计算机科学
COMPUTER SCIENCE

预训练语言模型的扩展模型研究综述

阿布都克力木·阿布力孜，张雨宁，阿力木江·亚森，郭文强，哈里旦木·阿布都克里木

引用本文

阿布都克力木·阿布力孜，张雨宁，阿力木江·亚森，郭文强，哈里旦木·阿布都克里木.预训练语言模型的扩展模型研究综述[J]．计算机科学, 2022, 49(11A): 210800125-12.

Abudukelimu ABULIZI, ZHANG Yu-ning, Alimujiang YASEN, GUO Wen-qiang, Abudukelimu HALIDANMU. Survey of Research on Extended Models of Pre-trained Language Models[J]. Computer Science, 2022, 49(11A): 210800125-12.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800125

预训练语言模型的扩展模型研究综述

阿布都克力木·阿布力孜[1,2]　张雨宁[1]　阿力木江·亚森[1]　郭文强[1]　哈里旦木·阿布都克里木[1,2]
1 新疆财经大学信息管理学院　乌鲁木齐 830012
2 新疆财经大学丝路经济与管理研究院　乌鲁木齐 830012
(keram1106@163.com)

摘　要　近些年,Transformer神经网络的提出,大大推动了预训练技术的发展。目前,基于深度学习的预训练模型已成为了自然语言处理领域的研究热点。自2018年底BERT在多个自然语言处理任务中达到了最优效果以来,一系列基于BERT改进的预训练模型相继被提出,也出现了针对各种场景而设计的预训练模型扩展模型。预训练模型从单语言扩展到跨语言、多模态、轻量化等任务,使得自然语言处理进入了一个全新的预训练时代。主要对轻量化预训练模型、融入知识的预训练模型、跨模态预训练语言模型、跨语言预训练语言模型的研究方法和研究结论进行梳理,并对预训练模型扩展模型面临的主要挑战进行总结,提出了4种扩展模型可能发展的研究趋势,为学习和理解预训练模型的初学者提供理论支持。

关键词:自然语言处理;预训练;轻量化;知识融合;多模态;跨语言
中图法分类号　TP391

Survey of Research on Extended Models of Pre-trained Language Models

Abudukelimu ABULIZI[1,2],ZHANG Yu-ning[1],Alimujiang YASEN[1],GUO Wen-qiang[1] and Abudukelimu HALIDANMU[1,2]
1 School of Information Management,Xinjiang University of Finance and Economics,Urumqi 830012,China
2 Institute of Silk Road Economy and Management,Xinjiang University of Finance and Economics,Urumqi 830012,China

Abstract　In recent years,the proposal of Transformer neural network has greatly promoted the development of pre-training technology. At present,pre-training models based on deep learning have become a research hotspot in the field of natural language processing. Since the end of 2018,BERT has achieved optimal results in multiple natural language processing tasks. A series of improved pre-training models based on BERT have been proposed one after another,and pre-training model extension models designed for various scenarios have also appeared. The expansion of pre-training models from single-language to tasks such as cross-language,multi-modality,and light-weighting has enabled natural language processing to enter a new era of pre-training. This paper mainly summarizes the research methods and research conclusions of lightweight pre-training models,knowledge-incorporated pre-training models,cross-modal pre-training language models and cross-language pre-training language models,as well as the main challenges faced by the pre-training model expansion model. In summary,four research trends for the possible development of extended models are proposed to provide theoretical support for beginners who learn and understand pre-training models.

Keywords　Natural language processing,Pre-training,Lightweight,Knowledge-incorporated,Cross-modal,Cross-language

1　引言

自然语言处理(Natural Language Processing,NLP)是计算机科学领域与人工智能领域的一个分支,旨在让机器能够理解人类的语言。随着NLP技术的发展以及计算力要求的不断增强,深层神经网络也不断被提出。为了充分训练深层模型参数并防止过拟合,往往需要更多的标注数据,但是监督的标注数据往往成本较大,因此需要充分利用网络中现存的大量无监督数据对模型进行训练[1]。在这种背景下,预训练技术被广泛地应用在自然语言处理领域,通过自监督的学习方法从大量的无监督数据中训练一个深层的预训练模型并迁移到下游具体任务有很好的指导作用。

预训练模型首先在计算机视觉领域ImageNet数据集上兴起。通过自监督学习从大规模任务无关的无标注语料库中训练深层的网络模型,得到一组训练参数,即一个词在某个特定上下文语境的语义表示,其本质可以被当成一种很好的初始化或正规化,旨在对文本的内隐知识进行表示。为了更好地学习语言表征,并将其在其他特定自然

基金项目:国家自然科学基金项目(61866035,61966033);2018年度自治区高层次人才引进项目(40050027);2018年度自治区科学技术厅天池博士项目(40050033);国家重点研发专项(2018YFC0825504)
This work was supported by the National Natural Science Foundation of China(61866035,61966033),2018 High-level Talented Person Project of Department of Human Resources and Social Security of Xinjiang Uyghur Autonomous Region(40050027),2018 Tianchi Ph.D Program Scientific Research Fund of Science and Technology Department of Xinjiang Uyghur Autonomous Region(40050033) and National Key Research and Development Program of China(2018YFC0825504).
通信作者:哈里旦木·阿布都克里木(abdklmhldm@gmail.com)

语言处理任务上进行微调，将其中深层的网络模型称为“预训练模型”。早期的预训练模型（如 word2vec[2] 和 GloVe 模型[3]）虽然可以提取词的上下文语义信息，但无法体现一个词在不同语境的不同含义。动态预训练模型（如 Cove[4]，ELMo[5]，GPT[6] 和 BERT[1] 等）对词动态提取符合其上下文语境的语义表示，该词表征向量为一个动态变量，可以根据上下文语境动态地更新语义信息，解决了一词多义的问题。

自 2018 年底 BERT 在多个自然语言处理任务中达到了最优效果（State Of The Art，SOTA）以来，一系列基于 BERT 改进的预训练模型相继被提出。这些模型进而推动了深度学习与预训练技术的发展浪潮，也出现了针对各种场景而设计的预训练模型扩展模型。

2 轻量化预训练模型

在自然语言处理工作中，虽然预训练＋微调的组合范式对下游任务性能有较大的提升[6]，但仍存在模型规模巨大、延迟性高的问题。自 BERT 模型提出后，许多预训练模型陆续被提出。这些模型加入了更多的预训练数据集，在更有难度的预训练任务上进行预训练，融入了更多的特定领域的知识，能训练更复杂的网络模型等，但是模型通常参数量大，推理时间长，难以部署在手机等边缘设备上。如 GPT-2[8]，GPT-3[9] 和 Switch-C[10] 等模型，其参数量高达 1.6 万亿，模型进行实时推理时，所耗费的时间、空间、金钱等成本较大，往往出现神经网络的“过度参数化”特性[11]，导致对设备的软硬件、内存和计算成本等要求也非常高。

图 1 给出了近年来预训练模型参数不断增加的发展趋势。例如，$BERT_{LARGE}$ 模型[1] 训练时只能访问谷歌的 TPU[13]，而 $BERT_{BASE}$ 模型需要进行一些优化才能进行训练[14]，即 BERT 要想在工业界尤其是资源有限的移动设备手机上部署是不切合实际的。随着深度神经网络在移动设备上的广泛应用，Michel 等[15] 指出，在预训练时大比例地去除 Transformer 里的注意力头数，不会影响模型的性能，因此减小模型尺寸和运行时延迟的需求越来越大。大型的预训练模型被提炼为一种轻量化的小模型是一种常见的压缩办法[16-17]，其试图在不损失精度的情况下减小模型的大小并加快模型的训练、推理速度[18]。

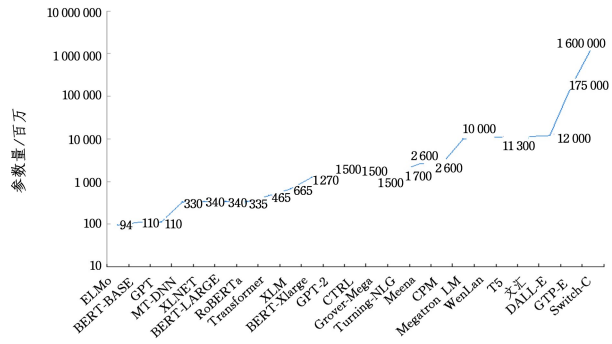

图 1 近年来预训练模型的参数对比

Fig. 1 Parameters comparison of large pre-training models

表 1 列出了轻量化小模型与 $BERT_{BASE}$ 模型的对比结果，主要从压缩和加速模型采取的方法、层数和参数量等方面进行了总结。可以看到，轻量化的小模型能够提高模型的训练效率。

