基于对抗正则化的自然语言推理

刘广灿 [1,2] 曹 宇 [1] 许家铭 [2] 徐 波 [2,3]

摘 要 目前自然语言推理 (Natural language inference, NLI) 模型存在严重依赖词信息进行推理的现象. 虽然词相关的判别信息在推理中占有重要的地位, 但是推理模型更应该去关注连续文本的内在含义和语言的表达, 通过整体把握句子含义进行推理, 而不是仅仅根据个别词之间的对立或相似关系进行浅层推理. 另外, 传统有监督学习方法使得模型过分依赖于训练集的语言先验, 而缺乏对语言逻辑的理解. 为了显式地强调句子序列编码学习的重要性, 并降低语言偏置的影响, 本文提出一种基于对抗正则化的自然语言推理方法. 该方法首先引入一个基于词编码的推理模型, 该模型以标准推理模型中的词编码作为输入, 并且只有利用语言偏置才能推理成功; 再通过两个模型间的对抗训练, 避免标准推理模型过多依赖语言偏置. 在 SNLI 和 Breaking-NLI 两个公开的标准数据集上进行实验, 该方法在 SNLI 数据集已有的基于句子嵌入的推理模型中达到最佳性能, 在测试集上取得了 87.60 % 的准确率; 并且在 Breaking-NLI 数据集上也取得了目前公开的最佳结果.

关键词 深度学习, 自然语言推理, 语言偏置, 对抗正则化

引用格式 刘广灿, 曹宇, 许家铭, 徐波. 基于对抗正则化的自然语言推理. 自动化学报, 2019, **45**(8): 1455−1463

DOI 10.16383/j.aas.c190076

Natural Language Inference Based on Adversarial Regularization

LIU Guang-Can[1,2] CAO Yu[1] XU Jia-Ming[2] XU Bo[2,3]

Abstract At present, natural language inference (NLI) models rely heavily on word information. Although the discriminant information related to the words plays an important role in inference, the inference models should pay more attention to the internal meaning of continuous text and the expression of language, and carry out inference through an overall grasp of sentence meaning rather than make shallow inference based on the opposition or similarity between individual words. In addition, the traditional supervised learning method makes the model rely too much on the language priori of the training set, and lacks the understanding of the language logic. In order to explicitly emphasize the importance of the learning sequence encoding and reduce the impact of language bias, this paper proposes a natural language inference method based on adversarial regularization. This method firstly introduces an inference model based on word encoding, which takes the word encoding in the standard inference model as input, and it can infer successfully only by using language bias. Then, through the adversarial training between the two models, the standard inference model can avoid relying too much on language bias. Experiments were carried out on two open standard datasets, SNLI and Breaking-NLI. On the SNLI dataset, the method achieves the best performance in existing inference models based on sentence embedding, and achieves 87.60 % accuracy in test set. And the inference model has achieved state-of-the-art result on the Breaking-NLI dataset.

Key words Deep learning, natural language inference (NLI), language bias, adversarial regularization

Citation Liu Guang-Can, Cao Yu, Xu Jia-Ming, Xu Bo. Natural language inference based on adversarial regularization. *Acta Automatica Sinica*, 2019, **45**(8): 1455−1463

自然语言推理 (Natural language inference, NLI) 又称为文本蕴含识别 (Recognizing textual entailment, RTE)[1−2], 是自然语言处理 (Natural language processing, NLP) 中一个重要的研究问题. 自然语言推理是一个确定两个或多个句子之间逻辑关系的任务, 例如: 给定一个前提 (Premise) 和一个假设 (Hypothesis), 目标是确定它们之间的逻辑关系是蕴涵、中立还是矛盾. SNLI[3] 和 Breaking-NLI[4] 等一系列高质量、大规模标准数据集的发布推动了自然语言推理的发展, 促进了大量相关研究[5−11], 表 1 展示了几个 SNLI 数据集中的例子. 目前基于神经网络的推理模型主要有两类: 一类侧重前提和假设分别进行句子嵌入, 然后使用分类器将

收稿日期 2019-01-30 录用日期 2019-06-02
Manuscript received January 30, 2019; accepted June 2, 2019
国家自然科学基金 (61602479), 中国科学院战略性先导科技专项基金 (XDB32070000), 北京脑科学专项基金 (Z181100001518006) 资助
Supported by National Natural Science Foundation of China (61602479), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB32070000), and the Beijing Brain Science Project (Z181100001518006)
本文责任编委 张军平
Recommended by Associate Editor ZHANG Jun-Ping
1. 哈尔滨理工大学自动化学院 哈尔滨 150080 2. 中国科学院自动化研究所 北京 100190 3. 中国科学院脑科学与智能技术卓越创新中心 上海 200031
1. School of Automation, Harbin University of Science and Technology, Harbin 150080 2. Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190 3. Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031

表 1 SNLI 数据集上的三个例子

Table 1 Three examples from the SNLI dataset

	Premise (前提)	Hypothesis (假设)	Label (标签)
	A soccer game with multiple males playing.	Some men are playing a sport.	Entailment
(译文)	一场有多名男子参加的足球比赛.	有些男人在做运动.	蕴涵
	A person on a horse jumps over a broken down airplane.	A person is training his horse for a competition.	Neutral
(译文)	一个人骑着马跳过了一架坏掉的飞机.	为了参加比赛, 一个人正在训练他的马.	中立
	A black race car starts up in front of a crowd of people.	A man is driving down a lonely road.	Contradiction
(译文)	一辆黑色赛车在一群人面前启动.	一个男人开着车行驶在荒凉的路上.	矛盾

其组合起来; 另一类不是分别处理两个句子, 而是使用交互注意力机制进行句子之间的交互. 本文关注基于句子嵌入的方法, 因为该方法没有限定要求两个句子, 可以延展到更多任务上.

