计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(9):104−111 [doi: 10.15888/j.cnki.csa.009933] [CSTR: 32024.14.csa.009933]

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

增强否定的自然语言推理①

曾碧卿[1,2], 罗智康[1], 陈威海[1]

[1](华南师范大学 人工智能学院, 佛山 528225)
[2](华南师范大学 阿伯丁数据科学与人工智能学院, 佛山 528225)
通信作者: 曾碧卿, E-mail: zengbiqing137@163.com

摘　要: 否定是自然语言中一种重要的表达方式, 对文本中表达的含义发挥着关键作用. 在自然语言推理中, 是否包含否定能够直接影响文本之间的语义关系. 然而, 当前的预训练模型在处理否定句时, 对语义关系判断的准确率会显著下降. 因此, 本文提出了一个增强预训练模型对自然语言推理中包含否定句的识别与理解的方法. 通过增强模型对文本中否定结构的注意力分数权重, 在不牺牲原有模型对肯定句处理的前提下, 有效提高了包含否定句的自然语言推理任务的准确率. 在针对否定的公开自然语言推理数据集上验证了本文方法的有效性.

关键词: 自然语言推理; 注意力机制; 预训练语言模型

引用格式: 曾碧卿,罗智康,陈威海.增强否定的自然语言推理.计算机系统应用,2025,34(9):104–111. http://www.c-s-a.org.cn/1003-3254/9933.html

Natural Language Inference via Negation Augmentation

ZENG Bi-Qing[1,2], LUO Zhi-Kang[1], CHEN Wei-Hai[1]

[1](School of Artificial Intelligence, South China Normal University, Foshan 528225, China)
[2](Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan 528225, China)

Abstract: Negation is an important expression form in natural language, playing a critical role in conveying textual meaning. In natural language inference, the presence of negation can directly affect semantic relationships between texts. However, current pre-trained models exhibit significant accuracy degradation in semantic relationship judgment when processing negative sentences. Therefore, this study proposes a method to enhance the ability of pre-trained models to recognize and comprehend negative sentences in natural language inference. By enhancing the attention score weights for negative structures in text, the proposed method significantly improves the accuracy of natural language inference tasks involving negative sentences without sacrificing the original performance on affirmative sentences of the model. The effectiveness of the proposed method has been verified on a public natural language inference dataset for negation analysis.

Key words: natural language inference (NLI); attention mechanism; pre-trained language model (PLM)

1　引言

自然语言推理 (natural language inference, NLI) 是自然语言处理研究中的一个重要领域, 旨在理解两段文本的语义并判断两者之间的关系[1]. 在自然语言推理中, 这两段文本通常被分别称为前提和假设, 两者之间的关系通常包括蕴含 (entailment)、矛盾 (contradiction) 和中立 (neutral). 自然语言推理在各种自然语言处理任务中发挥着至关重要的作用, 包括问答、阅读理解、文档摘要和关系提取. Transformer 是一种基于注意力机制的模型架构, 因具备良好的泛化性, 衍生出一系列

① 基金项目: 国家自然科学基金 (62076103); 广东省基础与应用基础研究基金 (2021A1515011171); 广州市基础研究计划基础与应用基础研究项目 (202102080282)
收稿时间: 2025-01-15; 修改时间: 2025-02-17; 采用时间: 2025-03-06; csa 在线出版时间: 2025-07-25
CNKI 网络首发时间: 2025-07-28

基于其架构的预训练模型, 如 BERT、GPT2 和 RoBERTa 等. 这些预训练模型使用大量的无标注文本进行预训练, 具备丰富的背景知识和强大的语义理解能力, 在处理特定任务时, 只需要针对特定数据集进行微调即可取得较好的性能. 这种预训练加微调的方法, 统一了多个自然语言处理任务, 成为如今主流的范式.

否定是自然语言中非常重要的表达方式, 也是形成逻辑关系的基本组成部分. 理解否定对于自然语言理解至关重要, 有助于提高诸如自然语言推理等任务的性能. Kassner 等人[2]在研究预训练语言模型 (PLM) 对句子语义的理解时提出两种探测任务, 否定 (negation) 探测任务和错误引导 (mispriming) 探测任务. 通过在 LAMA 数据集中的完形填空问题中插入了否定元素 (如“not”), 从而生成否定和非否定的完形填空问题. 如, 将“iOS is developed by [MASK]”修改为“iOS is not developed by [MASK]”. 实验结果显示, BERT 等模型在预测否定和非否定问题时, 预测结果均为“Apple”.