表 1 轻量化预训练模型与 $BERT_{BASE}$ 模型的对比

Table 1 Comparison of lightweight pretraining model with $BERT_{BASE}$ model

模型	方法	层数	参数量	速度	出版
$BERT_{BASE}$[1]	Basiline	12	110×10^6	×1	NAACL-HLT 2019
CompressingBERT[14]	剪枝	12	—	—	ACL 2020
ASH[15]		6	8.34%BERT	×1.175	NeurIPS 2019
Bert-of-theseus[19]	知识蒸馏	6	66×10^6	×1.94	Arxiv 2020
TinyBERT[20]		4	14.5×10^6	×9.4	EMNLP Findings 2020
DistilBERT[21]		6	6.6×10^6	×1.63	NeurIPS 2019
MobileBERT[22]		24	25.3×10^6	×3.73～1.64	ACL 2020
BERT-PKD[16]		3～6	$45.7\sim67\times10^6$	×4	EMNLP Findings 2019
PD[23]		6	7×10^6	×2	arXiv 2019
Extreme[24]		12	$1\sim19\times10^6$	×5.74	ICLR 2020
Q8BERT[25]	量化	12	—	—	NeurIPS 2019
Q-BERT[26]		12			AAAI 2020
FullyQT[27]		12			EMNLP 2020
MiniLM[28]	简化注意力	6/12	$33\times10^6/22\times10^6$	×2.7/×5.3	NeurIPS 2020
ALBERT[29]	结构优化	12	12×10^6	×5.6	ACL 2020
ELECTRA[30]		12	14×10^6	×8	ICLR 2020
DeeBERT[31]		12	—	×4	ACL 2020

2.1 模型压缩

2.1.1 剪枝技术

网络剪枝技术（Pruning）是一种帮助神经网络实现规模更小、效率更高的模型压缩方法，通过减少神经网络中的冗余结构和参数量进行加速模型的训练和推理[32]，主要包括权重大小、注意力头、网络层数、通道、神经元进行剪枝等，从而在性能损失不高的情况下实现模型轻量化、推理速度显著提高、能耗更低、存储模型所需要内存减少的目的，是减少 GPU 使用的可行方法，甚至可以将剪枝后的模型部署在智能手机上进行模型预测。剪枝一般分为非结构化剪枝和结构化剪枝。非结构化剪枝[33] 一般指无序的剪枝。直接裁剪将一些不重要的权重参数设置为 0，这种剪枝方法得到具有稀疏化的权重矩阵，一般只能压缩模型的大小，模型对推理速度的提升较为困难，对移动设备上要求的系统速度提升具有挑战性。结构化剪枝[33] 一般指对神经网络中的层数、通道数等进行裁剪，既可以缩短模型的大小，又可以加快模型的推理速度。大部分剪枝模型都是经过三阶段的流程：预训练、修剪、微调。Liu 等[34] 使用了三阶段的剪枝原理，先训练一个过度参数化的大模型，然后根据准则去除一些冗余的参数，对不重要的结构进行修剪，最后对剪枝后的模型进行微调，以恢复其所损失

的效果。而实验发现从零开始训练剪枝模型可以获得与微调相当甚至更好的性能,即证实了修剪后的框架对于模型的性能有所提高。从预先训练的权重中学习修剪后的模型结构也引发了讨论。基于此,Wang 等[35]提出了一种新的剪枝方法,去除了预训练阶段,首先对随机初始化的权值直接剪枝,然后将剪枝后的结构进行从头训练。实验证明了修剪后的模型结构可以直接从随机初始化的权值学习而不损失性能,也证明了预先训练的权重学习并不能得到一个有效的剪枝网络,从而限制了搜寻更多样化的模型结构。

基于 BERT 模型,研究者们提出了一些剪枝型轻量化 BERT 模型,并在常见的 GLUE[37] 基准上进行测评。例如,McCarley 等[38]提出了基于 BERT 的问题回答模型的结构化剪枝,在保证准确性损失最小情况下修剪前馈层和隐藏层的参数,提升问答系统的推理速度,研究了各种方法对 Transformer 的冗余结构进行修剪,并权衡了每种修剪方法的准确性-速度。Michel 等[39]利用简单的门控启发式,在机器翻译和自然语言推理任务上通过修剪多余的注意力头,几乎不会对性能产生影响,同时较 $BERT_{BASE}$ 模型提高了 17.5% 的推理速度。Gordon 等[40]采取一种更细粒度、更有效的量级权重裁剪方法,即通过将部分权重接近 0 的参数去除,进行模型压缩。实验发现,裁剪程度会影响训练前的损失以及下游任务的迁移能力,低水平的裁剪(30%~40%)不影响精度损失,中等程度和高程度的裁剪对损失和迁移能力影响较大,且推理速度较低水平剪枝有所下降。Frankl 和 Carbin[41]提出了彩票假设的非结构化剪枝方法。通过用原始网络的参数进行初始化,迭代寻找与原始网络测试准确度一样的子网络。

2.1.2 知识蒸馏技术

知识蒸馏(Knowledge Distillation,KD)是一种常见的模型压缩的方法,较早的知识蒸馏模型由 Hinton 等[42]提出并应用于分类任务上,其利用一种基于 teacher-student 框架的训练方法,将复杂度高、学习难度高的大网络模型(teacher)的特征表示迁移到复杂度低的小的模型(student)。其中,从老师模型输出的监督信息被称为知识(Knowledge),而学生从老师模型迁移学习到的监督信息被称为蒸馏(Distillation)。

随着大规模预训练模型的兴起,知识蒸馏在预训练模型中也得到了广泛的应用[43],这些模型均使用了基于 teacher-student 的结构。例如,Sanh 等[44]提出了更小、更快、更轻、更便宜的 DistilBERT 模型,并通过实验表明将 BERT 模型的大小减小到原来的 40% 左右,可以保留 97% 的语言理解能力,并提高 60% 的推理速度。Jiao 等[45]提出了两阶段的 TinyBERT 模型,在预训练和微调时均进行蒸馏,使小的学生网络(TinyBERT)不仅能学到大的教师网络(BERT)通用的知识,还能学到特定任务上的知识。与以往知识蒸馏不同,Sun 等[16]只使用教师网络最后一层的输出进行蒸馏,学生模型耐心地从教师模型的多个中间层学习增量式知识提取,在性能损失不高的情况下,使用小的网络来预估原始大模型的输出分布,以达到更好的精度和性能。以 Sanh 等人提出的 DistilBERT[21]模型为例,使用知识蒸馏的预训练预先训练过的更小的语言模型在下游任务上能够达到和大模型类似的性能,从而使模型在推理上达到更快、更轻的模型预算成本。对于监督学习的分类任务,将大模型神经网络输出的预测标签的不同概率称为软标签,将标签的正确分类称为硬标签。采用交叉熵损失函数最大化正确标签的概率进行大模型的学习,即让大模型预测正确标签的概率尽可能为 1,其他不合理的标签概率接近 0。而对于小的学生模型而言,学习到这种分类效果较为困难。为了让学生模型更好地学到老师模型的输出特征,在小模型神经网络的 SoftMax 引入了带有蒸馏的温度(T)参数:

$$q_i=\frac{\exp(z_i/T)}{\sum_i \exp(z_i/T)} \tag{1}$$

其中,T 控制输出分布的平滑度。T 越大,标签的分类分布越平缓,得到较软化的概率分布,其值介于 0~1 之间。在训练时,希望 T 值越大越好,这样学生模型能够学到越丰富的老师模型知识。在蒸馏模型训练时,学生模型首先对 S,$SoftMax(T=20)$ 的标签概率的输出与老师模型 $SoftMax(T=20)$ 的标签分类输出值进行软标签的损失函数($L1$)计算。在模型推理时,T 设置为 1 时,即小模型的分类结果会学习到类似于大模型的分类效果,恢复为标准的 $SoftMax$,将对 $SoftMax(T=1)$ 的标签分类输出值与原始标签进行硬标签的损失函数($L2$)计算。总的损失函数 L 为软标签($L1$)与硬标签($L2$)所对应的交叉熵的加权平均,软标签的交叉熵的加权系数越大,表明学习到老师模型的知识越多,知识蒸馏的效果越好,蒸馏模型的泛化能力越强。

