**Pre-trained Models for Natural Language Processing: A Survey**

近年来，预训练模型（PTMs）的出现将自然语言处理（NLP）带入了一个新的时代。在本次调查中，我们对NLP的PTMs进行了全面的回顾。本文首先简要介绍了语言表征学习及其研究进展。然后，我们从四个不同的角度根据分类法对现有的ptm进行了系统的分类。接下来，我们将描述如何使PTMs的知识适应下游任务。最后，对PTMs未来的研究方向进行了展望。本调查旨在为理解、使用和开发各种NLP任务的PTMs提供实践指导。

深度学习、神经网络、自然语言处理、预训练模型、分布式表示、单词嵌入、自监督学习、语言建模

**1 Introduction**

随着深度学习的发展，各种神经网络被广泛应用于解决自然语言处理（NLP）任务，如卷积神经网络（CNNs）[80，86，49]，递归神经网络（RNNs）[173，109]，基于图的神经网络（GNNs）[159，174，124]和注意机制[7184]。这些神经模型的优点之一是能够缓解特征工程问题。非神经NLP方法通常严重依赖于离散的手工特征，而神经方法通常使用低维密集向量（aka）分布式表示）隐式表示语言的语法或语义特征。这些表示在特定的NLP任务中学习。因此，神经方法使人们更容易开发各种NLP系统。

尽管神经模型在NLP任务中取得了成功，但与计算机视觉（CV）领域相比，其性能改进可能不太显著。主要原因是大多数受监控的NLP任务的当前数据集非常小（机器翻译除外）。深部神经网络通常具有大量的参数，使得它们在这些小的训练数据上过于拟合，在实际应用中不能很好地推广。因此，许多NLP任务的早期神经模型比较浅，通常只有1~3个神经层。

近年来，大量的研究表明，在大型语料库上建立的预训练模型（PTMs）可以学习通用语言表示，这有利于后续的NLP任务，避免从头开始训练新模型。随着计算能力的发展，深度模型（Transformer[184]）的出现，以及训练技能的不断提高，PTMs的体系结构也由浅到深的发展。第一代PTMs旨在学习好的单词嵌入。由于下游任务不再需要这些模型本身，对于计算科学来说通常是非常肤浅的，比如Skip Gram[129]和GloVe[133]。尽管这些预训练的嵌入可以捕获单词的语义，但它们是上下文无关的，不能捕获上下文中的高级概念，如多义消歧、句法结构、语义角色、回指等。第二代PTMs专注于学习上下文单词嵌入，如CoVe[126]、ELMo[135]、OpenAI-GPT[142]和BERT[36]。这些学习过的编码器仍然需要通过下游任务在上下文中表示单词。此外，还提出了不同的训练前任务，以学习不同目的的PTMs。

本次调查的贡献可概括如下：

1.全面审查。我们对NLP的PTMs进行了全面的回顾，包括背景知识、模型架构、训练前任务、各种扩展、自适应方法和应用。

2．新分类法。我们提出了一个用于NLP的PTMs分类法，它从四个不同的角度对现有PTMs进行分类：1）表示类型，2）模型架构；3）预训练任务类型；4）特定场景类型的扩展。

3.资源丰富。我们收集了大量关于PTMs的资源，包括PTMs的开源实现、可视化工具、语料库和纸质清单。

4.未来的方向。我们讨论并分析了现有PTMs的局限性。此外，我们还提出了未来可能的研究方向。

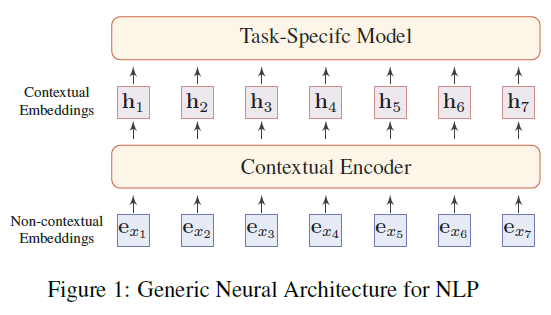
调查的其余部分安排如下。第2节概述了PTMs的背景概念和常用符号。第3节简要概述了PTMs，并阐明了PTMs的分类。第4节提供了PTM的扩展。第5节讨论如何将PTMs的知识转移到下游任务。第6节给出了PTMs的相关资源。第7节展示了跨各种NLP任务的应用程序集合。第8节讨论了当前的挑战并提出了未来的方向。第九节对论文进行了总结。

**2背景**

**2.1语言表征学习**

如Bengio等人所建议。[13] ，一个好的表示应该表示通用的优先级，这些优先级不是任务特定的，但是对于学习机器解决人工智能任务可能有用。在语言方面，一个好的表达应该捕捉隐藏在文本数据中的隐含的语言规则和常识，如词汇意义、句法结构、语义角色，甚至语用学。

分布式表示的核心思想是用低维实值向量描述文本的意义。而向量的每一个维度都没有对应的意义，而整体则代表一个具体的概念。图1展示了NLP的通用神经体系结构。词的嵌入有两种：非上下文嵌入和上下文嵌入。它们之间的区别在于是否根据单词出现的上下文动态地改变它的嵌入。



非上下文嵌入表示语言的第一步是将离散的语言符号映射到分布式嵌入空间中。形式上，对于vocabulary V中的每个单词（或子单词）x，我们使用查找表将其映射到向量，其中De是表示令牌嵌入维度的超参数。这些嵌入是在任务数据和其他模型参数上训练的。

这种嵌入有两个主要限制。第一个问题是嵌入是静态的。不管上下文如何，单词的嵌入总是一样的。因此，这些非语境嵌入无法模拟多义词。第二个问题是词汇量不足的问题。为了解决这个问题，字符级的单词表示或子单词表示在许多NLP任务中被广泛使用，比如CharCNN[87]、FastText[14]和Byte Pair Encoding（BPE）[154]。

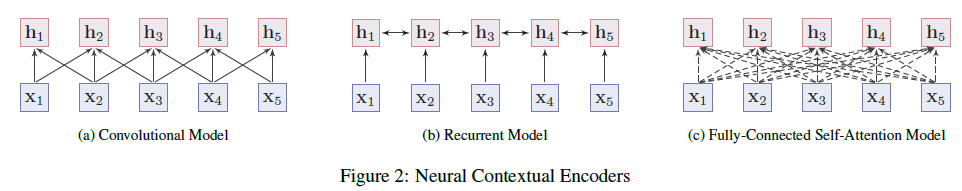
语境嵌入要解决多义词问题和词的语境依赖性，需要区分不同语境中词的语义。给定文本x1；x2……xT，其中每个标记是一个单词或子单词，xT的上下文表示依赖于整个文本。



其中fenc（）是神经编码器，如第2.2节所述，ht被称为上下文嵌入或动态嵌入由于中包含上下文信息，因此标记xt。

**2.2神经上下文编码器**

大多数神经上下文编码器可分为两类：序列模型和基于图的模型。图2说明了这些模型的体系结构。



**2.2.1 Sequence Models**

序列模型通常按顺序捕获单词的本地上下文。

卷积模型 卷积模型将单词嵌入到输入句子中，并通过卷积操作聚合来自其邻居的局部信息来捕获单词的含义[86]。

递归模型 递归模型捕获具有短记忆的单词的上下文表示，如LSTMs[64]和GRUs[23]。在实践中，双向LSTMs或GRUs用于从词的两边收集信息，但其性能往往受到长期依赖问题的影响。

**2.2.2 Non-Sequence Models**

非序列模型通过预先定义的词间树或图结构（如句法结构或语义关系）来学习上下文表示。一些流行的非序列模型包括递归NN[159]、TreeLSTM[174222]和GCN[88]。

尽管语言感知的图结构可以提供有用的归纳偏差，但是如何建立一个好的图结构也是一个具有挑战性的问题。此外，该结构在很大程度上依赖于专家知识或外部NLP工具，如依赖性解析器。

全连通自注意模型 在实际应用中，一种更为直接的方法是使用全连通图对每两个词的关系进行建模，让模型自己学习结构。通常，连接权值由自关注机制动态计算，隐式地表示单词之间的连接。全连接自关注模型的成功实例是Transformer[184]，它还需要其他补充模块，如位置嵌入、层规范化、剩余连接和位置前馈网络（FFN）层。

**2.2.3 Analysis**

序列模型学习具有局部偏误的词的上下文表示，难以捕捉词与词之间的长期交互作用。然而，序列模型通常易于训练，并且对于各种NLP任务都能获得良好的结果。

相反，作为一个实例化的全连接自注意模型，Transformer可以直接对序列中的每两个词之间的依赖关系进行建模，这一模型更强大，更适合于对语言的长距离依赖关系进行建模。

