计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(11):207−214 [doi: 10.15888/j.cnki.csa.008775]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进的 SemBERT 特征重组模型①

龚安靖, 陈红英

(华南师范大学 计算机学院, 广州 510631)
通信作者: 陈红英, E-mail: chenhy01@126.com

摘　要: SemBERT 模型实现了对 BERT 模型的提升, 但存在两个明显的缺陷. 第一, 模型获得向量表示的能力不足. 第二, 没有从特征的种类出发, 直接使用传统特征进行任务分类. 针对这两个缺陷, 提出了一种新的特征重组网络. 该模型在 SemBERT 内部添加自注意力机制, 外接特征重组机制, 得到更好的向量表示并且重新分配特征权重. 实验数据表明新的方法在 MRPC 数据集上比经典的 SemBERT 模型在 $F1$ 值上提高了 1%. 实现在小数据集上的明显提升, 并且超越了大多数优秀模型.

关键词: 特征重组; 向量表示; 自注意力机制; 特征权重; 深度学习; 自然语言处理

引用格式: 龚安靖,陈红英.改进的 SemBERT 特征重组模型.计算机系统应用,2022,31(11):207–214. http://www.c-s-a.org.cn/1003-3254/8775.html

Improved SemBERT-based Feature Reorganization Model

Gong An-Jing, CHEN Hong-Ying

(School of Computer, South China Normal University, Guangzhou 510631, China)

Abstract: Although the SemBERT model is an improved version of the BERT model, it has two obvious defects. One is its poor ability to obtain vector representation. The other is that it directly uses conventional features to classify tasks without considering the category of the features. A new feature reorganization network is proposed to address these two defects. This model adds a self-attention mechanism into the SemBERT model and obtains better vector representation with an external feature reorganization mechanism. Feature weights are also reassigned. Experimental data show that the $F1$ score of the new method on the Microsoft Research Paraphrase Corpus (MRPC) dataset is one percentage point higher than that of the classical SemBERT model. The proposed model has significantly improved performance on small datasets, and it outperforms most of the current outstanding models.

Key words: feature reorganization; vector representation; self-attention mechanism; feature weight; deep learning; natural language processing (NLP)

自然语言处理任务的关键就是学习序列的表示, 因此前人不断的改进模型以提升向量表示的能力. 早期的词袋模型, 例如使用 one hot 表示单词, 这种方法会形成稀疏矩阵, 过于低效. Mikolov 等[1] 提出了 Word-2Vec 模型, 通过无监督学习得到了词向量, 这是一种预训练模型, 提升了词向量的表示精度和效率. Shi 等[2] 提出双向 LSTM 网络, 提高模型捕捉上下文信息的能力, 得到更加精准的词向量表示. 但这种表示有明显的缺陷, 当句子的长度超出一个量级时, LSTM 会出现梯度消失的问题, 并且文本的局部信息表达能力有限. Mnih 等[3] 提出了在传统循环神经网络中加入注意力机制, 这使得模型可以更好地捕捉局部信息, 突出局部特征, 学会更好的向量表示. 虽然使用双向 LSTM 得到了前后文的信息, 但是严格意义来说, 时间片 t 的计算

① 基金项目: 广东省自然科学基金 (2020A1515010445)
收稿时间: 2022-02-20; 修改时间: 2022-03-21; 采用时间: 2022-03-30; csa 在线出版时间: 2022-07-07

© 中国科学院软件研究所 http://www.c-s-a.org.cn

依赖 t–1 时刻的计算结果, 而无法获取 t 后的计算结果, 所以模型得到的词向量表示并没有包括完整的上下文信息.

为解决这个问题, Vaswani 等[4] 提出了 Transformer 模型, 并提出了自注意力机制. Transformer 通过自注意力机制, 实现了真正意义上的获取前后文信息. 自注意力机制通过设置 Q、K、V 三种向量, 实现每个单词对一个句子所有单词的信息采集, 得到包含完整上下文信息的词向量. Devlin 等[5] 提出的 BERT 模型正是基于 Transformer 编码部分改进得到, 并且是现在大多数模型的核心编码器.