2.1.3 量化技术

最近,很多基于 Transformer 的预训练模型的参数量超过万亿个,尽管这些模型在某些 NLP 任务上达到了 SOTA,但是仍令人担忧。量化(Quantization)也是解决预训练模型高延时、高消耗的一种有效方法。其本质是使用低位精度进行参数的存储,并使低位的硬件操作来加速推理。早期,量化技术被广泛应用在计算机视觉领域,如用不同的量化方法[46-47]、混合精度量化[48-49]等技术来压缩模型的大小。目前自然语言处理领域中大部分模型都是在 BERT 模型的基础上量化,以将其部署在智能设备上。Devlin[1]提出的 110M 参数的 $BERT_{BASE}$ 和 334M 参数的 $BERT_{LARGE}$ 模型均使用 32 位浮点型表示其参数,这两种 BERT 模型都有很高的内存占用,并且在推断过程中需要大量的计算和宽带。对此,基于自然语言处理任务的预训练量化模型通常将浮点型存储为 32 位的 BERT 模型转化为整型存储一般为 8 位的模型压缩技术,也就是将一个原本用 32 位浮点型表示的权重向量压缩为用 8 位整型来表示这个权重向量,模型内存降低为 1/4 的同时,推理速度也会提高 4 倍左右,这样就可以使得通过大量数据预训练得到的预训练模型轻松部署在一些资源有限、低成本的嵌入式设备上,进而提升模型预测的效率。例如,Shen 等[26]提出的 Q-BERT 对 BERT 模型使用了一种新的量化方法,通过更精确的均值和方差测量方法,以实现更好的混合精度量化,可以在不显著增加硬件复杂性的情况下缓解精度下降问题。Zafrir 等[25]在微调阶段进行量化感知训练,以 4 倍压缩 BERT 的精度最小化损失。相比 32 位的浮点型 BERT 模型,量化的整型 Q8BERT 模型在 8 个不同的 NLP 任务上保持了 99% 的准确性,并加快了模型的推理速度。

2.1.4 权重共享技术

权重共享的本质是模型中的一些权重与模型中的其他参数共享相同的值,使用更少的参数量来解决 BERT 等大型预训练模型带来的内存限制、通信开销及模型退化等问题。

例如，$BERT_{LARGE}$简单地增加隐藏层的大小，反而降低了模型的性能。表2列出了隐藏层大小的改变对RACE数据集[49]性能的影响，实验发现将隐藏层大小增加至2048时会出现模型退化现象[29]。ALBERT[29]使用跨层参数共享，对每个注意力层使用相同的权重矩阵，防止参数随着网络层数增加而增加进而形成参数爆炸的问题。在不严重损失模型性能的条件下，这种减少模型参数量的模型压缩技术减少了内存消耗并提高了模型的运算效率。使用与$BERT_{LARGE}$[1]类似配置的ALBERT可使得模型参数减少至1/18，推理速度提升1.7倍。

表2 增加BERTLARGE隐藏层大小对RACE数据集上性能的影响

Table 2 Influence of increasing hidden size of BERTLARGE on RACE performance

模型	隐藏层大小	参数量	准确率/%
$BERT_{LARGE}$[51]	1024	334×10^6	72
$BERT_{xLARGE}$[51]	2048	1270×10^6	54.3

2.2 简化注意力

在预训练语言模型中，一般来说，自注意力模块的计算和存储复杂度相对较高，简化注意力也是轻量化预训练模型的重要方法。

以Transformer架构中的自注意力为例，简化注意力方法通常有以下几种。

(1)稀疏注意力，即将稀疏偏置引入到注意力计算，通过限制查询或键值对数量来减少注意力机制。OpenAI提出的Sparse Transformers[52]就是通过稀疏注意力来降低计算复杂度。

(2)使用原型和内存压缩，减少查询或键值记忆对的数量，以减小注意力矩阵的大小。例如，MiniLM[28]蒸馏了最后一层Transformer中的自注意力矩阵和Value矩阵，实现的模型效果比TinyBERT模型更好。

(3)带有先验的Attention，使用预先注意力分配来补充标准的自注意力机制。

(4)改进多头机制。例如通过剪枝、权值共享等模型压缩技术对多头注意力机制进行参数压缩。Sukhbaatar等提出的Adaptive Attention[53]方法就采用了这种思想简化注意力，降低矩阵运算量。

2.3 结构优化

模型结构设计对模型效率具有重要影响，优秀的轻量化模型结构可以大大提高模型参数的效率，甚至小模型效果可以超过大模型。模型结构优化主要运用低秩近似、矩阵分解、分组卷积、分解卷积等方法。例如，Lan等[29]提出的ALBERT模型对BERT的Embedding层进行矩阵分解并使用层间参数共享技术，此外还将BERT的下句预测任务修改为语句顺序预测任务，实现了模型轻量化，大大提高了模型效率。Clark等[30]提出了一种类似于生成-判别网络的ELECTRA模型，利用RTD(Replaced Token Detection)来对每个token进行预测，成功减少了模型所需算力。Xin等[31]提出的DeeBERT模型通过动态提前退出BERT，实现模型加速。澜舟科技-创新工场团队与上海交通大学、北京理工大学等单位联合研发的孟子轻量型模型刷新CLUE榜单，获得了第一。此模型基于语言学知识、知识图谱和领域数据增强等技术，从模型架构(包括基础层Embedding表示和交互层Attention机制)到预训练策略进行了全方位改进，使模型适应更多应用场景。

与对模型本身的结构进行优化不同，亚马逊团队提出了一种基于BERT的模型参数选择，使用完全多项式时间近似算法(FPTAS)对BERT进行优化，得到了BERT的最优参数子集——Bort[54]，实现了BERT轻量化。

除对预训练模型结构进行改进外，还有学者提出可以使用高效网络结构进行轻量化模型构建，例如Lite Transformer[55]，SqueezeNet，MobileNet[57]，ShuffleNet[58]以及Xception[59]等。

3 融入知识的预训练语言模型

目前大多数预训练模型都是通过句子的上下文语境来预测被掩盖的词，如BERT[1]，MASS[51]和XLM[61]等，但是这些模型缺少先验知识的背景，有些实体在预测时往往存在误差。例如，《格列佛游记》是英国作家“乔纳森·斯威夫特”创造的一部长篇讽刺小说，在模型不知道“格列佛游记”是一本书，“乔纳森·斯威夫特”是该书的作者时，很难通过上下文语境预测出“乔纳森·斯威夫特”或“格列佛游记”被掩盖的词。作者和作品这种关系被称作“先验知识”，模型知道的先验知识越多，预测出的结果越准确，即模型的泛化能力越强。

目前，一些模型也通过加入实体关系、知识图谱、场景图等显性或隐性地加入先验知识[62,65]，指导模型预测出更精确的结果。例如，Sun等[66]提出了一个利用知识掩盖的ERNIE$_{THU}$模型，ERNIE$_{THU}$没有直接地加入知识嵌入到词向量中，而是隐性地学习有关知识的信息和较长的语义依赖，如实体之间的关系、实体的属性等，引导词嵌入学习，在预训练时掩盖句子中的短语或者实体而非单个词语，这样模型可以含蓄地学到短语和实体之间的对齐关系。Liu等[67]提出的K-BERT模型，考虑到通用语言模型不能很好地解释各个领域的自然语言处理任务，整合了特定领域的知识进行预训练，很容易地将领域知识注入到模型中，不仅在特定任务上显著优于BERT，在公开领域也明显优于BERT。Peters等[68]提出的KnowBERT模型，将先验知识嵌入到大规模模型的通用方法，从而用结构化的知识增强它们的表示。

预训练模型还通过知识框架来改进知识遗忘的问题。例如Wang等[69]提出的K-Adpater包含一种adpater模块，每种预训练任务对应一个adapter模块，该模块蕴含了学习得到的相关的知识，在微调阶段通过采用不同的adapter模块即可引入相关知识。表3整理了融入知识的预训练语言模型，主要从训练形式、知识融入方法方面进行对比。

表3 融入知识的预训练语言模型的对比

Table 3 Comparison of pre-trained language models incorporating knowledge

模型	训练形式	融入方法
K-BERT[61]	融入先验知识	知识图谱
BERT-wwm[70]		修改MASK机制
KnowBERT[68]		修改MASK机制
SpanBert[71]		修改MASK机制
ZEN[72]		修改模型结构
ERNIE$_{THU}$[66]		修改模型结构
K-Adpater[69]	知识适配器	修改模型结构

4 跨模态预训练语言模型

早期,深度学习在计算机视觉领域也取得了一定的成果[73]。自从 Transformer 神经网络在自然语言处理任务上表现出惊人的效果,使用大量未标注的无监督数据进行预训练并在下游任务运用少量特定的标注数据进行监督学习的两阶段法便越来越受关注[74],使得 Transformer 神经网络也被尝试应用在计算机视觉语言多模态模型上。其中,多模态一般包括图像、视频、语音、文档、图表、对话等形式。本文主要对视觉文本预训练模型进行详细介绍,图 2 给出了自 Transformer 被提出以来,主要的多模态预训练模型的发展时间轴。

