外变双向编码器表示法：用于语言理解的深度双向外变网络的预训练

雅各布·德夫林 张明伟 李 肯顿 克里斯 蒂娜·图塔诺娃

谷歌人工智能语言

{Jacobdevlin，Mingweichang，Kentonl，Kristout}@google.com

抽象

我们引入了一种称为外变双向编码器表示法的新语言表示模型，它代表来自变压器的双向编码器表示。与最近的语言表示模型（Peters等人）不同。， 2018a; 拉德福德等.，2018），外变双向编码器表示法旨在通过在所有层中对左右上下文进行联合条件来预训练来自未标记文本的深度双向表示。因此，只需一个额外的输出层即可对预先训练的外变双向编码器表示法模型进行微调，从而为各种任务（例如问答和语言推理）创建最先进的模型，而无需对特定于任务的架构进行大量修改。

外变双向编码器表示法在概念上很简单，在经验上很强大。它在 11 项自然语言处理任务上获得了最先进的新结果，包括将 自然语言理解能力评估 分数提高到 80.5%（绝对改进 7.7%），多语式自然语言推理 准确率提高到 86.7%（绝对改进 4.6%）， 标准问题与答案数据集 v1.1 问答测试 F1 到 93.2（绝对改进 1.5 分）和 标准问题与答案数据集 v2.0 测试 F1 到 83.1（绝对改进 5.1 分）。

# 1 引言

语言模型预训练已被证明可有效改善许多自然语言处理任务（戴和乐,2015;彼得斯等.,2018年一;拉德福德等.,2018;霍华德和罗德,2018).这些包括句子级任务，例如自然语言推理（鲍曼等.,2015;威廉姆斯等.,2018） 和释义 （多兰和布罗克特,2005），旨在通过全面分析句子以及标记级任务（例如命名实体识别和问答）来预测句子之间的关系，其中模型需要在令牌级别生成细粒度输出（金相和德宗默尔德,2003;拉杰普尔卡尔等。,2016).

arXiv：1810.04805v2 [。.cs。 CL] 2019年5月24日

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，如嵌入式语言模型（Peters等人。，2018a），使用特定于任务的架构，其中包括预先训练的表示作为附加功能。微调方法，例如生成预训练转换器（开源人工智能 生成式预训练的外变网络模型）（Radford等人）。 ， 2018），引入了最少的任务特定参数，并通过简单地微调所有预训练参数来训练下游任务 。这两种方法在预训练期间共享相同的目标函数，它们使用单向语言模型来学习一般语言表示。

我们认为，当前的技术限制了预训练表示的能力，特别是对于微调方法。主要限制是标准语言模型是单向的，这限制了在预训练期间可以使用的体系结构的选择。例如，在 开源人工智能 生成式预训练的外变网络模型中，作者使用从左到右 的架构，其中每个令牌只能关注转换器自我注意层中的先前令牌（Vaswani等人）。， 2017）。对于句子级任务来说，这种限制是次优 的，并且在 将基于微调的方法应用于令牌级任务（如问答）时可能非常有害，因为从两个方向合并上下文至关重要。

在本文中，我们通过提出外变双向编码器表示法：来自变压器的双向编码器表示来改进基于微调的方法。受完形填空任务（Taylor，1953）的启发，外变双向编码器表示法通过使用“掩蔽语言模型”（有掩码的语言模型）预训练目标来缓解前面提到的单向性约束。屏蔽语言模型从输入中随机屏蔽一些标记，目标是仅根据其上下文预测屏蔽单词的原始词汇 ID。与从左到右的语言模型预训练不同，有掩码的语言模型目标使表示能够融合左右上下文，这使我们能够预训练深度双向转换器。除了屏蔽语言模型之外，我们还使用“下一句预测”任务来联合预训练文本对表示。我们论文的贡献如下：

* 我们证明了双向预训练对语言表示的重要性。与拉德福德等人不同。 （2018），它使用单向语言模型进行预训练，外变双向编码器表示法使用掩蔽语言模型来实现预训练的深度双向表示。这也与彼得斯等人形成鲜明对比。（2018a），它使用独立训练的从左到右和从右到左 LM 的浅层串联。
* 我们表明，预先训练的表示减少了对许多 大量工程 化的任务特定 架构的需求。外变双向编码器表示法是第一个基于微调的表示模型，可在大量句子级和令牌级任务上实现最先进的性能， 优于许多特定于任务的架构。
* 外变双向编码器表示法推进了11项自然语言处理任务的最新技术。代码和预训练模型可在 [https://github.com/ google-research/外变双向编码器表示法](https://github.com/google-research/bert)上找到[。](https://github.com/google-research/bert)

