软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006169]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于深度学习的语言模型研究进展*

王乃钰[1], 叶育鑫[1,3], 刘 露[2,3], 凤丽洲[4], 包 铁[1], 彭 涛[1,3]

[1](吉林大学 计算机科学与技术学院,吉林 长春 130012)

[2](吉林大学 软件学院,吉林 长春 130012)

[3](符号计算与知识工程教育部重点实验室(吉林大学),吉林 长春 130012)

[4](伊利诺伊大学芝加哥分校 计算机科学与技术系,美国 伊利诺伊州 芝加哥 60607)

通讯作者: 彭涛, E-mail: tpeng@jlu.edu.cn

摘 要: 语言模型旨在对语言的内隐知识进行表示,作为自然语言处理的基本问题,一直广受关注. 基于深度学习的语言模型是目前自然语言处理领域的研究热点,通过预训练-微调技术展现了内在强大的表示能力,并能够大幅提升下游任务性能. 本文围绕语言模型基本原理和不同应用方向,以神经概率语言模型与预训练语言模型作为深度学习与自然语言处理结合的切入点,从语言模型的基本概念和理论出发,介绍了神经概率与预训练模型的应用情况和当前面临的挑战,对现有神经概率、预训练语言模型及方法进行对比和分析. 我们又从新型训练任务和改进网络结构两方面对预训练语言模型训练方法进行详细阐述,并对目前预训练模型在规模压缩、知识融合、多模态和跨语言等研究方向进行概述和评价. 最后总结语言模型在当前自然语言处理应用中的瓶颈,对未来可能的研究重点做出展望.

关键词: 语言模型;预训练;深度学习;自然语言处理;神经语言模型

中图法分类号: TP391

中文引用格式: 王乃钰,叶育鑫,刘露,凤丽洲,包铁,彭涛. 基于深度学习的语言模型研究进展.软件学报,2020,32. http://www.jos.org.cn/1000-9825/6169.htm

英文引用格式: Wang NY, Ye YX, Liu L, Feng LZ, Bao T, Peng T. Language models based on deep learning: a review. Ruan Jian Xue Bao/Journal of Software, 2020,32 (in Chinese). http://www.jos.org.cn/1000-9825/6169.htm

Language Models Based on Deep Learning: A Review

WANG Nai-Yu[1], YE Yu-Xin[1,3], LIU Lu[2,3], FENG Li-Zhou[4], BAO Tie[1], PENG Tao[1,3]

[1](College of Computer Science and Technology, Jilin University, Changchun 130012, China)

[2](College of Software, Jilin University, Changchun 130012, China)

[3](Key Laboratory of Symbol Computation and Knowledge Engineering for Ministry of Education, Jilin University, Changchun 130012, China)

[4](Department of Computer Science, University of Illinois at Chicago, Chicago 60607, USA)

Abstract: Language model, to express implicit knowledge of language, has been widely concerned as a basic problem of natural language processing in which the current research hotspot is the language model based on deep learning. Through pre-training and fine-tuning techniques, language models show their inherently power of representation, also improve the performance of downstream tasks greatly. Around the basic principles and different application directions, this paper takes the neural probability language model and the pre-training language model as a pointcut for combining deep learning and natural language processing. We introduce the application as well as challenges of neural probability and pre-training model, which is based on the basic concepts and theories of language model. Then

* 基金项目: 国家自然科学基金(61872163, 61806084); 吉林省教育厅项目(JJKH20190160KJ)
Foundation item: National Natural Science Foundation of China (61872163, 61806084); Jilin Provincial Education Department Project(JJKH20190160KJ).
收稿时间: 2020-05-03; 修改时间: 2020-09-01; 采用时间: 2020-10-30; jos 在线出版时间: 2020-12-02

the existing neural probability, pre-training language model include their methods are compared and analyzed. In addition, we elaborate on the training methods of pre-training language model from two aspects of new training tasks and improved network structure. Meanwhile the current research directions of pre-training model in scale compression, knowledge fusion, multi-modality and cross-language are summarized and evaluated. Finally, this paper sums up the bottleneck of language model in natural language processing application, afterwards prospects for possible future research priorities.

Key words: language model; pre-training; deep learning; natural language processing; neural language model

1 引言

近年来,深度学习[1](Deep Learning,DL)被公认为是人工智能以及机器学习领域中研究最为深入和广泛的一个方向. 随着计算机理论的不断发展,深度学习几乎被应用于人工智能研究的各个领域,包括计算机视觉(Computer Vision,CV)[2-3]、自然语言处理(Nature Language Processing,NLP)[4-5]、推荐系统[6-7]、强化学习(Reinforcement learning,RL)[8-9]、语音识别[10-11]等. 深度学习是采用自动化特征表示与学习和深度神经网络(Deep Neural Network,DNN)的一系列机器学习算法集合. 目前对于深度学习的研究取得了一系列举世瞩目的成果,对包括推荐系统、疾病检测[12]、人机博弈[13-14]在内的诸多领域产生越来越重要的影响,并在计算机视觉和自然语言处理领域取得了革命性的成功.

深度学习最初起源于对人工神经网络的研究. 人工神经网络(Artificial Neural Network,ANN)[15]是LANDAHL 等人于 1943 年首先提出来的. 而后,感知机算法、Hopfield 神经网络[16]、玻尔兹曼机[17]、误差反向传播网络(Back Propagation Network,BP Network)[18]和径向基神经网络[19]等也相继被提出. 循环神经网络(Recurrent Neural Network,RNN)是深度学习以及语言模型领域一个相当重要的神经网络结构,这一结构的提出使得神经网络对于学习到的知识有了更深层次的记忆形式,对语言模型中长文本序列的建模提升显著,但是由于早期 RNN 网络结构存在梯度消失和梯度爆炸的问题,后来提出了长短期记忆网络(Long Short-Term Memory Network,LSTM)[20],LSTM 的问世在很大程度上改进了较长序列依赖的问题. 后续研究人员对 LSTM 进行了改进,提出双向长短期记忆网络(Bi-LSTM)[21],在一定程度上解决了对输入序列的双向信息建模的问题. LSTM 与后来提出的 Transformer 模型[22]被广泛应用于自然语言处理以及语言模型当中.

自然语言处理是语言学与计算机科学相交融的一个研究领域,其主要研究任务包括词性标注[23]、命名实体识别[24]、语义角色标注[25]、机器翻译[26]、自动问答[27]、情感分析[28]、文本摘要[29]、文本分类[30]、关系抽取[31]等. 自然语言作为伴随人类文明发展而不断变化的符号化系统,单词、句子以及段落间的关系难以人工量化,使得各项任务的性能提升受到了阻碍. 而深度学习以及神经网络其强大的表示学习能力和预测能力,与自然语言处理数据集所具有的高维度、无监督和数据量大的特点相契合. 自然语言处理本质上就是利用计算机融合语言学等其他领域的知识对自然语言的内隐知识进行建模和表示,从而进一步完成自然语言的理解和生成.

在自然语言处理研究的早期,文本一般采用独热(One-Hot)表示,产生的高维且稀疏的单词表示对模型训练和预测都带来了极大的困难. 因此,构建一种面向各下游任务的语言模型来建模文本序列,成为了自然语言处理中的基础课题. 对语言模型的研究实际上就是探究如何对语言内隐知识进行表示的过程. 语言模型是自然语言处理的一个核心问题. 在探索初期,研究人员融合计算语言学理论以及相应的领域知识,提出了基于规则的文法语言模型,但在规则语言模型的构建过程中,对研究人员的语言学知识和领域知识都提出了较高的要求,且存在以下缺陷:(1)文法规则可能脱离语言实际;(2)规则灵活性差,难以覆盖惯用语法及其他复杂的语言现象;(3)引入新规则时,需要考虑已有规则之间的联系,避免冲突[32]. 以上缺陷导致模型存在时间和人力成本高昂,难以迁移至不同领域的问题. 随着统计学习方法的不断发展,为了解决文法语言模型存在的上述问题,概率语言模型(或称统计语言模型)[33]应运而生,计算机根据基本假设,对语言模型的概率分布进行估算和推理. 而后,研究人员在概率语言模型的基础上融入神经网络,形成神经概率语言模型[34]. 然而神经概率语言模型在对自然语言建模的层次还不够深入,虽然在训练过程中捕获到了单词间、字符间的共现信息和单词语义,但面对

不同的输入,依然无法动态调整相应的编码表示. 对于单词在不同上下文中的语义角色、语法和语义变化情况等高层次信息没有进行表示.

随着更大规模神经网络的出现,在监督任务的数据集规模与神经网络模型参数量之间出现了严重的不平衡现象. 一方面语义层次更高的监督任务数据集的收集与标注将耗费巨大的人力物力和时间,另一方面巨大的网络参数量会导致严重的过拟合现象出现,模型面临极高的结构风险,甚至使得期望风险增大. 在大规模数据集上进行无监督预训练,可以缓解这种不平衡现象引发的一系列问题. 各类无监督学习和预训练方法工作的提出,推动了神经概率语言模型向预训练语言模型演进. 预训练语言模型采用自监督任务训练的方法完成构建,预训练过程实质上是对模型参数完成了初始化. 在面对下游任务时,该过程使模型加速收敛. 而且预训练过程中模型获得的丰富知识,降低了庞大参数规模带来的结构风险,因此取得了极佳的性能表现. 预训练语言模型中具有代表性的工作就是 BERT 模型[35],并自 BERT 提出之后,涌现出了一系列优异的模型,例如 FaceBook 的 RoBERTa、百度提出的 ERNIE、Google 提出的 T5、NVIDIA 提出的 MegatronLM 等,在部分下游任务中性能已经超越了人类. 一个具有较强表示能力和鲁棒性的语言模型将会对搜索引擎、多轮对话、知识图谱及语音助手等实际应用产生巨大的推动作用,语言模型的研究不仅具有关键的理论意义,从语言学角度来看其社会意义更是不言而喻.

本文的主要贡献如下:

1)介绍语言模型的原理以及应用情况并总结目前面临的挑战;

2)对比分析神经概率语言模型和预训练语言模型的发展情况和优缺点;

3)总结对比目前预训练语言模型的不同方法;

4)对基于深度学习的语言模型未来的研究趋势和重点进行了分析和展望.

2 语言模型简介

语言模型可以被认为是自然语言处理各下游任务的基石,其本质就是在回答一个问题:对于一个给定的文本序列,是否具有合理性并对其合理性进行量化. 对语言模型的研究经历了文法规则语言模型至概率语言模型,再发展为神经概率语言模型的过程. 伴随新式网络结构的出现,以及半监督学习、预训练思想的提出,预训练语言模型成为了当前语言模型研究的新热点. 在本节当中将主要介绍基于深度学习语言模型的基本理论和应用以及所面临的问题和挑战.

2.1 语言模型概念及基础理论

2.1.1 N 元语法模型[36]

N 元语法模型作为概率语言模型的理论基础,其思想对后续出现的神经概率语言模型和预训练语言模型都有着深远的影响,在语言模型领域有着举足轻重的地位,在语音识别、词性标注和机器翻译等领域应用广泛. 公式表示如下:

文本序列$[w_1, \dots, w_N]$,其中w_i表示一个单词,即计算下式:

$$p(w_1,\cdots,w_n)=\prod_{i=1}^{n} p(w_i \mid w_1,\cdots,w_{i-1}) \tag{1}$$

则对于给定上下文s中某一单词i的极大似然概率计算公式为:

$$p(w_i \mid s)=\frac{C(s,w_i)}{C(s)} \tag{2}$$

其中,$C(s,w_i)$为上下文s与单词i共同出现的次数,上下文s通常由几个单词组成. 以三元语法模型为例,$|s|$=2,当不考虑上下文时被称为一元语法模型. N 元语法模型是一种基于概率统计的建模算法,自 20 世纪 80 年代提出以来,在相当长的时间内都被作为语言模型的基础思想,该方法使用序列中每个单词概率的乘积表示整个文本序列出现的概率. 若 N 选取一个较大值,表示对序列中下一个单词出现的情况约束性更强,但也会导致得到的频率信息更加稀疏,并且使 N-gram 数目指数级增长,需要更强的平滑算法来消除这一影响;若 N 选取较小值,则

表示统计结果可靠性更高、泛化能力更好,但是会使约束性更弱.

总结来看,N-gram 语言模型虽然在多个领域得到了广泛的应用,但存在以下问题:(1)模型无法量化单词之间的相似度. 假设两个具有某种相似性的词“汽车”和“轿车”,如果“汽车”经常出现在某段单词序列之后,则模型会认为“轿车”出现在这段词后面的概率也比较大. 比如“白色的汽车”经常出现,那完全可以认为“白色的轿车”也可能经常出现. (2)N-gram 模型对长距离依赖问题难以建模,由于语料规模的限制使得 N 值更大的模型面临难以处理的稀疏问题,无法训练.

2.1.2 神经概率语言模型

在N元语法模型中,计算条件概率一般采用频次作商并归一化的方法,虽然研究人员提出了多种平滑算法,但依然面临着数据稀疏和维度灾难的问题. 对 N 元语法模型可以利用最大化对数似然,构造目标函数如下所示:

$$L = \sum \log(p(w_i \mid s)) \tag{3}$$

可见,$p(w_i|s)$实际上是w_i与s的函数,公式表示如下:

$$p(w_i \mid s) = F(w_i, s, \theta) \tag{4}$$

其中,参数θ为模型待定参数集,由此将计算所有 N 元语法的条件概率转化为最优化公式(3)表示的目标函数,并求解参数集θ. 在选取神经网络适当的情况下,参数集θ的规模可远小于 N 元语法中的参数量.

Bengio 模型是最早将神经网络应用于概率语言模型的工作,其模型由三层构成:输入层、隐藏层和输出层,有效避免了数据稀疏的问题.

模型根据当前单词的前$n-1$个单词作为输入,计算当前单词出现的概率:

$$P(w_t \mid w_{t-1}, \cdots, w_{t-n+1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{5}$$

$$y = b + Wx + U\tanh(d + Hx) \tag{6}$$

$$x = (w_{t-1}, \cdots, w_{t-n+1}) \tag{7}$$

$$L = \frac{1}{T}\sum_t \log f(w_t, w_{t-1}, \cdots, w_{t-n+1}; \theta) + R(\theta) \tag{8}$$

其中,W, U, H是神经网络的权重,b, d为偏置,y_i是每个输出单词i的非标准对数概率. 公式(8)为模型的损失函数.

Bengio 模型作为神经概率语言模型的开篇之作,提供了将神经网络融入概率语言模型的一种实现方法,并且由于神经网络本身的优势避免了数据稀疏和维度灾难的问题,亦不再需要构建各类平滑算法. 神经概率语言模型在处理相对长距离依赖问题时,能够比 N-gram 模型获得更好的预测精度. 在泛化能力方面也好于 N-gram 模型并且模型所需要学习的参数量远小于概率语言模型. 但不可避免的,神经概率语言模型依然一些问题,训练时采用固定窗口大小,这与人类可以使用大量的上下文信息进行预测是不一致的. 自然语言中文本序列的单词是时序相关的,但 Bengio 模型没有使用时序信息进行建模. 虽然参数量小于 N-gram 模型,但依然产生巨大的计算开销.

2.1.3 预训练语言模型

目前预训练语言模型主要基于以下四种建模思想:(1)双向语言模型[37](2)隐蔽语言模型[35](3)排序语言模型[38](4)编码器-解码器(Encoder-Decoder)框架[39].

双向语言模型:对一个单词序列$(w_1, w_2, \ldots, w_N)$,根据给定单词的上文计算该单词的概率为前向语言模型,公式如下:

$$p(w_1, w_2, \cdots, w_N) = \prod_{k=1}^{N} p(w_k \mid w_1, w_2, \cdots, w_{k-1}) \tag{9}$$

对应的后向语言模型表示如下:

$$p(w_1,w_2,\cdots,w_N)=\prod_{k=1}^{N}p(w_k\mid w_{k+1},w_{k+2},\cdots,w_N) \tag{10}$$

其优化目标为最大化两个方向的对数似然:

$$\sum_{k=1}^{N}(\log p(w_k\mid w_1,\cdots,w_{k-1};\theta_x,\overrightarrow{\theta_{network}})+\log p(w_k\mid w_{k+1},\cdots,w_N;\theta_x,\overleftarrow{\theta_{network}})) \tag{11}$$

其中,θ_x为输入单词的表示,$\overrightarrow{\theta_{network}}$与$\overleftarrow{\theta_{network}}$表示用于前向和后向建模的神经网络参数.

双向语言模型是最早被用于预训练模型建模的思想,选择某种网络作为特征抽取器,将两个方向上抽取到的文本表示拼接在一起,这种方法的代表就是 ELMo 模型和 GPT 模型,ELMo 能够同时建模单词的语法和语义表示,并且能够根据输入上下文的不同动态的改变多义词的表示. 在收敛性方面,由于预训练过程使得可以使用更小的训练数据达到更好的效果. 但在实现过程中仅进行两个方向上的简单拼接,未做深层次融合是较为遗憾的.