对自然语言推理广泛的研究使得很多复杂模型在基准数据集上取得了越来越高的表现, 但是最近的研究[11] 表明多数模型很少关注前提和假设的句义关系, 而是大量利用句子中个别词之间对立或相似等浅显关系进行推理作答, 更有甚者只是根据假设就可以进行推理. 可想而知这些推理模型很难应用到复杂的现实场景中, 它们根据句子中特定词之间的关系进行盲目推理, 比如根据前提中的 “expensive” 词和假设中的 “cheap” 词, 简单推理出两个句子是对立关系, 而实际上两句话描述的不是同一件事情, 正确的逻辑关系应该是中立. 推理模型过度依赖特定词, 说明模型只是抓住数据集中的语言偏置, 而不是依据前提和假设所表达的句义关系进行逻辑推理.

一种检测语言偏置对推理模型影响的方式是设计一个仅依赖词编码表示进行推理的模型 (为了方便描述, 本文使用 WIM (Word inference model) 表示仅依赖词编码表示进行推理的模型), 事实上 WIM 也可以作为一个标准的基线模型. 本文提出使用对抗正则化方法来降低语言偏置的影响, 具体方法是让一个标准的推理模型和这个只依赖词编码表示进行推理的对手进行博弈, 以减少语言偏置的影响. 在对抗机制下, 一方面训练 WIM, 使得该模型尽可能推理正确, 其中 WIM 模型的词编码表示是由标准推理模型提供; 另一方面训练标准推理模型, 调整它的词编码和句编码部分, 目的是在提高自身推理准确率的同时, 尽量降低 WIM 模型的性能. 在这种新颖的对抗正则化机制下, 优化自然语言推理模型.

本文提出的模型可以端到端训练, 而且扩展和延伸性比较强. 在 SNLI 和 Breaking-NLI 数据集上的实验结果表明了该方法的有效性: 本文提出的方法在 SNLI 数据集基于句子嵌入的推理模型中取得了最好的结果, 而且在 Breaking-NLI 数据集中也取得了领先的表现.

本文的主要贡献如下: 1) 通过多样信息整合, 多层级句子编码, 增强自然语言推理模型对句子的表示能力, 以探索更多语义信息. 2) 关注自然语言推理中的语言偏置现象, 并使用对抗正则化方法来解决这个问题, 此外该方法没有增加模型的参数, 不会增加模型测试时的复杂度. 3) 通过在 SNLI 和 Breaking-NLI 数据集上的实验表明本文提出方法的有效性, 模型推理表现取得了有效的提升.

1 相关工作

1.1 自然语言推理

目前句子嵌入在自然语言推理的众多方法中得到了广泛的应用, 这些方法背后的基本思想是分别对前提语句和假设语句进行编码, 然后将它们的句子表示结合起来使用神经网络进行分类, 具体结构如图 1 所示. 在已有的工作中, 很多研究工作使用卷积神经网络 (Convolution neural network, CNN) 和长短时记忆网络 (Long short-time memory, LSTM) 作为构建模块, 如 Liu 等[12] 提出基于双向长短时记忆网络 (Bidirectional LSTM, BiLSTM) 的句子编码结构, Mou 等[13] 提出基于树的 CNN 句子编码结构. 也有很多使用更加复杂的神经网络进行句子嵌入的研究工作, 如 Munkhdalai 等[14] 提出 NSE (Neural semantic encoder) 的记忆增强神经网络, 用于自然语言推理任务. 最近一些研究者开始探索应用于句子嵌入表示的自注意力机制. Shen 等[6] 提出 DiSAN 模型, 该模型没有使用 CNN 和循环神经网络 (Recurrent neural network, RNN), 而是完全依赖于研究者提出的多维注意力和双向自注意力机制. Shen 等[15] 提出 ReSAN (Reinforced self-attention network) 模型, 该模型使用强化学习将软注意力和硬注意力融合在一起. Im

等[16] 提出基于距离的自注意力网络模型, 该模型利用距离掩蔽来关注单词之间的距离, 从而对局部依赖关系进行建模. 此外, 还有研究者将胶囊网络中的动态路由机制应用到自然语言推理任务中[17], 并且取得了不错的效果. 虽然在自然语言推理中, 句子嵌入方法已经显示出其有效性, 但是也有多项研究表明, 将前提和假设句子对在句子编码期间联合处理, 关注它们之间的复杂交互, 模型会得到更好的结果. 然而, 这些交互式的方法不能在很多单个句子处理的任务上直接使用, 也不能像句子嵌入一样直接提供关于句子的语义理解. 本文选择基于句子嵌入的体系结构, 以便应用于更多 NLP 任务.