然而，由于Transformer结构重，模型偏差小，通常需要一个大的训练语料库，并且容易在小的或中等大小的数据集上过度拟合[142，53]。

目前，Transformer以其强大的容量成为PTMs的主流架构。

**2.3 Why Pre-training?**

随着深度学习的发展，模型参数的数目迅速增加。需要更大的数据集来充分训练模型参数并防止过度拟合。

然而，对于大多数NLP任务来说，构建大规模的标记数据集是一个巨大的挑战，因为注释成本非常昂贵，特别是对于语法和语义相关的任务。

相比之下，大规模的未标记语料库相对容易构建。为了利用巨大的未标记文本数据，我们可以首先从中学习一个好的表示，然后将这些表示用于其他任务。最近的研究表明，在大型未标注语料库上，通过从PTMs中提取表示，许多NLP任务的性能都有了显著提高。

预培训的优点可以总结如下：

一。在庞大的文本语料库上进行预训练可以学习通用的语言表示和帮助下游任务。

二。预训练可以提供更好的模型初始化，这通常会导致更好的泛化性能，并加快目标任务的收敛速度。

三。训练前可以看作是一种正则化，以避免对小数据的过度拟合[43]。

**2.4 A Brief History of PTMs for NLP**

预训练一直是学习深层神经网络参数的一种有效策略，然后对下游任务进行微调。早在2006年，深度学习的突破就出现在贪婪的分层无监督预训练和监督的微调之后[62]。在CV中，在庞大的ImageNet语料库上对模型进行预训练，然后在较小的数据上对不同的任务进行进一步的微调已经成为实际。这比随机初始化要好得多，因为模型学习一般的图像特征，然后可以用于各种视觉任务。

在NLP中，大规模语料库上的PTMs也被证明有助于从浅词嵌入到深神经模型的NLP下游任务。

**2.4.1第一代PTMs：预先训练的单词嵌入**

将单词表示为密集向量有很长的历史[60]。在神经网络语言模型（NNLM）的开创性工作[12]中引入了“现代”字嵌入。Collobert等人[26]结果表明，在未标记的数据上嵌入预先训练的单词可以显著提高许多NLP任务。为了解决计算复杂度问题，他们用成对排序任务来学习单词嵌入，而不是语言建模。他们的工作是首次尝试从未标记的数据中获取对其他任务有用的通用单词嵌入。Mikolov等人。[129]表明不需要深层神经网络来构建良好的单词嵌入。他们提出了两种浅层次的体系结构：连续的单词包（CBOW）和跳过语法（SG）模型。尽管简单，他们仍然可以学习高质量的单词嵌入来捕捉单词之间潜在的句法和语义相似性。Word2vec是这些模型中最流行的实现之一，它使预先训练的word嵌入可用于NLP中的不同任务。此外，GloVe[133]也是一个广泛使用的获取预训练单词嵌入的模型，它是通过从一个大的语料库中的全局单词-单词协同出现统计来计算的。

尽管经过预训练的单词嵌入在NLP任务中已经被证明是有效的，但是它们是上下文无关的，并且大多是由浅层模型训练的。当用于下游任务时，整个模型的其余部分仍然需要从头学习。

在同一时期，许多研究者还试图学习段落、句子或文档的嵌入，如段落向量[96]、跳过思想向量[89]、上下文向量[127]。与它们的现代继承者不同，这些句子嵌入模型试图将输入的句子编码为固定维向量表示，而不是每个标记的上下文表示。

**2.4.2第二代PTMs：预先训练的上下文编码器**

由于大多数NLP任务都超出了词的层次，因此在句子层次或更高层次上对神经编码器进行预训练是很自然的。神经编码器的输出向量也被称为上下文单词嵌入，因为它们根据上下文来表示单词语义。

Dai和Le[30]提出了NLP的PTM的第一个成功实例。他们使用语言模型（LM）或序列自动编码器对LSTMs进行初始化，发现预训练可以提高LSTMs在许多文本分类任务中的训练和泛化能力。Liu等人。[109]使用LM对共享LSTM编码器进行了预训练，并在多任务学习（MTL）框架下对其进行了微调。他们发现预训练和微调可以进一步提高多个文本分类任务的MTL性能。Ramachandran等人。[146]发现Seq2Seq模型可以通过无监督的预训练得到显著改善。编码器和解码器的权值由两种语言模型的预先训练权值初始化，然后用标记数据进行微调。除了用LM对上下文编码器进行预训练外，McCann等。[126]使用机器翻译（MT）从注意序列到序列模型预先训练深层LSTM编码器。通过预先训练的编码器输出的上下文向量（CoVe）可以提高各种常见NLP任务的性能。

由于这些前体PTMs，现代PTMs通常使用更大规模的语料库、更强大或更深层次的体系结构（例如Transformer）和新的预训练任务进行训练。

彼得斯等人。[135]带有双向语言模型（BiLM）的预训练2层LSTM编码器，由前向LM和后向LM组成。由预先训练的BiLM，ELMo（语言模型中的嵌入）输出的上下文表示，显示出对广泛的NLP任务带来了很大的改进。Akbik等人。[1] 用预先训练过的字符级LM嵌入上下文字符串捕获单词的意思。然而，这两个ptm通常被用作特征抽取器来生成上下文单词嵌入，并将其输入到下游任务的主模型中。它们的参数是固定的，其余的参数of主要的模型仍然是从零开始训练的。ULMFiT（Universal Language Model Fine tuning）[67]试图对文本分类（TC）的预训练LM进行微调，并在6个广泛使用的TC数据集上取得了最新的结果。ULMFiT包括三个阶段：1）一般域数据的预训练LM；2）目标数据的微调LM；3）目标任务的微调。ULMFiT还研究了一些有效的微调策略，包括区分微调、倾斜三角学习率和逐步解冻。

最近，非常深入的ptm在学习通用语言表示方面显示出了强大的能力：例如OpenAI-GPT（Generative Pre-training）[142]和BERT（Transformer的双向编码器表示）[36]。除了LM之外，还提出了越来越多的自监督任务（见3.1节），以使PTMs从大规模文本语料库中获取更多的知识。自ULMFiT和BERT以来，微调已经成为使PTMs适应下游任务的主流方法。

**3 Overview of PTMs**

ptm之间的主要区别在于上下文编码器的使用、预训练任务和目的。我们在第2.2节中简要介绍了上下文编码器的体系结构。在本节中，我们将重点介绍预培训任务，并给出PTMs的分类。

**3.1 Pre-training Tasks**

训练前的任务对学习语言的普遍表征至关重要。通常，这些训练前任务应该具有挑战性，并有大量的训练数据。在本节中，我们将训练前的任务归纳为三类：监督学习、非监督学习和自我监督学习。

一。监督学习（SL）是学习一个函数，该函数基于由输入输出对组成的训练数据将输入映射到输出。

二。无监督学习（UL）是从未标记的数据中发现一些内在的知识，如聚类、密度、潜在表示等。

三。自监督学习（SSL）是有监督学习和无监督学习的结合。SSL的学习模式与监督学习完全相同，但训练数据的标签是自动生成的。SSL的关键思想是以某种形式预测来自其他部分的输入的任何部分。例如，蒙面语言模型（MLM）是一个自我监督的任务，它试图预测给定剩余单词的句子中的蒙面单词。

在CV中，许多ptm是在像ImageNet这样的大型监督训练集上训练的。然而，在NLP中，大多数受监控任务的数据集都不足以训练出一个好的PTM。

唯一的例外是机器翻译（MT）。一个大规模的机器翻译数据集，WMT 2017，由超过700万个句子对组成。此外，MT是NLP中最具挑战性的任务之一，对其进行预训练的编码器可以使NLP下游的各种任务受益。作为一个成功的PTM，CoVe[126]是一个针对MT任务预先训练的编码器，它改进了许多常见的NLP任务：情绪分析（SST，IMDb）、问题分类（TREC）、蕴涵（SNLI）和问题回答（SQuAD）。