隐蔽语言模型(Masked Language Model,MLM):隐蔽语言模型是预训练模型当中最为常用的一种预训练目标任务,并在预训练语言模型的研究过程中衍生出了多种预训练目标任务,对于预训练语言模型的发展影响深远.这一目标最早被称为 *Cloze* 任务,以 BERT 模型中使用的隐蔽策略为例:选取输入序列中的 15%的元素作为待隐蔽的位置,待隐蔽位置中,其中 80%的位置被**[MASK]替换,**10%的位置使用其他元素替换,10%的位置不做改变. 这一模型引入了降噪自编码器的思想,迫使模型从人为加入的噪声中恢复原始的输入,从而学习共现信息. 最早在 BERT 模型中提出的这一训练目标任务,由于原始的 BERT 是面向英文领域,在迁移至中文领域时,随机对字符进行隐蔽,被隐蔽的字符之间缺乏联系,会导致模型丢失部分词语间的共现信息,后续提出的BERT-WWM模型对这个缺陷进行了改进. 另外若在预训练模型的构建中,仅使用MLM作为预训练目标任务,对于堆叠多层的 Transformer 结构难度较低,会导致模型无法有效学习,针对这个问题,后续的研究人员将生成对抗思想引入 MLM 任务中,进一步提升预训练任务难度.

排序语言模型(Permutation Language Model,PLM):排序语言模型同样是一种预训练目标任务,最早是在 XLNET 模型中提出来的,旨在融合自回归模型与自编码模型的优点. 对给定的输入序列$(w_1,w_2,\ldots,w_T)$,用Z_T表示输入序列所有可能的排列情况所组成的集合,用z_t表示一个排列$z\in Z_T$中的第t个元素,$z_{<t}$表示一个排列$z\in Z_T$中的前$t-1$个元素. 排序语言模型的目标函数形式化表示如下:

$$\max_{\theta}E_{Z\sim Z_T}\left[\sum_{t=1}^{T}\log p_\theta(x_{z_t}\mid x_{z_{<t}})\right] \tag{12}$$

上文中介绍的双向语言模型实质上属于自回归模型,即根据上文内容预测可能出现的下一个单词或根据下文内容预测上一个可能出现的单词. 自回归模型优点体现在文本摘要、机器翻译[40]等自然语言生成任务中性能更好,但缺点是只利用了上文或下文的信息,不能同时利用上文和下文的信息. 而以 BERT 为代表的采用隐蔽语言模型建模的方法可以被视为自编码模型,由于在训练阶段和微调阶段不一致的问题,导致采用这种思想建模的方法在自然语言生成任务中性能较低. 而 XLNET 中提出的排序语言模型在相当大的程度上改善了两种模型的缺陷,可以作为未来预训练目标任务构建的基本思路.

编码器-解码器框架: Encoder-Decoder 思想最早被用于机器翻译领域,而后被广泛应用于预训练语言模型的构建当中. 使用 Encoder-Decoder 架构的语言模型优势在于处理文本摘要和机器翻译两个任务上相对其他模型有更好的性能表现,但是由于模型由编码器和解码器两部分构成,模型规模一般较为庞大,需要巨大的算力支持.

在编码器部分,以 RNN 模型为例,对于给定的单词序列 $(w_1, w_2, \cdots, w_T)$,对每一个时间步 t,其隐藏状态 h_t 以下式给出:

$$h_t = f(h_{t-1}, x_t) \tag{13}$$

在输入序列的所有元素之后,RNN 的隐藏状态形成了一个中间语义表示 c. 在解码器部分,RNN 的隐藏状态按照下式计算:

$$h_t = f(h_{t-1}, y_{t-1}, c) \tag{14}$$

解码后输出的条件概率按照如下计算:

$$P(y_t \mid y_{t-1}, y_{t-2}, \cdots, y_1, c) = g(h_t, y_{t-1}, c) \tag{15}$$

其中,函数 g 一般为 softmax 函数.

编码器-解码器框架以最大化条件对数似然作为优化的目标函数:

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log p_{\theta}(y_n \mid x_n) \tag{16}$$

2.2 语言模型的应用

语言模型作为自然语言处理的基础,其生成的低维且稠密的单词分布式表示对于一系列下游任务的性能提升具有显著作用. 伴随着神经概率语言模型和预训练语言模型的快速发展,在文本分类、序列标注以及自动问答以及机器阅读理解等各类下游任务中都取得了更好的效果. 特别是预训练语言模型,在训练过程中学习到的丰富的语法和语义推理知识,对于机器翻译、问答系统等难度较高任务的性能改善更为显著.

2.2.1 神经概率语言模型的应用

对于分类任务,张志昌等人[41]提出了一种采用独立循环神经网络(Independently Recurrent Neural Network,IndRNN)和注意力机制的用户意图分类模型,以 Word2Vec[42]生成的词向量为输入,使用 IndRNN 对输入编码. 模型引入单词级注意力机制有效量化了领域词汇对意图类别的贡献,而且所采用的 IndRNN 在堆叠层次更深的情况下更易训练. 周俊佐等人[43]针对目前已有文本分类模型在人机对话意图分类中存在的性能优劣情况,提出一种混合意图分类模型,模型结构受到 GoogleNet 提出的 Inception 网络[44]启发. 混合模型分为三层:(1)第一层为词编码层;(2)第二层为句子编码层,使用 BiGRU 和 BiLSTM 作为编码器;(3)第三层为混合模型层,将第二层的输出分别输入 Capsule[45]、MFCNN[46]与 Attention[47]三种网络,完成分类. 该方法综合利用多种网络模型的输入与输出,获得了一定的性能提升.

杜慧等人[48]在 Word2Vec 中 CBOW 模型的基础上,对生成的词向量进行情感微调,得到同时包含语义和情感倾向的词向量,在微博情感分类任务中性能提升明显. 朱苏阳等人[49]针对情感分析中的情绪分析子任务,提出对抗式网络结构. 使用 Word2Vec 中的 Skip-gram 算法生成词向量输入模型,分别抽取极性、强度与可控性特征,并在三个维度间两两进行对抗性训练. 实验结果显示,在 EMOBANK 数据上三个维度的测试结果均有显著改进.

在机器翻译方面,刘宇鹏等人[50]提出一种层次化翻译模型,将层次化规则的归纳分为短语归纳和形式化规则归纳两部分完成,并在目标函数的构造过程中引入单词级语义错误、单词短语/规则语义错误和双语短语/规则语义错误三部分,使模型在训练过程中可以平衡三部分对目标函数的影响. 该方法使用基于 RNN 的神

经语言模型作为词向量生成模型. 实验结果表明模型的目标函数很好的平衡了不同错误情况之间的影响,并在训练过程中引入的双语对齐信息,获得了较好的性能提升.

在机器阅读理解任务中,梁小波等人[51]提出了一种基于双层自注意力机制的方法,模型分为单文档编码、多文档编码和答案预测三个部分. 在单文档编码部分中,对文档和问题的上下文信息使用 GRU 模型进行表示;使用上下文表示计算文档到问题和问题到文档两个方向上的注意力信息;在文档表示信息的自匹配问题中使用自注意力机制完成计算. 实验结果表明使用注意力机制在机器阅读理解中可以提升模型在创距离依赖问题上的表现,但是在两个方向注意力信息的融合过程中仅使用拼接和向量点乘的方法,融合方式较为简单,仍存在一定缺陷.

此外, Vijayakumar 等人[52]提出一种捕获语音和对应单词相关性的模型,以 Word2Vec 生成的向量作为嵌入层,在三个关于语音推理的下游任务:(1)基于文本的声音检索;(2)Foley 声音发现;(3)与语音相关的单词相关性评估,都取得了良好的表现.

目前,在神经概率语言模型的应用中,以 Word2Vec 模型生成的词向量作为模型输入,是目前主流的应用方法. 处理意图分类和情感分类等任务时,训练字符和单词级的词向量联合送入神经网络可以有效改善未登录词对性能的影响. 在实体关系抽取领域,Word2Vec 结合双向 LSTM 是目前较为通用的方法,但该方法性能提升遇到较大的瓶颈. 当神经概率语言模型在面对机器翻译、机器阅读理解等高层次的自然语言理解任务时,由于神经概率语言模型存在无法根据不同的上下文情况动态调整文本表示的问题,使得词语或短语在不同语境下的词性、词义、语义角色等信息的变化难以被表示,当语言模型这一基础问题存在较大的性能瓶颈时,在高层次上的网络结构改进只能是杯水车薪,无法满足在应用方面的要求. 并且在使用双向 LSTM 等循环神经网络构建深层模型时,会遇到梯度消失难以训练的问题. 这些都限制了神经概率语言模型在更广泛领域的应用.

2.2.2 预训练语言模型的应用

对于自然语言处理中的分类问题,Sun 等人[53]针对情感分析任务中的子课题:特定方面的情感分析. 使用 BERT 模型[35]作为特征抽取器,并对模型进行微调,在单句和句对两类输入的方面情感分析任务上都取得了相当大的性能提升. Karimi 等人[54]同样在这一任务中,做出了进一步改进,将对抗训练思想引入模型学习过程,并使用文献[55]提出的后训练 BERT 作为语言模型,在方面抽取和方面情感分析两个子任务上都取得了性能改进. Song 等人[56]提出使用 BERT 隐藏层中蕴含的知识以增强其在基于方面的情感分析任务中的表现,为了利用中间层的知识提出了两种池化策略,一种使用 LSTM 作为池化特征抽取器,一种使用注意力机制对从 Transformer 层中抽取的隐藏状态进行池化,获得了较为显著的分类效果改进.

Li 等人[57]将实体链接建模为分类问题,针对网络协议分析中的实体链接任务提出 PEL-BERT 模型,并将外部领域知识引入 BERT 模型当中,与直接在 BERT 上微调相比分类性能更好.

此外,在序列标注任务中,Tsai 等人[58]提出一种基于 BERT 面向多语言的序列标注模型,采用知识蒸馏方法,在多种低资源语言上的词性标注和形态属性预测两个任务上性能较好,并在推理时间上缩短了 27 倍.

对于问答系统领域,意图分类和槽位填充是其中的重要任务,这两个任务存在训练数据规模小、性能提升受到限制的难点,因此 Chen 等人[59]引入 BERT 模型,并对它们进行联合训练,相较于 RNN 模型,识别和填充的准确率均有显著提升. Gulyaev 等人[60]针对问答系统中的对话状态跟踪问题,提出了一种基于 BERT 的面向目标多任务对话跟踪器(GOaL-Oriented Multi-task BERT-based dialogue state tracker ,GOLOMB),在训练过程中联合学习对话跟踪过程中的多个子任务,将对话历史、可能的意图描述和槽位值共同输入到 BERT 中完成编码,在多个评价指标上表现良好.

Xu 等人[55]在机器阅读理解(Machine Reading Comprehension ,MRC)任务的基础上提出了评论阅读理解(Review Reading Comprehension ,RRC)任务,旨在从海量的消费者评论中获取信息,用以完成电子商务领域的问答任务,提出了一种后训练 BERT 算法,增强对于评论信息的抽取能力. 杨中成[61]将预训练语言模型融入机器译文质量评估这一任务当中,将预训练语言模型中提取出的机器译文特征与依存句法信息相融合,以

BERT[35]+LSTM+多层感知机作为模型架构,提出了一种句子级的机器译文质量评估方法.

自动问答、机器阅读理解以及目前测试预训练语言模型中常见的自然语言推理,都属于 NLP 领域中的高级任务,它们对于语言模型或网络结构的编码表示能力相对于分类和序列标注任务有着更高的要求. 从目前已有模型和方法来看,一些超大规模模型,已经在自动问答、机器阅读理解和自然语言推理任务中达到了超越人类的性能表现,这表明当前预训练语言模型的构建思路是有效的. 但是不可否认的是,无论是对这些大规模模型做何种方式的压缩,都会使模型在这些任务中的表现急剧劣化,这种情况要求研究人员在后续的改进思路中需要着重注意高层次语义语法信息的高效表示和无损压缩. 综上,预训练语言模型可以生成语义丰富的单词或句子表示,对于文本分类、序列标注等任务中,获得了巨大的性能提升. 在更高层次的意图分类、对话跟踪以及机器阅读理解任务上,预训练模型蕴含的语法和语义知识对其性能贡献显著. 并且面对多任务学习和低资源语言问题,与神经概率语言模型相比,知识表示和迁移能力更强.

2.3 语言模型优点、问题及挑战

对 N 元语法模型来说,其优点在于计算效率. 在 N 值较小时,对于算力的需求较低,虽然相应的会损失一部分共现信息,但是与后续提出的神经语言模型和预训练语言模型相比,训练速度依然是非常快的. 面临的问题在于随着上下文窗口大小的增加,其形成的 N-gram 子序列的数目成指数级增长,难以进行训练. 同时由于其仅捕获了有限个单词间的共现信息,对自然语言的结构层次不够深入,在句法、语义层面没有建模. 并且由于数据稀疏带来的问题,还需要引入一系列的平滑算法来减轻数据稀疏的影响.

而后提出的神经概率语言模型,使用神经网络对概率语言模型的参数进行估计,使得在扩大上下文窗口数目的同时降低了模型参数的规模,并且在神经网络的帮助下,语言模型不再需要持续改进平滑算法来缓解性能瓶颈的问题. 特别是 Word2Vec 模型[42],作为神经概率语言模型研究过程中的经典之作,它的提出不仅在语言模型领域有重要的意义. 由于训练目标是无监督的,一个数据量庞大的语料库就可完成训练,在训练过程的负采样技术对后续的语言模型中目标任务的研究提供了新的思路. 另一方面这一语言模型良好的表示能力和训练效率推动了下游任务研究的进一步发展.

虽然 Word2Vec 提出后,研究人员从不同方面对它进行了改进,例如将 Skip-gram、负采样技术与噪声对比估计方法相融合,或者将 Word2Vec 的框架用于捕获跨语言间的语义信息,实现低资源语言进行聚类以及分类任务[62],这些方法的提出都进一步挖掘了 Word2Vec 的潜力,但是并未能从根本上解决神经概率语言模型没有在句子级和语义级进行建模的问题,还面临着以下挑战:(1)由于其生成的向量表示与单词是一一对应的关系,一词多义的问题无法解决;(2)Word2Vec 产生的是一种全局单词表示,而忽略了单词在不同上下文情况下的语法和语义变化,表示能力存在不足.

在预训练语言模型特别是 BERT 模型[35]提出后,语言模型领域的研究进入了一个新时期,其采用的双向语言模型、隐蔽语言模型以及排序语言模型等理论,在更深层次上对自然语言中的语法语义信息完成了建模. 从基准测试结果来看,预训练语言模型的表示能力相比于神经概率语言模型有了质的提升,在某些任务上甚至超越了人类. 但另一方面,预训练语言模型规模庞大,其训练过程耗费大量的算力和时间,而且难以在低计算资源设备上进行部署,还面临着以下挑战:(1)由于预训练语言模型在预训练过程中需要大量的无监督文本数据,对于低资源语言不够友好;(2)由于预训练语言模型中采用的复杂网络结构,其解释性较低,网络结构中的哪些模块可以捕获何种信息尚不明确;(3)在模型压缩过程中,会导致语言模型在推理任务上的性能发生较大损失,而自然语言推理任务又是自然语言理解中一个非常关键的问题,如何在压缩中尽可能保留其推理能力亟待解决.

3 现有语言模型分类、对比及学习方法

语言模型是自然语言处理的核心问题,先后出现了文法语言模型、概率语言模型. 然而概率语言模型存在长距离依赖建模能力较差、无法解决一词多义问题以及高层语言特征没有表示的缺陷. 研究人员提出将深

度学习应用于语言模型,这一想法最先是徐伟于 2000 年发表的论文《Can Artificial Neural Networks Learn Language Models?》[63]中提出的,其研究指出神经网络方法在当时没有大规模的用于语言模型的研究主要有两个原因:(1)一些研究人员认为使用统计方法对自然语言进行建模是更合理的;(2)神经网络所需要的数据量是巨大的,训练效率较低. 其研究结果表明采用神经网络的语言模型在性能上优于当时已有的统计方法但是其计算成本是高昂的,因而在这一思想提出之后,后续一段时间内提出的方法主要针对这两方面进行了改进,一方面对神经网络的结构进行改进来提升性能和效率,另一方面针对计算成本对目标函数和梯度计算进行优化. 而后随着 LSTM 广泛应用和 Transformer 网络的提出,以及预训练思想、半监督思想的快速发展,预训练语言模型成为目前基于深度学习的语言模型中性能表现最为优异、研究最为广泛的一类语言模型.