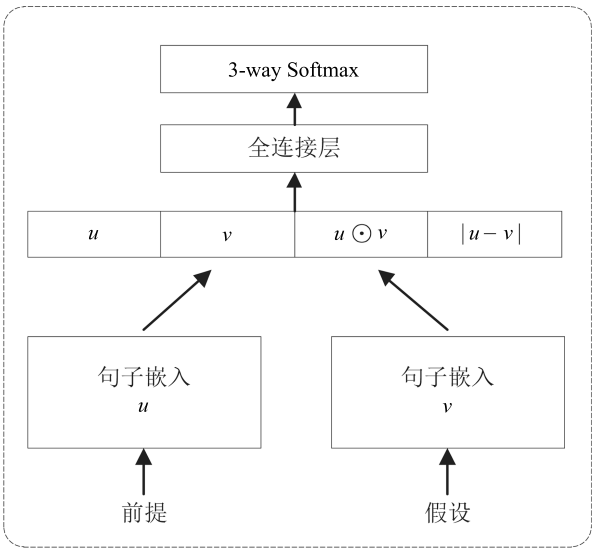

图 1　自然语言推理 (NLI) 整体结构框图

Fig. 1　The structure of natural language inference (NLI)

1.2　对抗学习

Goodfellow 等[18] 提出生成对抗网络 (Generative adversarial network, GAN) 作为一种学习数据分布的新方式. 生成对抗网络包含一个生成器 G 和一个判别器 D, G 和 D 在一个极小极大的博弈中被同步训练, 优化目标是达到纳什均衡

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim p_{\mathrm{data}}}[\log D(x)] + \mathrm{E}_{z\sim p_z}[\log(1-D(G(z)))] \tag{1}$$

其中, 生成器 G 根据从先验分布 p_z 中采样的隐含输入变量 z 来产生真实的数据, 以愚弄判别器 D. 另一方面, 判别器 D 是一个典型的二元分类器, 它试图去辨别它的输入数据是来自训练集还是来自生成器生成的集合. 生成对抗网络通过判别器为生成器提供损失梯度进行训练, 目的是学习一个生成模型, 使该模型的输出满足特定的分布 p_{data}.

生成对抗网络具有强大的模拟复杂分布的能力, 已受到广泛关注, 并且在图像和文本生成等领域演化出很多变体, 取得了大量令人瞩目的效果. 如针对对抗网络自身的改进 LSGAN[19] 和 WGAN[20], 对抗网络在图像生成上的应用 BicycleGAN[21] 和 DualGAN[22], 在文本生成上的应用 SeqGAN[23] 和 RankGAN[24] 等. 最近, 研究人员提出了其他对抗训练的策略[25−26], 以鼓励中间模型表示各种形式的不变性.

2　本文方法

图 2 是本文提出的基于对抗正则化的自然语言推理模型框图, 图中上半部分的标准 NLI 模型对应本文提出的增强的多层级表示推理模型 (Enhanced multi-level representations inference model, EMRIM), 下半部分的针对词编码的 NLI 对手对应前面提到的 WIM 模型. 其中 EMRIM 模型主要包括词编码器、句编码器、分类器三部分, 该模型通过增强的多层级编码结构探索丰富语言信息. 并且本文提出使用对抗正则化方法降低语言偏置的影响, 从而

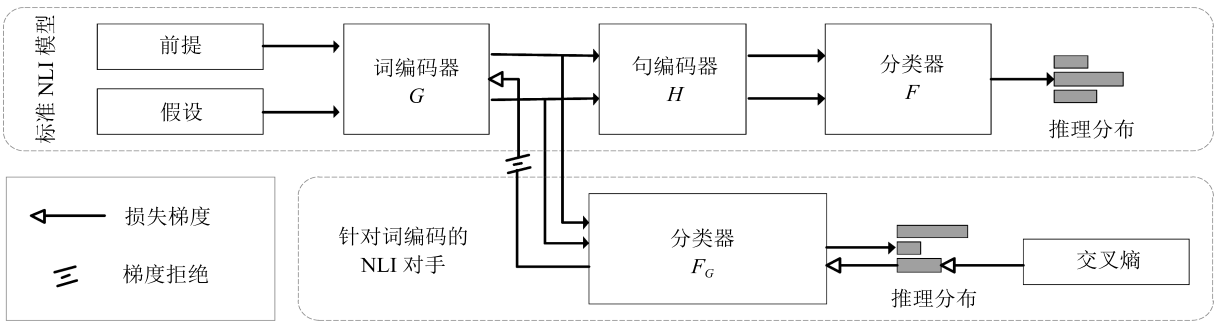

图 2　基于对抗正则化的自然语言推理模型结构框图

Fig. 2　The structure of natural language inference model based on adversarial regularization

进一步提升模型的推理能力. 本文从以下几个方面对提出的方法进行具体描述.

2.1 词编码器

丰富的表示信息在自然语言推理中扮演着重要的角色. 在我们的模型中, 我们将统筹多种类型的表示, 以更好地挖掘前提和假设句义信息, 这也是这项任务的基本组成部分. 首先将前提和假设中的每个单词转换成连续的表示形式, 对词信息进行融合和提取. 图 3 中展示了词编码的处理方式, 具体包含以下部分:

