Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Review

预训练语言模型及其应用

王海峰[a,*], 李纪为[b], Hua Wu[a], Eduard Hovy[c], Yu Sun[a]

[a] *Baidu Inc., Beijing 100193, China*
[b] *College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China*
[c] *Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

ARTICLE INFO

Article history:
Received 10 November 2021
Revised 8 March 2022
Accepted 5 April 2022
Available online 7 September 2022

关键词
预训练模型
自然语言处理

摘要

预训练语言模型（pre-trained language model, PTLM）在自然语言处理（natural language processing, NLP）领域取得了令人瞩目的成功，并由此引发了下游任务从监督学习到预训练-微调范式的转变。在此之后，一系列预训练模型的创新研究涌现出来。本文系统、全面地回顾了自然语言处理的代表性工作和最新进展，并按照类别系统性地介绍了自然语言处理领域的预训练模型。文中首先简要介绍了预训练模型以及不同的模型特点和框架；之后，介绍并分析了预训练模型的影响和挑战以及下游任务中的应用；最后，简要总结并阐述了预训练模型未来的研究方向。

1. 预训练模型的简要发展历程

预训练模型的概念和迁移学习（transfer learning）相关[1]。迁移学习的核心是将从一个或多个任务学习到的知识应用到新的任务中。传统的迁移学习采用有标注的数据进行模型的监督训练。这种方式在至少十年内被认为是一种实现迁移学习的常见做法。在深度学习领域中，通过在大量无标注数据中进行自我监督学习已经成为了目前迁移学习的主要实现方法。与传统方法不同的是，预训练可以通过无标注数据进行自我监督学习，并且通过微调或者少样本学习的方法将模型应用到下游任务。

在自然语言处理（NLP）领域中，语言模型通常是预训练的目标任务。语言模型的目标是根据给定的前面位置的词信息，预测下一个词[2–4]。神经语言模型的第一个里程杯[5]，是通过词的向量表示和前馈神经网络来建模连续n个词（n-gram）的概率。在此之后，深度学习主导了语言模型的发展。在早期的神经语言模型中，循环神经网络（recurrent neural network, RNN）得到广泛的采用[6–7]。在循环神经网络及其变体模型中，长短期记忆网络（long short-term memory, LSTM）[8]由于门机制而不容易受到梯度消失干扰因而受到更多研究者的关注。随着Transformer模型[9]的出现，一部分研究者基于Transformer结构，来构建表达力更强、更加高效的语言模型[10–14]。神经语言模型（如Word2Vec [15]和GloVe [16]）训练可以得到词的分布式表示[通常被称作词嵌入（word embeddings）]。使用训练后的分布式表示初始化深度学习模型

* Corresponding author.
E-mail address: wanghaifeng@baidu.com (H. Wang).

英文原文：Engineering 2023, 25(6): 51–65
引用本文：Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, Yu Sun. Pre-Trained Language Models and Their Applications. *Engineering*, https://doi.org/10.1016/j.eng.2022.04.024

中词嵌入层，已经是一种常用的技巧。这种方式可以显著提高下游任务的效果。比如，命名实体识别[16]、词性标注[17]和问答[18]等下游任务。

尽管使用词嵌入初始化的方式可以提高下游NLP任务的性能，但是它们仍然缺乏在不同语境中表达不同含义的能力。为了解决这个问题，研究者把完整的上下文信息加入到模型训练中，提出了语义感知模型。Dai和Le [19]使用未标注的数据来提高循环神经网络的序列学习的能力。这种方式在情感分析、文本分类等任务上取得了显著的性能提升。2017年，基于上下文的词向量（contextualized word vectors, CoVe）模型被提出。CoVe模型首先在机器翻译任务上进行预训练，之后再将编码器迁移到NLP下游任务中[20]。然而，这种方式仅仅在少量的机器翻译数据上进行了预训练，并且没有在所有的NLP任务上得到一致的性能提升。尽管如此，这些极具开创性的工作极大启发了后续基于上下文的预训练方法。

嵌入迁移的语言模型（embeddings from language models, ELMo）是另外一个预训练模型的开创性的工作。研究者通过首先利用双向的长短期记忆网络（bidirectional LSTM, Bi-LSTM）来学习单词的上下文表示，并且将预训练后的上下文向量直接应用到下游任务中[21]。这种方法在一系列自然语言处理任务中都取得了巨大的效果提升，其中包括机器问答、情感分析、语义角色标注、共指消歧和命名实体识别。

在这之后，一系列基于“预训练-微调”范式的预训练模型被提出。生成式预训练模型（generative pre-training, GPT）[22]首次提出使用单向Transformer结构做生成式语言模型预训练。作者通过实验验证了生成式预训练在下游任务上的巨大潜力。在GPT之后，文献[23]第一个利用双向Transformer模型作为模型的编码器，因此被称作双向基于Transformer的编码器表示模型（bidirectional encoder representations from transformers, BERT）。BERT模型通过多层神经网络对左右上下文进行建模来获得双向上下文语境。BERT提出了一种噪声自动编码预训练任务，被称作掩码语言模型（masked language modeling, MLM）。掩码语言模型类似于填空任务，是指模型基于上下文语境预测被掩盖位置的词。这种方式极大地提高了下游自然语言理解（natural language understanding, NLU）任务的性能。掩码语言模型的预训练标签通常来自于无标注数据自身。因此，这种方式也被称为自监督学习。通过利用互联网上的大规模无标注数据，模型可以通过预训练自动学习到语言的语法和语义表示。

预训练模型的巨大成功引发了研究者探索模型规模以及预训练技术的边界。其中，代表性的工作包括DeBERTa [24]、T5 [25]、GPT-3 [26]、CPM [27]、PanGu-α [28]和ERNIE 3.0 Titan [29]。大规模预训练模型（如GPT3）现在已经证明了在无样本和少量样本设置下的强大能力。通过几十个例子，GPT-3可以在SuperGLUE [30]取得与微调BERT相似的性能。GPT-3还可以生成高质量的创造性文本，以至于即使是人类也无法判断这些文本是否是由人类编写的。GPT-3的成功使得未来几十年里，人类可以使用GPT-3进行通用文本生成。在过去的数十年中，GPT-3所展现出来的能力被认为是不可能达到的。

通过结合知识来增强预训练模型的表达能力是另外一种预训练的方法[31]。一些研究者利用语言学知识，利用有标注或弱监督，来设计实体相关的预训练任务。例如，他们首先采用实体级或短语级掩码[31]和实体替换[32]等方式，打乱输入文本的实体边界，然后让模型预测被掩盖掉的实体内容。通过这种知识掩盖策略，模型可以更好地学习文本中的词汇、语法和语义信息。除此之外，将结构化知识与普通文本整合到预训练任务中也是其中一个研究方向。比如，K-BERT [33]、CoLAKE [34]、ERNIE-THU [35]、KnowBERT [36]、SenseBERT [37]、KEPLER [38]和ERNIE 3.0 [39]。ERNIE 3.0通过将知识融入到预训练模型中，在54个中文NLP标准测试集和一些英文基准测试（比如，SuperGLUE [30]）中取得了当时世界第一的结果。除此之外，K-Adapter [40]使用多个独立的适配器来完成不同的任务，更好地融合各种知识并且减轻了微调过程中经常出现灾难性遗忘的影响。知识融合极大地提高了非结构化文本与结构化文本之间的知识共享，因此大大提高了预训练模型的知识记忆和推理能力[39]。

然而，上述模型仅仅关注了数据资源丰富的语言，比如英语和中文。由此，很多资源紧缺的语言被忽视了。考虑到上述原因，很多研究者开始探索多语言预训练模型（multilingual pre-trained models, mPTM），通过将不同语言的语义表示限制在统一的向量空间中，实现来将资源丰富的语言知识迁移到资源短缺的语言中。受到BERT的启发，mBERT [41]提出使用多语言语料库，并采用多语言掩码语言模型（multilingual masked language modeling, MMLM）进行模型的预训练。从直观的角度来看，使用平行语料库更加有利于学习不同语种的跨语言表示学习。因此，XLM [42]利用双语句子对数据，并使用翻译语言模型（translation language modeling, TLM）作为目标进行预训练。采用TLM作为训练目标，可以鼓励模型将两种语言的表示对齐在一起，得到更好的跨语言表示。基于MMLM和TLM，研究者还训练了更多的多语言语言模

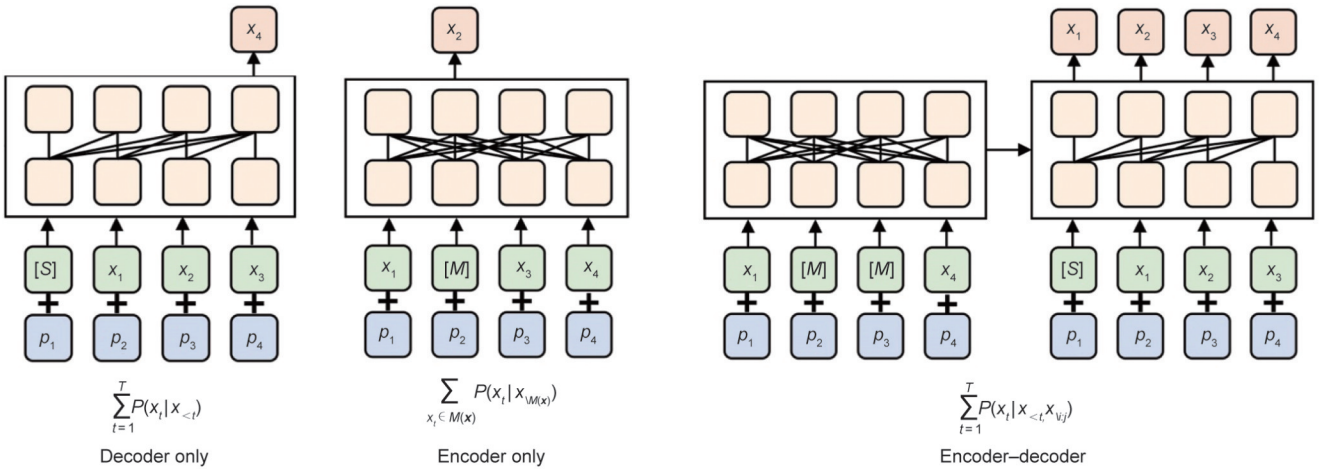

图1. 本图是几种预训练框架的示例。其中，$\boldsymbol{x}$代表原本的句子，$\boldsymbol{x}_t$ (t = 1,2,⋯,T)代表第t位置的词，T是序列长度，$M(\boldsymbol{x})$用于将词$\boldsymbol{x}$掩盖掉的特殊字符。S表示序列的起始标记嵌入。p_1、p_2、p_3和p_4表示第一至第四标记的位置嵌入。P是条件概率。i和j分别表示编码器输入标记的起始和结束索引。

型，比如XLM-R [43]、InfoXLM [44]和ERNIE-M [45]。这些研究表明，预训练多语言语言模型可以显著提高多语言NLP任务或者低资源语言任务的性能。

预训练模型在自然语言处理的成功，使得预训练技术迅速扩展到其他领域，如计算机视觉[46–48]和语音处理[49]。尽管基于自监督的预训练是目前自然语言处理领域最成功的迁移学习方式，但适用于计算机视觉领域的预训练方式是多样化的。在这其中，监督学习是计算机视觉任务中主要的预训练方式。Sun等 [48]发现使用大规模（噪声）有标数据集（如ImageNet [50]，JTF300M [48]）进行表示学习可以提高模型性能。通过学习视觉表示，模型在视觉下游任务的效果得到了显著的提升[48]。除此之外，一部分计算机视觉领域的研究者对自监督的模型预训练也展开了探索[51–56]。Doersch等 [53]中提出了一系列预测任务用来建模物品的视觉表示。Dosovitskiy等 [57]中使用Transformer的模型结构，在图像数据上利用掩码补丁预测来进行预训练。实验结果表明，预训练后的Transformer模型取得了与强对照卷积神经网络（convolutional neural network, CNN）模型相近的结果。