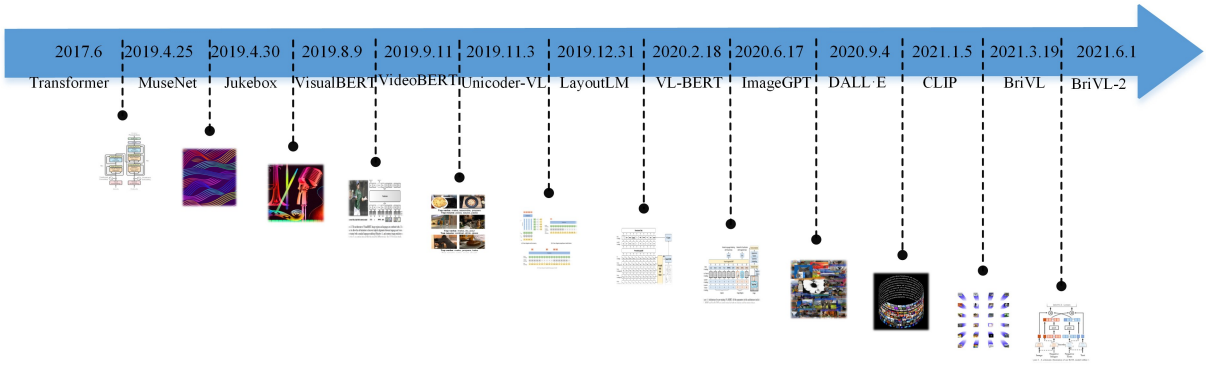

图 2 多模态预训练模型发展时间轴

Fig. 2 Timeline of multi-modal pre-training models development

为了适应多模态的场景学习,一系列多模态的预训练模型在大量的图文对的数据集上预训练,并在具体的下游任务进行微调/精调,进而提升了处理多模态任务的能力[75]。为了学习到视觉文本语言的对齐关系,多学者将自监督的 BERT 预训练模型推广到视频语言的理解与生成任务。Sun 等[76]将 BERT 模型进行扩展并对视频和文本数据进行建模,对于输入的视频数据,以 30 帧为一个视频片段进行采样,首先应用一个预先训练的 S3D 神经网络来提取其特征,提取出的特征向量通过聚类的方法进行离散化,进而在文本标记的基础上增加视觉标记,将其共同的表示标输入到预训练模型中对视频与文本信息进行建模。这种方式将视觉与文本信息进行简单的对齐并不能很好地建模更细粒度的多模态信息,通过聚类提取全局特征时,并不能捕捉到视频中的局部动作信息,视频中的连续动作可以加强对视频语义的理解。Zhu 等[77]从成对的视频和文本描述中提取全局和局部的特征信息作为模型的输入,全局的信号和局部的信号流均与文本信息流进行交互,这样模型不仅学到上下文语义信息,又能更细粒度地学到视觉文本之间的关系。在我国首个超大规模智能模型"悟道 1.0"项目中,Huo 等[78]首次发布了中文的 BriVL 多模态模型,以图文互检为目标,使用 3000 万对通用数据对,基于弱相关假设和对比学习的双塔模型使得多模态数据隐形对齐,其任务是给定一个中文描述(词/句子/诗词),模型自动生成一张弱相关的图像。该模型在 2017 年由创新工场、搜狗等单位联合主办的第一届全球挑战赛发布的中文图片描述数据集 AIC-ICC 中排名第一,与 UNITER[78]模型相比,其速度高达单塔模型的 20 倍。

4.1 对比学习的核心技术

在自然语言处理领域中,BERT,GPT 和 T5 等大规模预训练模型通过自监督学习,利用海量的无标签数据进行预训练,不断刷新了 NLP 的多种任务的 SOTA,证明了使用更多的数据集可以提升模型的性能。而在图像领域中,预训练主要用 ImageNet 对有标注的数据集进行有监督的训练,标注数据的成本较高,因此,学者对图像领域的研究也借鉴 BERT 等 NLP 预训练模型的思想,通过无监督方式训练大量的无标注数据来获得图像的特征知识,并将其迁移到下游任务中。图像领域的无监督学习分为生成式无监督学习和判别式无监督学习两种。相对于生成式,判别式无监督学习较为简单。判别式无监督学习中最为经典的技术是对比学习,其原则是让模型去学习文本的相似性和不相似性,模型可以在相同的向量空间尽可能逼近相似的词,尽可能远离不相似的词。对比学习的种类繁多,本文主要介绍一种基于负例的对比学习中的 SimCLR 模型[80],由于 SimCLR 结构对称,所以较为容易理解,之后的一些对比学习模型也是在此基础上进行改造的。

对比学习期望有一个学习模型,让图像样本映射到向量的空间,使相似的样本距离尽可能逼近,不相似的样本尽可能远离,通过构造正负例让模型真正学到图片不变性的信息。SimCLR 模型通过 InfoNCE 损失函数来防止模型的坍塌,对于一个图像样本 i 的损失函数定义如下:

$$L_i=-\log\left(\exp(S(z_i,z_i^+)/T)/\sum_{(j=0)}^{K}\exp(S(z_i,z_j)/T)\right) \tag{2}$$

其中,$\exp(S(z_i,z_i^+)/T)$为正例样本之间的相似度,$\sum_{(j=0)}^{K}\exp(S(z_i,z_j)/T)$为正例样本与负例样本之间的相似度,$T$ 为温度系数超参,主要使数据均匀分布。从上式中,可以看到正例样本对(z_i,z_i^+)的空间内距离越小,损失值越小,模型效果越好;负样本对(z_i,z_j)空间距离越大,损失值越小,模型效果越好。这种训练方式刚好满足了对比学习的初衷。

2021 年,OpenAI 将预训练语言模型转化为对视觉语言预训练模型的研究,提出了"下一代模型,或许可以针对文本输入,从而编辑和生成图片"的思路。基于此,OpenAI 官网提出了 CLIP 判别模型[81],给定一张图片,模型能够根据图片信息输出文本的描述,此过程相当于在图像上进行零样本学习。其使用了 4 亿个经过清洗的英文图文对,Transformer 对文本进行编码,ResNet 对图像进行编码,使用 SimCLR 对模型

进行训练，计算其样本内距相似度，使得相似的文本和图像表示更逼近，尽可能分开不相似的图像文本表示，将模型效率提升了4～10倍。同样，谷歌提出ALIGN框架[82]对图像文本的对齐表示进行建模，与CLIP不同的是，其数据集包含18亿的图像文本对进行预训练，并在零样本图文转换任务中取得了较好的性能，但效果次于CLIP模型和UNITER模型。

最近，国内悟道成员在BriVL多模态模型基础上共同发布了首个中英文多模态双塔BriVL-2多模态模型，其具有世界上最大的6.5亿通用图文对和53亿参数；对BriVL模型目标检测的网络结构进行了优化，基于DeepSpeed框架提出了首个支持大规模跨模态对比学习的预训练算法[51]；使用网格池化技术对图像特征进行提取，使用了高效的分布式多模态预训练，使用了数据并存、混合精度训练和零冗余优化器(ZeRO)三大技术来减少模型所占的空间；在训练时，分别考虑了ZeRO的第二和第三阶段，在使用MoCo作为学习策略时只使用了ZeRO的阶段二，在训练更大模型时，使用SimRLR的对比学习策略，应用ZeRO的阶段三，极大降低了模型加载所需要的显存。BriVL-2模型在AIC-ICC数据集上刷新了记录。

4.2 视觉语言的预训练任务

视觉和语言作为是人类感知世界并进行交互的两个重要方式，视觉语言相关的技术是人类研究非常重要的方向。视觉语言任务一般划分为理解式的视觉语言任务和生成式的视觉语言任务。其中，生成式的视觉语言一般作为视觉语言的预训练任务，包括图像/视频描述、文本图像生成等视觉语言任务。

对于大量的标注文本预训练数据集，含有高质量的标注图像-文本对的大规模数据集获取较为困难，目前被广泛使用的图像文本对数据集ConceptualCaptions[83]和SBU[84]被称为下游任务的领域外预训练数据集。UNITER模型[79]和Pixel-BERT[85]模型把MS-COCO和Visual-Genome数据集称为下游任务的领域内数据集，大部分模型使用了以上数据集进行预训练，如表4所列。

