史鑫陶, 韩韧. 基于改进 LM-BFF 的小样本文本分类[J]. 智能计算机与应用,2025,15(10):169-174. DOI:10.20169/j.issn.2095-2163.251026

基于改进 LM-BFF 的小样本文本分类

史鑫陶, 韩 韧
(上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘 要: LM-BFF 模型通过在文本中添加自然语言提示,实现了显著的小样本性能。然而模型存在过拟合预训练任务和数据的风险,导致与目标下游任务存在差距影响模型性能。为了克服模型对下游任务的差距,对其进行了改进,主要包括:采用 Noisy-tuning 通过添加矩阵扰动来更好地微调预训练语言模型;进行 2 次前向传播,通过最小化 2 次前向传播之间的差异来减少训练和推理的不一致性。实验结果表明,在多个分类数据集上性能均有提升。特别在 SST-2 和 SNLI 数据集上,准确率分别提升了 0.7%和 1.7%。

关键词: 自然语言提示; 小样本; Noisy-tuning; 预训练语言模型
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2025)10-0169-06

Small sample text classification based on improved LM-BFF

SHI Xintao, HAN Ren
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: The LM-BFF model achieves significant small-sample performance by adding natural language prompts to the text. However, there is a risk that the model will overfit the pre-training tasks and data, the performance of the model is affected. In order to overcome the gap between the model and the target downstream task, this paper introduces some improvements for LM-BFF: Noisy-tuning is employed to help better fine-tune pre-trained language models by adding matrix perturbations; R-drop is used to reduce inconsistencies by minimize the discrepancy between two consecutive forward passes during training and inference. Experimental results indicate that the proposed model shows improvements over LM-BFF across multiple classification datasets. Especially on SST-2 and SNLI dataset, accuracy is improved by 0.7% and 1.7%, respectively.

Key words: natural language prompts; few-shot; Noisy-tuning; pre-trained language model

0 引 言

文本分类是自然处理领域中的一个重要问题。随着 GPT-3[1]模型的出现,以及该模型在分类任务上展现出的优秀的小样本性能,在自然语言处理领域中得到了广泛重视与应用。通过给定一个自然语言提示和任务的一些演示,GPT-3 能够在不更新其底层语言的任何权重的情况下做出准确的预测。值得注意的是,GPT-3 由 175 B 参数组成,这使得在大多数实际项目中的使用颇具挑战性。因此本次研究在一个更实际的场景中研究小样本学习,在这个场景中使用较小的语言模型,例如 RoBERTa 等、以及少样本设置,本文中将用其来微调语言模型的权重[2]。为了实现特定任务的微调,基于提示的小样本微调已被广泛研究。基于提示的微调将下游任务重新表述为掩码语言建模问题,其中使用特定任务的模板在给定的提示上生成一个标签词。然而,构建最佳的提示需要专业领域的知识,手动提示可能是次优的[3]。

在基于提示微调的各种方法中,本文研究基于 LM-BFF[2]。该方法主要采用演示作为附加上下文的想法。其中,演示是通过在样本中选取相似的上下文与输入拼接而成,通过演示学习和提示来对小样本进行微调。然而由于小样本数据的有限性,导致预训练模型会过度拟合下游有限的标记数据。

基于之前的 LM-BFF 的设置,在多个 NLP 数据

作者简介: 史鑫陶(1998—),男,硕士研究生,主要研究方向:自然语言处理。
通信作者: 韩 韧(1980—),男,博士,副教授,主要研究方向:智能计算,边缘计算。Email: ren.han@usst.edu.cn。

收稿日期: 2023-12-29

集上的实验结果表明,本文所提出的模型可以带来更好的、更稳定的小样本性能[3]。本文的研究贡献总结如下:

(1)采用 Noisy-tuning、一种矩阵扰动方法。根据 PLM 的方差将具有不同强度的噪声添加到 PLM 中不同类型的参数矩阵中,从而有助于更好地来微调 PLM[4]。

