计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.20200626
59(1):118-126,2022

关于短文本匹配的泛化性和迁移性的研究分析

马新宇　范意兴　郭嘉丰　张儒清　苏立新　程学旗
(中国科学院网络数据科学与技术重点实验室(中国科学院计算技术研究所)　北京　100190)
(中国科学院大学　北京　100049)
(maxinyu17g@ict.ac.cn)

An Empirical Investigation of Generalization and Transfer in Short Text Matching

Ma Xinyu, Fan Yixing, Guo Jiafeng, Zhang Ruqing, Su Lixin, and Cheng Xueqi
(*CAS Key Laboratory of Network Data Science & Technology* (*Institute of Computing Technology, Chinese Academy of Sciences*), *Beijing* 100190)
(*University of Chinese Academy of Sciences*, *Beijing* 100049)

Abstract　Many tasks in natural language understanding, such as natural language inference, question answering, and paraphrasing can be viewed as short text matching problems. Recently, the emergence of a large number of datasets and deep learning models has made great success in short text matching. However, little study has been done on analyzing the generalization of these datasets across different text matching tasks, and how to leverage these supervised datasets of multiple domains to new domains to reduce the cost of annotating and improve their performance. In this paper, we conduct an extensive investigation of generalization and transfer across different datasets and show the factors that affect the generalization through visualization. Specially, we experiment with a conventional neural semantic matching model ESIM (enhanced sequential inference model) and a pre-trained language model BERT (bidirectional encoder representations from transformers) over 10 common datasets. We show that even BERT which is pre-trained on a large-scale dataset can still improve performance on the target dataset through transfer learning. Following our analysis, we also demonstrate that pre-training on multiple datasets shows good generalization and transfer. In the case of a new domain and few-shot setting, BERT which we pre-train on the multiple datasets first and then transfers to new datasets achieves exciting performance.

Key words　short text matching; generalization; transfer; few-shot; pre-trained language model

摘　要　自然语言理解中的许多任务,比如自然语言推断任务、机器问答和复述问题,都可以看作是短文本匹配问题.近年来,大量的数据集和深度学习模型的涌现使得短文本匹配任务取得了长足的进步,

收稿日期:2020-08-14;**修回日期**:2021-02-22
基金项目:国家自然科学基金项目(61722211,61773362,61872338,62006218,61902381);国家重点研发计划项目(2016QY02D0405);北京智源人工智能研究院项目(BAAI2019ZD0306);中国科学院青年创新促进会项目(20144310,2016102);重庆市基础科学与前沿技术研究专项项目(重点)(cstc2017jcjyBX0059);王宽诚教育基金会项目;联想-中科院联合实验室青年科学家项目
This work was supported by the National Natural Science Foundation of China (61722211, 61773362, 61872338, 62006218, 61902381), the National Key Research and Development Program of China (2016QY02D0405), the Project of Beijing Academy of Artificial Intelligence (BAAI2019ZD0306), the Youth Innovation Promotion Association CAS (20144310, 2016102), the Project of Chongqing Research Program of Basic Research and Frontier Technology (cstc2017jcyjBX0059), the K.C. Wong Education Foundation, and the Lenovo-CAS Joint Lab Youth Scientist Project.
通信作者:范意兴(fanyixing@ict.ac.cn)

然而,很少有工作去分析模型在不同数据集之间的泛化能力,以及如何在新领域中有效地利用现有不同领域中的大量带标注的数据,达到减少新领域的数据标注量和提升性能的目标.为此,重点分析了不同数据集之间的泛化性和迁移性,并且通过可视化的方式展示了影响数据集之间泛化性的因素.具体地,使用深度学习模型 ESIM(enhanced sequential inference model)和预训练语言模型 BERT(bidirectional encoder representations from transformers)在 10 个通用的短文本匹配数据集上进行了详尽的实验.通过实验,发现即使是在大规模语料预训练过的 BERT,合适的迁移仍能带来性能提升.基于以上的分析,也发现通过在混合数据集预训练过的模型,在新的领域和少量样本情况下,具有较好的泛化能力和迁移能力.

关键词 短文本匹配;泛化性;迁移性;少样本;预训练语言模型

中图法分类号 TP18

短文本匹配任务是指输入 2 段短文本,通常是句子级或者短语级,算法或者模型需要预测它们之间的匹配程度.自然语言理解中的许多任务,包括自然语言推断[1]、复述问题[2]、答案选择[3]、问答任务[4]都可以抽象成短文本匹配问题,例如自然语言推断需要判断 2 个句子间是否存在蕴含关系,复述问题需要判断 2 个句子是否语义相等,答案选择和问答任务都需要判断句子是否包含问题的答案.短文本匹配任务在现实中有着广泛的应用,例如在线聊天机器人中对话的匹配[5]、社区问答系统中相似问题的匹配[6],以及机器阅读理解系统[7]中问题和答案的匹配等.

