

BERT：用于语言理解的深度双向Transformer预训练

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova 谷歌AI语言
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

摘要

我们介绍了一种名为BERT的新语言表示模型，BERT代表来自Transformer的双向编码器表示。与近期的语言表示模型不同（Peters等人，2018a；Radford等人，2018），BERT旨在通过在所有层中联合调节左右上下文，从未标记文本中预训练深度双向表示。因此，预训练的BERT模型仅需添加一个额外的输出层进行微调，即可为广泛的任务（如问答和语言推理）创建最先进的模型，而无需对任务特定架构进行实质性修改。

BERT在概念上简单，但在实证中表现出强大的能力。它在十一个自然语言处理任务中取得了新的最先进成果，包括将GLUE分数提升至80.5%（绝对提升7.7个百分点）、MultiNLI准确率提升至86.7%（绝对提升4.6%）、SQuAD v1.1问答测试F1值达到93.2（绝对提升1.5分）以及SQuAD v2.0测试F1值达到83.1（绝对提升5.1分）。

1 引言

语言模型预训练已被证明能有效提升多种自然语言处理任务的性能（Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018）。这包括句子级任务（如自然语言推理（Bowman et al., 2015; Williams et al., 2018）和复述检测（Dolan and Brockett, 2005）），这类任务通过整体分析句子以预测其关联关系；同时也涵盖词元级任务（如命名实体识别与问答系统），此类任务要求模型在词元级别生成细粒度输出（Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016）。

将预训练语言表示应用于下游任务现有两种策略：*feature-based*和*fine-tuning*。基于特征的方法，如ELMo（Peters等人，2018a），采用包含预训练表示作为附加特征的特定任务架构。而微调方法，如生成式预训练Transformer（OpenAI GPT）（Radford等人，2018），则引入极少的任务特定参数，并通过直接微调*all*预训练参数在下游任务上进行训练。这两种方法在预训练阶段共享相同的目标函数，即使用单向语言模型来学习通用语言表示。

我们认为，当前的技术限制了预训练表征的潜力，尤其是在微调方法中。主要局限在于标准语言模型是单向的，这限制了预训练时可用的架构选择。例如，在OpenAI GPT中，作者采用了从左到右的架构，其中每个标记在Transformer的自注意力层（Vaswani等人，2017）中只能关注先前的标记。这种限制对于句子级任务来说并非最优，而在将基于微调的方法应用于标记级任务（如问答）时可能极为不利，因为在这些任务中，双向上下文的整合至关重要。

在本文中，我们通过提出BERT：来自Transformer的双向编码器表示，改进了基于微调的方法。BERT通过使用“掩码语言模型”（MLM）预训练目标来缓解之前提到的单向性限制，该目标受到完形填空任务（Taylor, 1953）的启发。掩码语言模型随机掩盖输入中的部分标记，其目标是预测被掩盖标记的原始词汇ID。

仅基于其上下文预测单词。与从左到右的语言模型预训练不同，MLM目标使表示能够融合左右上下文，这使我们能够预训练一个深度双向Transformer。除了掩码语言模型外，我们还使用“下一句预测”任务，联合预训练文本对表示。本文的贡献如下：

- 我们证明了双向预训练对于语言表征的重要性。与Radford等人（2018）使用单向语言模型进行预训练不同，BERT采用掩码语言模型来实现预训练的深度双向表征。这也与Peters等人（2018a）形成对比，后者仅将独立训练的左到右和右到左语言模型进行浅层拼接。
- 我们证明，预训练表征减少了对许多精心设计的任务特定架构的需求。BERT是首个基于微调的表征模型，在大量句子级和词元级任务上实现了最先进的性能，超越了众多任务特定的架构。
- BERT在十一个自然语言处理任务上取得了最先进的成果。代码和预训练模型可在<https://github.com/google-research/bert>获取。

2 相关工作

预训练通用语言表示有着悠久的历史，本节我们将简要回顾最广泛使用的方法。

2.1 基于无监督特征的方法

学习广泛适用的词语表示是几十年来一个活跃的研究领域，包括非神经方法（Brown等人，1992；Ando和Zhang，2005；Blitzer等人，2006）和神经方法（Mikolov等人，2013；Pennington等人，2014）。预训练的词嵌入是现代自然语言处理系统的核心组成部分，相比从零开始学习的嵌入能带来显著提升（Turian等人，2010）。为预训练词嵌入向量，研究者采用了从左到右的语言建模目标（Mnih和Hinton，2009），以及区分左右上下文中正确与错误词语的目标（Mikolov等人，2013）。

这些方法已被推广到更粗的粒度，例如句子嵌入（Kiros等人，2015；Logeswaran与Lee，2018）或段落嵌入（Le与Mikolov，2014）。为了训练句子表示，先前的研究采用了多种目标函数：对候选下一句子进行排序（Jernite等人，2017；Logeswaran与Lee，2018）、在给定前一句子表示的情况下从左到右生成下一句子的词语（Kiros等人，2015），或基于去噪自编码器的衍生目标（Hill等人，2016）。

ELMo及其前身（Peters等人，2017，2018a）从不同维度拓展了传统词嵌入研究。它们从左至右和从右至左的语言模型中提取*context-sensitive*特征，每个标记的上下文表示是左右双向表示的拼接。将上下文词嵌入与现有任务特定架构结合时，ELMo在多项重要NLP基准测试中实现了技术突破（Peters等人，2018a），包括问答系统（Rajpurkar等人，2016）、情感分析（Socher等人，2013）和命名实体识别（Tjong Kim Sang与De Meulder，2003）。Melamud等人（2016）提出通过LSTM从左右上下文预测单个单词的任务来学习上下文表示。与ELMo类似，他们的模型基于特征且非深度双向。Fedus等人（2018）研究表明完形填空任务可用于提升文本生成模型的鲁棒性。

2.2 无监督微调方法

与基于特征的方法一样，这一方向的首批研究仅从无标注文本中预训练词嵌入参数（Collobert和Weston，2008）。

最近，能够生成上下文标记表示的句子或文档编码器已通过无标签文本进行预训练，并针对有监督的下游任务进行微调（Dai and Le, 2015；Howard and Ruder, 2018；Radford et al., 2018）。这类方法的优势在于只需从头学习少量参数。至少部分得益于这一优势，OpenAI GPT（Radford et al., 2018）在GLUE基准测试（Wang et al., 2018a）的许多句子级任务中取得了当时最先进的结果。从左到右的语言模型—

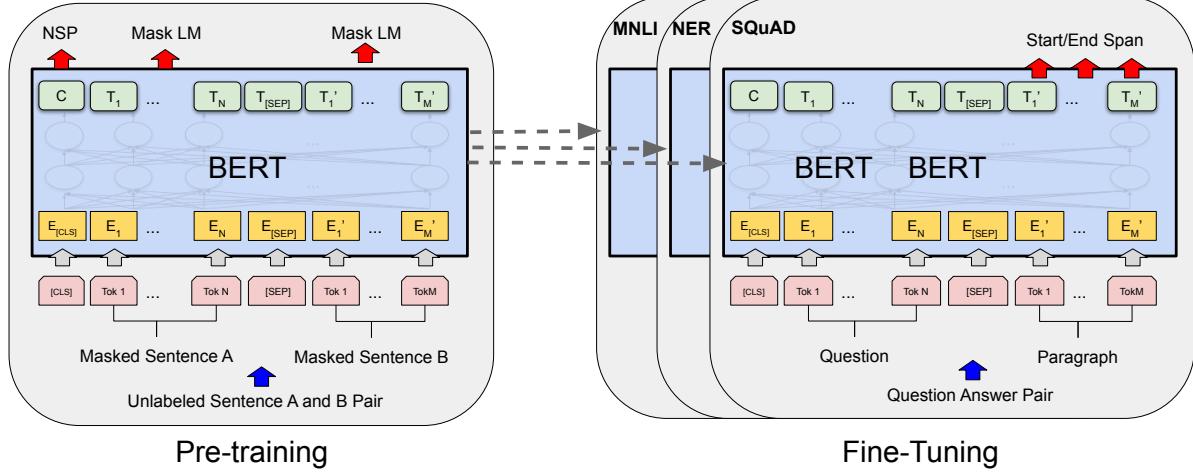


图1：BERT的整体预训练与微调流程。除了输出层外，预训练和微调使用相同的架构结构。相同的预训练模型参数被用于初始化不同下游任务的模型。在微调过程中，所有参数都会被调整。[CLS]是添加在每个输入样本前的特殊符号，而[SEP]是特殊的分隔符标记（例如用于分隔问题/答案）。