近期，对比学习已经被成功应用到视觉领域自监督预训练中。此外，对比预测编码[58]在语音、图像、文本和强化学习等各种场景中取得了很好的效果。文献[58–60]通过计算对比损失函数，使得数据增强后来自于同一张原始图片的两张不同图片相似度最大化，而不同原始图像的相似性最小化。近期，在视觉表示预训练任务上利用自然语言作为监督信号的方法也取得了很大进展[61]，并在图像分类和其他视觉任务中取得很好的效果。

预训练模型和方法同样适用于多模态领域。通过文本预训练模型与其他模态（比如，图像[62–65]、视频[66–67]和语音[68]）的融合，将预训练模型的应用范围拓展到了多模态领域。例如，文献[63]通过联合建模图片和文本的任务-已知性表示，显著地提高了在多个多模态任务上的性能。基于Transformer模型结构，预训练模型利用大规模图像-文本对数据，来对跨模态的语义对齐进行建模。对于视觉生成任务，DALL-E [69]和基于CLIP的生成[61]利用语言模型和视觉输入来生成引人注目的视觉场景。虽然多模态上最常用的预训练任务是MLM和遮蔽区域预测（masked region prediction, MRP），但Yu等[70]提出了知识增强的场景图预测，用来捕捉更详细语义的对齐。Gan等[71]将对抗性训练加入预训练过程，显著提高了模型的性能。Cho等[72]将多模态预训练制定为基于多模态上下文的统一语言建模任务。这些证据表明，预训练模型在人工智能（artificial intelligence, AI）社区中发挥着关键作用，有可能促进跨语音、计算机视觉和自然语言处理等研究领域的预训练框架的统一。

目前已经有一些关于预训练模型（pre-trained model, PTM）的调研总结论文。其中，一些文章聚焦于特定类型和应用的预训练模型。比如，基于Transformer的预训练语言模型（T-PTLMs）[73]、基于BERT的训练技术总结[74]、提示性学习[75]、数据增强[76]、文本生成[77]和对话设计[78]。另外一些文章则从全景的角度，对整个预训练模型的进展进行概括总结。例如，Ramponi和Plank [79]从早期的传统非神经方法介绍到目前的预训练语言模型。Qiu等 [80]从四个不同的角度，对已有的预训练语言模型进行了分类，并指出一些未来潜在研究方向。Bommasani等 [81]提出了基础模型的概念，提出了将不同子领域（如NLP、计算机视觉和语音）的预训练模型统一为一个模型的概念，并分析了它们在各种AI领域中的机遇和

挑战。Han等 [82]深入探讨了预训练模型的历史，揭示了预训练模型在AI发展进程中的重要地位。与之前工作不同的是，我们主要关注自然语言处理任务中的预训练模型：首先，我们详细分析了不同的预训练模型和规模化预训练模型的趋势，之后讨论它们对NLP领域的影响以及预训练模型的主要挑战；然后，本文针对预训练模型在工业应用中的观察和实践展开了详细的介绍。

在本文中，我们将首先在第2节中总结预训练模型的方法和类别，并在第3节中讨论预训练模型的影响和挑战。接下来的第4节中，我们将介绍预训练技术在工业界场景中的应用。最后，我们将对本文内容进行总结并对该领域的潜在研究方向展开讨论。

2. 预训练模型的方法

2.1. 预训练模型的不同框架和扩展

使用PTMs时，设计有效的训练方法以充分利用未注释的数据并协助下游微调非常重要。在本节中，我们简要介绍了目前一些广泛使用的预训练框架。图1总结了现有的流行预训练框架，可分为三类：仅使用Transformer解码器；仅使用Transformer编码器；使用Transformer解码器-编码器。以下是每个类别的简要描述，并在随后的小节中提供更详细的信息。

- 仅使用Transformer解码器的框架利用单向（从左到右）的Transformer解码器作为预训练骨干，并以单向自回归的方式预测标记。在这里，“自回归”是指基于历史标记预测当前标记，即当前标记左侧的部分序列。更具体地说，给定文本序列$\boldsymbol{x}=(x_1, x_2, x_3, \ldots, x_T)$[其中，$\boldsymbol{x}$是原始句子，$x_t$ (t=1, 2, ⋯, T)是第七个标记，T是序列长度]，自回归模型将输入文本序列的可能性分解为$p(\boldsymbol{x})=\prod_{t=1}^{T} p(x_t|x_{<t})$，其中，$p$是输入文本序列的可能性。
- 仅使用Transformer编码器的框架利用双向Transformer编码器，旨在恢复被随机屏蔽的标记，给定输入句子。
- 使用Transformer编码器-解码器的框架旨在通过在源侧屏蔽标记并在目标侧恢复标记来预训练序列生成模型。这些框架包括两类：①序列到序列编码器-解码器，包括具有单独参数的双向Transformer编码器和单向解码器；②统一编码器-解码器，其中双向Transformer编码器和从左到右的解码器同时预训练，具有共享的模型参数。

2.1.1. 仅使用Transformer解码器的框架

语言建模的目标是给定其历史，自回归地预测下一个标记。自回归的性质意味着每个位置的输入标记未来是不可见的，即每个标记只能关注前面的单词。生成式预训练模型（GPT）[22]是第一个使用Transformer解码器架构作为其主干的模型。给定一个单词序列作为上下文，GPT使用Transformer的屏蔽多头自注意力计算下一个单词的概率分布。在微调阶段，Transformer的预训练参数被设置为下游任务模型的初始化。GPT在BooksCorpus数据集上进行预训练，该数据集的大小几乎与1B Word Benchmark相同。它拥有数亿个参数，并在12个NLP数据集中的9个上改进了SOTA结果，展示了大规模PTMs的潜力。GPT-2 [83]遵循使用Transformer解码器的单向框架，使用了更大的WebText语料库进行训练，拥有15亿个模型参数。GPT-2在零样本设置下在8个测试语言建模数据集中的7个上达到了SOTA结果。GPT-3 [26]进一步增加了Transformer的参数到1750亿，并引入了上下文学习。GPT-2和GPT-3都可以在不进行微调的情况下应用于下游任务。它们通过扩大模型和数据集的规模实现了强大的性能。

单向语言建模仅对前面的语境进行自回归建模，忽略了后面的语境，这可能会降低下游任务的性能。为了解决这个问题，Yang等 [84]提出了置换语言建模（PLM）的方法，它对输入标记进行置换自回归建模。例如，“I love the movie”这个句子的置换可以是“I the movie love”。一旦选定了置换，置换后的句子的最后几个标记就是需要预测的目标。在上面的例子中，“love”这个标记是目标，依赖于可见的上下文“I the movie”。PLM的优点在于它可以充分利用上下文信息，为不同的被屏蔽标记建立依赖上下文的关系，包括前面和后面的单词。为了实现PLM，Yang等 [84]提出了一种新的双流自注意力机制，一个查询流用于计算查询向量，另一个内容流用于计算键/上下文向量。双流自注意力方法避免了可见上下文泄漏到被屏蔽的位置。

2.1.2. 仅使用Transformer编码器的框架

预训练的Transformer编码器，如BERT [23]，已成为NLP系统的标准。BERT使用Transformer作为骨干结构，并采用MLM框架。在预训练阶段，BERT随机替换单词，用特殊标记[MASK]来尝试基于上下文表示恢复损坏的单词。它还采用下一句预测（NSP）目标来捕捉两个句子之

间的语篇关系，这对于句子级任务（如问答）很有帮助。Devlin 等 [23]根据文献[85]将这个过程称为填空任务。BERT 是在 BooksCorpus（8 亿个单词）和英文维基百科（25 亿个单词）的组合上进行预训练的，在 17 个 NLP 任务上取得了显著的改进，甚至在一些下游任务中超过了人类的表现水平。然而，BERT 的缺点也很明显：因为[MASK]标记在微调期间在真实数据中不会出现，所以会在预训练和微调之间产生不匹配。为了解决这个问题，BERT 使用了一种新方法来屏蔽标记：在 15% 的随机位置中，只有 80% 被[MASK]标记替换，10% 保留为原始标记，10% 被替换为训练过程中的随机标记。这种屏蔽策略会导致模型需要更多的步骤来收敛，因为训练批次中只有 15% 的标记被预测。BERT 的另一个问题是它独立地预测标记，而没有考虑其他屏蔽标记。文献[86]提出的模型是一个统一的编码器-解码器模型，倾向于通过删除输入句子的文本跨度，并自回归地预测被屏蔽的跨度，从而缓解在屏蔽语言模型的预训练中同一跨度内的屏蔽标记独立假设问题。

在 BERT 成功之后，大量的研究工作投入到了 MLM 中。SpanBERT [87]设计了一种预测文本片段的模型，选择随机连续片段进行屏蔽，引入了一个片段边界预测目标，强制模型根据片段边界的结构信息来预测被屏蔽的文本片段。它通过将 BERT 中的 NSP 目标替换为单序列训练来获得更好的性能。SpanBERT 在问题回答和指代消解等与文本片段相关的任务上优于 BERT。类似于 SpanBERT，ERNIE [31]使用中文分词器获取短语信息，然后将 BERT 中的随机标记屏蔽替换为实体或短语屏蔽。ERNIE 还利用命名实体识别工具包识别实体边界，并在实体级别随机屏蔽标记，从而实现将外部知识集成到模型预训练中。

2.1.3. 使用 Transformer 编码器-解码器的框架

Transformer 编码器-解码器架构致力于自然语言生成（NLG）任务。与自然语言理解（NLU）专注于理解文本不同，NLG 旨在根据特定输入生成一种连贯、有意义且类人的自然语言表达。例如，机器翻译的目标是生成与给定源语言输入相同含义的目标语言句子；对于文本摘要，目标是生成输入文档的简短版本，以捕捉核心意义和观点。关键点是同时建模两个序列，一个用于输入，另一个用于输出。

文献[88]提出了一种用于语言生成的掩码序列到序列（MASS）学习方法，以预训练序列到序列模型。ASS 的基本思想是将带有掩码片段（即几个连续的标记）的句子作为输入，并在编码器表示的条件下预测掩码片段。这样，通过在源端进行掩码和在目标端进行预测，MASS 成功地将 Transformer 编码器框架转化为自回归框架。MASS 使用 WMT News Crawl 数据集的单语数据进行预训练，并在与直接使用注释数据训练的模型相比的机器翻译质量上取得了实质性的改进。

对 Transformer 编码器和解码器进行预训练会得到一个统一的模型，可以同时处理语言理解和语言生成。这个类别中的一种是标准的 Transformer 编码器-解码器模型，它并不共享统一的编码器和解码器组件。双向和自回归 Transformer（BART）[89]提出了与 MASS 类似的目标，但不同的是 MASS 掩盖了连续的一系列标记，即输入的 n-gram，而 BART 使用任意噪声函数破坏文本，即在不同位置掩盖/删除/替换/交换随机标记。BART 可以被视为上述两种架构的组合：源端的随机掩码策略使模型能够处理 NLU 任务，而整体的序列到序列预训练框架使模型能够推广到 NLG 任务。在新闻、图书、故事和网络文本的 160 GB 数据上进行预训练，BART 达到了与 RoBERTa [90]相当的结果，并在对话和抽象文本摘要上获得了新的 SOTA 结果。这个类别的另一种方法将编码器和解码器统一为相同的 Transformer 块。Dong 等[91]和 Bao 等[92]也提出了一个统一的语言模型预训练框架，用于 NLU 和生成任务。这些研究将自注意力矩阵划分为三个组件：双向组件、单向组件和序列到序列组件，分别代表单向、双向和序列到序列语言模型。实验证明，使用单一的预训练目标会带来性能提升。Du 等 [86]提出了文献[91]中报道的模型的变种，将掩码标记放置在未掩码标记的右侧，并进行自回归填充。Xiao 等 [93]在不同的粒度上屏蔽多个片段，以鼓励解码器更多地依赖编码器表示，从而增强编码器与解码器之间的相关性。Zhang 等[94]采用了不同的方法：首先，根据预定义的重要性标准从输入文档中删除一句话，然后基于剩余的上下文句子生成删除的句子。这种策略在句子级别执行自回归，促进对整个文档的理解和类似摘要的生成。在 12 个下游摘要任务上的实验展示了 SOTA 结果，显示了间隙句子预训练方法的有效性。

