

中文预训练模型研究进展

侯钰涛 阿布都克力木·阿布力孜 哈里旦木·阿布都克里木

新疆财经大学信息管理学院 乌鲁木齐 830012

(hyt1159871021@163.com)

摘要 近年来,预训练模型在自然语言处理领域蓬勃发展,旨在对自然语言隐含的知识进行建模和表示,但主流预训练模型大多针对英文领域。中文领域起步相对较晚,鉴于其在自然语言处理过程中的重要性,学术界和工业界都开展了广泛的研究,提出了众多的中文预训练模型。文中对中文预训练模型的相关研究成果进行了较为全面的回顾,首先介绍预训练模型的基本概况及其发展历史,对中文预训练模型主要使用的两种经典模型 Transformer 和 BERT 进行了梳理,然后根据不同模型所属类别提出了中文预训练模型的分类方法,并总结了中文领域的不同评测基准,最后对中文预训练模型未来的发展趋势进行了展望。旨在帮助科研工作者更全面地了解中文预训练模型的发展历程,继而为新模型的提出提供思路。

关键词: 中文预训练模型;自然语言处理;词向量;预处理;深度学习

中图法分类号 TP391

Advances in Chinese Pre-training Models

HOU Yu-tao, ABULIZI Abudukelimu and ABUDUKELIMU Halidanmu

School of Information Management, Xinjiang University of Finance and Economics, Urumqi 830012, China

Abstract In recent years, pre-training models have flourished in the field of natural language processing, aiming at modeling and representing the implicit knowledge of natural language. However, most of the mainstream pre-training models target at the English domain, and the Chinese domain starts relatively late. Given its importance in the natural language processing process, extensive research has been conducted in both academia and industry, and numerous Chinese pre-training models have been proposed. This paper presents a comprehensive review of the research results related to Chinese pre-training models, firstly introducing the basic overview of pre-training models and their development history, then sorting out the two classical models Transformer and BERT that are mainly used in Chinese pre-training models, then proposing a classification method for Chinese pre-training models according to model categories, and summarizes the different evaluation benchmarks in the Chinese domain. Finally, the future development trend of Chinese pre-training models is prospected. It aims to help researchers to gain a more comprehensive understanding of the development of Chinese pre-training models, and then to provide some ideas for the proposal of new models.

Keywords Chinese pre-training models, Natural language processing, Word embedding, Pre-training, Deep learning

1 引言

自然语言处理(Natural Language Processing, NLP)是计算机利用人类定义的算法对自然语言形式的输入进行加工处理的过程,旨在让计算机可以像人类一样理解和生成语言,具备如人类一样的听、说、读、写、问、答、对话、聊天等的能力,并利用已有知识和常识进行推理分析。自然语言处理技术的发展经历了从基于规则到基于统计的过程。随着深度学习的发展,图像、文本、声音、视频等不同形式的信息载体被自然语言处理技术突破,大量的神经网络被引入自然语言理解任务中,如循环神经网络^[1](Recurrent Neural Networks, RNN)、卷积

神经网络^[2](Convolutional Neural Networks, CNN)、注意力机制^[3](Attention Mechanism)等。在特定的自然语言处理任务中,神经网络可以隐性地学习到序列的语义表示与内在特征,因此,神经网络成为了解决复杂自然语言处理任务最有效的方法。随着计算力的不断增强,深度学习在自然语言处理领域中不断发展,分布式表示占据了主导地位,不仅在指定任务中可以端到端地学习语义表示,而且可以在大规模无标注的文本上进行自主学习,能更灵活地运用在各种下游任务中。然而,早期在有监督数据上训练浅层模型往往存在过拟合和标注数据不足等问题,在训练深层模型参数时,为了防止过拟合,通常需要大量的标注数据,但有监督的标注数据成本

到稿日期:2021-12-02 返修日期:2022-04-17

基金项目:国家自然科学基金(61866035,61966033)

This work was supported by the National Natural Science Foundation of China(61866035,61966033).

通信作者:哈里旦木·阿布都克里木(abdklmhldm@gmail.com)

较高,因此模型主要利用网络中现存的大量无监督数据进行训练。在此背景下,预训练技术被广泛地应用在自然语言处理领域。其中,最经典的预训练模型是 BERT^[4]模型,在多个自然语言处理任务中取得了最好结果(State of the Art, SOTA)。此后出现了一系列基于 BERT 的预训练模型,掀起了深度学习与预训练技术的发展浪潮。

随着国内外研究者在预训练模型方面的深入研究,目前已有很多关于预训练模型的综述,但缺少专门针对中文领域的相关综述。当前,中文预训练模型蓬勃发展并取得一定的成绩,因此,对现有研究成果进行全面的分析和总结非常必要。本文期望能为中文预训练相关领域的学者提供参考,帮助科研工作者了解目前的研究现状和未来的发展趋势。本文第 2 节概述预训练模型的基本情况;第 3 节主要介绍两种基本模型,即 Transformer 和 BERT;第 4 节根据不同模型的所属类别提出典型的中文预训练模型的分类方法,并汇总了中文预训练模型的相关资源;第 5 节梳理了中文领域的不同评测基准;最后总结全文并展望未来。

2 预训练模型

2.1 预训练模型发展史

从预训练语言模型的发展时间来看,可以将其分为静态预训练模型和动态预训练模型。2013 年,Mikolov 等^[5]在神经网络语言模型(Neural Network Language Model, NNLM)思想的基础上提出 Word2Vec,并引入大规模预训练的思路,旨在训练具有特征表示的词向量,其中包括 CBOW 和 Skip-Gram 两种训练方式。相比 NNLM 模型,Word2Vec 可以更全面地捕捉上下文信息,弥补 NNLM 模型只能看到上文信息的不足,提高模型的预测准确性,Word2Vec 极大地促进了深度学习在 NLP 中的发展。自 Word2Vec 模型被提出以来,一批训练词向量的模型相继涌现,例如,Glove^[6] 和 FastText^[7] 等模型均考虑如何得到文本单词较好的词向量表示,虽然对下游任务性能有所提升,但其本质上仍是一种静态的预训练模型。

2018 年,Peters 等^[8]提出的 ELMo 模型将语言模型带入动态的预训练时代。ELMo 模型采用双层双向的 LSTM^[9] 编码器进行预训练,提取上下文信息,并将各层词嵌入输入特定下游任务中进行微调。该模型不仅可以学习到底层单词的基础特征,而且可以学到高层的句法和语义信息。然而,ELMo 模型只能进行串行计算,无法并行计算,模型训练的效率较低;此外,该模型无法对长序列文本进行建模,常出现梯度消失等问题。而后,OpenAI 提出了 GPT(Generative Pre-training)^[10] 模型。与 ELMo 模型不同,GPT 采用 Transformer 深度神经网络,其处理长文本建模的能力强于 LSTM,仅使用 Transformer 解码器进行特征提取,在机器翻译等生成式任务上表现惊人,但这一特点也导致 GPT 只利用到了当前词前面的文本信息,并没有考虑到后文信息,其本质上依旧是一种单向语言模型。为了解决 GPT 等模型单向建模的问题,2018 年,Devlin 等^[4]提出了 BERT 模型,该模型是第一个基于 Trans-

former 的双向自监督学习的预训练模型,在英文语言理解评测基准^[11]榜单中的多个任务上达到了 SOTA 结果,此后出现了一大批基于 BERT 的预训练模型,大幅提升了下游自然语言处理任务的性能。中文预训练模型虽然起步较晚,但发展迅速,已经取得了一定成果,本文第 4 节将对其进行重点介绍。

2.2 研究中文预训练模型的原因

首先,中文和英文分别是世界上使用人数最多和范围最广的两种语言,然而在自然语言处理领域,英文预训练模型较为普遍,例如,以 BERT 为首及其后出现的大量预训练模型均是在单一语料英文数据集上进行训练,此外模型的设计理念也更适用于英文,比如分词方式及掩码方式等。其次,中文和英文语言本质上存在差异,它们的主要区别是,中文文本通常由多个连续的字符组成,词与词之间没有明显的分隔符。如果使用英文预训练模型去处理常见的中文任务,效果往往不佳。因此,为了推动中文领域自然语言处理技术和预训练模型在多语言任务方面的发展,构建以中文为核心的预训练模型势在必行。

3 Transformer 和 BERT

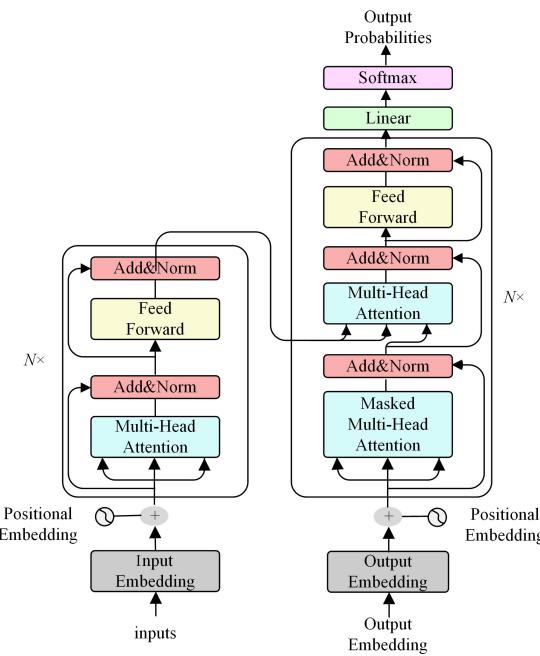
自 2021 年以来,中文预训练模型进入井喷式的发展阶段,其架构主要基于 Transformer 和 BERT 两种基础模型,本节主要介绍这两种模型。

3.1 Transformer

循环神经网络,特别是长短时记忆神经网络(Long Short-Term Memory,LSTM^[12])和门控递归神经网络(Gated Recurrent Unit,GRU^[13])可以很好地解决序列建模问题,利用循环细胞处理时间序列数据以及顺序数据,但循环神经网络的训练过程是迭代,只能进行串行计算,这种方式不仅效率较低,而且不能很好地捕捉双向语义之间的依赖问题。

为了解决以上神经网络存在的问题,Vaswani 等^[14]提出了一种新的序列到序列的神经网络模型 Transformer,该模型利用自注意力机制来提取长序列之间的语义信息,与传统神经网络的序列建模相比,Transformer 中的多头自注意力机制适合并行计算序列中每一个字符与其他字符的相关程度,将序列中不同位置的字符联系起来,从而对长序列的上下文语义信息进行建模。比起序列对齐的循环神经网络,使用 Transformer 的神经网络在神经语言建模中表现更佳,其应用的方式主要是先预训练语言模型,然后在下游任务上进行微调,以完成不同类型的任务。目前大部分预训练模型都在 Transformer 的基础上进行序列特征提取,大量实验表明 Transformer 模型是语言建模中最有效的深度神经网络结构,且这些预训练模型可以大幅改善各种自然语言处理任务的性能。