3.1 神经概率语言模型

神经概率语言模型的开山之作应属 Bengio 等人[34]的工作,模型联合学习词向量与分布表示的概率函数,从而避免数据稀疏问题. Mikolov 等人[64]提出循环神经网络语言模型,循环神经网络利用上下文的所有信息来预测下一个词,使用后向传播算法,可以达到很好的效果. C&W[65]提出 SENNA 模型,并给出了一种词向量的计算方法,将产生的单词表示用于一系列下游任务,如语义角色标注,词性标注,命名实体识别等. 2013 年 Mikolov 等人[42]提出 Word2Vec 模型,给出了两种计算单词分布式表示的方法,这一模型对自然语言处理领域产生了深远的影响,在很长一段时间内,使用 Word2Vec 产生的词向量作为嵌入层输入成为主流. 近年来神经概率语言模型不断深入和发展,为充分了解基于深度学习的语言模型研究进展,接下来分别介绍以上四种语言模型.

3.1.1 Bengio 模型

Bengio 模型[34]的基本原理在上文中已经介绍,此处不再赘述,后来针对上文提到的神经语言模型在训练和预测过程中高昂计算成本的问题,提出了采用蒙特卡洛采样方法来逼近梯度中的期望项以进一步降低计算复杂度[66]. 在研究过程中,重要性采样(Importance Sampling)的方法在逼近过程中取得了较好的效果,在计算效率上获得了 19 倍的提升. 后来又提出了自适应重要性采样方法(Adaptive Importance Sampling)[67],使得计算效率提升至 150 倍.

Mnih 等人[68]研究发现重要性采样方法在模型学习过程中存在稳定性不足的问题,自适应重要性采样方法实现困难且需要额外的内存存储自适应分布. Mnih 等人提出使用噪声对比估计方法(Noise-Constrastive Estimation,NCE)[69]作为替代,噪声对比估计方法旨在训练使用一个逻辑斯蒂回归分类器将真实分布的样本和噪声分布的样本区分开来,其优点在于学习过程中不会改变模型的稳定性. 实验结果表明这一方法可以用原有模型十分之一的训练时间来完成单词表示.

3.1.2 RNN 语言模型

Mikolov 等人[64]提出将循环神经网络应用于语言模型当中,该模型中使用的 RNN 网络结构被称为简单 RNN 或 Elman 网络[70],这种 RNN 的网络结构与后来研究人员提出的复杂 RNN 或 LSTM 等相比结构更为简单. 但是从另一角度来说,其对长距离序列依赖问题的处理能力,受限于其网络结构本身,也是相对较弱的. RNN 语言模型共包括输入层、隐藏层和输出层,计算过程形式化表示如下:

$$x(t) = w(t) + s(t-1) \tag{17}$$

$$s_j(t) = f(\sum_i x_i(t) u_{ji}) \tag{18}$$

$$y_k(t) = g(\sum_j^i s_j(t) v_{kj}) \tag{19}$$

其中, x 表示输入层, s 表示隐藏层, y 表示输出层, $w(t)$ 为当前的单词表示. f 为 sigmoid 函数, g 为 softmax 函数.

Bengio 等人[71]在 RNN 语言模型的基础上融入编码器-解码器框架,并首次将注意力机制引入神经机器翻译模型中. 在原始的编码器-解码器框架中,对于每个输入序列生成一个中间上下文向量 c. 这一框架的缺点在于对输入序列中的所有单词只使用一个中间上下文向量c表示,序列中的语义信息损失严重. 为解决这一缺点以及建模过程中的上下文语义贡献问题,Bengio 等人提出在编码器部分使用双向 RNN[72]进行编码,对两个方向的上下文进行建模. 在解码器部分,提出对隐藏状态□$_i$包含的信息对不同位置的单词语义贡献是不同的,并对这种贡献进行量化,公式表示如下:

$$p(y_i \mid y_1,\cdots,y_{i-1},x) = g(y_{i-1},s_i,c_i) \tag{20}$$

$$s_i = f(s_{i-1},y_{i-1},c_i) \tag{21}$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{22}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x}\exp(e_{ik})} \tag{23}$$

$$e_{ij} = a(s_{i-1},h_j) \tag{24}$$

式中s_i表示解码器 RNN 的隐藏状态,c_i表示上下文向量,□$_i$表示编码器 RNN 生成的隐藏状态,e_{ij}表示输入序列中第j个位置单词与输出序列中第i个单词的匹配程度,a表示一个前馈神经网络(该神经网络在模型训练过程中共同训练). 公式(20)给出了解码器部分的形式化表示,公式(22)用以计算每个隐藏状态对上下文向量的贡献. 模型在长句翻译任务中性能提升较大.

3.1.3 SENNA 模型

SENNA 模型[65]对给定上下文窗口中间的单词进行替换,并使模型对该单词和上下文的关系进行判断,来学习语料库中的上下文依赖关系,训练损失如下式所示:

$$\max(0,1-f(w,c)+f(\tilde{w},c)) \tag{25}$$

其中,c 为上下文窗口,$\tilde{w}$ 为将窗口中间的单词w替换后的单词,f表示不含 Softmax 层的神经网络.

SENNA 模型不仅学习到了预测单词上文知识,还将下文信息融入到了单词表示当中,并引入了 Okanohara 和 Tsujii 提出的负样本技术[73].

针对一词多义问题,Huang 等人[74]认为多义词其不同含义间可能差别较大,仅使用一个原型对单词进行表示是不充分的,在 Reisinger 和 Mooney 工作[75]的启发下引入多原型语言模型. 学习单词的多原型表示,按照以下步骤进行:首先针对每个词出现的位置设定一个固定大小的窗口,对窗口中的词求平均权重;然后使用 Spherical K-Means 聚类方法对窗口中的单词序列进行聚类;最后每个词在其所属的类别中被重新标记,用于训练类别中的词向量,多原型方法对单词相似度的计算公式如下:

$$AvgSimC(w,w') = \frac{1}{K^2}\sum_{i=1}^{k}\sum_{j=1}^{k} p(c,w,i)p(c',w',j)d(\mu_i(w),\mu_j(w')) \tag{26}$$

公式(26)中,$p(c,w,i)$为词w在给定上下文c的情况下属于类别i的概率,$\mu_i(w)$表示第i个类别中心点w,函数d为两个词之间相似度计算函数. 其实验结果表明,该语言模型在一词多义问题中取得了较好的效果,给后续的研究提供了思路.

3.1.4 Word2Vec 模型

Word2Vec 模型[42]的思想可以看作是对数线性模型和分层模型的结合,在引入 CBOW 和 Skip-gram 模型后,神经网络的结构就与对数线性模型的形式十分接近了,而后针对目标函数的计算复杂的问题,Mikolov 同样引入了层次 Softmax 层.

针对之前模型中计算复杂度主要来自于非线性隐藏层的问题,Word2Vec 选择继续采用之前工作[76]中提出的网络结构[77],提出了两种网络模型,图 1 为两种模型的结构示意图,一种名为连续词袋模型(CBOW),在该模型中移除了非线性隐藏层,投影层被所有单词共享,其训练目标是给定某一位置单词的上下文信息来预测这一位置的单词. 另一种网络模型为连续 Skip-gram 模型,其网络结构与 CBOW 类似,但训练目标不同,是通过给定一个单词预测其前后一定范围内的单词.

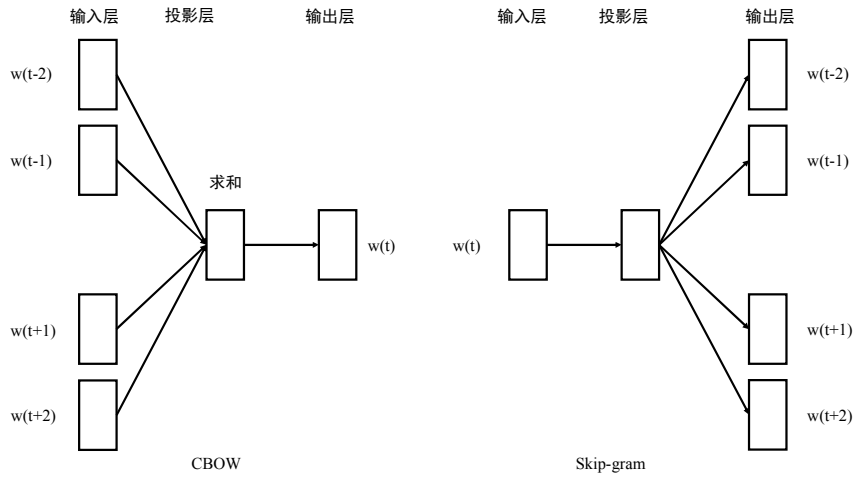

Fig.1 Word2Vec model schematic diagram

图 1 Word2Vec 模型示意图[42] 使用 PowerPoint 绘制

3.2 预训练语言模型

预训练语言模型是目前自然语言建模效果最好的一类语言模型,在 2018 年初,Peters 等人[37]就提出了 ELMo 模型,采用双向语言模型的思想,引入预训练过程. 而后 Radford 等人[78]提出 GPT 模型,使用 Transformer 作为模型的基本结构,并结合超大规模的无监督文本数据进行预训练,在自然语言生成任务中获得了显著的性能提升. BERT 问世后,一系列改进预训练模型被提出,本节将对预训练语言模型的提出和发展历程进行概述.

3.2.1 初期预训练语言模型

Dai 和 Le 在文献[79]中提出了两种使用无标签数据改进 RNN 语言模型性能的方法:第一种方法将训练目标设定为预测当前句子的下一个句子是什么;第二种方法来源于自编码(Autoencoder)思想,训练目标为通过由 RNN 组成的自编码器重构输入序列. 这一方法的提出对后续预训练语言模型中目标任务的设计产生了深远的影响.

在 BERT 模型提出之前就已经出现了几个具有代表性的工作:(1)Peters 等人[37]工作中,将预训练思想与双向语言模型相结合,使用 Jozefowicz 等人[80]采用的 CNN-BIG-LSTM 网络,并加入高速公路网络和层与层之间的残差连接作为建模双向语言模型的结构,其两个方向的 LSTM 之间参数是非共享的. ELMo 模型的思想本质上是一种自回归语言模型(Autoregressive LM),虽然采用了双向的 LSTM,但只是对两个方向的隐状态进行了简单拼接,没有进行更高层次的融合,效果提升仍有较大空间;(2)Radford 等人[78]提出的 GPT(Generative Pre-Training)模型就采用了无监督预训练-监督微调的两段式方法,使用堆叠 Transformer 作为 Decoder. GPT 模型同样是典型的自回归语言模型,虽然没有采用双向建模思想,但是在预训练的基础上使用 Transformer 作为基本结构,Transformer 在长距离的依赖问题上表现明显好于 LSTM,这也使得虽然 GPT 没有采用双向建模的思想,但性能表现依然超过了 ELMo 模型. 而后 OpenAI 团队继续对 GPT 进行扩展提出了 GPT-2 模型[81],将堆叠 Transformer 层数提升至 48 层,模型总参数量达到了 15 亿,并且将 Caruana 提出的多任务学习(Multitask Learning)[82]的思想融入其中. GPT-2 的问题在于没有改变其本质是自回归语言模型的问题,由于采用单向

Transformer 对上下文建模能力不足,其主要的性能提升来自于多任务预训练、超大规模数据集和超大规模模型的共同作用.

可以看到在预训练语言模型研究的初期,研究人员对上文提到的基本理论进行了融合. 因而 BERT 的横空出世也并非偶然,其实质上是前人研究思想的集大成者,在下一小节中将对 BERT 模型的原理和影响进行概述.

3.2.2 来自 Transformer 的双向编码表示(BERT)

Bidirectional Encoder Representations from Transformers(BERT)是 Devlin 等人[35]提出的工作,由于 GPT 模型仅使用了从左至右的单向语言模型,在语言建模的过程中某一个单词出现的分布不仅与其上文有关,也与下文有较大的关联,因此双向语言模型在 BERT 提出后成为后续方法的基本思想,并使用隐蔽语言模型和下一句预测(Next Sentence Prediction ,NSP)两个预训练目标,其中的 NSP 任务本质上来自于 Word2Vec 的负采样技术. BERT 采用堆叠多层双向 Transformer 作为网络基本结构,使用 WordPiece 分词器[83],模型输入由单词嵌入、分段嵌入和位置编码嵌入三部分组成.

BERT 的创新点在于使用双向 Transformer 作为特征抽取器,弥补了 ELMo 和 GPT 的遗憾,另一方面在预训练阶段引入的两个目标任务对于建模上下文表示以及共现信息有一定贡献. 但是其中的不足同样是不可忽视的:双向 Transformer 结构没有摆脱自编码模型的桎梏,其庞大的模型规模对于低计算资源的设备极不友好,难以部署和应用;预训练中的隐蔽语言建模会导致与微调阶段模型输入不一致,使得模型性能不能完全释放.

3.3 基于BERT的改进模型

在本节将对 BERT 提出后,研究人员对预训练模型的改进方法进行阐述和分析. 目前,对预训练语言模型的改进主要集中在两个方向:一方面是针对原 BERT 模型中的 MLM 以及 NSP 任务进行扩展或者替换;另一方面则是对模型的网络结构进行改进以使模型学习到更丰富的表示.

3.3.1 对训练目标进行改进

Cui 等人[84]提出 BERT-全词隐蔽(BERT-Whole Word Masking,BERT-WWM)模型,提出在中文领域 BERT 的预训练过程中如果按照原始的 MLM 训练目标,随机对字进行隐蔽会造成语义信息的损失,进而提出全词隐蔽训练目标. 在隐蔽过程中,按照词语级完成隐蔽,举例来看,假设输入序列为“使用语言模型来预测下一个词的概率”,原始 MLM 可能会隐蔽为“使用语言[MASK]型来[MASK]测下一个词的[MASK]率”,使用 WWM 后则变为“使用语言[MASK][MASK]来[MASK][MASK]下一个词的[MASK] [MASK]”. 这一训练目标可以在一定程度上迫使模型对词语内部的共现信息进行学习,进而获得表示能力更好的模型.

Yang 等人[38]提出 BERT 中存在训练和微调阶段输入不一致导致性能损失的问题,通过引入排序语言模型以及双流自注意力机制(Two-Stream Self-Attention)进行了改进. 在排序语言模型中,对输入序列采用不同的分解顺序,以使模型获取不同的上下文信息,同时为保证预训练和微调阶段的一致性,引入双流自注意力机制. 通过注意力隐蔽矩阵使模型在对当前位置的单词预测时,只能“看到”上下文和给定的位置信息,而在预测下一个位置的单词时,可以获得上一个单词位置信息和内容信息. 并且为了避免对输入序列全排列后进行预测带来的巨大计算开销和优化问题,在实现时,仅让模型预测顺序重排后序列的最后部分单词. 在网络结构方面,与已有工作不同,Yang 等人提出的模型采用 Transformer-XL[85]作为基本单元,是 Transformer 的改进模型,为了改进 Transformer 模型存在的上下文分段问题,即对源文档分段输入模型的过程中难以准确分句的问题,导致输入模型片段可能不是一个完整的句子,导致模型表示出现问题. Transformer-XL 提出在训练时引入记忆机制,将上一个分段编码后的表示存入记忆中,并加入当前分段的编码计算中. 这一模型的提出是排序语言模型在预训练语言模型中的首次应用,其结合了自回归思想与自编码思想的优点,使预训练模型从简单单向拼接朝着实质双向建模的方向演进,并且 Transformer-XL 结构的加入,使模型在长文本任务中性能提升显著.

Liu 等人[86]提出了一系列针对 BERT 训练过程中存在问题的改进模型 RoBERTa,其使用了更大的批量规模和无监督文本数据,在处理文本输入时,与 BERT 不同的是采用了字节对编码(Byte Pair Encoding,BPE)[87]进

行分词. 在目标任务中移除了 NSP 任务,并在采用动态隐蔽策略(每次的输入序列都使用不同的隐蔽模式,即使两次输入序列相同其隐蔽模式也不同). 其提出的 BERT 设计选择和训练策略,对后续模型的参数调整有显著帮助.

Joshi 等人[88]在 RoBERTa 的基础上提出了 SpanBERT 模型,同样采用了动态隐蔽的思想并去除 NSP 任务,同时还提出小段隐蔽(Span Mask)和小段边界目标任务(Span Boundary Objective ,SBO),即在隐蔽时对一定长度的单词进行隐蔽. 小段边界目标任务是给定被隐蔽小段两端的单词,通过两端的单词和被隐蔽部分的位置向量恢复被隐蔽的所有单词. 在训练时,借鉴了 RoBERTa 模型中提出的动态隐蔽策略,而不是在数据预处理时就进行隐蔽. 实验结果表明, 这一方法在 GLUE 基准测试中的自然语言推理和抽取式问答任务中效果提升显著. 在 XLNet 中,模型通过 PLM 来显式地学习被隐蔽词之间关系,而在 SpanBERT 中,这种关系的学习是通过被遮盖掉部分本身的强相关性,隐式地学习到的.

