用于语言理解的语义感知 BERT 模型

摘要

本文介绍了一个增强了语义感知的 BERT 模型，它通过将 BERT 模型与语义角色标注进行结合得到具有显式语义表示的微调 BERT，其在自然语言理解下游的自然语言推理、阅读理解等多个任务上的表现得到了提升。对 SemBERT 的模型结构进行了简单的介绍，概述了语义角色标注在 NLU 问题中起到的作用，并最终通过实验展示了引入的显式语义表示模块在几个基准数据集上的表现。

关键词：自然语言理解；自然语言推理；BERT；语义角色标注

1 引言

BERT 模型是一个基于 transformer 的预训练模型，能够用于预测句子中被掩盖的词或者预测句子的上下文关系。训练好的 BERT 模型能够用用于各种自然语言理解的下游任务。BERT 模型采用了简单的上下文特征来对目标进行表示和训练，但是很少在模型中考虑明确的上下文语义线索，这是 BERT 模型的主要局限性。SemBERT 的模型提出者 [11] 观察到之前模型产生的答案在语义上是不完整的，这表明之前的模型在 NLU 问题上存在上下文语义表示不足的问题。

自然语言理解任务（natural language understanding）需要对语言进行全面的理解并在此基础之上进行进一步的推理。自然语言理解任务包含了文本理解、关键词抽取、情感分析和语义角色标注等多个核心任务。自然语言推理研究的一个共同趋势是模型变得越来越复杂，使用堆叠的注意力机制和大量的语料库 [12]。

语义角色标注（semantic role labeling）以句子的谓词为中心，分析句子的各个成分与谓词之间的关系，并使用语义角色来表述其结构关系。自然语言理解任务与上下文语义分析的任务目的是相似的。语义角色标注能够发现句子的核心含义是何时何地何人做了何事，这与自然语言理解的任务目的是相匹配的。将语义角色标注与 BERT 模型进行结合能使 BERT 模型在自然语言理解任务上取得更好的语义表示，从而提升在自然语言理解任务上的性能。

在自然语言中，一个句子通常包含了多个语义结构。SemBERT 以细粒度的方式表示学习，包含了显式的语义表示，实现了更深的表示含义。SemBERT 模型在 BERT 模型的基础之上加入了明确的语境语义线索，其包含了三个部分：语义角色标注器、序列编码器和语义集成组件。SemBERT 相对 BERT 模型在 GLUE 基准和 SQuQD2.0 上都有着更好的性能。

2 相关工作

2.1 自然语言推理模型

由于能够从大量的未标注文本中获取单词的本地共现，分布式表示是用于自然语言处理模型的标准部分 [5]。但是，这些学习词向量的方法只涉及每个单词的单一的、独立于上下文的表示，但是忽略了句子级别的上下文编码。在最近的上下文语言模型，包括 ELMo、GPT、BERT 和 XLNet，它们通过加强上下文句子建模来填补这方面的空白。BERT 模型还引入了预测下一个句子的任务。

这些新的语言模型于传统的嵌入方法的主要技术改进是，它们专注于从语言模型中提取上下文敏感的特征。ELMo 有助于提升几个主要的 NLP 基准，包括 SQuAD 上的问答、情感分析 [9] 和命名实体识别 [8]。另一方面，BERT 在 GLUE 基准、MultiNLI 以及 SQuAD 上的语言理解任务上尤其有效 [2]。因此，SemBERT 模型使用预训练的 BERT 模型作为骨干编码器，遵循了这条提取上下文敏感特征的路线。

2.2 显式的上下文语义

在这篇文章发布之前的预训练上下文语言模型已经有一定程度上语言意义的加强了 [1]。但是根据文章作者的对 BERT 在 SQuAD 上语义不完整答案跨度的观察，BERT 模型使用的隐式语义可能不足以支持自然语言理解任务要求的强大上下文表示，所以模型引入了显式的语义表示。在语义角色标注中，语义关系通常被表示为谓词-参数结构。He 等人 [3] 提出了一个具有约束解码的深度网络 BiLSTM 架构，SemBERT 使用该架构作为基本的语义角色标注器，从而将语义角色标注集成到自然语言理解任务当中。