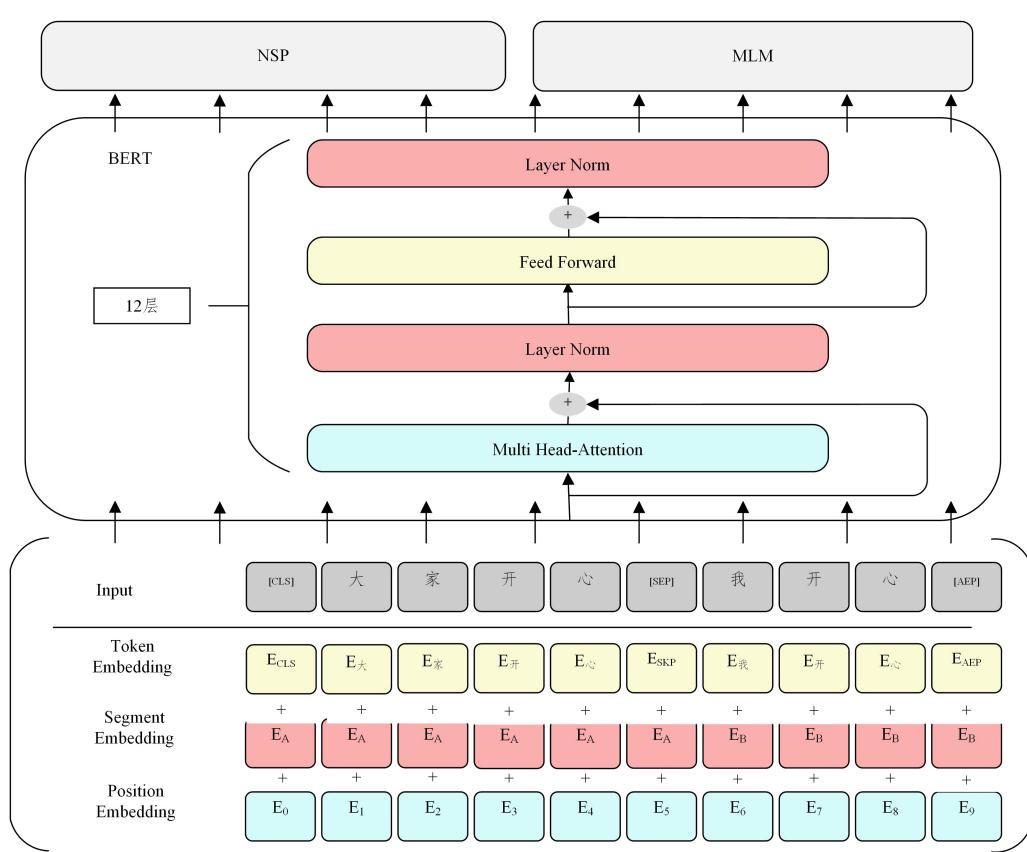
图 1 为典型的 Transformer 架构,该架构由 6 个结构相同的编码器和解码器堆叠而成。单个编码器由堆叠的自注意力层和前馈神经网络组成,解码器由堆叠的自注意力层、掩码注意力层和前馈神经网络组成。有关 Transformer 的详细细节介绍请参考文献[14]。

图 1 Transformer 示意图^[14]Fig. 1 Illustration of Transformer^[14]

3.2 BERT

BERT^[4] (Bidirectional Encoder Representations from Transformers)是由谷歌提出的一种面向自然语言处理任务的无监督预训练语言模型,由Transformer的双向编码器表示。BERT的架构如图2所示。

与传统基于自回归的语言建模方式不同,BERT引入了基于自编码(Auto-Encoding)的预训练任务,主要包括掩码语言模型(Masked Language Model, MLM)和下一个句子预测(Next Sentence Prediction, NSP)。MLM任务与完形填空类似,在具体的预训练中,用[mask]标记随机掩盖15%的token,虽然是双向语言模型,但[mask]标记在微调时并未出现,导致预训练和微调阶段出现不匹配的情况。为了解决此问题,训练数据生成器随机选择15%的token进行掩盖并预测,被选中的token中80%被[mask]标记替换,10%被随机词替换,10%不做改变。NSP预训练任务旨在判断两个句子是否来自同一个文本的连续句子,其本质是一个二分类任务,其中两个句子为连续句子的概率为50%(标记为IsNext),两个句子为非连续的概率也为50%(标记为NotNext)。但在后续研究过程中,NSP任务在提升模型性能方面并没有太大作用。有关BERT的详细细节请参考文献[4]。

图 2 BERT 示意图^[4]Fig. 2 Illustration of BERT^[4]

4 中文预训练模型分类

在自然语言处理领域,继Transformer和BERT出现之后,涌现出大量的预训练模型,这些模型主要针对英文领域,中文领域的研究起步较晚。但在近两年,中文预训练模型受

到广大学者的关注并取得了一定的研究成果。为了阐明现有的中文预训练模型,本节主要从以下6个方面对现有的预训练模型进行分类,图3展示了典型的中文预训练模型的分类图。

(1) 预训练模型的方法改进,主要包括掩码方式的转变、位置编码的转变、LN层的位置变化、MoE层的使用、多粒度

训练和其他改进。

(2)融入外部信息的预训练,主要包括命名实体、知识图谱、语言学知识和特定知识。

(3)关于多模态融合的预训练模型。

(4)侧重于高效计算的预训练,主要包括数据处理阶段、

预训练阶段以及技术优化。

(5)指特定领域的预训练,主要包括对话系统和其他领域的预训练模型。

(6)介绍一些其他变体,主要侧重于典型的英文预训练模型开源的中文版本。

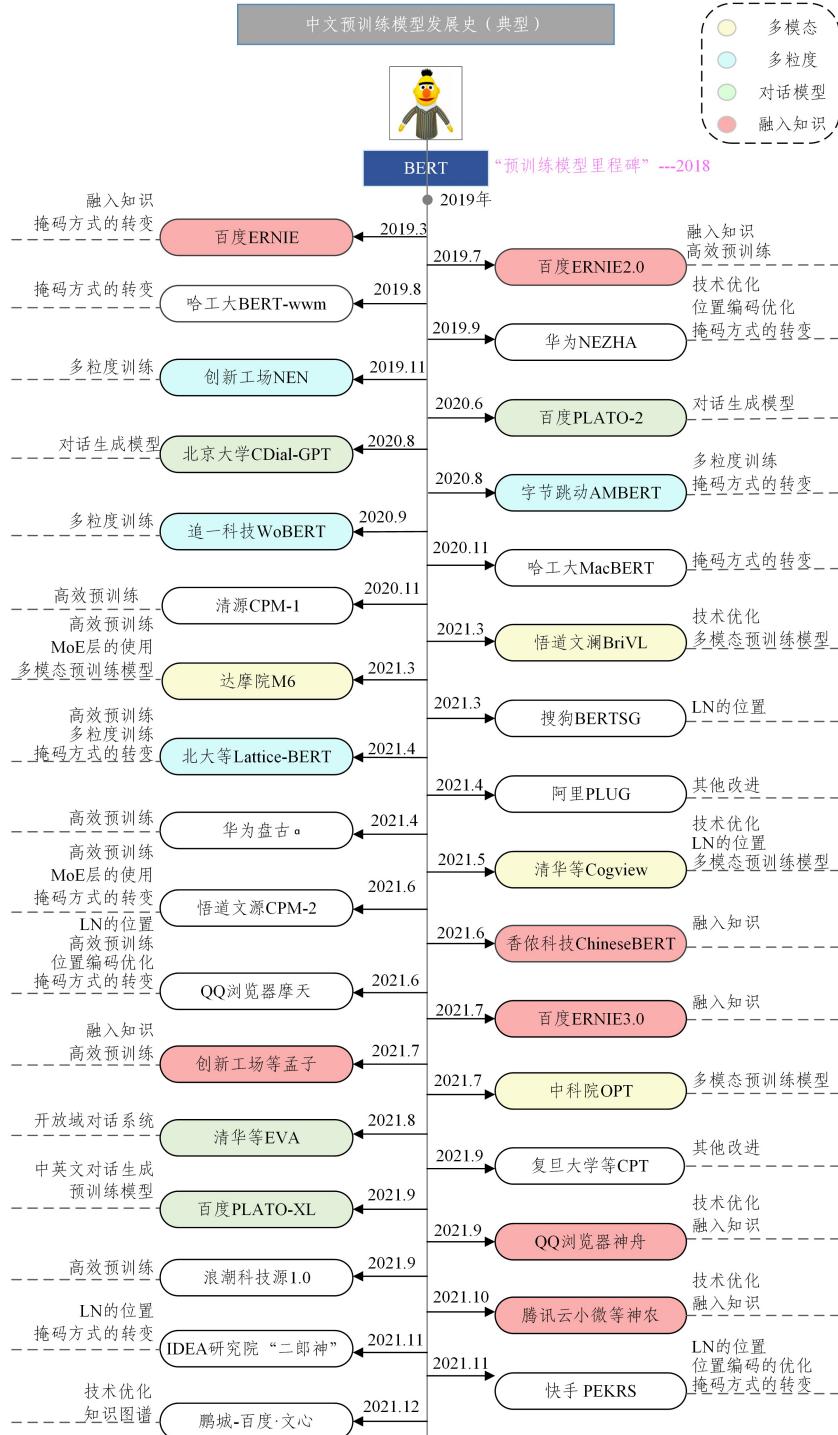


图3 中文预训练模型分类图

Fig. 3 Chinese pre-trained model classification chart

4.1 预训练模型的方法改进

4.1.1 掩码方式的转变

由于中文和英文语言本质上存在差异,因此中文预训练模型需要对掩码方式进行修改。BERT模型旨在对文本进行最小单元的切分,利用Wordpiece^[15]将单词划分为几个小片

段。在MLM任务中,掩码对象大多是不完整的单词,这种掩码方式可能会导致信息泄露,并且此现象在中文中更为明显。例如:“我喜欢打篮球[mask]。”模型很容易预测到掩码部分的词为“球”,原因是“篮”字可以给模型一定的提示。因此,谷歌官方进一步提出了进阶版的预训练任务:全词掩码(Whole

Word Masking, WWM)。该任务仅对掩码方式进行了修改,将最小的掩码单元由 WordPiece 子词转换为全词,即当一个字符被覆盖时,同属于该词语的其他字符都被覆盖,可以有效缓解 WordPiece 子词泄露的问题。如:“我喜欢打 [mask] [mask]。”这句话,预训练模型会将篮球这一个词语全部掩码,而不只是掩码一个字符,当模型在预测时可能会预测到“篮球”“排球”“游戏”等多种结果,使预测变得更富有挑战性。

在全词掩码预训练任务被提出之前,百度研究团队提出了基于知识增强的预训练模型 ERNIE^[16],该模型在 BERT 掩码方式的基础上进行修改,不仅采用了字级别的掩码方式,同时还对实体、短语级别都进行了掩码。将实体、短语作为掩码单元,一个单元通常由多个字符组成,掩码时同属一个单元的所有字符都被掩码,而不只是一个字符,其目的是通过这种方式来隐性地学习到知识和较长的语义依赖信息,这种掩码方式与全词掩码类似。

全词掩码预训练任务被提出之后,哈工大讯飞实验室将谷歌官方发布的 BERT-base 中文版本与全词掩码技术相结合,提出了 BERT-wwm,在中文数据集上进行训练,并将其与原始 BERT 和 ERNIE 以及他们提出的一些扩展模型(BERT-wwm-ext, RoBERTa-wwm-ext, RoBERTa-wwm-ext-large)进行实验对比,证实了全词掩码技术的有效性。之后,华为研究团队提出的 NEZHA^[17]模型在训练时采用 WWM 策略进行掩码,但与原策略不同的是,被掩码的中文字符约占序列总长度的 12%,随机替换的字符占 1.5%。