在自然语言推理任务中, 预训练模型也不能正确处理包含否定结构的实例. 具体表现为, BERT 等预训练语言模型经常将包含“not”或“no”等表示否定的词语但实际标签为中性或蕴含的文本对误分类为矛盾; 或把否定句当作肯定句处理. Hossain 等人[3]分析检查自然语言推理数据集中的样例, 发现数据集中肯定句与否定句的比例不平衡, SNLI 数据集[4]中包含否定结构的句子仅占 7.16%, MNLI 数据集[5]中否定句占比为 22.63%. 针对自然语言推理任务中否定句样例少的问题, 他们提出了一个新的测试集 NegNLI[6]. 该测试集在 SNLI、MNLI 和 RTE 数据集中分别选取 1 500 个样本, 并改写成否定句. BERT 等预训练模型在该测试集上的准确率相较于原数据集大幅下降, 在 NegNLI 的 SNLI 测试集上, BERT 的准确率仅为 49.10%, 与在 SNLI 测试集 89.30% 的准确率存在极大差异.

针对以上问题, 本文结合预训练模型, 提出一种两阶段对比学习的方法, 与自然语言推理中主流的集中于训练模型进行实验对比和研究分析, 主要贡献有以下 3 点.

(1) 提出一种否定增强方法, 能够根据现有数据集生成包含否定句的自然语言推理数据样本. 引入了 3 个否定的关系: 非蕴含、非中立和非否定, 用于增强否定表达的多样性.

(2) 设计一个注意力增强模块, 增强模型对句子中否定词的关注, 并通过实验证明该模块的有效性.

(3) 提出一个两阶段对比学习方法. 第 1 阶段对比学习通过结合否定增强的 3 个否定标签, 组成 3 组对立关系的两两对比. 第 2 阶段对比学习结合注意力增强模块, 进行原数据集上的 3 个关系的对比.

2 研究进展

自然语言推理是一个使用算法使计算机理解文本的语义并进行语义关系推断的任务. 本节从当前国内外关于自然语言推理数据集以及与否定问题相关的研究来进行表述.

自然语言推理的主要任务是理解两个文本的所表达的含义, 并根据语义推断两个文本之间的语义关系, 通常有蕴含、中立和矛盾 3 种关系. 完成自然语言推理任务要求模型或算法具有理解文本语义的能力, 并且还需要具备推理能力.

SNLI 数据集[4]是 Standford 提出的一个大规模自然语言推理数据集. 该数据集中的前提句来源于 Flickr 30k 数据集的图片描述, 假设句和标签则是人工生成. 研究者们根据 SNLI 数据集提出了许多衍生数据集. 例如, Camburu 等人[7]提出了 e-SNLI, 该数据集通过人工标注为每个样例标注解释和文本中的重点词. MNLI 数据集[5]是另一个代表性的大规模 NLI 数据集, 该数据集中的样例来自书信、新闻、帖子等多种文体, 用于评估模型在不同类型文本下的推理能力.

Devlin 等人[8]提出的基于 Transformer 编码器的预训练模型 BERT, 在多种不同自然语言处理任务中取得了优异的性能, 包括自然语言推理任务. RoBERTa、DeBERTa、T5 等预训练模型不断刷新自然语言推理任务的准确率. 至此预训练模型成为自然语言推理任务的首选模型. Hossain 等人[6]研究发现, 现有的大规模数据集 SNLI、MNLI 中缺乏足够的否定句样本, 因此提出了一个针对自然语言推理任务中否定句处理的 NegNLI 测试集. RoBERTa、DeBERTa 等预训练模型在该测试集上均表现不佳, 准确率大幅下降.

为解决自然语言推理的否定句问题, Hosseini 等人[9]提出了 BERTNOT 模型. BERTNOT 模型采用否定掩码语言模型的方式进行训练, 即前文提到的“iOS is not developed by [MASK]”方式. Helwe 等人[10]提出了 TINA 方法, 该方法采用 Unlikelihood 损失函数进行分类训练, 从减少错误分类的角度优化模型的准确率.

3 研究方法

本方法主要涉及 3 个部分, 模型概览如图 1 所示.

(1) 否定增强: 该方法能够利用现有数据集合成大量包含否定句的自然语言推理数据样本, 提高数据集中否定结构的占比.

(2) 注意力增强模块: 该模块采用多头注意力, 结合旋转位置编码, 优化模型对文本中否定词的注意力权重分布.

(3) 两阶段对比学习方法: 通过两次不同的对比学习任务, 增强模型对文本中否定语义的识别与理解.