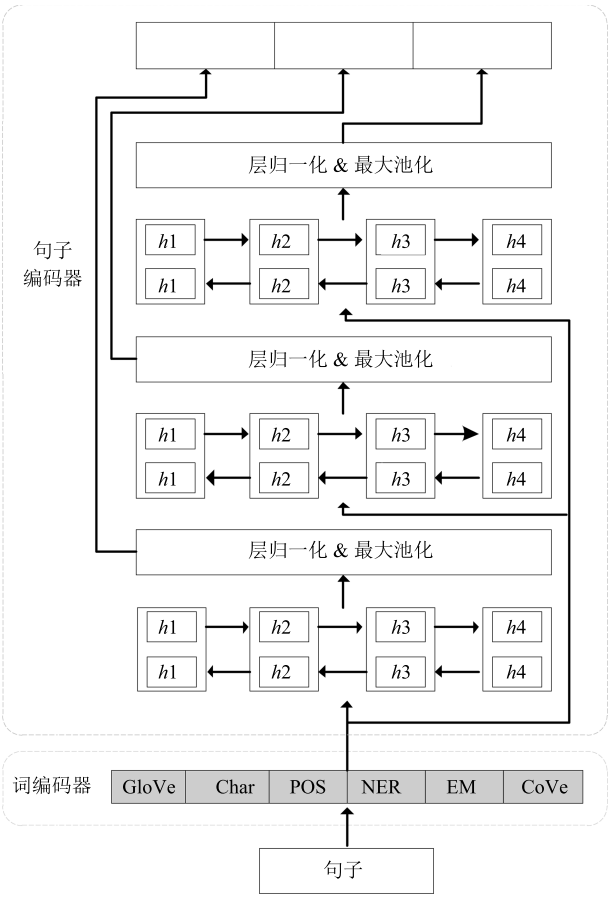

图 3 词编码器和句子编码器网络结构

Fig. 3 Word encoder and sentence encoder network structure

1) 词嵌入: 与之前方法的设置相似, 使用预训练的词向量 GloVe[27] 将每一个单词映射到向量空间.

2) 字符嵌入: 将卷积神经网络 (CNN) 应用到每个单词的字符上. 实践证明, 该方法对处理集外词 (Out of vocabulary, OOV) 有一定的帮助[28].

3) POS 和 NER 标签: 使用词性标注 (Part-of-speech, POS) 和命名实体识别 (Named-entity recognition, NER) 来获得单词的词性信息和实体信息, 然后每一个单词可以通过查表获得对应的 POS 嵌入表示和 NER 嵌入表示. 这种方法比常用的独热码包含更多信息.

4) 精确匹配 (Exact match, EM): 受机器阅读理解的启发, 使用 3 个二进制特征来表示这个词是否能与任何词准确匹配, 分别表示原始形式、小写形式和词干形式.

5) CoVe: 通过机器翻译[29] 得到词的上下文向量表示, 本文的模型对其进行降维处理, 以减少模型的参数量.

本文将前面提到的多种词信息串联起来使用, 这样不仅可以从更多角度获得词相关的表示信息, 而且为后续句子编码提供良好的基础表征, 以更准确地理解句子上下文含义, 从而做出合理的推理.

2.2 句子编码器

为了获得句子的语义信息, 将所有向量序列传递给使用 BiLSTM 和最大池化 (Max pooling) 的句子编码器. 输入一个长度为 T 的序列 $(w_1, w_2, w_3, \cdots, w_T)$, 双向长短时记忆网络的输出是 $(h_1, h_2, h_3, \cdots, h_T)$, 序列输出中的每一项计算如下:

$$\overrightarrow{h_t} = \overrightarrow{LSTM_t}(w_1, w_2, \cdots, w_T) \tag{2}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM_t}(w_1, w_2, \cdots, w_T) \tag{3}$$

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \tag{4}$$

接下来为了学习每个句子的整体表示, 对序列编码器隐藏层的输出应用最大池化处理, 得到与 h_t 同维度大小的向量

$$x = MaxPooling(h_1, h_2, h_3, \cdots, h_T) \tag{5}$$

先进的自然语言推理模型通常将句子编码器实现为多层结构, 鼓励模型模拟复杂函数, 同时捕获复杂的语言结构. 此外, 一些研究人员已经证实, 不同层能够提取不同类型的语法和语义信息[30]. 本文通过设置多层级结构, 探索每一层潜在的语义信息. 在推理模型中, 使用基于 BiLSTM 和 Max Pooling 的层次化句子编码器, 句子编码器包括三层, 每一层 BiLSTM 都是将原始输入语句序列作为输入; 而且, 除了第一层 BiLSTM 之外的其他 BiLSTM 层, 均使用前一层网络的最终状态来初始化其隐层状态. 对每一层 BiLSTM 的输出进行最大池化, 句子编码的最终输出是每一个最大池化层输出的串联拼接. 图 3 显示了具体的网络结构.

2.3 顶层分类器

句子编码器的输出是前提和假设的固定维度的向量表示 u 和 v, 然后将它们传递给顶层分类器. 在

自然语言推理任务中, 顶层分类器一般使用多层感知机 (Multilayer perceptron, MLP) 和 Softmax 函数来预测每个类别的概率. 本文以多种方式将这两个句子的表示聚合在一起, 并作为多层感知机的输入, 然后把多层感知机的输出传递给 Softmax 函数, 公式表示如下所示:

$$x = [u; v; u \odot v; |u - v|] \tag{6}$$

$$Output = \text{Softmax}(MLP(x)) \tag{7}$$

其中, $\odot$ 表示逐个对应元素相乘, 多层感知机包含两个带有修正线性单元 (Rectified linear unit, ReLU) 激活函数的隐层. 最后通过最小化带有 L2 正则项的多类交叉熵损失函数, 对整个模型进行端到端训练.