2.2. 放大预训练模型

NLP 的最新进展展示了使用数十亿参数扩展 PTMs 的有前途的趋势。OpenAI 研究人员训练了一个名为 GPT-3 的模型，该模型具有 1750 亿个参数[26]。GPT-3 在许多 NLP 数据集上实现了强大的性能，包括问答、机器翻译和三位数算术。GPT-3 表明，扩大语言模型可显著提高任务不可知性和少样本性能，有时甚至比之前的 SOTA 微调方法取得更好的结果[26]。尽管大型预训练模型是一个很有

前途的方向，但训练大规模PTMs是一项具有挑战性的任务，需要大量的训练数据和GPU资源。因此，高效的模型训练算法在扩大PTMs方面起着至关重要的作用。以下部分介绍了流行的大规模 PTMs 以及用于实现它们的训练方法。

2.2.1. 大规模预训练模型

表1 [24–28,39,95–102]总结了主流的大规模PTMs。近年来，PTMs的规模越来越大，从26亿到1750亿个参数不等。大规模预训练语言模型包含大量训练配方，包括指数级增加的可训练参数、预训练架构、知识增强、特定语言语料库和不同的预训练任务，以支持PTMs的十亿级训练。尽管这些模型的训练方法不同，但由于后者高效的并行计算性能，所有PTMs都使用transformers [9]作为标准主干。由于训练大规模模型需要大量无监督数据，因此扩展PTMs的研究主要集中在英文和中文等高资源语言上。

根据预训练架构中使用的不同设计，大规模PTMs通常可以分为三类（如第2.1节所述）：纯编码器、纯解码器和编码器-解码器。大多数大型PTMs仅利用纯解码器或编码器-解码器架构，而只有少数大型模型采用纯编码器设计。这是因为纯编码器模型不能很好地执行生成任务，如文本摘要和对话生成，而专为语言生成而设计的纯解码器模型不仅可以阐明NLG，还可以通过流行的提示技术阐明语言理解任务，如GPT-3 [26]。

- 大规模纯编码器模型采用双向transformer编码器来学习上下文表示；它们在NLU任务上表现出色。例如，DeBERTa 1.5B [24]由48个具有15亿个参数的transformer层组成，应用了分离的注意力机制并增强了掩码解码器以在 SuperGLUE [30]基准测试中超越人类性能。由于双向性使得模型无法直接用于NLG任务，DeBERTa训练了另一个版本的统一编

表1 大规模预训练语言模型总结

Model	Number of parameters	Model architecture	Knowledge learning	Language	Pre-training data	Training strategy	Training platform	Reference
$DeBERTa_{1.5B}$	1.5 billion	Encoder only	—	English	English data (78 GB)	—	PyTorch	[24]
T5	11 billion	Encoder–decoder (seq2seq)	—	English	C4 (750 GB)	Model/data parallelism	TensorFlow	[25]
GPT-3	175 billion	Decoder only	—	English	Cleaned CommonCrawl, WebText	Model parallelism	—	[26]
CPM	2.6 billion	Decoder only	—	Chinese	Chinese corpus (100GB)	—	PyTorch	[27]
PanGu-α	200 billion	Decoder only	—	Chinese	Chinese data (1.1 TB, 250B tokens)	MindSpore auto-parallel	MindSpore	[28]
ERNIE 3.0	10 billion	Encoder–decoder (unified)	√	Chinese, English	Chinese data (4 TB), English data	Model/pipeline/tensor parallelism	PaddlePaddle	[39]
Turing-NLG	17 billion	Decoder only	—	English	English data	DeepSpeed/ZeRO	—	[95]
HyperCLOVA	204 billion	Decoder only	—	Korean	Korean data	—	—	[96]
CPM-2	11 billion	Encoder–decoder (seq2seq)	—	Chinese, English	WuDao corpus (2.3 TB Chinese + 300 GB English)	—	PyTorch	[97]
CPM-2-MoE	198 billion	Encoder–decoder (seq2seq)	—	Chinese, English	WuDao corpus (2.3 TB Chinese + 300 GB English)	Mixture of Experts (MoE)	PyTorch	[98]
Switch transformers	1751 billion	Encoder–decoder (seq2seq)	—	English	C4 (750 GB)	MoE	TensorFlow	[99]
Yuan 1.0	245 billion	Encoder–decoder (unified)	—	Chinese	Chinese data (5 TB)	Model/pipeline/tensor parallelism	—	[100]
GLaM	1.2 trillion	Encoder only	—	English	English data (1.6 trillion tokens)	MoE/model parallelism	TensorFlow	[101]
Gropher	280 trillion	Decoder only	—	English	English data (10.5 TB)	Model/data parallelism	Jax	[102]

ZeRO: zero redundancy optimizer; MoE: mixture-to-expert.

表2 大规模多模态PTMs

Model	Number of parameters	Pre-training paradigm		Pre-training Data	Training parallelism	Training platform	Reference
		Denosing auto-encoder	Causal language model				
DALL-E	12 billion	×	√	250 million English text-image pairs	Mixed-precision training	PyTorch	[69]
CogView	4 billion	×	√	30 million English text-image pairs	—	PyTorch	[103]
M6	100 billion	√	×	1.9 TB images + 292 GB Chinese	MoE	—	[104]
ERNIE-ViLG	10 billion	√	√	145 million Chinese text-image pairs	Mixed-precision training	PaddlePaddle	[107]

码器-解码器来适应NLG任务。

- 纯解码器模型通过应用自回归掩码来使用transformer解码器，以防止当前标记关注未来标记。示例包括GPT-3 [26]、CPM [27]和PanGu-α [28]。这一系列PTMs旨在生成类似人类的文本。Turing-NLG [95]是一个170亿参数的语言模型，在语言模型基准测试中取得了强劲的性能。具有1750亿个参数的GPT-3可以惊人地编写欺骗人类读者的样本，证明大规模语言模型可以通过上下文学习显著推进少样本学习场景。除了英文大规模单语PTMs，还有中文、韩文等其他语言的模型。CPM [27]（26亿参数）和PanGu-α [28]（2000亿参数）是GPT-3的两个中国变体，而HyperCLOVA [96]是一个2040亿参数的韩国变体。
- 编码器-解码器模型可以进一步分为两类：①传统的序列到序列（seq2seq）编码器-解码器；②统一编码器-解码器。传统的序列到序列（seq2seq）编码器-解码器采用经典的transformer编码器-解码器架构进行预训练。最近的工作包括文本到文本转换的transformer（T5）[25]、多语言T5（mT5）[97]和经济高效的预训练语言模型（CPM-2）[98]。T5 [25]具有多达110亿个参数，通过以文本到文本的方式转换语言理解和生成任务，将NLP任务统一在一个框架中。作为T5的多语言变体，mT5 [97]拥有多达130亿个参数，将单语言数据扩展到101种人类语言，并在各种多语言基准测试中优于之前的SOTA结果。CPM-2 [98]具有110亿个参数，是一个在中文和英文上训练的双语模型，其混合专家（MoE）版本表示为CPM-2-MoE，具有1980亿个参数。该模型通过微调和提示展示了出色的通用语言智能。另一种编码器-解码器模型是统一的编码器-解码器框架，其中编码器-解码器架构共享相同的模块，并针对MLM和自回归语言建模应用不同的掩码策略。ERNIE3.0 [39]通过设计两个独立的理解和生成head来共同学习语言理解和生成，这两个head共享一个与任务无关的表示。作为ERNIE系列的第三代PTMs（100亿参数），ERNIE3.0结合了自回归因果语言模型和自编码模型的优点来训练大规模知识增强的PTMs。它在包括SuperGLUE [30]在内的各种NLP基准测试中的表现都超过了SOTA。这些方法表现出优越的性能，因为它们都倾向于将多个NLP任务统一在一个模型中，并使用不同种类的语料库或知识来提高性能。

上面提到的大多数大规模模型都是在没有整合知识的情况下在纯文本上训练的。因此，一些研究人员尝试将语言知识和世界知识等知识纳入PTMs。ERNIE 3.0在海量非结构化文本和知识图谱上预训练transformer，以学习词汇、句法和语义信息。它通过知识集成、短语掩码和命名实体掩码丰富了PTMs。

语言PTMs的巨大进步引起了对多模态预训练的研究兴趣[72,103–107]。表2 [69,103–104,107]列出了大规模多模态PTMs的详细信息。DALL-E [69]是GPT-3的120亿变体，它在2.5亿个英文文本图像对上进行训练，根据语言描述生成图像，从而提高零样本学习性能。ERNIE-ViLG [107]使用统一的生成预训练框架进行双向图像文本生成，将图像和文本生成制定为自回归生成任务。因此，它在文本到图像合成和图像描述等生成任务上优于以前的方法，该模型使用在1.45亿个高质量中文文本图像对上预训练的100亿参数模型。此外，多模态到多模态多任务巨型transformer（M6）[104]是一个拥有1000亿参数的transformer编码器，它在超过1.8 TB的图像和292 GB的中文文本上进行训练。M6在视觉问答、图像描述和中文图文匹配方面取得了出色的表现。除了对多模态任务的改进外，这些模型还可以提高单模态任务的性能，如文本分类、推理、摘要和问题生成[105]。这些结果表明，多模态预训练可以利用多模态信息来增强图像表示和文本表示，从而提高多模态任务和NLP任务的性能。

2.2.2. 大规模模型的高效训练

由于有限的GPU内存和无法承受的训练时间，PTMs

大小的指数增长对高效训练提出了巨大挑战。因此，利用有效的训练技术来加速大规模模型训练并非易事。

2.2.2.1. 密集模型

数据并行是一种简单的解决方案，它将不同的数据分区分配给多个机器，并在所有机器处复制相同的参数。但是，它通常会受到每个GPU批处理大小较小的影响。另一种解决方案是模型并行性，其中模型参数被分配给不同的机器。然而，传统的优化算法需要每个参数额外的内存来存储中间状态，这阻碍了模型大小的有效更新。流水线并行结合了模型并行和数据并行的优点，减少了低效的时间成本。Gpipe [108]使用一种新颖的批量分割流水线算法，首先将训练样本的小批量分成更小的微批量，然后在最后同时聚合梯度更新。Megatron-LM [109]是一种用于transformer网络的层内模型并行方法，它在自注意力和多层感知器（MLP）块上添加了一些同步原语。PTD-P [110]将跨多GPU服务器的流水线、张量和数据并行与一种新颖的交错流水线调度策略相结合，将吞吐量提高了10%以上。最近，Colossal-AI [111]实现了各种数据、流水线、序列和多张量并行的组合，用于大规模模型训练，这可能是训练密集模型的一个很好的选择。