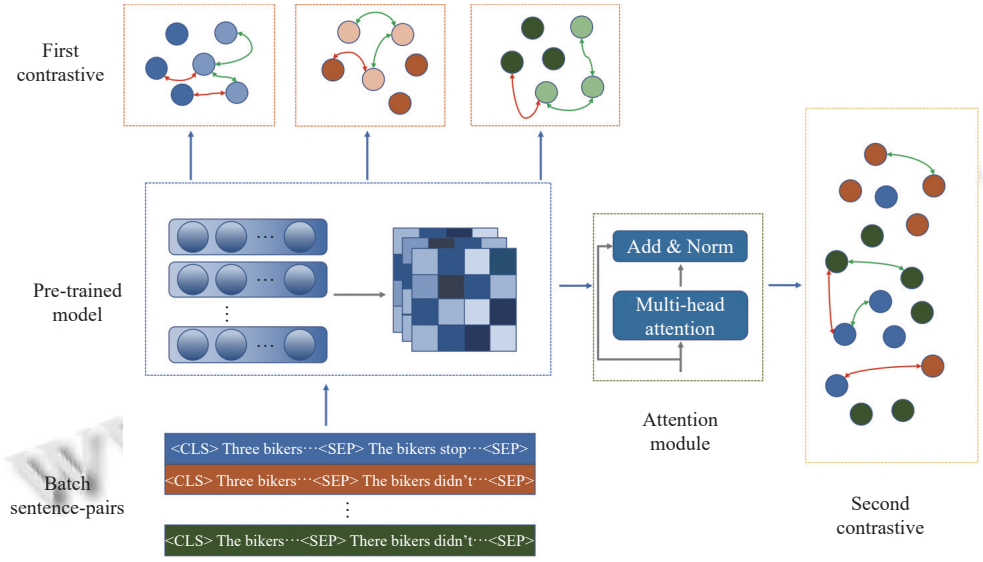

图 1 模型概览图

3.1 否定增强

否定增强是一种数据增强方法, 旨在提高自然语言推理数据集中包含否定结构的数据样本的数量. 本文使用 Spacy 工具包对数据集中的句子进行句法分析, 然后通过规则匹配句子中的动词、助动词等能够被否定改写的词语, 通过对这些词改写得到否定句. 对于句子“Nowhere in his confession did he mention the Monteagle letter”, 其匹配的改写规则如图 2 所示.

```
{
  "name": "aux before subj",
  "pattern": "{$;tag:/VB.*/}=A >/advmod|cc/ {word:/never|nobody|no|nothing|
             nowhere|neither|Never|Nobody|No|Nothing|Nowhere|Neither/}=npiword
             >/aux.*/ ({}= B $++ {}=subject) >/nsubj.*/ {}=subject ?>obj
             {tag:/NN.*/}=object",
  "actions": [
    {
      "type": "move",
      "to_move": "B",
      "anchor": "A",
      "position": "before"
    },
    {
      "type": "replace",
      "token": "",
      "to_replace": "npiword"
    }
  ]
}
```

图 2 改写规则示例

通过词性标签匹配到“mention”和“did”, 分别标记为“A”和“B”, 通过单词匹配到“Nowhere”, 将其标记为“npiword”. 该规则执行两个动作: 将“B”移动到“A”之前; 把“npiword”替换为空字符. 改写后得到其否定句“in his confession he did mention the Monteagle letter”.

得到前提和假设的否定句后, 将原始前提、原始假设、前提的否定、假设的否定两两组合得到新的句子对. 得到句子对后, 根据原始前提与原始假设的语义关系, 通过逻辑推导得到句子对的标签, 从而得到增强样本, 流程如图 3 所示.

否定增强还在原始 3 标签数据集中拓展得到 3 个否定标签: 非蕴含、非中立、非矛盾. 通过 3 个否定标签能够更加全面的表达前提与假设的语义关系.

3.2 注意力增强模块

由于预训练模型不能正确地识别出句子中的否定结构, 本文设计了一个注意力增强模块用于辅助模型识别句子中的否定结构. 该模块包含添加位置信息、注意力计算和归一化 3 个步骤. 预训练模型, 如 BERT 等, 会对输入文本进行特征编码, 得到长度为 N 的文本特征, 可以记为 $S_N = \{x_m\}_{m=1}^{N}$, 其中 $x_m \in R^d$ 表示序列中的第 m 个 token.