BERT 官方除了全词掩码任务外,还提出了 N-gram 掩码任务,其训练方式是将连续的 N-gram 文本进行掩码,让模型去预测被掩码的内容。相比全词掩码,这一任务需要进一步识别出短语级的文本边界信息。

以上两种掩码策略(WWM 和 N-gram)可以结合使用,例如,2020 年,哈工大讯飞实验室又提出了一种基于文本纠错的掩码语言模型 MacBERT^[18](MLM ascertainment, Mac),该模型采用 WWM 策略与 N-gram 掩码方式相结合的方式,其中词级别的 unigram 至 4-gram 的掩码比例分别是 40%, 30%, 20%, 10%。在预训练阶段,为了解决预训练和微调之间由于掩码方式造成的一致现象,他们使用相似词来替换 [mask] 标记,其中相似词从中文近义词工具包(Synonyms Toolkit)中获取。该掩码策略与原始 BERT 有些许不同,但总体都是对输入总序列的 15% 进行掩码,其中 80% 会被替换为相似词,10% 会被替换为随机词,10% 不做改变。

MLM 任务中的随机掩码方式是一种静态掩码,即每个样本只能进行一次掩码。静态掩码的方式降低了训练数据的多样性,数据的重复使用率有待提高。基于此,RoBERTa^[19]引入了动态掩码(Dynamic Masking)策略,该策略指掩码方法和位置会随着训练的进行实时变化,以此来保证同一个样本在不同的轮数下可以生成不同的掩码形式。当数据量较大时,动态掩码策略能有效提高数据的使用率。

近期,IDEA 研究院提出了一种轻量化模型“二郎神”^[20],该模型采用基于语言知识的掩码和动态掩码相结合的策略对中文进行分词,使模型可以学到基础的语言学知识。快手公司提出具有快手特色的预训练语言模型 PERKS^[21],该

模型采用多阶段多任务学习的方式使模型学到不同领域和不同粒度的信息。在第一阶段的训练过程中,其与“二郎神”的掩码方式类似,都采用动态全词掩码策略来学习基础的语言学知识;在第二阶段,将内外部语料库进行混合,利用关键实体识别技术对知识信息进行掩码,再引入句子级预测任务,以刻画句子级知识,并使用课程学习的方式逐步调整不同任务配比,让模型逐渐学习到更有难度的知识。经过前两个阶段的训练,模型就能获得较为全面的语言学知识。

此外,部分中文预训练模型还采用了其他的掩码策略。例如,北京大学等提出的一种中文多粒度预训练模型 Lattice-BERT^[22]采用整段掩码预测任务(Masked Segment Prediction, MSP)进行训练,原因是该模型在输入时是将单词和字符以词格图形式与文本一同输入到模型中,利用多粒度信息对句子建模,但这种方式会产生冗余信息。如果采用 MLM 任务中的随机掩码,模型就会从重叠的文本单元中预测到被掩码的词语,因而 Lai 等^[22]提出了 MSP 任务。该任务将一句话切分为多个段,前提是确保每段之间的字符都不重复,且在切分时保持段的长度最短,掩码时将整段信息都掩码,之后再预测掩码词,通过这种方式可以有效缓解信息泄露的问题。

“悟道·文源”团队提出的 CPM-2^[23]采用的是 MLM 任务的一种变体,该模型通过使用几种不同的特殊标记随机替换几个跨度(span)的字符来构造编码器的输入,然后利用解码器依次预测被替换的跨度词。例如,原句为“我要在新的一年里继续努力学习”,构建编码器输入的句子为“我要在[X]的一年里[Y]学习”,解码器输出为 “[X]新[Y]努力[Z]”,其中,[X],[Y],[Z]代表特殊标记,[X]和[Y]代表不同的跨度值,[Z]代表输出的结束标记。需要注意的是,此处特殊标记占总序列长度的 15%,被替换的跨度平均长度为 10。

QQ 浏览器搜索语义团队提出了中文预训练模型 Motian^[24],将 WWM 策略扩展为短语加词级别的掩码方案,该方案促使模型能从更长的上下文信息中预测被掩码的字符。为了减少预训练和微调之间不一致的现象,在掩码比例上,该团队成员对比了全随机词替换的方式,以一定概率分布的随机词替换和近义词替换等方式,进一步证实了使用一定概率分布的近义词替换效果更好。

百度提出了基于语音语义的文本纠错模型 MLM-phonetics^[25],此模型在使用特殊标记 [mask] 进行掩码的基础上,利用字音混淆词和混淆字符的拼音进行掩码。掩码比例占总序列长度的 20%,这 3 种掩码策略分别占比 40%, 30%, 30%, 有效缓解了预训练和下游任务输入字符上的差异,而且使模型融入了语音特征。表 1 列出了对以上几种掩码方式的总结。

表 1 掩码方式的转变

Table 1 Shift in masking approach

原始句子输入	学生的科研生活很充实
中文分词	学生的科研生活很充实
随机掩码输入	学 [mask] 的科研 [mask] 活很 [mask] 充实
WWM 输入	学生的 [mask] [mask] 生活 很 充 实
N-gram 掩码输入	学生的 [mask] [mask] [mask] [mask] 很 充 实
Mac 掩码输入	学生的 科研 状态 很 饱 满
MLM-phonetics 输入	学生的 [mask] 科研 生活 很 充 实

4.1.2 位置编码的转变

位置编码是对序列中词的位置进行编码,在一个句子中字符位置以及排列顺序极其重要,它是表达句子语义不可或缺的一部分。例如:“我借钱给你”和“你借钱给我”,这两句话包含相同的字符,然而字符位置和排列顺序的不同会导致语义的偏差。Transformer 没有采用本身具有顺序结构的循环神经网络,而是使用自注意力机制,先将输入的文本信息转换成对应的向量表示 $x = \{x_1, x_2, x_3, \dots, x_n\}$,同时融入每个词的位置信息,之后将其输入到 Transformer 中,通过自注意力机制来编码其上下文信息,进而学习词与词之间的依赖关系,最终得到每一向量的对应输出 $z = \{z_1, z_2, z_3, \dots, z_n\}$ 。具体计算方式为:首先将输入的词向量 x_i 通过 3 个不同的参数矩阵 $\mathbf{W}^q, \mathbf{W}^k, \mathbf{W}^v$ 映射为 3 个不同的新向量,即查询(Query)向量、值(Value)向量、键(Key)向量,分别将其定义为 q_i, k_i, v_i ,即第 i 个位置的词的 q, k, v 向量,其维度均为 512。查询向量是序列中某个位置的词给序列中其他位置的词进行匹配的结果;键向量是某个位置的词被序列中其他的词匹配的结果;值向量是某个位置的词被抽取出的词特征,输出是值的加权和。在具体实现时,以矩阵乘法的形式进行并行计算,使用 GPU 可加快模型的训练速度。输出 z_i 计算式如下:

$$q_i = \mathbf{W}^q x_i, k_i = \mathbf{W}^k x_i, v_i = \mathbf{W}^v x_i$$

$$e_{ij} = \frac{(q_i)(k_j)^T}{\sqrt{d_z}}$$

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^n \exp e_{ik}}$$

$$z_i = \sum_j^n \alpha_{ij} v_j$$

其中, e_{ij} 为输入元素的线性变换之间的缩放点积, α_{ij} 为使用 softmax 函数计算的位置 i 和 j 的隐藏状态之间的注意力分数。

在输入位置信息时,Transformer 模型采用绝对位置编码(Absolute Position Embedding, APE),即每一个位置都有一个固定的位置向量,计算式^[14]如下:

$$PE_{(pos, 2i)} = \sin(pos / 10000^{2i/d_{model}})$$

$$PE_{(pos, 2i+1)} = \cos(pos / 10000^{2i/d_{model}})$$

其中, pos 表示 $token$ 在句子序列中的实际位置, i 表示词向量的第 i 个维度, d_{model} 是向量的维度,向量偶数维采用正弦公式,奇数维采用余弦公式。

2018 年,Shaw 等^[26]对 Transformer 中的绝对位置编码做了进一步修改,提出了一种参数式相对位置编码(Relative Position Embedding, RPE),在此方案中,在计算注意力得分和权重值时各加入一个可训练的表示相对位置的参数。具体计算式^[26]如下:

$$z_i = \sum_{j=1}^n \alpha_{ij} (v_j + \alpha_{ij}^r)$$

$$e_{ij} = \frac{(q_i)(k_j + \alpha_{ij}^k)^T}{\sqrt{d_z}}$$

其中, α_{ij}^r 和 α_{ij}^k 为位置 i 和 j 的相对位置编码。此后,华为研究团队提出的 NEZHA^[17]在上述提出的相对位置编码基础上,

使用相对位置的正弦函数计算注意力得分和输出,其中, α_{ij}^y 和 α_{ij}^k 都是正弦函数,且在训练时保持不变。将 α_{ij}^y 和 α_{ij}^k 简写为 α_{ij} , α_{ij} 的计算式^[17]如下:

$$\alpha_{ij}[2k] = \sin((j-i)/(10000^{\frac{2+k}{d_h}}))$$

$$\alpha_{ij}[2k+1] = \cos((j-i)/(10000^{\frac{2+k}{d_h}}))$$

其中, d_h 指模型中每头的隐藏层大小, α_{ij} 指位置 i 和位置 j 的相对位置编码, k 是指词向量的第 k 个维度,此处分别表示第 $2k$ 和第 $2k+1$ 的维度。PERKS^[21]模型采用相对位置编码和绝对位置编码相结合的方式,显著加快了模型的训练速度。

4.1.3 LN 的位置变化

随着神经网络层数不断加深,训练的时间成本也不断增大,且在训练中会出现梯度爆炸和梯度消失等问题,种种原因导致模型训练变得非常困难。早先提出的批归一化^[27](Batch Normalization, BN)可以加速神经网络的训练,但需要设置恰当的 batchsize 才能有效地估计均值和方差,如果遇到比训练样本大的测试样本,批归一化将无法运行。为了克服批归一化的缺点,Ba 等提出了层归一化^[28](Layer Normalization, LN)技术,该方法对网络中的神经元输入进行缩放和中心化处理,以稳定深度神经网络的训练。具体计算为,设第 L 层的神经净输入为 $z(L)$,其均值 $\mu^{(L)}$ 和方差 $\sigma^{(L)2}$ 为:

$$\mu^{(L)} = \frac{1}{m^L} \sum_{m=1}^{m^L} z_m^{(L)}$$

$$\sigma^{(L)2} = \frac{1}{m^L} \sum_{m=1}^{m^L} (z_m^{(L)} - \mu^{(L)})^2$$

其中, m^L 指第 L 层神经元的数量,层归一化的定义为:

$$LN(z^L) = \frac{z(L) - \mu^{(L)}}{\sqrt{\sigma^{(L)2}} + \epsilon} \times \gamma + \beta$$

其中, γ 和 β 表示可学习的缩放和平移的参数向量。

在原始 Transformer 中,LN 层位于残差连接之后,称为 Post-LN。使用该架构进行训练时,由于最终的模型性能对于超参数的设置非常敏感,因此在训练过程中需要仔细调参才能获得最佳的模型性能,如 warm-up 学习率策略,但这种方法非常耗时。后续有研究者提出将 LN 层放在残差连接过程中,称为 Pre-LN。在该架构下,可去除 warm-up 学习率阶段,从而减少超参数的数量并缩短模型训练时间。