2.2.2.2. 稀疏模型

稀疏门控混合专家（MoE）模型[112]使用多个专家子网络的稀疏门控组合实现了超过1000倍的模型容量增量。通过利用集成机制，MoE使用门控单元来确定应激活哪些top-*k*子网络进行预测。

开关transformer [91]通过简化稀疏路由并用开关路由替换前馈全连接层，将PTMs的规模扩大到多达数万亿个参数，其中每个样本仅路由到一个专家。

2.2.2.3. 其他高效训练策略

内存高效优化的最新技术包括混合精度训练[113]和内存高效自适应优化。混合精度训练利用半精度浮点数而不损失模型精度，这几乎将内存需求减半。其他研究旨在提高内存效率的自适应优化。例如，零冗余优化器（ZeRO）[114]是Turing-NLG的催化剂，由ZeRO-DP和ZeRO-RP算法组成，分别旨在减少模型状态的内存占用和剩余内存消耗。首先，ZeRO-DP通过执行优化器状态分区、添加梯度分区和添加参数分区来优化优化器状态、梯度和参数。然后，ZeRO-R通过去除激活复制、预定义适当的临时缓冲区大小和主动内存管理来优化剩余内存。

3. 预训练模型的影响和挑战

3.1. NLP预训练模型的影响

预训练模型的出现给NLP领域带来了重大突破。在预训练模型出现之前，许多研究都集中在为特定NLP任务设计专门的模型，这些模型通常无法应用于其他任务。例如，文献[115]提出了用于文本分类的TextCNN模型，文献[8]提出了用于自然语言生成的LSTM模型。自出现以来，预训练模型因其在表示学习方面令人印象深刻的能力而开始作为NLP中的基础模型。这为NLP开辟了一个新的“预训练然后微调”范式。该范式可以充分利用未注释的数据来训练基础模型，然后使用有限的特定任务的标注数据对其进行微调。即使标注数据有限，下游NLP任务的性能也有很大提高。图2 [23,39,116–117]展示了在五个NLP任务基准上SOTA结果从未预训练的监督模型到预训练模型如BERT和ERNIE 3.0的演变。可以清晰看到预训练模型明显优于之前的非预训练模型，且知识增强的ERNIE 3.0在许多NLP任务上稳步超过BERT。另一个重要趋势是采用预训练模型来统一所有NLP任务。例如，T5 [25]以文本到文本的方式将语言理解和生成任务组合到一起，以序列到序列的预训练模型处理所有NLP任务。因此，NLP社区也见证了任务统一的新兴趋势。

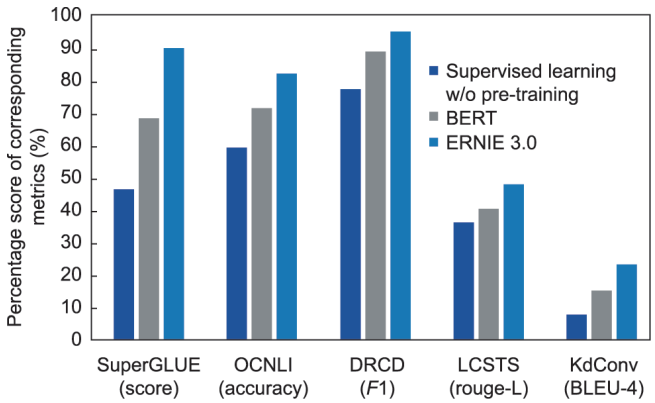

图2. 不同NLP基准测试中表示技术的演变转化。结果来自参考文献[23, 39,116–117]。SuperGLUE是一个由一系列困难的语言理解任务组成的NLU排行榜；OCNLI、DRCD、LCSTS和KdConv分别是自然语言推理、机器阅读理解、文本摘要和对话生成的评估语料库。w/o表示无。

GPT-3 [26]在零样本学习和少样本学习方面表现出了良好的性能。与GPT-3一起，一种新的利用提示进行训练的方法[118]被提出，重新制定了任务的范式。通过引入新范式以更好地利用预训练模型，预训练后用提示调优的方法引领了一种新趋势。不同于通过微调使预训练模型适应下游任务，下游任务被预定义为“槽填充”任务：给定一个人为设计的带有空槽的句子模板，让预训练模型学习

填充这些模板空槽。该框架已经被证明是强大的，它使语言模型能够适应少样本及零样本场景。因此，这一技术也在NLP社区引起了广泛的关注。我们一般从以下三个方面来描述预训练语言模型的影响：自然语言理解、自然语言生成和对话。对于对话，预训练模型专注于响应式生成。鉴于其有大量的相关工作，我们将对话作为一个单独的类别。

3.1.1. 自然语言理解

自然语言理解在NLP中是一个广泛的主题，其包括许多任务，如命名体识别、情感分析、文档归类、阅读理解、语义匹配、自然语言推理和信息抽取。表3 [39,116–117,119–120]比较了使用和不使用预训练技术的模型在四种不同NLU任务上的性能。可以看出，经过预训练的模型明显优于未经过预训练的模型。因此，预训练模型已经成为NLU任务中的标准主干。许多研究人员已经使用预训练模型来提供与任务无关的表示，然后设计特定于任务的架构或目标来提升NLU性能。例如，BERTGCN [121]结合了BERT的表示能力和图卷积网络（GCN）的直推式学习以提高文本分类性能，使其准确率提高了约4%。

为了比较预训练模型在NLU任务上的性能，研究人员在两个基准测试GLUE和SuperGLUE上上传他们的结果。这些预训练模型如今在两个排行榜上的表现已经优于人类。此外，mBERT [41]、XLM [42]、mT5 [97]、ERNIE-M [45]等多语言模型使用统一模型来表示多种语言，从而可以在不同语言之间共享学习信息。该技术缓解了低资源语言中的数据稀疏问题，并减少了为每种特定语言训练专门语言模型的需求。这种新范式正在改变NLP研究的重点，从为多语言任务设计专门的模型到研究如何在这些任务中使用预训练模型。

3.1.2. 自然语言生成

NLG任务，如文本摘要、问题生成和数据到文本生成，在NLP中都非常具有挑战性。由于巨大的搜索空间，预训练模型时代之前的方法因标注数据不足和参数有限，很难生成流畅、连贯和信息丰富的文本。如表4 [94,122–125]所示，预训练模型在提高NLG任务的性能方面发挥了关键作用。大规模预训练模型从未标注数据中自动学习单词组合和句子表达，显著提高了模型在语言生成方面的流畅性、连贯性和信息性。ERNIE-GEN [93]使用一种增强的多流序列到序列预训练和微调框架，并结合逐跨生成任务来生成连续的实体，在五个有代表性的NLG任务上取得了新的SOTA结果。研究人员和从业者还在生成任务上预训练特定任务的transformer模型，如MASS [88]和PEGASUS [94]。更具体地说，MASS采用编码器-解码器框架在给定句子剩余部分的情况下重建句子片段，在没有机器翻译预训练的情况下实现了比基准有显著提升。PEGASUS用于预训练具有精心设计的预训练目标的大规模编码器-编码器模型，该模型在所有12个文本摘要任务上均取得了SOTA效果。随着模型规模的增长，预训练模型逐渐展现出显著的创意写作能力。GPT-3、HyperCLOVA、ERNIE 3.0等模型仅通过零样本学习就能够生成文章、问答、小说和程序代码。生成文本的质量有时可以与人类编写的文本相媲美。例如，人类在区分真新闻和GPT-3生成的假新闻时的准确率仅为52%。

3.1.3. 对话

在过去的几年中，几个具有代表性的对话生成模型已经使用从社交媒体（包括Twitter、Reddit、微博、百度贴吧）收集的类人对话数据进行了预训练。基于通用语言模型GPT-2 [83]，DialoGPT [126]已经使用Reddit评论进行

表3 在NLU任务中，有无预训练的SOTA结果

NLU task	Sentiment analysis SST-2 binary classification (accuracy)	Natural language inference OCNLI (*F*1)	Nested named entity recognition GENIA (*F*1)	Machine reading comprehension DRCD (*F*1)
SOTA w/o pre-training	93.2	59.80	74.80	78.03
SOTA w/ pre-training	97.5	82.75	83.75	95.84

Results are from Refs. [39,116–117,119–120]. w/: with; SST-2: Stanford Sentiment Treebank v2; OCNLI: Original Chinese Natural Language Inference; DRCD: Delta Reading Comprehension Dataset.

表4 在NLG任务中，有无预训练的SOTA结果

NLG task	Text summarization AESLC (ROUGE-L)	Dialogue generation KdConv-film (BLEU-4)	Question generation SQuAD 1.1 (BLEU-4)	Data-to-text generation WebNLG (BLEU)
SOTA w/o pre-training	23.44	5.40	15.87	63.69
SOTA w/ pre-training	36.51	74.44	25.41	66.07

Results are from Refs. [94,122–125]. ESLC: English Skills Learning Center; BLEU: bilingual evaluation understudy; ROUGE-L: recall-oriented understudy for gisting evaluation-longest common subsequence.

了响应式生成的训练。Meena [127]将网络参数增大到2.6亿，并在训练过程中使用更多社交媒体对话，从而显著提高响应质量。为了减轻大型语料库中不受欢迎的有毒或偏见特征，Blender [128]进一步使用人工标注好的数据集对预训练模型进行微调，并强调了参与、共情和个性等理想的对话技巧。为了缓解开放域聊天中的安全响应问题，PLATO [129]将离散潜在变量编码进transformer以生成各式各样的响应。此外，PLATO-2 [130]通过课程学习进一步扩展了PLATO的中文和英文响应式生成。DSTC9挑战[131]表明，PLATO-2在多个会话任务中表现出卓越的性能，包括开放域聊天、基于知识的对话和面向任务的对话。最近，PLATO-XL [132]被扩展到110亿参数量，并进行多方感知预训练以更好地区分社交媒体对话中的角色。其他规模较小的中文对话预训练模型包括Cdial-GPT [133]、ProphetNet-X [134]以及EVA [135]。

通过这些大规模对话预训练模型，困扰传统端到端神经模型方法的一些问题[136–137]得到了显著缓解，包括回答流畅性和上下文相关性方面的不足。此外，与依赖复杂框架的现有聊天机器人（如Mitsuku [138]和XiaoIce [139]）相比，这些对话预训练模型在多轮对话中表现出卓越的性能，尤其是在参与性和人性化方面。

3.2. 研究关键挑战

尽管预训练模型显著提高了NLP任务的性能，但它们仍然存在一些关键挑战，如可解释性、鲁棒性、推理能力和大规模预训练模型的部署。本节描述了这些挑战，希望未来可以在这些方向上投入更多的努力。

3.2.1. 可部署性

预训练模型的一个趋势是容量的大幅增加。自从GPT [22]和BERT [23]发布以来，预训练模型的参数数量和预训练数据的大小都呈指数级增长。例如，最大版本的GPT-3 [26]需要3.64×10^3 petaflopdays的总训练计算量，导致总次数约为3.14×10^{23} flops，成本为数百万美元。模型规模的快速增长引起了对规模和可部署性之间权衡的关注。已经提出了两种类型的策略来解决这个问题：①大型预训练模型仅通过API调用用作基础模型，类似于GPT-3模型的使用方式。这种策略可以有效地使用预训练模型，并避免在每个设备上部署模型，但极大地限制了模型的应用范围。②大型模型被压缩为较小的模型[140]，以便进行潜在的部署。典型的压缩技术包括模型压缩和知识蒸馏。遗憾的是，现有的压缩技术无法将超大型预训练模型（如GPT-3）压缩到适合部署在单个GPU或终端设备（如笔记本电脑或手机）上的程度。因此，为了让更多的用户可以使用大型预训练模型，必须对模型压缩进行更进一步的研究。另一个前景光明的方向是使用参数高效化技术，如prompttuning [141–146]，以减少部署的内存预算；这仍然是一个有待进一步探索的广阔领域。