近年来,短文本匹配取得了显著进步,一方面得益于深度学习的发展,深度学习模型在计算机视觉和自然语言处理领域已经几乎显著超越或者碾压所有传统基于规则的方法和机器学习技术,尤其是近年来以 BERT(bidirectional encoder representations from transformers)[8]为代表的首先在大规模无监督语料上学习通用表达的预训练模型,使得自然语言处理(natural language processing, NLP)中的很多任务得到了全面的性能提升.然而,深度学习模型通常需要大量的标注数据来进行训练,现实中往往没有很多标注数据,并且标注数据的成本很高;同时预训练模型一直被认为有较好的泛化能力,但是通过我们的实验发现,在大多数情况下,BERT 的泛化能力是比较差的.另一方面大量公开的数据集也对短文本匹配任务的发展起到了非常重要的促进作用.比如被广泛使用的 GLUE benchmark[9]是很多自然语言理解模型评价的标准数据集集合,10 个任务中的 9 个都可以抽象成短文本匹配问题.问答领域也发布了很多通用的数据集,比如 SQuAD[10],Ms marco[11],尤其是 Ms marco,它的标注数据量达到了百万级.然而,对于这些现有的有标注数据集,很少有研究去分析它们是否能泛化到一个新的数据集,以及如何利用这些现有的大量有标注数据到新的领域中去,达到减少新领域的数据标注量和提升性能的目标.

本文的主要目标为分析短文本匹配数据集之间的泛化性和迁移性,为未来利用迁移学习的方式在新领域提升短文本匹配模型的性能提供有价值的指导.具体来说,我们将分析目标分解为 3 个具体的研究问题.

问题 1. 不同短文本匹配数据集之间的泛化关系是怎么样的,与哪些因素有关?

问题 2. 是否能利用其他领域的数据集提升目标领域数据集的性能?

问题 3. 基于以上分析,如何有效利用其他领域的标注数据来提升新领域的性能?

对于问题 1,我们首先在源数据集上进行训练,然后在目标数据集上直接测试性能(泛化),为了规避模型对泛化能力的影响,我们在所有实验中都使用了 2 个深度学习模型进行对照,一个是传统的深度学习模型 ESIM[12],另外一个是预训练语言模型 BERT[8].然后通过力导向图算法可视化数据集之间的关系.对于问题 2,我们首先在源数据集上进行预训练,然后在目标数据集上微调再测试性能(迁移),同时对数据集之间的迁移关系进行了定量的分析,即给定不同的源数据集的数据量,观察迁移带来的性能提升效果.对于问题 3,基于对数据集之间泛化性和迁移性的分析,我们提出了一种简单可行的办法,可以提升模型的泛化能力和迁移能力,并在新领域并且只有很少标注样本的情况下,取得了很好的效果.

本文的主要贡献有 3 个方面:

1) 分析研究了不同数据集上的模型泛化能力.

发现模型的泛化能力主要与数据集的类型和来源有关,并且类型的影响要大于数据集的来源.

2) 分析研究了不同数据集上的模型迁移能力.发现数据集的迁移能力和泛化能力的趋势往往是不一致的;并且即使是预训练模型 BERT,在目标数据集很大的情况下(几十万标注样本),合适的迁移仍能带来性能的提升.

3) 基于对泛化性和迁移性的分析,提出通过将数据集进行平均混合的方式,可以提高模型的泛化能力和迁移能力.实验表明 BERT 在此混合数据集预训练之后,在新的领域并且只有 100 个标注样本的情况下,达到了堪比原始千级和万级标注数据量的效果.

1 任务描述

给定数据集 $S=\{(T_1,T_2,r)_i^N\}$,其中 T_1 和 T_2 是 2 段短文本,r 是标签,表示 2 段短文本的匹配程度,一般是二分类.在自然语言推断中,T_1 和 T_2 分别表示前提和假设,r 代表 2 段文本是否有蕴含关系;在复述问题中,T_1 和 T_2 表示 2 个问题,r 代表 2 段问题是否语义相同;在答案选择和转换后的问答问题中,r 代表后一个句子是否是前一个句子的答案.虽然这 3 种匹配任务的定义不同,但是他们都可以抽象成:

$$r=F(T_1,T_2),$$

其中 F 代表不同的匹配模型.

我们选择了 10 个常用的短文本匹配数据集来进行此次分析实验,这些数据集除了 SNLI[13] 和 MNLI[14] 为三分类,其余全部为二分类.我们按照数据集规模,以 10 万数据量为分界线,将数据集分为大数据集和小数据集.在接下来的实验中,为了消除数据规模的影响,我们固定大数据集的数据量,每个大数据集只随机选取 10 万的数据进行实验.表 1 是关于这些数据集的一些特征描述,包括数据集的数据量、数据集来源和类型.

我们首先介绍大数据集:

SNLI[13] 是斯坦福自然语言推断数据集,它是根据图像数据集人工标注的,需要判断 2 个句子是蕴含、矛盾或者中立关系.

MNLI[14] 是多类型自然语言推断数据集,数据集中的数据来自多种领域,例如演讲、小说和政府报告等.它需要判断 2 个句子是蕴含、矛盾或者中立关系.

QNLI 是由问答数据集 SQuAD[10] 转换成的自然语言推断数据集,数据主要来自于英文维基百科.该数据集需要判断问题的答案是否被另外一个句子所包含.