2.4 对抗正则化方法

1) 标准推理模型: 给定数据集 $D = \{p_i, q_i, a_i\}$, 其中包含前提句 $p_i \in \mathcal{P}$、假设句 $q_i \in \mathcal{Q}$、推理标签 $a_i \in \mathcal{A}$ 三部分, 自然语言推理任务就是从前提和假设句子中推理出它们的逻辑关系. 为了描述方便, 定义词编码器的操作为 G, 定义句子编码器为 H, 最后的分类层为 F, p 和 q 为数据集 D 中某样本的两个句子, 所以我们的推理模型可以表示为, 首先这两个句子通过词编码器分别得到表示 g_u 和 g_v

$$g_u = G(p) \tag{8}$$

$$g_v = G(q) \tag{9}$$

然后输出的结果经过句编码器的处理得到句子表示 u 和 v

$$u = H(g_u) \tag{10}$$

$$v = H(g_v) \tag{11}$$

最后将两者的句子表示传递给顶层分类器预测逻辑关系

$$P(\mathcal{A}|p, q) = F(u, v) \tag{12}$$

现有的自然语言推理模型一般都遵循类似的模式, 通过标准的交叉熵函数进行训练, 通过优化参数最小化损失函数

$$\mathcal{L}_{NLI}(G, H, F) = \mathbb{E}_{\mathcal{P},\mathcal{Q},\mathcal{A}}[-\log(P(a_i|p_i, q_i))] \tag{13}$$

2) WIM: 对 NLI 中关于词的语言偏置强弱直观的度量是模型仅从词编码就可以预测答案的能力. 我们将这个模型形式化为一个映射 F_G, 如上所述, 我们假设 F_G 是可微的, 并把从标准推理模型获得的词编码作为输入, 以便 F_G 可以进行预测

$$P_{F_G}(\mathcal{A}|p, q) = F_G(g_u, g_v) \tag{14}$$

将这个模型参数化为与顶层分类器相似的结构, 只是为了便于后续处理. 在其基础上加入了最大池化层. 如上所述, 该模型可以用交叉熵函数进行训练

$$\mathcal{L}_G(G, F_G) = \mathbb{E}_{\mathcal{P},\mathcal{Q},\mathcal{A}}[-\log(P_{F_G}(a_i|p_i, q_i))] \tag{15}$$

3) 对抗正则化减少语言偏置: 如图 2 所示, 本文将标准推理模型和只依赖词编码的推理模型设置为对抗状态, 引入对抗正则化的方法优化自然语言推理模型. 其中只依赖词编码的推理模型为了推理成功, 需要学习训练数据集中的语言偏置, 但是因为这种语言偏置忽略了句义信息, 导致标准推理模型推理错误. 为了减少语言偏置, 将两个模型设置为对抗状态, 通过修改词编码部分来降低只依赖词编码模型的表现; 同时强化句子编码部分, 以捕获更多上下文信息和语义信息, 从而达到在提升标准推理模型推理表现的同时减少对语言偏置的依赖的目的. 可以将这两个模型的对立关系描述为

$$\min_{G,H,F} \max_{F_G} (\mathcal{L}_{NLI}(G, H, F) - \lambda \mathcal{L}_G(G, F_G)) \tag{16}$$

基于对抗正则化的自然语言推理模型的训练过程如下, 首先训练只依赖词编码的推理模型, 该模型的训练目标是最小化其对应的交叉熵损失函数, 但是词编码器 $G(\cdot)$ 不会根据这个梯度信息更新, 这个操作对应了图 2 中的梯度拒绝部分. 潜在地, 这迫使分类器 F_G 要基于标准推理模型给出的词编码表示尽可能好地进行推理. 然后训练更新标准推理模型, 该模型的梯度信息来自于两部分: 一部分是标准推理模型本身对应的交叉熵损失函数; 另一部分来自于只依赖词编码的推理模型负的加权的交叉熵损失函数, 其中分类器 F_G 的参数是不更新的, 分类器只是起到梯度传递的作用. 最后这两个训练过程进行交替训练更新, 通过不断对抗博弈, 以到达理想的纳什均衡状态.

我们使用正则化系数 λ 来调控推理模型的性能和语言偏置的权衡. λ 取值较小表明较少的正则化发生, 标准推理模型继续学习语言偏置. 另一方面, 当 λ 取值较大时, 表示去除较多语言偏置, 可能导致标准推理模型和只依赖词编码的模型的表现都不好; 此外权重过大会加重对词编码的影响, 以至于词编码器没有能力学习合理的词表示, 从而进一步影响句子表征等高层表示. 所以要设置合适的权重来权衡两者之间的重要性.

3 实验

3.1 数据集

我们在 SNLI 和 Breaking-NLI 数据集上验证

本文的方法.

SNLI (Stanford natural language inference)[3] 数据集大约有 57 万人工标注的句子对, 该数据集比其他同类数据集大两个数量级. 其中前提数据来源于 Flickr30k 语料库中的字幕, 而假设句数据和标签是人工合成的. 数据集提供的标签分别是 "entailment", "neutral", "contradiction", "–". 其中 "–" 表示注释者之间无法达成共识. 遵照 Bowman 等[3] 提出的方式删除标签为 "–" 的句子对, 然后生成训练集、验证集和测试集.

Breaking-NLI[4] 数据集是一个自然语言推理的测试集, 包括 8 193 个前提和假设句子对, 其中前提和假设只是有一个词或短语被替换了, 其他成分是相同的. 该数据集被用来测试自然语言推理模型, 推理模型需要一定的词汇和世界知识才能实现合理的表现.