3.2.2. 模型的可信度

预训练模型的另一个挑战是其可信度，主要涉及其可解释性[147]和鲁棒性[148]。尽管预训练模型在各种任务中实现了SOTA性能，但它们如何做出决策有时对人类来说是模糊的，这使得预训练模型难以应用于模型可解释性至关重要的领域，如医疗保健和法律[149]。因此，人们对解释深度神经模型越来越感兴趣[150]。特别是，许多研究旨在了解预训练模型在其表示中学习了什么[151]。

一些关于深度神经模型可信度的研究已经发表。这些包括：对预训练模型的语言结构分析[152]，其目的是分析预训练语言模型所学习的语言知识，并试图理解其成功的原因；模型行为分析[153]，用多个测试集评估模型的鲁棒性和可靠性；事后解释分析[154]，旨在为深度模型的预测提供可理解的解释。

尽管在这一领域已经做了很多研究，但为了建立可靠的系统，必须解决以下挑战：①用于NLP任务的通用解释方法（现有的解释方法是为分类任务设计的）；②模型预测与所学知识或提取解释之间的因果分析；③可解释性综合评价平台，包括评价数据和评价指标。

3.2.3. 常识知识和推理

大规模预训练模型被发现编码一些常识性知识[155]。然而，为了挖掘在预训练模型中学习到的常识性知识，需要设计适当的探索任务，例如，将关系知识提取任务制定为填空语句的完成，以检验预训练模型的知识学习能力[156]。虽然预训练模型从文本中学习了一些知识，但仍然有大量的知识不能仅从文本中获得。一个可能的方向是让模型从多模态视觉输入和文本输入中学习这种知识。

除了常识知识之外，其他研究也在质疑预训练模型是否具有推理能力。例如，Talmor等[157]设计了不同的任务来评估预训练模型的推理能力。研究人员将预训练从微调中分离出来，发现大多数预训练模型的推理能力都很差，这表明现有的预训练模型缺乏推理能力。为了缓解这一问题，一个可能的方向是将先验知识集成到预训练模型中，以指导模型隐式地学习推理规则。

3.2.4. 模型安全性

预训练模型的一个严重问题是它们容易受到对抗性例

子的攻击，当输入中注入扰动时，可能会误导模型产生特定的错误预测[158]。这种易感性使预训练模型面临安全问题：模型很容易受到第三方的对抗模式攻击，在实际应用中造成不可挽回的损失。除了对抗性攻击之外，另一种形式的攻击，即后门攻击，也是对预训练模型的威胁。与通常在神经模型推理过程中发生的对抗性攻击不同，后门攻击在训练过程中入侵模型[159]。如果一个模型是故意在后门数据上训练的，那么用户在涉及隐私和安全问题的应用程序中使用这个模型将是极其危险的。未来的工作可以致力于提高预训练模型对对抗性攻击的鲁棒性。为了应对后门攻击，模型应该能够在输入中检测到可以激活后门攻击的触发器并移除触发器，从而增强模型的安全性。

4. 预训练模型的应用

4.1. 应用的平台和工具包

PTMs 由于其普适性已成为 NLP 中的基础模型。许多研究人员已经开发了一系列开源工具包和平台，以更好地利用 PTMs。这些工具包和平台通常包含各种 PTMs、微调工具和模型压缩工具。

4.1.1. 工具包

当研究人员提出一个新的预训练语言模型时，通常会开源相应的工具包以供使用。这种工具包通常提供了基于特定模型的下游任务开发代码，因此缺乏通用性。典型的工具包包括 google-research/bert [160]、PaddlePaddle/ERNIE [161] 和 PCL-Platform.Intelligence/PanGu-α [162]。这些工具包提供了一系列开源的 PTMs，如 BERT、ERNIE 和 PanGu-Alpha，以及源代码和训练数据。例如，ERNIE 工具包不仅提供 ERNIE 的源代码、训练数据和预训练模型，还提供了一些增强的 ERNIE 系列模型，如 ERNIE-Doc [163] 和 ERNIE-ViL [70]。为了将 ERNIE 模型部署到在线服务中，ERNIE 工具包还提供了模型压缩工具。

随着 PTMs 的广泛发布，了解如何在统一的工具包中使用这些模型已成为迫切的需求。在这种背景下，通用 NLP 应用程序的工具包已经开发出来。典型的工具包包括 HuggingFace/Transformers [164]、Fairseq [165] 和 PaddleNLP [166]。PTMs 以用户友好的方式集成到此类通用工具包中。以 HuggingFace 为例，该工具包集成了各类 PTMs 的代码和下游应用（包括分类、生成、摘要、翻译和问答等）的开发代码。

4.1.2. 平台

除了工具包，平台为用户提供了定制 PTM 服务的功能。这些平台可以为开发人员提供建立模型和将其部署到在线服务的工具。例如，百度文心[167]是一个旨在促进 PTMs 使用的平台。该平台满足有经验的开发人员和初学者的需求。它使开发人员能够轻松地构建自己的模型，因为用户只需要提供数据和模型配置。它还为经验丰富的开发人员提供工具包，以训练他们为应用程序量身定制的模型。其他平台如 AliceMind [168]提供类似的服务，它们没有明显的区别。OpenAIAPI [169]是另一种仅基于 PTMs 开发应用程序的平台。OpenAIAPI 基于 GPT-3 [26]，它提供特定的高级功能，如英法翻译、语法纠正、问答、广告生成和产品名称生成。

4.2. 应用

PTMs 已经广泛应用于实际应用程序，包括文档智能、内容创作、虚拟助手和智能搜索引擎。下面，我们将描述 PTMs 在每个领域中的应用。

4.2.1. 文档智能

PTMs 的一项广泛研究的应用是文档智能，其中包括情感分析、新闻分类、反垃圾邮件检测和信息提取。情绪分析被广泛用于识别情绪极性，如舆论，用于市场研究、品牌声誉分析和社交媒体影响。Garg 和 Chatterjee [170]中的研究人员提出使用 PTMs 对 Twitter 提要的情感进行分析，其标签采用三个类别的值，即正面、中性和负面。AlQahtani [171]中提出将数据挖掘技术与 PTMs 相结合来分析产品的客户评价。最近，Singh 等[172]使用 PTMs 分析了公众对冠状病毒对社交生活的影响的看法。Chen 和 Sokolova [173]提出分析公众在流行社交媒体平台上对新冠病毒肺炎相关消息中的情感，用户在该平台上分享他们的故事以寻求其他用户的支持，尤其是新冠病毒肺炎大流行期间。实验结果表明，PTMs 在情感极性分类方面可以实现显著的性能提升，证明了预训练模型的有效性。

新闻分类和反垃圾邮件检测也可以建模为分类任务。Ding 等 [163]应用 PTMs 将新闻分类为极左或极右的立场。Liu 等 [174]提出发布在 Arxiv.org 上的论文分为 11 个类别，包括数学和计算机科学等。Jwa 等 [175]使用 BERT 通过分析新闻标题与正文之间的关系来检测假新闻。

文档信息提取在工业界被广泛应用。许多 AI 云服务包含信息提取工具[176]，如谷歌 AI 云、百度 AI 云、阿里巴巴 AI 云。在这些服务中，百度建立了基于 PTM 的平台 TextMind，用于文档信息提取应用，包括费用报销的收据

分析、简历信息提取、财务报表分析、合同分析和法律判决分析。全球最大的在线家居零售商 Wayfair 也应用 BERT 从客户留言中提取信息。

文档图像理解是文档智能领域的另一个重要研究课题，用于自动读取、理解和分析商业文件。一系列多模态文档 PTMs [177]已经被提出，用于联合建模商业文件中文本、图像和布局信息之间的交互，用于许多文档图像理解任务，如收据理解、文档图像分类和文档信息提取。Applica 提出了一种考虑布局、图形和文本的解决方案，以便在金融服务、保险服务、生命科学等复杂业务流程中提取精确答案。

4.2.2. 内容创作

内容创作任务通常被设计用于验证最近提出的大规模模型[22]的性能。例如，Narrativa 应用 GPT-2 从客户提供的几个词中实现内容自动化，并生成高质量广告内容[178]。GPT-2 已经证明其为电子商务生成内容的能力，以将人类从繁重的任务中解放出来。微软也展示了预训练生成模型 Turing-NLG 对于自动建议推荐[95]的益处。此外，许多研究人员基于 GPT-3 构建了各种演示应用程序，包括广告生成、AI 文案、书籍撰写、代码生成、客户服务等。对于视觉内容创作，预训练多模态生成模型如 DALL-E [69]、CogView [103]和 ERNIE-ViLG [107]大大提高了生成图像的质量和保真度。CogView 的结果证明了该模型在工业时装设计等单一领域生成高质量图像的能力，因此该模型已部署在在线时装生产中。

除了这些工业应用外，研究人员还展示了 PTMs 在创意写作方面的潜在能力，包括诗歌生成[179]、歌词生成[27]、电子邮件自动完成[180]、TO-DO 生成[181]、句子和段落的自动完成甚至是长篇小说的生成[22]。尽管 PTM 表现出强大的生成能力，但越来越多的人对生成模型产生了担忧，包括隐私和版权。

4.2.3. 虚拟助手

如今，许多应用程序都采用了虚拟助手。典型应用包括智能音箱，如亚马逊的 Alexa [182]和百度的小度[129]。这些应用使用了 PTMs，并表明 PTMs 在智能音箱中可以提供出色的口语理解和语音识别[183]能力。通过 PTMs 带来的好处，这些智能音箱可以响应天气预报查询、点播歌曲以及语音控制智能家居设备。此外，智能音箱可以与人类就广泛的话题进行交谈，从而建立更紧密、更稳定的用户与系统之间的关系。除了在智能音箱中使用 PTMs，PTMs 还被部署在基于手机的虚拟助手中，如 Siri 和 GoogleAssistant。例如，NDTV [184]表明，PTMs 可以提高交互质量，而 Vicent [185]表明，PTMs 可以用于智能客服机器人来识别客户情绪。

随着 PLMs 在虚拟助手中应用越来越广泛，聊天机器人生成的响应也越来越接近人类。例如，微软提出了一种基于 PLM 的模型 DialoGPT，它从 Reddit 的评论历史中学习，可以流畅地回复用户。谷歌也建议使用 PLMs 开发可以“随便聊”[127]的聊天机器人应用程序。为使机器人更像人类，Facebook 将 PLM 应用于一系列对话聊天机器人，命名为 Blender 和 Blender2.0 [128]。不久之后，百度提出了基于 PLM 的模型 PLATO-XL [132]，进一步推动聊天机器人的性能并达到 SOTA 的人类评估和自动评估指标。由于 PTMs 的性能提高，这些应用程序在与用户的交互中可以非常稳健[186]。

4.2.4. 智能搜索

除了上述应用之外，PTMs 还广泛应用于搜索引擎中。谷歌已经将 PTMs 应用于其谷歌搜索中，并取得了显著的改进[187]。百度也应用了 PTMs，ERNIE 2.0 [188]和 ERNIE 3.0 [39]作为其骨干，通过将文本编码为密集表示来支持语义匹配，从而在百度搜索中获得更好的检索性能[189]。Facebook [190]揭示了一个用于个性化系统的统一嵌入框架，并指出他们未来的工作将包含 PTMs。

为了满足对多媒体内容搜索不断增长的需求，可以通过利用多模式 PTMs 来增强图像和视频搜索引擎的性能。例如，WenLan [106]开发了两个基于图像文本匹配的真实应用程序，从而展示了多模态预训练的强大功能。