Ms marco[11] 是一个大规模的微软阅读理解数据集,我们从原始的数据集中抽取了 12 万数据并转换为问题和答案所在句的匹配问题,该数据集的数据来自网页文档.

QQP[15] 是复述问题数据集,该数据集主要来自问答网站 Quora,它需要判断 1 个句子对是否是重复问题,是二分类任务.

小数据集主要有:

WNLI[16] 是一个自然语言推断数据集,数据主要来自于小说,训练集数据量约为 1 000.

RTE[17] 是二分类自然语言推断数据集,数据主要来自于网上新闻,训练集数据量约为 3 000.

SciTail[18] 是自然语言推断数据集,数据主要来自于科学考试中的多项选择题,训练集数据量约为 3 万.

MRPC[2] 是一个复述问题数据集,数据主要来自在线新闻,训练集数据量约为 4 000.

WikiQA[19] 是一个问答数据集,数据主要来自必应的搜索问题和维基百科的文档,训练集数据量约为 3 万.

Table 1 Statistics of Different Matching Datasets

表 1 相关数据集描述

规模	数据集	数据量	来源	类型
大数据集	SNLI	57 万	网页文档	自然语言推断
	MNLI	43.3 万	网页文档	自然语言推断
	QNLI	13 万	维基百科	自然语言推断
	Ms marco	12 万	维基百科	问答
	QQP	40 万	网页文档	复述问题
小数据集	WNLI	0.1 万	书籍	自然语言推断
	RTE	0.3 万	网页新闻	自然语言推断
	SciTail	2.7 万	书籍	自然语言推断
	MRPC	0.4 万	网页新闻	复述问题
	WikiQA	3 万	维基百科	问答

2 模 型

我们使用了 2 种常用的深度学习模型进行分析实验:一种是传统的深度学习模型 ESIM[12],另外一种是预训练语言模型 BERT[8],2 个模型主要用来进行对照.

2.1 ESIM 模型

ESIM[7]是短文本匹配中效果较好的模型之一,它模型结构简单,但是高效,且在短文本匹配中比较通用,它主要由 3 个模块组成.

1) 表达层.使用双向长短记忆网络 LSTM[20]编码 2 个句子的词向量,提取词表达.

2) 特征提取层.由局部注意力机制[21]对 2 个句子进行对齐操作.

3) 匹配层.使用池化层和全连接计算匹配得分.我们使用了 Spacy 分词工具对文本进行分词和 Glove300d[22]初始化词向量.

2.2 BERT 模型

BERT[8]是预训练语言模型,它首先在大量的无标注文档上进行预训练,然后在下游任务上进行微调,它在多种自然语言理解任务上达到了目前最好的效果.模型的结构使用了 Transformer[21]的编码器部分,主要的组成部分包括了自注意力、LayerNorm[23]等.输入由 2 个句子 s_1 和 s_2 拼接而成,[CLS]+s_1+[SEP]+s_2+[SEP],输出为线性全连接层,即分类层.实验中句子最大长度设置为 128,BERT 的分词使用自带的 WordPiece[24]分词方法,一个单词有可能会被分成多个子词,我们使用谷歌开源的 BERT-base-uncased 版本的模型进行实验.

3 实验及分析

为了对比公正,我们严格控制了实验流程,确保所有数据集的训练和测试流程一致,尽量不引入其他变量,只微调部分超参.我们将发布代码,供研究社区使用.ESIM 模型在训练过程中采用了 early-stop 技术,使用 Adam[25]优化器进行优化,使用 Glove300d 初始化词向量;BERT 模型对大数据集学习率设置为 2E−5,小数据集学习率选取 1E−5,2E−5,3E−5,4E−5,5E−5 中效果最好的,训练 5 次取平均值.

3.1 模型是否可以泛化到新的数据集

深度学习模型目前在单个数据集上已经取得了很好的性能,我们很自然地想到是否能从一个数据集直接泛化到一个新的数据集,不做任何训练就能达到很好的性能.我们首先在 5 个大的数据集上进行训练,由于这些数据集规模不一,为了消除数据量的影响,我们对每个大数据集只取 10 万的样本进行训练.另外我们也随机从这 5 个大数据集中随机取 2 万的样本,组成一个新的数据集 Multi-100K (MT100K).在源数据集上训练完模型之后,我们在其他所有数据集上直接进行测试,不在目标数据集上再进行训练.由于 SNLI 和 MNLI 是三分类自然语言推断数据集,我们将其标签从(蕴含、中立和矛盾)合并成 2 类(蕴含和不蕴含)以进行泛化实验.