表4 多模态预训练模型的结构对比

Table 4 Structure comparison of different multi-modal pretraining models

模型	特征提取	输入形式	架构	预训练数据集	是否开源
VisualBERT[88]	R-CNN[89]	图像	单流	In domain	否
ERNIE-VIL[86]	R-CNN	图像	双流	Out of domain+in domain	否
VILBERT[90]	R-CNN	图像	双流	Out of domain	否
VLBERT[91]	R-CNN	图像	单流	Out of domain	是
Unicoder-VL[92]	R-CNN	图像	单流	Out of domain	否
UNITER[79]	R-CNN	图像	单流	Out of domain+In domain	是
PixelBERT[93]	ResNet-50[94]	图像	双流	Out of domain+In domain	否
ImageBERT[95]	R-CNN	图像	单流	Out of domain	否
ViLLA[96]	R-CNN	图像	单流	Out of domain+In domain	是
LXMERT[97]	R-CNN	图像	单流	In domain	是
UNIMO[98]	R-CNN	图像	单流	Out of domain+In domain	是
Oscar[99]	R-CNN	图像	单流	Out of domain+In domain	是
VLP[100]	R-CNN	图像	单流	Out of domain	是
12-in-1[101]	R-CNN	图像	单流	Out of domain+In domain	否
videoBERT[76]	S3D	视频	双流	Out of domain+In domain	否
CBT[102]	S3D	视频	单流	Out of domain	否
LSF[103]	SDResNet50	视频	单流	Out of domain	是
ActBERT[77]	CNN+ResNet-101[94]	视频	单流	HowTo100M	否

对不同的视觉语言任务设计不同的预训练网络结构，通常都是有效地将视觉与文本信息进行融合，但多模态信息语义的对齐却具有挑战性。大多数预训练模型忽略了构建视觉和语言的详细语义对齐的重要性，因此训练的模型不能很好地表示一些真实场景所需的细粒度语义。为了使得视觉文本获得更细粒度的语义信息，ERNIE-VIL模型[86]受ERNIE1.0模型掩盖文本实体的预训练任务的启发[87]，引入了结构化的场景图，分别从对象、对象的关系、对象的属性构建三元组作为图像与文本的先验知识，更好地指导图像文本细粒度的语义对齐。

4.3 视觉语言的下游任务

理解式的视觉任务主要作为下游任务评估模型，以理解视觉和语言两种模态的输入信号，并建立视觉和语言之间的关联为目标，包括视觉问答任务(Visual Question Answering，VQA)[104]、视觉常识推理(Visual Commonsense Reasoning，VCR)[105]、引用表达式理解任务(Referring Expression Comprehension)和跨模态检索(Cross-modal Retrieval)等任务。其中，VQA任务[86]旨在根据指定的图片输出自然语言问题的答案，常见的数据集VQA2.0包含204k图像和关于这些图像的110万个问题[106]，强调图片文本和图片对齐的关系。VCR任务[107]旨在评估多模态模型常识的推理能力，这个整体任务(Q→AR)被分解为视觉问答任务(Q→A)和答案论证(QA→R)两个子任务。跨模态检索任务也是多模态领域经典的视觉语言任务，在生活中利用较为广泛，旨在衡量图像与文本的匹配关系。

例如，在搜索引擎中输入自然语言的描述搜索最相关的图片。图文跨模态检索任务主要包括图片检索文本和文本检索图片，常见的数据集是Flickr30k[108]。引用表达式理解任务主要根据给定的图片描述选择最符合的图片区域，对图像中的对象进行定位，经常采用RefCOCO+数据集[109]进行评估。表5对比了不同的视觉语言模型在下游任务的表现。

表 5　视觉文本预训练模型在下游任务上的表现

Table 5　Performance of vision-text pre-training models in downstream tasks

(单位:%)

数据集	衡量标准	Unicoder-VL$_{BASE}$	ViLBERT$_{BASE}$	UNITER$_{BASE}$	VLBERT$_{BASE}$	ERNIE-VIL$_{LARGE}$
VQA2.0[106]	*Accuracy*	—	70.55	71.56	71.79	73.78
VCR[107]	(Q→A)	72.6	72.42	—	75.5	78.52
	(QA→R)	74.5	74.47		77.9	83.37
	(Q→AR)	54.5	54.04		58.9	65.81
IRFlickr30k[108]	*Recall*@1	71.50	58.20	—	—	75.1
	Recall@5	90.90	84.90			93.42
	Recall@10	94.90	91.52			96.26
RefCOCO+[109]	*Accuracy*	—	72.34	72.78	72.59	74.24

5　跨语言预训练语言模型

自然语言处理任务上的预训练一般使用特定的语言训练,并不能直接应用于其他语言,这对于机器理解标注语料稀少的小语种(如维吾尔语、哈萨克语等)是一个巨大的挑战。因此构建一个统一的语言模型去理解多种语言,是近年来自然语言处理领域研究的热点。跨语言的众多研究将预训练模型扩展到跨语言任务中,以提升下游任务的性能。跨语言预训练模型涵盖了大约 100 种语言,其中包括 XLM-R 模型[61]、RemBERT 模型[110]和 InfoXLM 模型[111],其他模型见表 6 跨语言权威的 Xtreme 榜单[112]。

表 6　跨语言模型在 Xtreme 榜单上的性能对比

Table 6　Performance comparsion of Cross-language models in Xtreme

(单位:%)

模型	参与单位	分数	文本分类	结构化预测	阅读理解	语义检索
—	Human	93.3	95.1	97.0	87.8	—
ERNIE-M[113]	百度	80.9	—	—	—	87.9
ULRv2[114]+StableTune	微软图灵团队 微软亚洲研究院	80.7	88.8	75.4	72.9	89.3
VECO[115]	DAMO NLP Team 阿里巴巴	77.2	87.0	70.4	68.0	88.1
FILTER[116]	微软	77.0	87.5	71.9	68.5	84.4
X-STILTs[117]	纽约大学	73.5	83.9	69.4	67.2	76.5
XLM$_{LARGE}$[117]	XTREME Team Alphabet CMU	68.2	82.8	69.0	62.3	61.6
mBERT[118]	XTREME Team Alphabet CMU	59.6	73.7	66.3	53.8	47.7
RemBERT[110]	谷歌研究院 DeepMind	56.1	84.1	73.3	68.6	—
XLM[66]	XTREME Team Alphabet CMU	55.8	75.0	65.6	43.9	44.7

表 6 对比了跨语言模型在涵盖了 12 个语系 40 种语言数据集上的测评结果,包括文本分类、结构化预测、语义检索和阅读理解这 4 类自然语言处理任务的 9 个数据集。其中跨语言的预训练模型主要分为两类:判别式模型和生成式模型。对于判别式语言模型,类似于自然语言理解任务,旨在提升模型推理判断两个句子之间关系的跨语言模型的性能。

最近,预先训练的跨语言模型 mBERT[1]和 XLM[66]通过跨语言的 Masked 语言模型(MLM)联合对多种语言上的大型 Transformer 模型[119]进行预训练,可以很好地提升自然语言理解性能。Lample 和 Conneau[119]提出翻译语言模型 TLM,从平行语料库中对齐的跨语言数据输入到模型中,并在自然语言推理数据集标准上达到 SOTA,进一步展示了预先训练的跨语言模型可以提升机器翻译和自然语言生成任务;随后又提出了一系列跨语言预训练模型,但这些预训练均在维基百科数据集上进行预训练,但维基百科中的低资源语言数据相对有限。实验证明,增大模型或预训练数据集可以显著提升性能[120]。Conneau 等[61]提出的 XLM-R 模型将语料库中的低资源语言的数据量平均增加了两个数量级,使用 2.5T 爬虫数据集[121],其中含有 100 种语言,并首次对数据集高低资源数据的分布进行了权衡,以用最适度的算力提升自然语言处理任务的性能。

考虑到语言的全球化,大多以 BERT 为基础的跨语言预训练模型均在英文上进行训练。Huang 等[122]将跨语言、多模型和多任务融合在一起,提出了跨语言多模型的 M3P 模型,利用跨语言的迁移能力,让模型在一种语言上训练,把其学到的知识迁移到其他语言上,提高在其他语言上的性能。大量特定的 BERT 模型已经被训练用于英语以外的语言上[123]。例如,Antoun 等[124]对 BERT 模型在大规模的阿拉伯语数据上进行预训练,达到了和 BERT 在英文数据上训练类似的结果,该模型在测试的大多数阿拉伯语自然语言处理任务中取得了最先进的性能。Wilie 等[125]对低资源的印度尼西亚语进行自然语言理解任务的预训练,并首次提出了一个用于印度尼西亚语自然语言理解(IndoNLU)任务的训练、评估和基准测试的庞大资源。生成式语言模型主要是为提升生成式任务而提出的一系列跨语言模型。Song 等[126]提出的 MASS 模型,基于隐蔽的序列到序列预训练的编码器-解码器语言生成方法,编码器输入连续被[MASK]标记代替的片段,解码器使用自回归的方法预测出被掩盖的片段。与 MASS 模型类似,mBART[127]和 XNLG[128]使用基于序列到序列的编码器-解码器架构。为了打破平行语料库大小对模型性能的限制,尤其是低资源语言,Ouyang 等[113]提出的 ERNIE-M 模型将跨语言的表示与单语语料库对齐,在 96 种语言的单语语料库上生成伪平行句对,以学习不同语言之间的语义对齐,增强了跨语言模型的语义建模。