在这一部分中，我们将介绍一些在现有PTMs中广泛使用的训练前任务。我们可以把这些任务看作是自我监督的学习。表1还总结了它们的损失函数。

**3.1.1 Language Modeling (LM)**

NLP中最常见的无监督任务是概率语言建模（LM），它是一个经典的概率密度估计问题。虽然LM是一个普遍的概念，但在实践中，LM常常特别指自回归LM或单向LM。

给定一个文本序列x1:T=[x1；x2； ；xT]，其联合概率p（x1:T）可以分解为



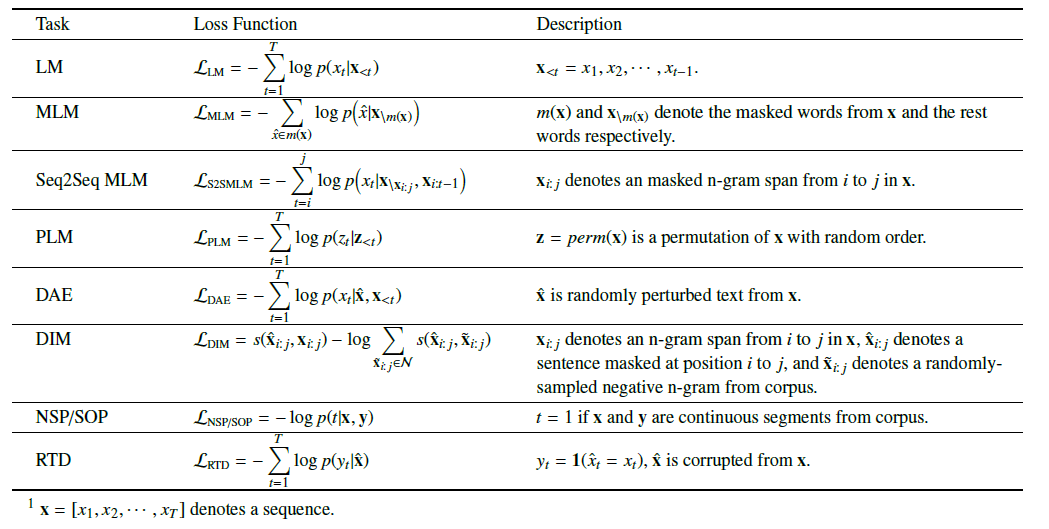
其中x0是表示序列开始的特殊标记。

条件概率可以由给定语言上下文的词汇上的概率分布来建模。文中用神经编码器fenc（）对上下文进行建模，条件概率为



其中gLM（）是预测层。

给定一个庞大的语料库，我们可以用最大似然估计（MLE）对整个网络进行训练。单向LM的一个缺点是，每个令牌的表示只对向左上下文令牌及其本身进行编码。然而，更好的文本上下文表示应该从两个方向编码上下文信息。一种改进的解决方案是双向LM（BiLM），它由两个单向LM组成：从左到右的前LM和从右到左的后LM。对于BiLM，Baevski等人[6] 提出了前塔操作左向右LM，后塔操作右向左LM的双塔模型。



**3.1.2 Masked Language Modeling (MLM)**

蒙面语言建模（Masked language modeling，MLM）是Taylor[178]在文献中首次提出的一种完形填空任务。Devlin等人[36]将此任务改编为一个新的训练前任务，以克服标准单向LM的缺点。松散地说，MLM首先从输入句子中屏蔽掉一些代币，然后训练模型，用其余的代币预测被屏蔽的代币。但是，这种预训练方法将在预训练阶段和微调阶段之间创建不匹配，因为掩码令牌不会在微调阶段出现。根据经验，为了解决这个问题，Devlin等人[36]在80%的时间内使用特殊的[MASK]标记，在10%的时间内使用随机标记，在10%的时间内使用原始标记执行掩蔽。

序列MLM（Seq2Seq-MLM）通常作为分类问题来解决。我们将屏蔽序列输入到一个神经编码器，该编码器的输出向量进一步输入到softmax分类器中以预测屏蔽令牌。或者，我们可以使用编码器-解码器（又名Seq2Seq）用于MLM的体系结构，在该体系结构中，编码器被馈送一个屏蔽序列，解码器以自动回归的方式依次生成屏蔽令牌。我们称这种MLM为顺序MLM（Seq2SeqMLM），在MASS[160]和T5[144]中使用。Seq2Seq MLM有利于Seq2Seq式的下游任务，如问答、总结、机器翻译等。

增强型屏蔽语言建模（E-MLM）同时，有多个研究提出了不同增强型的MLM以进一步改进BERT。RoBERTa[117]通过动态掩蔽改进了BERT，而不是静态掩蔽。

UniLM[39，8]将掩码预测的任务扩展到三种语言建模任务：单向、双向和序列到序列预测。XLM[27]对并行双语句子对的连接执行MLM，称为翻译语言建模（Translation Language Modeling，TLM）。Span BERT[77]用随机邻接词掩蔽和跨边界目标（SBO）代替MLM，将结构信息集成到预训练中，这就要求系统基于跨边界预测掩蔽跨。此外，StructBERT[193]还引入了跨阶恢复任务，以进一步合并语言结构。

另一个丰富MLM 的方法是吸收外部知识（见第4.1节）。

**3.1.3 Permuted Language Modeling (PLM)**

尽管MLM任务在岗前培训中得到了广泛的应用，但杨等表示[209]声称，当模型应用于下游任务时，MLM预训练中使用的一些特殊标记，如[MASK]不存在，导致预训练和微调之间存在差距。为了克服这个问题，排列语言建模（PLM）[209]是一个预训练目标以取代MLM。简言之，PLM是一个关于输入序列随机排列的语言建模任务。置换是从所有可能的置换中随机抽取的。然后选择置换序列中的部分标记作为目标，并根据剩余标记和目标的自然位置训练模型来预测这些目标。注意，这种排列不影响序列的自然位置，只定义了标记预测的顺序。在实际应用中，由于收敛速度慢，只能预测置换序列中的最后几个令牌。针对目标感知表示，引入了一种特殊的双流自关注。

**3.1.4去噪自动编码器（DAE）**

去噪自动编码器（DAE）接受部分损坏的输入，旨在恢复原始的未失真输入。特定于语言，一个序列到序列的模型，比如标准的转换器，被用来重建原始文本。损坏文本[100]有几种方法：

（1）标记屏蔽：从输入中随机采样标记，并用[MASK]元素替换它们。

（2） 令牌删除：从输入中随机删除令牌。与标记屏蔽不同，模型需要确定丢失输入的位置。

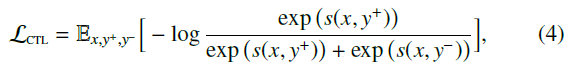
（3） 文本填充：与SpanBERT一样，对多个文本跨距进行采样，并用单个[MASK]标记替换。每个跨距长度从泊松分布（=3）中得出。模型需要预测一个跨度中缺少多少令牌。

（4） 句子排列：根据句号将文档分成句子，并将这些句子按随机顺序排列。

（5） 文档旋转：随机均匀地选择一个标记，并旋转文档，使其以该标记开头。模型需要确定文档的实际起始位置。

**3.1.5对比学习（CTL）**

对比学习[153]假设一些观察到的文本对在语义上比随机抽取的文本更相似。学习文本对（x；y）的得分函数s（x；y）以最小化目标函数：



其中（x；y+）是相似的一对，且可能与x不同。y+和通常称为正样本和负样本。分数函数s（x；y）通常由可学习神经编码器以两种方式计算：或。

CTL背后的理念是“通过比较学习”。与LM相比，CTL通常具有较小的计算复杂度，因此是PTMs理想的替代训练准则。

Collobert等人。[26]提出了区分真假短语的成对排序任务。该模型需要预测一个合法短语比一个不正确短语的得分更高，而不正确短语是通过将其中心词替换为一个随机词获得的。

Mnih和Kavukcuoglu[131]利用噪声对比估计（NCE）[55]对单词嵌入进行训练，训练二元分类器来区分真假样本。在著名的word2vec嵌入中也使用了NCE的思想[129]。

在下面的段落中，我们将简要介绍一些最近提出的CTL任务。

Deep InfoMax（DIM） deepinfomax（DIM）[63]最初是针对图像提出的，它通过最大化图像表示和图像局部区域之间的相互信息来提高表示的质量。

孔等人[90]将DIM应用于语言表征学习。序列x的全局表示被定义为上下文编码器fenc（x）输出的第一个标记（假设是句子符号的特殊开头）的隐藏状态。席席席的目的是为FEC（X:J）t FEC（席席：J）分配比FEC（X:J）t FEC（X:J）更高的分数，其中表示n- GRAM2跨度，从I到j在x，x X:j表示在位置I到j的掩蔽的句子，和XI:J表示从语料库中随机抽取的负n元。