3 本文方法

3.1 模型架构

SemBERT 模型包含了 3 个结构。图 1概述了模型的框架。输入序列通过 BERT 字段标记器切分为了子词，然后通过卷积层将子词表示转换为单词级别，获得上下文单词表示。同时，输入序列也被传递给语义角色标注器来获得有多个明确语义的谓词派生结构。最后将上下文表示和语义嵌入进行串联得到了联合表示。

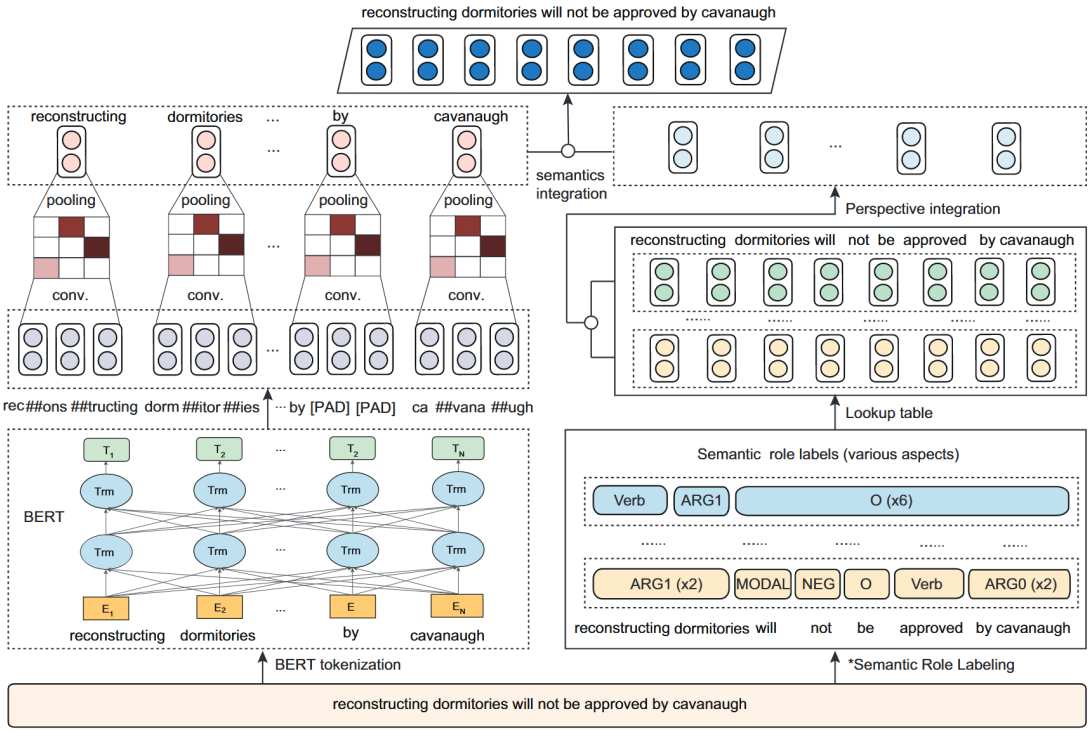

图 1. SemBERT 模型框架

3.2 语义标注器

模型使用了预先训练的语义标注器，架构句子标注为若干个语义序列。模型采用的标注器为语义角色的 PropBank 样式 [6]，对输入序列的每一个子词进行语义标注。在 PropBan 中，每个谓词都被分配了一个或者多个语义角色，用于描述谓词和其论元之间的关系。*ARG0* 通常表示具有原型施事特征的论元，*ARG1* 表示具有受事特征的论元，*ARG3* 表示施事来源，*ARG4* 则表示定语。如图 2所示，同一个句子可能有多个不同的谓语-论元结构。示例句子 [*resonstrucfting dormitories will not be approved by cavanaugh*] 存在两种语义结构。文章提供了经过语义角色标注的数据集以及一个语义角色标注器来对原始数据进行处理。