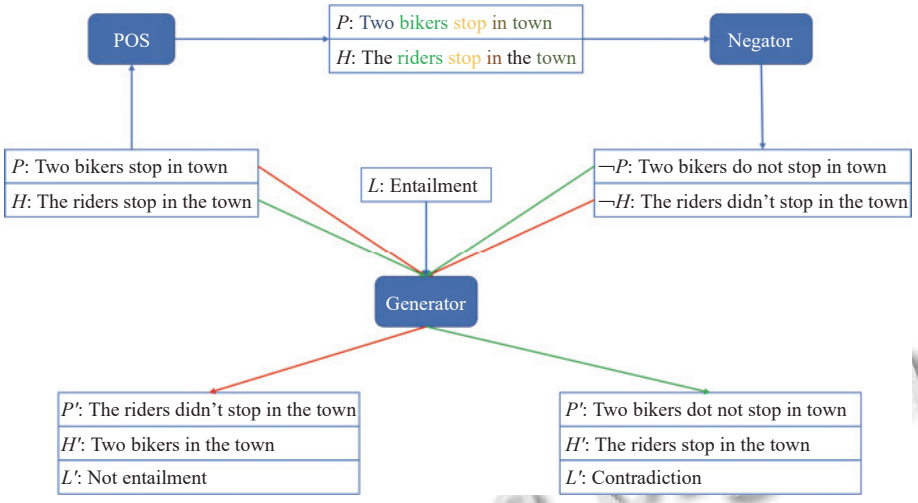

图 3 否定增强的流程

注意到文本中的否定结构通常与被否定的部分相对靠近, 本模块引入旋转位置编码 (RoPE)[11]增强文本中 token 之间的位置信息. 对于每个 token, 其嵌入向量可以表示为$x_m=[x_{m,1},x_{m,2},\cdots,x_{m,d}]$, 其中 d 是偶数. 将x_m划分为$d/2$对二维向量$(x_{m,2k-1},x_{m,2k})$, 其中$k=1,2,\cdots,d/2$. x_m经过 RoPE 编码后得到x'_m, 计算过程可表示为:

$$\begin{pmatrix} x'_{m,2k-1} \\ x'_{m,2k} \end{pmatrix}=\begin{pmatrix} \cos\theta_m & -\sin\theta_m \\ \sin\theta_m & \cos\theta_m \end{pmatrix}\begin{pmatrix} x_{m,2k-1} \\ x_{m,2k} \end{pmatrix} \tag{1}$$

其中, $\theta_m=\dfrac{m}{10000^{2k/d}}$.

注意力计算采用 Transformer[12]中的点积多头注意力计算方法. 在进行注意力计算前, 首先对输入进行线性变换得到$Q=W^Q S$, $K=W^K S$, $V=W^V S$. 其中, W^Q、W^K、W^V是可学习的权重矩阵. 随后对Q、K进行位置编码得到Q'、K', 序列中任意两个 token 的点积可以表示为:

$$q'_m\hat{k}'_n=q_m k_n*(\cos(m-n)\theta+i\sin(m-n)\theta) \tag{2}$$

$$q'_m=q_m*(\cos m\theta+i\sin m\theta) \tag{3}$$

$$\hat{k}'_m=k_n*(\cos n\theta-i\sin n\theta) \tag{4}$$

其中, $*$表示元素相乘. 因此模块中多头注意力的第j个头的注意力计算可以表示为:

$$head_j=Attention(Q,K,V)_j=Softmax\left(\frac{Q'\widehat{K'}^{\mathrm{T}}}{\sqrt{d_j}}\right)V \tag{5}$$

其中, d_j表示第j个头的维度, 将每个头的注意力计算拼接得到整体的注意力. 拼接各个头的注意力计算结果后, 再通过一个线性层进行投影, 并使用残差连接得到注意力增强模块的计算结果. 计算步骤可以表示为:

$$T=Norm(S+Concat[head_1,head_2,\cdots,head_n]) \tag{6}$$

其中, *Norm* 表示 RMSNorm 归一化操作, *Concat* 表示拼接操作, S 是预训练模型提取的文本序列特征.

3.3 两阶段对比学习

对比学习是计算机视觉领域中常用的方法, 也应用于自然语言推理[13,14]任务中. 对比学习的核心思想是在数据的向量特征空间中, 为每个目标样本指定对应的正样本和负样本, 正样本通常与目标样本具有相同的标签, 负样本通常与目标样本的标签不同. 通过在特征空间中减小正样本与目标样本之间的距离, 增大负样本与目标样本之间的距离, 使得模型能够理解样本之间的特征表示. 本文使用一个两阶段对比学习方法来提高预训练模型对文本中否定语义的理解和处理.

在前文提到的否定增强中, 本文引入了 3 个否定标签: 非蕴含、非中立与非假设. 这 3 个否定标签可以与原数据集中的 3 个标签: 蕴含、中立与假设组成 3 组对立关系, 第 1 阶段的对比学习基于这 3 组对立关系进行. 在这一阶段, 本方法只区分这 3 组关系的肯定与否定, 由于否定标签是一个更加宽泛的表达, 因此, 这 3 组关系会将特征向量投影到 3 个独立的特征空间进行表达. 这一阶段的对比学习相似度计算使用点积相似度, 计算公式表示为:

$$sim(S_i,S_p)=\frac{S_i S_p}{\sqrt{d}} \tag{7}$$

$$l_{i,p}=\frac{e^{sim(S_i,S_p)/\tau_1}}{\sum_{k\in I,k\neq i}e^{sim(S_i,S_p)/\tau_1}} \tag{8}$$

$$L_1 = \sum_{i \in I} \frac{1}{|P|} \sum -\log l_{i,p} \tag{9}$$

其中, I 表示一个批次的样本数据, i 是 I 中的一个样本, S 表示预训练模型提取的文本序列特征, $p \in P$ 表示样本 i 的正样本, τ_1 表示温度.