虽然 BERT 模型取得了一系列的成果, 但是其编码时只考虑语境信息, 缺少结构化语义信息, 限制了模型的表示能力. 语义角色标注技术提供一种句子级语义分析的方法, 如格文法, 可以得到句法结构与语义之间的关系, 为模型添加了更丰富的语义信息. FrameNet[6] 和 PropBank[7] 提出形式化的语义框架, 可以实现将语义关系表现为谓词-参数结构, 并且解析谓词-参数结构, 是现在常用的语义角色标注框架.

Ganin 等[8] 引入了梯度反转层 (GRL) 来获取公共特征. 公共特征是指在语义信息中, 对于分类是没有明确指示作用的, 例如: 电脑、手机、音响这些词在情感分类中, 起不到明确指示作用. 而纯净特征是相对于公共特征的概念, 如果信息对于分类标签有着很明确的指示作用, 如: 完美、开心表达了积极的分类语义, 糟糕、难过表达了消极的分类语义, 这类特征称之为纯净特征. 通过公共特征实现将域适应嵌入到学习表示的过程中, 使最终的分类决策对域的变化具有更强的鉴别性和不变性, 而不受两个域之间位移的影响.

BERT 模型出现后, 自然语言处理任务的研究进入了一个新的阶段. 近两年, 许多以 BERT 为基础的优秀模型被提出. SemBERT 模型[9] 将语义角色标注方法与 BERT 模型相结合, 增加了显式语义结构, 获得了更丰富的语义信息. ELECTRA[10] 提出将 BERT 模型中 mask 标记换成替代词, 这样做既解决了预训练和微调之间的不匹配问题, 又解决了每批次训练数据中只有 15% 的标记被预测导致数据利用率不足的问题. XLNet[11] 提出使用自回归模型代替自编码模型, 使用双流注意力机制, 并且引入了 transformer-xl 机制, 避免了 BERT 模型中 mask 方法带来的弊端. LIMIT-BERT[12] 提出了语言学信息的多任务 BERT, 采用半监督学习策略, 提供与语言模型训练相同的大量语言任务数据, 不仅提高了语言学任务的性能, 而且还受益于正则化效应和一般表征的语言学信息, 实现提升适应新的任务和领域的能力. SAIN[13] 的推理模块将显式语境化语义作为补充输入, 通过注意机制实现对语义线索的一系列推理, 提高了模型的理解性能. 除此之外, Qin 等 [14] 提出了使用纯净特征完成分类的方法. 以 BERT 为特征提取器并借助梯度反转层, 提取了公共特征, 然后在公共特征上做映射, 获得纯净特征, 并送入分类层完成任务.

然而这些方法大部分都只是提升了模型对文本内容的理解, 提高了模型的向量表示能力, 但是却没有从特征的种类出发, 而是直接使用传统特征或者纯净特征进行任务分类, 没有根据实际任务改变纯净特征和公共特征的权重以实现特征重组, 从而得到更合适的特征向量.

因此, 本文基于 SemBERT 模型, 使用两种思路构建模型. 一方面, 增加自注意力机制来提升模型向量表示的能力. 另一方面, 添加特征重组层, 提取公共特征和纯净特征, 再通过注意力机制改变共同特征与纯净特征的权重, 完成对模型的构建.

1 模型方法

图 1 描述了模型框架. 由特征提取部分和特征重组部分组成. 图 2 所示的 P-net 和 C-net 是特征提取部分, 内部结构完全一样 (讲解模型时将以 P-net 为例), 用于提取传统特征 (包含了公共特征和纯净特征). 特征重组部分则由梯度反转层 (GRL), 特征映射层 (OPL), 简单注意力机制 (Simple_Attention) 组成. 梯度反转层用于提取公共特征, 特征映射层用于提取纯净特征, 简单注意力机制用于重组共同特征和纯净特征.