为了进一步提高搜索引擎的性能，研究人员最近越来越关注多语言搜索引擎模型。多语言模型使用多语言语料库进行预训练以学习跨语言信息[191]。多语言模型最显著的优势是它们的跨语言可迁移性，从而提高在资源稀缺的语言上的表现。

5. 结论和未来工作

PTMs 可以充分利用未标注数据进行自监督学习，在 NLP 领域已经成为基础模型，显著提高下游 NLP 任务的性能。PTMs 的出现为 NLP 开辟了一种新的“预训练然后微调”的范式。随着模型参数的增加，PTMs 在零样本学习或少样本学习方面表现出了良好的性能。它们在 NLP 中的成功正在引发更多针对 PTMs 在计算机视觉、语音处理以及多模态理解和生成等其他领域的研究，揭示了它们作为这些领域基础模型的潜力。

尽管PTMs在NLP中取得了巨大的成功，但要实现通用人工智能仍有很长的路要走。首先，PTMs是不易理解的黑匣子。由于transformer模型的非线性，它们的可解释性和鲁棒性还有待探索。因此，在我们完全理解其原理之前，很难使用PTM做出可靠的决策和推理。研究PTM的不确定性是值得投入大量精力的。此外，目前的多模态和多语言预训练[192]仍处于早期阶段。统一多模式和多语言预训练将成为有待进一步探索的令人兴奋的趋势，这可能会提高这些资源稀缺任务的性能。另一个有前途的方向是将先验知识纳入PTMs中，以提高它们的推理能力和效率。现有的知识预训练工作，如K-BERT [33]和ERNIE3.0 [39]，已将知识三元组注入预训练或微调中。然而，PTMs在常识意识和推理方面表现出的能力有限，需要进一步改进。虽然大规模PTMs已经展示了强大的泛化能力，但如何有效地部署它们仍然是一个悬而未决的问题。针对需要低延迟的应用程序，PTM的模型压缩仍然是一个有前途的方向。现有的模型压缩方法包括蒸馏[193]、剪枝[194]、量化[195]等。然而，如何高效地构建具有可部署推理时间的大规模PTMs仍然是一个持续的挑战。此外，设计更高效的架构来代替或改进变压器仍然是一个开放性问题。

总之，要使PTMs能够拥有做出可靠的决策和进行可靠的规划这些人工智能的基本要素，仍有很长的路要走。需要提出和开发更加高效和强大的神经网络。幸运的是，在实际应用中使用PTMs不断提供更多的数据并应对新的挑战，有望促进新的预训练方法的快速发展。

Compliance with ethics guidelines

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Bahl LR, Jelinek F, Mercer RL. A maximum likelihood approach to continuous speech recognition. IEEE Trans Pattern Anal Mach Intell 1983; PAMI-5(2): 179–90.

[2] Thrun S, Pratt L. Learning to learn. Cham: Springer; 1998.

[3] Nadas A. Estimation of probabilities in the language model of the IBM speech recognition system. IEEE Trans Acoust Speech Signal Process 1984; 32(4): 859–61.

[4] Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. Comput Speech Lang 1999;13(4):359–94.

[5] Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. J Mach Learn Res 2003;3:1137–55.

[6] Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012); 2012 Sep 9 – 13; Portland, OR, USA. 2012. p. 194–7.

[7] Mikolov T, Zweig G. Context dependent recurrent neural network language model. In: Proceedings of 2012 IEEE Spoken Language Technology Workshop (SLT); 2012 Dec 2–5; Miami, FL, USA. 2012. p. 234–9.

[8] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997; 9(8):1735–80.

[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA. 2017. p. 5998–6008.

[10] Shazeer N, Cheng Y, Parmar N, Tran D, Vaswani A, Koanantakool P, et al. Mesh-TensorFlow: deep learning for supercomputers. In: Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018); 2018 Dec 3–8; Montréal, QC, Canada; 2018.

[11] Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019); 2019 Jul 28–Aug 2; Florence, Italy. 2019. p. 2978–88.

[12] Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. 2020. arXiv:2004.05150.

[13] Press O, Smith NA, Lewis M. Shortformer: better language modeling using shorter inputs. 2020. arXiv:2012.15832.

[14] Press O, Smith NA, Levy O. Improving transformer models by reordering their sublayers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020); 2020 Jul 5–10; online. 2020. p. 2996–3005.

[15] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th Conference on Neural Information Processing Systems (NIPS 2013); 2013 Dec 5–10; Lake Tahoe, NV, USA. 2013. p. 3111–9.

[16] Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 Oct 25 – 29; Doha, Qatar; 2014. p.1532–43.

[17] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach Learn Res 2011;12:2493–537.

[18] Xiong C, Zhong V, Socher R. DCN+ : mixed objective and deep residual coattention for question answering. In: Proceedings of the 6th International Conference on Learning Representations (ICLR 2018); 2018 Apr 30 – May 3; Vancouver, BC, Canada; 2018.

[19] Dai AM, Le QV. Semi-supervised sequence learning. In: Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2015); 2015 Dec 7–12; Montréal, QC, Canada. 2015. p. 3079–87.

[20] McCann B, Bradbury J, Xiong C, Socher R. Learned in translation: contextualized word vectors. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA; 2017.

[21] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018 Jun 1–6; New Orleans, LA, USA; 2018. p. 2227–37.

[22] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. San Francisco: OpenAI; 2018.

[23] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2 – 7; Minneapolis, MN, USA; 2019. p. 4171–86.

[24] He P, Liu X, Gao J, Chen W. DeBERTa: decoding-enhanced BERT with disentangled attention. In: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021); 2021 May 3–7; Vienna, Austria; 2021.

[25] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2019;21(140):1–67.

[26] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020); 2020 Dec 7 – 12; online. 2020. p. 1877–901.

[27] Zhang Z, Han X, Zhou H, Ke P, Gu Y, Ye D, et al. CPM: a large-scale

generative Chinese pre-trained language model. AI Open 2021;2:93–9.
[28] Zeng W, Ren X, Su T, Wang H, Liao Y, Wang Z, et al. PanGu-a: large-scale autoregressive pretrained Chinese language models with auto-parallel computation. 2021. arXiv:2104.12369.
[29] Wang S, Sun Y, Xiang Y, Wu Z, Ding S, Gong W, et al. ERNIE 3.0 Titan: exploring larger-scale knowledge enhanced pre-training for language understanding and generation. 2021. arXiv:2112.12731.
[30] Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, et al. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 9–14; Vancouver, BC, Canada. 2019. p. 3266–80.
[31] Sun Y, Wang S, Li Y, Feng S, Chen X, Zhang H, et al. ERNIE: enhanced representation through knowledge integration. 2019. arXiv:1904.09223.
[32] Xiong W, Du J, Wang WY, Stoyanov V. Pretrained encyclopedia: weakly supervised knowledge-pretrained language model. In: Proceedings of the 8th International Conference on Learning Representations (ICLR 2020); 2020 Apr 26–30; Addis Ababa, Ethiopia; 2020.
[33] Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, et al. K-BERT: enabling language representation with knowledge graph. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York City, NY, USA. Palo Alto: AAAI Press; 2020. p. 2901–8.
[34] Sun T, Shao Y, Qiu X, Guo Q, Hu Y, Huang X, et al. CoLAKE: contextualized language and knowledge embedding. In: Proceedings of the 28th International Conference on Computational Linguistics; 2020 Dec 8–13; online. 2020. p. 3660–70.
[35] Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: enhanced language representation with informative entities. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019); 2019 Jul 28–Aug 2; Florence, Italy. 2019. p. 1441–51.
[36] Peters ME, Neumann M, Logan IV RL, Schwartz R, Joshi V, Singh S, et al. Knowledge enhanced contextual word representations. In: 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 Nov 3–7; HongKong, China. 2019. p. 43–54.
[37] Levine Y, Lenz B, Dagan O, Ram O, Padnos D, Sharir O, et al. SenseBERT: driving some sense into BERT. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020); 2020 Jul 5–10; online. 2020. p. 4656–67.
[38] Wang X, Gao T, Zhu Z, Zhang Z, Liu Z, Li J, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation. Trans Assoc Comput Linguist 2021;9:176–94.
[39] Sun Y, Wang S, Feng S, Ding S, Pang C, Shang J, et al. ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. 2021. arXiv:2107.02137.
[40] Wang R, Tang D, Duan N, Wei Z, Huang X, Ji J, et al. K-Adapter: infusing knowledge into pre-trained models with adapters. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021); 2021 Aug 1–6; online. 2021. p. 1405–18.
[41] Wu S, BetoDredze M., Bentz, Becas: the surprising cross-lingual effectiveness of BERT. In: 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 Nov 3–7; HongKong, China. 2019. p. 833–44.
[42] Conneau A, Lample G. Cross-lingual language model pretraining. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 8–14; Vancouver, BC, Canada. 2019. p. 7057–67.
[43] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020); 2020 Jul 5–10; online. 2020. p. 8440–51.
[44] Chi Z, Dong L, Wei F, Yang N, Singhal S, Wang W, et al. InfoXLM: an information-theoretic framework for cross-lingual language model pretraining. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2021 Jun 6–11; online. 2021. p. 3576–88.
[45] Ouyang X, Wang S, Pang C, Sun Y, Tian H, Wu H, et al. ERNIE-M: enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2021 Nov 7–11; online. 2021. p. 27–38.
[46] Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, et al. DeCAF: a deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014); 2014 Jun 21–26; Beijing, China. 2014. p. 647–55.
[47] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2014 Jun 23–28; Columbus, OH, USA. 2014. p. 580–7.
[48] Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. 2017. p. 843–52.
[49] Schneider S, Baevski A, Collobert R, Auli M. Wav2vec: unsupervised pretraining for speech recognition. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association (InterSpeech 2019); 2019 Sep 15–19; Graz, Austria. 2019. p. 3465–9.
[50] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2009 Jun 20–25; Miami, FL, USA. 2009. p. 248–55.
[51] Mahajan D, Girshick R, Ramanathan V, He K, Paluri M, Li Y, et al. Exploring the limits of weakly supervised pretraining. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. 2018. p. 181–96.
[52] Zhai X, Kolesnikov A, Houlsby N, Beyer L. Scaling vision transformers. 2021. arXiv:2106.04560.
[53] Doersch C, Gupta A, Efros AA. Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. 2015. p. 1422– 30.
[54] Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles. In: Proceedings of the European Conference on Computer Vision (ECCV); 2016 Oct 8–16; Amsterdam, NetherlandsThe. 2016. p. 69–84.
[55] Misra I, van der Maaten L. Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 14–19; online. 2020. p. 6707–17.
[56] Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations. In: Proceedings of the 6th International Conference on Learning Representations (ICLR 2018); 2018 Apr 30–May 3; Vancouver, BC, Canada; 2018.
[57] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. In: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021); 2021 May 3–7; Vienna, Austria; 2021.
[58] Van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. 2018. arXiv:1807.03748.
[59] He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 14–19; online. 2020. p. 9729–38.
[60] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning (ICML 2020); 2020 Jul 12–18; online. 2020. p. 1597–607.
[61] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021); 2021 Jul 18–24; online. 2021. p. 8748–63.
[62] Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, et al. Scaling up visual and vision – language representation learning with noisy text supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021); 2021 Jul 18–24; online. 2021. p. 4904–16.
[63] Lu J, Batra D, Parikh D, Lee S. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 9–14; Vancouver, BC, Canada. 2019. p. 13–23.
[64] Tan H, Bansal M. LXMERT: learning cross-modality encoder representations from transformers. In: 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 Nov 3–7; Hong Kong, China; 2019.
[65] Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: a simple and