与第 1 阶段的对比学习宽泛地区分每组对立关系不同, 第 2 阶段的对比学习回归到原始数据集中 3 个标签的对比, 旨在提高模型对句子否定结构的识别, 并准确判断前提与假设的语义关系. 这一阶段的对比学习相似度计算采用余弦相似度, 计算公式为:

$$l_{i,p} = \frac{e^{\cos(T,T_p)/\tau_2}}{\sum_{k \in I, k \neq i} e^{\cos(T,T_p)/\tau_2}} \tag{10}$$

$$L_2 = \sum_{i \in I} \frac{1}{|P|} \sum -\log l_{i,p} \tag{11}$$

其中, I 表示一个批次的样本数据, i 是 I 中的一个样本, T 表示注意力模块输出的文本序列特征, $p \in P$ 表示样本 i 的正样本, τ_2 表示温度.

最后, 将两个阶段对比学习的损失相加得到最终的两阶段对比损失:

$$Loss = L_1 + L_2 \tag{12}$$

4 实验

4.1 数据集

本文在 NegNLI 数据集上评估本文的模型方法, 这是一个专用于测试自然语言推理模型对处理包含否定结构的句子的公共测试集. Hossain 等人[6]在 SNLI、MNLI 和 RTE 数据集中分别选择了 1 500 条样本进行否定改写构建了该数据集, 对于 SNLI 和 MNLI, 这 3 种标签每个标签的数据各有 500 条. 数据集的统计信息见表 1. 本文的训练集采用 SNLI 和 MNLI 的训练集, 并在这两个训练集上使用否定增强获得新样本. 为防止新样本被包含在测试集中, 对于新生成的样本会与所有测试数据计算词重叠度, 若新生成的样本中的词语与任意一条测试数据的词语重叠度达到 60%, 该样本将被舍弃. 生成数据与原始的训练数据一起组成训练集和验证集, 两者按 9:1 随机切分.

表 1 数据集的统计信息

标签	$SNLI_{dev}$	$SNLI_{neg}$	$MNLI_{m}$	$MNLI_{mm}$	$MNLI_{neg}$
蕴含	3 329	500	3 379	3 463	500
中立	3 335	500	3 123	3 129	500
矛盾	3 378	500	3 213	3 240	500
合计	9 842	1 500	9 815	9 832	1 500

4.2 基线模型

实验超参数参考 BERTNOT[9]论文中的设置, 对于 SNLI 数据集, 学习率设置为 1E–5, 权重衰退设置为 0.1; 对于 MNLI 数据集, 学习率设置为 2E–5, 权重衰退设置为 0. 其他实验超参数相同, 如表 2 所示.

表 2 实验超参数

参数	数值
文本最大长度	512
注意力模块头数量	3
Epoch	3
批大小	32
Dropout率	0.1
第1阶段对比的温度	0.08
第2阶段对比的温度	0.05

为验证本文方法的有效性, 本文选取了 Encoder-Only、Decoder-Only 和 Encoder-Decoder 这 3 类不同结构的预训练模型, 验证本文方法对改进预训练模型处理否定句的有效性. 此外还与 BERTNOT 和 TINA 两个方法进行了对比. 选取的预训练模型具体如下.

BERT[8]: 基于 Transformer 编码器架构的双向编码模型. 它在两个任务上进行预训练: 掩码语言模型 (MLM) 和下一句预测 (NSP).

RoBERTa[15]: 作为 BERT 的一种变体, 其架构与 BERT 相似, 但在许多自然语言处理任务上实现了更好的性能. 与 BERT 相比, 它在更大的数据集上使用更大的批次进行了更长时间的预训练, 并且仅针对 MLM 任务训练, 去除了 BERT 中的 NSP 任务.

ALBERT[16]: 是一种轻量化的 BERT 架构, 通过分解 Embedding 层的参数和跨层参数共享来减少参数量, 并使用句子顺序预测任务替代了 NSP 任务.

GPT2[17]: 仅由 Transformers 解码器架构组成的模型. 它的预训练任务是预测下一个词, 即根据已有的文本序列预测接下来的一个词.

XLNet[18]: 是一种基于 Transformer 的模型, 该模型在称为排列语言建模的任务上进行预训练. 排列语言建模是指通过训练模型对所有可能的句子中词语排列进行学习, 以捕捉双向上下文.