3.2 实验设置

本文在实验中使用预先训练好的 300 维的 GloVe 840B 词向量来初始化词嵌入向量, 词嵌入中的集外词使用 $[-0.1, 0.1]$ 随机初始化, 在模型训练期间词嵌入向量被不断更新, 以学习适合 NLI 任务的更有效的表示. 我们使用 Spacy 对单词进行标记并生成 POS 和 NER 标签, POS 和 NER 的嵌入维度分别是 26 和 20. 所有 BiLSTM 的隐层大小设置为 250, 为了避免过拟合, 在层间使用 dropout[31] 和层归一化 (Layer normalization)[32] 处理方法. 使用 Adam[33] 算法优化模型参数, 并设置学习率为 0.0001, 权重衰减为 1×10^{-8}. 设置批次大小为 32, 以进行更多探索. 在对抗训练过程中, 两个模型交替训练的频率为 1 : 1. 在所有方法中都是使用 500 维的 BiLSTM (250 维前向 LSTM + 250 维后向 LSTM).

3.3 实验结果

表 2 显示了使用句子嵌入方法的不同模型在 SNLI 训练集和测试集的结果. 我们使用以下几种方法进行实验对比:

1) BiLSTM_MP: 该模型的词编码器使用本文提出的多信息融合编码方式, 但是句编码器使用了简单堆叠的三层 BiLSTM 网络, 并根据最后一层 BiLSTM 的输出进行最大池化处理, 最后经过顶层分类器得到推理结果.

2) BiLSTM_MP + AR: 该方法是在 BiLSTM_MP 基础上使用对抗正则化.

表 2 不同方法在 SNLI 上的实验结果 (%)

Table 2 Experimental results for different methods on SNLI (%)

对比方法	模型	训练准确率	测试准确率
Mou 等[13] (2015)	300D Tree-based CNN encoders	83.3	82.1
Liu 等[12] (2016)	600D (300 + 300) BiLSTM encoders	86.4	83.3
Liu 等[12] (2016)	600D BiLSTM encoders with intra-attention	84.5	84.2
Conneau 等[34] (2017)	4096D BiLSTM with max-pooling	85.6	84.5
Shen 等[6] (2017)	Directional self-attention network encoders	91.1	85.6
Yi 等[7] (2018)	300D CAFE (no cross-sentence attention)	87.3	85.9
Im 等[16] (2017)	Distance-based Self-Attention Network	89.6	86.3
Kim 等[35] (2018)	DRCN (-Attn, -Flag)	91.4	86.5
Talman 等[36] (2018)	600D HBMP	89.9	86.6
Chen 等[37] (2018)	600D BiLSTM with generalized pooling	94.9	86.6
Kiela 等[38] (2018)	512D Dynamic Meta-Embeddings	91.6	86.7
Yoon 等[17] (2018)	600D Dynamic Self-Attention Model	87.3	86.8
Yoon 等[17] (2018)	Multiple-Dynamic Self-Attention Model	89.0	87.4
本文方法	BiLSTM_MP	89.46	86.51
本文方法	EMRIM	92.71	87.36
本文方法	BiLSTM_MP + AR	89.02	86.73
本文方法	EMRIM + AR	93.26	**87.60**

3) EMRIM: 该方法是第 2 节提出的增强的多层级表示推理模型.

4) EMRIM + AR: 在 EMRIM 中加入对抗正则化方法.

表 2 显示了本文实验结果与 SNLI 官方排行榜结果, 根据实验对比, 本文提出的 EMRIM 方法达到了 87.36 % 的准确率, 已经接近排行榜中的最好结果 87.4 %, 这说明在推理模型中使用多种类型信息增强的词编码器和多层级的句编码器, 确实可以提取更丰富更准确的语义表示, 从而利于模型推理. 当分别为标准推理模型 BiLSTM_MP 和 EMRIM 增加只依赖词编码进行推理的对抗模型之后, 在不断博弈的进化过程中, 两个标准模型的推理性能进一步提升, BiLSTM_MP + AR 比 BiLSTM_MP 高出约 0.22 % 的准确率, EMRIM + AR 比 EMRIM 高出约 0.24 % 的准确率. 这表明了本文提出的对抗正则化方法的有效性: 该方法可以减少标准推理模型对语言偏置的依赖, 避免依据词间浅显的关系进行盲目推理; 而是强调语义理解, 通过对句义的整体把握做出选择. 需要注意的是对抗正则方法没有增加标准推理模型的参数量, 并且不会增加模型测试时的复杂度.

表 3 是不同方法在 Breaking-NLI 测试集上的实验结果[4], 这些模型都是在 SNLI 数据集上训练, 然后在 Breaking-NLI 数据集上测试. 实验发现在 SNLI 测试集上表现不错的 ESIM 模型, 在这个测试集上的性能急剧下降. 本文提出的 EMRIM + AR 模型在该测试集上取得了目前公开的最高准确率, 这说明本文提出的模型具有良好的词汇知识和世界知识; 通过应用对抗正则化方法, 推理模型在理解词汇的同时, 关注句义表达, 整体把握推理需求, 做出合理推理.