实验结果如表 2 所示,表 2 中值为准确率,其中上表是 ESIM 模型结果,下表是 BERT 模型结果,虚线右侧是大数据集,虚线左侧是小数据集,行代表源数据集,列代表目标数据集,SELF 行表示在目标数据集上训练和测试,MT100K 行表示在 MT100K

Table 2 Generalization Experimental Results

表 2 泛化实验结果 %

模型	数据集	SciTail	WNLI	RTE	MRPC	WikiQA	SNLI	MNLI	QNLI	Ms marco	QQP
ESIM	SNLI	62.8	46.5	54.9	46.7	86.5	—	**73.4**	50.9	50.0	67.5
	MNLI	67.6	43.7	**63.2**	58.0	**87.0**	**80.8**	—	51.2	49.9	66.8
	QNLI	60.7	53.5	47.7	69.5	76.7	62.8	63.6	—	**74.0**	64.0
	Ms marco	**72.7**	**56.3**	52.0	**69.9**	81.5	64.1	63.9	**74.9**	—	**68.2**
	QQP	62.1	**56.3**	50.5	58.6	86.9	67.4	66.9	53.2	51.8	—
	MT100K	74.7	46.5	62.1	69.5	83.7	—	—	—	—	—
	SELF	82.6	57.8	57.0	69.0	88.0	90.4	81.8	82.4	83.8	88.1
BERT	SNLI	75.8	43.7	70.0	60.9	85.8	—	**82.5**	51.1	50.3	70.5
	MNLI	77.2	42.3	**73.3**	57.3	87.4	**86.7**	—	50.5	50.1	**72.5**
	QNLI	69.3	**57.8**	46.6	**71.3**	**87.6**	59.8	64.5	—	**82.3**	66.1
	Ms marco	**82.7**	43.7	57.8	67.7	79.5	58.0	51.8	**77.1**	—	67.9
	QQP	73.9	47.9	53.5	65.6	86.8	68.8	69.5	54.0	52.3	—
	MT100K	79.4	52.3	70.8	65.4	82.2	—	—	—	—	—
	SELF	94.9	56.3	67.5	85.2	91.4	92.8	88.4	91.3	92.7	88.3

注:黑体值为此数据集泛化最好的结果;“—”代表这些值不可用.

数据集上训练,然后在目标数据集上测试,由于MT100K是从大数据集中取了部分数据,所以它对大数据集不做泛化实验.我们发现模型在不同数据集上的泛化能力非常差.对于ESIM模型,在源数据集上训练,在目标数据集上测试,比在目标数据集上训练和测试(SELF行)的性能平均下降14.1个百分点,即使我们取每个数据集上泛化性能最好的结果,性能也平均下降了6.1个百分点;对于BERT模型,在源数据上训练,在目标数据集上测试与SELF行相比,性能下降了18.7个百分点,即使取每个数据集上泛化能力最好的,性能也下降了7.5个百分点.因此,模型在训练过的数据集上发生了过拟合现象.虽然BERT相比于ESIM模型下降的稍多,但是整体泛化性能还是要比ESIM高2.5个百分点,这得益于BERT在大规模语料上预训练以及模型本身的能力.

3.2 数据集之间的泛化关系与什么因素有关

从3.1节实验中,我们发现数据集之间的泛化性能差异很大,我们在此进一步分析影响数据集之间泛化能力的因素.具体地说,我们使用力导向图算法[26]进行可视化,其中节点代表数据集,形状表示数据集的类型,颜色表示数据集的来源,边我们采用一个数据集在另外一个数据集上的泛化性能来代表.假设P_{ij}表示BERT在数据集S_i上训练和在S_j上测试的性能,P_i表示只在数据集S_i上训练和测试的性能.如果我们在S_i和S_j上训练和测试是双向的,那么2个节点之间的力为$F(S_i,S_j)=\frac{P_{ij}}{P_j}+\frac{P_{ji}}{P_i}$;如果在$S_i$上训练,只在$S_j$上测试,那么$F(S_i,S_j)=\frac{2P_{ij}}{P_j}$.

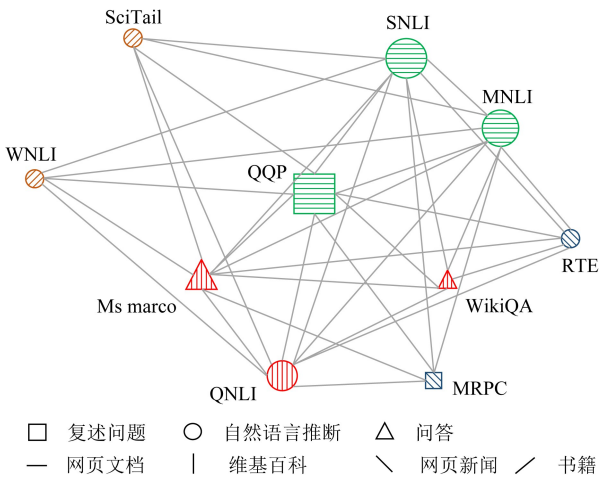


Fig. 1 Visualization of similarity between datasets

图1 数据集之间相似度的可视化

结果如图1所示,我们发现这些数据集很自然地按照任务类型(形状内的纹理)和数据集来源(形状)聚集.泛化性最好的2个数据集一定是类型相同,来源相同,例如图1中右上角区域的SNLI和MNLI,左下角区域Ms marco和QNLI,因为QNLI是从问答数据集中转变过来的,训练数据更像是问答类型,所以它和问答数据集的泛化性能较好,尤其是和Ms marco数据集.因此,数据集之间的泛化能力主要与数据集的类型和来源有关.