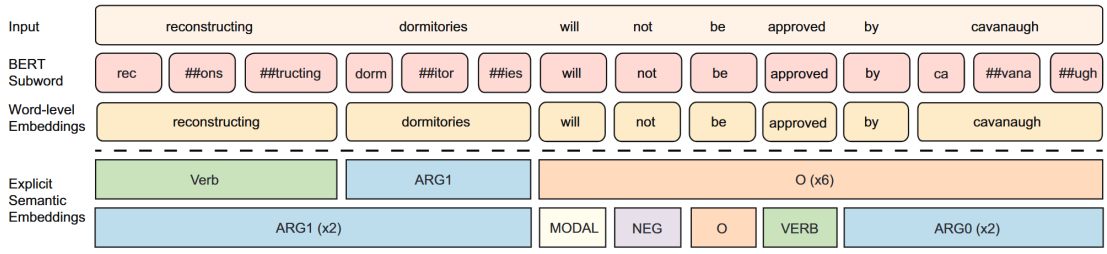

图 2. 语义角色标注

3.3 文本嵌入

原始文本和语义角色标签被表示为嵌入向量输入到预训练的 BERT 之中。单词序列先被分为子词，通过 Transformor 捕获子词的上下文信息来生成上下文嵌入。对于与每个谓词相关的 m 个标记序列，有 $T = \{t_1, ..., t_m\}$，其中 t_i 包含 n 个标记，记为 $\{label_1^i, label_2^i, ..., label_n^i\}$

语义标签是单词级别的，长度等同于原始的句子长度。将语义信号视为嵌入，将标签映射到向量，从而得到 $\{v_1^i, v_2^i, ..., v_n^i\}$，并输入一个 $BiGRU$ 层，从而获得潜在空间的 m 个标签序列的表示。

$$e\left(t_i\right) = BiGRU\left(v_1^i, v_2^i, ..., v_n^i\right) \tag{1}$$

L_i 表示对应子词的标签序列，输入到全连接层获得维度为 d 的精炼联合表示。

$$e'\left(L_i\right) = W_2\left[e\left(t_1\right), e\left(t_2\right), ..., e\left(t_m\right)\right] + b_2 \tag{2}$$

$$e^t = \left\{e'\left(L_1\right), ..., e'\left(L_n\right)\right\} \tag{3}$$

3.4 集成模块

由于预训练的 BERT 得到的是基于子词的序列，但是语义标签是单词级别的。所以在集成之前需要对序列的大小进行对其。模型使用最大池化的 CNN 获得单词级别的表示。假设单词 x_i 由子词 $[s_1, s_2, ..., s_l]$ 组成，其中 l 表示单词的子词字数。BERT 中的子词 s_j 表示为 $e\left(s_j\right)$。首先使用 Conv1D 层，$e' = W_1\left[e\left(s_i\right), e\left(s_(i+1)\right), ..., e\left(e_(i+k-1)\right)\right] + b_1$，其中 W_1 和 b_1 为可训练参数，k 是核大小。然后将 ReLU 和最大池化的神经网络用于 x_i 的输出嵌入序列：

$$e_i^* = ReLU\left(E_i'\right), e\left(x_i\right) = MaxPooling\left(e_1^*, ..., e_(l-k+1)^*\right) \tag{4}$$

词序列的整体表示为 $e^w = \left\{e\left(x_1\right), ..., e\left(x_n\right)\right\} \in \mathbb{R}^{n \times d_w}$

3.5 数据集介绍

文章在 10 个基准数据集上对 SemBERT 模型进行评估。其中包括 GLUE 基准测试 [10]、SNLI 以及 SQuAD2.0 数据集 [4]。这些数据集中包含了阅读理解、语义相似度、情感分类等多种下游任务。

一般语言评估（Ganeral Language Understanding Evaluation）是一个包含九个不同的自然语言理解任务的基准数据集，能用于检测模型对特定语言现象的理解。GLUE 基准包含一个用于评估和比较模型的在线平台。GLUE 的九项任务包含两个单句输入的内容：ColA 、SST-2，三项涉及语义相似性的任务：MRPC、STS-B 和 QQP，四个自然语言理解任务：MNLI、QNLI、RTE 和 WNLI。GLUE 基准包含任务如表1所示。基准测试以总体的平均性能来进行衡量。