当前较为流行的预训练模型都采用 Pre-LN 策略。例如,搜狗团队在 BERT 基础上提出了 BERTSG^[29],在模型结构方面摒弃了 Post-LN 架构,该团队成员认为在训练超大模型时,学习率优化不当会导致收敛效果变差,因此 BERTSG 采用 Pre-LN,大大提升了训练效率。近期,QQ 浏览器团队提出的 Motian,IDEA 研究院提出的“二郎神”以及快手研究团队提出的 PERKS,这 3 种轻量化模型都使用了 Pre-LN 技术来加速模型的收敛,其参数量虽然仅有十亿左右,但在多项中文任务上取得了最优结果。

清华大学等联合发布了中文多模态预训练模型 Cogview^[30],该模型在 Pre-LN 基础上继续改进,Ding 等认为文本和图像在预训练时只依靠 Pre-LN 可能导致深层值的爆炸,因此提出了 Sandwich-LN。该模型在残差分支处的全连接层

与注意力机制前后各添加一个层归一化,对中间特征值进行正则化,研究证实了 Sandwich-LN 可使训练过程更加稳定。3 种不同 LN 层的位置如图 4 所示。

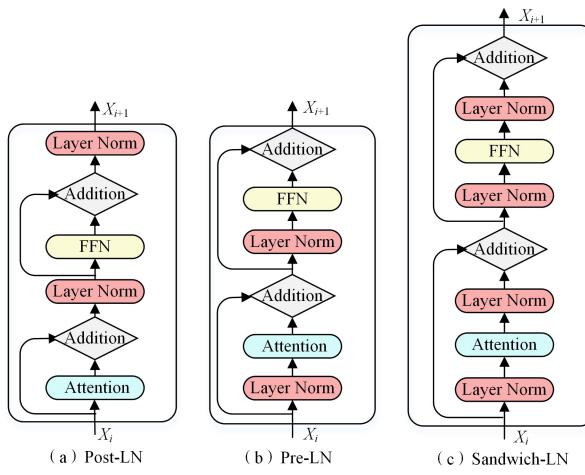


图 4 LN 的位置^[30]

Fig. 4 Location of LN^[30]

4.1.4 MoE 层的使用

在训练数据足够的情况下,增加模型的参数量可以获得更好的预测结果,这一观点已在多个领域的研究工作中得到证实。因此,研究者们致力于开发超大规模的预训练模型。模型规模不断增大,训练数据不断增加,导致训练成本也成倍增加,且受限于算力资源,大规模预训练模型的训练与推理面临巨大的挑战。为了解决这一问题,谷歌提出了一种通用神经网络的组件:稀疏门控混合专家层(Sparsely-Gated Mixture-of-Experts Layer, MoE^[31]),该组件通过在网络中引入多个专家来减少需要激活的神经元数量,以此提升模型的计算效率。该网络由 n 个专家(Expert)和一个决定激活哪位专家工作的门网络(Gating Network)组成。其中,每位专家网络均由一个简单的前馈神经网络构成,它们架构相同,但参数彼此独立;门网络由一个全连接层和一个 softmax 层组成。MoE 层^[31]如图 5 所示。

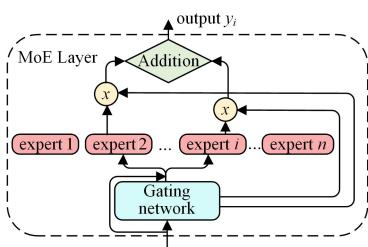


图 5 MoE 示意图^[31]

Fig. 5 Illustration of MoE^[31]

图 5 中门控网络激活第 2 个专家网络和第 i 个专家网络,在给定输入 x 的情况下,MoE 层的输出 y 的计算式如下:

$$y = \sum_{i=1}^n g(x)_i e(x)_i$$

其中, $g(x)_i$ 指门控网络的输出, $e(x)_i$ 指第 i 个专家网络的输出。

MoE 层是扩充模型容量的一种途径,但随着专家数量的

增多,模型参数量也会增大,这时,存储需求可能会超过 GPU 容量,会给模型训练和推理带来困难;其次,专家在工作时,会出现几位专家工作而其他专家围观的现象,没有做到合理分配。为了使训练好的 MoE 层更容易应用到下游任务中,“悟道·文源”团队提出的 CPM-2^[23]在模型推理阶段借助一种深度学习推理框架,通过参数卸载(Offload)和动态调度等策略,同时利用内存和显存来保存模型参数,解决了存储需求过大和专家工作负载不均衡等问题。在该推理框架下,使用单个 GPU 就能高效推理出具有数千万参数的 MoE。

阿里达摩院提出了万亿级别多模态模型 M6^[32],该团队成员对 MoE 层进行分析后认为影响模型质量最重要的问题不是负载不均衡,而是激活的专家个数和专家容量。专家容量定义^[33]为:

$$c = \frac{k \cdot t}{n} \cdot \alpha$$

其中, t 指一个批次(batch)里面 token 的数量, n 指专家的总数量, t/n 指正常情况下每位专家平均的 token 数量, k 指 top- k routing, α 指容量系数,一般选取 1 或 1.5,以此来限定专家能接受的最高 token 范围。以上参数会影响计算量和最后模型的效果,显然激活的专家个数越多,效果越好,但使用传统 top- k routing 方式训练的模型效率大幅下降,为此, Yang 等^[33]提出了 Expert Prototyping 模型,将专家分为 k 个不同的组,每个组都采用 top-1 机制,最终为 k top-1 稀疏激活,图 6 给出了 2 top-1 的 Expert Prototyping 的示意图。M6 模型在阿里自研的 Whale 框架中实现了 MoE,且支持专家并行,该方法有利于提高训练速度。

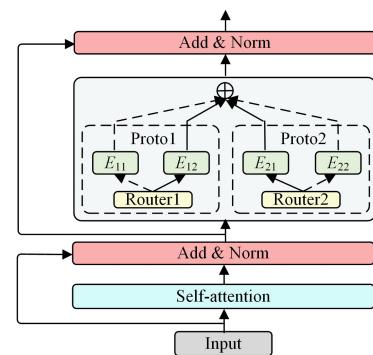


图 6 2 Top-1 Routing 专家原型图^[33]

Fig. 6 2 Top-1 Routing for expert prototyping^[33]

4.1.5 多粒度训练

原始 BERT 模型采用细粒度进行分词,但由于中英文语法的差异,该分词方式会丢失词的部分信息,通过调整输入序列的词颗粒度可以达到优化模型的效果。表 2 列出了单粒度和多粒度的输入对比。

表 2 单粒度和多粒度的输入对比

Table 2 Comparison of single- and multi-granularity inputs

原始句子	他的科研生活很充实
单粒度输入	他的科研生活很充实
多粒度输入	他的科研生活很充实

基于这一思想,创新工场提出了一种中文文本编码器

ZEN(a BERT-based Chinese (Z) text encoder Enhanced by N-gram representations)^[34],该模型引入了经典的N-gram算法来增强中文信息的对齐,在训练集和测试集显性地加入了大粒度的文本知识,有利于模型的语义建模。目前,ZEN模型已经更新到ZEN2.0^[35]版本,此模型更多关注跨领域和跨语言方面的研究。

字节跳动实验室提出了一种多粒度的BERT模型,被称为AMBERT^[36]。该模型使用两种编码器分别处理细粒度和粗粒度的token序列,每个编码器结构与BERT相同,除嵌入层外,其他层参数共享。该模型在中英文数据集做了大量实验,结果表明,在中、英两语中,AMBERT模型的性能明显优于单粒度的BERT模型,且在中文中的表现更佳。此后,腾讯看点等机构的研究者提出了一种采用多粒度输入信息的预训练方法LICHEE^[37]。文献[37]主要将此模型与AMBERT模型进行对比,Guo等认为,与BERT相比,AMBERT虽然性能有所提升且没有降低推理速度,但其推理成本增加了一倍左右,原因在于AMBERT模型采用两个具有共享参数的编码器分别对细粒度和粗粒度token编码。而LICHEE仅在嵌入层融合对输入文本的多粒度信息,并没有改变预训练模型的结构,因此不会带来额外的推理开销。该方法可以应用到各种预训练语言模型中,且经过大量实验证明,该方法能显著提升语言模型的表示能力,并普遍适用于下游自然语言理解任务。

北大和阿里利用中文的多粒度表示提出了预训练语言模型Lattice-BERT^[22],该模型将单词和字符以词格图形式与文本一同输入到模型中,利用多粒度信息对句子建模,其中词格图(Lattice Graph)是一个有向无环图,包含句子中字和词的所有信息,使模型能在预训练阶段学到字和词的全部信息,提高训练效率,词格图如图7所示。

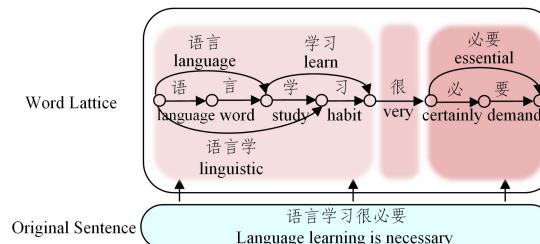


图7 词格图^[22]

Fig. 7 Schematic diagram of word patterns^[22]

追一科技公司提出的关于词颗粒度方面的中文预训练模型WoBERT^[38]对BERT的分词方式进行了修改,在Tokenizer中加入“前分词(Pre-tokenize)”操作,更好地将中文词从文本中划分出来,并保存在字典中。具体操作如下:首先将中文词语加入到字典vocab.txt中;然后使用Pre-tokenize对输入文本进行操作,得到 $\{t_1, t_2, t_3, \dots, t_n\}$,之后遍历每一个 t_i ,如果 t_i 在词表中,则保存下来,否则将 t_i 使用BERT本身的 tokenize函数再次进行操作;最后将每个 t_i 的 tokenize结果按顺序拼接起来,作为最终结果。实验结果表明,该模型在保证其性能的情况下,速度提升明显。

4.1.6 其他改进

阿里达摩院提出了一种具备理解和生成能力的中文预训练模型PLUG^[39],该模型采用编码器和解码器双向建模的方式进行训练,编码器和解码器部分由先前达摩院自研的自然语言理解模型StructBERT^[40]和自然语言生成模型PALM^[41](Pre-training an Autoencoding&autoregressive Language Model)构成。其中StructBERT构建词级别和句子级别的语言信息,以此加强模型对语义、语法的学习能力;PALM模型结合了自回归和自编码两种预训练方法,引入MLM任务来增强编码器的表征能力,以此提升解码器的生成能力。

复旦大学和之江实验室联合提出了中文预训练模型CPT(Chinese Pre-trained Unbalanced Transformer)^[42]。为了使模型兼顾自然语言理解和生成能力,其结构由一个深层次的共享编码器和两个浅层次的解码器组成。其中,共享编码器类似于Transformer编码器,旨在捕获表达语言理解和生成的通用语义表征,一个浅层次解码器类似于Transformer的编码器,另一个类似于Transformer解码器,分别用于理解和生成。这种不平衡的Transformer架构能节省计算和存储成本,加快文本生成的推理速度。CPT的结构如图8所示。