# 2 相关工作

预训练通用语言表示的历史由来已久，我们将在本节中简要回顾使用最 广泛的 方法。

2.1 基于无监督特征的方法

几十年来，学习广泛适用的单词表示一直是研究的一个活跃领域，包括非神经（Brown等人）。， 1992; 安藤和张，2005; 布利策等.，2006）和神经（Mikolov等人。， 2013; 彭宁顿等.，2014）方法。预训练词嵌入是现代自然语言处理系统不可或缺的一部分，与从头开始学习的嵌入相比提供了显着的改进（Tarian等人。，2010）。为了预训练词嵌入向量，使用了从左到右的语言建模目标（Mnih and Hinton，2009），以及在左右上下文中区分正确和不正确单词的目标（Mikolov et al.，2013）。

这些方法已被推广到更粗糙的粒度，例如句子嵌入（Kiros et al.， 2015; Logeswaran和Lee，2018）或段落嵌入（Le和Mikolov，2014）。为了训练句子表示，先前的工作已经使用目标对候选的下一个句子进行排名（Jernite et al.， 2017; Logeswaran和Lee，2018），从左到右生成下一个句子单词，给定前一个句子的表示（Kiros et al.，2015），或去噪自动编码器派生目标（Hill等人。，2016）。

嵌入式语言模型及其前身（Peters等人。 ， 2017， 2018a） 将传统词嵌入研究推广到不同的维度。它们 从从左到右和从右到左的语言模型中提取上下文相关特征。每个标记的上下文表示形式是从左到右和从右到左表示的串联。当将上下文词嵌入与现有的特定于任务的架构集成时，嵌入式语言模型推进了几个主要自然语言处理基准的最新技术（Peters等人。，2018a），包括问答（Rajpurkar等人。，2016），情感分析（Socher等人。，2013），以及命名实体识别（Tjong Kim Sang和De Meulder，2003）。梅拉默德等. （2016）提出通过一项任务来学习上下文表示，以使用长短期记忆模型从左右上下文中预测单个单词。 与嵌入式语言模型类似，它们的模型是基于特征的，而不是深度双向的。 费杜斯等. （2018）表明，完形填空任务可用于提高文本生成模型的稳定性。

2.2 无监督微调方法

与基于特征的方法一样，第一个方向只从未标记的文本中预先训练单词嵌入参数（Collo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的文本中预先训练，并针对受监督的下游任务进行了微调（Dai and Le，2015; 霍华德和罗德，2018年; 拉德福德等.，2018）。这些方法的优点是需要从头开始学习的参数很少。至少部分由于这一优势，开源人工智能 生成式预训练的外变网络模型（Radford等人）。，2018）在自然语言理解能力评估基准的许多句子级任务上取得了以前最先进的结果（Wang等人。，2018a）。从左到右的语言模型-

|  |
| --- |
| 图 1：外变双向编码器表示法 的整体预训练和微调程序。除了输出层之外，在预训练和微调中都使用相同的架构。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，所有参数都会被微调。[CLS] 是在每个输入示例前面添加的特殊符号，而 [SEP] 是一个特殊的分隔符标记（例如分隔 ques- |

答案/答案）。

和自动编码器的目标函数，用于预训练这样的模型（Howard and Ruder， 2018; 拉德福德等.， 2018; 戴和乐， 2015）。

2.3 从监督数据迁移学习

也有研究表明，具有大型数据集的监督任务（例如自然语言推理）可以有效转移（Conneau 等人）。， 2017）和机器翻译（McCann et al.， 2017）。计算机视觉研究还证明了从大型预训练模型进行迁移学习的重要性，其中有效的方法是微调使用ImageNet预先训练的模型（Deng等人）。， 2009; 约辛斯基 等.， 2014）。

# 3 外变双向编码器表示法

我们将在本节中介绍外变双向编码器表示法及其详细实现。我们的框架中有两个步骤：*预训练*和*微调*。在预训练期间，模型在不同的预训练任务中针对未标记的数据进行训练。为了进行微调，首先使用预先训练的参数初始化外变双向编码器表示法模型，并使用 来自下游任务的标记数据对所有参数进行微调。每个下游任务都有单独的微调模型，即使它们使用相同的预训练参数进行初始化。图 1 中的问答 示例将作为本节的运行示例。