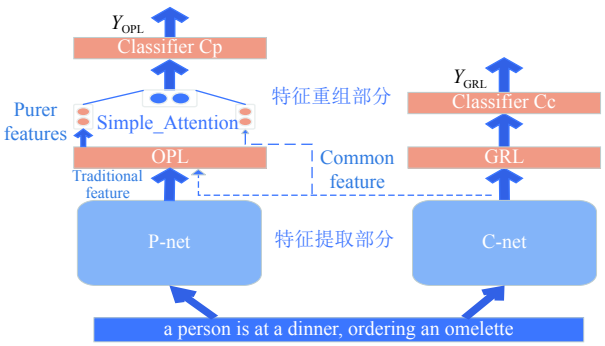

图 1 改进的 SemBERT 特征重组模型

© 中国科学院软件研究所 http://www.c-s-a.org.cn

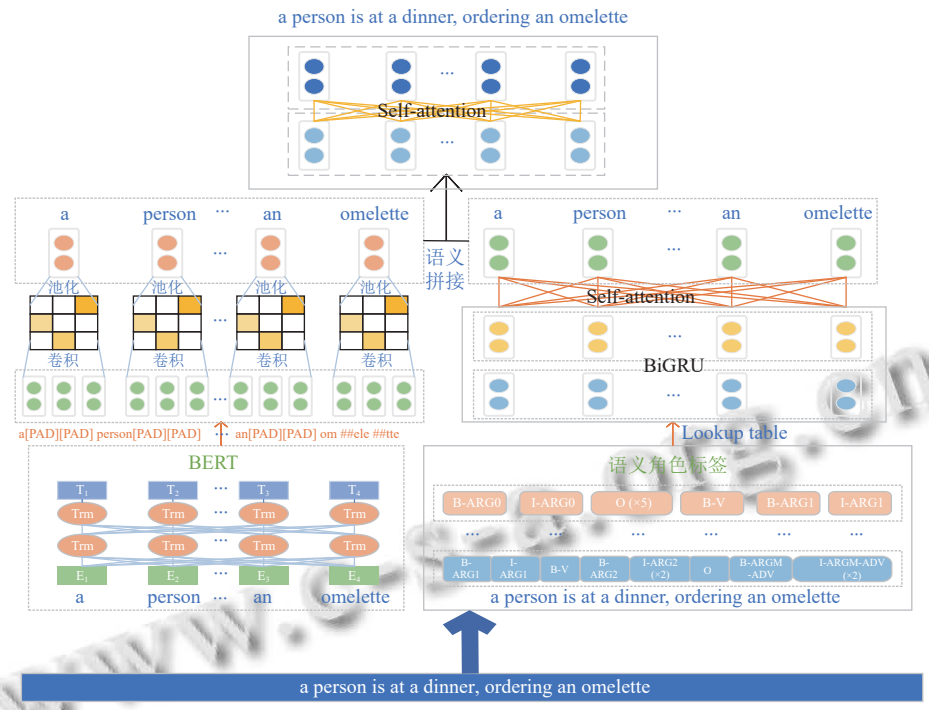

图 2 特征提取器 (P-net/C-net)

P-net 可以处理多个序列输入, BERT 将序列进行分词处理, 得到向量表示, 然后再通过卷积神经网络卷积池化合并成原始词, 得到左边的向量表示 (为了去对齐语义标签部分).

与此同时, 右边部分也将通过语义角色标注在句子中找到谓词, 然后进行参数标记, 得到语义标签. 接着将标注的序列嵌入到向量表示, 经过 BiGRU 层后, 进行一个自注意力机制 (这也是我们与 SemBERT 模型的重要区别), 获得一个更好的向量表示, 最后与左边部分的向量表示进行拼接, 将拼接的结果进行自注意力机制, 形成了联合表示. 再将联合表示送入到模型的特征重组部分, 进入到 GRL 和 OPL 后, 得到公共特征和纯净特征, 再将纯净特征与公共特征进行简单自注意力, 得到重组后的特征向量, 也就完成了下游任务的最终表示. 其整体更新过程如算法 1 所示.

算法 1. 特征重组模型算法

输入: D: 包含 n 个句子对的数据集
输出: PD (释义|D)

1) 通过 get_train/dev/test/label _ex-amples() 函数, 获得训练集, 验证集, 测试集和标签集;
2) 通过 convert_examples_to_features() 和 transform_tag_features() 函数, 加载数据, 并且进行数据预处理;
3) **For** each iteration b=1, 2, ⋯, M **do**;
4) **C-net 部分:**
5) 将数据送入模型特征提取器;
6) 使用 BERT 加载每个分词对应的 embedding, 得到嵌入矩阵 X_i;
7) 用 padding 将每个词的分词填充到相同的长度;
8) 使用卷积+池化将分词合并成原始词 $e(x_i)$;
9) 将预训练好的语义映射为向量 V_i;
10) 将向量 V_i 送入 BiGRU 层和 self-attention 层得到 $e(v_i)$;
11) 将 $e(x_i)$ 和 $e(v_i)$ 两边数据进行拼接, 然后送入 self-attention 层得到传统特征向量 H_i;
12) 将特征向量 $H1_i$ (第 1 次为传统特征, 之后为公共特征) 经过 GRL;
13) 执行分类;
14) **P-net 部分:**
15) 重复步骤 5)–步骤 11) 得到传统特征向量 H_i;
16) 使用传统特征 H_i 和公共特征 $H1_i$ 做特征映射, 得到纯净特征 $H2_i$;
17) 将纯净特征 $H2_i$ 和公共特征 $H1_i$ 送入到 Simple_Attention 层进行特征重组得到重组后的向量 $final_v_i$;
18) 执行分类;
19) C-net, P-net 参数更新;
20) **End for**