3.6 任务描述

GLUE 中的九个任务除了 STS-B 是回归任务以外，其余均为分类任务。MNLI 数据集有三个类别，其余两个分类任务有两个类别。实验部分实现了 SemBERT 在 MNLI 上的评估。MNLI 数据集包含两个语句：前提语句和假设语句。模型任务是判断前提语句包含假设（contradiction）、与假设矛盾（entailment）或者两者都不（neutral）。前提语句的收集来源达数十种，其中包括转录的语音、小说以及政府报告。评价的标准是预测准确率。表2是一个 MNLI 的示例数据：SQuAD（Stanford Question Answering Dataset） [7] 是斯坦福大学发布的阅读理解数据集。数据集将 SQuAD1.1 中的 10 万个数据与 5 万个不可回答的问题结合起来。该数据集要求系统在尽可能的情况下回答问题，并在没有段落支持时放弃回答。

表 1. GLUE 基准数据集任务描述

任务	训练集大小	任务	领域
单句输入任务			
CoLA	8.5k	可接收性	杂项
SST-2	67k	情感	电影评论
相似性与释义任务			
MRPC	3.7k	释义	新闻
STS-B	7k	文本 sim 值	杂项
QQP	364k	释义	在线问答
推理任务			
MNLI	393k	自然语言推理	杂项
QNLI	108k	问答/自然语言推理	维基百科
RTE	2.5k	自然语言推理	杂项
WNLI	634	coref./自然语言推理	小说类书籍

表 2. MNLI 示例数据

前提	标签	假设	来源
The Old One always comforted Ca'daan, except today.	*neutral*	Ca'daan knew the Old One very well.	小说
yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	*contradiction*	August is a black out month for vacations in the company.	信件
At the other end of entailment Pennsylvania Avenue, people began to line up for a White House tour.	*entailment*	People formed a line at the end of Pennsylvania Avenue.	911 报告

SNLI 是一个自然语言推理的数据集。自然语言推理任务要求系统阅读一对句子并判断其含义之间的关系。除去 GLUE 中的 3 个数据集（MNLI、QNLI 和 RTE），文章还在 SNLI 数据集上进行了评估。SNLI 与 MNLI 的结构类似，包含了 27 万个句子。表3是 SNLI 的数据示例。

表 3. SNLI 示例数据

句子 1	**句子 2**	**标签**
Two women are embracing while holding to go packages.	The sisters are hugging goodbye while holding to go packages after just eating lunch.	*neutral*
Two women are embracing while holding to go packages.	Two woman are holding packages.	*entailment*
Two women are embracing while holding to go packages.	The men are fighting outside a deli.	*contradiction*

4 复现细节

4.1 与已有开源代码对比

在实验中对比了不同超参数下的模型性能，进行了多次实验对超参数进行调整以获得最好的模型表现。在复现代码的过程中遇到了许多挑战。由于 allennlp 包存在的报错问题，在复现的过程中难以使用给定的 srl 模型进行语义分割。实验中配置了多个不同的环境来解决使用的库的问题，但是未能解决此问题。因此实验只能使用给出的已标注的数据集来进行实验。在实验之前，我阅读了多篇与自然语言推理、BERT 模型以及语义角色标注相关的论文，在此基础之上才更好地理解 SemBERT 模型中各个模块的工作原理，也能更好地理解原文作者在 BERT 模型中加入显式语义线索的必要性。在实验的过程中，由于模型参数过大，实验中遇到多次内存不足的问题，因此需要在硬件限制和模型性能之间做出一定取舍，在此限制的基础之上得到了训练的最佳参数以及最佳的准确率。

4.2 实验环境设置

实验使用了 1.12.1 版本的 torch，1.13 版本的 cuda 以及 3.10.15 版本的 python。实验平台为 Windows10 系统的 PyCharm2022。实验使用了 BERT 的预训练权重，微调程序与 BERT 模型也是相同的。实验主要测试了自然语言理解数据集，包括 MNSI 和 SNLI。