6　挑战与展望

2017 年 Transformer 神经网络的提出,推动了自然语言处理领域的技术发展,开启了一个全新的预训练时代。预训练模型的发展也推动了其拓展模型的发展,轻量化、融入

知识、多模态、跨语言等模型都取得了阶段性的成果，但由于目前相关技术仍处于探索阶段，因此仍面临一些挑战。下面对未来可能的研究方向进行展望。

(1)对轻量化预训练模型的研究

自2018年以来，大规模的预训练语言模型相继被提出，参数量达数百万甚至到万亿个，往往出现“过度参数化”的现象，从而导致高内存占用和高计算成本。对于工业化的需求，这些模型很难部署到移动端，虽然在设备上实时操作这些模型有潜力支持新奇和有趣的语言处理应用程序，但这些模型不断增长的计算和内存需求可能会阻碍其广泛使用。

目前已有一些预训练模型采用多种方法融合方式进行预训练。例如，Sun等[129]提出的MobileBERT模型，将注意力机制改进后的模型进行量化和蒸馏。实验证明，虽然量化将MobileBERT模型压缩为了1/4，但模型的性能几乎没有下降。因此模型压缩、简化注意力和结构优化的深度结合，是今后模型轻量化值得研究的方向。

(2)对融入知识的预训练语言模型研究

由于预训练模型缺乏一定的常识性，可以将知识图谱、场景图等世界知识融入到预训练模型中。研究表明[130]，引入知识的确可以提升自然语言处理任务的性能。目前，知识图谱的构建准确度并不高，缺少形式化知识体系系统，目前的知识系统中往往缺少动作、状态、逻辑关系描述等。早期通过完全监督的人类手工标注三元组构建知识图谱(如Freebase[131]和WordNet[132]知识图谱)。随着深度学习的发展，NELL系统[133]通过半监督的方法，利用机器学习实现知识图谱的自动构建，但错误率增加10倍，无监督学习的知识图谱将成为今后的研究热点。但深层结构化的知识图谱的研究目前仍处于探索阶段，引入知识图谱等外部知识对模型仅仅起到辅助作用，并不能让模型有真正的常识。如何使模型真正具有常识，也是人工智能未来的研究方向。

(3)对跨模态预训练模型的研究

跨模态的预训练模型还有巨大探索空间。首先，构建高质量的多模态跨语言的数据集对预训练也是十分关键的，将直接影响模型的性能和泛化能力。其次，对图、文、视频数据相结合的预处理、多模态预训练任务的设计和多模态信息的融合等还需要进一步探讨；将大数据与富知识融合，并通过与图像等跨模态信息进行交互，显著提升以自然语言为核心的中文语义理解能力，突破多模态数据融合的难题，从“形合意迷”到“合意合”，实现看图说话，生成可控文本等复杂的多模态任务。

(4)对跨语言预训练模型的研究

目前预训练语言模型大多都是在英文语料库进行预训练的(如英文的维基百科等)，在其他语言上进行微调，进而解决引文的自然语言问题[134]。而引入中文或其他低资源语言的预训练模型甚少。为了促进语言的多样性和全球化，我国智源研究院和清华大学单位联合提出的CPM模型[135]，是我国首个以中文为核心的大规模的预训练模型，但中文的预训练模型仍处于初级探索阶段，探索更具通用能力的深度语言理解的预训练模型是值得关注的。

结束语 Transformer神经网络和BERT模型的提出掀起了预训练模型的浪潮。本文主要对预训练模型的拓展模型即轻量化预训练模型、融入知识的预训练语言模型、跨模态与训练语言模型、跨语言预训练模型4种模型的研究现状及其技术应用进行了梳理，最后结合4种预训练模型提出了4种未来可能的研究方向，为学习预训练模型的初学者提供一些指导性帮助。

参 考 文 献

[1] DEVLIN J, CHANG MW, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of Conference on Computational Linguistics: Human Language Technologies. 2019: 4171-4186.

[2] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013: 3111-3119.

[3] PENNINGTON J, SOCHER R, MANNING CD. GloVe: Global vectors forword representation[C]// Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[4] MCCANN B, BRADBURY J, XIONG C M, et al. Learned in translation: Contextualized word vectors[C]// Proc. of the 31st International Conference on Neural Information Processing Systems. 2017: 6297-6308.

[5] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations [C] // Proc. of Conference on the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018: 2227-2237

[6] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. [2021-07-03]. https://openai. com/blog/language-unsupervised/

[7] BAEVSKI A, EDUNOV S, LIU Y H, et al. Cloze-driven pretraining of self-attention networks[C]// Proc. of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[8] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[EB/OL]. [2021-07-03]. https://www. cs. ubc. ca/~ amuham01/LING530/papers/radford2018 improving. pdf.

[9] BROWN TB, MANN B, RYDER N, et al. Language models are few-shot learners [C] // Advances in Neural Information Processing Systems 33(NeurIPS2020). 2020: 1877-1901.

[10] FEDUS W, ZOPH B, SHAZEER N. Switch Transformers: Scaling to Trillion Parameter Models with Simple an deficient Sparsity [J]. arXiv: 2101. 03961, 2021.

[11] BA J, CARUANA R. Do deep nets really need to be deep? [C]// Proc. of the 27th international Conference on Neural Information Processing Systems. 2014: 2654-2662.

[12] DENTON M L, ZAREMBA W, BRUNA J, et al. Exploiting linear structure within convolutional networks for efficient evaluation[C]// Proc. of the 27th International Conference on Neural Information Processing Systems. 2014: 1269-1277.

[13] GORDON M A, DUH K, ANDREWS N. Compressing Bert: Studying the effffects of weight pruning on transfer learning [J]. arXiv: 2002. 08307, 2020.

[14] SOHONI N S, ABERGER C R, LESZCZYNSKI M, et al. Low-

memory neural network training: A technical report[C]// Proc. of Conference on annual event of the European Federation of Corrosion. 2019.

[15] MICHEL P, LEVY O, NEUBIG G. Are sixteen heads really better than one? [C]// Proc. of Thirty-third Conference on Neural Information Processing Systems. 2019: 14014-14024.

[16] SUN S Q, CHENG Y, GAN Z, et al. Patient knowledge distillation for BERT model compression[C]// Proc. of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 4323-4332

[17] WANG N Y, YE Y X, LIU L, et al. Language models based on deep learning: A review[J]. Ruan Jian Xue Bao/Journal of Software, 2021, 32(4): 1082-1115

[18] BUCILA C, CARUANA R, NICULESCU-MIZIL A. Model compression[C]// Proc. of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006: 535-541.

[19] XU C W, ZHOU W, G E T, et al. Bert-of-theseus: Compressing bert by progressive module replacing [J]. arXiv: 2002. 02925, 2020.

[20] JIAO X Q, YIN Y C. Tiny BERT: Distilling BERT for natural language understanding[C]// Findings of the Association for Computational Linguistics: EMNLP. 2020: 4163-4174.

[21] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaperandlighter [J]. arXiv: 1910. 01108, 2019.

[22] SUN Z Q, YU H K. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices[C]// Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2158-2170.

[23] TURC I, CHANG M W, LEE K, et al. Well-read students learn better: The impact of student initialization on knowledge distillation [J]. arXiv: 1908. 08962, 2019.

[24] ZHAO S Q, GUPTA R. Extreme Language Model Compression with Optimal Subwords and Shared Projections [J]. arXiv: 1909. 11687, 2019.

[25] ZAFRIR O, BOUDOUKH G, IZSA K, et al. Q8bert: Quantized 8 bit bert[C]// Proc. of Thirty-third Conference on Neural Information Processing Systems. 2019.

[26] SHEN S, DONG Z, YE J Y, et al. Q-bert: Hessian based ultra low precision quantization of bert[C]// Proc. of AAAI. 2020: 8815-8821.

[27] PRATO G, CHARLAIX E, REZAGHOLIZADEH M. Fully quantized transformer for machine translation[C]// Proc. of the Conference on Empirical Methods in Natural Language Processing: Findings. 2020: 1-14.