performant baseline for vision and language. 2019. arXiv:1908.03557.
[66] Sun C, Myers A, Vondrick C, Murphy K, Schmid C. VideoBERT: a joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. 2019. p. 7464–73.
[67] Sun C, Baradel F, Murphy K, Schmid C. Learning video representations using contrastive bidirectional transformer. 2019. arXiv:1906.05743.
[68] Chuang YS, Liu CL, Lee H, Lee L. SpeechBERT: an audio-and-text jointly learned language model for end-to-end spoken question answering. In: Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech 2020); 2020 Oct 25 – 29; Shanghai, China. 2020. p. 4168–72.
[69] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-shot text-toimage generation. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021); 2021 Jul 18–24; online. 2021. p. 8821–31.
[70] Yu F, Tang J, Yin W, Sun Y, Tian H, Wu H, et al. ERNIE-ViL: knowledge enhanced vision – language representations through scene graphs. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence; 2021 Feb 2–9; online. Palo Alto: AAAI Press; 2021. p. 3208–16.
[71] Gan Z, Chen YC, Li L, Zhu C, Cheng Y, Liu J. Large-scale adversarial training for vision-and-language representation learning. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020); 2020 Dec 7–12; online. 2020. p. 6616–28.
[72] Cho J, Lei J, Tan H, Bansal M. Unifying vision-and-language tasks via text generation. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021); 2021 Jul 18–24; online. 2021. p. 1931–42.
[73] Kalyan KS, Rajasekharan A, Sangeetha S. AMMUS: a survey of transformerbased pretrained models in natural language processing. 2021. arXiv:2108.05542.
[74] Kaliyar RK. A multi-layer bidirectional transformer encoder for pre-trained word embedding: a survey of BERT. In: Proceedings of 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence); 2020 Jan 29–31; Noida, India. 2020. p. 336–40.
[75] Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. 2021. arXiv:2107.13586.
[76] Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. 2021. arXiv:2111.01243.
[77] Li J, Tang T, Zhao WX, Wen JR. Pretrained language models for text generation: a survey. 2021. arXiv:2105.10311.
[78] Zaib M, Sheng QZ, Zhang W. A short survey of pre-trained language models for conversational AI—a new age in NLP. In: Proceedings of the Australasian Computer Science Week Multiconference (ACSW' 20); 2020 Feb 3 – 7; Melbourne, VIC, Australia. 2020.
[79] Ramponi A, Plank B. Neural unsupervised domain adaptation in NLP—a survey. In: Proceedings of the 28th International Conference on Computational Linguistics; 2020 Dec 8–13; onine. 2020. p. 6838–55.
[80] Qiu XP, Sun TX, Xu YG, Shao YF, Dai N, Huang XJ. Pre-trained models for natural language processing: a survey. Sci China Technol Sci 2020; 63(10): 1872–97.
[81] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. 2021. arXiv:2108.07258.
[82] Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, et al. Pre-trained models: past, present and future. AI Open 2021;2:225–50.
[83] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. San Francisco: OpenAI; 2019.
[84] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. XLNet: generalized autoregressive pretraining for language understanding. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 9–14; Vancouver, BC, Canada. 2019. p. 5754–64.
[85] Taylor WL. "Cloze procedure": a new tool for measuring readability. J Mass Commun Q 1953;30(4):415–33.
[86] Du Z, Qian Y, Liu X, Ding M, Qiu J, Yang Z, et al. GLM: general language model pretraining with autoregressive blank infilling. 2021. arXiv:2103.10360.
[87] Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O. SpanBERT: improving pre-training by representing and predicting spans. Trans Assoc Comput Linguist 2020;8:64–77.
[88] Song K, Tan X, Qin T, Lu J, Liu TY. MASS: masked sequence to sequence pretraining for language generation. In: Proceedings of the 36th International Conference on Machine Learning (ICML 2019); 2019 Jun 9–15; Long Beach, CA, USA. 2019. p. 5926–36.
[89] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020); 2020 Jul 5–10; online. 2020. p. 7871–80.
[90] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. 2019. arXiv:1907.11692.
[91] Dong L, Yang N, Wang W, Wei F, Liu X, Wang Y, et al. Unified language model pre-training for natural language understanding and generation. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 9–14; Vancouver, BC, Canada. 2019. p. 13042–54.
[92] Bao H, Dong L, Wei F, Wang W, Yang N, Liu X, et al. UniLMv2: pseudo-masked language models for unified language model pre-training. In: Proceedings of the 37th International Conference on Machine Learning (ICML 2020); 2020 Jul 12–18; online. 2020. p. 642–52.
[93] Xiao D, Zhang H, Li Y, Sun Y, Tian H, Wu H, et al. ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI); 2021 Jan 7 – 15; Yokohama, Japan. 2021. p. 3997–4003.
[94] Zhang J, Zhao Y, Saleh M, Liu P. PEGASUS: pre-training with extracted gapsentences for abstractive summarization. In: Proceedings of the 37th International Conference on Machine Learning (ICML 2020); 2020 Jul 12– 18; online. 2020. p. 11328–39.
[95] Rosset C. Turing-NLG: a 17-billion-parameter language model by Microsoft [Internet]. Redmond: Microsoft; 2020 Feb 13 [cited 2021 Nov 4]. Available from: https://www. microsoft. com/en-us/research/blog/turing-nlg-a-17- billion-parameter-language-model-by-microsoft/.
[96] Kim B, Kim HS, Lee SW, Lee G, Kwak D, Hyeon JD, et al. What changes can large-scale language models bring? Intensive study on HyperCLOVA: billionsscale Korean generative pretrained transformers. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2021 Nov 7–11; online. 2021. p. 3405–24.
[97] Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, et al. mT5: a massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2021 Jun 6 – 11; online. 2021. p. 483–98.
[98] Zhang Z, Gu Y, Han X, Chen S, Xiao C, Sun Z, et al. CPM-2: large-scale costeffective pre-trained language models. 2021. arXiv:2106.10715.
[99] Fedus W, Zoph B, Shazeer N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. 2021. arXiv: 2101.03961.
[100] Wu S, Zhao X, Yu T, Zhang R, Shen C, Liu H, et al. Yuan 1.0: large-scale pretrained language model in zero-shot and few-shot learning. 2021. arXiv: 2110.04725.
[101] Du N, Huang Y, Dai AM, Tong S. Lepikhin D, Xu Y, et al. GLaM: efficient scaling of language models with mixture-of-experts. 2021. arXiv: 2112.06905.
[102] Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, et al. Scaling language models: methods, analysis & insights from training gopher. 2021. arXiv: 2112.11446.
[103] Ding M, Yan Z , Hong W, Zheng W, Zhou C, Yin D, et al. CogView: mastering text-to-image generation via transformers. 2021. arXiv: 2105.13290.
[104] Lin J, Men R, Yang A, Zhou C, Ding M, Zhang Y, et al. M6: a Chinese multimodal pretrainer. 2021. arXiv:2103.00823.
[105] Li W, Gao C, Niu G, Xiao X, Liu H, Liu J, et al. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021); 2021 Aug 1–6; online. 2021. p. 2592–607.
[106] Huo Y, Zhang M, Liu G, Lu H, Gao Y, Yang G, et al. WenLan: bridging vision and language by large-scale multi-modal pre-training. 2021. arXiv:2103.06561.
[107] Zhang H, Yin W, Fang Y, Li L, Duan B, Wu Z, et al. ERNIE-ViLG: unified generative pre-training for bidirectional vision-language generation. 2021. arXiv:2112.15283.
[108] Huang Y, Cheng Y, Bapna A, Firat O, Chen D, Chen M, et al. GPipe: efficient training of giant neural networks using pipeline parallelism. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 9–14; Vancouver, BC, Canada. 2019. p. 103–12.
[109] Shoeybi M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B. Megatron-LM: training multi-billion parameter language models using model parallelism. 2019. arXiv:1909.08053.
[110] Narayanan D, Shoeybi M, Casper J, LeGresley P, Patwary M, Korthikanti V,

et al. Efficient large-scale language model training on GPU clusters using megatron-LM. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC 21); 2021 Nov 14–19; St. Louis, MO, USA; 2021.

[111] Bian Z, Liu H, Wang B, Huang H, Li Y, Wang C, et al. Colossal-AI: a unified deep learning system for large-scale parallel training. 2021. arXiv:2110.14883.

[112] Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. In: Proceedings of the 5th International Conference on Learning Representations (ICLR 2017); 2017 Apr 24–26; Toulon, France; 2017.

[113] Narang S, Diamos G, Elsen E, Micikevicius P, Alben J, Garcia D, et al. Mixed precision training. In: Proceedings of the 6th International Conference on Learning Representations (ICLR 2018); 2018 Apr 30 – May 3; Vancouver, BC, Canada; 2018.

[114] Rajbhandari S, Rasley J, Ruwase O, He Y. ZeRO: memory optimizations toward training trillion parameter models. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC 20); 2020 Nov 9–19; Atlanta, GA, USA; 2020.

[115] Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 Oct 25 – 29; Doha, Qatar. 2014. p. 1746–51.

[116] Hu H, Richardson K, Xu L, Li L, Kübler S, Moss L. OCNLI: original Chinese natural language inference. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; online. 2020. p. 3512–26.

[117] Shao CC, Liu T, Lai Y, Tseng Y, Tsai S. DRCD: a Chinese machine reading comprehension dataset. 2018. arXiv:1806.00920..

[118] Schick T, Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; 2021 Apr 19–23; online. 2021. p. 255–69.

[119] Gray S, Radford A, Kingma DP. GPU kernels for block-sparse weights. 2017. arXiv:1711.09224.

[120] Lin H, Lu Y, Han X, Sun L. Sequence-to-nuggets: nested entity mention detection via anchor-region networks. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019); 2019 Jul 28–Aug 2; Florence, Italy. 2019. p. 5182–92.

[121] Lin Y, Meng Y, Sun X, Han Q, Kuang K, Li J, et al. BertGCN: transductive text classification by combining GCN and BERT. 2021. arXiv: 2105.05727.

[122] Zhang R, Tetreault J. This email could save your life: introducing the task of email subject line generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019); 2019 Jul 28–Aug 2; Florence, Italy. 2019. p. 446–56.

[123] Zhou H, Zheng C, Huang K, Huang M, Zhu X. KdConv: a Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020); 2020 Jul 5–10; online. 2020. p. 7098–108.

[124] Cho J, Seo M, Hajishirzi H, et al. Mixture content selection for diverse sequence generation. 2019. arXiv:1909.01953.

[125] Ribeiro LFR, Zhang Y, Gardent C, Gurevych I. Modeling global and local node contexts for text generation from knowledge graphs. Trans Assoc Comput Linguist 2020;8:589–604.

[126] Zhang Y, Sun S, Galley M, Chen YC, Brockett C, Gao X, et al. DialoGPT: largescale generative pre-training for conversational response generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL 2020); 2020 Jul 5–10; online. 2020. p. 270–8.

[127] Adiwardana D, Luong MT, So DR, Hall J, Fiedel N, Thoppilan R, et al. Towards a human-like open-domain chatbot. 2020. arXiv:2001.09977.

[128] Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y, et al. Recipes for building an open-domain chatbot. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; 2021 Apr 19–23; online. 2021. p. 300–25.

[129] DuerOS [Internet]. Beijing: Baidu; c2017 [cited 2021 Nov 4]. Available from: https://dueros.baidu.com/en/index.html.

[130] Bao S, He H, Wang F, Wu H, Wang H, Wu W, et al. PLATO-2: towards building an open-domain chatbot via curriculum learning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACLIJCNLP 2021); 2021 Aug 1–6; online. 2021. p. 2513–25.

[131] Gunasekara C, Kim S, D'Haro LF, Rastogi A, Chen YN, Eric M, et al. Overview of the ninth dialog system technology challenge: DSTC9. 2020. arXiv:2011.06486.