(2)将同一数据样本前向传播 2 次、即 Dropout 两次[5],得到 2 个差异很小的概率分布,通过在原来的交叉熵损失中加入这 2 个分布的对称 Kullback-Leibler(KL)散度损失来进行参数更新。

1 基于提示的微调

1.1 问题设置

在提示设置下基于小样本进行微调,已经被广泛研究用于中等大小的预训练语言模型。如 BRET[6] 和 RoBERTa[7]。例如,PET[8] 将下游任务重新制定为掩码语言建模问题并进行基于梯度的优化。AutoPrompts[9] 使用梯度引导搜索作为离散的标记来创建提示。还有一些工作是使用提示标记的连续向量,称为软提示。包括 Lester 等学者[10] 提出的由可学习的连续嵌入组成的软提示,同时还冻结了预训练模型的权重。Gu 等学者提出了预训练提示:通过在预训练阶段添加软提示来获得更好的初始化。Gao 等学者[2] 也探索了将演示结合的 LM-BFF,其中的演示是通过在相似输入示例上取消掩码提示来构建的。

假设访问预训练模型 L, 希望对带有标签空间 Y 的任务 D 进行微调。对于任务的训练集 D_{train}, 假设每类有 K 个训练示例,使得示例总数为:

$$K_{\text{tot}} = K \times |Y|, D_{\text{train}} = \{(x_{\text{in}}^i, y^i)\}_{i=1}^{K_{\text{tot}}} \quad (1)$$

对于模型的选择和参数的调整,假设开发集 D_{dev} 的大小与小样本训练集相同,即 $|D_{\text{dev}}| = |D_{\text{train}}|$。研究中取 L = RoBERTa-large, K = 16。

NLP 中的小样本学习范式主要包括:

(1)半监督学习[11]。给出了一组未标记的示例。

(2)元学习[12]。给出了一组辅助任务。

(3)中间训练[13]。给出了相关的中间任务。

通过对可用资源做出最小假设来偏离这些设置:研究中只假设一些带注释的示例和预训练的语言模型。对 4 个单句和 2 个句子对英语任务进行了系统研究,包括来自 GLUE 基准[14] 测试的任务、SNLI[15] 和其他几个流行的句子分类任务(MR、CR、Subj)。对于单句任务,目标是根据输入句子 $x_{\text{in}} = x_1$ 进行预测,例如预测电影评论是否为正。对于句子对任务,目标是取一对输入句子 $x_{\text{in}} = (x_1, x_2)$ 并预测两者间的关系,使其真正接近小样本设置。众所周知,对小数据集进行微调可能会受到不稳定性的影响,并且在给定新的数据拆分的情况下,结果可能会发生巨大的变化。为了解决这个问题,从数据集中随机采样 5 个 D_{train} 和 D_{dev} 并测量各自的平均性能,对数据集进行采样可以更稳健地衡量性能,并更好地估计方差。主要在 SST-2 和 SNLI 进行试点实验。

1.2 分类问题

给定一个掩码语言模型 L, 首先将输入 x_{in} 转换为 token(词例)序列 $\tilde{x}$, 语言模型 L 则将 $\tilde{x}$ 映射到隐藏向量序列 $\{h_k \in R^d\}$。在标准微调过程中,通常取 $\tilde{x}_{\text{single}} = [\text{CLS}]\ x_1\ [\text{SEP}]$, $\tilde{x}_{\text{pair}} = [\text{CLS}]\ x_1\ [\text{SEP}]\ x_2\ [\text{SEP}]$。对于下游具有标签空间 Y 的分类任务,通过最大化正确标签的对数概率来训练特定任务的头部 Softmax ($\boldsymbol{W}_o \boldsymbol{h}_{[\text{CLS}]}$), 其中 $\boldsymbol{h}_{[\text{CLS}]}$ 是[CLS]的隐藏向量, $\boldsymbol{W}_o \in R^{|Y| \times d}$ 是在微调开始时引入的一组随机初始化参数。类似地,对于回归任务,引入 $w_o \in R^d$ 并优化 $w_o \cdot \boldsymbol{h}_{[\text{CLS}]}$ 和黄金标签之间的均方误差。任何一种情况下,引入的新参数的数量可能很大。例如,一个简单的二分类任务将为 RoBERTa-large 模型引入 2 048 个新参数。这就使得从少量标注数据(例如 32 个示例)中学习具有一定的挑战性。