1.1 语义角色标签

在数据预处理过程中, 使用预先训练的语义标记器将句子的每个单词注释上标签. 对于一个特定的句子, 会有多个谓词-参数结构. 如图 3 所示, 对于序列“a person is at a diner, ordering an omelette”, 会被语义标记器, 增添两组语义角色标签. 这两组语义角色标签最后

© 中国科学院软件研究所 http://www.c-s-a.org.cn

会随着编码层一起进行向量表示, 丰富语义, 提供更多的文本信息.

1.2 编码

P-net 的 BERT 部分将原始文本序列送到预先训练的 BERT 模型. 输入句子 $X = \{x_1, \cdots, x_n\}$ 是长度为 n 的单词序列, 它首先被切成很多个子词. 然后, BERT 内部的 Transformer 通过自注意力机制获取子词的上下文信息, 得到每个子词的词向量表示.

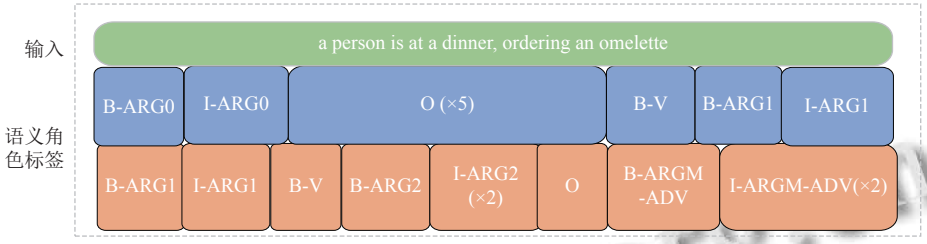

图 3 语义角色标签示例

语义标签部分, 每个谓词有 m 组标注序列, $T = \{t_1, \cdots, t_m\}$, 其中第 i 组标注序列 t_i 包含 n 个标签, 表示为 $\{label^i_1, label^i_2, \cdots, label^i_n\}$. 标签是单词级的, 而不是分词级, 所以长度等于句子 X 的原始长度 n. 将语义信息进行嵌入, 将这些标签映射到向量 $\{v^i_1, v^i_2, \cdots, v^i_n\}$, 并送入到 BiGRU 层获得 m 组标签序列在空间中的标签表示:

$$e(t_i) = BiGRU(v^i{}_1, v^i{}_2, \cdots, v^i{}_n) \tag{1}$$

其中, $0 < i \leqslant m$. 再对 $e(t_i)$ 使用自注意力机制获得更好的表示 $e_a(t_i)$:

$$e_a(t_i) = Attention(e(t_i)) \tag{2}$$

对于 m 组标签, 让 L 代表单词 X_i, $e(L_i) = \{e_a(t_1), \cdots, e_a(t_m)\}$, 将 m 组标签的得到的结果进行拼接, 并将它们送到全连接层, 以获得更加精简的表示 $e'(L_i)$:

$$e'(L_i) = W_1 \cdot e(L_i) + b_1 \tag{3}$$

其中, W_1 和 b_1 都是可训练的参数, 整个序列的表示则为 e^t:

$$e^t = \{e'(L_1), \cdots, e'(L_n)\} \tag{4}$$

1.3 整合

当序列经过了左边部分 BERT 层的编码和右边部分语义的显式标注、BiGRU 层与自注意力层的信息提取, 接着需要将两边的信息进行整合. 由于 BERT 模型基于子词的训练, 而语义标注模块是基于原始词, 所以两边的序列长度不同, 需要一个对齐的过程. 如果将语义标注的序列进行拆分与左边对齐, 将会失去意义, 例如像标注的逻辑主体, 开始的位置等这种信息进行拆分, 不符合逻辑. 所以选择把子词合并成原始词的方法. 将每个单词的子单词进行分组, 并且使用卷积加池化的方式来获得原始词的表示. 以一个词为例, 假设单词 X_i 由一系列子词 $[s_1, s_2, \cdots, s_l]$ 组成. 其中, l 为单词 X_i 的子单词数.