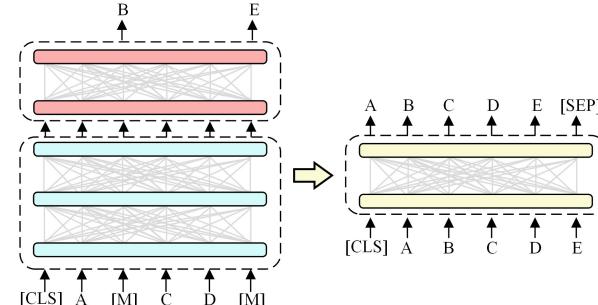


图8 CPT结构图^[42]

Fig. 8 Diagram of CPT structure^[42]

4.2 融入知识的预训练

此前提出的预训练模型(BERT, RoBERTa等)大多通过无监督训练任务(如MLM)从大规模的语料库中学习语义表示。尽管这些模型取得了巨大的成功,但缺少相关的先验知识,很难学习到更丰富的语义信息,在预测时会出现误差。例如,《红楼梦》是中国古代作家曹雪芹所撰写的一部长篇小说。将“红楼梦”或者“曹雪芹”进行掩码,在模型不知道“红楼梦”是一本书、“曹雪芹”是该书作者时,很难通过上下文信息预测到掩码词。作品与作者之间的实体关系即可称为先验知识,融合先验知识对于中文而言尤为重要。目前在中文领域,已有部分模型可以通过隐性或者显性的方式学习到更丰富的语义信息。本文主要从以下几个方面进行介绍。

4.2.1 命名实体

命名实体一般指具有特殊意义的实体,是了解文本真实语义的重要信息,例如,“乌鲁木齐是新疆的省会”,如果将“乌鲁木齐”替换为“太原”,虽然该句子在语法层面是成立的,但在事实层面却是错误的。对于掩码输入“[mask]”是新疆的

省会”这句话而言,正确的预测结果是乌鲁木齐市,而不是其他实体名称,通过训练,模型可以从预训练文本中学到“乌鲁木齐”的实体知识,以及“乌鲁木齐”和“新疆”两个实体之间的语义关系。为了能使模型的预测更准确,可以从外部引入相应的实体信息进行训练。百度研究人员提出了基于知识增强的ERNIE^[46]模型,该模型引入了大量的论坛对话类数据,对语料库中的词、实体、实体关系进行语义建模,可以隐式地学习到相关知识和具有较长语义依赖性的信息。

4.2.2 知识图谱

虽然一些常识性的知识可以通过大规模语料训练来隐性地将部分知识储存在模型中,但仅依靠这部分知识无法支撑模型对下游任务的处理,还需要引入外部知识来增强模型的训练效果。在预训练模型中,外部知识常用三元组的形式表示,利用实体关系将两种不同实体连接起来,收集大量三元组即可构成知识图谱。有关知识图谱的简单示例如图9所示。外部知识以知识图谱的形式与文本一同输入到预训练模型中,加强模型对知识的理解。

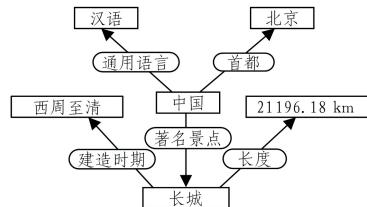


图 9 知识图谱样例

Fig. 9 Sample knowledge graph

百度提出的ERNIE3.0^[43]模型继承了ERNIE2.0中持续学习和多任务学习的方式,可从大规模语料库中学到句法和语义知识,但缺乏世界知识。为了弥补这一缺陷,ERNIE3.0引入大规模知识图谱,使模型在零样本或少样本的条件下也能达到良好的性能。最近ERNIE3.0又取得进一步的突破,成为全球首个知识增强的千亿大模型,参数量高达2600亿,新模型被命名为鹏城·文心^[44](ERNIE 3.0 Titan)。

腾讯QQ浏览器团队提出了“神舟”^[45]模型,该模型引入了基于搜索的知识图谱数据,以此来增强模型对知识的理解能力,并使用3种知识性任务(远监督关系分类、三元组-文本Mask预测和同类实体替换判别)对模型进行训练,实验结果表明,这3种任务均能使模型有效地学习到知识。

4.2.3 语言学知识

语言学知识指语言本身具备的最基本的知识,包括字形和拼音、语义和语法等信息。

香依科技研究团队从汉字的两大特性,即字形和拼音角度出发,提出了融合拼音与字符信息的中文预训练模型Chinese BERT^[46]。该团队成员认为,汉语是一种符号语言,中文字符本身就包含了一些额外的语义信息,而此前提出的中文预训练语言模型缺少该方面的研究。Chinese BERT是第一个引入字形和拼音信息的模型。融合字形与拼音信息的嵌入如图10所示,其中字符嵌入是根据中文的不同字体从视觉特征中获得字符的语义信息。拼音嵌入是学习中文发音的重要特征,有效缓解了中文中普遍存在的多音字现象。

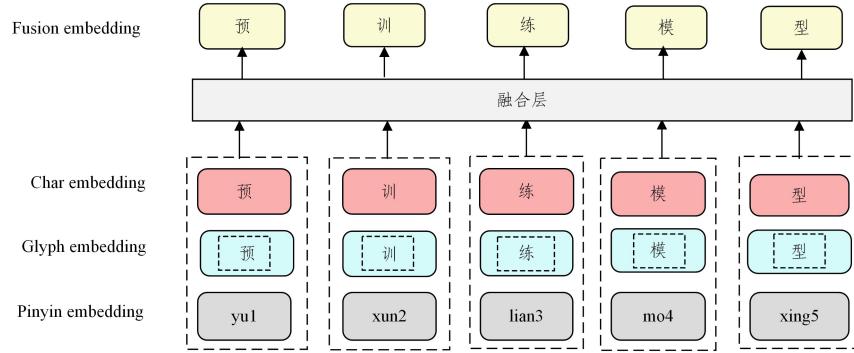


图 10 融合字形和拼音的嵌入^[46]

Fig. 10 Embedding of glyphs and pinyin^[46]

澜舟科技-创新工场等提出了一种轻量化模型孟子^[47],该模型在嵌入层中引入词性标注与语义角色等具有语言学特征的信息,在注意力层引入基于句法的约束信息,以此提高模型对语言学知识的建模能力。

4.2.4 特定知识

腾讯云小微与腾讯AI Lab团队联合提出了一种基于知识的中文预训练模型“神农”^[48],该团队成员致力于将知识融入到模型中,从两方面对知识的处理进行建模,分别是通用型和任务型。通用型知识主要指常识性知识和现有知识,覆盖范围广泛;任务型知识主要指特定场景下的知识。此外,将文章中的推理知识也作为通用知识融入预训练过程中,如对比、转折、因果等指示性的虚词,有助于更好地理解语义信息。

4.3 多模态融合的预训练

模态是指数据的表现形式,主要包括视频、音频、图像和文本等。现实生活中,人们周围充斥着各种模态的信息,当前的互联网音视频资源占比较大,纯文本信息只能覆盖互联网资源的一小部分,更加丰富的音视频信息并没有被充分利用。随着预训练模型的不断发展,研究不能仅局限于文本、视觉等单模态领域,应该对多模态的交互领域进行探索。

目前,多模态预训练模型的研究大多针对图像-文本领域,例如,悟道·文澜团队等提出了图文互检双塔预训练模型BriVL^[49]。以往的大多数多模态研究主要根据图像、文本之间的强语义关系进行显性地建模,但这种建模方式在现实生活中效果往往不佳。因此,该团队成员利用隐性建模的方式

进行多模态预训练,以便更好地理解图像和文本信息。在此基础上,BriVL选择了双塔结构作为基本架构,先将图像和文本信息分别输入两个独立的图像和文本编码器,然后再将图像和文本嵌入拼接在一起。由于双塔结构较为简单,因此进一步地引入对比学习的方式来提高深度神经网络的表示能力,以实现高效的图文互检。

阿里达摩院提出的多模态预训练模型 M6^[32]实现了单模态和多模态数据的统一预训练。除图像到文本生成外,还设计出一个文本引导图像生成的下游任务可以生成高质量图像。该模型参数量高达万亿级别,通过采用专家并行策略及多种优化技术,可以大幅降低能源的消耗。近期,M6 模型规模已扩展至十万亿级别,是目前最大的中文多模态预训练模型。

CogView^[30]是由清华大学等联合发布的一个大规模多模态模型,该模型主要研究文本到图像的生成,使用 VQ-VAE^[50] 和 GPT 作为基本架构。训练分为两个部分,首先利用 VQ-VAE 将图像信息序列化表示为一系列的 token 序列,将图像信息和文本信息一同输入到 GPT 模型中,然后通过 Transformer 解码器来实现生成任务。

除以上多模态模型以外,中国科学院自动化研究所提出的多模态模型 OPT^[51]是全球第一个包含“图文音”3 种模态的预训练模型,其基本架构由 3 个单模态编码器(用于处理每种不同的模态)、1 个跨模态编码器(用于处理 3 种模态之间的相关性)以及 2 个跨模态解码器(用于生成文本和图像)构成,通过这种方式使模型在各种跨模态理解和

生成任务中取得良好性能。

4.4 高效计算的预训练

近年来,预训练模型发展迅速,参数量越来越大,增大神经网络虽然能够提高模型准确率,但也会增大训练模型占用的内存和计算复杂度,因此本节主要讨论如何提高训练效率。

4.4.1 数据预处理阶段

清华大学和智源研究所联合提出的 CPM^[52]模型在数据预处理阶段构建出一个新的子词词汇表,它是一个包含字符和词的子词表,同时在词与词之间引入了额外的分词符,使用特殊标记表示,使得数据预处理更加适合中文的文本规则,以提升模型性能。

最近,浪潮人工智能研究院提出了具有 2450 亿的巨量模型源 1.0^[53],该模型不仅参数量大,而且构建了全球最大的中文数据集。构建如此大的语料库非常困难,因此在数据预处理阶段采用自研的数据过滤系统(Massive Data Filtering System,MDFS)进行处理,主要包括 3 个阶段:数据采集、粗筛和精筛。通过这种方式可获得高质量的数据集,为模型的训练打好坚实的基础。

4.4.2 预训练阶段

ERNIE2.0^[54]在预训练阶段采用持续学习和多任务学习的方式,具体步骤为:首先在基础任务上训练模型,其次持续地引入新任务对模型进行优化,每个新任务都在前一个任务训练的基础上进行,以确保模型不会遗忘先前学到的知识,从而提高训练效率。可持续学习的预训练框架如图 11 所示。

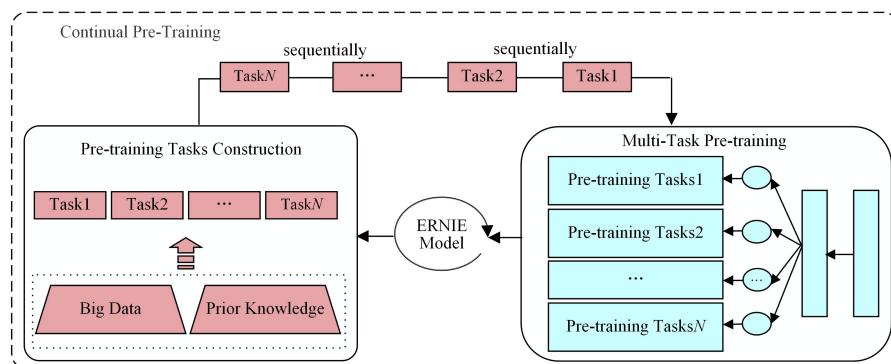


图 11 可持续学习的预训练框架^[54]