外变双向编码器表示法的一个显着特点是其跨不同任务的统一架构。预训练架构和最终下游架构之间的差异很小。

模型架构 外变双向编码器表示法的模型架构是基于Vaswani等人描述的原始实现的多层双向外变编码器。（2017）并在Tensor2tensor库中发布。[[1]](#footnote-1) 由于外变网络的使用已经变得普遍，并且我们的实现与原始实现几乎相同，因此我们将省略模型架构的详尽背景描述，并请读者参考Vaswani等人。 （2017）以及“带注释的外变网络”等优秀指南。[[2]](#footnote-2)

在这项工作中，我们将层数（即外变网络块）表示为*L*，隐藏大小表示为*H*，将注意自身的数量表示为*A*。[[3]](#footnote-3)我们主要报告两种模型大小的结果：外变双向编码器表示法 基础（L=12，H=768，A=12，总参数= 110M）和外变双向编码器表示法 大型（L = 24，H = 1024，A = 16，总参数= 340M）。

出于比较目的，选择外变双向编码器表示法基础具有与开源人工智能 生成式预训练的外变网络模型相同的模型大小 。然而，至关重要的是，外变双向编码器表示法的外变网络使用的双向注意自身，而生成式预训练的外变网络的外变网络使用的约束注意自身，其中每个令牌只能关注其左侧的上下文。[[4]](#footnote-4)

输入/输出表示 为了使外变双向编码器表示法处理各种下游任务，我们的输入表示能够在一个标记序列中明确表示单个句子和一对句子（例如，问题，答案）。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”是指外变双向编码器表示法的输入文字序列，可以是单个句子或两个句子打包在一起。

我们使用字片嵌入（Wu等人。 ，2016）与30，000词汇记号。每个序列的第一个标记始终是特殊分类标记 （[CLS]）。与此记号对应的最终隐藏状态用作分类任务的聚合序列表示形式。句子对被打包成一个序列。我们以两种方式区分句子。首先，我们用一个特殊的记号（[SEP]）将它们分开。其次，我们为每个标记添加一个学习嵌入，指示它属于句子 A 还是句子 B。如图 1 所示，我们将输入的嵌入层表示为 *E*，特殊 [CLS] 记号的最终隐层向量表示为 *C*  ∈ R H，第 i 个输入文字的最终隐层向量表示为 *Ti* ∈ R*H*。

对于给定的文字，其输入表示是通过对相应的文字、片段和位置嵌入层求和来构造的。这种结构的可视化如图 2 所示。

3.1 预训练外变双向编码器表示法

与彼得斯等人不同。 （2018a）和Radford等人。 （2018），我们不使用传统的从左到右或从右到左的语言模型来预训练外变双向编码器表示法。相反，我们使用本节中描述的两个无监督任务来预训练外变双向编码器表示法。此步骤如图 1 的左侧所示。

任务#1：有掩码的语言模型 直观地说，有理由相信深度双向模型比从左到右的模型或从左到右和从右到左的模型的浅层串联更强大。不幸的是，标准条件语言模型只能从左到右*或*从右到左进行训练，因为双向条件反射将允许每个单词间接“看到自己”，并且该模型可以在多层上下文中轻松预测目标单词。

前者通常被称为“外变网络编码器”，而仅左上下文版本被称为“外变网络解码器”，因为它可用于文本生成。

为了训练深度双向表示法网络，我们只需随机屏蔽一定比例的输入文字，然后预测这些屏蔽文字。我们将此过程称为“有掩码的语言模型”（有掩码的语言模型），尽管它在文献中通常被称为*完形填空*任务（Taylor，1953）。在这种情况下，对应于掩码文字的最终隐藏向量被馈送到词汇表上的输出软最大中，就像在标准语言模型中一样。在我们所有的实验中，我们随机屏蔽了每个序列中所有 字片 的 15%。 与去噪自动编码器相反（Vincent等人。 ，2008），我们只预测被屏蔽的单词，而不是重建整个输入。