语言建模和自编码器目标已被用于此类模型的预训练（Howard和Ruder, 2018; Radford等人, 2018; Dai和Le, 2015）。

2.3 从监督数据中进行迁移学习

已有研究显示，从拥有大型数据集的监督任务（如自然语言推理（Conneau等人, 2017）和机器翻译（McCann等人, 2017））中可以实现有效的迁移。计算机视觉研究同样证明了从大型预训练模型进行迁移学习的重要性，其中一种有效方法是对基于ImageNet（Deng等人, 2009; Yosinski等人, 2014）预训练的模型进行微调。

3 BERT

本节将介绍BERT及其具体实现。我们的框架包含两个步骤：*pre-training*和*fine-tuning*。在预训练阶段，模型通过不同的预训练任务在无标注数据上进行训练。在微调阶段，BERT模型首先用预训练参数初始化，随后使用下游任务的标注数据对所有参数进行微调。每个下游任务都有独立的微调模型，即使它们初始化的预训练参数相同。图1中的问答示例将作为本节的运行案例。

BERT的一个显著特点是其跨不同任务的统一架构。有微

预训练架构与最终下游架构之间的微小差异。

模型架构 BERT的模型架构是一个基于Vaswani等人（2017）所描述并在tensor2tensor库中发布的原始实现的多层双向Transformer编码器。¹由于Transformer的使用已变得普遍，且我们的实现与原始版本几乎相同，我们将省略对模型架构的详尽背景描述，并建议读者参考Vaswani等人（2017）以及《The Annotated Transformer》等优秀指南。²

在本工作中，我们将层数（即Transformer块）表示为 L ，隐藏层大小表示为 H ，自注意力头数表示为 A 。³我们主要报告两种模型规模的结果：BERT_{BASE} ($L=12$, $H=768$, $A=12$ 、总参数量=1.1亿)，以及BERT_{LARGE} ($L=24$, $H=1024$ 、 $A=16$ 、总参数量=3.4亿)。

BERT_{BASE}被选择为与OpenAI GPT具有相同的模型大小以便进行比较。然而，关键区别在于，BERT Transformer采用双向自注意力机制，而GPT Transformer则使用受限的自注意力机制，其中每个标记只能关注其左侧的上下文。⁴

¹<https://github.com/tensorflow/tensor2tensor>

²<http://nlp.seas.harvard.edu/2018/04/03/attention.html>

³In all cases we set the feed-forward/filter size to be $4H$, i.e., 3072 for the $H = 768$ and 4096 for the $H = 1024$.

⁴We note that in the literature the bidirectional Trans-

输入/输出表示 为了使BERT能够处理各种下游任务，我们的输入表示方法能够在一个标记序列中明确表示单个句子和句子对（例如，〈问题，答案〉）。在本研究中，“句子”可以是任意连续的文本片段，而非严格意义上的语言学句子。“序列”指的是输入到BERT的标记序列，它可以是单个句子，也可以是打包在一起的两个句子。

我们使用WordPiece嵌入（Wu等人，2016），词汇量为30,000个词元。每个序列的第一个词元始终是一个特殊的分类标记（[CLS]）。该标记对应的最终隐藏状态被用作分类任务的聚合序列表示。句子对被合并成一个单一序列。我们通过两种方式区分句子：首先，用特殊标记（[SEP]）分隔它们；其次，为每个词元添加一个可学习的嵌入，以指示其属于句子A还是句子B。如图1所示，我们将输入嵌入表示为 E ，特殊[CLS]标记的最终隐藏向量表示为 $C \in \mathbb{R}^H$ ，而 i^{th} 输入词元的最终隐藏向量表示为 $T_i \in \mathbb{R}^H$ 。

对于给定的标记，其输入表示是通过将对应的标记嵌入、段嵌入和位置嵌入相加来构建的。这种构建的可视化展示可见于图2。

3.1 预训练BERT

与Peters等人（2018a）和Radford等人（2018）不同，我们未使用传统的从左到右或从右到左语言模型对BERT进行预训练。相反，我们采用本节描述的两个无监督任务对BERT进行预训练。该步骤展示在图1的左侧部分。

任务 #1：掩码语言模型 直观上，我们有理由相信，深度双向模型严格优于从左到右的模型或从左到右与从右到左模型的浅层拼接。然而，标准的条件语言模型只能以从左到右 or 或从右到左的方式进行训练，因为双向条件作用会使每个词间接地“看到自身”，导致模型在多层上下文中可以轻易预测目标词。

former is often referred to as a “Transformer encoder” while the left-context-only version is referred to as a “Transformer decoder” since it can be used for text generation.

为了训练深度双向表示，我们只需随机掩盖一定比例的输入标记，然后预测这些被掩盖的标记。尽管文献中常将此方法称为Cloze任务（Taylor, 1953），我们仍将其称为“掩码语言模型”（MLM）。在此过程中，对应于掩码标记的最终隐藏向量会像标准语言模型那样，被输入词汇表上的输出softmax层。在所有实验中，我们随机掩盖每个序列中15%的WordPiece标记。与去噪自编码器（Vincent等，2008）不同，我们仅预测被掩盖的词汇而非重构整个输入。

尽管这使我们能够获得一个双向预训练模型，但一个缺点是我们在预训练和微调之间造成了不匹配，因为[MASK]标记在微调期间不会出现。为了缓解这个问题，我们并不总是用实际的[MASK]标记替换“被遮蔽”的词语。训练数据生成器会随机选择15%的标记位置进行预测。如果选择了第 i 个标记，我们将第 i 个标记替换为：(1) 80%的情况下使用[MASK]标记；(2) 10%的情况下使用随机标记；(3) 10%的情况下保持第 i 个标记不变。随后， T_i 将用于通过交叉熵损失预测原始标记。我们在附录C.2中比较了此过程的不同变体。

任务 #2：下一句预测（NSP） 许多重要的下游任务，如问答（QA）和自然语言推理（NLI），都基于理解两个句子之间的relationship，而语言建模无法直接捕捉这种关系。为了训练一个能够理解句子关联的模型，我们针对一个二值化的*next sentence prediction*任务进行预训练，该任务可以轻松从任何单语语料库中生成。具体来说，在为每个预训练示例选择句子A和B时，50%的情况下B是实际紧随A的下一句（标记为IsNext），另外50%的情况下B是语料库中的随机句子（标记为NotNext）。如图1所示， C 被用于下一句预测（NSP）。⁵ 尽管这一任务设计简单，但我们在第5.1节中证明，针对该任务的预训练对QA和NLI都非常有益。⁶

⁵The final model achieves 97%-98% accuracy on NSP.

⁶The vector C is not a meaningful sentence representation without fine-tuning, since it was trained with NSP.