Dong 等人[89]提出的 UNILM 模型,这一模型可以看作是前述方法的集大成之作,在预训练目标任务中,采用了三种语言模型目标任务:(1)双向语言模型,即 BERT 模型所采用的方式;(2)单向语言模型(Unidirectional LM),这一目标在 ELMo 和 GPT 模型中被采用;(3)序列到序列语言模型,这一目标任务是由 Song 等人[90]提出的. 通过自注意力层的掩码机制,UNILM 可以在预训练过程中同时完成三种目标任务的训练. 在训练中,UNILM 采用 SpanBERT 中提出的小段隐蔽策略,模型的损失函数由以上三种预训练任务的损失函数共同构成,并且为了保持损失函数各部分贡献的一致性,三种目标任务在训练时训练相同的时间. 多种目标任务的建模和参数共享使得语言模型在自然语言理解和生成任务上都具有较好的泛化能力.

Sun 等人[91]在 ERNIE 的基础上提出了 2.0 版本,提出了多达 7 种预训练任务,覆盖单词级、句子级以及语义级三个方面,并将持续学习的思想加入到这一框架中,使上一个训练任务中的知识可以被保留,进而让模型获得更长距离的记忆. 在模型架构方面,ERNIE 同样使用 Transformer Encoder 作为编码器,但是与 BERT 的模型输入不同的是,ERNIE 还引入了任务嵌入,使模型在持续学习过程中能够区分不同任务. 类似地,Wang 等人[92]同样提出了单词级以及句子级的训练目标. 单词级训练目标在 MLM 的基础上引入了排序语言模型[38]的思想,输入序列中的单词一部分被隐蔽,而后选择序列中任意 trigrams 作为子序列,将 trigrams 中的单词打乱顺序后输入模型,使模型学习如何恢复 trigrams 的顺序. 而且 Wang 等人还对 NSP 任务也进行了改进,对于给定的句子对(S_1, S_2),将预测目标分为三类:(1)S_2是S_1的下一个句子;(2)S_2是S_1的上一个句子;(3)S_2随机采样自另一个文档,与S_1没有上下文关系. 上述改进使模型在自然语言推理任务上性能提升明显.

最近 Clark 等人[93]提出了 ELECTRA 模型,为解决 MLM 任务难度低,无法让模型深层次地捕获语义信息的问题,模型采用生成对抗网络(Generative Adversarial Network,GAN)[94]的思想,由一个生成器(Generator)和一个判别器(Discriminator)组成,采用 MLM 作为生成器,训练中生成器将被隐蔽的单词恢复,输入到判别器中,判断每个单词是原始输入的还是由生成器生成的,通过这样的方式完成学习. 实验结果表明,在相同模型规模和数据量的情况下,这一目标任务更加高效,表示能力更强. 在训练过程中还进行了权值共享,模型的损失函数用下式表示:

$$\min_{\theta_G,\theta_D}\sum_{x\in X} L_G(x,\theta_G)+\lambda L_G(x,\theta_D) \tag{27}$$

实际上 ELECTRA 模型可以看作是带有负采样的 CBOW 模型的大规模版本,将 CBOW 模型中的 BOW 编码器替换为了 Transformer,将基于 unigram 的负采样替换为了基于 MLM 生成器的负采样方法,并且把目标任务定义为判断输入单词是否来自数据的二分类任务. ELECTRA 模型与上文中论述的方法相比,最显著的区别在于预训练目标任务从生成式的学习方法改进为了判别式的学习方法,模型不再需要恢复被隐蔽的字符或者单词,而是判断没有字符或单词是否被替换过,使模型具有更高的计算效率和参数效率.

此外,Raffel 等人[95]将迁移学习的思想融入预训练语言模型中,将所有自然语言问题转换为文本-文本(text-to-text)形式. 使用隐蔽语言模型思想作为预训练目标,在对输入序列隐蔽时采用 Joshi 等人[88]工作提出

的小段隐蔽策略,形成了一个面向各种下游任务的训练框架,由于其庞大的模型规模和预训练数据集,在某些任务中取得了目前非常好的效果.

目前,对预训练模型的改进,主要集中在提出新训练目标任务的方向上. Liu 等人[86]工作中提出的一系列训练策略和微调方法,被广泛应用于后续的研究中. 文献[93]的工作中首次将生成对抗的思想引入预训练过程中. 由于对 GAN 的研究还不深入,训练过程中存在梯度难以在生成器和判别器间传递的问题,还需要新的损失函数或训练方法的研究. Sun 等人[91]和 Raffel 等人[95]的工作表明了,持续学习和迁移学习的思想有助于模型在不同类任务中同时获得性提升.

3.3.2 对网络结构进行改进

原始的 BERT 模型采用的是 Seq2Seq 框架中 Encoder 部分,Song 等人[90]提出隐蔽序列到序列预训练方法(Masked Sequence to Sequence,MASS),在训练中对编码器的输入序列随机隐蔽长度为k的连续片段,通过解码器部分将被隐蔽片段恢复,损失函数如下:

$$L(\theta;\chi)=\frac{1}{|\chi|}\sum_{x\in\chi}\log P(x^{u:v}\mid x^{\backslash u:v};\theta) \tag{28}$$

其中,χ为训练集,x为输入序列,$x^{u:v}$表示x中从位置u到v的一段,$x^{\backslash u:v}$表示x中第u到第v个位置被隐蔽.

由于 BERT 只训练一个编码器用于自然语言理解,而 GPT 模型训练的是一个解码器. 在自然语言生成任务中,以上两种模型只能分开预训练编码器和解码器,因此编码器-注意力-解码器没有被联合训练,注意力机制也不会被预训练,而解码器对编码器的注意力机制在这类任务中非常重要,因此 BERT[35]和 GPT[78] 在语言生成任务中表现弱于 MASS.

同样,Lewis 等人[96]提出的 BART 模型也采用了编码器-解码器的结构,编码层采用双向 Transformer,其本质依然是降噪自编码器的思想. 在预训练目标任务中,使用了 5 种加入噪声的模式:(1)单字隐蔽;(2)单字删除;(3)跨度隐蔽;(4)句子重排;(5)文档重排. 在编码器部分,序列在输入编码器前就已经进行了隐蔽,经过编码器编码后,解码器根据编码器输出的编码表示和未被隐蔽的序列恢复原始序列. 一系列噪声模式的加入使得 BART 在序列生成和自然语言推理任务上的表现提升明显.

3.3.3 模型对比

Table 1 Comparison of Improved Method

表 1 BERT 的改进方法比较

模型	模型结构	预训练目标任务	GLUE
人类	-	-	87.1
BERT	多层双向 Transformer	MLM+NSP	80.5
RoBERTa	多层双向 Transformer	动态 MLM+全句采样	88.5
SpanBERT	多层双向 Transformer	小段 MLM+小段边界	82.8
XLNET	多层双向 Transformer	排序语言模型	89.5
UNILM	多层 Transformer	单向+双向+Seq2Seq	80.8
ERNIE	多层双向 Transformer	单词+句子+语义三级	83.6/90.1
ALICE	多层双向 Transformer	MLM+单词重排+句子重排	87.0

ELECTRA	多层 Transformer + 生成对抗+权值共享	生成对抗式替换检测目标	88.1
T5	多层双向 Transformer +Seq2Seq	小段 MLM	89.7
BART	多层 Seq2Seq Transformer + cross-attention	MLM+单词删除+句子重排+文档轮换+文本填充	-
MTDNN	多层双向 Transformer 共享层 + 特定任务层	MLM+NSP	82.7

Table 2 Performance Comparison of Improved Models

表 2 改进模型的性能比较

模型	CoLA	SST-2	MPRC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	WNLI
人类	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0/92.8	91.2	93.6	95.9
$BERT_{base}$	52.1	93.5	88.9/-	85.8/-	71.2/-	84.6/83.4	90.5	66.4	-
$BERT_{large}$	60.5	94.9	89.3/-	86.5/-	72.1/-	86.7/85.9	92.7	70.1	-
RoBERTa	67.8	96.7	92.3/89.8	92.9/91.9	74.3/90.2	90.8/90.2	98.9	88.2	89.0
SpanBERT	64.3	94.8	90.9/87.9	89.9/89.1	71.9/89.5	88.1/87.7	94.3	79.0	65.1
XLNET	70.2	97.1	92.9/90.5	**93.0/92.6**	74.7/90.4	90.9/90.9	99.0	88.5	92.5
UNILM	61.1	94.5	90.0/-	-/87.7	71.7/-	87.0/85.9	92.7	70.9	65.1
$ERNIE_{original}$	63.5	95.6	87.4/90.2	91.2/90.6	**90.1**/73.8	88.7/88.8	94.6	80.2	67.8
$ERNIE_{latest}$	**72.2**	**97.5**	93.0/90.7	92.9/92.5	75.2/90.8	91.2/90.6	98.0	90.4	**94.5**
ALICE $v2_{larg}$	71.3	97.1	**93.9/91.9**	93.0/92.5	74.8/91.0	90.7/90.4	**99.2**	87.4	91.8
$ELECTRA_{lar}$	68.2	-	89.6/-	91.0/-	-/90.1	-/90.1	95.4	83.6	-
T5	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	**92.0/91.7**	96.7	**92.5**	93.2
BART	62.8	-	-/90.4	-	-/**92.5**	89.9/90.1	94.9	87.0	-
MT-DNN	62.5	95.6	91.1/88.2	89.5/88.8	72.7/89.6	86.7/86.0	93.1	81.4	65.1

MRPC: F1/accuracy, STS-B: Pearson/Spearmanr correlation,QQP: F1/accuracy, MNLI: matched/mistached accuracies

表中$ERNIE_{original}$和$ERNIE_{latest}$均为百度公司提出的模型

表 1 总结和对比了目前主流预训练语言模型所使用的模型结构、预训练目标任务以及所得的 GLUE 测试分数. 表 2 则给出了主流模型在 GLUE 中各下游任务上的详细性能表现. 由表 1 和表 2 可以看到,截止至目前[1],GLUE 得分最高的 ERNIE(百度)模型[91],与其最初发布的性能相比,在 CoLA、RTE 和 WNLI 三个任务上获得了巨大的提升,提升幅度最大达到了 26.7,这与其整体的模型设计是分不开的. ERNIE(百度)模型采用了持续学习的思想,并可以不断引入更多以及更新颖的预训练任务,其最早提出的 3 个层次共 7 项预训练目标任务就从不同方面捕获单词级、句子级以及语义级的内隐信息,同时还引入了种类更为丰富的无监督训练语料. XLNET[38]、ELECTRA[93]以及 T5 模型[95]的优异性能表现也证明了,更大规模的数据对于模型的提升是有显著作用的. 从 STS-B 任务各模型的表现来看,XLNET 与 ALICE[92]两个模型取得了最为出色的成绩,而两个模型在预训练目标中均采用了排序语言模型的思想,这表明顺序重排目标有利于模型学习自然语言中语义相似度问题中的内隐信息.

从 WNLI 任务来看,除 ERNIE(百度)外,表现最好的模型分别是 T5、XLNET 和 ALICE. T5 模型是通过以

1 2019 年 12 月 29 日

110 亿参数为代价获得 93.2 分的成绩的,其参数量更小的版本性能表现均显著低于 XLNET 和 ALICE 的. 而且值得注意的是,在同为自然语言推理任务的 MNLI 和 RTE 测试中,XLNET、ALICE 和 BART[96]模型的得分十分接近. 在 MPRC 和 QQP 两个任务中,也可以观测到上述现象,XLNET 与 ALICE 的表现均较为优异且分数相近,而且 BART 方法中提出的 5 种训练目标中也有类似重排的任务,可知排序思想在训练模型的推理能力上作用显著.

从 T5 模型在 CoLA 即语言接受性任务的表现以及在该任务中 T5 模型的纵向对比来看,更大规模的参数量使得 T5 在 CoLA 任务上从 41.0 提升至了 70.8,这一过程以庞大的参数规模为代价,而与 XLNET 与 ALICE 横向对比可以看到,更大规模的参数量和无监督训练数据,并不能使 T5 模型在该任务中获得进一步有效的性能提升,由于这一测试任务的特殊性,在未来研究人员应当从更为新型的预训练目标或在顺序重排任务的基础上进行改进,以获得提升.

总结来看,更大的参数量和数据集规模对模型的性能提升是具有积极意义的,但是其提升幅度有限. 排序思想在语言接受性和自然语言推理任务上贡献显著,在未来研究针对推理任务的语言模型时,使用排序思想的预训练目标将会获得较好的性能. 持续学习思想和更多层次的预训练任务是目前预训练语言模型改进的关键.

4 预训练语言模型的数据集和基准测试

目前主流预训练语言模型都采用预训练-微调两阶段的思路进行应用,因而在预训练阶段需要大规模的无监督数据对模型进行预训练. 在英文领域,由于国外对预训练语言模型的研究起步较早,因此无论是预训练数据集还是基准测试都已经较为完善. 但在中文领域中,因为研究起步较晚,预训练数据集和基准测试任务还未能形成固定形式,不同模型间采用的训练数据和测试任务不尽相同.

4.1.1 常用预训练数据集

在英文领域中,预训练阶段主要采用以下四个数据集:

BooksCorpus[97]:图书语料库最初是研究人员为研究句子相似度而从网络上爬取收集形成的数据集. 该数据集中一共包含有 16 个种类的 11038 本图书,总单词数达到了 9.8 亿个,词汇表数目为 131 余万个,该语料库与英文维基百科(English Wikipedia)语料是当前主流方法中最为常用的语料数据.

English Wikipedia:英文维基百科数据是由维基百科官方定期更新和发布的,其格式为 web 文本原始数据,需进行预处理. 目前多数模型采用的是共计 25 亿个单词的版本.

Giga5[98]:全称为 English gigaword fifth edition,这一数据集是语言数据联合会(Linguistic Data Consortium,LDC)所提出的,共包含有来自法新社、美联社、纽约时报和新华社的 400 万篇新闻文章.

ClueWeb 2012-B[99] :该语料数据是在 ClueWeb09 的基础上扩展而来的,由 7.33 亿个英文网页构成,主要来源于对 Web 网页、推特链接和维基旅行的爬取.

在中文领域中,由于提出的方法还较少,所采用的训练数据集种类还不统一,BERT-WWM 模型[84]中采用的是维基百科定期发布的 Wikipedia dump 语料,预处理后共计 1360 万行文本. 在 ERNIE(百度)模型[91]中则使用了百科、新闻、对话和信息检索领域的语料.

4.1.2 中文和英文基准测试

在中文的基准测试任务中,同样由于研究起步较晚的原因,并未形成类似于英文中 GLUE 等成型的测试数据集,对基准测试数据的选择还存在一定分歧,主要有 7 类测试任务,名称和任务内容由表 3 给出:

Table 3 Chinese Benchmark Task

表 3 中文基准测试任务

测试任务	任务内容
机器阅读理解	属于文档级建模任务,旨在从给定的文本中提取连续的片段以回答问题

命名实体识别	识别各种实体,包括人名,位置名和组织机构名等
自然语言推理	确定两个句子或两个单词间语义关系(蕴含、矛盾和中立)
情感分析	分析句子蕴含的情绪是积极、消极或中性的
语义相似度	基于两个句子的含义或语义内容的相似性来判断意图是否相同
问答	为问题选择对应回答
文档分类	对整篇文章所属的类别进行区分

目前,以上 7 类测试任务常用数据集如下:

机器阅读理解(Machine Reading Comprehension,MRC):CMRC 2018[100],DRCD [101],DuReader[102],CJRC[2]

命名实体识别(Named Entity Recognition,NER): MSRA-NER (SIGHAN 2006)[103] ,People Daily[3]

自然语言推断(Natural Language Inference,NLI):XNLI[104]

情感分析(Sentiment Analysis,SA):ChnSentiCorp[4],Sina Weibo[5]

语义相似度(Semantic Similarity,SS):LCQMC[105],BQ Corpus[106]

问答(Question Answering,QA):NLPCC-DBQA[6]

文档分类(Document Classification,DC):THUCNews[107]

Table 4 Chinese Benchmark Dataset

表 4 中文基准测试任务数据

测试任务	类型	训练集	开发集	测试集	标签数	评价指标
CMRC 2018	MRC	10K	3.2K	4.9K	-	EM/F1
DRCD	MRC	27K	3.5K	3.5K	-	EM/F1
DuReader	MRC	271.5K	10K	-	-	EM/F1
CJRC	MRC	10K	3.2K	3.2K	-	EM/F1
MSRA-NER	NER	45K	2.3K	4.6K	7	F1
People Daily	NER	51K	4.6K	2.3K	7	F1
XNLI	NLI	392K	2.5K	2.5K	3	Accuracy
ChnSentiCorp	SA	9.6K	1.2K	1.2K	2	Accuracy
Sina Weibo	SA	100K	10K	10K	2	Accuracy
LCQMC	SS	240K	8.8K	12.5K	2	Accuracy
BQ Corpus	SS	100K	10K	10K	2	Accuracy
NLPCC-DBQA	QA	182K	41K	82K	2	F1
THUCNews	DC	50K	5K	10K	10	Accuracy

下面对表 4 中给出的数据集做简要介绍:

CMRC 2018**:**中文机器阅读理解(Chinese Machine Reading Comprehension)是用于机器阅读理解测试的抽取式理解数据集,由中国中文信息学会、科大讯飞和哈尔滨工业大学共同发布的.