虽然这允许我们获得双向预训练模型，但缺点是我们在预训练和微调之间造成了不匹配，因为在微调过程中不会出现 [MASK] 记号。为了缓解这种情况，我们并不总是将“屏蔽”字替换为实际的 [MASK] 记号。训练数据生成器随机选择 15% 的文字位置进行预测。如果选择了第 i 个文字，我们将第 i 个文字替换为 （1） 80% 的时侯使用 [MASK] 记号 （2） 10% 的时侯是随机记号 （3） 10% 的时侯未更改的第 *i* 个令牌。然后，*Ti* 将用于预测具有交叉熵损失的原始文字。我们在附录C.2中比较了该程序的变体。

任务#2：下一句的预测（下一句的预测） 许多重要的下游任务，如问答（问答）和自然语言推理（自然语言推理）都是基于理解两个句子之间的关系，这是语言建模无法直接捕获的。为了训练一个理解句子关系的模型，我们预先训练了一个二值化的*下一个句子预测*任务，该任务可以从任何单语语料库轻松生成。具体来说，当为每个预训练示例选择句子 A 和 B 时，50% 的时侯 B 是 A 之后的实际下一个句子（标记为 IsNext），50% 的时侯是语料库中的随机句子（标记为 NotNext）。如图 1 所示，*C* 用于下一句的预测 （下一句的预测）。[[5]](#footnote-5) 尽管它很简单，但我们在第 5.1 节中证明，针对此任务的预训练对 问答 和 自然语言推理 都非常有益。 [[6]](#footnote-6)

|  |
| --- |
| 图 2：外变双向编码器表示法 输入表示。输入的嵌入是文字的嵌入层的总和，片段的 |

嵌入层 和位置的嵌入层。

下一句的预测任务与Jernite等人使用的表示学习目标密切相关。（2017）和Logeswaran和Lee（2018）。但是，在以前的工作中，只有句子的嵌入层被转移到下游任务，其中外变双向编码器表示法传输所有参数以初始化结束任务模型参数。

预训练数据 预训练过程在很大程度上遵循了现有的语言模型预训练文献。对于预训练语料库，我们使用BooksCorpus（800M字）（Zhu等人。 ，2015）和英语维基百科（2，500M字）。对于维基百科，我们只提取文本段落，忽略列表、表格和标题。使用文档级语料库而不是像十亿字基准这样的打乱句子级语料库至关重要（Chelba等人）。，2013）以提取长连续序列。

3.2 微调外变双向编码器表示法

微调很简单，因为转换器中的 自注意力 机制允许外变双向编码器表示法通过交换适当的输入和输出来对许多下游任务进行建模 - 无论它们涉及单个文本还是文本对。对于涉及文本对的应用程序，一种常见的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）; 徐 等. （2017）.相反，外变双向编码器表示法使用自我注意机制来统一这两个阶段，因为对具有自我注意的串联文本对进行编码有效地包括 两个句子之间的双向交叉注意力。

对于每个任务，我们只需将 特定于任务的 输入和输出插入外变双向编码器表示法，并端到端地微调所有参数。在输入时，预训练中的句子 A 和句子 B 类似于 （1） 释义中的句子对，（2） 蕴涵中的假设-前提对，（3） 问答中的问题段落对，以及

（4）文本分类或序列标记中的退化文本∅对。在输出端，文字表示法被馈送到输出层，用于 文字级 任务，例如序列文字或问答，而 [CLS] 表示被馈送到输出层进行分类，例如蕴涵或情感分析。

与预训练相比，微调相对便宜。 论文中的所有结果都可以在单个（云端张量处理单元）Cloud TPU上最多复制1小时，或者在GPU（图形处理单元）上复制几个小时，从完全相同的预训练模型开始。7  我们将在第 4 节的相应小节中描述特定于任务的详细信息。更多详情请参阅附录A.5。

# 4 实验

在本节中，我们展示了 11 个 自然语言处理 任务的 外变双向编码器表示法 微调结果。

4.1 自然语言理解能力评估

自然语言理解能力评估（自然语言理解能力评估）基准（Wang等人）。， 2018a）是各种自然语言理解任务的集合。自然语言理解能力评估数据集的详细说明包含在附录 B.1中。

为了微调 自然语言理解能力评估，我们表示第 3 节中所述的输入序列（对于单个句子或句子对），并使用对应于 第一个输入标记 （[CLS]） 的最终隐藏向量 C ∈ R *H* 作为聚合表示。微调期间引入的唯一新参数是分类图层权重 *W* ∈ R K×*H*，其中 *K* 是标注的数量。我们用 *C* 和 *W* 计算标准分类损失，即 log（软最大（*CWT* ））。