替换的令牌检测（RTD） 替换的令牌检测（RTD）与NCE相同，但根据令牌周围的上下文预测是否替换该令牌。

负抽样CBOW（CBOW-NS）[129]可以看作是RTD的一个简单版本，其中负抽样是从具有简单建议分布的词汇中随机抽取的。

ELECTRA[24]通过使用生成器替换序列的一些标记来改进RTD。生成器G和鉴别器D按照两个步骤进行训练：（1）只训练具有MLM任务的生成器执行n1步；（2）使用生成器的权重初始化鉴别器的权重。然后对鉴别器进行n2步的鉴别力训练，保持G不变。这里，判别任务指示是否将输入令牌替换为G。生成器是在预训练后抛出的，只有鉴别器会在下游任务上进行微调。

RTD也是解决失配问题的另一种方法。网络在训练前会看到[蒙版]，但在对下游任务进行微调时不会看到。

类似地，WKLM[202]替换实体级别的单词，而不是标记级别的单词。具体来说，WKLM用相同类型的其他实体的名称替换实体提及，并训练模型以区分实体是否已被替换。

下一句预测 标点是文本数据的自然分隔符。因此，利用这些方法构建预训练方法是合理的。下一句预测（NSP）[36]就是一个很好的例子。顾名思义，NSP训练模型是为了区分两个输入句子是否是训练语料中的连续段。具体来说，在为每个预训练例子选择句子对时，50%的时间，第二个句子是第一个句子的实际下一个句子，50%的时间，它是来自语料库的随机句子。通过这样做，它能够教会模型理解两个输入语句之间的关系，从而有利于对这些信息敏感的下游任务，如问答和自然语言推理。

然而，NSP任务的必要性受到了后续工作的质疑[77209117 93]。Yang等人。[209]发现NSP任务的影响不可靠，而Joshi等人。[77]发现无NSP丢失的单句训练优于有NSP丢失的句子对训练。此外，Liu等人。[117]对NSP任务进行了进一步的分析，结果表明，当使用单个文档中的文本块进行训练时，删除NSP丢失匹配项或稍微提高下游任务的性能。

为了更好地模拟句子间的连贯性，ALBERT[93]用句子顺序预测（SOP）损失代替了NSP损失。正如Lan等人推测的那样。[93]，NSP将主题预测和一致性预测融合在一个任务中。因此，模型可以仅仅依靠简单的任务，即主题预测来进行预测。与NSP不同，SOP使用同一文档中的两个连续段作为正示例，使用相同的两个连续段作为负示例，但它们的顺序交换为负示例。因此，ALBERT在各种下游任务上始终优于BERT。

StructBERT[193]和BERTje[33]也将SOP作为他们的自我监督学习任务。

**3.1.6其他**

除上述任务外，还有许多其他辅助性的培训前任务，旨在整合事实知识（见第4.1节）、改进跨语言任务（见第4.2节）、多模态应用（见第4.3节）或其他特定任务（见第4.4节）。

**3.2 PTMs的分类**

为了阐明NLP中现有PTMs之间的关系，我们建立了PTMs的分类法，从四个不同的角度对现有PTMs进行分类：

1。表示类型：根据用于下游任务的表示，我们可以将ptm分为非上下文模型和上下文模型。

2。架构：PTMs使用的主干网，包括LSTM、Transformer编码器、Transformer解码器和全Transformer架构。“Transformer”是指标准的编码器-解码器架构。“Transformer编码器”和“Transformer解码器”分别指标准Transformer架构的编码器和解码器部分。它们的区别在于，解码器部分使用带有三角形矩阵的掩蔽自我注意来防止令牌进入其未来（右）位置。

三。培训前任务类型：PTMs使用的培训前任务类型。我们已经在第3.1节中讨论过了。

四。扩展：为各种场景设计的PTMs，包括知识丰富的PTMs、多语言或语言专用PTMs、多模型PTMs、领域专用PTMs和压缩PTMs。我们将在第4节中特别介绍这些扩展。

图3显示了分类法以及一些相应的代表性ptm。此外，表2更详细地区分了一些有代表性的ptm。

**3.3模型分析**

由于PTMs的巨大成功，了解PTMs所获取的知识种类以及如何从中获取知识是非常重要的。在分析语言知识和世界知识的文献中，有大量的文献存储在预先训练好的非语境和语境嵌入中。

**3.3.1非上下文嵌入**

首先探讨了静态词嵌入的知识种类。Mikolov等人[130]发现用神经网络语言模型学习的单词表示法能够捕捉语言中的语言规律，单词之间的关系可以用一个特定于关系的向量集来表征。进一步的类比实验[129]证明了skip-gram模型产生的词向量既能捕捉句法关系又能捕捉语义关系，如vec（“China”）！-vec（“Beijing”）=vec（“Japan”）！-vec（“Tokyo”）。此外，他们还发现词向量的成分性，例如vec（“德国”）+vec（“首都”）与vec（“柏林”）很接近。鲁宾斯坦等人受到这些作品的启发。[151]发现分布词表示法擅长预测分类属性（如狗是动物），但无法学习属性属性（如天鹅是白色）。类似地，Gupta等人。[54]表明word2vec嵌入隐式地编码实体的引用属性。分布式词向量和一个简单的有监督模型可以学习以合理的精度预测实体的数字和二进制属性。

**3.3.2上下文嵌入**

大量的研究对语境嵌入中不同类型的知识进行了探索和归纳。一般来说，知识分为两类：语言知识和世界知识。

语言知识为了研究PTMs中的语言知识，设计了一系列的探究任务。Tenney等人。[180]，Liu等人。[108]发现BERT在词性标注和成分标注等句法任务中表现良好。然而，与简单的句法任务相比，BERT在语义和细粒度的句法任务方面还不够好。

此外，Tenney等人。[179]分析了BERT层在不同任务中的作用，发现BERT以与NLP管道中相似的顺序解决任务。此外，在BERT中还存在主谓一致性知识和语义角色知识。此外，Hewitt和Manning[59]，Jawahar等人。[72]，Kim等人。[85]提出了几种从BERT中提取依赖树和组成树的方法，证明了BERT对语法结构的编码能力。Reif等人。[148]探索了BERT内部表征的几何结构，并找到了一些证据：1）语言特征似乎在单独的语义和句法子空间中表现出来；2）注意矩阵包含语法表征；3）BERT在很好的水平上区分了词义。

除了语言知识外，PTMs还可以存储训练数据中的世界知识。探索世界知识的一种简单方法是用“填空”完形填空语句来查询BERT，例如，“Dante出生在[MASK]中”。Petroni等人。[138]通过从多个知识源手动创建单令牌完形填空语句（查询）来构造LAMA（语言模型分析）任务。他们的实验表明，BERT包含了与传统信息抽取方法相竞争的世界知识。因为查询的简单性LAMA，Jiang [74]认为LAMA只是测量语言模型所知道的下限，并提出更先进的方法来生成更多的高效查询。尽管喇嘛的发现令人惊讶，但它也受到了后续工作的质疑[141，82]。类似地，一些研究从BERT中归纳出下游任务的关系知识[15]和常识知识[32]。

**4 PTMs的扩展**

**4.1知识丰富的PTMs**

PTMs通常从通用的大规模文本语料库中学习通用语言表示，但缺乏特定领域的知识。将来自外部知识库的领域知识整合到PTM中被证明是有效的。外部知识的范围从语言[94，83，136，191]、语义[99]、常识[52]、事实[214，136，111，202，195]到特定领域的知识[58，111]。

一方面，可以在训练前注入外部知识。早期的研究[197217201205]集中在知识图形嵌入和单词嵌入的联合学习上。自BERT以来，一些辅助的训练前任务被设计成将外部知识融入到深度PTMs中。

LIBERT[94]通过一个附加的语言约束任务将语言知识整合在一起。Ke等人。[83]整合每个词的情感极性，将MLM延伸到标签感知MLM（LA-MLM）。因此，他们提出的模型SentiLR在多个句子和方面级别的情感分类任务中实现了最先进的性能。Levine等人。[99]提出了SenseBERT，它经过预先训练，不仅可以预测蒙面代币，还可以预测它们在WordNet中的超感。ERNIE（THU）[214]将预先训练在知识图上的实体嵌入与文本中相应的实体提及集成起来，以增强文本表示。类似地，KnowBERT[136]将BERT与实体链接模型联合训练，以端到端的方式合并实体表示。Wang等人。[195]提出了开普勒，它联合优化了知识嵌入和语言建模的目标。这些工作通过实体嵌入注入知识图的结构信息。相反，K-BERT[111]显式地将从KG中提取的相关三元组注入到句子中，以获得BERT的扩展树形输入。此外，熊等人。[202]采用实体替换识别，以鼓励模型更加了解事实知识。然而，这些方法大多在注入知识时更新PTMs的参数，从而避免了在注入多种知识时的灾难性遗忘。为了解决这个问题，K-Adapter[191]通过为不同的预训练任务独立地训练不同的适配器来注入多种类型的知识，从而允许持续的知识注入。