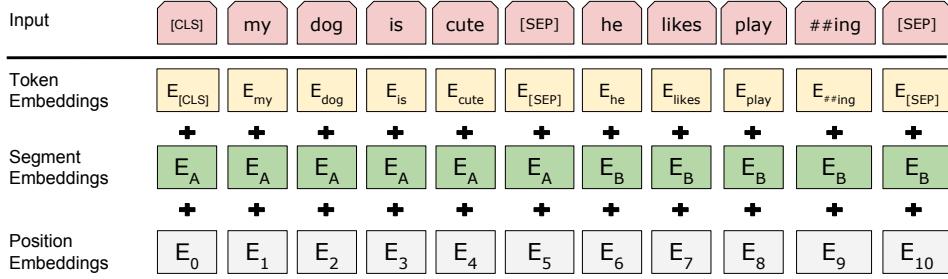


图2：BERT输入表示。输入嵌入是词元嵌入、分段嵌入和位置嵌入的总和。

NSP任务与Jernite等人（2017）以及Logeswara n和Lee（2018）使用的表示学习目标密切相关。然而，在先前的工作中，仅将句子嵌入迁移至下游任务，而BERT则迁移所有参数以初始化终端任务模型参数。

预训练数据 预训练过程主要遵循现有关于语言模型预训练的文献。对于预训练语料库，我们使用了BooksCorpus（8亿词）（Zhu等人，2015年）和英文维基百科（25亿词）。对于维基百科，我们仅提取文本段落，忽略列表、表格和标题。为了提取长连续序列，使用文档级语料库而非如Billion Word Benchmark（Chelba等人，2013年）这类经过打乱的句子级语料库至关重要。

3.2 微调BERT

微调过程十分直接，因为Transformer中的自注意力机制让BERT能够通过替换适当的输入和输出，来建模多种下游任务——无论涉及单个文本还是文本对。对于包含文本对的应用，一种常见模式是在应用双向交叉注意力之前独立编码文本对，例如Parikh等人（2016）和Seo等人（2017）的研究。而BERT则利用自注意力机制统一这两个阶段，因为通过自注意力对拼接文本对进行编码时，实际上已包含两个句子之间的*bidi-directional*交叉注意力。

对于每个任务，我们只需将任务特定的输入和输出插入BERT，并端到端地微调所有参数。在输入层面，预训练中的句子A和句子B分别对应：（1）释义任务中的句子对，（2）蕴含任务中的假设-前提对，（3）问答任务中的问题-段落对，以及

（4）文本分类或序列标注中的一种退化文本- \emptyset 对。在输出端，词元表示被送入输出层以执行词元级任务，如序列标注或问答；而[CLS]表示则被送入输出层以进行分类任务，如蕴含判断或情感分析。

与预训练相比，微调的成本相对较低。本文中的所有结果都可以从完全相同的预训练模型出发，在单个Cloud TPU上至多1小时内复现，或在GPU上花费数小时完成。⁷我们在第4节的相应小节中描述了任务特定的细节。更多细节可在附录A.5中找到。

4 实验

在本节中，我们展示了BERT在11项NLP任务上的微调结果。

4.1 GLUE

通用语言理解评估（GLUE）基准（Wang等人，2018a）是一个包含多样化自然语言理解任务的集合。GLUE数据集的详细描述见附录B.1。

为了在GLUE上进行微调，我们按照第3节所述的方式表示输入序列（针对单句或句对），并使用与第一个输入标记（[CLS]）对应的最终隐藏向量 $C \in \mathbb{R}^H$ 作为聚合表示。微调过程中引入的唯一新参数是分类层权重 $W \in \mathbb{R}^{K \times H}$ ，其中 K 是标签数量。我们使用 C 和 W 计算标准分类损失，即 $\text{log}(\text{softmax}(CW^T))$ 。

⁷For example, the BERT SQuAD model can be trained in around 30 minutes on a single Cloud TPU to achieve a Dev F1 score of 91.0%.

⁸See (10) in <https://gluebenchmark.com/faq>.

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

表1：GLUE测试结果，由评估服务器（<https://gluebenchmark.com/leaderboard>）评分。每个任务下方的数字表示训练样本的数量。“平均”列与官方GLUE分数略有不同，因为我们排除了有问题的WNLI数据集。⁸ BERT和OpenAI GPT是单模型、单任务系统。QQP和MRPC任务报告F1分数，STS-B任务报告斯皮尔曼相关系数，其余任务报告准确率分数。我们排除了使用BERT作为其组成部分的条目。

我们使用32的批次大小，并在所有GLUE任务的数据上进行了3个轮次的微调。针对每个任务，我们在开发集上选择了最佳的微调学习率（从5e-5、4e-5、3e-5和2e-5中选取）。此外，对于BERT_{LARGE}，我们发现微调在小数据集上有时不稳定，因此我们进行了多次随机重启，并在开发集上选择了最佳模型。在随机重启过程中，我们使用相同的预训练检查点，但执行不同的微调数据洗牌和分类层初始化。⁹

结果如表1所示。BERT_{BASE}和BERT_{LARGE}在所有任务上均以显著优势超越所有系统，相较于先前的最优技术分别实现了4.5%和7.0%的平均准确率提升。需注意的是，除注意力掩码机制外，BERT_{BASE}与OpenAI GPT的模型架构几乎完全相同。在规模最大且最常被引用的GLUE任务MNLI上，BERT取得了4.6%的绝对准确率提升。根据官方GLUE排行榜¹⁰截至撰写时的数据，BERT_{LARGE}获得80.5分，而OpenAI GPT的得分为72.8。

我们发现，BERT_{LARGE}在所有任务上的表现都显著优于BERT_{BASE}，尤其是在训练数据非常有限的任务中。模型规模的影响将在第5.2节进行更深入的探讨。

4.2 SQuAD v1.1

斯坦福问答数据集（SQuAD {v*}1.1）是一个包含10万条众包问答对的数据集（Rajpurkar等人，2016）。给定一个问题及一篇来自

⁹The GLUE data set distribution does not include the Test labels, and we only made a single GLUE evaluation server submission for each of BERT_{BASE} and BERT_{LARGE}.

¹⁰<https://gluebenchmark.com/leaderboard>

包含答案的维基百科文章，任务是在文章中预测答案文本的范围。

如图1所示，在问答任务中，我们将输入问题与文本段落表示为单个打包序列，其中问题使用A嵌入表示，段落使用B嵌入表示。我们仅在微调阶段引入起始向量 $S \in \mathbb{R}^H$ 和结束向量 $E \in \mathbb{R}^H$ 。词语*i*作为答案跨度起始位置的概率，通过计算 T_i 与 S 的点积后，对段落中所有词语进行softmax归一化得到： $P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$ 。答案跨度结束位置的计算采用类似公式。候选跨度从位置*i*到位置*j*的得分定义为 $S \cdot T_i + E \cdot T_j$ ，并以 $j \geq i$ 作为预测依据选取最高得分的跨度。训练目标为正确起始位置与结束位置的对数似然之和。我们采用5e-5的学习率和32的批大小进行3个周期的微调。

表2展示了排行榜前列的条目以及顶尖已发表系统的结果（Seo等人，2017；Clark与Gardner，2018；Peters等人，2018a；Hu等人，2018）。SQuAD排行榜上的顶尖结果目前尚无最新的公开系统描述可用，¹¹且这些系统在训练时允许使用任何公开数据。因此，我们在系统中采用了适度的数据增强方法，即先在TriviaQA（Joshi等人，2017）上进行微调，再对SQuAD进行微调。

我们表现最佳的集成系统比排行榜上的顶级系统高出+1.5个F1值，而作为单一系统则高出+1.3个F1值。实际上，我们的单一BERT模型在F1分数上甚至超越了顶级的集成系统。在没有经过TriviaQA微调的情况下——

¹¹QANet is described in Yu et al. (2018), but the system has improved substantially after publication.