7例如，外变双向编码器表示法 标准问题与答案数据集 模型可以在单个云端张量处理单元上训练大约 30 分钟，以实现 91.0% 的开发 F1 分数。

8见 https://自然语言理解能力评估benchmark.com/faq（10）中的。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 系统 | 锰里-（米/毫米） | QQP | 秦丽 | SST-2 | 可乐 | STS-B | 微软研究院释义语料库 | 即食 | 平均 | |  | 392k | 363k | 108k | 67k | 8.5千米 | 5.7千米 | 3.5千米 | 2.5千米 | - | | 预开放AI 下 | 80.6/80.1 | 66.1 | 82.3 | 93.2 | 35.0 | 81.0 | 86.0 | 61.7 | 74.0 | | Bi长短期记忆模型+嵌入式语言模型+Attn | 76.4/76.1 | 64.8 | 79.8 | 90.4 | 36.0 | 73.3 | 84.9 | 56.8 | 71.0 | | 开源人工智能 生成式预训练的外变网络模型 | 82.1/81.4 | 70.3 | 87.4 | 91.3 | 45.4 | 80.0 | 82.3 | 56.0 | 75.1 | | 伯特基地 | 84.6/83.4 | 71.2 | 90.5 | 93.5 | 52.1 | 85.8 | 88.9 | 66.4 | 79.6 | | 伯特·大 | 86.7/85.9 | 72.1 | 92.7 | 94.9 | 60.5 | 86.5 | 89.3 | 70.1 | 82.1 |   表1：由评估服务器（[https://自然语言理解能力评估benchmark.com/leaderboard](https://gluebenchmark.com/leaderboard)）评分的自然语言理解能力评估测试结果。每个任务下方的数字表示训练示例的数量。“平均”列与官方 自然语言理解能力评估 分数略有不同，因为我们排除了有问题的 W自然语言推理 集。8 外变双向编码器表示法和开源人工智能 生成式预训练的外变网络模型是单模型，单任务。报告QQP和微软研究院释义语料库的F1分数，报告STS-B的Spearman相关性，并报告其他任务的准确性分数。我们排除使用 外变双向编码器表示法 作为其组件之一的条目。 |

我们使用 32 的批大小，并对所有 自然语言理解能力评估 任务的数据微调 3 个时期。对于每个任务，我们在 开发阶段集合上选择了最佳的微调学习率（在 5e-5、4e-5、3e-5 和 2e-5 中）。

此外，对于外变双向编码器表示法大型，我们发现微调有时在小数据集上不稳定，因此我们运行了几次随机重启，并在开发集中选择了最佳模型。通过随机重启，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器层初始化。[[7]](#footnote-7)

结果如表1所示。

外变双向编码器表示法基础和外变双向编码器表示法大型在所有任务上的表现都大大优于所有系统，与现有技术相比，平均精度分别提高了4.5%和7.0%。请注意，

外变双向编码器表示法基础 和 开源人工智能 生成式预训练的外变网络模型在模型架构方面几乎相同，除了注意力掩码。对于最大和最广泛报道的自然语言理解能力评估任务，有掩码的自然语言推理，外变双向编码器表示法获得了4.6%的绝对精度改进。关于官方

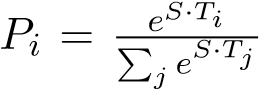
自然语言理解能力评估排行榜第10名，外变双向编码器表示法大型获得80.5分，而开源人工智能 生成式预训练的外变网络模型在撰写本文之日获得72.8分。

我们发现外变双向编码器表示法大型在所有任务中都明显优于外变双向编码器表示法基础，尤其是那些训练数据很少的任务。模型大小的影响在第 5.2 节中进行了更彻底的探讨。

4.2 标准问题与答案数据集 v1.1

斯坦福问答数据集（标准问题与答案数据集 v1.1）是100k众包问答对的集合（Rajpurkar 等人。,

2016）。 给定一个问题和维基百科中包含答案的段落，任务是预测段落中的答案文本跨度。

如图 1 所示，在问答任务中，我们将输入问题和段落表示为单个打包序列，问题使用 A 嵌入，段落使用 B 嵌入。在微调过程中，我们只引入起始向量 *S*  ∈ R H 和结束向量 *E* ∈ R*H* 。文字i 是答案范围开始的概率计算为 *Ti* 和 *S*  之间的点积，后跟 段落中所有文字的 软最大：。