表 3 不同方法在 Breaking-NLI 上的测试结果

Table 3 Experimental results for different methods on Breaking-NLI

模型	测试准确率 (%)
Decomposable Attention[39]	51.9
Residual-Stacked-Encoder[40]	62.2
ESIM[8]	65.6
KIM[41]	83.5
EMRIM	88.37
EMRIM + AR	**89.96**

注意到在对抗训练过程中, 随着标准推理模型在 SNLI 测试集上的表现提升, 只依赖词编码进行推理的模型的性能上升到一定程度之后不再增加, 而且有稍微下降的趋势. 这表明对抗优化策略执行得很好, 这也是和我们的直觉是一致的.

表 4 是权重 λ 对 SNLI 测试集推理准确率的影响. 根据权重和准确率的变化趋势, 可以得到以下分析. 在较高的权值下, 基于词编码的大部分判别信息都已经丢失, 即标准推理模型是通过牺牲自己的性能, 从而降低了只作用于词编码模型的性能, 但是事实上在推理中根据词信息进行判别还是占有一定重要地位的, 不应完全忽略; 另外, 权重过大也导致模型底层学习不到合理的词向量表示, 继而影响模型高层网络对句子编码能力和推理能力. 在权值较小时, 标准推理模型的性能相较之前也没有明显提升, 毕竟完全根据词中的判别信息进行推理是片面的, 因为忽略了对句子内容的整体理解和把握, 会导致模型的推理脱离了对应的描述场景, 从而难于做出正确抉择. 只有兼顾词中表达的判别信息和句义分析这两方面, 自然语言推理模型才会做出正确的推理.

表 4 权重 λ 对 NLI 准确率的影响

Table 4 Impact of weight λ on NLI accuracy

权重值	测试准确率 (%)
0.5	86.90
0.25	87.14
0.10	87.60
0.05	87.35
0.01	87.39

4 结束语

本文提出增强的多层级表示推理模型, 通过多样信息整合和多层级句子编码, 增强模型对句子的表示能力, 探索更多语义信息. 在标准推理模型中引入对抗正则化方法, 通过标准推理模型和只依赖词编码进行推理的模型进行博弈训练, 以减少语言偏置对推理模型的影响, 使模型能够基于上下文进行有效推理. 在 SNLI 和 Breaking-NLI 数据集上的实验结果验证了本文方法的有效性. 在未来的研究工作中, 我们希望将该方法应用到更多任务中去.

References

1 Guo Mao-Sheng, Zhang Yu, Liu Ting. Research advances and prospect of recognizing textual entailment and knowledge acquisition. *Acta Automatica Sinica*, 2017, **40**(4): 889−910
(郭茂盛, 张宇, 刘挺. 文本蕴含关系识别与知识获取研究进展及展望. 计算机学报, 2017, **40**(4): 889−910)

2 Ren Han, Feng Wen-He, Liu Mao-Fu, Wan Jing. Recognizing textual entailment based on inference phenomena. *Jour-*

nal of Chinese Information Processing, 2017, **31**(1): 184−191
(任函, 冯文贺, 刘茂福, 万菁. 基于语言现象的文本蕴涵识别. 中文信息学报, 2017, **31**(1): 184−191)

3 Bowman S R, Angeli G, Potts C, Manning C D. A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015. 632−642

4 Glockner M, Shwartz V, Goldberg Y. Breaking nli systems with sentences that require simple lexical inferences. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers). Melbourne, Australia: Association for Computational Linguistics, 2018. 650−655

5 Gong Y, Luo H, Zhang J. Natural language inference over interaction space. In: Proceedings of the 6th International Conference on Learning Representations. 2018.

6 Shen T, Zhou T, Long G, Jiang J, Pan S, Zhang C. DiSAN: directional self-attention network for RNN/CNN-free language understanding. In: Proceedings of the 32nd AAAI Conference on Articifical Intelligence. New Orleans, USA: AAAI, 2018. 5446−5445

7 Yi T, Luu A T, Siu C H. Compare, compress and propagate: enhancing neural architectures with alignment factorization for natural language inference. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018. 1565−1575

8 Chen Q, Zhu X, Ling Z H, Wei S, Jiang H, Inkpen D. Enhanced LSTM for natural language inference. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017. 1657−1668

9 Zhang K, Lv G, Wu L, Chen E, Liu Q, Wu H, Wu F. Image-enhanced multi-level sentence representation net for natural language inference. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM). Singapore, Singapore: IEEE, 2018. 747−756

10 Camburu O M, Rocktäschel T, Lukasiewicz T, Blunsom P. e-SNLI: natural language inference with natural language explanations. In: Proceedings of the 32nd Conference on Neural Information Processing Systems. Montréal, Canada: NeurIPS, 2018. 9560−9572

11 Poliak A, Naradowsky J, Haldar A, Rudinger R, Van Durme B. Hypothesis only baselines in natural language inference. In: Proceedings of the 7th Joint Conference on Lexical and Computational Semantics. Vancouver, Canada, 2018. 180−191

12 Liu Y, Sun C, Lin L, Wang X. Learning natural language inference using bidirectional LSTM model and inner-attention. arXiv preprint arXiv: 1605.09090, 2016.

13 Mou L, Men R, Li G, Xu Y, Zhang L, Yan R, et al. Natural language inference by tree-based convolution and heuristic matching. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016. 130

14 Munkhdalai T, Yu H. Neural semantic encoders. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, 2017. 397−407

15 Shen T, Zhou T, Long G, Jiang J, Wang S, Zhang C. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 4345−4352

16 Im J, Cho S. Distance-based self-attention network for natural language inference. arXiv preprint arXiv: 1712.02047, 2017.