[28] WANG W H, WEI F R, DONG L, et al. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers [J]. Advances in Neural Information Processing System, 2020, 33: 5776-5788.

[29] LAN Z Z, CHEN M D. Albert: Alite bert for self-supervised learning of language representations [J]. arXiv: 1909. 11942, 2019.

[30] CLARK K, LUONG M T, LE Q V, et al. ELECTRA: Pre-training text encoders as discriminators rather than generators[C]// Proc. of the Int'l Conference on Learning Representations. 2019.

[31] XIN J, TANG R, LEE J, et al. DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference[C]// Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2246-2251.

[32] HE Y, ZHANG X, SUN J. Channel pruning for accelerating very deep neural networks[C]// Proc. of the IEEE International Conference on Computer Vision. Venice, 2017: 1389-1397.

[33] HAN S, MAO H, DALLY W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding [C]// Proc. of International Conference on Learning Representations. 2016.

[34] LUO J H, WU J, LIN W. Thinet: A filter level pruning method for deep neural network compression[C]// Proc. of the IEEE International Conference on Computer Vision. 2017: 5058-5066.

[35] LIU Z, SUN M, ZHOU T, et al. Rethinking the value of network pruning[C]// Proc. of International Conference on Learning Representations. 2019.

[36] WANG YL, ZHANG XL, XIE LX, et al. Pruning from Scratch [J]. arXiv: 1909. 12579v1, 2019.

[37] 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.

[38] MCCARLEY J S, CHAKRAVARTI R, SIL A. Structured Pruning of a BERT-based Question Answering Model [J]. arXiv: 1910. 06360v2, 2020.

[39] MICHEL P, LEVY O, NEUBIG G. Are sixteen heads really better than one? [C]// Proc. of Thirty-third Conference on Neural Information Processing Systems. 2019: 14014-14024.

[40] GORDON M, DUH K, ANDREWS N. CompressingBERT: Studying the Effffects of Weight Pruning on Transfer Learning [J]. arXiv: 2002. 08307, 2020.

[41] FRANKLE J, CARBIN M. Thelottery ticket hypothesis: Finding sparse, trainable neural networks[C]// Proc. of the Seventh International Conference on Learning Representations. 2019.

[42] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[C]// Proc. of Deep Learning Workshop on NIPS. 2014.

[43] WANG W, WEI F, DONG L, et al. Minilm: Deepself-attention distillation for task-agnostic compression of pre-trained transformers [J]. arXiv: 2002. 10957, 2020.

[44] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaperandlighter [J]. arXiv: 1910. 01108, 2019.

[45] JIAO X Q, YIN Y C. Tiny BERT: Distilling BERT for natural language understanding[C]// Findings of the Association for Computational Linguistics: EMNLP. 2020: 4163-4174.

[46] KRISHNAMOORTHI R. Quantizing deep convolutional networks for efficient inference: Awhitepaper [J]. arXiv: 1806. 08342, 2018.

[47] ZHANG D, YANG J, YE D, et al. LQ-nets: Learned quantization for highly accurateand compact deep neural networks[C]// Proc. of the 15th European Conference on Computer Vision. 2018: 365-382.

[48] DONG Z, YAO Z, GHOLAMI A M. et al. HAWQ: Hessian

AWare Quantization of Neural Networks with Mixed-Precision [C]// Proc. of International Conference on Computer Vision(ICCV). 2019.

[49] WU B, WANG Y, ZHANG P, et al. Mixed Precision Quantization of ConvNets via Diffffferentiable Neural Architecture Search [C]// Proc. of ICLR. 2019.

[50] LAI G K, XIE Q Z, LIU H X, et al. Race: Large-scale reading comprehension dataset from examinations[C]// Proc. of Empirical Methods in Natural Language Processing. 2017: 785-794.

[51] SHOEYBI M, PATWARY M, PURI R. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism[J]. arXiv: 1909. 08053, 2019.

[52] CHILD R, GRAY S, RADFORD A, et al. Generating Long Sequences with Sparse Transformers[J]. arXiv: 190410509, 2019.

[53] SUKHBAATAR S, GRAVE E, BOJANOWSKI P, et al. Adaptive Attention Span in Transformers[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[54] WYNTER A D, PERRY D J. Optimal Subarchitecture Extraction For BERT[J]. arXiv: 2010. 10499, 2020.

[55] WU Z, LIU Z, LIN J, et al. Lite Transformer with Long-Short Range Attention[J]. arXiv: 2004. 11886, 2020.

[56] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0. 5 mb model size[J]. arXiv: 1602. 07360, 2016.

[57] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [J]. arXiv: 1704. 04816, 2017.

[58] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2018.

[59] CHOLLET F. Xception: Deep Learning with Depthwise Separable Convolutions[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2017.

[60] HE P C, LIU X D, GAO J F. Deberta: decoding-enhanced bert with disentangled attention [J]. arXiv: 2006. 03654, 2020.

[61] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[C]// Proc. of the 58th Annual Meeting of the Association for Computational Linguistic. 2020: 8440-8451.

[62] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]// Proc. of the 26th International Conference on Neural Information Processing Systems. 2013: 2787-2795.

[63] XIN J, ZHU H, HAN X, et al. Putitback: Entity typing with language model enhancement[C]// Proc. of Conference on Empirical Methodsin Natural Language Processing. 2018: 993-998.

[64] YAGHOOBZADEH Y, SCHÜTZE H. Multilevelre presentations for fifine-grained typing of knowledge base entities[C]// Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017: 578-589.

[65] YAMADA I, SHINDO H, TAKEDA H, et al. Joint learning of the embedding of words and entities for named entity disambiguation[C]// Proc. of the 20th SIGNLL Conference on Computational Natural Language Learning(CoNLL). 2016: 250-259.

[66] SUN Y, WANG S H, LI Y K, et al. ERNIE2. 0: A continual pre-training framework for language understanding[C]// Proc. of AAAI. 2019.

[67] LIU WJ, ZHOU P, ZHAO Z, et al. K-BERT: Enabling language representation with knowledge graph[C]// Proc. of AAAI. 2019.

[68] PETERS ME, NEUMANN M, LOGAN IVRL, et al. Knowledge enhanced contextual word representations[C]// Proc. of Conference on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conference on Natural Language Processing (EMNLPIJCNLP). 2019: 43-54.

[69] WANG R, TANG D, DUAN N, et al. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters[J]. arXiv: 2002. 01808, 2020.

[70] CUI Y, CHE W, LIU T, et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. arXiv: 1906. 08101, 2019.

[71] JOSHI M, CHEN D, LIU Y, et al. SpanBERT: Improving Pre-training by Representing and Predicting Spans[J]. arXiv: 1907. 10529, 2019.

[72] DIAO S, BAI J, SONG Y, et al. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations[C]// Findings of the Association for Computational Linguistics(EMNLP 2020). 2020.

[73] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classifification with deep convolutional neural networks[C]// Proc. of the Neural Information Processing Systems. 2012: 1106-1114.

[74] CARON M, BOJANOWSKI P, JOULIN A, et al. Deep clustering forun supervised learning of visual features[C]// Proc. of Computer Visio-ECCV. 2018: 139-156.

[75] HAN K, XIAO A, WU E H, et al. Transformer in Transformer [J]. Advances in Neural Information Processing System, 2021, 34: 15908-15919.

[76] SUN C, MYERS A, VONDRICK C, et al. VideoBERT: A joint model for video and language representation learning[C]// Proc. of the IEEE Int'l Conference on Computer Vision. 2019: 7464-7473.

[77] ZHU L C, YANG Y. ActBERT: Learning Global-Local Video-Text Representations[C]// Proc. of CVPR. 2020: 8746-8755.

[78] HUO Y Q, ZHANG M L, LIU G Z. WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training [J]. arXiv: 2103. 06561v5, 2021.

[79] CHEN Y C, LI L, YU L, et al. Uniter: Learning universal image-text representations [J]. arXiv: 1909. 11740, 2019.

[80] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representation [C]// Proc. of the 37th International Conference on Machine Learning(ICML 2020). 2020: 1575-1585.

[81] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[82] JIA C, YANG YF, XIA Y, et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision [C]// International Conference on Machine Learning. PMLR, 2021: 4904-4916.

[83] SHARMA P,DING N,GOODMANS,et al. Conceptual captions: A cleaned, hypernymed, imagealt-textdataset for automatic image captioning[C]// Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. 2018:2556-2565.

[84] ORDONEZ V,KULKARNI G,BERG T L. Im2text: Describing images using 1 million captioned photographs[C] // Proc. of Neural Information Processing Systems. 2011:1143-1151.

[85] HUANG Z H,ZENG Z Y,LIU B,et al. Pixel-BERT: Aligning Image Pixel swith Text by Deep Multi-Modal Transformers [J]. arXiv:2004.00849,2020.