类似的公式用于答案范围的末尾。候选人从位置 *i* 到位置 *j*  的分数定义为 *S*·*Ti* + *E*·*T* j，以及将 *j* ≥ *i* 用作预测的最大评分范围 。训练目标是正确开始和结束位置的对数似然估计的总和。我们对 3 个 epoch 进行微调，学习率为 5e-5，批量大小为 32。

表 2 显示了排行榜条目以及已发布系统排名靠前的结果（Seo 等人。， 2017; 克拉克和加德纳，2018 年; 彼得斯等.， 2018a; 胡等.，2018）。标准问题与答案数据集 排行榜的排名靠前的结果没有最新的公共系统描述，[[8]](#footnote-8)并且允许在训练其系统时使用任何公共数据。因此，我们通过首先对Trivia问答进行微调（Joshi等人）。，2017）用于对标准问题与答案数据集进行微调。

我们表现最好的系统在整体系统中比顶级排行榜系统高出 +1.5 F1（渔夫1分布），在单个系统中比顶级排行榜系统高出 +1.3 F1。事实上，我们的单一外变双向编码器表示法模型优于顶级集成系统

就 F1得分而言。没有微调 琐事问答 的数据

系统

开发

测试

在

F1

在

F1

顶级排行榜系统（2018年12月10日）

人类- - 82.3 91.2

#1 套装 - nlnet- - 86.0 91.7

#2 在一起 - 问答Net - - 84.5 90.5

|  |  |
| --- | --- |
| 发表  双达夫+嵌入式语言模型（单）- 85.6-85.8 | |
| R.M.读者（合奏） | 81.2 87.9 82.3 88.5 |
| 我们  伯特基地 （单曲） | 80.8 88.5- - |
| 伯特·大 （单曲） | 84.1 90.9- - |
| 伯特·大 （合奏） | 85.8 91.8- - |
| 外变双向编码器表示法大型 （Sgl.+ 琐事问答） | 84.2 91.1 85.1 91.8 |
| 外变双向编码器表示法大型 （Ens.+ 琐事问答） | 86.2 92.2 87.4 93.2 |

表 2： 标准问题与答案数据集 1.1 结果。外变双向编码器表示法集成是7x系统，使用不同的训练前检查点和微调种子。

|  |  |  |
| --- | --- | --- |
| 系统 | 开发 测试  在 F1中 在F1 中 | |
| 顶级排行榜系统（2018年12月10日）  人类 86.3 89.0 86.9 89.5  #1 单 - MIR-MRC （F-NET） - -74.8 78.0 | | |
| #2 单曲 - nlnet - | | - 74.2 77.1 |
| 出版的 梦想 （合奏） - | | - 71.4 74.9 |
| 单品+（单） - | | 71.4 74.4 |
| 我们  伯特大 （单） 78.7 8 | | 1.9 80.0 83.1 |

表 3： 标准问题与答案数据集 2.0 结果。我们排除使用 外变双向编码器表示法 作为其组件之一的条目。

，我们只损失了0.1-0.4 F1，仍然远远超过所有现有系统。[[9]](#footnote-9)

4.3 标准问题与答案数据集 v2.0

标准问题与答案数据集 2.0 任务扩展了 标准问题与答案数据集 1.1 问题定义，允许提供的段落中不存在简短答案的可能性，从而使问题更加现实。

我们使用一种简单的方法来扩展此任务的 标准问题与答案数据集 v1.1 外变双向编码器表示法 模型。我们将无答案的问题视为具有以 [CLS] 标记开头和结尾的答案区间。开始和结束答案区间位置的概率空间将扩展为包括 [CLS] 标记的位置。为了预测，我们比较无答案区间的分数： *s*null = *S*·*C* + *E*·*C （小写s是分数，E这里是结束向量）*到最佳非空区间的分数

|  |  |  |
| --- | --- | --- |
| 系统 | 开发 | 测试 |
|  |  |  |
| 人类（专家）† | - | 85.0 |
| 人类（5 个注释）† | - | 88.0 |