17 Yoon D, Lee D, Lee S K. Dynamic self-attention: computing attention over words dynamically for sentence embedding. arXiv preprint arXiv: 1808.07383, 2018.

18 Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 2014 Conference on Advances in Neural Information Processing Systems. Montreal, Canada: Curran Associates, Inc., 2014. 2672−2680

19 Mao X, Li Q, Xie H, Lau R Y, Wang Z, Smolley S P. Least squares generative adversarial networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2813−2821

20 Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: ICML, 2017. 214−233

21 Zhu J Y, Zhang R, Pathak D, Darrell T, Efros A A, Wang O, et al. Toward multimodal image-to-image translation. In: Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, USA: NIPS, 2017. 465−476

22 Yi Z, Zhang H, Tan P, Gong M. Dualgan: unsupervised dual learning for image-to-image translation. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2849−2857

23 Yu L T, Zhang W N, Wang J, Yu Y. Seqgan: sequence generative adversarial nets with policy gradient. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI 2017. 2852−2858

24 Lin K, Li D, He X, Zhang Z, Sun M T. Adversarial ranking for language generation. In: Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, CA, USA: NIPS, 2017. 3155−3165

25 Tzeng E, Hoffman J, Saenko K, Darrell T. Adversarial discriminative domain adaptation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA: IEEE, 2017. 7167−7176

26 Lample G, Zeghidour N, Usunier N, Bordes A, Denoyer L. Fader networks: manipulating images by sliding attributes. In: Proceedings of the 31st Conference in Neural Information Processing Systems. Long Beach, CA, USA: NIPS, 2017. 5967−5976

27 Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014. 1532−1543

28 Yang Z, Dhingra B, Yuan Y, Hu J, Cohen W W, Salakhutdinov R. Words or characters? fine-grained gating for reading comprehension. In: Proceedings of the 5th International Conference on Learning Representations. 2017.

29 McCann B, Bradbury J, Xiong C, Socher R. Learned in translation: contextualized word vectors. In: Proceedings of the 31st Conference in Neural Information Processing Systems. Long Beach, CA, USA: NIPS, 2017. 6294−6305

30 Anastasopoulos A, Chiang D. Tied multitask learning for neural speech translation. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA: Association for Computational Linguistics, 2018. 82−91

31 Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv: 1207.0580, 2012.

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

33 Kingma D P, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv: 1412.6980, 2014.

34 Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017. 670−680

35 Kim S, Hong J H, Kang I, Kwak N. Semantic sentence matching with densely-connected recurrent and co-attentive information. arXiv preprint arXiv: 1805.11360, 2018.

36 Talman A, Yli-Jyrä A, Tiedemann J. Natural language inference with hierarchical bilstm max pooling architecture. arXiv preprint arXiv: 1808.08762, 2018.

37 Chen Q, Ling Z H, Zhu X. Enhancing sentence embedding with generalized pooling. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: The COLING 2018 Organizing Committee, 2018. 1815−1826

38 Kiela D, Wang C, Cho K. Dynamic meta-embeddings for improved sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018. 1466−1477

39 Parikh A, Täckström O, Das D, Uszkoreit J. A decomposable attention model for natural language inference. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA: Association for Computational Linguistics, 2016. 2249−2255

40 Nie Y, Bansal M. Shortcut-stacked sentence encoders for multi-domain inference. In: Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP. Copenhagen, Denmark: Association for Computational Linguistics, 2017. 41−45

41 Chen Q, Zhu X, Ling Z H, Inkpen D, Wei S. Neural natural language inference models enhanced with external knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Computational Linguistics, 2018. 2406−2417

刘广灿 哈尔滨理工大学自动化学院硕士研究生. 主要研究方向为语音分离, 自然语言理解与生成.
E-mail: c1240754278@163.com
(**LIU Guang-Can** Master student at the School of Automation, Harbin University of Science and Technology. His research interest covers speech separation, natural language understand and generation.)

曹 宇 哈尔滨理工大学副教授. 2009 年获得哈尔滨工业大学博士学位. 主要研究方向为模式识别, 机器视觉和机器人. E-mail: cyhit@163.com
(**CAO Yu** Associate professor at Harbin University of Science and Technology. He received his Ph. D. degree from Harbin Institute of Technology in 2009. His research interest covers pattern recognition, machine vision, and robot.)

许家铭 中国科学院自动化研究所副研究员. 主要研究方向为语音处理与听觉注意, 智能问答和对话, 深度学习和强化学习. 本文通信作者.
E-mail: jiaming.xu@ia.ac.cn
(**XU Jia-Ming** Associate professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers speech processing and auditory attention, question and answering and dialog system, deep learning, and reinforcement learning. Corresponding author of this paper.)

徐 波 中国科学院自动化研究所所长, 研究员. 中国科学院脑科学与智能技术卓越创新中心副主任. 长期从事人工智能研究. 主要研究领域包括类脑智能, 类脑认知计算模型, 自然语言处理与理解, 类脑机器人. E-mail: xubo@ia.ac.cn
(**XU Bo** Professor, president of the Institute of the Automation, Chinese Academy of Sciences, and deputy director of the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences. His research interest covers brain-inspired intelligence, brain-inspired cognitive models, natural language processing and understanding, and brain-inspired robotics.)