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
BERT _{BASE} (Single)	80.8	88.5	-	-
BERT _{LARGE} (Single)	84.1	90.9	-	-
BERT _{LARGE} (Ensemble)	85.8	91.8	-	-
BERT _{LARGE} (Sgl.+TriviaQA)	84.2	91.1	85.1	91.8
BERT _{LARGE} (Ens.+TriviaQA)	86.2	92.2	87.4	93.2

表2: SQuAD 1.1结果。BERT集成系统包含7个使用不同预训练检查点与微调种子的系统。

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-	-	71.4	74.4
Ours				
BERT _{LARGE} (Single)	78.7	81.9	80.0	83.1

表3: SQuAD 2.0结果。我们排除了使用BERT作为其组件之一的条目。

在调优数据上，我们仅损失0.1-0.4 F1值，仍以显著优势超越所有现有系统。¹²

4.3 SQuAD v2.0

SQuAD 2.0任务扩展了SQuAD 1.1的问题定义，允许在提供的段落中可能不存在简短答案，从而使问题更加贴近现实。

我们采用一种简单的方法来扩展SQuAD v1.1 BERT模型以完成此任务。我们将没有答案的问题视为其答案跨度起始和结束于[CLS]标记。答案跨度起始和结束位置的概率空间被扩展至包含[CLS]标记的位置。在预测时，我们将无答案跨度的得分： $s_{\text{null}} = S \cdot C + E \cdot C$ 与最佳非空跨度的得分进行比较。

¹²The TriviaQA data we used consists of paragraphs from TriviaQA-Wiki formed of the first 400 tokens in documents, that contain at least one of the provided possible answers.

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
BERT _{BASE}	81.6	-
BERT _{LARGE}	86.6	86.3
Human (expert) [†]	-	85.0
Human (5 annotations) [†]	-	88.0

表4: SWAG开发和测试准确率。[†]人类表现是通过100个样本测量的，如SWAG论文中所述。

$\hat{s}_{i,j} = \max_{j \geq i} S \cdot T_i + E \cdot T_j$ 我们预测当 $\hat{s}_{i,j} > s_{\text{null}} + \tau$ 时存在非空答案，其中阈值 τ 是在开发集上为最大化F1值而选定的。该模型未使用TriviaQA数据。我们以5e-5的学习率和48的批量大小进行了2个周期的微调。

与先前排行榜条目及顶尖已发表工作 (Sun等人, 2018; Wang等人, 2018b) 相比的结果展示于表3，其中排除了使用BERT作为组件的系统。我们观察到相较于先前最佳系统有+5.1 F1值的提升。

4.4 SWAG

对抗生成情境 (SWAG) 数据集包含113k个句子对补全示例，用于评估基于常识的推理能力 (Zellers等人, 2018)。给定一个句子，任务是从四个选项中选择最合理的后续内容。

在SWAG数据集上进行微调时，我们构建四个输入序列，每个序列包含给定句子（句子A）与一个可能后续句（句子B）的拼接。引入的唯一任务特定参数是一个向量，其与[CLS]标记表示C的点积表示每个选项的得分，该得分通过softmax层进行归一化处理。

我们对模型进行了3个周期的微调，学习率为2e-5，批处理大小为16。结果如表4所示。BERT_{LARGE}的表现优于作者基线ESIM+ELMo系统+27.1%，并超过OpenAI GPT 8.3%。

5 消融研究

在本节中，我们对BERT的多个方面进行了消融实验，以更好地理解它们的相对重要性。此外

Tasks	Dev Set				
	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
BERT _{BASE}	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

表5：使用BERT_{BASE}架构对预训练任务进行消融实验。“No NSP”表示训练时不包含下一句预测任务。“LTR & No NSP”表示像OpenAI GPT一样，作为从左到右的语言模型进行训练，且不包含下一句预测。“+ BiLSTM”表示在微调阶段，在“LT R + No NSP”模型顶部添加一个随机初始化的BiLSTM。

消融研究可在附录C中找到。

5.1 预训练任务的影响

我们通过使用与BERT_{BASE}完全相同的预训练数据、微调方案和超参数来评估两个预训练目标，以此证明BERT深度双向性的重要性：

无NSP: 一种双向模型，采用“掩码语言模型”(MLM)进行训练，但不包含“下一句预测”(NSP)任务。

LTR & 无NSP: 一种仅使用标准从左到右(LTR)语言模型进行训练的纯左上下文模型，而非掩码语言模型。在微调阶段同样应用了仅左向约束，因为取消该约束会导致预训练与微调不匹配，从而降低下游任务性能。此外，该模型在预训练阶段未使用NSP任务。该模型可直接与OpenAI GPT进行对比，但采用了我们更大的训练数据集、输入表示方法及微调方案。

我们首先考察了NSP任务带来的影响。在表5中，我们发现移除NSP会显著降低QNLI、MNLI和SQuAD 1.1的性能。接着，我们通过比较“无NSP”与“仅左向语境 & 无NSP”来评估训练双向表示的影响。左向语境模型在所有任务上的表现均差于MLM模型，其中在MRPC和SQuAD任务上性能下降尤为明显。

对于SQuAD数据集，直观上可以明显看出，LTR模型在词元预测方面表现会很差，因为词元级别的隐藏状态缺乏右侧上下文。为了切实尝试增强LTR系统，我们在其顶部添加了一个随机初始化的BiLSTM。这确实显著提升了SQuAD上的结果，但

结果仍然远不如预训练的双向模型。BiLSTM在GLUE任务上的表现反而有所下降。

我们认识到，也可以像ELMo那样训练独立的LTR和RTL模型，并将每个标记表示为两个模型的拼接。然而：(a) 这比单一的双向模型成本高一倍；(b) 对于像QA这样的任务来说，这并不直观，因为RTL模型无法根据问题来调整答案；(c) 这严格来说不如深度双向模型强大，因为后者能在每一层同时利用左右上下文。

5.2 模型规模的影响

在本节中，我们探讨模型规模对微调任务准确性的影响。我们训练了多个具有不同层数、隐藏单元和注意力头数量的BERT模型，同时保持其他超参数和训练流程与先前描述的一致。

在选定的GLUE任务上的结果如表6所示。在此表中，我们报告了5次随机重启微调的平均开发集准确率。可以看出，更大的模型在所有四个数据集上都带来了严格的准确率提升，即使是对于仅有3600个标注训练样本、且与预训练任务存在显著差异的MRPC数据集也是如此。另一个可能令人惊讶的发现是，我们能够在已经相对现有文献相当庞大的模型基础上取得如此显著的改进。例如，Vaswani等人(2017)研究中探索的最大Transformer结构为(L=6, H=1024, A=16)，编码器包含1亿参数；而我们在文献中找到的最大Transformer结构为(L=64, H=512, A=2)，包含2.35亿参数(Ai-Rou等人，2018)。相比之下，BERT_{BASE}包含1.1亿参数，BERT_{LARGE}则包含3.4亿参数。

长期以来，人们已经认识到增加模型规模会持续提升机器翻译和语言建模等大规模任务的性能，这一点通过表6所示的训练数据LM困惑度得以验证。然而，我们相信这是首个令人信服地证明：只要模型经过充分预训练，即使扩展到极大规模，也能在极小规模任务上带来显著提升的研究。Peters等人(2018b)的研究表明