解决这个问题的一种方法是基于提示的微调,其中语言模型 L 负责完成自然语言提示。例如,可以使用输入 x_1(例如,"A total waste of time")来完成二元情感分类任务为:

$$x_{\text{prompt}} = [\text{CLS}]\ x_1\ \text{It was}\ [\text{MASK}] \cdot [\text{SEP}] \quad (2)$$

接着,让语言模型 L 决定更适合填充[MASK]的是"great"(正)、还是"terrible"(负)。

对于分类任务,首先定义 $M: y \to v$ 是任务标签空间到词汇表中单个单词的映射。然后对于每个 x_{in}, 假设 $x_{\text{prompt}} = T(x_{\text{in}})$ 是一个掩码语言建模(MLM)输入,其中包含一个[MASK]标记。通过这种方式,可以将研究任务视为 MLM,并将预测类别 $y \in Y$ 的概率建模为:

$$p(y \mid x_{\text{in}}) = p([\text{MASK}] = M(y) \mid x_{\text{prompt}}) = \frac{\exp(w_{M(y)} \cdot h_{[\text{MASK}]})}{\sum_{y' \in Y} \exp(w_{M(y')} \cdot h_{[\text{MASK}]})} \quad (3)$$

其中，$h_{[\mathrm{MASK}]}$ 表示[MASK]的隐藏向量，w_v 表示对应于 $v \in V$ 的 pre-Softmax 向量。当监督示例 $\{(x_{\mathrm{in}}, y)\}$ 可用时，可以 L 对进行微调来最小化交叉熵损失。这种方法重用了预训练的权重并且不引入任何新参数，还减少了预训练与微调之间的差距，使其在少样本场景中更有效。

1.3　自动生成模板

本次研究将模板和标签词两者统称为提示。之前是手工设计模板和标签词。这通常需要领域的专业知识和一定的试错实验。表 1 总结了本次研究为实验中的数据集设计的手动模板和标签词。这些模板和标签词是通过直觉设计的，并且考虑了之前相关文献中使用的格式。

不同的提示可以导致准确性出现巨大的差异。具体说，当模板固定时，标签词与“语义类别”匹配程度越高，准确率也会越高。使用的相同的标签词，即使模板出现了微小的变化也可能会造成巨大的差异。因此设计好的模板和标签词是极为重要的。

表 1　手动模板和标签词

Table 1　Manual templates and tag words

数据集	模板	标签词
SST-2	$<S_1>$ It was [MASK] .	positive: great, negative: terrible
MR	$<S_1>$ It was [MASK] .	positive: great, negative: terrible
CR	$<S_1>$ It was [MASK] .	positive: great, negative: terrible
Subj	$<S_1>$ This is [MASK] .	subjective: subjective, objective: objective
MNLI	$<S_1>$? [MASK], $<S_2>$	entailment:Yes,netural: Maybe, contradiction: No
SNLI	$<S_1>$? [MASK], $<S_2>$	entailment:Yes,netural: Maybe, contradiction: No

2　实验方法

2.1　Noisy-tuning

研究中采用 Noisy-tuning，与直接对下游任务数据上的 PLM 进行微调的标准微调不同，Noisy-tuning 是在微调前添加少量的噪声来扰动 PLM 参数。在 Noisy-tuning 中是使用矩阵扰动方法，并根据不同参数矩阵的标准差将不同强度的均匀噪声添加到不同的参数矩阵中[4]。