Fig. 11 Pre-training framework for sustainable learning^[54]

智源等研究团队提出了基于中文法律文本的预训练模型 Lawformer^[55],该团队成员认为全注意力机制的计算复杂度过高,传统预训练模型在处理长文档时会受到限制,因此采用 Longformer^[56]作为基本编码器,结合滑动窗口注意力机制、扩张滑动窗口注意力机制和全局注意力机制的方式来编码长序列。通过以上 3 种稀疏注意力模式以降低计算复杂度。

目前,深度学习和预训练模型技术不断发展,现有的深度学习框架在大规模预训练阶段面临着巨大挑战,因此,一些大型科技公司为训练超大规模的预训练模型研发了专门的深度学习框架。例如,华为研究团队提出了具有千亿参数量的大模型盘古 α^[57],该模型主要依托华为自研的 MindSpore 深度学习框架进行训练,支持全自动并行策略,通过数据并行、模型并行、流水线(Pipeline)并行等技术,缩短模型训练时间。

悟道·文源团队提出的 CPM-2^[23]在预训练阶段采用 TDS(Tsinghua DeepSpeed)加速框架和知识继承的方式进行预训练。在 TDS 加速框架中,Transformer 将编码器、解码器和编码器加解码器 3 种不同的网络统一起来,可以加速任意的网络模型。此外,CPM-2 模型支持 3 种不同类型的加速方法,包括数据并行、流水线并行和模型并行。知识继承的目的是希望模型能从数据中学习到有用的知识,利用已有模型信息再进行下一阶段的学习,具体分为 3 个阶段。第一阶段,在纯中文文本语料库上预训练,打好中文基础;第二阶段为多语言预训练阶段,主要在中英混合语料库上训练,调整中英文文本比例,缓解遗忘问题;第三阶段使用第二阶段训练好的中英模型初始化对应的 MoE 模型,增强模型的学习能力。

4.4.3 技术型优化

通常,神经网络在训练过程中主要使用单精度(FP32)来存储和计算,然而模型的权重通常会落在有限的范围内,而半精度(FP16)可以减少内存占用并能加快训练速度,但在半精度浮点数下进行训练可能会出现浮点截断和溢出等问题。因此,Micikevicius 等^[58]提出了一种混合精度训练法来帮助解决上述问题。混合精度在保证模型训练精度的前提下,利用半精度浮点数来加速模型训练。具体来说,将训练过程中产生的权重、激活、梯度利用 FP16 进行存储,同时拷贝一份 FP32 的权重参数,用于优化时的更新。在每次迭代训练中,将存储在 FP16 格式中的权重、激活、梯度进行前向和后向传播。混合精度训练引入了权重备份(Weight Backup)、损失放大(Loss Scaling)和精度累加(Precision Accumulated) 3 种技术。上文中提到的中文预训练模型 NEZHA, Motian, CPM-2, M6 等都有采用混合精度训练策略。

悟道·文源团队提出的 CPM-2^[23] 模型,除了上述预训练阶段的优化外,在模型微调方面探索了全参微调、提示微调以及先提示微调后全参微调 3 种微调机制,其中提示微调(Prompt Tuning)^[59] 指预训练加提示的范式,与传统微调(Fine Tuning)的区别在于,其不需要模型去适应不同的下游任务,而是让复杂的下游任务在提示的帮助下更好地适配预训练模型。

4.5 特定领域和特定任务的预训练

4.5.1 开放域对话系统

近年来,大规模预训练语言模型的蓬勃发展促进了开放领域对话系统的研究。例如,百度在 2019 年 10 月预发布通用领域的对话生成模型 PLATO,在 2020 年提出了超大规模对话生成预训练模型 PLATO-2^[60],该模型将训练分为两个阶段。首先,在一对一映射的简化框架下训练粗粒度生成模型来学习回复生成;其次,训练一个包含潜在变量的细粒度生成模型和一个评估模型,用于不同的回复生成和响应一致性估计。2021 年 9 月份,百度又取得新的突破,发布了 PLATO-XL^[61],该模型的参数规模首次突破百亿,是当前最大的中英文对话预训练生成模型,并且在第十届对话系统挑战赛上的多个任务中获得冠军。

由于缺乏公开的大规模高质量中文对话数据集,中文对话领域的研究较为滞后,因此 Wang 等^[62]构建了包含 1200 万对话的中文对话数据集 LCCC,并在此数据集上训练出大规模中文对话生成模型 CDial-GPT。该模型首先在包含不同话题的小说数据集上进行训练,生成中文小说 GPT 预训练模型;其次,使用 LCCC 中的对话数据对该模型继续训练,最终生成中文对话模型 CDial-GPT,并且与 CPM 模型进行了对比,原因是 CPM 的架构与 GPT 类似,且这两种模型同时引入对话生成任务,实验结果表明 CPM 的性能明显优于 CDial-GPT。近期,清华大学在开放域对话生成方面的研究又取得了新的突破,和智源研究所一同合作提出了大规模生成式的开放域中文对话系统 EVA^[63],Zhou 等认为,和英文的开放域对话系统相比,预训练模型仍受到对话数据和模型大小的限制,因此提出了包含 28 亿参数的 EVA 模型,是目前最大的中文预训练对话模型,同时构建出目前最大的中文对话数据集

(WDC Dialogue),其中包含 14 亿上下文响应对。

4.5.2 其他领域的预训练

受通用领域预训练语言模型(PLM)成功的启发,许多研究人员致力于将预训练模型应用到其他的领域或任务中。例如,智源等研究团队提出了基于中文法律文本的预训练模型 Lawformer^[55],用于解决法律任务,该模型包含数千万个刑事和民事案件文档。

熵简科技人工智能实验室提出了基于金融领域语料训练的 BERT 模型 FinBERT^[64],该模型采用和原始 BERT 相同的架构,包含 base 和 large 两个版本。为了使模型能更好地学习到金融领域词句的语义特征,引入了两种有监督的学习任务,即研报行业分类和财经新闻的金融实体识别。

2020 年初新冠肺炎疫情爆发,在线教育站在发展的顺风口极大地推动了教育领域人工智能技术的发展,然而在线教育领域的自然语言处理受到了数据的限制,其原因是,从有关教育音频和视频中转录的文本数据存在很多错误,而且教育领域中包含很多专有词汇。为了解决这些问题,好未来 AI 中台机器学习团队收集了大量教育领域的中文数据,并在此基础上构建出教育领域首个在线教学中文预训练模型 TAL-EduBERT^[65],其模型结构与 BERT 类似。

北京理工大学在 BERT-wwm 模型的基础上提出了 GuwenBERT^[66],该模型在古文数据集中训练而成,有 base 和 large 两种版本,模型训练后能自动帮助文言文断句,还能学习被掩码的词语。清华大学人工智能研究所自然语言处理与社会人文计算研究中心在 GitHub 上开源了中国古典诗词预训练模型 BERT-CCPoem^[67],该模型是在中国古典诗词的语料库 CCPoem-Full v1.0 中训练而成。

随着生物医学领域网络数据资源的快速增长,对生物医药学领域的研究变得愈发重要。阿里巴巴研究团队提出了针对生物医学领域的预训练模型 MC-BERT^[68],并发布了中文生物医学语言评测基准(ChineseBLUE)。该基准目前已更新(在后文 5.4 小节具体介绍)。目前的预训练模型大多在文本数据上进行训练,较少涉及一些需要对结构化的表格数据进行处理的任务。英文领域已有一些针对表格领域的探索,但中文领域关于表格的预训练还处于空白状态。基于此,阿里达摩院研究团队提出了预训练表格模型 SDCUP^[69],是中文领域首个表格类型的预训练模型。

4.6 其他模型

除上述由中国学者提出的中文预训练模型外,还有部分典型的英文预训练模型在中文语料库上训练且开源的中文版本。这些模型主要分为 3 类。第一类侧重于自然语言理解,主要包括 BERT_base, RoBERTa_zh, ALBERT_zh, Chinese-XLNET, Chinese-ELECTRA, Chinese_RoFormer, Struct-BERT_ch, large; 第二类侧重于自然语言生成,主要包括 GPT2-ml, Chinese-Transformer-XL, Chinese_T5, T5-PEGASUS, Chinese-BART; 第三类是通用型,主要包括 Chinese UniLM, Chinese_Simbert 和 Chinese_RoFormer-sim。以上模型大部分由国外著名的科技公司提出,小部分由追一科技公司提出,表 3 总结了这些基础模型所开源的中文版的源地址、公开语料库、中文评测基准和其他资源(注意:部分模型会有多种变体,此处只整理其中一个,其他变体见其他资源中的汇总信息)。

表3 训练模型相关资源汇总

Table 3 Summary of resources related to pre-trained models

资源	模型	源地址
BERT_base ^[4]		https://github.com/google-research/bert
RoBERTa_zh ^[19]		https://github.com/brightmart/roberta_zh
Albert_zh ^[20]		https://github.com/brightmart/albert_zh
Chinese-XLNet ^[71]		https://github.com/ymcui/Chinese-XLNet
Chinese-ELECTRA ^[72]		https://github.com/ymcui/Chinese-ELECTRA
Chinese-Roformer ^[73]		https://github.com/ZhuiyiTechnology/roformer
StructBERT. ch. large ^[40]		https://github.com/alibaba/AliceMind/tree/main/StructBERT
GPT2-ml		https://github.com/imcaspar/gpt2-ml
Chinese-Transformer-XL ^[74]		https://github.com/THUDM/Chinese-Transformer-XL
ChineseT5 ^[75]		https://github.com/dbiir/UER-py
T5 PEGASUS ^[76]		https://github.com/ZhuiyiTechnology/t5-pegasus
Chinese BART ^[77]		https://github.com/fastnlp/CPT
Chinese UniLM ^[78]		https://github.com/YunwenTechnology/Unilm
Chinese Simbert ^[79]		https://github.com/ZhuiyiTechnology/pretrained-models
Chinese RoFormer-sim ^[80]		https://github.com/ZhuiyiTechnology/roformer-sim
语料库		
zh. wikipedia		https://dumps.wikimedia.org/zhwiki/latest/
LCCC ^[62]		https://github.com/thu-coai/CDial-GPT
CLUECorpus2020 ^[81]		https://github.com/CLUEbenchmark/CLUECorpus2020
Chinese WPLC		https://git.openi.org.cn/PCL-Platform.Intelligence/Chinese_WPLC
WuDaoCorpora 2.0		https://data.wudaoai.cn
THUAIPOet Datasets		https://github.com/THUNLP-AIPoet/Datasets
评测基准		
中文语言理解测评基准(CLUE) ^[82]		https://github.com/CLUEbenchmark/CLUE
中文多模态测评基准(MUGE)		https://tianchi.aliyun.com/muge
中文医学语言理解测评基准(CBLUE) ^[83]		https://github.com/CBLUEbenchmark/CBLUE
智源指数(CUGE) ^[84]		cuge.baai.ac.cn
其他相关资源		
典型的中文预训练模型汇总		https://github.com/lonePatient/awesome-pretrained-chinese-nlp-models
公开的中文语料库汇总		https://github.com/brightmart/nlp_chinese_corpus
中文词向量汇总		https://github.com/Embedding/Chinese-Word-Vectors
中文自然语言处理资源库		https://github.com/fighting41love/funNLP