BART[19]: 一个由类似 BERT 的编码器块和类似 GPT 的解码器块组成的序列到序列模型. 预训练任务包括在应用不同的噪声函数 (如标记遮挡、标记填充和句子排列) 后, 将损坏的文本重建为其原始形式.

ModernBERT[20]: 它使用卷积神经网络来表示图像和文本, 并通过跨模态比较模块学习文本和图像之间

的差异. 然后, 一个语义关联模块将文本和图像特征映射到一个共享的语义空间.

4.3 主要结果

实验通过与基线模型进行比较来评估本文方法的有效性. 实验结果如表 3 所示, 表中数据分别为准确率与 Macro-$F1$ 值, *表示数据引用自原文献. 后缀为 TINA 的模型表示该模型使用了 TINA 方法, 后缀为 Ours 的模型表示该模型使用了本文的方法.

通过对 Encoder-Only、Decoder-Only 和 Encoder-Decoder 这 3 类模型进行实验, 实验结果表明本文的方法能够增强这 3 类模型在自然语言推理任务中包含否定句时, 对句子对进行语义关系推理的准确率. 并且与 TINA 方法相比, 本文的方法在以肯定句为主的数据集上, 准确率的损失更小.

表 3 本文方法与其他方法的准确率/Macro-$F1$ 结果比较 (%)

模型	SNLI		MNLI		
	$SNLI_{dev}$	$SNLI_{neg}$	$MNLI_{m}$	$MNLI_{mm}$	$MNLI_{neg}$
BERTNOT*	89.00/—	45.96/—	84.31/—	—/—	60.89/—
BERT	90.12/90.08	50.46/50.31	83.29/83.20	83.84/83.71	63.47/60.80
BERT-TINA	89.62/89.60	68.13/68.02	81.23/80.73	81.84/81.76	68.23/67.91
BERT-Ours	90.51/90.32	72.40/68.39	83.12/82.67	83.07/82.83	72.26/72.22
RoBERTa	91.37/91.34	56.73/56.70	85.87/85.79	85.74/85.73	65.26/65.13
RoBERTa-TINA	90.86/90.73	65.16/65.15	85.16/85.02	84.86/84.79	69.13/68.99
RoBERTa-Ours	91.42/91.40	73.42/73.23	85.30/85.12	85.82/85.70	73.43/74.32
GPT2	88.12/88.08	47.47/47.26	80.86/80.63	81.23/81.10	62.33/62.15
GPT2-TINA	87.60/87.58	54.47/54.29	80.52/80.45	80.71/80.69	67.13/66.92
GPT2-Ours	88.47/88.41	67.76/67.74	80.74/80.71	81.10/80.92	69.26/69.24
XLNet	90.91/90.90	53.93/53.90	85.13/85.06	85.32/85.26	65.45/65.38
XLNet-TINA	90.62/90.53	66.72/66.65	84.81/84.72	84.63/84.55	68.92/68.91
XLNet-Ours	91.24/91.20	69.73/69.73	84.83/84.76	84.42/84.13	72.93/72.87
BART	90.93/90.86	56.86/56.81	84.81/84.74	84.96/84.86	66.36/66.13
BART-TINA	90.84/90.77	68.87/68.76	84.01/83.92	83.96/83.88	69.34/69.28
BART-Ours	91.21/91.17	72.14/72.13	84.38/83.32	84.57/84.21	70.87/70.85
ALBERT	89.67/89.65	51.73/51.67	82.87/82.86	82.87/82.82	63.36/63.01
ALBERT-Ours	89.81/89.76	70.20/70.03	82.93/82.89	83.24/83.19	71.21/71.19
ModernBERT	91.52/91.52	55.73/55.66	87.73/87.65	88.12/88.01	64.80/64.46
ModernBERT-Ours	91.21/90.70	72.67/70.8	86.72/86.62	87.17/87.03	73.26/73.05

4.4 消融实验

为了研究本方法中每个部分对模型性能的影响, 本文进行了消融实验, 测试了去除某些组件后模型的性能. 具体来说, 本文以 BERT 为骨架模型, 研究了以下几种消融实验的变体.

(1) 去除注意力增强模块: 该变体去除了注意力增强模块, 同时由于第 2 阶段对比学习依赖注意力增强模块, 因此该变体也去除了第 2 阶段的对比学习.

(2) 仅去除第 1 阶段对比学习: 仅去除了方法中第 1 阶段的对比学习.

(3) 仅去除第 2 阶段对比学习: 仅去除了方法中第 2 阶段的对比学习.

(4) 去除两阶段对比学习: 去除第 1 阶段和第 2 阶段的对比学习.