DRCD**:**Delta 阅读理解数据集(Delta Reading Comprehension Dataset,DRCD)同样是是提取式阅 读理解数据集,由三角洲研究所(Delta Research Institute)发布.

DuReader:是一个大规模的现实世界中文数据集,用于机器阅读理解和问答,由百度在 ACL 2018 上发布. 数据集中的所有问题均采样自现实中匿名用户的搜索请求,相应的回答由人工生成.

CJRC:包含有是/否问题,无答案问题和跨度提取问题. 数据来自于中国法律判决文件.

MSRA-NER(SIGHAN 2006):由微软亚洲研究院发布.

People Daily:爬取自人民日报官方网站.

XNLI:是目前被广泛使用的用于 NLI 任务的数据集.

2 http://cail.cipsc.org.cn

3 https://github.com/ProHiryu/bert-chinese-ner/tree/master/data

4 https://github.com/pengming617/bert_classification

5 https://github.com/SophonPlus/ChineseNlpCorpus/

6 http://tcci.ccf.org.cn/conference/2016/dldoc/evagline2.pdf

ChnSentiCorp:包括多个领域的评论文本,例如酒店,书籍和电子产品.

Sina Weibo:收集自新浪微博,包含积极和消极两种情感极性.

NLPCC-DBQA:于 2016 年在 NLPCC 上发布.

LCQMC:是由哈尔滨工业大学在 COLTING 2018 上发布的.

BQ Corpus:由哈尔滨工业大学和微众银行在 EMNLP2018 上发布的.

THUCNews:由新浪新闻的文章构成,包含有多种类别,由清华大学自然语言处理实验室发布.

在英文领域中,对预训练语言模型的基准测试已经形成了规模,目前来看主要由三种基准测试组成:GLUE、SQuAD 和 RACE

GLUE:通用语言理解评估(General Language Understanding Evaluation,GLUE)是目前在预训练语言模型的测试中最为常用的评测数据集,其一共包含有 9 项任务,具体如下:

CoLA:语言接受度语料库(The Corpus of Linguistic Acceptability)[108]包含来自 23 种语言的 10657 个句子,CoLA 通常用于判断句子是否符合语法规范.

SST-2:斯坦福情感树(The Stanford Sentiment Treebank)[109]包含 9645 条电影评论,并带有情感倾 向注释.

MNLI:多类型自然语言推理(Multi-genre Natural Language Inference)[110]由 43 万个带有文本蕴含信息注释的句子对组成,通常用于文本推理任务.

RTE:识别文本蕴含(Recognizing Textual Entailment)[111]是类似于 MNLI 的语料库,同样常用于自然语言推理任务.

WNLI:Winograd 自然语言推理(Winograd Natural Language Inference)[112]是捕获两个段落之间共指信息的数据集.

QQP:Quora 问题对(Quora Question Pairs)[7],由超过 40 万个句子对组成,抽取自 Quora 中的问答社区.

MRPC:Microsoft 释义研究语料库(Microsoft Research Paraphrase Corpus)[113]包含从互联网新闻中提取的 5800 个句子对,任务类型与 QQP 任务类似.

STS-B:文本语义相似度基准测试(The Semantic Textual Similarity Benchmark)[114]其中的文本来自图片标题,新闻标题和论坛.

QNLI:问题自然语言推理(Question Natural Language Inference)[115]其任务是判断给定的文本对是否是问题-回答.

GLUE 中各测试数据集的情况由表 5 给出:

Table 5 GLUE Dataset

表 5 GLUE 测试数据

测试任务	类型	训练集	开发集	测试集	标签数	评价指标
CoLA	Acceptability	8.5K	1K	1K	2	Matthews corr
SST-2	Sentiment	67K	872	1.8K	2	Accuracy
MNLI	NLI	393K	20K	20K	3	Accuracy
RTE	NLI	2.5K	276	3K	2	Accuracy
WNLI	NLI	634	71	146	2	Accuracy
QQP	Paraphrase	364K	40K	391K	2	Accuracy/F1
MPRC	Paraphrase	3.7K	408	1.7K	2	Accuracy/F1
STS-B	Similarity	7K	1.5K	1.4K	1	Pearson/Spearman corr
QNLI	QA/NLI	108K	5.7K	5.7K	2	Accuracy

SQuAD:斯坦福问答数据集(Stanford Question Answering Dataset)是一个大规模的机器阅读理解数据集,包含有两个任务. SQuAD1.1[115]中成对给出问题与对应的回答,数据集共包含 10 万个样本,而 SQuAD2.0[116]

7 https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

中则加入了无回答问题,并将规模扩充至 15 万个.

RACE: RACE 数据集[117]包含来自初中和高中英语考试中抽取的近 10 万个问题,对应的回答由专家给出,是机器阅读理解数据集中最具挑战性的一个. RACE 中文本平均长度大于 300,比其他阅读理解数据集(如 SQuAD)序列更长.

目前,在预训练语言模型的训练数据和基准测试方面,英文领域中训练数据和测试任务已经形成了较为完善的规模,研究人员提出的不同方法所使用的无监督数据和基准测试基本上是一致的,有利于后续研究人员比较其性能表现和进一步分析. 而在中文领域,由于研究起步较晚,无监督数据和基准测试仍不完善,尚未形成相对固定的数据集. 不同方法间的性能比较和分析较为困难,还需要研究人员收集和处理以形成较为规范的训练和测试数据集. 在下一节当中将对未来基于深度学习语言模型的研究趋势进行分析和展望.

5 预训练语言模型的扩展方法

在构建语言模型时,首先对模型在大规模的语料库上进行预训练过程,最早在 Word2Vec 模型[42]中就已经得到了应用,并且表明预训练过程对于语言模型的鲁棒性有着较大的帮助. BERT 模型[35]一经问世,其在 GLUE 基准测试中的优异表现,使得无监督文本预训练结合下游任务微调成为了目前神经语言模型的主流思路. 不仅如此,在该模型提出后,基于深度学习语言模型的研究成为了 NLP 领域的新热潮. 从对预训练模型的蒸馏、量化、剪枝,探索大规模语言模型在边缘计算设备上部署的可能性,到一系列多模态、跨语言模型的提出,对预训练语言模型的不断研究,推动 NLP 应用朝着高性能、高鲁棒性、高可部署性的方向发展起到了至关重要的作用. 本节旨在概述当前预训练语言模型中先有的变体模型以及与其他领域的融合方法.

5.1 模型压缩方法

自 GPT、BERT 等一系列使用 Transformer 作为特征抽取器的语言模型出现,其巨大的模型规模使得语言模型的训练和预测都面临着计算资源和时间的高度消耗,导致语言模型在边缘设备和低计算资源设备上的部署和应用难以实现,限制了预训练技术的实际应用中的作用. 目前,预训练语言模型规模压缩主要有三种方法:(1)知识蒸馏;(2)参数量化;(3)网络剪枝. 本节将对其进行概述.

5.1.1 知识蒸馏

知识蒸馏[118]实质上就是一个“教学”过程,旨在将大规模网络模型T中蕴含的知识转移至小规模模型S中,使模型S尽可能模仿T的行为,记f^T和f^S为两模型的行为函数,则知识蒸馏的目标就是最小化如下目标函数:

$$L_{KD}=\sum_{x\in X}L(f^s(x),f^T(x)) \tag{29}$$

其中,$L(\cdot)$为计算两模型行为差别的损失函数,x表示模型输入,X表示训练集.

Tang 等人[119]首先对 BERT 的蒸馏进行了探索,将 BERT 中的特定任务知识迁移到了一个单层双向 LSTM 当中,使用均方差(Mean-Squared-Error,MSE)作为蒸馏目标函数,并采用多种数据增强方法对训练集扩充以保证知识蒸馏的高效性. 在测试中取得了与 ELMo 相媲美的成绩且参数量相较于 ELMo 缩小了 98 倍,推理速度提升了 15 倍,这也表明了在大规模模型的指导下,浅层网络模型依然具有较强的学习和建模能力;Sun 等人[120]提出 Patient Knowledge Distillation(PKD)方法,即从“教师”模型的隐藏层中抽取内隐知识而不只是让“学生”模型模仿其输出,使用两种策略:一种是 PKD-Last,使用“教师”模型的最后k层中蕴含的知识;另一种是 PKD-Skip,将“教师”模型中每k层中的知识进行抽取和蒸馏. 图 2 展示了上述两种蒸馏策略,图中 CE Loss 表示交叉熵损失,DS Loss 表示两模型输出距离函数,PT 为隐藏层距离函数. 结果表明 PKD 方法可以有效减少由于蒸馏带来的性能损失,但是较为遗憾的是这一方法模型压缩和推理时间上取得的提升不是很显著. Turc 等人[121]提出一种预训练蒸馏(Pre-trained Distillation,PD)方法,首先使用无监督数据在小规模“学生”模型上进行预训练,而后使用与监督数据分布类似的迁移数据将“教师”模型知识蒸馏至“学生”模型,最后使用监督数据进行微调,这一方法是蒸馏思想在预训练语言模型中又一有效的探索. 与之前工作不同的是,其在预训练阶段就对模型进行

蒸馏,将对后续的研究以更多的启发;Jiao 等人[122]对 PKD 方法进行了更深一步的研究并提出了 TinyBERT 模型,对预训练和微调过程中涉及到的嵌入层、注意力层、隐藏层和预测层均进行了蒸馏操作,并采用数据增强方法对微调阶段使用的监督数据进行扩充,在尽可能保证性能损失较小的同时让模型规模以及推理时间均有显著提升.

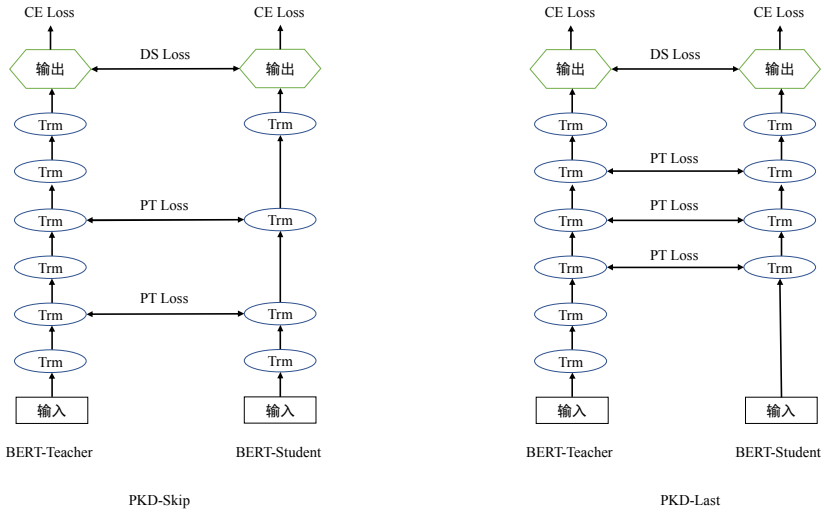

Fig.2 two policy of PKD model schematic diagram

图 2 PKD 方法两种策略示意图[120]

Liu 等人[123]针对 MT-DNN 模型[124]进行了知识蒸馏的研究,由于 MT-DNN 是一种多任务集成学习的预训练语言模型,集成学习方法在泛化能力上有着较好的表现,但是其巨大的规模和训练时间无法进行线上部署,对其进行蒸馏后的小模型可以保留其良好的泛化能力并易于线上部署. 对每个任务训练由多个神经网络形成的集成学习模型,作为“教师”模型,将训练集中的正确标注结果命名为 hard targets,相应的将“教师”模型生成一系列预测结果称为 soft targets,使用 hard targets 与 soft targets 联合训练“学生”模型,以保证其泛化能力不受模型规模的影响,其结果表明在某些任务上蒸馏后的模型表现好于未蒸馏的模型. Yang 等人[125]也采用了类似思路,将多任务学习与知识蒸馏相结合提出了两阶段多教师知识蒸馏方法(Two-stage Multi-teacher Knowledge Distillation,TMKD). 在问答系统任务中使用多个关联任务以及不同的超参数训练多个“教师”模型,将“教师”模型中的知识分为预训练-微调两个阶段蒸馏至“学生”模型中 ,该方法在推理速度上有显著提升. Liu 等人[126]结合前述工作的思想,使用共享层预训练和多任务学习方法,将 BERT 蒸馏至带注意力机制的双向 LSTM 上,提出 BNN(Bi-attentive Student Neural Network)模型,在推理时间和性能上做到了较好的平衡.

除上述提到的知识蒸馏模型外,Xu 等人[127]针对目前知识蒸馏方法存在“学生”模型对“教师”模型质量依赖程度较高的问题,对知识蒸馏的思想进行了扩展,提出 Theseus 压缩方法. 在训练过程中对原始模型中的 Transformer 层进行替换,并减少替换后 Transformer 层数,完成规模压缩. 实验结果显示,其表示能力保留程度达到了 98%,但在模型压缩程度方面还有一定差距.

5.1.2 参数量化

量化(Quantization)实际上就是通过将模型中高精度(32bit 等)矩阵转换为低精度(8bit 等),在矩阵运算以及激活函数等部分中使用低精度来加速推理时间和降低模型对存储空间的要求. 量化方法在计算机视觉领域已经有了较为广泛的应用和深入研究,但在 NLP 领域,由于 GPT、BERT 等语言模型提出之前,语言模型的大小和推理速度是相对可接受的,因此量化方法在预训练语言模型中的工作不够深入.

Cheong 等人[128]最先对 Transformer 结构的量化方法进行了研究,对两种主流的量化算法,K-means 量化算法[129]和二值量化算法[130]在 Transformer 中的实现进行了探索. 在 K-means 量化算法中,将权重矩阵使用聚类

后的簇心索引替换,并使用表格来映射索引和值. 在二值量化方法中,保留原始权重,并在前向计算时将权重中的实际值用两个隐蔽值进行替换. 在实验中 K-means 方法保留了 98. 43%的性能表现并获得了 5.85 倍的压缩效果,二值化方法性能损失较为严重,相应获得了巨大的压缩比例. 但比较遗憾的是,以上两种方法都属于“伪量化”方法,在实际运算中依然使用完整精度进行计算,因此在推理速度上没有提升.

Shen 等人[131]基于 Hessian 信息提出了一种分组处理的混合精度量化方法,使用 Hessian 信息对 BERT 中的各层行为进行分析,提出基于最高特征值的均值和方差的灵敏度测量方法. 在分组处理中,对多头注意力层中的矩阵进行分组,为不同组设定不同的量化范围和查找表,在权重矩阵上获得了 13 倍的压缩率,激活层和嵌入层缩小了 4 倍同时性能损失仅为 2. 3%. Zafrir 等人[132]使用对称线性量化方法,将权重矩阵和激活函数量化至 8bit 整数,并对全连接层和入层中的所有矩阵乘法操作进行了量化,内存空间缩小了四倍且保留了 99%的性能表现. 这一方法在模型规模和推理速度上均获得了提升且性能损失最小,也表明了在预训练语言模型的量化研究上,将量化方法应用至计算过程当中将是提升推理速度的主要方法,但是对量化策略仍需要更深层次的探索.

5.1.3 网络剪枝

剪枝(Pruning)方法旨在对大规模模型中的权重连接、神经元或者权重矩阵取其精华去其糟粕,将不活跃或对模型影响低的结构予以去除,在保证性能的条件下降低模型规模. 目前主要有三种思路的剪枝方法:(1)权重连接剪枝:根据权重大小或其他判定条件,去除层与层之间部分神经元连接,通过对加速稀疏矩阵的运算来完成模型加速;(2)神经元剪枝:通过打分函数,对网络中神经元对输出的贡献进行量化,去除贡献低的神经元以压缩模型规模;(3)权重矩阵剪枝:其主要思路与以上两种方法类似,不再赘述.

Voita 等人[133]对 Transformer 中多头注意力机制的作用和冗余性进行了研究,基于 Louizos 等人[134]对网络权重剪枝的方法,对 Transformer 中的注意力头采用一种基于随机门和宽松L_0乘法项的方法进行剪枝. 在英语-俄语翻译任务中,即使去除了 48 个头中的 38 个,在基准测试中性能的影响也微乎其微. Michel 和 Levy[135]同样对这一问题进行了研究,提出一种基于贪心算法的剪枝策略,这一方法表明在去除 20%-40%的注意力头时对性能的影响较低,在某些层中甚至只是用单头注意力就可以达到未剪枝前的效果. Fan 等人[136]提出 LayerDrop 方法,在训练阶段使用全规模的网络并对多头注意力层中的权重随机 Drop,在测试阶段对网络中的层数按照不同的策略随机 Drop,该方法的性能损失较低,但是其模型的压缩程度和推理速度的提升较为有限. McCarley[137]对前人的工作进行了更深入改进,对 BERT 中的注意力头、隐藏层规模、嵌入层规模都进行了剪枝,提出门替换方法,使用不同的策略控制注意力层、隐藏层、嵌入层中激活的神经元个数,在保证性能损失较小的情况下,使推理时间和空间占用都减少了近 50%.