(1) 将来自 BERT 的子词 s_j 表示为 $e(s_j)$, 使用一个 Conv1D 层:

$$e'_i = W_2[e(s_i), e(s_{i+1}), \cdots, e(s_{i+k-1})] + b_2 \tag{5}$$

其中, W_2 和 b_2 是可训练参数, k 是核大小.

(2) 将 ReLU 和 max pooling 应用到 x_i 的输出嵌入序列中:

$$e''_i = ReLU(e'_i) \tag{6}$$

$$e(x_i) = MaxPooling(e''_1, \cdots, e''_{l-k+1}) \tag{7}$$

经过上述操作, 序列 X 的表示为 $e^w = \{e(x_1), \cdots, e(x_n)\} \in R^{n \times dw}$, dw 表示词向量维度. 实现了把子词合并成原始词, 完成了与语义标注部分的对齐.

(3) 将左右部分连接起来得到向量 $h = e^w || e^t$, 并将 h 送入到自注意力机制中:

$$H = Attention(h) \tag{8}$$

最后, 将得到的向量 H 作为文本信息的传统特征向量表示送入到特征重组机制中.

1.4 特征重组

把 P-net 模型得到的 H 称为 H_1, C-net 得到的 H 称为 H_2, 两个向量得到的过程是一样的. 现在需要将 H_2 送入到 GRL 层. 这部分网络便是通过梯度反转层来学习公共特征.

GRL 是在梯度反向传播的过程中让梯度反转, 求导值添加负号, 让它不向最优解靠拢. 之所以这样做是

© 中国科学院软件研究所 http://www.c-s-a.org.cn

因为, 公共特征通常是分类器无法区分的特征, 其提取过程如下:

(1) 通过数学公式, 梯度反向层被两个不相容的方程定义, 式 (9) 和式 (10) 描述了其正向和反向传播行为:

$$GRL(x) = x \tag{9}$$

$$\frac{\partial GRL(x)}{\partial x} = -\lambda I \tag{10}$$

其中, x 为向量表示, I 为梯度, λ 为超参数.

(2) 将特征向量 H_2 通过 GRL 层处理后, 得到 $GRL(H_2) = \widetilde{H}_2$. 然后再将 $\widetilde{H}_2$ 送到分类器中:

$$Y_{\text{GRL}} = Softmax(\widetilde{H}_2 \cdot W_3 + b_3) \tag{11}$$

$$Loss_c = -\begin{pmatrix} Y_{\text{truth}} \cdot \log(Y_{\text{GRL}}) + \\ (1 - Y_{\text{truth}}) \cdot \log(1 - Y_{\text{GRL}}) \end{pmatrix} \tag{12}$$

其中, W_3, b_3 为权重, Y_{truth} 是样本标签, 式 (12) 描述了预测值 $f(x)$ 与真实值 Y 的不一致程度. 分类器进行分类后, 再通过反向传播和优化函数得到公共特征的向量表示.

当完成了右边的公共特征提取, 将 $\widetilde{H}_2$ 送到特征映射层完成纯净特征的提取. 简单的思路就是, 将 H_1 投影到 $\widetilde{H}_2$ 的正交方向上. 与公共特征正交的特征空间是完全纯净的, 借助特征映射层 (OPL) 来完成这一目标:

(1) 将 H_1 在 $\widetilde{H}_2$ 上做投影, 得到了 H_1':

$$H_1' = Proj(H_1, \widetilde{H}_2) \tag{13}$$

其中, $Proj$ 是一个投影公式:

$$Proj(x, y) = \frac{x \cdot y}{|y|} \cdot \frac{y}{|y|} \tag{14}$$

其中, x, y 都是向量.