4.3 实验参数

实验包含多个超参数的设置。根据使用的数据集不同，参数需要设置是否进行语义角色标注。实验的初始学习率设置为 {8e-6，1e-5，2e-5，3e-5}，预热率为 0.1，L2 权重衰减为 0.01，batch size 大小在 {16, 24, 32} 中进行选择。根据不同任务的具体情况，在 [2,5] 中设置任务的最大迭代数。任务的最大长度设置为 200。进行多次实验来选择最佳的实验参数。

5 实验结果分析

在实验中使用了 SNLI 数据与预训练好的 SemBERT 模型。数据集使用标注好语义角色的 SNLI 数据集以及 MNLI 数据集。表4和表5是原文给出的实验结果。其中，SemBERT 在 MNLI 和 BERT 与 SemBERT 在 SNLI 上的性能是文章的实验结果，其余结果均为排行榜或其他文章的数据。

表 4. 原文实验结果

方法	**MNLI** m/mm(acc)
ALBERT	91.3/91.0
RoBERTa	90.8/90.20
XLNET	90.2/89.8
BiLSTM+ELMo+Attn	76.4/76.1
GPT	82.1/81.4
GPT on STILTs	80.8/80.6
MT-DNN	86.7/86.0
BERT$_{\text{BASE}}$	84.6/83.4
BERT$_{\text{LARGE}}$	86.7/85.9
SemBERT$_{\text{BASE}}$	84.4/84.0
SemBERT$_{\text{LARGE}}$	87.6/86.3

表 5. SNLI 实验结果

方法	**Dev**	**Test**
DRCN	-	90.1
SJRC	-	91.3
MT-DNN	92.2	91.6
BERT$_{\text{BASE}}$	90.8	90.7
BERT$_{\text{LARGE}}$	91.3	91.1
SemBERT$_{\text{BASE}}$	91.2	91.0
SemBERT$_{\text{LARGE}}$	92.3	91.6

在复现实验中，使用的模型为用 SNLI 数据集进行预训练的 SemBERT 模型。数据集为原文作者提供的已进行语义标注的 SNLI 数据集以及 MNLI 数据集。最终在 SNLI 数据集上的准确率为 85.86%，损失为 0.3678。在 MNLI 数据集上的准确率为 69.64%，损失为 0.7273。

对比文章中的模型性能，实验在 SNLI 模型上的表现与原文有所出入，在 MNLI 数据集上的表现则比较差。模型结果的差异一方面是因为超参数的选择有一定的不同。比如 BERT 模型的选择、最大序列长度以及迭代样本量的选择。由于实验环境的限制，实验不能选择性能最佳的超参数，导致了模型性能的下降。在 MNLI 数据集上性能的降低是由于使用了用 SNLI 进行预训练的 SemBERT 模型。

6 总结与展望

本文介绍了引入显式语义表示的 BERT 得到的 SemBERT 模型。介绍了模型的基础架构和其三个模块的运作原理。阐述了语义角色标注在自然语言理解的问题所起到的作用。并展示了在自然语言理解的几个基准数据集上模型所取得的性能提升。未来在对 SRL 标注器进行一定的改进后，能够将更多的数据集用于此模型。此模型揭示了显式的上下文语义在自然语言理解任务中所起到的重要作用。在后续研究中，如果遇到 NLU 问题，可以引入语义角色标注来提高模型在自然语言理解任务上的性能。

参考文献

[1] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of bert's attention. In Tal Linzen, Grzegorz Chrupala, Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[3] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. Deep Semantic Role Labeling: What Works and What's Next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, 2017.

[4] Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[5] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher

J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119, 2013.

[6] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 2005.

[7] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL, 2003.

[9] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.

[10] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[11] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-Aware BERT for Language Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9628–9635, 2020.

[12] H. Zhao, X. Zhang, and C. Kit. Integrative Semantic Dependency Parsing via Efficient Large-scale Feature Selection. *Journal of Artificial Intelligence Research*, 46:203–233, 2013.