将 PLM 中的参数矩阵表示为 $[\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_N]$，其中 N 是参数矩阵类型的数量。将参数矩阵 $\boldsymbol{W}_i$ 的扰动表示为 $\widetilde{\boldsymbol{W}}_i$，计算如下：

$$\widetilde{\boldsymbol{W}}_i = \boldsymbol{W}_i + U\left(-\frac{\lambda}{2}, \frac{\lambda}{2}\right) \times \mathrm{std}(\boldsymbol{W}_i) \tag{4}$$

其中，std 表示标准偏差；函数 $U(a,b)$ 表示均匀分布噪声范围为 a 到 b；λ 表示控制相对噪声强度的超参数。在具有较高方差的 PLM 中，Noisy-tuning 参数将随着噪声增强而有所增加。此外，在 PLM 中对于一些常量矩阵，不会被扰动，因其标准差是 0，可以确保常数矩阵不会被额外的噪声意外激活。

2.2　R-drop

对于小样本数据存在的过拟合现象，引入 dropout 并将数据进行 2 次前向传播，从中选取 2 次子模型的输出，接着通过最小化 2 次输出的 Kullback-Leibler(KL)散度使 2 个子模型的输出达到一致，从而减少训练和推理之间的不一致性[5]。具体流程如图 1 所示。

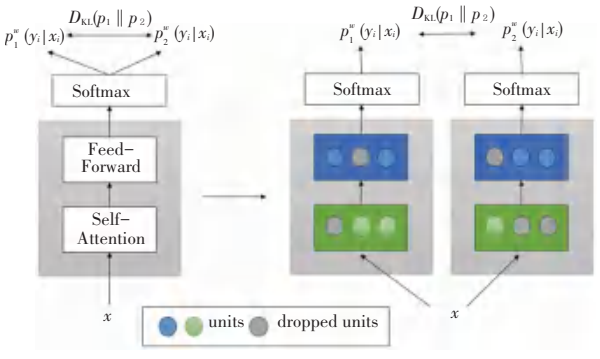

图 1　2 次前向传播的总体框架

Fig. 1　The overall framework of the two forward propagations

给定训练集 $D_{\mathrm{train}} = \{(x_i, y_i)\}_{i=1}^{n}$，训练的目标是学习一个模型。其中，$n$ 表示训练样本的数量，(x_i, y_i) 表示标记数据对。这里，x_i 表示输入数据，y_i 表示标签。映射函数的概率分布记为 $P^w(y \mid x)$。2 个分布 P_1 和 P_2 之间 KL 散度用 $D_{\mathrm{KL}}(P_1 \| P_2)$ 表示。参见图 1 右侧，对于相同的数据对 (x_i, y_i)，$P_1^w(y_i \mid x_i)$ 和 $P_2^w(y_i \mid x_i)$ 的分布是不同的，通过最小化这 2 个输出分布之间的双向 Kullback-Leibler(KL)散度来对模型预测进行正则化。

3　实验结果及分析

在本节中，本文在 6 个真实的数据集下进行试

验,并选取 SST-2 和 SNLI 作为试点实验,展示了实验细节。研究中,使用精确度作为评估指标来验证模型性能。

3.1 数据集

(1)MR[16]:是一个二分类的电影评论数据集,包括 10 662 个样本,分别为 5 331 个正面样本和 5 331 个负面样本。

(2)SST-2[17]:是对 MR 的一个扩充,是一个二分类数据集,包括训练集、开发集和测试集。大小分别为:6 920、872 和 1 821。

(3)斯坦福自然语言推理(SNLI)数据集:被广泛应用于自然语言推理(NLI)。该数据集由 550 152、10 000 和 10 000 个句子对组成,分别用于训练、开发和测试。每对使用 3 个标签之一进行注释:中立、包含和矛盾。