另一方面，我们可以将外部知识融入到预先训练的模型中，而无需从头开始再训练。作为一个例子，K-BERT[111]允许在对下游任务进行微调时注入事实知识。Guan等人。[52]使用常识知识库ConceptNet和ATOMIC来增强故事生成的GPT-2。Yang等人。[207]提出了一种用于机器阅读理解的知识文本融合模型，以获取相关的语言知识和事实知识。

此外，洛根四世等人。[119]和Hayashi等人。[57]将语言模型分别扩展到知识图语言模型（KGLM）和潜在关系语言模型（LRLM），这两种模型都允许基于知识图的预测。这些新的KG条件语言模型显示出潜在的预训练能力。

**4.2多语言和特定语言的PTM**

**4.2.1 Multilingual PTMs**

跨语言语言理解（XLU）早期的研究主要集中在多语言单词嵌入的学习上[45，123，158]，它在一个语义空间中表示来自多种语言的文本。然而，这些方法通常需要语言之间的（弱）对齐。

多语种BERT（mBERT）是由MLM预先训练与共享词汇和重量的维基百科文本从前104种语言。每个训练样本都是单语文档，没有专门设计的跨语言目标，也没有任何跨语言数据。尽管如此，mBERT在跨语言泛化方面的表现还是出人意料地好[140]。K等人。[79]表明语言之间的词汇重叠在跨语言成功中起着微不足道的作用。

XLM[27]通过合并跨语言任务translation language modeling（TLM）改进了mBERT，TLM在并行双语句子对的连接上执行MLM。Unicoder[68]进一步提出了三种新的跨语言预训练任务，包括跨语言单词恢复、跨语言释义分类和跨语言屏蔽语言模型（XMLM）。

XLM RoBERTa（XLM-R）[28]是一个规模化的多语言编码器，在大量的训练数据，2.5TB清除100种不同语言的CommonCrawl数据。XLM-RoBERTa的训练前任务只是单语MLM。XLM-R在多个跨语言基准（包括XNLI、MLQA和NER）上取得了最新的成果。

跨语言语言生成（XLG） 多语言生成是从输入语言中生成不同语言文本的一种任务，如机器翻译和跨语言摘要。

与多语种分类的PTMs不同，多语种生成的PTMs通常需要对编码器和解码器进行联合预训练，而不是只关注编码器。

MASS[160]在多种语言上使用Seq2Seq-MLM对Seq2Seq模型进行了预训练，并对无监督的NMT进行了显著的改进。XNLG[19]对跨语言自然语言生成进行两阶段的预训练。第一阶段用单语MLM和跨语言MLM（XMLM）任务对编码器进行预训练。第二阶段使用单语DAE和跨语言自动编码（XAE）任务对解码器进行预训练，同时保持编码器固定。实验结果表明，XNLG在跨语言问题生成和跨语言抽象总结方面具有一定的优势。mBART[118]是BART[100]的一个多语种扩展，它与Seq2Seq去噪自动编码器（DAE）任务一起在25种语言的大规模单语语料库上对编码器和解码器进行预训练。实验表明，mBART在各种机器翻译（MT）任务中产生了显著的性能提升。

4.2.2特定语言的PTMs

尽管多语言PTMs在许多语言上表现良好，但最近的研究表明，在单一语言上训练的PTMs明显优于多语言结果[125、95、186]。

对于没有明确的词边界的汉语来说，对更大粒度的词表示（29，37，198）和多粒度的词表示（170，171）进行建模已经取得了很大的成功。Kuratov和Arkhipov[92]使用转移学习技术将多语种的PTM适应于俄语的单语PTM。此外，还发布了一些针对不同语言的单语ptm，如法语的camenbert[125]和FlauBERT[95]，芬兰语的FinBERT[186]，荷兰语的BERTje[33]和robert[35]，阿拉伯语的AraBERT[4]。

**4.3 Multi-Modal PTMs**

观察PTMs在许多NLP任务中的成功，一些研究集中在获得PTMs的跨模态版本上。这些模型中的大多数都是为一般的视觉和语言特征编码而设计的。这些模型是在大量的跨模态数据（如带字幕的视频或带字幕的图像）上进行预训练，结合扩展的预训练任务，充分利用多模态特征。通常，基于视觉的MLM、蒙蔽视觉特征建模和视觉语言匹配等任务被广泛应用于多模态预训练中，如VideoBERT[165]、VisualBERT[103]、ViLBERT[120]。

**4.3.1 Video-Text PTMs**

VideoBERT[165]和CBT[164]是视频和文本的联合模型。为了获得用于预训练的视觉和语言标记序列，视频分别由基于CNN的编码器和off the shelf语音识别技术进行预处理。单个Transformer编码器对处理后的数据进行训练，学习视频字幕等下游任务的视觉语言表示。此外，UniViLM[122]提出引入生成任务以进一步预训练在下游任务中使用的解码器。

**4.3.2 Image-Text PTMs**

除了视频语言预训练的方法外，还有一些工作在图像-文本对上引入PTMs，以适应视觉问答（VQA）和视觉常识推理（VCR）等下游任务。有几种模型分别采用两个独立的编码器来表示图像和文本，如ViLBERT[120]和LXMERT[175]。而VisualBERT[103]、B2T2[2]、VLBERT[163]、Unicoder VL[101]和UNITER[17]等其他方法则提出了单流统一Transformer。尽管这些模型结构不同，但在这些方法中引入了类似的训练前任务，如MLM和图像-文本匹配。为了更好地利用视觉元素，在经过预训练的变换器编码之前，通过应用RoI或包围盒检索技术将图像转换成区域序列。

**4.3.3 Audio-Text PTMs**

此外，有几种方法探索了音频文本对上PTMs的可能性，如SpeechBERT[22]。本研究尝试建立一个端到端的语音问答（SQA）模型，使用一个单一的Transformer编码器对音频和文本进行编码，然后在语音和文本语料库中使用MLM进行预训练，并对问答进行微调。

**4.4特定领域和特定任务的PTM**

大多数公共可用的ptm都是在通用领域语料库（如Wikipedia）上进行培训的，这些语料库将其应用限制在特定的领域或任务上。近年来，一些研究提出了基于专业语料库的PTMs训练方法，如BioBERT[98]用于生物医学文本，SciBERT[11]用于科学文本，ClinicalBERT[69,3]用于临床文本。

除了对特定领域的PTM进行预训练外，一些工作还试图使可用的预训练模型适应目标应用，例如生物医学实体规范化[73]、专利分类[97]、进度注释分类和关键词提取[176]。

还提出了一些面向任务的预训练任务，如用于情感分析的SentiLR[83]中的情感标签感知MLM、用于文本摘要的间隙句子生成（GSG）[212]和用于不流畅检测的噪声词检测[192]。

**4.5模型压缩**

由于ptm通常由至少数亿个参数组成，因此很难将其部署到现实应用程序中的在线服务和资源受限的设备上。模型压缩[16]是一种减小模型尺寸和提高计算效率的潜在方法。压缩PTMs的方法有五种：（1）模型剪枝，它去除了不太重要的参数；（2）权重量化，它使用较少的比特来表示参数；（3）在相似的模型单元之间共享参数；（4）知识蒸馏[61]，它训练一个较小的学生模型，从原始模型的中间输出中学习，（5）模块替换，用更紧凑的替换品替换原始PTMs的模块。

表3给出了一些有代表性的压缩ptm的比较。

**4.5.1模型修剪**

模型剪枝是指去除部分神经网络（如权值、神经元、层次、通道、注意头等），从而达到减小模型大小、加快推理时间的目的。

Gordon等人。[51]探讨了修剪的时机（例如，在训练前修剪，在下游微调后修剪）和修剪制度。Michel等人。[128]和Voita等人。[187]试图剪掉Transformer组中所有的自我注意的头。

**4.5.2量化**

量化是指将较高精度的参数压缩到较低精度。沈等人的作品。[156]和Zafrir等人。[211]只关注这个领域。注意，量化通常需要兼容的硬件。

**4.5.3 Parameter Sharing**

另一个著名的减少参数数量的方法是参数共享，它广泛应用于CNNs、RNNs和Transformer中[34]。ALBERT[93]采用跨层参数共享和因子化嵌入参数化方法来降低PTMs的参数。虽然参数的数目大大减少，但ALBERT的训练和推理时间甚至比标准的BERT还要长。

通常，参数共享并不能提高推理阶段的计算效率。

**4.5.4知识提炼**

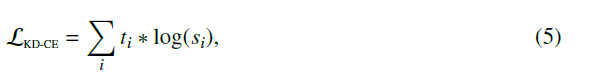
知识蒸馏（KD）[61]是一种压缩技术，它训练一个称为学生模型的小模型来再现一个称为教师模型的大模型的行为。