3.3 源数据集规模越大,其泛化能力是否越强

从实验中,我们发现MT100K的泛化性能已经接近在目标数据集泛化最好的性能,平均性能仅下降3%.为此,我们基于混合数据集进一步分析源数据集规模大小对其泛化能力的影响.具体地,我们从5个大数据集中等量抽取4万、6万、8万和10万数据进行训练,然后将其泛化到小的数据集.

实验结果如表3所示,我们可以看到增大源数据集规模之后,除了WNLI数据集,其他4个数据集的性能都有所提升,平均性能提高了6个百分点.尤其是MT500K,在SciTail数据集上比MT400K提升了约2个百分点.增大源数据集规模之后,WNLI数据集性能下降的一个可能原因是WNLI的训练和测试集过于小,其中测试集只有几百个测试样本,与预训练数据集的规模差异过大,导致模型在源数据集上出现了过拟合现象.因此一般情况下,源数据集规模越大,泛化性能越强.

Table 3 Generalization Results of Different Sizes of Mixed Datasets on Small Datasets

表3 不同规模的混合数据集在小数据集上的泛化结果 %

数据集	SciTail	WNLI	RTE	MRPC	WikiQA
MT100K	79.4	**52.3**	70.8	65.4	82.2
MT200K	77.4	43.7	72.2	62.0	82.0
MT300K	78.5	42.5	73.7	65.1	85.2
MT400K	79.8	43.7	73.3	65.8	85.7
MT500K	**85.2**	43.7	**75.1**	**67.5**	**86.6**

注:黑体值表示最优值.

3.4 不同数据集之间的泛化和迁移趋势是否一致

在这个实验中,我们分析不同数据集之间的迁移是否和泛化具有相同的趋势.如果趋势一致,我们就可以直接选择泛化能力强的数据集,用来提升目标领域数据集的性能.具体地,我们在5个大数据集上进行预训练,然后在其他所有数据集上进行迁移测试,并分析迁移后的性能与之前分析的泛化性能

的一致性.同泛化实验一样,我们保证训练流程一致,只微调部分参数,尤其 BERT 这种对小数据集敏感的模型,我们实验多次取平均值.

实验结果如表 4 所示,我们发现了数据集之间泛化和迁移的趋势往往是不一致的.从源数据到目标数据的组合共有 45 种,而 ESIM 和 BERT 模型泛化和迁移趋势一致的分别只有 3 个和 2 个.虽然泛化实验计算量很小,但是我们不能直接从泛化实验的趋势推广到迁移实验.另外,从实验结果看,BERT 模型比 ESIM 模型的迁移效果要高很多.因此,预训练模型的泛化能力要比传统的深度学习模型强,并且迁移能够带来比泛化更大的性能提升.

Table 4 Results of Transfer Experiment

表 4 迁移实验结果 %

模型	数据集	SciTail	WNLI	RTE	MRPC	WikiQA	SNLI	MNLI	QNLI	Ms marco	QQP
ESIM	SNLI	87.5	**59.2**	63.9	76.4	89.7	—	**72.0**	82.9	88.5	84.3
	MNLI	**88.6**	46.5	66.1	**79.5**	89.7	83.1	—	82.9	88.5	85.1
	QNLI	88.0	56.3	65.3	77.8	89.5	**83.7**	70.8	—	**89.3**	**85.2**
	Ms marco	87.7	50.7	**66.8**	77.9	**90.0**	83.0	70.8	**83.0**	—	85.1
	QQP	85.2	54.9	62.1	75.7	89.6	81.6	68.8	82.2	88.3	—
	MT100K	87.9	56.3	64.6	76.6	89.8	—	—	—	—	—
	SELF	82.6	57.8	57.0	69.0	88.0	82.7	70.4	82.5	87.9	84.3
BERT	SNLI	95.3	56.3	**74.4**	85.7	91.8	—	81.6	90.0	92.7	88.1
	MNLI	95.3	56.3	72.9	**86.1**	91.1	**88.3**	—	90.0	**93.1**	87.8
	QNLI	95.7	56.3	72.9	85.2	92.2	87.9	81.5	—	92.9	88.2
	Ms marco	94.9	56.3	70.8	85.3	**92.4**	87.8	**81.7**	**90.9**	—	**88.3**
	QQP	**95.8**	**62.0**	72.9	85.3	90.9	87.6	81.2	90.9	92.8	—
	MT100K	95.5	56.3	75.8	84.8	92.4	—	—	—	—	—
	SELF	94.9	56.33	67.5	85.2	91.4	87.7	81.0	91.3	92.7	88.3

注:黑体值为此数据集泛化最好的结果;"—"代表这些值不可用.

3.5 迁移能够带来多少性能的提升

我们在此实验中,进一步分析经过在源数据集预训练之后的模型,迁移到目标数据集上能够带来多少性能提升,尤其是预训练语言模型 BERT,是否能在目标数据集很大的情况下,还能有性能上的提升.