(4)多体裁自然语言推理(MNLI)[18]数据集:是一个 433 k 句子对的集合,这些句子对带有文本包含标签。语料库是 SNLI 的扩展,涵盖了广泛的口语和书面语体裁。

(5)Customer Reviews(CR)[19]:是各种产品的客户评论,任务是预测正面/负面评论。

(6)Subjectivity 主观性数据集(Subj)[20]:任务是将句子分类为主观或客观的。

3.2 实验设置

3.2.1 超参数选择

本文研究中使用了 2.1 节中设置的模型 L = RoBERTa-large,并且从数据集的训练集中每个类取 K 个训练示例组成训练集 D_{train},对于开发集和测试集采用同样的策略,这里采用 K = 16,并随机选取示例组成 5 个训练集、开发集以及测试集。对于网格搜索,从{1e-5,2e-5,5e-5}的学习率和{2,4,8}的批量大小中获取最佳的设置。对 SST-2 数据集进行试点实验,同时还使用提前停止来避免过拟合。在每次试验中,对模型进行了 1 000 步的训练,每 100 步验证性能,并采用最佳检查点来保存训练后模型的权重。

3.2.2 实验结果

研究中比较了 3 种方法,包括:没有任何噪声的 Noisy-tuning、具有矩阵高斯噪声的 Noisy-tuning 和具有矩阵均匀噪声的 Noisy-tuning。在 SST-2 和 SNLI 上进行试点,实验结果如图 2 所示。此外分析发现,添加均匀噪声优于高斯噪声,这可能是因为高斯噪声的范围更广,一些极值可能会影响模型性能。因此,在 Noisy-tuning 中使用矩阵均匀噪声。

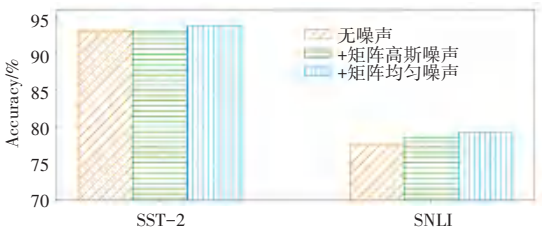

图 2 不同噪声类型的影响

Fig. 2 Effect of different noise types

进一步研究了 Noisy-tuning 中最重要的超参数、即 λ 的影响,这是控制相对噪声强度的。从 $\lambda \in \{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$ 进行实验,结果如图 3、图 4 所示。主要在 SST-2 和 SNLI 上进行试点实验。