(2) 当完成了 H_1 在 $\widetilde{H}_2$ 上的投影得到 H_1' 后, 由数学知识可得到纯净的分类向量 $\widetilde{H}_1$:

$$\widetilde{H}_1 = H_1 - H_1' \tag{15}$$

最后使用一个简单的注意力机制将公共特征和纯净特征重组.

(1) 将纯净特征向量和公共特征向量分别送入到线性层, 各自得到一个权重:

$$w_p = (\widetilde{H}_1 \cdot W_4 + b_4) \tag{16}$$

$$w_c = (\widetilde{H}_2 \cdot W_5 + b_5) \tag{17}$$

其中, W_4、b_4 、W_5、b_5 是可训练参数

(2) 把两个权重拼接, 然后 $Softmax$, 于是就得到了两个向量各自对应的权重:

$$w = Softmax([w_p, w_c]) \tag{18}$$

(3) 将对应的权重和向量相乘, 得到了重新分配后的纯净特征向量和公共特征向量:

$$att_p = w[0] \cdot \widetilde{H}_1 \tag{19}$$

$$att_c = w[1] \cdot \widetilde{H}_2 \tag{20}$$

(4) 最后把它们相加得到了重组后的向量:

$$final = att_p + att_c \tag{21}$$

(5) 得到这重组后的向量再送入分类器中:

$$Y_{\text{OPL}} = Softmax(final \cdot W_6 + b_6) \tag{22}$$

$$Loss_p = -(Y_{\text{truth}} \cdot \log(Y_{\text{OPL}}) + (1 - Y_{\text{truth}}) \cdot \log(1 - Y_{\text{OPL}})) \tag{23}$$

其中, W_6 和 b_6 是可训练参数, Y_{truth} 是样本标签. 式 (23) 述了预测值 $f(x)$ 与真实值 Y 的不一致程度.

2 实验

2.1 实验设置

实验基于 SemBERT 的 PyTorch 框架实现. 其中, 使用的是基础型 BERT, 将预训练好的 BERT 权值, 送入下游任务进行微调. 学习率初始值设置为 1E–5, 预热学习率为 0.1, L2 正则化为 0.01. 批量大小设置为 24. 设置的 epoch 值为 10. 语义标签的组数, 设置为 3. SRL 的向量维度为 64. 使用这些参数, 获得了最好的实验结果.

2.2 数据集

数据集选取 Glue[15] 的微软研究院释义语料库 (Microsoft research paraphrase corpus, MRPC), 相似性和释义任务, 是从在线新闻源中自动抽取句子对, 并人工注释句子对中的句子是否在语义上等效. 类别并不平衡, 其中 68% 的正样本, 使用 $F1$ 值 (平衡 F 分数) 为评价准则, 其中 $F1$ 的定义如下所示:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{24}$$

其中, $precision$ 是精准率, $recall$ 是召回率, 定义如下:

$$precision = \frac{TP}{TP+FP} \tag{25}$$

$$recall = \frac{TP}{TP+FN} \tag{26}$$

其中, TP 表示预测为 1, 实际为 1, 预测正确; FP 表示

© 中国科学院软件研究所 http://www.c-s-a.org.cn

预测为 1, 实际为 0, 预测错误; *FN* 表示预测为 0, 实际为 1, 预测错误.

该数据集的样本个数: 训练集 3 668 个, 开发集 408 个, 测试集 1 725 个.

2.3 实验结果

表 1 显示了实验结果, 在 MRPC 数据集上, 本文模型比 SemBERT 模型在 *F*1 值上提升了 1%, 从 88.2% 提升到了 89.2%, 这个结果也超过了之前的优秀模型. 因此可以说明, 在无法获得大量数据的时候, 可以使用两种方法提升模型的能力: 第一, 通过自注意力机制获得更丰富的语义和更好的向量表示; 第二, 通过特征重组机制, 修改纯净特征与公共特征的权重, 让得到的数据特征表示更加准确.

表 1 数据集的基准测试结果和本文模型的数据结果 (结果来自于 GLUE 官网) (%)

模型	MRPC (*F*1)
BiLSTM+ELMo+Attn	84.9
GPT[16]	82.3
GPT on STILTs	87.7
BERT	88.9
SemBERT	88.2
本文	89.2

2.4 消融实验和分析

为了分别研究这两个方法带来的效果, 进行了消融实验, 来验证这两个方法是否都能提高模型的性能. 除此之外, 进行一组直接使用纯净特征完成任务的实验, 因为纯净特征直接表达了句子感情, 例如, 我很开心, 用“开心”就可以表达整个句子的感情, 所以可能会让任务完成的更高效. 作为对比, 再添加一个直接使用共同特征完成任务的实验. 消融实验使用的仍然是 MRPC 数据集, 结果如表 2 所示. 为了达到更加直观的效果, 所有实验的参数 (包括完整模型) 都使用了相同的大小.