从实验结果表 4 中,我们可以看到 ESIM 模型在 5 个大数据集上迁移的平均性能比在源数据集本身进行训练和测试(SELF 行)提升 2.2 个百分点,如果选取每个数据集上泛化最好的性能,则提升 3.5 个百分点;对于 BERT 模型,在 5 个小数据集上如果选取泛化最好的性能,则提升 3 个百分点,在 5 个大数据集上如果选取每个数据集上泛化最好的性能,仍能提升 0.3 个百分点.所以,即使是 BERT 这种预训练模型,合适的迁移仍然能带来性能的提升.

3.6 目标领域数据量的大小对性能的影响

我们在此实验中分析,在目标领域中不同数据量的情况下对应的迁移性能的变化,观察模型能否在少样本的情况下取得不错的效果.具体来说,我们首先使用在 5 个大的目标数据集上迁移能力最好的大的源数据集上进行预训练,然后不断增加目标数据集的数据量,比如对于 ESIM 模型和 SNLI 数据集,我们先在 QNLI 上预训练,然后不断增加 SNLI 的数据量,测试其性能.由于 MT100K 良好的泛化能力和迁移能力,我们也将其加入到了 BERT 模型下的实验.实验结果如图 2 所示,图 2(a)为 ESIM 模型下的实验结果,图 2(b)为 BERT 模型下的实验结果.图 2 中正方形标记的曲线代表只在目标数据集上的训练和测试;三角形标记曲线代表首先在迁移效果最好的数据集上预训练,然后在目标数据集上训练;圆形标记曲线代表先在 MT100K 上训练,然后在目标数据集上训练和测试.实验结果表明,在迁移效果最好的数据集上进行预训练,不管对于 ESIM 模型还是 BERT 模型,在几乎所有实验中都要比只在数据集本身训练和测试要高,证明了合适的迁移会给数据集带来性能上的提升.根据统计,我们发现在源数据集上进行预训练之后,仅需 37%的目标数据集量,就能达到目标数据集 95%的性能.

此外,由图 2(b)所示,在 MT100K 上进行预训练的结果曲线比在迁移效果最好的源数据集训练的曲线更加平滑,性能也有所提升,进一步验证了多个数据集的融合能提高模型的迁移能力.