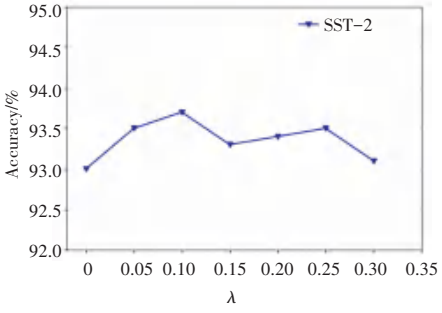

图 3 噪声强度 λ 的影响

Fig. 3 Effect of noise intensity λ

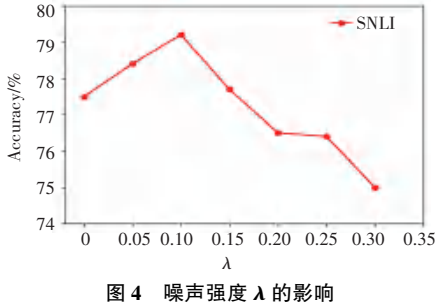

图 4 噪声强度 λ 的影响

Fig. 4 Effect of noise intensity λ

本文发现,当 λ 太大或太小时,性能都不是最优的。这是因为 λ 太大时,PLM 中有用的预训练知识可能会被随机噪声淹没。λ 太小时,PLM 很难进行参数空间探索,从而克服过拟合问题。所以选取 $\lambda = 0.1$ 作为后续实验的超参数设置。接下来,本文比较了在 SST-2 和 SNLI 上使用和不使用 Noisy-tuning 进行提示微调的准确度对比。对于 SST-2 和 SNLI 数据集,测量了 5 个不同的随机采样 D_{train} 和 D_{dev} 拆分的性能。结果如图 5、图 6 所示。分析可知,本文模型的性能是优于基线模型的,添加的 Noisy-tuning 使得扰动的 PLM 具有较低的过拟合预训练任务的风险,因此具有更好的泛化能力。

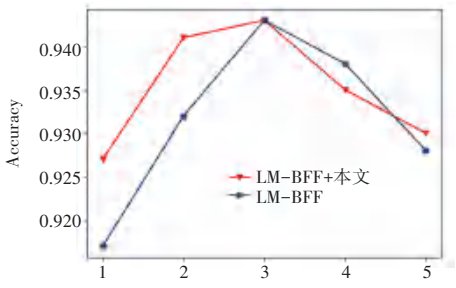

图 5　SST-2 上模型的性能对比

Fig. 5　Comparison of the performance of the models on SST-2

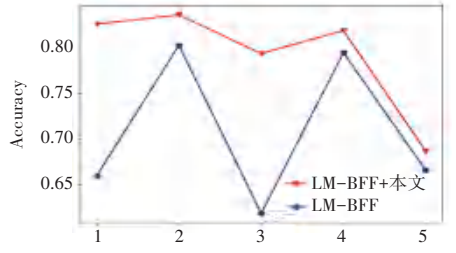

图 6　SNLI 上模型的性能对比

Fig. 6　Comparison of the performance of the models on SNLI

在其他几个数据集上进行实验，实验主要结果见表 2。

3.3　结果分析

本文实验主要是基于 RoBERTa-large，FT 表示标准微调(Fine-tuning)，并设置 $K = 16$，auto 表示模板是自动生成的。主要比较了 2 种基线：小样本设置中的标准微调和基于提示+演示的小样本微调。表 2 显示了使用单个模板的实验结果。首先，采用基于提示和演示的微调在小样本数据集上的效果明显优于标准微调；其次，采用 Noisy-tuning 在多个任务上是优于基线的，进一步表明本文模型在基于提示的微调上具有更好的性能，对于微调下游任务具有更好的泛化性。

3.4　消融实验

为了论证 Noisy-tuning 的有效性，在本文的模型基础上去掉了 Noisy-tuning 进行相关的实验，即先对 LM-BFF 添加 R-Drop 进行实验，然后再添加矩阵均匀噪声做对比实验。实验结果见表 3。从表 3 中可以看出在 R-Drop 的基础上添加 Noisy-tuning 可以有效地提高模型的精准度。

表 2　RoBERTa-large 的实验结果

Table 2　Experimental results in RoBERTa-large

模型	SST-2(acc)	MR(acc)	CR(acc)	MNLI(acc)	SNLI(acc)	Subj(acc)
Fine-tuning	81.4(3.8)	76.9(5.9)	75.8(3.2)	45.8(6.4)	48.4(4.8)	90.8(1.8)
LM-BFF(auto)	93.0(0.6)	87.7(1.4)	91.0(0.9)	70.0(3.6)	77.5(3.5)	91.4(1.8)
LM-BFF(auto)+本文	93.7(0.6)	89.3(0.7)	91.6(0.6)	70.6(2.9)	79.2(4.2)	91.5(1.7)