在增加预训练双向语言模型规模从两层到四层对下游任务的影响方面，结果不一；而Melamud等人（2016）曾简要提及，将隐藏层维度从200增至600有所助益，但继续提升至1000并未带来进一步改善。这两项先前研究均采用基于特征的方法——我们推测，当模型直接在下游任务上进行微调，且仅使用极少量随机初始化的附加参数时，即使下游任务数据量非常有限，任务专用模型仍能受益于更庞大、表达能力更强的预训练表征。

5.3 基于BERT的特征方法

迄今为止展示的所有BERT结果都采用了微调方法，即在预训练模型基础上添加简单的分类层，并针对下游任务对所有参数进行联合微调。然而，基于特征的方法——从预训练模型中提取固定特征——具有特定优势。首先，并非所有任务都能轻易用Transformer编码器架构表示，因此需要添加任务特定的模型架构。其次，预算算训练数据的高成本表征具有显著计算优势：只需计算一次昂贵的数据表征，即可基于该表征使用更轻量的模型进行大量实验。

在本节中，我们通过将BERT应用于CoNLL-2003命名实体识别（NER）任务（Tjong Kim Sang和De Meulder, 2003）来比较这两种方法。在BERT的输入中，我们使用保留大小写的WordPiece模型，并包含数据提供的最大文档上下文。遵循标准做法，我们将其表述为标注任务，但不使用条件随机场（CRF）。

Hyperparams			Dev Set Accuracy			
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

表6：BERT模型规模的消融实验。#L = 层数；#H = 隐藏层大小；#A = 注意力头数量。“LM (ppl)”指掩码语言模型在预留训练数据上的困惑度。

System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	93.1
Fine-tuning approach		
BERT _{LARGE}	96.6	92.8
BERT _{BASE}	96.4	92.4
Feature-based approach (BERT _{BASE})		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

表7：CoNLL-2003命名实体识别结果。超参数通过开发集选定。所报告的开发集和测试集分数是使用这些超参数进行5次随机重启后的平均值。

在输出层中，我们使用第一个子标记的表示作为输入，通过NER标签集上的标记级分类器进行处理。

为了消融微调方法，我们采用基于特征的方法，通过从一个或多个层提取激活值*without*，而不微调BERT的任何参数。这些上下文嵌入被用作随机初始化的两层768维BiLSTM的输入，然后接入分类层。

结果如表7所示。BERT_{LARGE}与最先进方法相比表现出竞争力。最佳性能的方法将预训练Transformer顶部四个隐藏层的标记表示进行拼接，其结果仅比微调整个模型低0.3个F1值。这表明BERT对于微调和基于特征的方法都同样有效。

6 结论

最近，基于语言模型的迁移学习带来的实证改进表明，丰富的无监督预训练是许多语言理解系统不可或缺的一部分。特别是，这些成果使得即使是低资源任务也能从深度单向架构中受益。我们的主要贡献在于进一步将这些发现推广到深度*bidirectional*架构，使得同一个预训练模型能够成功应对广泛的自然语言处理任务。

参考文献

- Alan Akbik、Duncan Blythe 和 Roland Vollgraf。2018。用于序列标注的上下文字符串嵌入。收录于 *Proceedings of the 27th International Conference on Computational Linguistics*, 第 1638–1649 页。
- Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo 和 Llion Jones。2018。基于更深层次自注意力的字符级语言建模。arXiv preprint arXiv:1808.04444。
- 安藤理惠和张彤。2005年。从多任务和无标签数据中学习预测结构的框架。*Journal of Machine Learning Research*, 6(11月):1817–1853。
- Luisa Bentivogli、Bernardo Magnini、Ido Dagan、Hoà Trang Dang 和 Danilo Giampiccolo。2009年。第五届PASCAL文本蕴含识别挑战赛。收录于 *TAC-NIST*。
- John Blitzer, Ryan McDonald 和 Fernando Pereira。2006年。基于结构对应学习的领域自适应。载于 *Proceedings of the 2006 conference on empirical methods in natural language processing*, 第120–128页。计算语言学协会。
- Samuel R. Bowman、Gabor Angeli、Christopher Potts 和 Christopher D. Manning。2015。一个用于学习自然语言推理的大规模标注语料库。载于 *EMNLP*。计算语言学协会。
- 彼得·F·布朗、彼得·V·德索萨、罗伯特·L·默瑟、文森特·J·德拉彼得拉与詹妮弗·C·赖。1992年。基于类的自然语言n元模型。*Computational linguistics*, 18(4): 467–479。
- Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 任务 1：语义文本相似性多语言与跨语言重点评估。载于 *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 第 1–14 页，加拿大温哥华。计算语言学协会。
- Ciprian Chelba、Tomas Mikolov、Mike Schuster、齐格、Thorsten Brants、Phillipp Koehn 与 Tony Robinson。2013年。用于衡量统计语言建模进展的十亿词基准测试。arXiv preprint arXiv:1312.3005。
- Z. Chen, H. Zhang, X. Zhang, and L. Zhao. 2018. Quora问题对。
- 克里斯托弗·克拉克和马特·加德纳。2018年。简单有效的多段落阅读理解。于 *ACL*。
- 凯文·克拉克、Minh-Thang Luong、克里斯托弗·D·曼宁与郭去疾。2018。基于交叉视角训练的半监督序列建模。发表于 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 第1914–1925页。
- Ronan Collobert 和 Jason Weston。2008。自然语言处理的统一架构：采用多任务学习的深度神经网络。于 *Proceedings of the 25th international conference on Machine learning*, 第160–167页。ACM。
- Alexis Conneau、Douwe Kiela、Holger Schwenk、Loïc Barrault 与 Antoine Bordes。2017。基于自然语言推理数据的通用句子表示监督学习。收录于 *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 第670–680页，丹麦哥本哈根。计算语言学协会。
- Andrew M Dai 和 Quoc V Le。2015。半监督序列学习。于 *Advances in neural information processing systems*, 第3079–3087页。
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: 一个大规模分层图像数据库。于 *CVPR09*。
- William B Dolan 与 Chris Brockett。2005。自动构建句子复述语料库。发表于 *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*。
- William Fedus、Ian Goodfellow 和 Andrew M Dai。2018。Maskgan：通过填充 $\{v^*\}$ 实现更好的文本生成。arXiv preprint arXiv:1801.07736。
- Dan Hendrycks 与 Kevin Gimpel。2016。通过高斯误差线性单元连接非线性与随机正则化器。CoRR, abs/1606.08415。
- Felix Hill、Kyunghyun Cho 和 Anna Korhonen。2016。从未标注数据中学习句子的分布式表示。于 *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*。计算语言学协会。
- Jeremy Howard 与 Sebastian Ruder。2018。面向文本分类的通用语言模型微调。发表于 *ACL*。计算语言学协会。
- 胡明浩、彭宇星、黄震、邱锡鹏、韦福如、周明。2018。用于机器阅读理解任务的强化记忆网络。发表于 *IJCAI*。
- Yacine Jernite、Samuel R. Bowman 与 David Sontag。2017。基于语篇的目标用于快速无监督句子表示学习。CoRR, abs/1705.00557。