表 2 消融实验结果 (%)

模型	MRPC (*F*1)
SemBERT+ self-attention +Project	88.3
SemBERT+ self-attention +Common	79.9
SemBERT+self-attention	88.7
SemBERT+ Restructuring Feature	88.6
本文模型	89.2

首先, 去掉了简单注意力机制部分, 直接使用纯净特征来完成任务, 得到的 *F*1 值为 82.3%, 仅提升了 0.1%. 这与猜想不一致, 按照之前的分析, 纯净特征是可以取得不错的效果. 但是忽略了如果只使用纯净特征, 可能会使得句子不完整, 语义不清. 例如, 我开心地哭了. 直接使用纯净特征, 就剩下“开心, 哭”, 这两个词, 让模型无法区分情感. 所以尽管共同特征不重要, 却是不可缺少的一部分, 否则会使得句子失去完整语义, 让模型得不到很好的提升.

然后, 也尝试直接使用公共特征来完成任务, 得到的实验结果 *F*1 值仅为 79.9%, 比基础模型 SemBERT 低了 8.3%, 比本文模型低了 9.3%, 这个结果和本文猜想一致, 所有类别都有的特征无法起到好的指示作用, 例如, 情感分类中的“你我他”, 这些词是很难起到很好的指示作用的. 因此根据这两个消融实验, 就说明了使用简单的注意力机制将纯净特征和公共特征重新组合的必要性, 也说明了模型的合理性.

其次, 去掉了特征重组部分, 使用了 SemBERT+ self-attention 组合, 这里没有拆分两个自注意力机制, 而是把他们看做一个整体. 因为, 这两个自注意力机制都是用来获得更好的向量表示, 都属于第一个方法. 通过实验的结果, 自注意力机制的确能够提高模型的性能, 在 MRPC 数据集上提升了 0.5%. 这个结果分析可以得到两个结论, 第一, 自注意力机制起到作用, 经过之前的分析, 能够确定自注意力机制帮助模型更好地获得向量表示, 更好地获得文本信息; 第二, 更好的向量表示可以帮助模型提升对文本的理解, 帮助提高模型的性能.

接着, 保留了特征重组的部分, 去除了自注意力机制, 研究改变纯净特征和公共特征的权重比, 对模型产生的影响. SemBERT + Restructuring feature 组合在 MRPC 上取得了 88.6% 的 *F*1 值, 提升了 0.4%. 这个结果分析可以得到 3 个结论, 第一, 特征重组机制起到了积极作用. 第二, GRL 和 OPL 能够提取共同特征、纯净特征, 简单的注意力机制也能实现改变纯净特征和共同特征的权重比, 并且帮助模型实现重组特征得到更合适的特征向量; 第三, 纯净特征与公共特征对任务起到的作用应该是不同的, 改变纯净特征和共同特征的比重确实能提高模型的性能.

最后, 将 5 个模型的训练损失值, 验证损失值, 还有验证集的 *F*1 结果画成了折线图进行比较, 如图 4–图 6 所示. 重点关注横坐标 epoch_2、epoch_3、epoch_4 对应的结果, 因为那是几个模型取得最好结果的 epoch 值, 可以发现完整模型的 3 个结果均不是最好的

© 中国科学院软件研究所 http://www.c-s-a.org.cn

那个, 尤其是在训练损失值是最高的. 虽然参数最多, 但能在测试集得到最好的结果, 这说明完整模型最不容易过拟合, 可以间接体现模型的有效参数更多, 这也体现了模型的优越性.