表 3　不同模型对比

Table 3　Comparison of different models

模型	SST-2(acc)	MR(acc)	CR(acc)	MNLI(acc)	SNLI(acc)	Subj(acc)
LM-BFF	93.0(0.6)	87.7(1.4)	91.0(0.9)	70.0(3.6)	77.5(3.5)	91.4(1.8)
LM-BFF+RD	93.5(0.5)	89.1(1.1)	91.7(0.7)	69.7(2.2)	78.8(4.2)	90.6(0.8)
LM-BFF+RD+Noisy+tuning	93.7(0.6)	89.3(0.7)	91.6(0.6)	70.6(2.9)	79.2(4.2)	91.7(1.7)

4　结束语

本文主要采用 Noisy-tuning 通过在 PLM 的参数矩阵上添加噪声扰动以及结合 R-drop 来进行更好的基于提示的微调。结果表明在多个任务上优于基线：SST-2、MR、CR、SNLI、MNLI、Subj。但是仍然存在不足之处：目前的分类是基于相对简单的数据集；都是基于短文本来进行分类的。未来工作的研究重点主要包括：

(1)对文档级别进行应用分析是否有效。

(2)选取相对复杂的数据集分析噪声扰动对其的影响。

(3)大多数自动生成的模板是合理、且语法正确的，对于标签词在某些情况下可能违反直觉，目前尚不清楚为什么语言模型选择了这些词，后期拟将这些内容作为研究课题展开讨论。

参考文献

[1] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[2] GAO Tianyu, FISCH A, CHEN Danqi. Making pre-trained language models better few-shot learners[J]. arXiv preprint arXiv, 2012.15723, 2020.

[3] PARK E, JEON D, KIM S, et al. LM-BFF-MS: Improving

few-shot fine-tuning of language models based on multiple soft demonstration memory [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). ACL,2022: 310-317.

[4] WU Chuhan, WU Fangzhao, QI Tao, et al. Noisytune: A little noise can help you finetune pretrained language models better[J]. arXiv preprint arXiv,2202.12024,2022.

[5] LIANG Xiaobo, WU Lijun, LI Juntao, et al. R-drop: Regularized dropout for neural networks[J]. Advances in Neural Information Processing Systems, 2021,34: 10890-10905.

[6] DEVLIN J, CHANG Mingwei, LEE K,et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv,1810.04805,2018.

[7] LIU Yinyhan, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv, 1907.11692,2019.

[8] SCHICK T, SCHÜTZE H. Exploiting cloze questions for few shot text classification and natural language inference [J]. arXiv preprint arXiv,2001.07676,2020.

[9] SHIN T, RAZEGHI Y, ROBERT L L I V, et al. Autoprompt: Eliciting knowledge from language models with automatically generated prompts[J]. arXiv preprint arXiv,2010.15980,2020.

[10] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv, 2104.08691,2021.

[11] XIE Qizhe, DAI Zihang, HOVY E, et al. Unsupervised data augmentation for consistency training[J]. arXiv preprint arXiv, 1904.12848,2020.

[12] BAO Yujia, WU Menghua, CHANG Shiyu, et al. Few-shot text classification with distributional signatures [J]. arXiv preprint arXiv, 1908.06039,2020.

[13] PHANG J, FÉVRY T, BOWMAN S R. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks [J]. arXiv preprint arXiv,1811.01088, 2018.

[14] WANG A, SINGH A, MICHAEL J,et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding[J]. arXiv preprint arXiv,1804.07461,2018.

[15] BOWMA N, SAMUEL R, ANGELI G, et al. A large annotated corpus for learning natural language inference[C]// Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. ACL,2015:632-642.

[16] PANG Bo, LEE L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales [C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL,2005:115-124.

[17] SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C]// Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. ACL,2013:1631-1642.

[18] WILLIAMS A, NANGIA N, BOWMAN S R. A broad-coverage challenge corpus for sentence understanding through inference[J]. arXiv preprint arXiv,1704.05426,2017.

[19] HU Mingqing, LIU Bing. Mining and summarizing customer reviews [C]// Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM, 2004:168-177.

[20] PANG B, LEE L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts [C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. ACL,2004: 271-278.