Mandar Joshi、Eunsol Choi、Daniel S Weld和Luke Zettlemoyer。2017。Triviaqa：一个用于阅读理解的大规模远程监督挑战数据集。于*ACL*。Ryan Kiros、Yukun Zhu、Ruslan R Salakhutdinov、Richard Zweig、Raquel Urtasun、Antonio Torralba和Sanja Fidler。2015。Skip-thought向量。于*Advances in neural information processing systems*, 第3294–3302页。Quoc Le和Tomas Mikolov。2014。句子和文档的分布式表示。于*International Conference on Machine Learning*, 第1188–1196页。Hector J Levesque、Ernest Davis和Leora Morgenstern。2011。Winograd模式挑战。于*Aaaai spring symposium: Logical formalizations of commonsense reasoning*, 第46卷, 第47页。Lajanugen Logeswaran和Honglak Lee。2018。一种学习句子表示的高效框架。于*International Conference on Learning Representations*。Bryan McCann、James Bradbury、Caiming Xiong和Richard Socher。2017。在翻译中学习：上下文化的词向量。于*NIPS*。Oren Melamed、Jacob Goldberger和Ido Dagan。2016。context 2vec：使用双向LSTM学习通用上下文嵌入。于*CoNLL*。Tomas Mikolov、Ilya Sutskever、Kai Chen、Greg S Corrado和Jeff Dean。2013。词和短语的分布式表示及其组合性。于*Advances in Neural Information Processing Systems 26*, 第3111–3119页。Curran Associates, Inc. Andriy Mnih和Geoffrey E Hinton。2009。一种可扩展的层次化分布式语言模型。于D. Koller、D. Schuurmans、Y. Bengio和L. Bottou编辑, *Advances in Neural Information Processing Systems 21*, 第1081–1088页。Curran Associates, Inc. Ankur P Parikh、Oscar Täckström、Dipanjan Das和Jakob Uszkoreit。2016。一种用于自然语言推理的可分解注意力模型。于*EMNLP*。Jeffrey Pennington、Richard Socher和Christopher D. Manning。2014。Glove：用于词表示的全局向量。于*Empirical Methods in Natural Language Processing (EMNLP)*, 第1532–1543页。Matthew Peters、Waleed Ammar、Chandra Bhagavatula和Russell Power。2017。使用双向语言模型的半监督序列标注。于*ACL*。Matthew Peters、Mark Neumann、Mohit Iyyer、Matt Gardner、Christopher Clark、Kenton Lee和Luke Zettlemoyer。2018a。深度上下文化的词表示。于*NAACL*。

Matthew Peters、Mark Neumann、Luke Zettlemoyer 和 Wen-tau Yih。2018b。剖析上下文词嵌入：架构与表示。于*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 第1499–1510页。Alec Radford、Karthik Narasimhan、Tim Salimans 和 Ilya Sutskever。2018。通过无监督学习提升语言理解能力。技术报告, OpenAI。Pranav Rajpurkar、Jian Zhang、Konstantin Lopyrev 和 Percy Liang。2016。SQuAD：面向文本机器理解的10万+个问题。于*Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 第2383–2392页。Minjoon Seo、Aniruddha Kembhavi、Ali Farhadi 和 Hannaneh Hajishirzi。2017。用于机器理解的双向注意力流。于*ICLR*。Richard Socher、Alex Perelygin、Jean Wu、Jason Chuang、Christopher D Manning、Andrew Ng 和 Christopher Potts。2013。基于情感树库语义组合性的递归深度模型。于*Proceedings of the 2013 conference on empirical methods in natural language processing*, 第1631–1642页。Fu Sun、Linyang Li、Xipeng Qi 和 Yang Liu。2018。U-Net：针对无法回答问题的机器阅读理解。*arXiv preprint arXiv:1810.06638*。Wilson L Taylor。1953。完形填空程序：一种衡量可读性的新工具。*Journalism Bulletin*, 30(4):415–433。Erik F Tjong Kim Sang 和 Fien De Meulder。2003。CONLL-2003共享任务介绍：语言无关的命名实体识别。于*CoNLL*。Joseph Turian、Lev Ratinov 和 Yoshua Bengio。2010。词表示：一种简单通用的半监督学习方法。于*Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, 第384–394页。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017。注意力机制就是全部所需。于*Advances in Neural Information Processing Systems*, 第6000–6010页。Pascal Vincent、Hugo Larochelle、Yoshua Bengio 和 Pierre-Antoine Manzagol。2008。通过去噪自编码器提取与组合鲁棒特征。于*Proceedings of the 25th international conference on Machine learning*, 第1096–1103页。ACM。Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel Bowman。2018a。GLUE：一个多任务基准测试与分析平台

用于自然语言理解。在 *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* 中, 第353至355页。

王伟、严明和吴晨。2018b。用于阅读理解和问答的多粒度分层注意力融合网络。于 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*。计算语言学协会。

亚历克斯·沃斯塔特、阿曼普里特·辛格和塞缪尔·R·鲍曼。2018年。神经网络可接受性判断。*arXiv preprint arXiv:1805.12471*。

Adina Williams、Nikita Nangia与Samuel R Bowman。2018。一个用于通过推理理解句子的广覆盖挑战语料库。于NAACL。

吴永辉、Mike Schuster、陈智峰、Quoc V Le、Mohammad Norouzi、Wolfgang Macherey、Maxim Krikun、曹原、高勤、Klaus Macherey 等。2016。谷歌的神经机器翻译系统：弥合人类与机器翻译之间的差距。*arXiv preprint arXiv:1609.08144*。

Jason Yosinski、Jeff Clune、Yoshua Bengio 与 Hod Lipson。2014。深度神经网络中的特征可迁移性如何？载于 *Advances in neural information processing systems*, 第 3320–3328 页。

Adams Wei Yu、David Dohan、Minh-Thang Luong、Rui Zhao、Kai Chen、Mohammad Norouzi 和 Quoc V Le。2018。QANet：结合局部卷积与全局自注意力机制用于阅读理解。于 *ICLR*。

Rowan Zellers、Yonatan Bisk、Roy Schwartz 和 Yejin Choi。2018。Swag：一个用于基础常识推理的大规模对抗性数据集。发表于 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*。

Yukun Zhu、Ryan Kiros、Rich Zemel、Ruslan Salakhutdinov、Raquel Urtasun、Antonio Torralba 和 Sanja Fidler。2015。对齐书籍与电影：通过观看电影与阅读书籍实现故事化视觉解释。发表于 *Proceedings of the IEEE international conference on computer vision*, 第 19–27 页。

《BERT：用于语言理解的深度双向Transformer预训练》附录

我们将附录组织为三个部分：

- BERT的额外实现细节在附录A中给出；

- 我们实验的更多细节在附录B中给出；并且
- 额外的消融研究见附录C。

我们针对BERT进行了额外的消融研究，包括：– 训练步数的影响；以及 – 不同掩码处理方法的消融分析。

BERT 的额外细节

A.1 预训练任务示意图

我们在下面提供预训练任务的示例。

掩码语言模型与掩码过程假设未标记的句子是“my dog is hairy”，在随机掩码过程中我们选择了第4个标记（对应“hairy”），我们的掩码过程可以进一步通过以下方式说明：

- 80% 的情况下：将单词替换为 [MASK] 标记，例如，my dog is hairy → my dog is [MASK]
- 10% 的情况下：用随机词替换该词，例如，我的狗毛茸茸的 → 我的狗是苹果
- 10% 的情况下：保持单词不变，例如，my dog is hairy → my dog is hairy。这样做的目的是使表征偏向实际观察到的单词。

该方法的优势在于，Transformer编码器不知道哪些词将被要求预测，或者哪些词已被随机词替换，因此它必须为每个输入标记保留一个分布式的上下文表示。此外，由于随机替换仅发生在所有标记的1.5%（即15%的10%），这似乎不会损害模型的语言理解能力。在C.2节中，我们评估了该程序的影响。