在这里，教师模型可以是许多模型的集合，并且通常经过良好的预训练。与模型压缩不同，蒸馏技术通过一些优化目标从固定的教师模型学习小学生模型，而压缩技术旨在搜索稀疏的体系结构。

一般而言，蒸馏机理可分为三类：（1）软目标概率蒸馏；（2）其他知识蒸馏；（3）其他结构蒸馏：

（1） 从软目标概率中提取。Bucilua等人[16] 结果表明，使学生近似于教师模型，可以将知识从教师传递到学生。一种常用的方法是逼近教师模型的logits。

DistilBERT[152]将学生模型训练为教师软目标概率的蒸馏损失：



其中ti和si分别是教师模型和学生模型估计的概率。

从软目标概率中提取的信息也可用于任务特定模型，如信息检索[121]和序列标记[181]。

（2） 从其他知识中提炼。从软目标概率出发，将教师模型看作一个黑箱，只关注其输出。此外，分解教师模式和更多知识的提炼可以给学生模式带来改进。

TinyBERT[75]使用嵌入输出、隐藏状态和自关注分布执行逐层蒸馏。MobileBERT[172]还使用软目标概率、隐藏状态和自我注意分布执行逐层蒸馏。MiniLM[194]从教师模型中提取了自我注意分布和自我注意价值关系。另外，其他模型通过多种方法提取知识。Sun等人。[169]引入了“patient”教学学生机制，Liu等人。[113]利用KD改进了预先训练的多任务深神经网络。

（3） 蒸馏到其他结构。一般来说，学生模型的结构与教师模型的结构是相同的，只是层的尺寸较小，隐藏的尺寸较小。然而，从Transformer到RNN[177]或CNN[20]不仅可以减少参数，而且可以简化模型结构，从而降低计算复杂度。

**4.5.5 Module Replacing**

模块替换是一种有趣而简单的减小模型尺寸的方法，它用更紧凑的替代品替换了原PTMs的大模块。Xu等人。[203]在著名的思想实验“忒修斯之船”的推动下，提出了忒修斯压缩算法，该算法用较少参数的模块逐步替换源模型中的模块。与KD不同，Theseus压缩只需要一个特定于任务的丢失函数。特修斯的BERT压缩模型在保持源模型98%以上性能的同时，速度提高了1.94。

**4.5.6其他**

除减少模型尺寸外，还有其他方法在资源有限的实际场景中提高PTMs的计算效率。Liu等人。[112]提出了一种实用的速度可调BERT，即FastBERT，它可以通过样本自适应机制动态地减少计算步骤。

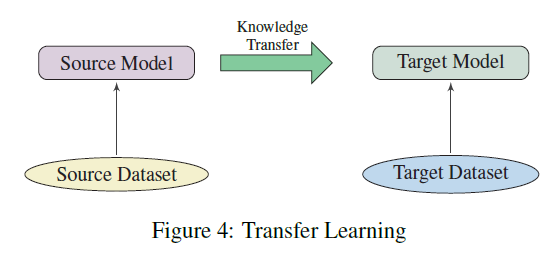
**5 Adapting PTMs to Downstream Tasks**

尽管PTMs从一个庞大的语料库中获取通用的语言知识，但是如何有效地将它们的知识适应于下游任务仍然是一个关键问题。

**5.1 Transfer Learning**

转移学习[132]是将源任务（或领域）中的知识调整到目标任务（或领域）中。图4展示了转移学习。

自然语言处理中的迁移学习有许多类型，如领域适应、跨语言学习、多任务学习等。使PTMs适应下游任务是一种顺序转移学习任务，在该任务中，任务被顺序地学习，目标任务标记了数据。



**5.2 How to Transfer?**

要将PTM的知识转移到下游NLP任务，我们需要考虑以下问题：

**5.2.1选择合适的训练前任务、模型结构和语料库**

不同的ptm通常在同一个下游任务上有不同的效果，因为这些ptm使用不同的预训练任务、模型架构和语料库进行训练。

（1） 目前，语言模型是最受欢迎的预训练任务，能够更有效地解决一系列NLP问题[143]。然而，不同的训练前任务有自己的偏好，对不同的任务有不同的评价。例如，NSP任务[36]使PTM理解两个句子之间的关系。因此，PTM可以帮助下游任务，如问答（QA）和自然语言推理（NLI）。

（2） PTM的体系结构对下游任务也很重要。例如，虽然BERT有助于大多数自然语言理解任务，但很难生成语言。

（3） 下游任务的数据分布应近似于PTMs。目前，有大量的o-the-shelf型ptm，它们同样可以方便地用于各种领域特定或语言特定的下游任务。

因此，在给定目标任务的情况下，选择合适的预训练任务、体系结构和语料库来训练PTMs一直是一个很好的解决方案。

**5.2.2选择合适的层**

给定一个预先训练好的深度模型，不同的层应该捕获不同类型的信息，如词性标记、解析、长期依赖、语义角色、共指。对于基于RNN的模型，Belinkov等人。[10] 以及Melamud等人[127]表明，在多层LSTM编码器中从不同层学习的表示有利于不同的任务（例如，预测POS标记和理解词义）。对于基于Transformer的PTM，Tenney等人。[179]found BERT代表了传统NLP管道的步骤：基本句法信息出现在网络的早期，而高级语义信息出现在更高的层次。

设表示具有l层的预训练模型的第l层表示，g（）表示目标任务的特定任务模型。

选择表示有三种方法：

a） 仅嵌入。一种方法是只选择预先训练好的静态嵌入，而模型的其余部分仍然需要从头训练才能完成新的目标任务。他们无法获取更有用的更高层次的信息。单词嵌入只在捕捉单词的语义方面有用，但我们也需要理解更高层次的概念，如词义。

b） 顶层。最简单、最有效的方法是将顶层的表示形式输入到特定于任务的模型g（H（L））中。

c） 所有层。更灵活的方法是自动选择软版本中的最佳层，如ELMo[135]：



其中是层l的softmax规格化权重，是一个标量，用于缩放通过预训练模型输出的向量。混合表示被输入到任务特定的模型g（rt）中。

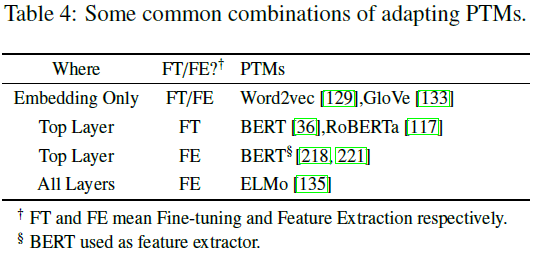
**5.2.3调还是不调？**

目前，有两种常用的模型转换方法：特征提取（预训练参数被冻结）和微调（预训练参数被解冻和微调）。

在特征提取方法中，将预先训练好的模型看作off-the-shelf特征提取器。此外，公开内部层非常重要，因为它们通常对最可转移的表示进行编码[137]。

虽然这两种方法都能显著地提高NLP任务的效率，但特征提取方法需要更复杂的任务特定体系结构。因此，与特征提取方法相比，微调方法对于许多不同的下游任务通常更通用和方便。

表4给出了一些常用的自适应ptm组合。



**5.3微调策略**

随着PTMs深度的增加，PTMs捕捉到的表示使得下游任务更加容易。因此，整个模型的任务特定层是简单的。自ULMFit和BERT以来，微调已成为PTMs的主要自适应方法。然而，微调的过程往往是脆弱的：即使使用相同的超参数值，不同的随机种子也可能导致实质上不同的结果[38]。

除了标准微调之外，还有一些有用的微调策略。

两级微调 另一种解决方案是两级传递，它引入了预训练和微调之间的中间阶段。在第一阶段，PTM被转换成一个由中间任务或语料库微调的模型。在第二阶段，传输的模型将微调到目标任务。Sun等人。[167]表明，对相关领域语料库进行“进一步的预训练”可以进一步提高BERT的能力，并在8个广泛研究的文本分类数据集上取得了最新的性能。Phang等人。[139]和Garg等人。[48]引入了与目标任务相关的中间监督任务，这给BERT、GPT和ELMo带来了很大的改进。Li等人[106]还使用了两阶段转移来预测故事的结尾。提出的TransBERT（transferable BERT）不仅可以从大规模的无标记数据中传递一般的语言知识，还可以从各种语义相关的有监督任务中传递特定类型的知识。