5 中文领域的评测基准

5.1 为什么建立中文领域的评测基准

首先,从使用人数上看,中国人口占世界人口的五分之一,人数庞大,因此中文是世界上使用人数最多的语言;其次,从语言体系上看,中文与英文差异较大;最后,从数据集角度出发,中文领域公开可用的数据集较少,此前提出的中文预训练模型在英文评测基准上评估,无法完全体现出模型性能。当下预训练模型的发展极其迅速,英文领域的评测基准已步入成熟阶段,而中文领域的缺失必然会导致技术落后,因此中文领域的评测基准必不可少。本节主要介绍4种不同的评测基准。

5.2 中文语言理解测评基准

中文语言理解测评基准^[82](Chinese Language Understanding Evaluation Benchmark, CLUE)用于评估中文预训练模型在不同任务上的性能。由于英文预训练模型发展迅速,此前已出现英文自然语言理解评测基准,主要包括GLUE和SuperGLUE。这两种评测基准数据规范,体积庞大,可以在多个任务上全方位考验模型性能,几乎所有预训练模型都以在这两种评测基准上实现SOTA结果为目标。然而,该评测基准更适用于英文,中文评测并不适用,原因是中文在语言上与英文或者其他印欧语系差异较大,因此自然语言处理领域的众多研究者联合开展CLUE项目,希望为科研工作者提供一个高质量衡量模型的平台,以促进预训练模型在中文语言处理能力上的提升。表4列出了CLUE中包含的9个自然语言理解任务。

表4 测评基准的任务汇总

Table 4 Summary of tasks for measurement benchmark

任务名	训练/验证/测试集数量	任务类型	评估指标
		CLUE 的 9 个自然语言处理任务	
TNEWS	53.3k/10k/10k	今日头条中文新闻(短文)分类	Accuracy
IFLYTEK	12.1k/2.6k/2.6k	长文本分类	Accuracy
CLUEWSC2020	1244/304/290	代词消歧	Accuracy
AFQMC	34.3k/4.3k/3.9k	蚂蚁金融语义相似度	Accuracy

(续表)

CSL	20k/3k/3k	论文关键词识别	Accuracy
OCNLI	50k/3k/3k	自然语言推理	Accuracy
CMRC 2018	10k/3.4k/4.9k	中文阅读理解任务	EM
ChiD	577k/23k/23k	成语阅读理解填空	Accuracy
C3	11.9k/3.8k/3.9k	中文多选阅读理解	Accuracy
CBLUE 的 8 个医疗语言理解任务			
CMeEE	15k/5k/3k	中文医学命名实体识别	Micro F1
CMeIE	14.4k/3.6k/4.5k	中文医学文本实体关系抽取	Micro F1
CHIP-CDN	6k/2k/10.2k	临床术语标准化任务	Micro F1
CHIP-STS	16k/4k/10k	临床试验筛选标准短文本分类	Macro F1
CHIP-CTC	22.9k/7.7k/10k	平安医疗科技疾病问答迁移学习	Macro F1
KUAKE-QIC	6.9k/1.9k/1.9k	医疗搜索检索词意图分类	Accuracy
KUAKE-QTR	24.1k/2.9k/5.4k	医疗搜索查询词-页面标题相关性	Accuracy
KUAKE-QQR	15k/1.6k/1.6k	医疗搜索查询词-查询词相关性	Accuracy

5.3 中文医学语言理解测评基准

中文医学语言理解测评基准^[83](Chinese Biomedical Language Understanding Evaluation Benchmark, CBLUE)是由中国中文信息学会医疗健康与生物信息处理委员会所发布,包括数据集、基准模型和排行榜,以此促进人工智能技术在医疗领域的发展。

5.4 中文多模态测评基准

近年来,预训练技术的成功实践不仅推动着自然语言处理领域的快速发展,而且也推动着多模态领域的研究。然而,现存的多模态评测基准以英文为主,中文领域缺少多模态的评测基准。考虑到中文多模态领域的快速发展,阿里达摩院等在计算机视觉委员会的协助下推出了大规模中文多模态的评测基准(Multimodal Understanding and Generation Evaluation Benchmark, MUGE)。目前,这一评测基准刚刚被提出,正处于起步发展阶段,只公布了3种任务,如表5所列。未来也会增加更多的多模态任务和数据集,希望能为科研工作者提供支持。

表5 MUGE的3个多模态任务

Table 5 Three multimodal tasks of MUGE

任务名	任务类型	描述
E-Commerce IC (Image Caption)	图像描述生成	根据给定图片生成相应文字描述
E-Commerce T2I (Text to Image)	文本到图像生成	根据给定文本生成相应图片
Multimodal Retrieval Dataset	多模态检索	考察模型对图文理解和匹配的能力

5.5 智源指数

2021年末,智源研究院与其他单位联合研制并推出一种全面均衡的机器中文语言能力评测基准智源指数^[85](Chinese Language Understanding and Generation Evaluation, CUQE),该指数具有以下特点:1)层次化基准框架,以人类语言能力为参照,它涵盖7种重要语言能力、17个主流任务、19个代表性数据集;2)多层次得分策略,基于层次结构框架,除总分数以外,CUQE还提供不同级别模型性能的评估,包括数据集、任务、语言能力的性能。CUQE能帮助研究者更多地关注模型本身的改进,提升对模型发展的指导性。

6 研究趋势与展望

中文预训练模型已在多个领域实现商业化落地,并展现

出一定的市场潜力,取得了长足发展,但也存在较多挑战,例如预训练模型规模和性能之间的平衡问题;如何构建更加通用型的预训练模型;如何突破现有多模态和轻量化模型的瓶颈;如何构建融入更多中文特色的预训练模型等。本文主要从以下几个方面对未来进行展望。

6.1 规模

随着以BERT和GPT等为代表的大型预训练模型的出现,逐渐掀起了预训练模型朝大规模方向发展的浪潮。大量的研究表明,模型参数量越大,训练数据量越多的预训练模型表现更出色。中文领域存在众多大规模预训练模型,如源1.0参数2457亿,训练数据集达5000GB;ERNIE 3.0 Titan参数2600亿;中文多模态模型M6参数量已经扩展至十万亿级别。目前预训练模型还未达到模型的性能极限,增大模型参数量和训练数据仍是提高模型性能最有效的手段,探索超大规模预训练模型的道路还将继续,也需要更加注重模型的创新性、训练的低碳化和应用的高效性。

然而,训练超大规模的模型仍存在很大挑战。首先,使用最大GPU也不可能在内存中拟合所有参数;其次,算法优化不足会耗费极长的训练时间;最后,搭建超大规模模型会带来巨大的成本,让学术界和小型科技公司望而却步。如何在模型性能和成本之间取得平衡也是当前学者探索的另外一条道路,如探索轻量化的预训练模型。近期腾讯提出的“神农”、澜舟科技提出的“孟子”及IDEA研究院提出的“二郎神”等轻量化模型,仅以十亿左右的参数量就在部分任务上达到了SOTA结果,因此探索轻量化模型势在必行。

6.2 融入外部信息

预训练模型在部分任务上已无限接近人类,甚至超越人类,然而,其对知识的掌握依旧不足,如何让预训练模型真正理解并运用知识是一个值得长期研究的课题,尤其是中华民族上下五千年形成的文化知识颇多,比如“常识性知识”和“特定领域的知识”等。特定领域的知识可以帮助模型挖掘不同领域特有的知识,如果能够将特定领域的行业知识与模型结合起来训练,不仅可以将预训练模型更广泛地应用到不同的下游任务,在各行各业中实现良好的产业落地,而且可以与脑科学、心理学、神经学等其他学科融合,更好地发展人工智能,服务人类生活。

除了融入知识信息之外,还可以从中文字形和字音等方面考虑。因为中文语言的特殊性,其字符的符号也包含一些

额外信息,这些额外信息能增强中文自然语言的表现力,如ChineseBERT^[46]模型中提出将中文字形和拼音信息融入预训练模型中,以此增强模型对中文语料的建模能力,但这一方向的研究还相对较少,仍有待完善。

6.3 多模态领域

现实世界离不开语言,语言离不开语音和视觉信息,类似于人的感觉器官:眼、耳、嘴,任何一样的缺失都会影响生活。当前,互联网音视频资源占比较大,纯文本信息只能覆盖互联网资源的一小部分,更加丰富的音视频信息并没有被充分利用,因此预训练模型必然朝着多模态的趋势发展。目前,多模态预训练模型的研究大多只考虑了两种模态,图像文本或者视频文本,而音频信息大多被忽视。中文预训练模型起步虽晚,但成绩斐然。中科院自动化所提出了全球首个图文音(视觉-文本-语音)三模态的预训练模型OPT^[51],该模型同时具备跨模态理解与生成的能力。通过上述分析可知,多模态的研究拥有很大的发展空间。

结束语 本文主要围绕中文预训练模型的研究现状进行概述。从模型规模上看,中文预训练模型的发展正处于两条道路上。一是朝着超大规模预训练模型的方向发展;二是寻求轻量化模型的发展。从外部信息来看,大多数的预训练模型都融入了各种知识,预训练与先验知识的深度融合刻不容缓。从高效训练上看,现有模型都在不断地探索更加高效的训练方式。从多模态的角度上看,中文多模态预训练模型的发展正处于上升阶段,正朝着更多模态、更加通用的方向发展。从特定领域的模型来看,预训练模型可应用于多种领域,具有较大的发展潜力。综上所述,中文预训练模型虽然取得了不可忽视的成绩,但还有更大的发展空间,未来将朝着更大规模、更加高效、适用更多领域的方向发展。

参 考 文 献

- [1] LIU P F, QIU X P, HUANG X J. Recurrent neural network for text classification with multi-task learning[C]// Proceedings of the 2016 Conference on IJCAI. 2016: 2073-2879.
- [2] KRIZHEVSKY A, SUSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. London: MIT Press, 2012: 1097-1105.
- [3] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. arXiv: 1409.0473v7, 2014.
- [4] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019: 4171-4186.
- [5] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv: 1301.3781v1, 2013.
- [6] PENNINGTON J, SOCHER R, MANNING C D. GloVe: Global Vectors for Word Representation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Proces-
- sing(EMNLP). 2014: 1532-1543.
- [7] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of Tricks for Efficient Text Classification[J]. arXiv: 1607.01759, 2016.
- [8] PETERS M, NEUMANN M, LYNNER M, et al. Deep Contextualized Word Representations[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. 2018: 2227-2237.
- [9] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting[J]. arXiv: 1506.04214, 2015.
- [10] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[OL]. [2022-04-15]. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.
- [11] WANG A, SINGH A, MICHAEL J, et al. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding[J]. arXiv: 1804.07461, 2018.
- [12] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [13] CHO K, MERRIENBOER B V, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1724-1734.
- [14] VASWANI A, SHAZER N, PARMAR N, et al. Attention Is All You Need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.
- [15] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv: 1609.08144, 2016.
- [16] SUN Y, WANG SH, LI Y K, et al. ERNIE: enhanced representation through knowledge integration[J]. arXiv: 1904.09223, 2019.
- [17] WEI J, REN X, LI X, et al. NEZHA: Neural Co-ntextualized Representation for Chinese Language Understanding[J]. arXiv: 1909.00204, 2019.
- [18] CUI Y, CHE W, LIU T, et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing[J]. arXiv: 2004.13922, 2020.
- [19] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. [J]. arXiv: 1907.11692, 2019.
- [20] ERLANGSHEN Pre-training model [OL]. [2021-11-15]. <https://huggingface.co/IDEA-CCNL/Erlangshen-1.3B>.
- [21] PEKRS Pre-training model [OL]. [2021-11-16]. https://mp.weixin.qq.com/s/r85W7T26vy6_IIRAWY1ZKA.
- [22] LAI Y, LIU Y, FENG Y, et al. Lattice-BERT: Leveraging Multi-Granularity Representations in Chinese Pre-Trained Language Models[J]. arXiv: 2104.07204, 2021.
- [23] ZHANG Z, GU Y, HAN X, et al. CPM-2: Large-Scale Cost-Effective Pre-Trained Language Models[J]. arXiv: 2106.10715, 2021.