与标准语言模型训练相比，掩码语言模型仅对每个批次中15%的标记进行预测，这表明模型可能需要更多的预训练步骤。

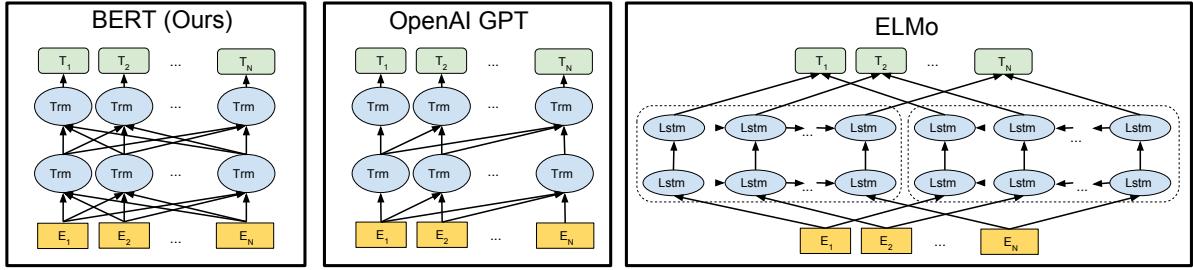


图3：预训练模型架构的差异。BERT使用双向Transformer。OpenAI GPT使用从左到右的Transformer。ELMo使用独立训练的从左到右和从右到左LSTM的拼接来为下游任务生成特征。在这三者中，只有BERT表示在所有层中同时以左右上下文为条件。除了架构差异外，BERT和OpenAI GPT是微调方法，而ELMo是基于特征的方法。

收敛。在C.1节中，我们证明MLM的收敛速度确实略慢于从左到右的模型（后者预测每个词元），但MLM模型带来的实证改进远远超过了训练成本的增加。

下一句预测 下一句预测任务可以通过以下示例进行说明。

```
输入 = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP] 标签 = IsNext
输入 = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP] 标签
= NotNext
```

A.2 预训练流程

为了生成每个训练输入序列，我们从语料库中采样两段文本（尽管它们通常比单个句子长得多，但也可以更短，我们仍将其称为“句子”）。第一句接收A嵌入，第二句接收B嵌入。在50%的情况下，B是实际接在A之后的下一句，另外50%的情况下它是一个随机句子，这是为“下一句预测”任务而设计的。采样时确保组合长度不超过 ≤ 512 个标记。在应用Word Piece分词后，以15%的统一掩码率进行语言模型掩码处理，且不对部分词片段作特殊考虑。

我们以256个序列的批次大小（256个序列 * 512个标记 = 128,000个标记/批次）训练了1,000,000步，这大约相当于40

在33亿单词的语料库上训练多个周期。我们采用Adam优化器，学习率为 $1e-4$ ， $\beta_1 = 0.9$ ， $\beta_2 = 0.999$ ，L2权重衰减为0.01，学习率在前10,000步进行预热，之后线性衰减。所有层使用0.1的dropout概率。遵循OpenAI GPT的方法，我们使用gelu激活函数（Hendrycks和Gimpel, 2016）而非标准的relu。训练损失由掩码语言模型似然的均值与下一句预测似然的均值相加构成。

BERT_{BASE}的训练在4个Cloud TPU上以Pod配置（共16个TPU芯片）进行。¹³ BERT_{LARGE}的训练在16个Cloud TPU（共64个TPU芯片）上完成。每次预训练需要4天时间。

较长的序列成本不成比例地高昂，因为注意力机制的计算复杂度与序列长度呈平方关系。为了在我们的实验中加速预训练，我们先用128的序列长度预训练模型90%的步骤，然后用512的序列长度训练剩余10%的步骤，以学习位置嵌入。

A.3 微调程序

对于微调，大多数模型超参数与预训练时保持一致，除了批大小、学习率和训练周期数。dropout概率始终保持在0.1。最佳超参数值因任务而异，但我们发现以下取值范围在所有任务中都能取得良好效果：

- 批次大小：16、32

¹³<https://cloudplatform.googleblog.com/2018/06/Cloud-TPU-now-offers-preemptible-pricing-and-global-availability.html>

- 学习率 (Adam) : 5e-5、3e-5、2e-5
- 训练轮数: 2、3、4

我们还观察到，大型数据集（例如10万+个带标签的训练样本）对超参数选择的敏感度远低于小型数据集。微调通常非常快，因此合理做法是直接对上述参数进行穷举搜索，并选择在开发集上表现最佳的模型。

A.4 BERT、ELMo 与 OpenAI GPT 的对比

这里我们研究了近期流行的表示学习模型之间的差异，包括ELMo、OpenAI GPT和BERT。模型架构之间的对比在图3中直观展示。需要注意的是，除了架构差异外，BERT和OpenAI GPT采用微调方法，而ELMo则是基于特征的方法。

与BERT最为接近的现有预训练方法是Open AI GPT，后者在大规模文本语料上训练了一个从左到右的Transformer语言模型。实际上，BERT的许多设计决策都刻意保持与GPT高度一致，以便两种方法能在最小变量下进行对比。本文的核心论点是：双向性特征及第3.1节提出的两项预训练任务构成了实证改进的主要来源，但我们也注意到BERT与GPT在训练方式上还存在若干其他差异：

- GPT在BooksCorpus（8亿单词）上训练；BERT在BooksCorpus（8亿单词）和维基百科（25亿单词）上训练。
- GPT在微调时引入了句子分隔符（[SEP]）和分类器标记（[CLS]）；而BERT在预训练期间就学习了[SEP]、[CLS]以及句子A/B嵌入。
- GPT以32,000词的批量大小训练了100万步；BERT以128,000词的批量大小训练了100万步。
- GPT在所有微调实验中均采用5e-5的相同学习率；BERT则选择在开发集上表现最佳的任务特定微调学习率。

为了分离这些差异的影响，我们在5.1节进行了消融实验，结果表明大部分改进实际上源于两个预训练任务及其所实现的双向性。

A.5 不同任务上的微调示例

图4展示了在不同任务上对BERT进行微调的示意图。我们的任务特定模型通过将BERT与一个额外的输出层结合而形成，因此只需从头学习最少量的参数。在这些任务中，(a)和(b)是序列级任务，而(c)和(d)是词元级任务。图中， E 代表输入嵌入， T_i 代表词元*i*的上下文表示，[CLS]是分类输出的特殊符号，[SEP]则是用于分隔非连续词元序列的特殊符号。

B 详细实验设置

B.1 GLUE基准实验的详细描述。

我们在表1中的GLUE结果来自<https://gluebenchmark.com/排行榜>和<https://blog.openai.com/language-unsupervised>。GLUE基准包含以下数据集，其描述最初由Wang等人（2018a）总结：

MNLI多类型自然语言推理是一项大规模众包蕴含关系分类任务（Williams等人，2018）。给定一对句子，其目标是预测第二句话相对于第一句话是*entailment*、*contradiction*还是*neutral*。

QQP Quora问题对是一项二元分类任务，其目标是判断Quora上提出的两个问题在语义上是否等价（Chen等人，2018年）。

QNLI（问题自然语言推理）是斯坦福问答数据集（Rajpurkar等人，2016）的一个版本，该数据集已被转化为二元分类任务（Wang等人，2018a）。正例是包含正确答案的（问题，句子）对，负例则来自同一段落中不包含答案的（问题，句子）对。