结果如表 4 所示. 如预期的那样, 完整方法 (包括所有组件) 获得了最佳性能. 此外, 去除注意力增强模块 (变体 1) 会导致性能大幅下降. 这表明注意力增强模块可能有效注意到了文本否定结构, 该模块与两阶段对比学习相结合能够提高模型对否定的理解. 仅去除第 1 阶段对比学习 (变体 2) 和仅去除第 2 阶段对比学习 (变体 3) 均会对模型性能有一定影响, 当同时去除两个阶段的对比学习 (变体 4) 后, 模型的准确率下降较大. 这表明两阶段的对比学习相结合对模型理解否定有较大的贡献.

表 4 准确率的消融实验结果 (%)

方法以及变体	$SNLI_{neg}$	$MNLI_{neg}$
完整方法	72.40	72.26
去除注意力增强模块	62.51	65.28
仅去除第1阶段对比学习	71.12	70.81
仅去除第2阶段对比学习	70.06	69.58
去除两阶段对比学习	67.40	67.31

为了探讨温度两阶段对比学习的影响, 我们在 SNLI 数据集上对两个对比学习选取不同的温度进行了实验, 实验结果如图 4 所示. 随着温度的提高, 模型的性能先提高后下降. 温度过高会使模型对负样本的

关注降低, 而温度过低会使模型只能关注到区别较大的负样本. 本方法中的两个对比学习对正负样本定义和相似度计算不同, 导致两个对比学习的最佳温度有所不同, 根据实验结果, 最终两个对比学习分别选取 0.08 和 0.05 作为温度参数.

图 4 不同温度下模型的准确率 (%)

此外, 为了进一步研究本方法的贡献, 我们还对 SNLI 数据集中句子对特征表示进行了 t-SNE 可视化, 如图 5 和图 6 所示. 图 5 中是 BERT 原始模型的输出, 图 6 是本方法中注意力增强模块的输出. 可以观察到, 使用注意力增强模块和对比学习后, 模型能够将同一类别正样本的数据映射到特征空间中相近的位置, 同时将不同类别的负样本映射到更远的位置. 这也很好地解释了变体 1 中去除注意力增强模块后准确率大幅下降的原因.

4.5 案例分析

为了进一步解释模型工作的原因, 我们可视化了模型的注意力分布, 以说明本方法能够预训练模型在自然语言推理中对否定句识别和理解. 如图 7 所示, 原始模型对更加注重分类头的信息, 并未对文本中的否定词加以关注; 而经过了本文的方法处理之后, 模型能够更多地关注文本中表达否定的词语. 此外, 对于更复杂的双重否定与隐含否定场景, 如图 7 中第 3 个示例和第 4 个示例, 本文的方法也使模型能关注句子中表达否定的词语.