表 4：SWAG 开发和测试精度。 †正如SWAG论文中所报告的那样，人类的表现是用100个样本来衡量的。

*s*ˆ i*，*j = max*j*≥*iS*·*Ti* + *E*·*TJ*.当 s ˆ*i，j > s*null + τ 时，我们预测一个非空答案，其中在开发集上选择阈值 *τ* 以最大化 F1。我们没有在这个模型中使用琐事问答数据。我们对 2 个 轮回 进行了微调，学习率为 5e-5，批量大小为 48。

结果与之前的排行榜条目和发表最多的作品进行了比较（Sun等人。， 2018; 王等.表3所示，不包括使用外变双向编码器表示法作为其组件之一的系统。我们观察到 +5.1 F1 比之前的最佳系统有所改进。

4.4 对抗生成的情境

对抗生成的情境（SWAG）数据集包含113k个句子对完成示例，这些示例评估了有根据的常识推理（Zellers等人。，2018）。给定一个句子，任务是在四个选项中选择最合理的延续。

在对 对抗生成的情境 数据集进行微调时，我们构建了四个输入序列，每个序列都包含给定句子的串联（句子 A）和可能的延续（句子 B）。引入的唯一特定于任务的参数是一个向量，其带有 [CLS] 标记表示 *C*  的点积表示每个选择的分数，该分数使用 软最大 层进行归一化。

我们对模型进行了 3 个 轮回 的微调，学习率为 2e-5，批量大小为 16。结果如表4所示。外变双向编码器表示法大型的性能比作者的基线ESIM + 嵌入式语言模型系统高出+27.1%，开源人工智能 生成式预训练的外变网络模型高出8.3%。 5 消融研究

在本节中，我们对外变双向编码器表示法的许多方面进行了消融实验，以便更好地了解它们的相对重要性。附加

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 开发集 | | | | |
| 任务 | 姆利-米  （累计） | 秦丽  （累计） | 微软研究院释义语料库  （累计） | SST-2 S  （累计） | 四  （F1） |
| 伯特基地 | 84.4 | 88.4 | 86.7 | 92.7 | 88.5 |
| 没有 下一句的预测 | 83.9 | 84.9 | 86.5 | 92.6 | 87.9 |
| LTR & No 下一句的预测 | 82.1 | 84.3 | 77.5 | 92.1 | 77.8 |
| + 比利斯特姆 | 82.1 | 84.1 | 75.7 | 91.6 | 84.9 |

表 5：使用 外变双向编码器表示法基础  架构的预训练任务消融。“No 下一句的预测”在没有下一个句子预测任务的情况下进行训练。“LTR & No 下一句的预测”被训练为没有下一句话预测的从左到右的LM，就像开源人工智能 生成式预训练的外变网络模型一样。“+ Bi长短期记忆模型”在微调期间在“LTR + No 下一句的预测”模型之上添加了一个随机初始化的Bi长短期记忆模型。

消融研究可以在附录C中找到。

5.1 预训练任务的影响

我们通过使用与外变双向编码器表示法 基础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，从而证明了外变双向编码器表示法深度双向性的重要性：

No 下一句的预测：使用“有掩码的语言模型（有掩码的语言模型）训练但没有“下一句预测”（下一句的预测）任务的双向模型。

LTR & No 下一句的预测：一种仅左上下文模型，使用标准的从左到右（LTR）LM而不是有掩码的语言模型进行训练。仅左约束也应用于微调，因为删除它会导致预训练/微调不匹配，从而降低下游性能。此外，此模型是在没有 下一句的预测 任务的情况下预先训练的。这与 开源人工智能 生成式预训练的外变网络模型 直接相当，但使用我们更大的训练数据集、输入表示和微调方案。

我们首先考察下一句的预测任务带来的影响。在表 5 中，我们表明删除 下一句的预测 会显著损害 Q自然语言推理、M自然语言推理 和 标准问题与答案数据集 1.1 的性能。接下来，我们通过比较“No 下一句的预测”和“LTR & No 下一句的预测”来评估训练双向表示的影响。LTR模型在所有任务上的表现都比有掩码的语言模型模型差，微软研究院释义语料库和标准问题与答案数据集上的下降幅度很大。

对于 标准问题与答案数据集，直观地很清楚，LTR 模型在令牌预测方面表现不佳，因为令牌级别的隐藏状态没有右侧上下文。为了真诚地尝试加强LTR系统，我们在顶部添加了一个随机初始化的Bi长短期记忆模型。这确实显着改善了标准问题与答案数据集上的结果，但结果仍然比预训练的双向模型差得多。Bi长短期记忆模型损害了自然语言理解能力评估任务的性能。

