**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

**摘要**

我们引入了一个新的语言表示模型BERT，它代表了来自Transformers的双向编码器表示。与最近的语言表示模型（Peters等人，2018a；Radford等人，2018）不同，BERT的设计目的是通过在所有层中的左、右上下文上联合调节，从未标记文本中预处理深层的双向表示。其结果是，预先训练的BERT模型只需一个额外的输出层就可以进行微调，从而为广泛的任务（如问答和语言推理）创建最新的模型，而无需对特定于任务的体系结构进行大量修改。

Bert概念简单，经验强大。它在11个自然语言处理任务中获得了最新的结果，包括将GLUE分数提高到80.5%（绝对提高7.7%），多NLI准确度提高到86.7%（绝对提高4.6%），班1.1问答测试F1至93.2（1.5分绝对改善），班2.0测试F1至83.1（5.1分绝对改善）。

**1 引言**

语言模型预训练对于改善许多自然语言处理任务是有效的（Dai and Le，2015；Peters et al.，2018a；Radford et al.，2018；Howard and Ruder，2018）。这些任务包括句子级任务，如自然语言推理（Bowman等人，2015；Williams等人，2018）和复述（Dolan和Brockett，2005），其目的是通过对句子进行整体分析来预测句子之间的关系，以及标记级任务，如命名实体识别和问答，模型需要在令牌级别生成细粒度输出（Tjong Kim Sang和De Meulder，2003；Rajpurkar等人，2016）。

将预先训练的语言表示应用于下游任务有两种现有策略：基于特征和微调。基于特征的方法，例如ELMo（Peters等人，2018a），使用特定于任务的体系结构，其中包括作为附加特征的预先训练的表示。微调方法，如生成式预训练Transformers（OpenAI GPT）（Radford等人，2018），引入最小任务特定参数，并通过简单微调所有预训练参数来训练下游任务。这两种方法在训练前都有相同的目标函数，即使用单向语言模型来学习一般的语言表征。

我们认为目前的技术限制了预训练表示的能力，特别是对于微调方法。主要的限制是标准语言模型是单向的，这限制了可以在预培训期间使用的体系结构的选择。例如，在OpenAI GPT中，作者使用了一个左至右的架构，其中每个令牌只能处理Transformer的self-attention层中以前的令牌（Vaswani等人，2017）。这样的限制对于句子级任务来说是次优的，并且当将基于精细调整的方法应用于诸如问答之类的令牌级任务时可能会非常有害，而在这些任务中，从两个方向合并上下文是至关重要的。

在本文中，我们改进了基于微调的方法，提出了BERT：来自变压器的双向编码器表示。BERT在完形填空任务（Taylor，1953）的启发下，通过使用一个“蒙面语言模型”（MLM）的预训练目标来缓解前面提到的单向性约束。屏蔽语言模型随机屏蔽输入中的一些标记，目的是仅基于上下文预测屏蔽的原始词汇。与左-右语言模型的预训练不同，MLM的目标使表现融合了左和右上下文，这使得我们可以预先训练一个深层的双向转换器。除了蒙面语言模型外，我们还使用了一个“下一句预测”任务，该任务联合预处理文本对表示。本文的贡献如下：

我们证明了双向预训练对于语言表达的重要性。与Radford等人不同。（2018）使用单向语言模型进行预训练，BERT使用屏蔽语言模型实现预训练的深度双向表示。这也与Peters等人的研究结果相反。（2018a），它使用了从左到右和从右到左的独立训练LMs的浅串联。

我们表明，预先训练的表示减少了对许多高度工程化的特定任务体系结构的需求。BERT是第一个基于精细调整的表示模型，它在一组大型句子级和令牌级任务上实现了最新的性能，其性能优于许多特定于任务的体系结构。

伯特提出了11项NLP任务的最新技术。代码和预先培训的模型可在https://github.com/google-research/bert.

**2 相关的工作**

一般语言表征的培训有很长的历史，我们在本节简要回顾了最广泛使用的方法。

**2.1无监督的基于特征的方法**

学习广泛适用的单词表示法几十年来一直是一个活跃的研究领域，包括非神经方法（Brown et al.，1992；Ando and Zhang，2005；Blitzer et al.，2006）和神经方法（Mikolov et al.，2013；Pennington et al.，2014）。经过预训练的单词嵌入是现代NLP系统的一个组成部分，与从头学习的嵌入相比，它提供了显著的改进（Turian等人，2010）。为了预训练单词嵌入向量，使用了从左到右的语言建模目标（Mnih和Hinton，2009），以及在左右上下文中区分正确和错误单词的目标（Mikolov等人，2013）。

这些方法被推广到更粗的颗粒，例如句子嵌入（Kiros等人，2015；Logeswaran和Lee，2018）或段落嵌入（Le和Mikolov，2014）。为了训练句子表示，先前的工作使用目标对候选的下一个句子进行排序（Jernite等人，2017；Logeswaran和Lee，2018），左到右生成下一个句子单词，给出上一个句子的表示（Kiros等人，2015），或去噪自编码器派生的目标（Hill等人，2016）。

ELMo及其前身（Peters et al.，2017，2018a）从不同的维度概括了传统的单词嵌入研究。它们从左到右和从右到左的语言模型中提取上下文相关的特征。每个标记的上下文表示是从左到右和从右到左表示的连接。当将上下文单词嵌入与现有的特定任务架构相结合时，ELMo提出了几个主要NLP基准（Peters et al.，2018a）的最新技术，包括问答（Rajpurkar et al.，2016）、情绪分析（Socher et al.，2013）和命名实体识别（Tjong Kim Sang and De Meulder，2003）。Melamud等人。（2016）建议通过一项任务学习语境表征，使用LSTMs从左右语境预测单个单词。与ELMo相似，它们的模型是基于特征的，并且没有深度的双向性。Fedus等人。（2018）表明完形填空任务可以用来提高文本生成模型的健壮性。