Guo 等人[138]提出了重加权近端剪枝(Reweighted Proximal Pruning ,RPP)方法,这一方法将重加权L_1最小化方法[139]与近端算法[140]相融合,相较于 NIP(New Iterative Pruning)方法达到了在 59. 3%的剪枝率的情况下没有增加性能损失的表现.

除以上提到的蒸馏、量化和剪枝方法,矩阵分解以及参数共享同样是常用的模型轻量化方法,Lan 等人[141]提出的 ALBERT 模型中,就对嵌入层进行了矩阵分解并对所有层的参数进行了共享,去掉了原始 BERT 中的 NSP 任务,并提出了新的段落顺序预测目标任务. 其参数量的压缩效果显著,推理速度的加速程度较为明显,并且这一工作也证明了 NSP 目标任务对语言模型的建模帮助不足,替换预训练目标任务有助于性能的提升.

5.1.4 模型试验比较

由于目前在网络剪枝和参数量化方面,研究还不深入,已有方法对模型的性能测试还未形成固定模式,不同方法间使用的测试数据各不相同,难以对其性能进行定量的比较和分析,但是知识蒸馏在预训练语言模型中的应用已经较为深入和成熟,方法间性能测试部分基本上使用了相同的下游任务,有利于性能的比较和分析,在本节当中,将对主流知识蒸馏模型间的性能进行比较,并对改进后的性能表现进行分析.

根据表 6 和表 7 的对比,ALBERT-xxlarge 模型[141]虽然参数量达到了 233M,但与 BERT-xlarge 模型 1270M 的参数量及其优异的性能表现相比,可以看出 ALBERT 模型所提出的方法,即嵌入层矩阵分解以及跨层参数共

享并通过多任务联合训练,在多个下游任务中表现超越了人类,并在 QNLI 任务中取得了惊人的准确率,其性能表现甚至超过了很多未进行轻量化处理的模型. 综合来看,ALBERT 模型在各类自然语言推理任务上的表现已经接近人类,而其他知识蒸馏模型则相对原始 BERT 有着较大的性能损失. 这也显示了目前的知识蒸馏方法,尽管使用了不同形式的蒸馏策略以保证其性能,但是依然使模型的推理能力发生了劣化,要求研究人员需要对蒸馏策略进行改进,或者探索模型中影响推理能力的网络结构保留其规模. 另外,表 7 中斜体加粗的指标表示除 ALBERT 外各知识蒸馏模型在不同任务上表现最好的方法,可以看到 TinyBERT[122]以 7.5 倍的模型压缩率,在 MNLI、QNLI 以及 MPRC 四个任务上取得了与最好性能相当接近的表现,在 MNLI、QNLI 以及 MPRC 上与最好性能相差分别仅为 0.3/0.4,1.3 和 0.4,并在 QQP 数据集上取得了好于 BERT 模型的效果. 上述提到的任务基本上均为自然语言推理任务,这表明了 TinyBERT 中所使用的预训练-微调两阶段蒸馏方法,可以相对较好的保留模型的自然语言推理能力. 其消融实验也证明,特定任务蒸馏策略使模型推理能力得到了保留,并且表明对 Transformer 层的蒸馏是至关重要的,其性能贡献非常巨大. BERT-PD[121]以及上述提到的方法亦证明了对预训练阶段的知识蒸馏,对保留模型表示推理是必要的,此外各方法在 SST-2 任务上的表现可以看出,蒸馏方法不会使模型在分类预测任务上的能力有较为显著的降低,即使模型的规模被压缩至原始 BERT 的十分之一,其在情感分析中的性能依然是可接受范围内的. 还应当注意的是,除 ALBERT 外,知识蒸馏方法在 CoLA 任务上的性能降低也是非常巨大的,这也是未来研究人员需要探索的方向.

Table 6 Comparison of Knowledge Distillation Method

表 6 知识蒸馏模型方法比较

模型	压缩方式	参数量
$BERT_{base}$	-	110M(1.0)
ALBERT	嵌入层参数分解 跨层参数共享	12M(9.2x)
TinyBERT	Transformer 层蒸馏 两阶段学习蒸馏	14.5M(7.5x)
DistillBERT	标准知识蒸馏	52.2M(2.1x)
Distill BiLSTM	标准知识蒸馏 BERT→单层 BiLSTM	**10.1M(10.8x)**
BERT-PKD	潜在知识蒸馏	67.0M(1.6x)
BERT-PD	预训练知识蒸馏	67.5M(1.6x)
BNN	共享层预训练 多任务学习+数据增强	10.2M(10.8x)
BERT-TMKD	预训练知识蒸馏 特定任务微调蒸馏	45.7M(2.4x)

Table 7 Performance Comparison of Knowledge Distillation Language Models

表 7 知识蒸馏语言模型性能表现对比

模型	参数量	MNLI	QQP	SST-2	QNLI	MPRC	RTE	CoLA	STS-B
人类	-	92.0	59.5/80.4	97.8	91.2	86.3/80.8	93.6	66.4	92.7/92.6
$BERT_{base}$	110M (1.0)	84.6	71.2/-	93.5	90.5	88.9/-	66.4	52.1	85.8/-
$ALBERT_{xxl}$	233M (0.5x)	**91.3**	**74.2/90.5**	**97.1**	**99.2**	**93.4/91.2**	**89.2**	**69.1**	**92.5/92.0**
TinyBERT	14.5M (7.5x)	82.5	***71.3/89.2***	***92.6***	87.7	86.4/81.2	62.9	***43.3***	***81.2/79.9***
Distill BERT	52.2M (2.1x)	78.9	68.5/-	91.4	85.2	82.4	54.1	32.8	76.1/-
Distill BiLSTM	**10.1M (10.8x)**	73.0	68.2/88.1	90.7	-	-	-	-	-
BERT-PKD	67.0M (1.6x)	81.5	70.7/88.9	92.0	***89.0***	85.0/79.9	65.5	24.8	-
BERT-PD	67.5M (1.6x)	***82.8***	70.4/88.9	91.8	88.9	***86.8/81.7***	65.3	-	-
BNN	10.2M (10.8x)	78.6	70.7/88.6	91.0	85.4	85.4/79.7	***67.3***	-	80.9/80.9
BERT-TM KD	45.7M (2.4x)	80.4	-	-	86.4	-	67.5	-	-

MRPC: F1/accuracy, STS-B: Pearson/Spearmanr correlation,QQP: F1/accuracy, MNLI: matched/mistached accuracies

5.2 预训练语言模型与知识

将外源知识引入预训练语言模型以获得表示能力更丰富鲁棒性更强的语言模型,是目前语言模型研究领域的一个新课题. 值得注意的是,Google 公司提出的知识图谱技术在信息检索及其他交叉领域已经展现出了巨大的影响力. 引入外源知识后形成的丰富表示将对下游任务中的情感分析、文本分类、关系抽取等任务带来提升.

在非直接实体注入方面,Levine 等人[142]认为 BERT 出现后的一系列自监督目标任务的改进方法,在运行方式上是相似的. 但显然,语言中部分单词的含义具有多重性,Levine 等人的工作在单词语义层面提出一种隐蔽语义训练任务,模型需要预测缺失单词的含义. 在 WordNet 的辅助下完成自监督学习. 该工作是具有启发性的,对后续预训练语言模型训练目标的改进给出了一种新方向. 针对预训练语言模型中获取的知识仅来自于无监督文本的问题,Lauscher 等人[143]提出将现实中的语义相似性信息融入语言模型中,模型联合训练 MLM、NSP 与词汇关系分类(Lexical Relation Classifier,LRC)三个目标任务,让模型学习语言中的同义词和上下位词知识,提升模型表示能力.

清华大学的 Zhang 等人[144]提出 Enhanced Language Representation with Informative Entities (ERNIE)模型,将知识图谱中的实体知识与 BERT 模型相结合,提出了结构化知识编码以及异构信息融合问题的解决方法. 为了对结构化信息进行编码,采用知识嵌入方法(如 TransE)先对知识图谱进行编码,然后作为 ERNIE 模型的输入. 在训练过程中,提出新的目标任务,对输入文本中的实体随机隐蔽,模型从知识图谱中选择正确的实体,完成实体与知识图谱的对齐. Liu 等人[145]对知识图谱引入预训练语言模型后存在的异构嵌入空间和知识噪声的问题提出了解决方法,将知识图谱中的三元组注入输入文本中生成句子树,然后把句子树转化为向量表示输入带隐蔽矩阵的 Transformer 中,缓解了异构表示空间带来的问题. 该工作所提出带隐蔽矩阵的 Transformer 可以有效

减低知识噪声对文本语义的影响.

类似地,Wang 等人[146]将知识嵌入技术引入预训练语言模型中. 但是该模型的创新点在于,将知识嵌入任务与语言模型目标任务联合训练,并引入了负采样技术,加速模型收敛.

Peters 等人[147]同样对将结构化知识编码到 BERT 中进行了探索,其提出知识注意力与重语境化(Knowledge Attention and Recontextualization ,KAR)方法,将 KAR 方法设计为一个单独的处理层,接受上一层的输出作为输入,将输入影射至实体维度中再进行实体链接,并采用多层感知机进行重语境化,将输出继续送入 Transformer 中. 在实体分类、实体识别等任务中性能表现较好.

5.3 多模态方法

BERT 的问世不仅对 NLP 领域产生了巨大的影响,还出现了一系列的变种模型,多模态(Cross-Modal)学习的研究自 BERT 提出后成为了一个研究新热点,目前而言主要涉及视觉和文本两种模态的融合. 就现在多模态学习在预训练语言模型中的研究进展而言,在网络结构上可大致分为两类:一类直接将视觉和文本流进行跨模态预训练;另一类则是先对两个模态编码,使用编码后的表示进行跨模态融合.

5.3.1 直接跨模态学习

Sun 等人[148]提出 VideoBERT 模型,这一模型是最先对预训练语言模型在多模态学习任务中提出的工作,在训练过程中使用 YouTube 网站采样到的视频与对应字幕作为无监督数据,使用向量量化(vector quantization)方法对采样到的视频帧进行表示,将隐蔽模型作为目标任务(对输入图片转化后的嵌入向量同样进行隐蔽)跨模态训练.

Alberti 等人[149]提出 B2T2 方法,采用预训练后的 ResNet-152[150]作为视觉流的特征抽取器,并使用分块检测后的结果作为输入而不是视频帧,提出早期-晚期二阶段模态融合方法,在早期融合阶段,没有将单独的图像作为序列输入,而是在原始输入中被隐蔽的单词位置输入了该词提到的图像区域块特征.

Li 等人[151]提出的 Unicoder-VL 模型实际上可以看作是以上提到的工作(VideoBERT 和 B2T2)的融合,其网络结构与预训练任务基本与 VideoBERT 相同,视觉流则采用 Faster R-CNN 识别后的区域块特征作为模型输入,与 B2T2 相同,在隐蔽目标任务中,使用被隐蔽词对应的视觉表示作为输入. 融合方法在部分任务中有性能提升,但是其创新性一般,仅是两个工作的简单融合,对模态融合方式以及预训练目标任务均没有改进. Su 等人[152]在 Unicoder-VL 的基础上提出了 VL-BERT 模型,其模型结构基本一致,但是在输入中将视觉流中的完整表示与文本的嵌入表示一同输入到模型中.

Chen 等人[153]提出了 UNITER 模型,在该工作中建模方式与 VideoBERT 方式相同,将抽取后的视觉流与文本流同时输入 Transformer 中,在更大规模的数据集上进行了预训练,并引入三种训练目标任务:隐蔽语言模型、隐蔽区域建模、图文匹配任务. 这一方法的性能提升主要来自于丰富的预训练目标以及大规模的预训练数据集,在网络结构以及融合方式上基本无改进.

5.3.2 编码后跨模态学习

Lu 等人[154]针对模态融合与预训练的问题进行了改进,提出 ViLBERT 模型,对图片的处理没有采用向量量化的方法而是用监督数据在 Faster R-CNN 网络[155]上进行预训练,在模态融合阶段提出双流 Co-attention Transformer 层,交换视觉流和文本流中的 Key 和 Value 矩阵以完成联合训练. 引入隐蔽多模态学习任务和多模态对齐任务,在一系列多模态下游任务中取得了显著提升. Sun 等人[156]继续对之前提出的 VideoBERT 进行了改进,认为向量量化会损失视觉流输入中有效的信息,限制模态融合的性能,提出对比双向 Transformer(Contastive Bidirectional Transformer,CBT),对使用 S3D 框架[157]处理后的视觉流输入使用对比噪声估计方法进行预训练. Tan 等人[158]提出 LXMBERT 模型,与 ViLBERT 相比,LXMBERT 在视觉流编码后加入了 Encoder 网络,在预训练过程中引入多达 5 种预训练目标任务,性能提升显著.

5.3.3 模型方法比较

表 8 对比了目前跨模态语言模型的结构、视觉流输入形式、预训练目标任务以及对应微调时的下游任

务. 根据对比结果,将图片区域分割识别后作为视觉流的输入是目前以及未来主流的视觉流处理方式,从目前性能表现最好的 UNITER 模型与 LXMERT 模型对比来说,是否对视觉流和文本流使用 Transformer 进行编码和投影是存在争议的. LXMERT 在某些任务中表现好于 UNITER,但是 UNITER 在预训练过程中加入了丰富的目标任务. 同时 UNITER 在 Large(即更大规模的 Transformer 以及堆叠层数)模型的测试中,面对各项任务中的表现均好于当前已有方法,亦表明现阶段跨模态语言模型的性能提升,主要来自于规模更大类型更丰富的预训练数据集,以及捕获模态间内在交互方式的预训练目标任务. 这其中的原因是跨模态领域的数据集还不能像纯文本模态的语言模型,拥有大量的无监督数据可供模型学习,另一方面与对 BERT 的一系列改进方法类似,层次更深入的目标任务可以使模型学习不同模态间的内在联系.

Table 8 Comparison of Cross-Modal Method

表 8 跨模态方法比较

模型	模型结构	视觉流格式	预训练任务	下游任务
VideoBERT	一个跨模态 Transformer	视频帧	句子-图片对齐 隐蔽语言模型 隐蔽视频-单词预测	零次学习动作分类 视频字幕
CBT	两个单模态 Transformer + 一个跨模态 Transformer	视频帧	句子-图片对齐 隐蔽语言模型 隐蔽视频特征恢复	动作预测 视频字幕
ViLBERT	一个单模态 Transformer + 一个跨模态 Transformer	图片区域分割	句子-图片对齐 隐蔽语言模型 隐蔽视频特征分类	视频问题回答+视频常识推理+图片重构+零次学习+图片重构
B2T2	一个跨模态 Transformer	图片区域分割	句子-图片对齐 隐蔽语言模型	视频常识推理
LXMERT	两个单模态 Transformer + 一个跨模态 Transformer	图片区域分割	句子-图片对齐+隐蔽语言模型+隐蔽视频特征分类+隐蔽视频特征恢复+视频问答	视频问题回答 自然语言视觉推理
Unicoder-VL	一个单模态 Transformer	图片区域分割	句子-图片对齐 隐蔽语言模型 隐蔽视频特征分类	图片-文本重构 零次学习图片文本重构
VL-BERT	一个单模态 Transformer	图片区域分割	隐蔽语言模型 隐蔽视频特征分类	视频问答 视频常识推理
UNITER	一个跨模态 Transformer	图片区域分割	隐蔽语言模型 隐蔽区域建模 图片-文本匹配	视频问答 视频常识推理 自然语言视频推理

5.4 跨语言方法

由于相当一部分自然语言的文本数据是低资源的,无法获得大规模的无监督文本数据,自预训练语言模型提出后,研究人员开始探索预训练模型能否学习不同语言间的一致性,使模型在低资源语言的下游任务上仍保持良好的性能表现.

紧随 multilingual BERT 之后,Lample 等人[159]也提出了跨语言模型(cross-lingual language model,XLM),模型中所有语言的字典通过 BPE 方式生成,并被所有语言共享. 在预训练目标任务中除隐蔽语言模型外,Lample 等人提出了翻译语言模型(Translation Language Modeling ,TLM),将对齐后的不同语言输入模型中,使模型学习语言间对齐的表示. 该工作提出的 TLM 目标任务对后续研究影响深远,几乎成为了跨语言模型中的基本目标任务. Wu 等人[160]在 XLM 模型的基础上,未对模型结构进行修改,仅在 Encoder 部分中的一些层进行了参数共享;而后 Conneau 等人[161]在 XLM 模型上进行了扩展,将语言数量扩展至 100 种,并对训练数据进

行了大规模扩充,结果表明语言数目和训练数据规模同样可能限制跨语言模型的发挥.

Eisenschlos 等人[162]在 ULMFiT 模型的基础上提出了 MultiFiT 模型,并将原模型中 LSTM 替换为 QRNN[163],获得了 16 倍的速度提升,在生成子词字典时采用了文献[164]的方法,相较于 BPE 更为灵活,并加入了标签平滑算法(label smoothing)[165]和单周期余弦策略[166],将 Bootstrapping 方法[167]作为跨语言训练策略.