多任务微调 刘等。[114]在多任务学习框架下对BERT进行了微调，这表明多任务学习和预训练是互补的技术。

带有额外适应模块的微调 微调的主要缺点是其参数效率：每个下游任务都有自己的微调参数。因此，一种较好的解决方案是在原参数不变的情况下，在PTMs中注入一些可微调的自适应模块。

Stickland和Murray[162]为单共享BERT模型配备了额外的任务特定适应模块，即投射注意层（PALs）。与PALs共享的BERT在GLUE基准上分别匹配经过微调的模型，参数大约减少了7倍。同样，Houlsby等人。[66]通过添加适配器模块修改了预先训练的BERT的体系结构。适配器模块产生一个紧凑且可扩展的模型；它们只为每个任务添加一些可训练的参数，并且可以添加新的任务，而不必重新访问以前的任务。原始网络的参数保持不变，从而产生高度的参数共享。

其他人 被广泛使用的集成模型的成功所激励，Xu等人。[206]采用自集成和自蒸馏两种目标机制改进了BERT的微调，可以在不利用外部资源或显著降低训练效率的情况下提高BERT在下游任务上的性能。他们在一个单一的训练过程中集成了集成和升华。教师模型是一个集合模型，通过参数平均前几个时间步的几个学生模型。逐步解冻[67]不是同时对所有层进行微调，而是一种从顶层开始逐步解冻ptm层的有效方法。Chronopoulou等人。[21]提出了一种简单的解冻方法：顺序解冻，先对随机初始化的任务特定层进行微调，然后对PTM的隐藏层进行解冻，最后对嵌入层进行解冻。

Li和Eisner[104]使用可变信息瓶颈压缩ELMo嵌入，同时只保留有助于目标任务的信息。

一般来说，以上的工作表明，通过更好的微调策略可以进一步刺激PTMs的效用。

6.PTM的资源

在线提供了许多便携式测试系统的相关资源。表5提供了一些流行的存储库，包括第三方实现、纸质列表、可视化工具和PTMs的其他相关资源。

此外，还有一些关于NLP的PTMs的好的调查论文[196110150]。

**7 Applications**

在这一部分中，我们总结了PTMs在几个典型NLP任务中的应用。

**7.1一般评价基准**

对于NLP社区来说，一个重要的问题是如何用一个可比较的指标来评估PTMs。因此，大规模的基准是必要的。

通用语言理解评估（GLUE）基准[190]是9个自然语言理解任务的集合，包括单句分类任务（CoLA和SST-2）、成对文本分类任务（MNLI、RTE、WNLI、QQP和MRPC）、文本相似性任务（STSB）和相关排序任务（QNLI）。GLUE benchmark是为评估模型的健壮性和泛化而精心设计的。GLUE不为测试集提供标签，而是设置评估服务器。

然而，由于近年来的进展极大地侵蚀了GLUE基准的净空，一个新的基准被称为SuperGLUE[189]被提出。与GLUE相比，SuperGLUE具有更具挑战性的任务和更多样的任务格式（例如，coreference resolution和question answering）。最先进的PTM列在相应的排行榜4）5）中。

**7.2 Question Answering**

问答（QA）或更狭义的概念机阅读理解（MRC）是NLP社区的一个重要应用。从简单到困难，有三种类型的QA任务：单轮抽取QA（SQuAD）[145]、多轮生成QA（CoQA）[147]和多跳QA（HotpotQA）[208]。

伯特创造性地将抽取的QA任务转换为预测答案的起始跨度和结束跨度的跨度预测任务[36]。之后，PTM作为预测跨度的编码器已经成为一个竞争的基线。对于采掘QA，Zhang等人。[215]提出了一种可追溯的读卡器架构，并用PTM（如ALBERT）初始化编码器。对于多轮生成QA，Ju等人。[78]提出了“PTM+对抗性训练+理论基础标注+知识提炼”模型。对于多跳QA，Tu等人。[182]提出了一个可解释的“选择、回答和解释”（SAE）系统，PTM在选择模块中充当编码器。

通常，所提出的QA模型中的编码器参数是通过PTM初始化的，其他参数是随机的初始化。最先进的模型列在相应的排行榜上。

**7.3情绪分析**

BERT通过简单地微调SST-2（一个广泛用于情绪分析（SA）的数据集）而优于先前的最新模型。Bataa和Wu[9]将BERT与转移学习技术结合起来，实现了日本SA的最新水平。

尽管他们在简单的情绪分类方面取得了成功，但是直接将BERT应用到基于方面的情绪分析（ABSA）中（ABSA是一个细粒度的SA任务）并没有显示出显著的改进[166]。为了更好地利用BERT的强大表示，Sun等人。[166]通过将ABSA从一个句子分类任务转换为一个句子对分类任务来构造一个辅助句子。Xu等人。[204]提出了将BERT从源域和任务调整到ABSA域和任务的后训练。此外，Rietzler等人。[149]通过分析具有ABSA性能的跨域岗位培训行为，扩展了[204]的工作。Karimi等人。[81]表明对抗性训练可以进一步提高训练后的BERT的成绩。Song等人。[161]增加了一个额外的池模块，可以实现为LSTM或注意机制，以利用BERT中间层进行ABSA。此外，Li等人。[105]共同学习面向端到端ABSA的方面检测和情感分类。SentiLR[83]从SentiWordNet中获取词性标记和先验情感极性，并采用标签感知的MLM方法，利用引入的语言知识来获取句子级情感标记和词级情感转移之间的关系。SentiLR在多个句子和方面级别的情感分类任务中实现了最先进的性能。

对于情绪转移，Wu等人。[200]基于伯特提出的“掩模和填充”。在蒙版步骤中，该模型通过蒙版情感标记将情感与内容分离。在填充步骤中，它使用BERT和目标情感嵌入来填充蒙面位置。

**7.4命名实体识别**

命名实体识别（NER）在信息抽取中起着重要的作用，在许多NLP下游任务中发挥着重要的作用。在深度学习中，大多数的内质网方法都在序列标记框架中。句子中的实体信息将被转换成标签序列，一个标签对应一个单词。该模型用于预测每个单词的标签。自从ELMo和BERT在NLP中显示出他们的能力以来，关于NER的预训练模型还有很多工作要做。

Akbik等人。[1] 使用一个预先训练的字符级语言模型为NER生成单词级嵌入。TagLM[134]和ELMo[135]使用预先训练的语言模型的最后一层输出和每层输出的加权和作为单词嵌入的一部分。Liu等人。[107]采用分层剪枝和紧密连接的方法，加速了ELMo对内生关系的推断。Devlin等人。[36]使用第一个BPE的BERT表示来预测没有CRF的每个单词的标签。Pires等人。[140]通过多语言BERT实现了零射击。蔡等人。[181]利用知识蒸馏，在单个CPU上为NER运行一个小的BERT。此外，BERT还用于领域特定的NER，如生物医学[56，98]等。

**7.5机器翻译**

机器翻译（Machine Translation，MT）是自然语言处理（NLP）领域的一个重要课题，吸引了众多研究者的关注。几乎所有的神经机器翻译（NMT）模型共享编码器解码器框架，其首先将编码器的输入令牌编码为隐藏的表示，然后从解码器中解码目标语言中的输出令牌。Ramachandran等人。[146]发现通过使用预先训练好的两种语言模型的权重初始化编码器和解码器，可以显著地改进编码器-解码器模型。Edunov等人。[41]使用ELMo在NMT模型中设置单词嵌入层。本文采用一种预先训练好的语言模型进行源词嵌入初始化，提高了英语-土耳其语和英语-德语NMT模型的性能。

考虑到BERT在其他NLP任务中的出色性能，研究如何将BERT合并到NMT模型中是很自然的。Conneau和Lample[27]试图通过一个多语种的预先训练的BERT模型初始化整个编码器和解码器，并表明在无监督的机器翻译和英语-罗马尼亚语监督的机器翻译上可以实现显著的改进。同样，Clinchant等人。[25]设计了一系列不同的实验来检验在NMT模型的编码器部分使用BERT的最佳策略。他们使用BERT作为编码器的初始化，取得了一些改进。此外，他们发现这些模型可以在域外数据集上获得更好的性能。Imamura和Sumita[70]提出了一种用于NMT的两级BERT微调方法。

在第一阶段，编码器由一个预先训练好的BERT模型初始化，它们只在训练集上训练解码器。第二阶段，整个NMT模型在训练集上进行联合微调。实验结果表明，该方法优于直接对整个模型进行微调的单级微调方法。除此之外，朱等人。