[132] Bao S, He H, Wang F, Wu H, Wang H, Wu W, et al. PLATO-XL: exploring the large-scale pre-training of dialogue generation. 2021. arXiv:2109.09519.

[133] Wang Y, Ke P, Zheng Y, Huang K, Jiang Y, Zhu X, et al. A large-scale Chinese short-text conversation dataset. In: Proceedings of the 9th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC 2020); 2020 Oct 14–18; Zhengzhou, China. 2020. p. 91–103.

[134] Qi W, Gong Y, Yan Y, Xu C, Yao B, Zhou B, et al. ProphetNet-X: large-scale pretraining models for English, Chinese, multi-lingual, dialog, and code generation. 2021. arXiv:2104.08006.

[135] Zhou H, Ke P, Zhang Z, Gu Y, Zheng Y, Zheng C, et al. EVA: an open-domain Chinese dialogue system with large-scale generative pre-training. 2021. arXiv: 2108.01547.

[136] Vinyals O, Le Q. A neural conversational model. 2015. arXiv:1506.05869.

[137] Serban I, Sordoni A, Bengio Y, Courville A, Pineau J. Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence; 2016 Feb 12–17; Phoenix, AZ, USA. Palo Alto: AAAI Press; 2016. p. 3776–83.

[138] Worswick S. "Mitsuku wins loebner prize 2018!" [Internet]. Medium; 2018 Sep 13 [cited 2021 Nov 4]. Available from: https://medium.com/ pandorabots-blog/mitsuku-wins-loebner-prize-2018-3e8d98c5f2a7.

[139] Zhou L, Gao J, Li D, Shum HY. The design and implementation of XiaoIce, an empathetic social chatbot. Comput Linguist 2020;46(1):53–93.

[140] Xin J, Tang R, Lee J, Yu Y, Lin J. DeeBERT: dynamic early exiting for accelerating BERT inference. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020); 2020 Jul 5 – 10; online. 2020. p. 2246–51.

[141] Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, Laroussilhe QD, Gesmundo A, et al. Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning (ICML 2019); 2019 Jun 9–15; Long Beach, CA, USA. 2019. p. 2790–9.

[142] Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021); 2021 Aug 1 – 6; online. 2021. p. 4582–97.

[143] Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021); 2021 Aug 1 – 6; online. 2021. p. 3816–30.

[144] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2021 Nov 7 – 11; online. 2021. p. 3045–59.

[145] Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, et al. GPT understands, too. 2021. arXiv:2103.10385.

[146] Han X, Zhao W, Ding N, Liu Z, Sun M. PTR: prompt tuning with rules for text classification. 2021. arXiv:2105.11259.

[147] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017. arXiv:1702.08608.

[148] Wallace E, Feng S, Kandpal N, Gardner M, Singh S. Universal adversarial triggers for attacking and analyzing NLP. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP); 2019 Nov 3–7; HongKong, China. 2019. p. 2153–62.

[149] Fort K, YesCouillault A., care! we Results of the ethics and natural language processing surveys. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016); 2016 May 23–28; Portorož, Slovenia. 2016. p. 1593–600.

[150] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. In: Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014); 2014 Apr 14–16; Banff, AB, Canada; 2014.

[151] Hewitt J, Manning CD. A structural probe for finding syntax in word representations. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human LanguageTechnologies; 2019Jun2–7;Minneapolis,MN,USA.2019. p. 4129–38.

[152] Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019); 2019 Jul 28 – Aug 2; Florence, Italy.

2019. p. 3651–7.
[153] Linzen T, Dupoux E, Goldberg Y. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Trans Assoc Comput Linguist 2016;4:521–35.
[154] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12–17; DiegoSan, CA, USA. 2016. p. 1135–44.
[155] Davison J, Feldman J, Rush AM. Commonsense knowledge mining from pretrained models. In: 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 Nov 3–7; HongKong, China. 2019. p. 1173–8.
[156] Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y, et al. Language models as knowledge bases? In: 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 Nov 3–7; HongKong, China. 2019. p. 2463–73.
[157] Talmor A, Elazar Y, Goldberg Y, Berant J. oLMpics-on what language model pre-training captures. Trans Assoc Comput Linguist 2020;8:743–58.
[158] Morris JX, Lifland E, Yoo JY, Grigsby J, Jin D, Qi Y. TextAttack: a framework for adversarial attacks, data augmentation, and adversarial training in NLP. 2020. arXiv:2005.05909.
[159] Jia J, Liu Y, Gong NZ. BadEncoder: backdoor attacks to pre-trained encoders in self-supervised learning. 2021. arXiv:2108.00352.
[160] Devlin J. Google-research/bert [Internet]. GitHub; 2018 Oct 11 [cited 2021 Nov 4]. Available from: https://github.com/google-research/bert.
[161] Baidu Ernie Team. Paddlepaddle/ernie [Internet]. GitHub; 2019 Apr 19 [cited 2021 Nov 4]. Available from: https://github.com/PaddlePaddle/ERNIE.
[162] Huawei. Pcl-platform. intelligence/pangu-alpha [Internet]. San Francisco: OpenAI; 2021 Apr 26 [cited 2021 Nov 4]. Available from: https://git.openi. org.cn/PCL-Platform.Intelligence/PanGu-Alpha.
[163] Ding S, Shang J, Wang S, Sun Y, Tian H, Wu H, et al. ERNIE-Doc: a retrospective long-document modeling transformer. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021); 2021 Aug 1–6; online. 2021. p. 2914–27.
[164] Huggingface [Internet]. Hugging Face; 2020 Apr 26 [cited 2021 Nov 4]. Available from: https://huggingface.co.
[165] Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, et al. FAIRSEQ: a fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Demonstrations); 2019 Jun 2–7; Minneapolis, MN, USA. 2019. p. 48–53.
[166] Baidu PaddlePaddle Team. Paddlepaddle/paddlenlp [Internet]. GitHub; 2020 Nov 16 [cited 2021 Nov 4]. Available from: https://github. com/PaddlePaddle/PaddleNLP.
[167] Wenxin ernie [Internet]. Beijing: Baidu; c2021 [cited 2021 Nov 4]. Available from: https://wenxin.baidu.com.
[168] Alibaba Damo Academy. AliceMind [Internet]. Aliyuncs; c2021 [cited 2021 Nov 4]. Available from: https://alicemind.aliyuncs.com.
[169] Openai API [Internet]. San Francisco: OpenAI; c2021 [cited 2021 Nov 4]. Available from: https://openai.com/api.
[170] Garg Y, Chatterjee N. Sentiment analysis of twitter feeds. In: Proceedings of the 3rd International Conference on Big Data Analytics (BDA 2014); 2014 Dec 20–23; New Delhi, India. 2014. p. 33–52.
[171] AlQahtani ASM. Product sentiment analysis for amazon reviews. Int J Comput Sci Inf Technol 2021;13(3):15–30.
[172] Singh M, Jakhar AK, Pandey S. Sentiment analysis on the impact of coronavirus in social life using the BERT model. Soc Netw Anal Min 2021;11:33.
[173] Chen Z, Sokolova M. Sentiment analysis of the COVID-related r/Depression posts. 2021. arXiv:2108.06215.
[174] Liu Y, Liu J, Chen L, Lu Y, Feng S, Feng Z, et al. ERNIE-SPARSE: learning hierarchical efficient transformer through regularized self-attention. 2022. arXiv:2203.12276.
[175] Jwa H, Oh D, Park K, Kang JM, Lim H. exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (BERT). Appl Sci 2019;9(19):4062.
[176] Soares LB, FitzGerald N, Ling J, Kwiatkowski T. Matching the blanks: distributional similarity for relation learning. 2019. arXiv:1906.03158.
[177] Wang Z, Xu Y, Cui L, Shang J, Wei F. LayoutReader: pre-training of text and layout for reading order detection. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2021 Nov 7–11; online. 2021. p. 4735–44.
[178] gpt-2-for-the-advertising-industry [Internet]. San Francisco: OpenAI; 2017 Aug 1 [cited 2021 Nov 4]. Available from: https://www.narrativa.com/gpt-2- for-the-advertising-industry.
[179] Agarwal R, Kann K. Acrostic poem generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; online. 2020. p. 1230–40.
[180] Lee DH, Hu Z, Lee RKW. Improving text auto-completion with next phrase prediction. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2021 Nov 7 – 11; online. 2021. p. 4434–8.
[181] Mukherjee S, Mukherjee S, Hasegawa M, Awadallah AH, White R. Smart todo: automatic generation of to-do items from emails. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020); 2020 Jul 5–10; online. 2020. p. 8680–9.
[182] What are Alexa Built-in Devices? [Internet]. Seattle: Amazon; c2010 – 2023 [cited 2021 Nov 4]. Available from: https://developer. amazon. com/alexavoice-service.
[183] Mari A. Voice commerce: understanding shopping-related voice assistants and their effect on brands. In: Proceedings of the International Media Management Academic Association Annual Conference; 2019 Oct 4–6; Doha, Qatar; 2019.
[184] Google assistant update speech recognition name pronunciation BERT smart speakers [Internet]. NDTV; 2021 Apr 29 [cited 2021 Nov 4]. Available from: https://gadgets. ndtv. com/apps/news/google-assistant-update-speechrecognition-name-pronunciation-bert-smart-speak.
[185] Vincent J. The future of AI is a conversation with a computer [Internet]. New York City: The Verge; 2021 Nov 1 [cited 2021 Nov 4]. Available from: https:// www. theverge. com/22734662/ai-language-artificial-intelligence-futuremodels-gpt-3-limitations-bias/.
[186] Meet the AI powering today's smartest smartphones [Internet]. San Francisco: Wired; 2017 Aug 1 [cited 2021 Nov 4]. Available from: https:// www. wired. com/sponsored/story/meet-the-ai-powering-todays-smartestsmartphones.
[187] Nayak P. Understanding searches better than ever before [Internet]. Google; [cited 2021 Nov 4]. Available from: https://blog.google/products/search/ search-language-understanding-bert/.
[188] Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, et al. ERNIE 2.0: a continual pretraining framework for language understanding. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence; 2020 Feb 7 – 12; New York City, NY, USA. Palo Alto: AAAI Press; 2020. p. 8968–75.
[189] Liu Y, Lu W, Cheng S, Shi D, Wang S, Cheng Z, et al. Pre-trained language model for web-scale retrieval in Baidu Search. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 21); 2021 Aug 14–18; online. 2021. p. 3365–75.
[190] Huang JT, Sharma A, Sun S, Xia L, Zhang D, Pronin P, et al. Embedding-based retrieval in Facebook Search. In: Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 20); 2020 Jul 6–10; online. 2020. p. 2553–61.
[191] Yu P, Fei H, Li P. Cross-lingual language model pretraining for retrieval. In: Proceedings of the Web Conference; 2021 Apr 19–23; online. 2021. p.1029– 39.
[192] Ni M, Huang H, Su L, Cui E, Bharti T, Wang L, et al. M3P: learning universal representations via multitask multilingual multimodal pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 19–25; online. 2021. p. 3977–86.
[193] Sanh V, Debut L, Chaumond J, DistilBERTWolf T., a distilled version of BERT: smaller, faster, cheaper and lighter. 2019. arXiv:1910.01108.
[194] Gordon MA, Duh K, Andrews N. Compressing BERT: studying the effects of weight pruning on transfer learning. In: Proceedings of the 5th Workshop on Representation Learning for NLP; 2020 Jul 9; Seattle, WA, USA. 2020. p. 143–55.
[195] Kim S, Gholami A, Yao Z, Mahoney MW, Keutzer K. I-BERT: integer-only BERT quantization. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021); 2021 Jul 18–24; online. 2021. p. 5506–18.