Huang 等人[168]提出 Unicoder 模型,模型结构同样基于 XLM 模型,在预训练目标任务上,提出了三种新的任务:(1)跨语言单词恢复(Cross-lingual Word Recovery),这一任务旨在让模型学习如何对齐两个语言的单词;(2)跨语言含义分类(Cross-lingual Paraphrase Classification),让模型判断输入的两种语言的含义是否相同;(3)跨语言隐蔽语言模型.

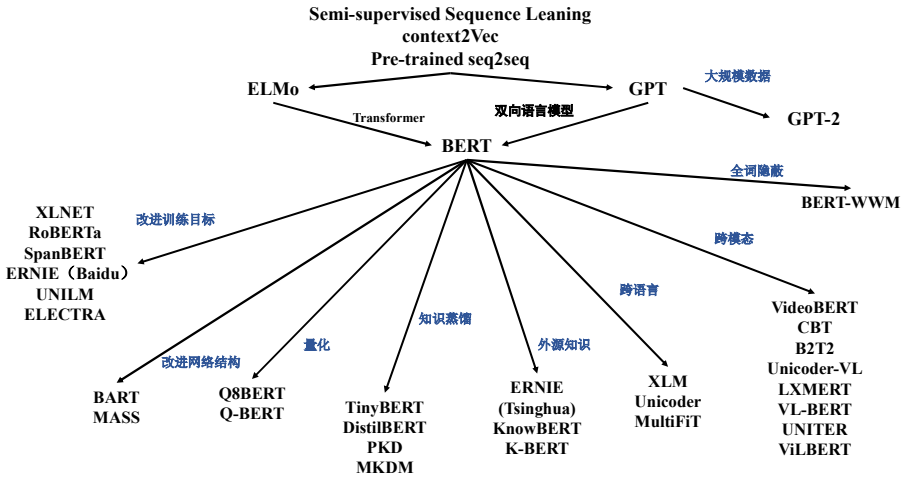

Fig.3 development of pretraining language model

图 3 预训练语言模型发展历程示意图

图 3 展示了目前预训练语言模型发展过程,本文在 Wang 等人[8]工作的基础上进一步对预训练语言模型的发展过程和发展方向进行了梳理和分类. 在研究早期,半监督学习、端到端框架以及预训练等思想的提出,为预训练语言模型奠定了理论基础. 而后 ELMo 模型[37]创新性的采用了双向语言模型,但没有使用 Transformer 网络,不同的是,GPT 模型[78]使用 Transformer 作为特征抽取器,但却只进行了单向语言建模. BERT[35]作为两者的集大成者,同时提出了两种训练目标,取得了革命性的进展,在语言模型领域掀起了研究热潮,一系列基于 BERT 针对不同方向的工作被提出. 从图 3 来看,研究人员对其应用的各个方面进行了探索,上文中总结了预训练语言模型的主要发展历程和方向. 关于模型轻量化,已经取得了不错的进展,TinyBERT[122]以及 ALBERT[141]等模型的提出证明了模型轻量化与性能损失之间是可以达到平衡并进一步提升的. 但是对于通过参数量化进行模型压缩这一方法,量化后的模型性能损失较为严重,更好的量化策略亟待研究. 而对于跨语言、多模态融合以及知识融合方面,由于表示能力如此优秀的模型(即 BERT)仅提出一年多的时间,相关的研究还不深入. 用于训练和测试的数据集还比较匮乏,不同语言不同模态的知识相融合的方法还比较单一. 对于外源知识,解决原始语言模型和外源知识表示空间不一致的方法还不够好,仍然存在诸多需要解决的问题. 在模型网络结构和目标任务改进方面,已经有不少相当优秀的模型和方法被提出,例如 XLNET[38]、ERNIE(Baidu)[91]以及 ELECTRA[93]等,给后续的研究提供了启发. 但是不可否认预训练语言模型虽表现优异,依然远没有达到真正的人工智能所需要的水平,还面临着诸多挑战和争议.

8 https://github.com/thunlp/PLMpapers

6 基于深度学习的语言模型研究趋势展望

近年来,基于深度学习的预训练语言模型已引起研究人员的广泛关注,特别是2018年以来,国内外对预训练语言模型的研究成果呈井喷式出现,对诸多下游任务的性能表现提升显著,在某些任务中甚至超过了人类的表现. 对于预训练语言模型面临的挑战,研究人员已经针对一些领域提出了解决方案. 在本节当中将对预训练语言模型的未来研究趋势进行展望.

6.1 对模型轻量化的研究

随着目前移动设备算力的进一步提升,将性能优异的预训练模型在低计算资源设备和边缘设备上面部署成为了可能,未来对语言模型进行压缩的方法,较大可能依然是上文中提到的蒸馏、量化和剪枝三个方向. 在知识蒸馏方向,一方面可以对现有的蒸馏目标函数进行优化,以更细粒度的形式度量“教师”模型与“学生”模型之间知识迁移的程度. 另一方面,文献[119]提出的工作也为知识蒸馏研究提供了思路:(1)可将大规模模型蒸馏至更为简单的网络结构,甚至是支持向量机(Support Vector Machine,SVM)或者逻辑斯蒂回归模型;(2)可将注意力机制引入到蒸馏过程中. 文献[122]的工作也表明,在蒸馏过程中,注意力层、全连接层以及嵌入层中的内隐知识也可以通过构建目标函数进行迁移,在上文的实验对比中也可以看到,如何最大程度降低蒸馏过程对模型自然语言推理能力的损失也是未来知识蒸馏模型研究的一个重要方向. 在参数量化方面,由于超低精度的量化会带来巨大的性能损失,因此更好的量化策略以及混合精度量化、多阶段量化,根据语言模型各层对量化精度要求的不同,采用适合的量化策略,将是未来参数量化的主要方向. 此外,参数量化与模型剪枝的融合也将是语言模型轻量化的一个重要思路.

6.2 多模态融合语言模型研究

语言模型的多模态融合在未来一段时间内仍将以视觉流和文本流的融合为主,跨模态的融合特征抽取器现阶段多数采用的是交叉 Transformer,其融合方式较为单一,注意力机制在融合过程中应有的性能没有完全发挥. 另一方面,当前已有工作的网络结构主要分为两种:一种是直接的跨模态融合;另一种是对视觉流和文本流编码后再进行跨模态融合. 文献[158]提出的工作证明了,对两种模态编码后融合的整体思路是更为有效的. 如何最大限度保留编码后的信息,提出融合表示能力更强的网络结构,以及将计算机视觉领域的先进方法引入模型中,将会是多模态融合语言模型的主要研究趋势.

6.3 跨语言融合语言模型研究

为解决低资源语言的自然语言理解问题,研究人员提出了跨语言融合的预训练语言模型,通过将高资源语言与低资源语言联合训练,以抽取不同种语言之间内在语义的一致性,完成模型训练知识的迁移. 目前,跨语言模型的研究仍处于早期阶段,大部分工作主要集中在预训练目标任务的改进这一方向. 对不同语言的编码、融合方式以及网络结构上的改进还不深入,文献[161]的工作证明多语言监督数据集生成和增强算法对于跨语言模型的性能提升是有显著作用的,未来该方向仍存在巨大的改进空间.

6.4 与知识图谱融合的语言模型研究

知识图谱在信息检索领域已经展现出了巨大潜力,作为自然语言处理、信息检索和知识表示领域的交叉课题,将知识图谱中丰富的内在知识和内部推理形成的信息融入预训练语言模型当中,是进一步增强预训练语言模型自然语言理解和推理能力的重要思路. 在知识图谱或知识库信息的融入过程中,其两者表示空间不一致,是研究人员亟待解决的一个问题. 另一方面,文献[144]的工作也提出,由于较早语言模型相对类 BERT 模型在训练和推理速度上有着天然的优势,将外源知识引入 ELMo 等早期预训练模型以增强其语言推理和理解能力,可以使预训练模型在部署和应用层面有更广阔的空间. 另外,通过外源知识的补充和引导,对预训练数据进行启发式指导下的数据增强,是未来知识图谱与语言模型融合的一个切入点. 目前的各类工作已经证明,更丰富更大规模的预训练数据对语言模型的表示能力有着直接的提升作用,这一思路也将是外源知识在预训练语言模型领域的一个应用方向.

6.5 基于新网络结构的语言模型研究

预训练语言模型能取得如此举世瞩目的成就,与 Transformer 结构以及自注意力机制的提出密不可分. 最近,Moradshahi 等人[169]提出了 HUBERT 模型,在模型中加入了张量积表示(Tensor-Product Representation,TPR)层,将 BERT 表示分解为内容和形式两部分,在模型的自然语言推理能力上取得了较大提升. 另外,文献[170]提出单头注意力 RNN(Single Headed Attention RNN,SHA RNN),在压缩模型规模的同时性能损失极低. 这些工作都表明无论是预训练-微调框架中的编码解码部分,还是 Transformer 结构都存在着改进的空间,尤其是从 Transformer 的并行性和规模压缩这一思路进行探索,作为预训练语言模型的基础结构,新的注意力机制、改进 Transformer 以及先前工作与 Transformer 的结合都将是从根本上改进语言模型性能的研究方向.

6.6 预训练语言模型可解释性的研究

虽然深度学习在语言模型领域中应用广泛且效果显著,在语言模型的研究中脱离深度学习的帮助已经成为了几乎不可能完成的事情,但是无论是神经网络还是目前从语言模型中提取出的单词表示,其可解释性一直都是被部分学者质疑和研究的问题. 语言模型训练后生成的稠密实数向量表示,目前还难以解释其每一维度的含义是什么,为什么预训练语言模型可以在多种任务上同时获得卓越的性能表现,其内在的交互作用机制,目前尚未明了. 但是已经一些研究被提出,文献[171]就通过对 BERT 中自注意力层权重编码,对 BERT 如何捕获各种语言学信息的过程进行分析和可视化;文献[172]就以 BERT 在问答任务上的微调过程为切入点,对模型中的隐状态进行了分析,结果表明 BERT 可以将特定任务的信息隐式地合并到单词表示中. 在未来研究中,对预训练语言模型的内在机理以及注意力机制的交互方式进行分析和解释,会是主要的研究方向.

7 结束语

语言模型,被视为自然语言处理领域的一个基础课题,一直以来是自然语言处理领域的研究热点. 基于深度学习的语言模型,无论是神经概率语言模型还是预训练语言模型,在多个方向上取得了令人瞩目的进展,并推动了下游任务的性能提升. 通过对以上语言模型研究代表方法的梳理,本文认为基于深度学习的语言模型研究具有重要的意义:(1)神经概率语言模型特别是 Word2Vec,在 NLP 研究早期,对序列标注、文本分类等任务产生了重要的推动作用;(2)预训练语言模型以新的思路完成自然语言建模,在多类下游任务中取得了超越人类的性能表现,这将促使基于语言模型的各类应用更具智能性;(3)在对预训练模型的改进中,涌现出了多种表示能力强、计算效率高的新型网络结构,这些网络可以被迁移至其他任务或领域推动进一步性能提升;(4)目前,已经有一部分具有代表性的预训练模型框架被提出,框架融合了迁移学习和持续学习的思想,可以同时在多类下游任务上获得性能改进,对它们的研究将从根本上促进人工智能的发展.

本文还对基于深度学习的语言模型目前面临的挑战、已有解决方案以及未来发展趋势进行了阐述. 梳理了近年来神经概率语言模型的发展脉络,进一步对当前预训练语言模型的发展情况、研究方向进行了概述、分析和评价. 最后对基于深度学习的语言模型在未来轻量化、多模态、跨语言以及可解释性等方向的研究趋势进行了展望. 期待能有更多研究人员参与到语言模型的研究工作当中,也希望本文能对国内有关基于深度学习的语言模型的研究提供一些帮助.

References:

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. nature, 2015, 521(7553): 436-444.

[2] Durand T, Mehrasa N, Mori G. Learning a deep convnet for multi-label classification with partial labels. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 647-657.

[3] Li Y, Chen X, Zhu Z, et al. Attention-guided unified network for panoptic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7026-7035.

[4] Wang Q, Li B, Xiao T, et al. Learning deep transformer models for machine translation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1810-1822.

[5] Fu T J, Li P H, Ma W Y. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1409-1418.

[6] Yu T, Shen Y, Jin H. An visual dialog augmented interactive recommender system. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019: 157-165.

[7] Xiao W, Zhao H, Pan H, et al. Beyond personalization: social content recommendation for creator equality and consumer satisfaction. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 235-245.

[8] Lim S H, Xu H, Mannor S. Reinforcement learning in robust markov decision processes. Advances in Neural Information Processing Systems. 2013: 701-709.

[9] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529.

[10] Kim S, Dalmia S, Metze F. Gated embeddings in end-to-end speech recognition for conversational-context fusion. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1131-1141.

[11] Hosseini-Kivanani N, Vásquez-Correa J C, Stede M, et al. Automated cross-language intelligibility analysis of Parkinson's disease patients using speech recognition cechnologies. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. 2019: 74-80.

[12] Chen X, Liu P, Sun Y, et al. Research on disease prediction models based on imbalanced medical data sets. Chinese Journal of Computers,2019,42(03):596-609(in Chinese with English abstract).

[13] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. nature, 2016, 529(7587): 484.

[14] Tian Y, Zhu Y. Better computer go player with neural network and long-term prediction. arXiv preprint arXiv:1511.06410, 2015.

[15] LANDAHL H D, MCCULLOCH W S, PITTS W. A statistical consequence of the logical calculus of nervous nets. Bulletin of Mathematical Biology, 1943, 5(4):135-137.

[16] Hopfield J J. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 1982, 79(8): 2554-2558.

[17] Aarts E H L, Korst J H M. Simulated annealing and Boltzmann machines: a stochastic approach to combinatorial optimization and neural computing. Wiley-Interscience series in discrete mathematics and optimization, 1989.

[18] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. Cognitive modeling, 1988, 5(3): 1.

[19] Broomhead D S, Lowe D. Radial basis functions, multi-variable functional interpolation and adaptive networks[R]. Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

[20] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation, 1997, 9(8): 1735-1780.

[21] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks, 2005, 18(5-6): 602-610.

[22] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems. 2017: 5998-6008.

[23] Garimella A, Banea C, Hovy D, et al. Women's syntactic resilience and men's grammatical luck: gender-bias in part-of-speech tagging and dependency parsing. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3493-3498.

[24] Liu T, Yao J, Lin C Y. Towards improving neural named entity recognition with gazetteers. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5301-5307.

[25] Cai R, Lapata M. Syntax-aware semantic role labeling without parsing. Transactions of the Association for Computational Linguistics, 2019, 7: 343-356.

[26] Wang Q, Li B, Xiao T, et al. Learning deep transformer models for machine translation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1810-1822.

[27] Kim H, Bansal M. Improving visual question answering by referring to generated paragraph captions. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3606-3612.

[28] He R, Lee W S, Ng H T, et al. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 504-515.

[29] Nallapati R, Zhou B, dos Santos C, et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond. Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. 2016: 280-290.

[30] Haj-Yahia Z, Sieg A, Deleris L A. Towards unsupervised text classification leveraging experts and word embeddings. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 371-379.

[31] Xu S, Li P, Kong F, et al. Topic Tensor Network for Implicit Discourse Relation Recognition in Chinese. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 608-618.

[32] Huang J. Study of application of a language model combining statistics and rules in Chinese input method[D]. XiDian University,2008 (in Chinese with English abstract).

[33] Bahl L R, Jelinek F, Mercer R L. A maximum likelihood approach to continuous speech recognition[M]// Readings in speech recognition. Morgan Kaufmann Publishers Inc. 1990.

[34] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. Journal of machine learning research, 2003, 3(Feb): 1137-1155.

[35] Devlin J, Chang M W, 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, Volume 1 (Long and Short Papers). 2019: 4171-4186.

[36] Forney G D. The viterbi algorithm. Proceedings of the IEEE, 1973, 61(3): 268-278.

[37] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 2227-2237.

[38] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019: 5754-5764.

[39] Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1724-1734.

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

[41] Zhang ZC, Zhang ZW, Zhang ZM. User intent classification based on IndRNN-Attention. Journal of Computer Research and Development, 2019,56(07):1517-1524(in Chinese with English abstract).

[42] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[43] Zhou JZ, Zhu ZK, He ZQ, et al. Hybrid neural network models for human-machine dialogue intention classification. Ruan Jian Xue Bao/Journal of Software, 2019,30(11):3313−3325 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5862.htm

[44] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[45] Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules[C]. In: Advances in Neural Information Processing Systems. MIT Press, 2017. 3856−3866.

[46] Li C, Chai YM, Nan XF, Gao ML. Research on problem classification method based on deep learning. Computer Science, 2016, 43(12):115−119 (in Chinese with English abstract).

[47] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]. In: Advances in Neural Information Processing Systems. MIT Press, 2014. 2204−2212.