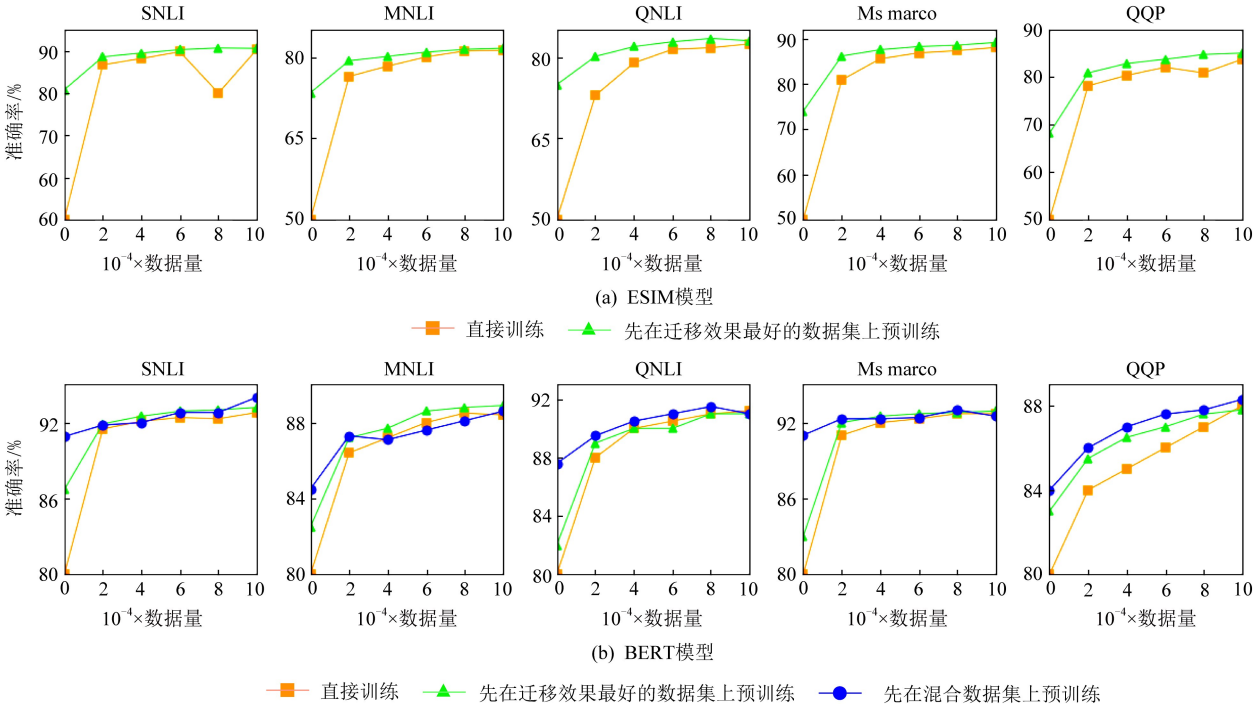


Fig. 2 Learning curves of five large datasets

图 2 5 个大数据集的学习曲线

3.7 混合后的数据集在新领域和小样本情况下的性能

基于对不同领域数据集之间泛化性和迁移性的分析,我们发现预训练模型 BERT 在平均混合的数据集上预训练之后,有着更好的泛化能力和迁移能力.为此,我们探索在少量样本的情况下,将 MT100K 应用到新领域的表现.我们选择 4 个小数据集 SciTail,WNLI,RTE,MRPC 进行实验,在这 4 个小数据中随机取 100 个样例.我们先使用 BERT 模型在 MT100K 上进行预训练,然后再在这 4 个小数据集上 100 个样例进行微调和评价.

实验结果如表 5 所示,其中 MT100K 行是在 MT100K 数据上预训练之后,再在 4 个小数据集上 100 个样本进行微调,最后进行测试;SELF 行代表在小数据集上的全量数据进行训练和测试.我们发现在少量样本(只有 100 个)情况下,MT100K 的平均性能只比全量的小数据集低 4.1 个百分点.因此在短文本匹配领域,对于新领域的数据集,我们可以首先将其他领域的数据集等量混合,然后使用 BERT 模型在混合的数据集预训练之后,再进行少量的目标样本标注和训练,就可以达到不错的效果.

Table 5 Experimental Results of MT100K on 100 Samples

表 5 MT100K 在 100 个样本上的实验结果 %

数据集	SciTail	WNLI	RTE	MRPC
MT100K	87.3	56.3	71.8	72.0
SELF	94.9	56.3	67.5	85.2

4 相关工作

深度学习方法尤其是预训练方法在短文本匹配问题取得了目前最好的效果,但是这些模型在某个数据集训练之后很难泛化到其他的数据集,尤其因为短文本匹配包含多种匹配任务[1-4],这种任务之间的差距使得泛化更加困难.Liu 等人[27]结合了多任务学习和预训练方法,训练之后可以在多个数据集上取得较好的效果,并且证明了经过多任务学习,模型的领域迁移能力有很大的提升.Raffel 等人[28]将文本匹配问题和阅读理解等其他问题全部抽象成“text-to-text”问题,通过超大规模的无标注数据的预训练,提升模型的泛化能力.Yogatama 等人[29]尝试通过在多个数据集上预训练得到更加通用的表达.本文主要受 Talmor[30]的工作启发,它们从机器

阅读的领域出发,研究了不同的机器阅读理解数据集的泛化性和迁移性,我们尝试将其推广到短文本匹配领域.

5 总 结

本文在10个数据集上,使用2种深度学习模型对短文本匹配的泛化性和迁移性进行了详尽的分析实验.实验结果表明:1)深度学习模型的泛化能力很差,会在训练过的数据集发生过拟合,即使是BERT这种预训练模型;2)影响泛化能力的因素主要有匹配的类型和匹配数据集的来源;3)实验结果显示不同数据集之间泛化性和迁移性趋势通常情况下并不保持一致;4)通过迁移,模型能在小数据集上有较大的提升,即使是BERT这种预训练模型,合适的迁移也能在目标数据集带来提升;5)将数据集进行简单的混合能带来更好的泛化能力和迁移能力,在将其应用到新的领域和少量样本的情况下,模型取得的效果非常好.总结来看,本文的分析工作对未来利用现有数据集迁移到新的领域,并且在减少目标数据标注量和提升性能上提供了有价值的指导.

作者贡献声明:马新宇负责所有的实验及分析,以及文章的撰写;范意兴参与了论文想法的讨论、论文逻辑的梳理,还设计各个实验和修改论文等;郭嘉丰参与了论文想法的讨论、论文摘要和前言的修改,以及部分实验的分析;张儒清参与了论文想法的讨论、论文逻辑的梳理和实验的设计,修改了论文的实验部分;苏立新是论文想法的提出者,并在实验过程中给予了大量的指导;程学旗参与了论文想法的讨论,确定了论文的分析思路,提供了服务器进行大量的实验.

参 考 文 献

[1] MacCartney B, Galley M, Christopher D M. A phrase-based alignment model for natural language inference [C] //Proc of the 2008 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2008: 802-811

[2] Dolan W B, Brockett C. Automatically constructing a corpus of sentential paraphrases [C] //Proc of the 3rd Int Workshop on Paraphrasing (IWP2005). Stroudsburg, PA: ACL, 2005: 9-16

[3] Yang Yi, Yih W, Meek C. Wikiqa: A challenge dataset for open-domain question answering [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 2013-2018

[4] Brill E, Dumais S, Banko M. An analysis of the AskMSR question-answering system [C] //Proc of the 2002 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2002: 257-264

[5] Huang Jizhou, Zhou Ming, Yang Dan. Extracting chatbot knowledge from online discussion forums [C] //Proc of the 2002 Int Joint Conf on Artificial Intelligence. Amsterdam: Elsevier, 2007: 423-428