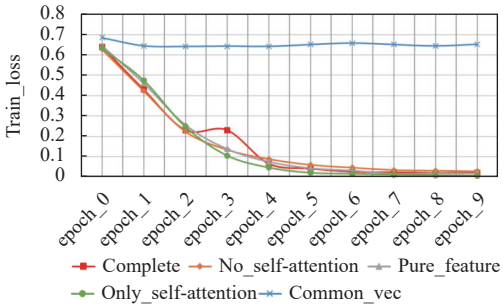

图 4 消融实验 5 个模型的训练损失值变化曲线

将部分模型与完整模型进行对比, 明显发现这两个组件叠加在一起会产生更好的效果. 通过消融实验也能证明, 本文提升模型的两种方法的正确性, 并且自注意力机制和特征重组也能起到优化向量表示和改变特征比重的作用.

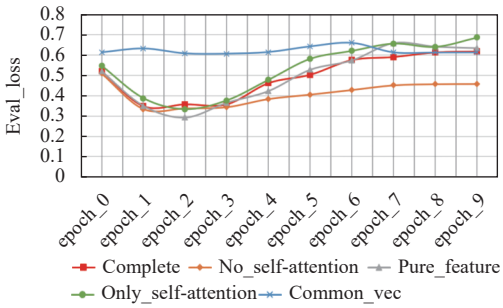

图 5 消融实验 5 个模型的验证损失值变化曲线

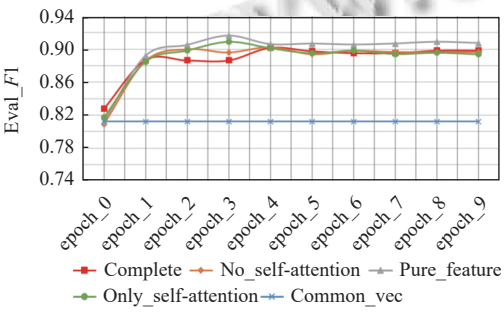

图 6 消融实验 5 个模型的验证集结果 $F1$ 值变化曲线

3 结论

本文提出了一种新的特征重组网络. 该方法从更好地获取语义信息、向量表示和重新分配特征权重出发, 在 SemBERT 内部添加自注意力机制, 外接特征重组机制, 实现了在 MRPC 数据集上对 SemBERT 的提高. 这证明了方法的正确性, 自注意力机制和特征重组机制的有效性. 未来的研究将继续从两个角度出发进行改进, 在特征重组方面, 目前使用改进的 SemBERT+GRL 提取公共特征, 会使得模型过于复杂, 参数量过大, 所以尝试寻求一个新的特征提取器, 让模型可以更加简洁高效的提取公共特征. 除此之外, 尝试去改善特征重组的方法, 现在使用的是一个简单的注意力机制, 下一步希望找到更好的方法去重组公共特征和纯净特征, 使得重组的特征更加有效. 从语义信息的角度出发, 希望能够进一步优化特征提取器, 让模型获得更好的向量的表示, 提取到更丰富的传统特征.

参考文献

1 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Proceedings of the 1st International Conference on Learning Representations. Scottsdale, 2013.

2 Shi XJ, Chen ZR, Wang H, *et al*. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 802–810.

3 Mnih V, Heess N, Graves A, *et al*. Recurrent models of visual attention. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2204–2212.

4 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.

5 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.

6 Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet project. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Montreal: ACL, 1998. 86–90.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

7 Palmer M, Gildea D, Kingsbury P. The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, 2005, 31(1): 71–106. [doi: 10.1162/0891201053630264]

8 Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR, 2015. 1180–1189.

9 Zhang ZS, Wu YW, Zhao H, *et al*. Semantics-aware BERT for language understanding. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020. 9628–9635.

10 Clark K, Luong MT, Le QV, *et al*. ELECTRA: Pre-training text encoders as discriminators rather than generators. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.

11 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, 2019. 517.

12 Zhou JR, Zhang ZS, Zhao H, *et al*. LIMIT-BERT: Linguistics informed multi-task BERT. Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020. ACL, 2020. 4450–4461.

13 Zhang SL, Zhao H, Zhou JR, *et al*. Semantics-aware inferential network for natural language understanding. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020. 14437–14445.

14 Qin Q, Hu WP, Liu B. Feature projection for improved text classification. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 8161–8171.

15 Wang A, Singh A, Michael J, *et al*. GLUE: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of the 7th International Conference on Learning Representations. New Orleans: OpenReview.net, 2019.

16 Radford A, Narasimhan K, Salimans T, *et al*. Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. 2018.

(校对责编: 牛欣悦)

www.c-s-a.org.cn

© 中国科学院软件研究所 http://www.c-s-a.org.cn