[48] Du H, Xu XK, Wu DY, et al. A sentiment classification method based on sentiment-specific word embedding. Journal of Chinese Information Processing,2017,31(03):170-176(in Chinese with English abstract).

[49] Zhu SY, Li SS, Zhou GD. Multi-dimensional emotion regression via adversarial neural network. Ruan Jian Xue Bao/Journal of Software, 2019,30(7):2091-2108 (in Chinese with English abstract). http.//www.jos.org.cn/1000-9825/5838.htm

[50] Liu YP, Ma CG, Zhang YN. Hierarchical machine translation model based on deep recursive neural network[D]Chinese Journal of Computers,2017,40(04):861-871(in Chinese with English abstract).

[51] Liang XB, Ren FL, Liu YK, et al. N-Reader: Machine reading comprehension based on double layers of self-attention. Journal of Chinese Information Processing, 2018,32(10):130-137(in Chinese with English abstract).

[52] Vijayakumar A, Vedantam R, Parikh D. Sound-word2vec: learning word representations grounded in sounds. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 920-925.

[53] Sun C, Huang L, Qiu X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 380-385.

[54] Karimi A, Rossi L, Prati A, et al. Adversarial training for aspect-based sentiment analysis with BERT. arXiv preprint arXiv:2001.11316, 2020.

[55] Xu H, Liu B, Shu L, et al. BERT post-training for review reading comprehension and aspect-based sentiment analysis. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 2324-2335.

[56] Song Y, Wang J, Liang Z, et al. Utilizing BERT intermediate layers for aspect based sentiment analysis and natural language inference. arXiv preprint arXiv:2002.04815, 2020.

[57] Li S, Cui W, Liu Y, et al. PEL-BERT: A joint model for protocol entity linking. arXiv preprint arXiv:2002.00744, 2020.

[58] Tsai H, Riesa J, Johnson M, et al. Small and practical bert models for sequence labeling. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3623-3627.

[59] Chen Q, Zhuo Z, Wang W. Bert for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909, 2019.

[60] Gulyaev P, Elistratova E, Konovalov V, et al. Goal-oriented multi-task bert-based dialogue state tracker. arXiv preprint arXiv:2002.02450, 2020.

[61] Yang ZC. Quality estimation of machine translation using pre-training language model[D]. Beijing Jiaotong University,2019(in Chinese with English abstract).

[62] Wolf L, Hanani Y, Bar K, et al. Joint word2vec networks for bilingual semantic representations. Int. J. Comput. Linguistics Appl., 2014, 5(1): 27-42.

[63] Xu W, Rudnicky A. Can artificial neural networks learn language models?. Sixth International Conference on Spoken Language Processing. 2000.

[64] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model. Eleventh annual conference of the international speech communication association. 2010.

[65] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. Journal of machine learning research, 2011, 12(Aug): 2493-2537.

[66] Bengio Y, Senécal J S. Quick training of probabilistic neural nets by importance sampling. AISTATS. 2003: 1-9.

[67] Bengio Y, Senécal J S. Adaptive importance sampling to accelerate training of a neural probabilistic language model. IEEE Transactions on Neural Networks, 2008, 19(4): 713-722.

[68] Mnih A, Teh Y W. A fast and simple algorithm for training neural probabilistic language models. Proceedings of the 29th International Coference on International Conference on Machine Learning. 2012: 419-426.

[69] Gutmann M, Hyvärinen A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010: 297-304.

[70] Elman J L. Finding structure in time. Cognitive science, 1990, 14(2): 179-211.

[71] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, 2015.

[72] Schuster M, Paliwal K K. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 1997, 45(11): 2673-2681.

[73] Okanohara D, Tsujii J. A discriminative language model with pseudo-negative samples. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007: 73-80.

[74] Huang E H, Socher R, Manning C D, et al. Improving word representations via global context and multiple word prototypes. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012: 873-882.

[75] Reisinger J, Mooney R J. Multi-prototype vector-space models of word meaning. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 109-117.

[76] Mikolov T, Kopecky J, Burget L, et al. Neural network based language models for highly inflective languages. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2009: 4725-4728.

[77] Mikolov T. Language modeling for speech recognition in czech. Ph. D. dissertation, Masters thesis, 2007.

[78] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[OL]. [2019-09-30] https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf

[79] Dai A M, Le Q V. Semi-supervised sequence learning. Advances in neural information processing systems. 2015: 3079-3087.

[80] Jozefowicz R, Vinyals O, Schuster M, et al. Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410, 2016.

[81] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI Blog, 2019, 1(8).

[82] Caruana R. Multitask learning. Machine learning, 1997, 28(1): 41-75.

[83] Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.

[84] Cui Y, Che W, Liu T, et al. Pre-Training with Whole Word Masking for Chinese BERT. arXiv preprint arXiv:1906.08101, 2019.

[85] Dai Z, Yang Z, Yang Y, et al. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2978-2988.

[86] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[87] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 1715-1725.

[88] Joshi M, Chen D, Liu Y, et al. Spanbert: Improving pre-training by representing and predicting spans. arXiv preprint arXiv:1907.10529, 2019.

[89] Dong L, Yang N, Wang W, et al. Unified language model pre-training for natural language understanding and generation. Advances in Neural Information Processing Systems. 2019: 13042-13054.

[90] Song K, Tan X, Qin T, et al. MASS: Masked sequence to sequence pre-training for language generation. International Conference on Machine Learning. 2019: 5926-5936.

[91] Sun Y, Wang S, Li Y, et al. Ernie 2.0: A continual pre-training framework for language understanding. arXiv preprint arXiv:1907.12412, 2019.

[92] Wang W, Bi B, Yan M, et al. StructBERT: Incorporating language structures into pre-training for deep language understanding. arXiv preprint arXiv:1908.04577, 2019.

[93] Clark K, Luong M T, Le Q V, et al. ELECTRA: Pre-training text encoders as discriminators rather than generators. International Conference on Learning Representations. 2019.

[94] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in neural information processing systems. 2014: 2672-2680.

[95] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683, 2019.

[96] Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.

[97] Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. Proceedings of the IEEE international conference on computer vision. 2015: 19-27.

[98] Parker R, Graff D, Kong J, et al. English gigaword fifth edition, linguistic data consortium. Google Scholar, 2011.

[99] Callan J, Hoy M, Yoo C, et al. Clueweb09 data set. 2009.

[100] Cui Y, Liu T, Che W, et al. A span-extraction dataset for chinese machine reading comprehension. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 5886-5891.

[101] Shao C C, Liu T, Lai Y, et al. Drcd: a chinese machine reading comprehension dataset. arXiv preprint arXiv:1806.00920, 2018.

[102] He W, Liu K, Liu J, et al. DuReader: A chinese machine reading comprehension dataset from real-world applications. Proceedings of the Workshop on Machine Reading for Question Answering. 2018: 37-46.

[103] Levow G A. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 108-117.

[104] Conneau A, Rinott R, Lample G, et al. XNLI: Evaluating cross-lingual sentence representations. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2475-2485.

[105] Liu X, Chen Q, Deng C, et al. Lcqmc: A large-scale chinese question matching corpus. Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1952-1962.

[106] Chen J, Chen Q, Liu X, et al. The bq corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 4946-4951.

[107] Li J, Sun M. Scalable term selection for text categorization. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007: 774-782.

[108] Warstadt A, Singh A, Bowman S R. Neural network acceptability judgments. Transactions of the Association for Computational Linguistics, 2019, 7: 625-641.

[109] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1631-1642.

[110] Williams A, Nangia N, Bowman S. A broad-coverage challenge corpus for sentence understanding through inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 1112-1122.

[111] Bentivogli L, Clark P, Dagan I, et al. The fifth PASCAL recognizing textual entailment challenge. TAC. 2009.

[112] Levesque H, Davis E, Morgenstern L. The winograd schema challenge. Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning. 2012.

[113] Dolan W B, Brockett C. Automatically constructing a corpus of sentential paraphrases. Proceedings of the Third International Workshop on Paraphrasing (IWP2005). 2005.

[114] Cer D, Diab M, Agirre E, et al. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017: 1-14.

[115] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100,000+ Questions for machine comprehension of text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2383-2392.

[116] Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for SQuAD. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018: 784-789.

[117] Lai G, Xie Q, Liu H, et al. RACE: Large-scale reading comprehension dataset from examinations. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 785-794.

[118] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[119] Tang R, Lu Y, Liu L, et al. Distilling task-specific knowledge from BERT into simple neural networks. arXiv preprint arXiv:1903.12136, 2019.

[120] Sun S, Cheng Y, Gan Z, et al. Patient knowledge distillation for BERT model compression. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 4314-4323.

[121] Turc I, Chang M W, Lee K, et al. Well-read students learn better: The impact of student initialization on knowledge distillation. arXiv preprint arXiv:1908.08962, 2019.

[122] Jiao X, Yin Y, Shang L, et al. TinyBERT: Distilling BERT for natural language understanding. arXiv preprint arXiv:1909.10351, 2019.

[123] Liu X, He P, Chen W, et al. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. arXiv preprint arXiv:1904.09482, 2019.

[124] Liu X, He P, Chen W, et al. Multi-Task deep neural networks for natural language understanding. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4487-4496.

[125] Yang Z, Shou L, Gong M, et al. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. Proceedings of the 13th International Conference on Web Search and Data Mining. 2020:690-698.

[126] Liu L, Wang H, Lin J, et al. Attentive student meets multi-task teacher: improved knowledge distillation for pretrained models. arXiv preprint arXiv:1911.03588, 2019.

[127] Xu C, Zhou W, Ge T, et al. Bert-of-theseus: Compressing bertby progressive module replacing. arXiv preprint arXiv:2002.02925, 2019.

[128] Cheong R, Daniel R. Transformers. zip: Compressing transformers with pruning and quantization[OL]. [2019-10-28] https://web. stanford. edu/class/cs224n/reports/custom/15763707. pdf

[129] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. 4th International Conference on Learning Representations. 2016.

[130] Hubara I, Courbariaux M, Soudry D, et al. Binarized neural networks. Advances in neural information processing systems. 2016: 4107-4115.

[131] Shen S, Dong Z, Ye J, et al. Q-bert: Hessian based ultra low precision quantization of bert. arXiv preprint arXiv:1909.05840, 2019.

[132] Zafrir O, Boudoukh G, Izsak P, et al. Q8bert: Quantized 8bit bert. arXiv preprint arXiv:1910.06188, 2019.

[133] Voita E, Talbot D, Moiseev F, et al. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5797-5808.

[134] Louizos C, Welling M, Kingma D P. Learning sparse neural networks through L_0 regularization. arXiv preprint

arXiv:1712.01312, 2017.

[135] Michel P, Levy O, Neubig G. Are sixteen heads really better than one?[C]/ Advances in Neural Information Processing Systems. 2019:14014-14024.

[136] Fan A, Grave E, Joulin A. Reducing transformer depth on demand with structured dropout. arXiv preprint arXiv:1909.11556, 2019.

[137] McCarley J S. Pruning a BERT-based question answering model. arXiv preprint arXiv:1910.06360, 2019.

[138] Guo F M, Liu S, Mungall F S, et al. Reweighted proximal pruning for large-scale language representation. arXiv preprint arXiv:1909.12486, 2019.

[139] Candes E J, Wakin M B, Boyd S P. Enhancing sparsity by reweighted ℓ 1 minimization. Journal of Fourier analysis and applications, 2008, 14(5-6): 877-905.

[140] Parikh N, Boyd S. Proximal algorithms. Foundations and Trends® in Optimization, 2014, 1(3): 127-239.

[141] Lan Z, Chen M, Goodman S, et al. ALBERT: A lite BERT for self-supervised learning of language representations. International Conference on Learning Representations. 2019.

[142] Levine Y, Lenz B, Dagan O, et al. SenseBERT: Driving Some Sense into BERT. arXiv preprint arXiv:1908.05646, 2019.

[143] Lauscher A, Vulic I, Ponti E, et al. Specializing Unsupervised Pretraining Models for Word-Level Semantic Similarity. arXiv preprint arXiv:1909.02339, 2019.

[144] Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced language representation with informative entities. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1441-1451.

[145] Liu W, Zhou P, Zhao Z, et al. K-BERT: Enabling language representation with knowledge graph. arXiv preprint arXiv:1909.07606, 2019.

[146] Wang X, Gao T, Zhu Z, et al. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. arXiv preprint arXiv:1911.06136, 2019.

[147] Peters M E, Neumann M, Logan R, et al. Knowledge enhanced contextual word representations. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 43-54.

[148] Sun C, Myers A, Vondrick C, et al. Videobert: A joint model for video and language representation learning. Proceedings of the IEEE International Conference on Computer Vision. 2019: 7464-7473.

[149] Alberti C, Ling J, Collins M, et al. Fusion of detected objects in text for visual question answering. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 2131-2140.

[150] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks. European conference on computer vision. Springer, Cham, 2016: 630-645.

[151] Li G, Duan N, Fang Y, et al. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. arXiv preprint arXiv:1908.06066, 2019.

[152] Su W, Zhu X, Cao Y, et al. Vl-bert: Pre-training of generic visual-linguistic representations. 8th International Conference on Learning Representations.2020.

[153] Chen Y C, Li L, Yu L, et al. Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740, 2019.

[154] Lu J, Batra D, Parikh D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems. 2019: 13-23.

[155] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. 2015: 91-99.

[156] Sun C, Baradel F, Murphy K, et al. Learning video representations using contrastive bidirectional transformer. arXiv preprint

arXiv: 1906.05743,2019.

[157] Xie S, Sun C, Huang J, et al. Rethinking spatiotemporal feature learning for video understanding. arXiv preprint arXiv:1712.04851, 2017, 1(2): 5.

[158] Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.

[159] Conneau A, Lample G. Cross-lingual language model pretraining. Advances in Neural Information Processing Systems. 2019: 7057-7067.

[160] Wu S, Conneau A, Li H, et al. Emerging cross-lingual structure in pretrained language models. arXiv preprint arXiv:1911.01464, 2019.

[161] Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116, 2019.

[162] Eisenschlos J, Ruder S, Czapla P, et al. MultiFiT: Efficient multi-lingual language model fine-tuning. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 5706-5711.

[163] Bradbury J, Merity S, Xiong C, et al. Quasi-recurrent neural networks. 5th International Conference on Learning Representations. 2017.

[164] Kudo T. Subword regularization: Improving neural network translation models with multiple subword candidates. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 66-75.

[165] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[166] Smith L N. A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820, 2018.

[167] Ruder S, Plank B. Strong baselines for neural semi-supervised learning under domain shift. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1044-1054.

[168] Huang H, Liang Y, Duan N, et al. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 2485-2494.

[169] Moradshahi M, Palangi H, Lam M S, et al. HUBERT untangles BERT to improve transfer across NLP tasks. arXiv preprint arXiv:1910.12647, 2019.

[170] Merity S. Single Headed Attention RNN: Stop thinking with your head. arXiv preprint arXiv:1911.11423, 2019.

[171] Kovaleva O, Romanov A, Rogers A, et al. Revealing the dark secrets of BERT. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 4356-4365.

[172] van Aken B, Winter B, Löser A, et al. How does BERT answer questions?: A layer-wise analysis of transformer representations. Proceedings of the 28th ACM International Conference on Information and Knowledge Management. ACM, 2019: 1823-1832.

附中文参考文献:

[12] 陈旭,刘鹏鹤,孙毓忠,沈曦,张磊,王晓青,孙晓平,程伟.面向不均衡医学数据集的疾病预测模型研究.计算机报,2019,42(3):596-609.

[32] 黄珺. 统计和规则相结合的语言模型在中文输入法中的应用研究[D].西安电子科技大学,2008.

[41] 张志昌,张珍文,张治满.基于 IndRNN-Attention 的用户意图分类.计算机研究与发展,2019,56(7):1517-1524.

[43] 周俊佐,朱宗奎,何正球,陈文亮,张民.面向人机对话意图分类的混合神经网络模型.软件学报,2019,30(11):3313−3325.

http://www.jos.org.cn/1000-9825/5862.htm

[46] 李超,柴玉梅,南晓斐,高明磊.基于深度学习的问题分类方法研究.计算机科学,2016,43(12):115−119.

[48] 杜慧,徐学可,伍大勇,刘悦,余智华,程学旗.基于情感词向量的微博情感分类.中文信息学报,2017,31(3):170-176.

[49] 朱苏阳,李寿山,周国栋.基于对抗式神经网络的多维度情绪回归.软件学报,2019,30(7):2091-2108. http.//www.jos.org.cn/1000-9825/5838.htm

[50] 刘宇鹏,马春光,张亚楠.深度递归的层次化机器翻译模型.计算机学报,2017,40(04):861-871.

[51] 梁小波,任飞亮,刘永康,潘凌峰,侯依宁,张熠,李妍.N-Reader:基于双层 Self-attention 的机器阅读理解模型.中文信息学报,2018,32(10):130-137.

[61] 杨中成. 融合预训练语言模型的机器译文质量评估[D].北京交通大学,2019.