**2.2无监督微调方法**

与基于特征的方法一样，第一种方法仅在这个方向上工作，即从未标记文本中预训练单词嵌入参数（Collobert和Weston，2008）。

最近，产生上下文标记表示的句子或文档编码器已经从未标记文本中进行了预训练，并针对受监督的下游任务进行了微调（Dai和Le，2015；Howard和Ruder，2018；Radford等人，2018）。这些方法的优点是几乎不需要从头学习参数。至少部分由于这一优势，OpenAI GPT（Radford et al.，2018）从GLUE benchmark（Wang et al.，2018a）在许多句子级任务上取得了先前的最新成果。从左到右语言模型和自动编码器目标已用于此类模型的预培训（Howard and Ruder，2018；Radford et al.，2018；Dai and Le，2015）。

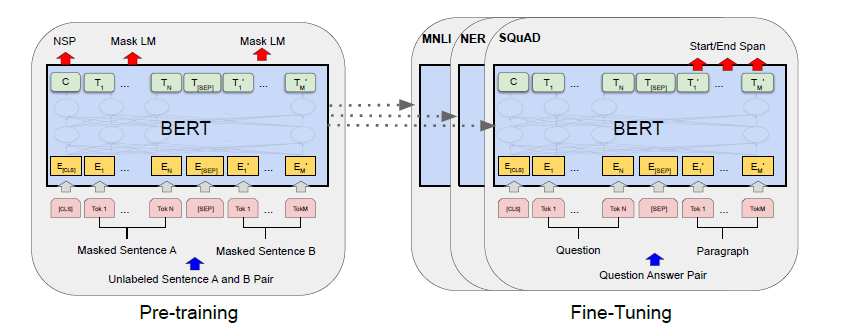


图1:BERT的总体预培训和微调程序。除了输出层之外，在预训练和微调中都使用相同的体系结构。相同的预训练模型参数用于初始化不同下游任务的模型。在微调过程中，所有参数都会微调。[CLS]是在每个输入示例前面添加的特殊符号，[SEP]是特殊的分隔符标记（例如分隔问题/答案）。

**2.3监督数据的迁移学习**

也有研究表明，在大数据集的监督任务中，例如自然语言推理（Conneau等人，2017年）和机器翻译（McCann等人，2017年）也有有效的迁移。计算机视觉研究还证明了从大型预训练模型进行转移学习的重要性，其中一个有效的方法是对使用ImageNet预训练的模型进行微调（Deng等人，2009；Yosinski等人，2014）。

**3 BERT**

在本节中，我们将介绍BERT及其详细实现。在我们的框架中有两个步骤：预培训和微调。在预训练过程中，该模型在不同的预训练任务上对未标记的数据进行训练。对于精细调整，首先使用预先训练的参数初始化BERT模型，然后使用来自下游任务的标记数据对所有参数进行精细调整。每个下游任务都有单独的微调模型，即使它们是用相同的预训练参数初始化的。图1中的问答示例将作为本节的运行示例。

BERT的一个显著特点是其跨不同任务的统一架构。预先训练的体系结构和最终的下游体系结构之间的差别很小。

**模型架构** BERT的模型架构是基于Vaswani等人描述的原始实现的多层双向Transformer。（2017）并在tensor2tensor库中发布。由于Transformer的使用已变得普遍，并且我们的实现几乎与原始版本相同，因此我们将省略对模型架构的详尽背景描述，并请读者参考Vaswani等人（2017）以及优秀的指南，如“注释Transformer”。

在这项工作中，我们将层数（即，Transformer块）表示为L，将隐藏大小表示为H，将自聚焦头的数目表示为A。3我们主要报告两种模型大小的结果：（L=12，H=768，A=12，总参数=110M）和（L=24，H=1024，A=16，总参数=340M）。

为了进行比较，选择了与OpenAI GPT具有相同模型大小的BERTBASE。然而，关键的是，BERT变换器使用双向自我注意，而GPT变换器使用约束自我注意，其中每个令牌只能关注其左侧的上下文。

**输入/输出表示** 使BERT能够处理各种下游任务，我们的输入表示能够在一个标记序列中明确地表示一个句子和一对句子（例如，h Question，Answer i）。在这项工作中，“句子”可以是连续文本的任意跨度，而不是实际的语言句子。“序列”是指输入到BERT的令牌序列，它可以是一个句子，也可以是打包在一起的两个句子。

我们使用词条嵌入（Wu et al.，2016）和30000个标记词汇表。每个序列的第一个标记总是一个特殊的分类标记（[CLS]）。与此标记对应的最终隐藏状态用作分类任务的聚合序列表示。句子对被打包成一个序列。我们用两种方法来区分句子。首先，我们用一个特殊的标记（[SEP]）将它们分开。其次，我们将学习嵌入添加到每个标记中，以指示它是属于句子a还是句子B。如图1所示，我们将输入嵌入表示为E，将特殊[CLS]标记的最终隐藏向量表示为，将第i个输入标记的最终隐藏向量表示为。

对于给定的令牌，其输入表示是通过对相应的令牌、段和位置嵌入求和来构造的。这个结构的可视化如图2所示。