[221]建议使用预先训练的BERT作为额外的内存，以促进NMT模型。具体地说，它们首先编码通过预先训练的BERT输入令牌，并使用最后一层的输出作为额外的内存。然后，NMT模型可以通过编码器和解码器的每一层中的额外注意模块来访问存储器。它们在有监督、半监督和无监督的机器翻译方面都有显著的改进。

MASS（屏蔽序列到序列预训练）[160]利用Seq2Seq MLM来联合预训练编码器和解码器，而不是仅仅预训练编码器。在实验中，这种方法可以超越Conneau和Lample[27]在无监督的MT和英语-罗马尼亚语-监督的MT上提出的BERT式的预训练，不同于MASS，mBART[118]，BART[100]的多语种扩展，结合Seq2Seq去噪自动编码器（DAE）任务，对25种语言的大规模单语语料库中的编解码器进行预训练。实验表明，在句子和文档两个层次上，mBART都能显著提高机器翻译的有监督和无监督能力。

**7.6 Summarization**

近年来，以产生一个较短的文本来保留较长文本的大部分意义，引起了NLP界的关注。自PTM的广泛应用以来，任务得到了显著的改进。Zhong等人。[218]引入可转移知识（如BERT）进行总结，并超越了以往的模型。Zhang等人。[213]尝试预先训练一个文档级模型，该模型预测句子而不是单词，然后将其应用于下游任务，如摘要。更详细地说，张等人。[212]设计了一个用于预训练的Gap语句生成（GSG）任务，其目标是从输入中生成类似摘要的文本。此外，刘和拉帕塔[116]提出了BERTSUM。BERTSUM包含了一个新的文档级编码器，以及一个用于抽取摘要和抽象摘要的通用框架。在编码器帧中，BERTSUM通过插入多个[CLS]标记来扩展BERT，以学习句子表示。为了提取摘要，BERTSUM堆叠了几个中间内容转换层。为了进行抽象总结，BERTSUM提出了一种采用新的微调调度的两阶段微调方法。Zhong等人。[219]提出了一种新的摘要级框架MATCHSUM，并将抽取摘要概念化为一个语义文本匹配问题。他们提出了一个连体BERT架构来计算源文档和候选摘要之间的相似性，并且仅使用BERT的基本版本就在CNN/DailyMail（44.41 in ROUGE-1）上获得了最新的结果。

**7.7对抗性攻击和防御**

深层神经模型容易受到对抗性例子的影响，这些例子可能会误导模型产生特定的错误预测，并对原始输入产生不可察觉的干扰。在CV中，对抗性攻击和防御被广泛研究。然而，由于语言的离散性，它对文本仍然具有挑战性。文本对抗性样本的生成需要具备这样的品质：（1）对人类判断不易察觉，但对神经模型有误导作用；（2）语法流利，语义与原始输入一致。Jin等人。[76]以对抗性的例子成功地攻击了文本分类和文本蕴涵方面的微调BERT。华莱士等人。[188]定义的通用对抗触发器，当与任何输入连接时，它可以诱导模型产生特定目的的预测。一些触发器甚至会导致GPT-2模型产生种族主义文本。Sun等人[168]显示伯特在拼写错误方面并不强大。

PTMs还具有产生对抗性样本的巨大潜力。Li等人[102]提出的BERT攻击是一种基于BERT的高质量、高效的攻击。他们在下游任务中将BERT与另一个微调的BERT进行对比，并成功地误导目标模型进行错误的预测，在成功率和扰动百分比方面均优于最新的攻击策略，同时生成的对抗样本流畅且语义保持。

此外，针对PTMs的对抗性防御也很有前景，这将提高PTMs的鲁棒性，使其对对抗性攻击免疫。

对抗训练的目的是通过最小化嵌入空间中保持标签扰动的最大风险来提高泛化能力。最近的研究[220115]表明，对抗性的预训练或微调可以提高NLP的PTMs的泛化性和鲁棒性。

**8 Future Directions**

虽然PTMs已经证明了其在各种NLP任务中的能力，但是由于语言的复杂性，仍然存在着挑战。在本节中，我们提出了未来PTMs的五个方向。

（1） PTMs的上限目前，PTMs还没有达到其上限。目前大部分的PTMs可以通过更多的训练步骤和更大的语料库来进一步改进。

通过增加模型的深度，可以进一步提高NLP的技术水平，例如威震天LM[157]（83亿参数，72个Transformer层，隐藏尺寸为3072和32个注意头）和Turing-NLG9（170亿参数，78个Transformer层，隐藏尺寸为4256和28个注意头）。

通用语言教学法一直是我们学习语言内在普遍知识（甚至世界知识）的追求。然而，这种PTMs通常需要更深层次的体系结构、更大的语料库和具有挑战性的训练前任务，这进一步导致了更高的训练成本。然而，训练大型模型也是一个具有挑战性的问题，需要更复杂、更科学的训练技术，如分布式训练、混合精度训练、梯度累积训练等，因此，更实际的方向是设计更科学的模型结构、自监督的预训练任务、优化器，以及使用现有硬件和软件的培训技能。伊莱克特拉[24]是朝这个方向发展的一个很好的解决方案。

（2） PTMs的体系结构 Transformer已被证明是一种有效的预训练体系结构。然而，Transformer的主要限制是它的计算复杂度，这是二次输入长度。受gpu内存的限制，目前大多数ptm无法处理大于512个令牌的序列。打破这一限制需要改进Transformer的结构，如Transformer XL[31]。因此，寻找更有效的PTMs模型体系结构对于获取更大范围的上下文信息非常重要。

深部结构的设计具有挑战性，我们可以寻求一些自动化方法的帮助，例如神经结构搜索（NAS）[223]。

（3） 面向任务的预训练和模型压缩 在实践中，不同的下游任务对PTMs的能力提出了不同的要求。PTMs与下游任务的差异通常表现在两个方面：模型结构和数据分布。较大的差异可能导致PTMs的益处可能微不足道。例如，文本生成通常需要一个特定的任务来预先训练编码器和解码器，而文本匹配则需要为句子对设计的预先训练任务。

此外，虽然更大的ptm通常可以带来更好的性能，但实际问题是如何在特殊情况下利用这些巨大的ptm，例如低容量设备和低延迟应用程序。因此，我们可以为下游任务仔细设计特定的模型架构和预训练任务，或者从现有的PTMs中提取部分特定于任务的知识。

与其从头开始训练面向任务的ptm，我们可以使用现有的通用ptm，通过使用模型压缩等技术来教授它们（参见第4.5节）。尽管CV[18]中对CNNs的模型压缩进行了广泛的研究，但对NLP的PTMs的压缩才刚刚开始。Transformer的全连接结构也使模型压缩更具挑战性。

4） 知识转移除了微调，微调是目前PTMs将知识转移到下游任务的主要方法，但其不足之处在于其参数效率：每个下游任务都有自己的微调参数。一种改进的解决方案是固定PTMs的原始参数，并为特定任务添加小的微调自适应模块[16266]。因此，我们可以使用共享PTM来服务多个下游任务。实际上，从PTMs中挖掘知识可以更加灵活，例如特征提取、知识提取[210]、数据扩充[199、91]，使用PTMs作为外部知识[138]。希望有更多的科学方法。

（5） PTMs的可解释性和可靠性尽管PTMs达到了令人印象深刻的性能，但其深刻的非线性结构使得决策过程高度不透明。

可解释人工智能（XAI）[5]是近年来人工智能领域的研究热点。与图像的CNNs不同，由于类似转换器的体系结构和语言的复杂性，解释PTMs比较困难。

为了分析PTMs中包含的语言和世界知识，我们做了大量的工作（见第3.3节），这有助于我们在一定程度上理解这些pmt。

然而，许多关于模型分析的工作都依赖于注意机制，注意对可解释性的客观性仍然存在争议[71155]。

此外，PTM也容易受到敌对攻击（见第7.7节）。随着PTMs在生产系统中的广泛应用，PTMs的可靠性也成为一个备受关注的问题。对PTMs的对抗性攻击的研究有助于我们通过充分暴露PTMs的脆弱性来了解PTMs的能力。PTMs的对抗性防御也很有前景，它可以提高PTMs的鲁棒性，使其对对抗性攻击免疫。

总的来说，PTMs作为许多NLP应用中的关键组件，其可解释性和可靠性在许多方面还有待进一步探讨，这有助于我们理解PTMs是如何工作的，并为更好地使用和进一步改进PTMs提供了指导。

**9.结论**

在本研究中，我们对NLP的PTMs进行了全面的概述，包括背景知识、模型架构、训练前任务、各种扩展、自适应方法、相关资源和应用。在现有PTMs的基础上，我们从四个不同的角度提出了PTMs的新分类法。我们还提出了PTMs未来可能的研究方向。