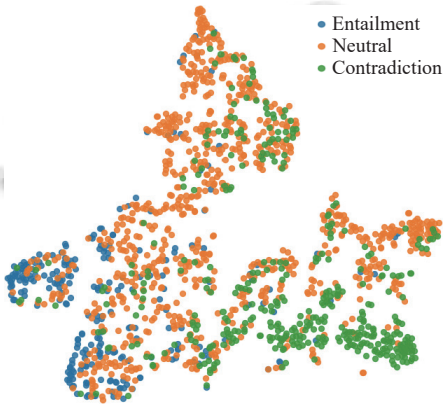

图 5 原始模型的 t-SNE 可视化

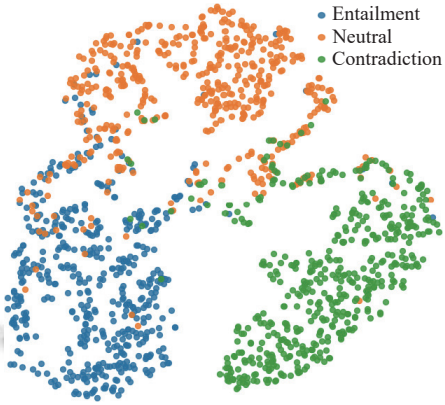

图 6 本方法的 t-SNE 可视化

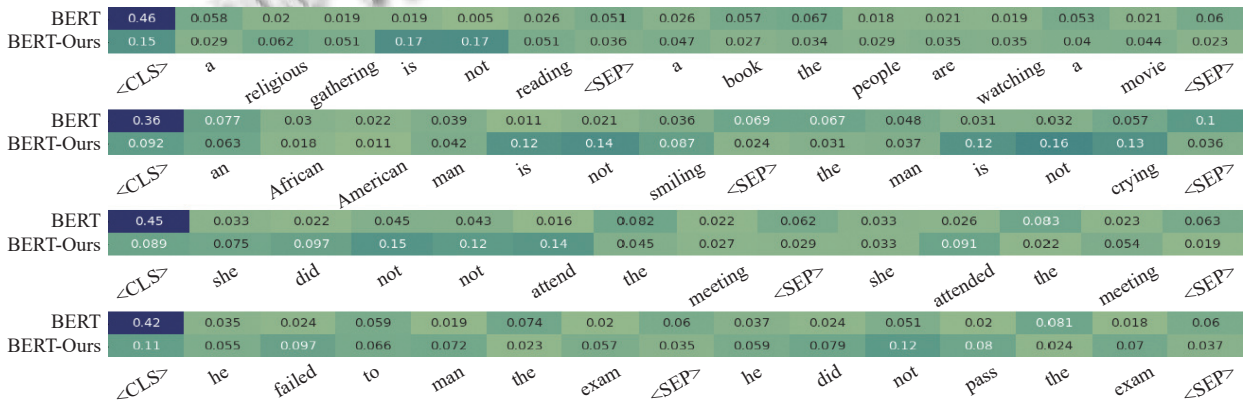

图 7 部分句子中词语的注意力权重

5 总结

本文提出了一种自然语言推理否定增强方法. 通过设计一个注意力增强模块加强模型对句子中否定结构的识别, 并使用两阶段的对比学习方法提高预训练模型对文本中否定语义的理解, 本文方法能够在不牺牲模型对肯定句的处理能力的前提下, 大幅提高模型对自然语言推理中的否定句的理解, 从而提高模型对包含否定句的自然语言推理任务的准确率. 实验结果表明, 本文的方法处理否定句的准确率优于其他基准模型和方法.

本研究的一个不足之处是本文的方法仅在分类类型的自然语言推理数据集上评估了本文方法的有效性. 然而, 尽管大多数自然语言推理数据集是三分类数据集, 但部分自然语言推理数据集不属于分类任务, 如问答类型的自然语言推理任务. 未来, 我们的工作将探索其他非分类类型的数据集.

参考文献

1 Sadat M, Caragea C. MSciNLI: A diverse benchmark for scientific natural language inference. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Mexico City: ACL, 2024. 1610–1629.

2 Kassner N, Schütze H. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. arXiv:1911.03343, 2019.

3 Hossain M, Hamilton K, Palmer A, *et al*. Predicting the focus of negation: Model and error analysis. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 8389–8401.

4 Bowman SR, Angeli G, Potts C, *et al*. A large annotated corpus for learning natural language inference. arXiv:1508.05326, 2015.

5 Williams A, Nangia N, Bowman SR. A broad-coverage challenge corpus for sentence understanding through inference. arXiv:1704.05426, 2017.

6 Hossain M, Kovatchev V, Dutta P, *et al*. An analysis of natural language inference benchmarks through the lens of negation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 9106–9118.

7 Camburu OM, Rocktäschel T, Lukasiewicz T, *et al*. e-SNLI: Natural language inference with natural language explanations. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 9560–9572.

8 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. Minneapolis: ACL, 2019. 4171–4186.

9 Hosseini A, Reddy S, Bahdanau D, *et al*. Understanding by understanding not: Modeling negation in language models. arXiv:2105.03519, 2021.

10 Helwe C, Coumes S, Clavel C, *et al*. TINA: Textual inference with negation augmentation. Proceedings of the 2022 Findings of the Association for Computational Linguistics. Abu Dhabi: ACL, 2022. 4086–4099.

11 Su JL, Ahmed M, Lu Y, *et al*. RoFormer: Enhanced Transformer with Rotary Position Embedding. Neurocomputing, 2024, 568: 127063. [doi: 10.1016/j.neucom.2023.127063]

12 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: Curran Associates Inc., 2017. 6000–6010.

13 Li S, Hu XM, Lin L, *et al*. Pair-level supervised contrastive learning for natural language inference. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 8237–8241.

14 Li SA, Hu XM, Lin L, *et al*. A multi-level supervised contrastive learning framework for low-resource natural language inference. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1771–1783. [doi: 10.1109/TASLP.2023.3270771]

15 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.

16 Lan ZZ, Chen MD, Goodman S, *et al*. ALBERT: A lite BERT for self-supervised learning of language representations. arXiv:1909.11942, 2019.

17 Radford A, Wu J, Child R, *et al*. Language models are unsupervised multitask learners. OpenAI Blog, 2019, 1(8): 9.

18 Yang ZL, Dai ZH, Yang YM, *et al*. XLNet: Generalized autoregressive pretraining for language understanding. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 517.

19 Lewis M, Liu YH, Goyal N, *et al*. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461, 2019.

20 Warner B, Chaffin A, Clavié B, *et al*. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv:2412.13663, 2024.

(校对责编: 王欣欣)