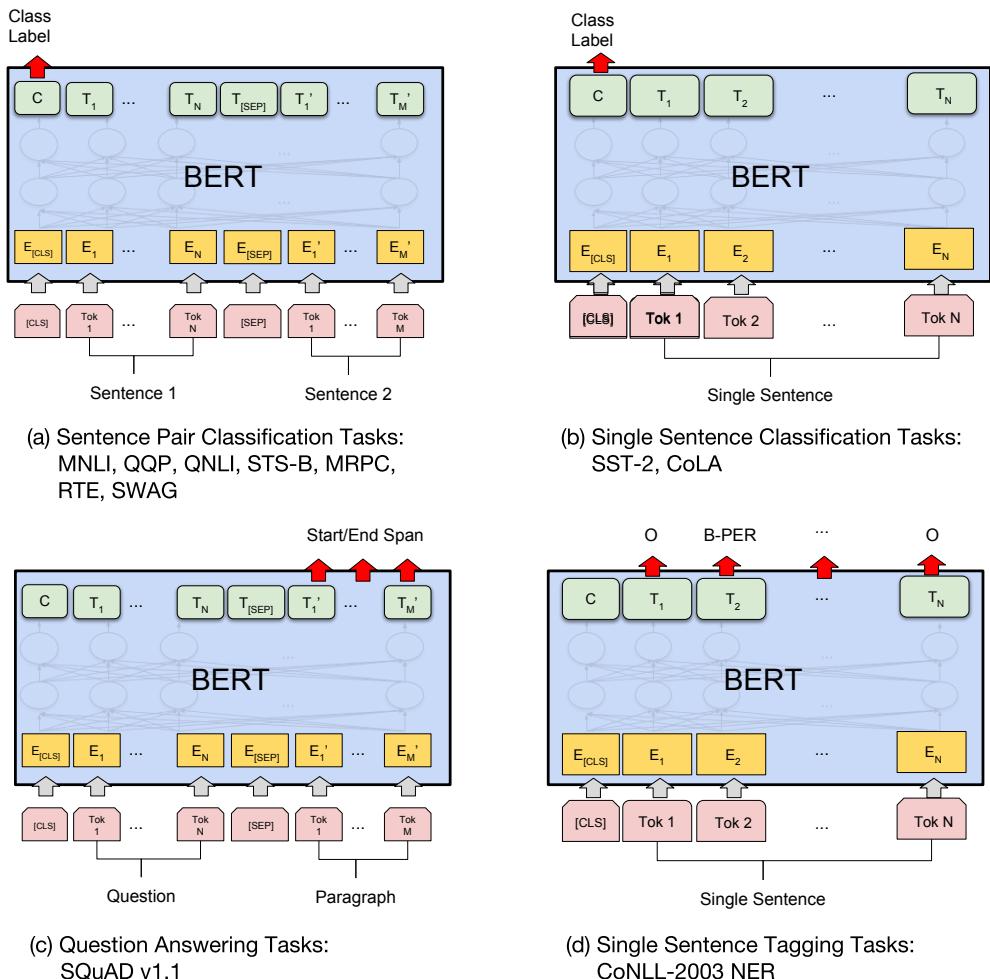


图4：在不同任务上微调BERT的示意图。

SST-2 斯坦福情感树库是一个二元单句分类任务，由从电影评论中提取的句子组成，并包含人工标注的情感标签（Socher et al., 2013）。

CoLA（语言可接受性语料库）是一个二元的单句分类任务，其目标是预测一个英语句子在语言学上是否“可接受”（Warstadt等人，2018年）。

STS-B 语义文本相似性基准测试集是一个从新闻标题和其他来源收集的句子对集合（Cer等人，2017年）。这些句子对通过1到5的评分进行标注，以表示两个句子在语义上的相似程度

◦

MRPC 微软研究释义语料库由从在线新闻源自动提取的句子对组成，并经过人工标注。

对于句子对中的句子是否语义等价（Dolan和Brockett, 2005年）。

RTE (Recognizing Textual Entailment, 文本蕴含识别) 是一项与MNLI类似的二元蕴含任务，但其训练数据量要少得多 (Bentivogli等人, 2009年)¹⁴。

WNLI Winograd NLI 是一个小型自然语言推理数据集 (Levesque等人, 2011)。GLUE 网页指出该数据集的构建存在问题, {v*} 并且所有提交至 GLUE 的已训练系统的表现都低于预测多数类别的 65.1% 基线准确率。因此，为公平对待 OpenAI GPT，我们排除了该数据集。在我们的 GLUE 提交结果中，始终预测多

¹⁴Note that we only report single-task fine-tuning results in this paper. A multitask fine-tuning approach could potentially push the performance even further. For example, we did observe substantial improvements on RTE from multi-task training with MNLI.

¹⁵<https://gluebenchmark.com/faq>

jority class.

C 额外的消融研究

C.1 训练步数的影响

图5展示了从经过 k 步预训练的检查点进行微调后的MNLI开发集准确率。这使我们能够回答以下问题：

1. 问题：BERT是否真的需要如此大量的预训练（128,000词/批次 * 1,000,000步）才能达到较高的微调精度？

答案：是的，与训练50万步相比，BERT_{BASE}在训练100万步时，在MNLI上获得了近1.0%的额外准确率。

2. 问题：由于每个批次中只有15%的单词被预测，而不是每个单词都被预测，MLM预训练是否比LTR预训练收敛得更慢？

答案：MLM模型的收敛速度确实略慢于LTR模型。然而，在绝对准确度方面，MLM模型几乎立即开始超越LTR模型。

C.2 不同掩码程序的消融研究

在3.1节中我们提到，BERT在使用掩码语言模型（MLM）目标进行预训练时，对目标标记采用了混合掩码策略。以下是一项消融研究，旨在评估不同掩码策略的效果。

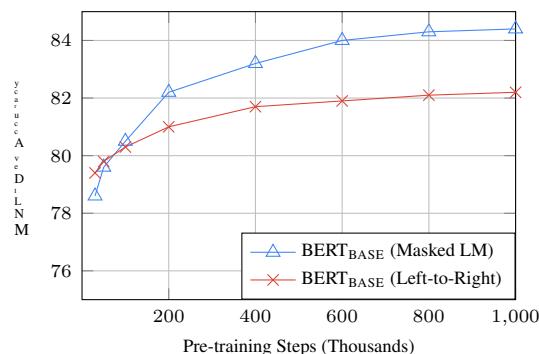


图5：训练步数的消融实验。这展示了从预训练了 k 步的模型参数开始微调后，MNLI的准确率。x轴表示 k 的值。

需要注意的是，掩码策略的目的是减少预训练与微调之间的不匹配，因为[MASK]符号在微调阶段从未出现。我们报告了MNLI和NER的Dev结果。对于NER，我们同时报告了微调和基于特征的方法，因为我们预计这种不匹配在基于特征的方法中会被放大，因为模型将没有机会调整表示。

MASK	Masking Rates			Dev Set Results		
	SAME	RND	MNLI	Fine-tune	NER	Feature-based
80%	10%	10%	84.2	95.4	94.9	
100%	0%	0%	84.3	94.9	94.0	
80%	0%	20%	84.1	95.2	94.6	
80%	20%	0%	84.4	95.2	94.7	
0%	20%	80%	83.7	94.8	94.6	
0%	0%	100%	83.6	94.9	94.6	

表8：不同掩码策略的消融实验。

结果如表8所示。表中，MASK表示我们将目标标记替换为[MASK]符号以进行MLM；SAME表示我们保持目标标记不变；RND表示我们将目标标记替换为另一个随机标记。

表格左侧的数字代表了在MLM预训练期间使用的特定策略的概率（BERT使用80%、10%、10%）。论文的右侧部分代表了开发集的结果。对于基于特征的方法，我们将BERT的最后4层连接起来作为特征，这在第5.3节中被证明是最佳方法。

从表中可以看出，微调对不同掩码策略表现出惊人的鲁棒性。然而正如预期，在将基于特征的方法应用于NER时，仅使用MASK策略会产生问题。有趣的是，仅使用RND策略的表现也远逊于我们的策略。