[86] YU F,TANG J J,YIN W C,et al. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph [J]. arXiv:2006.16934,2020.

[87] VINCENT P,LAROCHELLE H,BENGIO Y O,et al. Extracting and composing robust features withde noising autoencoders [C]// Proc. of the 25th International Conference on Machine Learning. 2008:1096-1103.

[88] LI L H,ATSKAR M Y,YIN D,et al. Visualbert: A simple and performant baseline for vision and language [J]. arXiv: 1908.03557,2019.

[89] REN S Q,HE K M,GIRSHICK R,et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems. 2015:91-99.

[90] LU J,BATRA D,PARIKH D,et al. VilBERT: Pretraining task-agnostic visio linguistic representations for vision-and-language tasks[C] // Advances in Neural Information Processing Systems. 2019:13-23.

[91] SU W,ZHU X,CAO Y,et al. Vl-BERT: Pre-training of generic visual-linguistic representations[C]// Proc. of the 8th Int'l Conference on Learning Representations. 2020.

[92] LI G,DUAN N,FANG Y,et al. Unicoder-vl: Auniversal encoder for vision and language by cross-modal pre-training[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:11336-11344.

[93] HUANG Z H,ZENG Z Y,LIU B,et al. Pixel-BERT: Aligning Image Pixel swith Text by Deep Multi-Modal Transformers [J]. arXiv:2004.00849,2020.

[94] HE K M,ZHANG X Y,REN S Q,et al. Deep Residual Learning for Image Recognition[C] // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

[95] QI D,SU L,SONG J,et al. Image BERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data [J]. arXiv:2001.07966,2020.

[96] GAN Z,CHEN Y C,LI L J,et al. Large-scale adversarial training forvision-and-language representation learning[J]. Advances in Neural Information Processing Systems,2020,33:6616-6628.

[97] TAN H,BANSAL M. Lxmert: Learning cross-modality encoder representations fromtrans formers[C]// Proc. of Conference on Empirical Methods in Natural Language Processing. 2019.

[98] LI W,GAO C,NIU G C,et al. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning [J]. arXiv:2012.15409v1,2020.

[99] LI XJ,YIN X,LI CY,et al. Oscar: Object semantics aligned pre-training for vision-language tasks[C]// Proc. of the European Conference on Computer Vision. 2020:121-137.

[100] ZHOU L W,PALANGI H,ZHANG L,et al. Unifified vision language pre-training for image captioning and VQA[C]// Proc. of the SemEval workshop at ACL. 2017:13041-13049.

[101] LU JS,GOSWAMI VE,ROHRBACH M,et al. 12-in-1: Multi-task vision and languagere presentation learning[C]// Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[102] SUN C,BARADE LF,MURPHY K,et al. Contrastive bidirectional transformer for temporal representation learning [J]. arXiv:1906.05743,2019.

[103] LI TH,LI M. Learning spatiotemporal features via video and text paird is crimination [J]. arXiv:2001.05691,2020.

[104] ANTOL S,AGRAWAL A,LU J,et al. Vqa: Visual question answering[C] // Proc. of the IEEE international Conference on computer vision. 2015:2425-2433.

[105] ZELLERS R,BISK Y O,FARHADI A,et al. From recognition to cognition: Visual commonsense reasoning[C] // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:6720-6731.

[106] GOYAL Y,KHOT T,SUMMERS-STAY D,et al. Making the v in vqa matter: Elevating the role of image understanding invisual question answering[C]// Proc. of Computer Vision and Pattern Recognition(CVPR). 2017.

[107] ZELLERS R,BISK Y O,FARHADI A,et al. From recognition to cognition: Visual commonsense reasoning[C] // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:6720-6731.

[108] PLUMMER B A,WANG L W,CERVANTES C M,et al. Flickr30k entities: Collecting region-to-phrase correspondences for riche image-to-sentence models[C]// Proc. of ICCV. 2015.

[109] KAZEMZADEH S,ORDONEZ V,MATTEN M,et al. Refer it game: Referring to objects in photographs of natural scenes [C]// Proc. of Conference on Empirical Methods in Natural Language Processing(EMNLP). 2014:787-798.

[110] CHUNG H W,FEVRY T,TSAI H,et al. Rethinking embedding couplingin pre-trained language models[C]// Proc. of ICLR 2021. 2021.

[111] CHI Z W,DONG L,WEI F R,et al. INFOXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training [J]. arXiv:2007.07834v1,2020.

[112] ZHAO S Q,GUPTA R. Extreme Language Model Compression with Optimal Subwords and Shared Projections [J]. arXiv: 1909.11687,2019.

[113] OUYANG X,WANG S H,PANG C,et al. ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora [J]. arXiv:2012.15674,2021.

[114] TIWARY S,ZHOU M. T-ULRv2[EB/OL]. 2020. https://www.microsoft. com/en-us/research/blog/microsoft-turing-universal language-representation-model-t-ulrv2-tops-xtreme-leader-board/?lang=frca.

[115] DAI Z,YANG Z,YANG Y,et al. Transformer-XL: Attentive language models beyond a fixed-length context[C]// Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:2978-2988.

[116] FANG Y W,WANG S H,GAN Z,et al. FILTER: An Enhanced Fusion Method for Cross-lingual Language Understanding[C]//

Proc. of Association for the Advancement of Artifificial Intelligence. 2020.

[117]PHANG J, HTUT P M, PRUKSACHATKUN Y, et al. English Intermediate-Task Training Improves Zero-Shot Cross-Lingual Transfer Too [J]. arXiv: 2005. 13013v1, 2020.

[118]PIRES T, SCHLINGER E, GARRETTE D, et al. How multilingual is Multilingual BERT? [J]. arXiv: 1906. 01502v1, 2019.

[119]VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need[C]// Proc. of the 31st Conference on Neural Information Processing Systems. 2017: 5998-6008.

[120]DELOBELLE P, WINTERS T, BERENDT B. RobBERT: a Dutch RoBERTa-based language model [J]. arXiv: 2001. 06286, 2020.

[121]WENZEK G, LACHAUX M E, CONNEAU A, et al. Ccnet: Extracting high quality monolingual datasets from web crawldata [J]. arXiv: 1911. 00359, 2019.

[122]HUANG H Y, SU L, QI D. M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pretraining [J]. arXiv: 006. 02635v1, 2020.

[123]RUDER, SEBASTIAN. ML and NLP Research High lights of 2020 [EB/OL]. http://ruder. io/research-highlights-2020, 2021.

[124]ANTOUN W, BALY F, HAJJ H. AraBERT: Transformer-based Model for Arabic Language Understanding[C]// Proc. of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offffensive Language Detection. 2020: 9-15.

[125]WILIE B, VINCENTIO K, WINATA G I. IndoNLU: Benchmark and Resources for Evaluating Indonesian Natura lLanguage Understanding[C]// Proc. of the 1st Conference for the Asia-Pacifific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. 2020: 843-857.

[126]SONG K, TAN X, QIN T, et al. MASS: Masked sequence to sequence pre-training for language generation[C]// Proc. of the Int'l Conference on Machine Learning. 2019: 5926-5936.

[127]LIU Y H, GU JT, GOYAL N, et al. Multilingual denoising pre-training for neural machine translation [J]. arXiv: 2001. 08210, 2020.

[128]CHI Z W, DONG L, WEI F R. Cross-lingual natural language generation via pre-training[C]// Proc. of AAAI. 2020: 7570-7577.

[129]SUN Z Q, YU H K. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices[C]// Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2158-2170.

[130]LIU Z Y, SUN M S, LIN Y K, et al. Knowledge representation learning: A review[J]. Journal of Computer Research and Development, 2016, 53(2): 247-261.

[131]BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: acollaboratively created graph database for structuring human knowledge[C]// Proc. of the ACM SIGMOD International Conference on Management of Data. 2008: 1247-1250.

[132]MILLERG A. WordNet: alexical database for English[J]. Communications of the ACM. 1995, 38: 483.

[133]MITCHELL T, COHEN W, HRUSCHKA E, et al. Never Ending Language Learning[C]// Proc. of the Conference on Artifificial Intelligence. 20151: 103-115.

[134]LIANG X B, REN F L, LIU Y K, et al. N-reader: Machine reading comprehension based on double layers of self-attention[J]. Journal of Chinese Information Processing, 2018, 32(10): 130-137

[135]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.

Abudukelimu ABULIZI, born in 1983, Ph. D, lecturer, is a member of China Computer Federation. His main research interests include cognitive neuroscience, artificial intelligence and big data mining.

Abudukelimu HALIDANMU, 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.