[6] Li Baoli, Liu Yandong, Ram A, et al. Exploring question subjectivity prediction in community QA [C] //Proc of the 31st Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2008: 735-736

[7] Chen Danqi, Fisch A, Weston J, et al. Reading Wikipedia to answer open-domain questions [C] //Proc of the 55th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2017: 1870-1879

[8] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C] //Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171-4186

[9] Wang A, Singh A, Michael J, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding [C] //Proc of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Stroudsburg, PA: ACL, 2018: 353-355

[10] Rajpurkar P, Zhang Jian, Lopyrev K, et al. SQuAD: 100000+ questions for machine comprehension of text [C] //Proc of the 2016 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 2383-2392

[11] Bajaj P, Campos D, Craswell N, et al. Ms marco: A human generated machine reading comprehension dataset [J]. arXiv preprint, arXiv:1611.09268, 2016

[12] Chen Qian, Zhu Xiaodan, Ling Zhenhua, et al. Enhanced LSTM for natural language inference [C] //Proc of the 55th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2017: 1657-1668

[13] Bowman S, Angeli G, Potts C, et al. A large annotated corpus for learning natural language inference [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 632-642

[14] Williams A, Nangia N, Bowman S. A broad-coverage challenge corpus for sentence understanding through inference [C] //Proc of the 2018 Conf of the North American Chapter of the ACL: Human Language Technologies. Stroudsburg, PA: ACL, 2018: 1112-1122

[15] Wang Zhiguo, Hamza W, Florian R. Bilateral multi-perspective matching for natural language sentences [C] //Proc of the 26th Int Joint Conf on Artificial Intelligence. Amsterdam: Elsevier, 2017: 4144-4150

[16] Levesque H, Davis E, Morgenstern L. The winograd schema challenge [C] //Proc of the 13th Int Conf on Principles of Knowledge Representation and Reasoning. New York: ACM, 2012: 552-561

[17] Giampiccolo D, Magnini B, Dagan I, et al. The third pascal recognizing textual entailment challenge [C] //Proc of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Stroudsburg, PA: ACL, 2007: 1-9

[18] Khot T, Sabharwal A, Clark P. SciTail: A textual entailment dataset from science question answering [C] // Proc of the AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 41-42

[19] Yang Y, Yih W, Meek C. Wikiqa: A challenge dataset for open-domain question answering [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 2013-2018

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

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] //Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 5998-6008

[22] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation [C] //Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1532-1543

[23] Ba J L, Kiros J R, Hinton G E. Layer normalization [J]. arXiv preprint, arXiv:1607.06450, 2016

[24] Wu Yonghui, Schuster M, Chen Zhifeng, et al. Google's neural machine translation system: Bridging the gap between human and machine translation [J]. arXiv preprint, arXiv: 1609.08144, 2016

[25] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint, arXiv:1412.6980, 2014

[26] Fruchterman T M J, Reingold E M. Graph drawing by force-directed placement [J]. Software: Practice and Experience, 1991, 21(11): 1129-1164

[27] Liu Xiaodong, He Pengcheng, Chen Weizhu, et al. Multi-task deep neural networks for natural language understanding [C] //Proc of the 57th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2019: 4487-4496

[28] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. arXiv preprint, arXiv:1910.10683, 2019

[29] Yogatama D, d'Autume C M, Connor J, et al. Learning and evaluating general linguistic intelligence [J]. arXiv preprint, arXiv:1901.11373, 2019

[30] Talmor A, Berant J. Multi Q A: An empirical investigation of generalization and transfer in reading comprehension [C] // Proc of the 57th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2019: 4911-4921

Ma Xinyu, born in 1995. PhD candidate. His main research interests include information retrieval and natural language understanding.

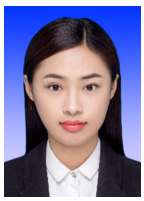
马新宇,1995 年生.博士研究生.主要研究方向为信息检索和自然语言理解.

Fan Yixing, born in 1990. Assistant professor. His main research interests include data mining and information retrieval.

范意兴,1990 年生.助理研究员.主要研究方向为数据挖掘、信息检索.

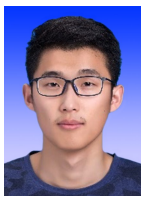
Guo Jiafeng, born in 1980. Professor. His main research interests include data mining and information retrieval. (guojiafeng@ict.ac.cn)

郭嘉丰,1980 年生.研究员.主要研究方向为数据挖掘、信息检索.



Zhang Ruqing, born in 1994. Assistant professor. Her main research interest is natural language processing. (zhangruqing@ict.ac.cn)

张儒清,1994 年生.助理研究员.主要研究方向为自然语言处理.



Su Lixin, born in 1992. PhD. His main research interests include information retrieval and question answering. (sulixin17b@ict.ac.cn)

苏立新,1992 年生.博士.主要研究方向为信息检索和问答.

Cheng Xueqi, born in 1971. Professor. His main research include social computing, information retrieval and data mining. (cxq@ict.ac.cn)

程学旗,1971 年生.研究员.主要研究方向为社会计算、信息检索与数据挖掘.