- [24] MOTIAN Pre-training model [OL]. [2021-06-24]. <https://mp.weixin.qq.com/s/HQL0Hk49UR6kVNtrvcXEGA>.
- [25] ZHANG R,PANG C,ZHANG C,et al. Correcting Chinese Spelling Errors with Phonetic Pre-Training[C]// Findings of the Association for Computational Linguistics: ACL-IJCNLP. 2021: 2250-2261.
- [26] SHAW P,USZKOREIT J,VASWANI A. Self-Attention with Relative Position Representations[J]. arXiv:1803.02155,2018.
- [27] IOFFE S,SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]// International Conference on Machine Learning. 2015;448-456.
- [28] BA J L, KIROS J R, HINTON G E. Layer normalization[J]. arXiv:1607.06450,2016.
- [29] BERTSG Pre-training model [OL]. [2021-03-15]. <https://baijiahao.baidu.com/s?id=1695185167027662850&wfr=spider&for=pc>.
- [30] DING M,YANG Z,HONG W,et al. CogView: Mastering Text-to-Image Generation via Transformers[J]. arXiv:2105.13290, 2021.
- [31] SHAZEE N,MISHOSEINI N,MAZIARZ K,et al. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer[J]. arXiv:1701.06538,2017.
- [32] LIN J,MEN R,YANG A,et al. M6: A Chinese Multimodal Pre-trainer[J]. arXiv:2103.00823,2021.
- [33] YANG A,LIN J,MEN R,et al. M6-T: Exploring Sparse Expert Models and Beyond[J]. arXiv:2105.15082, 2021.
- [34] DIAO S Z,BAI J X,SONG Y,et al. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations[C]// Findings of the Association for Computational Linguistics. 2020: 4729-4740.
- [35] SONGY,ZHANGT,WANGY,etal. ZEN 2.0: Continue Training and Adaption for N-gram Enhanced Text Encoders[J]. arXiv: 2105.01279,2021.
- [36] ZHANG X,LI P,LI H. AMBERT: A Pre-Trained Language Model with Multi-Grained Tokenization [J]. arXiv:2008.11869, 2020.
- [37] GUO W,ZHAO M,ZHANG L,et al. LICHEE: Improving Language Model Pre-Training with Multi-Grained Tokenization[J]. arXiv:2108.00801,2021.
- [38] WoBERT Pre-training model [OL]. [2020-09-18]. <https://ke-xue.fm/archives/7758>.
- [39] PLUG Pre-training model [OL]. [2021-04-19]. https://mp.weixin.qq.com/s/-aV6Hh-BFoW41HQop_Z02w.
- [40] WANG W,BI B,YAN M,et al. StructBERT: Incorporating Language Structures into Pre-Training for Deep Language Understanding[J]. arXiv:1908.04577,2019.
- [41] BI B,LI C,WU C,et al. PALM: Pre-training an Autoencoding & Autoregressive Language Model for Context-conditioned Generation[J]. arXiv:2004.07159,2020.
- [42] SHAO Y,GENG Z,LIU Y,et al. CPT: A Pre-Trained Unbalanced Transformer for Both Chinese Language Understanding and Generation[J]. arXiv:2109.05729,2021.
- [43] SUN Y,WANG S,FENG S,et al. ERNIE 3.0: Large-Scale Knowledge Enhanced Pre-Training for Language Understanding and Generation[J]. arXiv:2107.02137,2021.
- [44] WANG S,SUN Y,XIANG Y,et al. ERNIE 3.0 Titan: Exploring Larger-scale Knowledge Enhanced Pre-training for Language Understanding and Generation[J]. arXiv:2112.12731, 2021.
- [45] SHEN Z. Pre-training model [OL]. [2021-09-30]. <https://www.jiqizhixin.com/articles/2021-09-30-2>.
- [46] SUN Z,LI X,SUN X,et al. ChineseBERT: Chinese Pretraining Enhanced by Glyph and Pinyin Information[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(Volume 1: Long Papers). 2021: 2065-2075.
- [47] ZHANG Z,ZHANG H,CHEN K,et al. Mengzi: Towards Lightweight yet Ingenious Pre-Trained Models for Chinese[J]. arXiv:2110.06696,2021.
- [48] SHEN N. Pre-training model [OL]. [2021-10-20]. https://mp.weixin.qq.com/s/coW_OlBRA4lwVLZaRyxO9Q.
- [49] HUO Y,ZHANG M,LIU G,et al. WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training [J]. arXiv:2103.06561,2021.
- [50] OORD A,VINYALS O,KAVUKCUOGLU K. Neural discrete representation learning[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017:6309-6318.
- [51] LIU J,ZHU X,LIU F,et al. OPT: Omni-Perception Pre-Trainer for Cross-Modal Understanding and Generation[J]. arXiv: 2107.00249,2021.
- [52] ZHANG Z,HAN X,ZHOU H,et al. CPM: A Large-Scale Generative Chinese Pre-Trained Language Model[J]. arXiv: 2012.00413,2020.
- [53] WU S,ZHAO X,YU T,et al. Yuan 1.0: Large-Scale Pre-Trained Language Model in Zero-Shot and Few-Shot Learning [J]. arXiv:2110.04725,2021.
- [54] SUN Y,WANG S,LI Y,et al. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5):8968-8975.
- [55] XIAO C,HU X,LIU Z,et al. Lawformer: A Pre-Trained Language Model for Chinese Legal Long Documents [J]. arXiv: 2105.03887,2021.
- [56] BELTAGY I,PETERS M E,COHAN A. LongFormer: The Long-Document Transformer[J]. arXiv:2004.05150,2020.
- [57] ZENG W,REN X,SU T,et al. PanGu-\$\alpha\$: Large-Scale Autoregressive Pretrained Chinese Language Models with Auto-Parallel Computation[J]. arXiv:2104.12369,2021.
- [58] MICIKEVICIUS P,NARANG S,ALBEN J,et al. Mixed Precision Training[J]. arXiv:1710.03740,2017.
- [59] LESTER B,AL-RFOU R,CONSTANT N. The power of scale for parameter-efficient prompt tuning[J]. arXiv:2104.08691, 2021.
- [60] BAO S,HE H,WANG F,et al. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning[J]. arXiv: 2006.16779,2020.

- [61] BAO S,HE H,WANG F,et al.PLATO-XL:Exploring the Large-Scale Pre-Training of Dialogue Generation[J].arXiv:2109.09519,2021.
- [62] WANG Y,KE P,ZHENG Y,et al.A Large-Scale Chinese Short-Text Conversation Dataset[J].arXiv:2008.03946,2020.
- [63] ZHOU H,KE P,ZHANG Z,et al.EVA:An Open-Domain Chinese Dialogue System with Large-Scale Generative Pre-Training[J].arXiv:2108.01547,2021.
- [64] LIU Z,HUANG D,HUANG D,et al.FinBERT:A Pre-trained Financial Language Representation Model for Financial Text Mining[C]//Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence.2021:4513-4519.
- [65] TAL-EduBERT Pre-training model [OL].[2021-01-26].<https://github.com/tal-tech/edu-bert>.
- [66] GUWENBERT Pre-training model [OL].[2021-08-31]<https://github.com/ethan-yt/guwenbert>.
- [67] BERT-CCPoem Pre-training model [OL].[2021-7-5]<https://github.com/THUNLP-AIPoet/BERT-CCPoem>.
- [68] ZHANG N,JIA Q,YIN K,et al.Conceptualized Representation Learning for Chinese Biomedical Text Mining[J].arXiv:2008.10813,2020.
- [69] HUI B,SHI X,GENG R,et al.Improving Text-to-SQL with Schema Dependency Learning[J].arXiv:2103.04399,2021.
- [70] LAN Z,CHEN M,GOODMAN S,et al.ALBE-RT:A Lite BERT for Self-Supervised Learning of Language Representations[J].arXiv:1909.11942,2019.
- [71] YANG Z,DAI Z,YANG Y,et al.XLNet:Generalized Autoregressive Pretraining for Language Understanding[J].arXiv:1906.08237,2019.
- [72] CLARK K,LUONG M T,LE Q V,et al.ELEC-TRA:Pre-Training Text Encoders as Discriminators Rather Than Generators[J].arXiv:2003.10555,2020.
- [73] SU J,LU Y,PAN S,et al.RoFormer:Enhanced Transformer with Rotary Position Embedding[J].arXiv:2104.09864,2021.
- [74] DAI Z,YANG Z,YANG Y,et al.Transformer-XL:Attentive Language Models Beyond a Fixed-Length Context[J].arXiv:1901.02860,2019.
- [75] RAFFEL C,SHAZEER N,ROBERTS A,et al.Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer[J].arXiv:1910.10683,2019.
- [76] ZHANG J,ZHAO Y,SALEH M,et al.PEGASU-S:Pre-Trai-
- ning with Extracted Gap-Sentences for Abstractive Summarization[J].arXiv:1912.08777,2019.
- [77] LEWIS M,LIU Y,GOYAL N,et al.BART:Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.2020:7871-7880.
- [78] DONG L,YANG N,WANG W,et al.Unified Language Model Pre-Training for Natural Language Understanding and Generation[J].arXiv:1905.03197,2019.
- [79] SU J L.SimBERT Pretraining model [OL].[2020-05-18].<https://www.spaces.ac.cn/archives/7427>.
- [80] SU J L.RoFormer-Sim Pretraining model [OL].[2021-06-11].<https://www.spaces.ac.cn/archives/8454>.
- [81] XU L,ZHANG X,DONG Q.CLUECorpus2020:A Large-Scale Chinese Corpus for Pre-Training Language Model[J].arXiv:2003.01355,2020.
- [82] XU L,HU H,ZHANG X,et al.CLUE:A Chinese Language Understanding Evaluation Benchmark[J].arXiv:2004.05986,2020.
- [83] ZHANG N,CHEN M,BI Z,et al.CBLUE:A Chinese Biomedical Language Understanding Evaluation Benchmark[J].arXiv:2106.08087,2021.
- [84] YAO Y,DONG Q,GUAN J,et al.CUGE:A Chinese Language Understanding and Generation Evaluation Benchmark[J].arXiv:2112.13610,2021.



HOU Yu-tao, born in 1998, postgraduate, is a student member of China Computer Federation. Her main research interests include natural language processing and so on.



ABUDUKELIMU Abulizi, born in 1978, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include artificial intelligence and natural language processing.

(责任编辑:喻藜)