 将 A 和 B 进行最大池化后分别得到两个文本的向量表示 $v1$ 和 $v2$, 则两个文本的匹配结果的计算如下:

$$y' = F([v1; v2; v1 \circ v2; |v1 - v2|]) \tag{10}$$

其中, y'表示两个文本的匹配结果的预测值, $v1 \circ v2$ 表示 $v1$ 和 $v2$ 对应元素逐个相乘, 强调两个文本之间相同之处, 而$|v1–v2|$强调两个文本之间不同之处, F 函数表示将这 4 个向量的拼接向量输入到全连接神经网络后再经过 LogSoftmax 分类器处理输出匹配结果的预测值.

3 实验与分析

3.1 数据集和模型参数

LCQMC 数据集由哈工大 (深圳) 智能计算研究中心提供, 包含来自多个领域的 260 068 个中文问句对, 相同询问意图的句子对标记为 1, 否则为 0, 并预先将其切分为了训练集 (238 766 对)、验证集 (8802 对) 和测试集 (12 500 对), 其中同义对和非同义对的比例接近 1:1.

本次实验设置的主要模型参数如表 1 所示.

表 1 模型参数设置表

参数	取值
编码器层数	2
注意力头数	8
隐藏层维数	768
多头自注意力的dropout	0.1
模型优化器	Adam
最大序列长度	25
批大小batch_size	512

3.2 评估指标

本文的 ISTM 模型所应用的短文本匹配属于二分类范畴, 我们采用 $F1$ 值和准确率 Acc 作为评估模型效果的指标, 以 $F1$ 值为主, 以准确率为辅. $F1$ 值由精确率 P 和召回率 R 得到, 相关的计算公式如下:

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{13}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

3.3 模型对比实验

为了验证本文模型效果, 基于同一数据集, 选取多个经典的文本匹配模型进行实验对比.

LSTM-DSSM[19]: 针对 DSSM 不能够很好地捕捉上下文信息的问题, Palangi 等提出使用 LSTM 替换 DSSM 的深度神经网络, 以捕捉到文本的上下文信息.

Transformer-DSSM[20]: 为了更强的并行计算能力和更好的特征提取能力, 赵梦凡提出使用 Transformer 编码组件替换 DSSM 的深度神经网络.

为了增强对比效果, 本文还类似地加入了 RNN-DSSM、BiRNN-DSSM、GRU-DSSM、BiGRU-DSSM 和 BiLSTM-DSSM 作为对比实验. 各模型的实验对比结果如表 2 所示, 其中模型采用的文本向量化方式默认为 BERT, 否则为 Word2Vec.

表 2 模型对比实验结果

序号	模型	P	R	$F1$	Acc
1	RNN-DSSM	80.4	91.2	85.5	84.5
2	BiRNN-DSSM	81.2	90.8	85.7	84.9
3	GRU-DSSM	79.9	91.8	85.4	84.3
4	BiGRU-DSSM	75.5	**95.5**	84.3	82.3
5	LSTM-DSSM	78.5	93.3	85.3	83.9
6	BiLSTM-DSSM	78.5	93.7	85.5	84.1
7	Word2Vec-BiLSTM-DSSM	67.4	91.0	77.5	73.5
8	Transformer-DSSM	82.3	90.3	86.3	85.7
9	ISTM	**83.7**	90.1	**86.8**	**86.3**

从实验结果可以看出, 本文提出的 ISTM 模型的 $F1$ 值可达到 86.8%, 准确率可达到 86.3%, 优于其他模型. 由实验 6 和实验 7 可知, BERT 的表现比 Word2Vec 更加优秀, 以至于实验 1 至实验 5 所代表的其他 RNN 模型的 $F1$ 值和准确率都大幅领先于实验 7. 由实验 1 至实验 8 可知, Transformer 编码器相对于 RNN 拥有更好的特征提取能力. 由实验 8 和实验 9 可知, ISTM

模型的多层次信息交互确实提升了短文本匹配的效果, 主要体现在 *F*1 值和准确率的提升. ISTM 模型取得更好的匹配效果, 其原因在于:

(1) 使用 BERT 模型进行文本向量化, 能够解决一词多义的问题;

(2) Transformer 编码器拥有更为优秀的特征提取能力;

(3) 多层次信息交互使得两个文本都获得丰富的交互信息, 这对短文本匹配效果有着较好的提升.

4 结束语

针对短文本匹配问题, 本文提出了一种基于 Transformer 改进的短文本匹配模型 ISTM, 该模型利用 BERT 实现文本向量化表示, 在 DSSM 模型架构的基础上引入 Transformer 编码器作为特征提取器, 并增加交互注意力层, 使得模型自身拥有获取多层次交互信息的能力. 从对比实验结果来看, 本文提出的 ISTM 模型能够有效提升短文本匹配的效果.

参考文献

1 庞亮, 兰艳艳, 徐君, 等. 深度文本匹配综述. 计算机学报, 2017, 40(4): 985–1003.

2 Li H, Xu J. Semantic Matching in Search. Hanover: Now Publishers Inc., 2014.113. [doi: 10.1561/9781601988058]

3 王瑛, 何启涛. 智能问答系统研究. 电子技术与软件工程, 2019, (5): 174–175.

4 陈鑫, 李伟康, 洪宇, 等. 面向问句复述识别的多卷积自交互匹配方法研究. 中文信息学报, 2019, 33(10): 99–108, 118. [doi: 10.3969/j.issn.1003-0077.2019.10.012]

5 李冠宇, 张鹏飞, 贾彩燕. 一种注意力增强的自然语言推理模型. 计算机工程, 2020, 46(7): 91–97.

6 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. arXiv: 1310.04546v1, 2013.

7 Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 1532–1543.

8 Peters ME, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans: Association for Computational Linguistics, 2018. 2227–2237.

9 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.

10 Yang ZL, Dai ZH, Yang YM, *et al*. XLNet: Generalized autoregressive pretraining for language understanding. arXiv: 1906.08237, 2019.

11 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

12 Huang PS, He XD, Gao JF, *et al*. Learning deep structured semantic models for web search using clickthrough data. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco: ACM, 2013. 2333–2338.

13 Liu X, Chen QC, Deng C, *et al*. LCQMC: A large-scale Chinese question matching corpus. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018. 1952–1962.

14 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473, 2014.

15 Luong T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015. 1412–1421.

16 Yang ZC, Yang DY, Dyer C, *et al*. Hierarchical attention networks for document classification. Proceedings of NAACL-HLT 2016. San Diego: Association for Computational Linguistics, 2016. 1480–1489.

17 夏瑜潞. 循环神经网络的发展综述. 电脑知识与技术, 2019, 15(21): 182–184.

18 Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv: 1607.06450, 2016.

19 Palangi H, Deng L, Shen Y, *et al*. Semantic modelling with long-short-term memory for information retrieval. arXiv: 1412.6629, 2014.

20 赵梦凡. 基于 Transformer 的文本语义相似度算法研究 [硕士学位论文]. 湘潭: 湘潭大学, 2020.