我们认识到，也可以像嵌入式语言模型那样训练单独的LTR和RTL模型，并将每个令牌表示为两个模型的串联。然而：（a） 这比单一的双向模式贵一倍;（b） 这比单一的双向模式贵一倍。（b）这对于像问答这样的任务来说是不直观的，因为RTL模型将无法根据问题来决定答案;（c） 严格来说，它不如深度双向模型强大，因为它可以在每一层同时使用左上下文和右上下文。

5.2 模型尺寸的影响

在本节中，我们将探讨模型大小对微调任务准确性的影响。我们训练 了许多 具有不同数量的层、隐藏单元和注意力头的 外变双向编码器表示法 模型，同时使用与前面描述的相同的超参数和训练过程。

所选自然语言理解能力评估任务的结果如表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阶段数据集 准确性。我们可以看到，更大的模型会导致所有四个数据集的严格准确性提高，即使对于只有 3，600 个标记训练样本的 微软研究院释义语料库，并且 与训练前任务有很大不同。也许令人惊讶的是，我们能够 在相对于现有文献已经相当大的模型之上实现如此显着的改进。例如，Vaswani等人探索的最大外变网络模型。（2017）是（L=6，H=1024，A=16）（神经网络的层数=6，每层神经元的数量=1024，激活函数=16），编码器参数为100M，我们在文献中发现的最大变压器是（L=64，H=512，A=2），参数为235M（Al-Rfou et al.，2018）。相比之下，外变双向编码器表示法 基础包含110M参数，大型外变双向编码器表示法包含340M参数。

人们早就知道，增加模型大小将导致大规模任务（如机器翻译和语言建模）的持续改进，表 6 所示的保留训练数据的 LM 困惑性证明了这一点。然而，我们相信这是第一个令人信服地证明，只要模型经过充分预训练，扩展到极端模型大小也会导致非常小尺度的任务的大幅改进的工作。彼得斯等. （2018b）对将预训练的双向语言模型大小从两层增加到四层的下游任务影响提出了不同的结果，Melamud等人。 （2016）顺便提到，将隐藏维度大小从200增加到600有所帮助，但进一步增加到1，000并没有带来进一步的改进。 这两个先前的工作都使用了基于特征的方法——我们假设，当模型直接在下游任务上进行微调并且只使用非常少量随机初始化的附加参数时，任务特定模型可以从更大、更具表现力的预训练表示中受益，即使下游任务数据非常小时。

5.3 基于特征的外变双向编码器表示法方法

到目前为止，所有外变双向编码器表示法结果都使用了微调方法，其中将简单的分类层添加到预训练的模型中，并且所有参数都在下游任务上联合微调。然而，从预训练模型中提取固定特征的基于特征的方法具有一定的优势。首先，并非所有任务都可以用 外变神经网络模型 编码器架构轻松表示，因此需要添加特定于任务的模型架构。其次，预先计算一次昂贵的训练数据表示，然后在这种表示之上使用更便宜的模型运行许多实验，具有主要的计算优势。在本节中，我们将外变双向编码器表示法应用于公共的负对数似然估计-2003命名实体识别（命名实体识别）任务来比较这两种方法（Tjong Kim Sang和De Meulder，2003）。在外变双向编码器表示法的输入中，我们使用保留大小写的字片模型，并包括数据提供的最大文档上下文。按照标准做法，我们将其表述为标记任务，但不使用 条件随机场

超参数 开发集精度

#L#H #A LM （ppl） M自然语言推理-m 微软研究院释义语料库 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：外变双向编码器表示法 模型大小的消融。#L = 层数;#H = 隐藏大小;#A = 注意力头数。“LM （ppl）”是保留训练数据的掩蔽 LM 困惑度。

系统 开发 F1 测试 F1

|  |  |  |
| --- | --- | --- |
| 嵌入式语言模型 （Peters et al.， 2018a） | 95.7 | 92.2 |
| CVT（克拉克等人。， 2018） | - | 92.6 |
| CSE（Akbik et al.， 2018） | - | 93.1 |
| 微调方法  伯特·大 | 96.6 | 92.8 |
| 伯特基地 | 96.4 | 92.4 |
| 基于特征的方法（外变双向编码器表示法基础）嵌入 | 91.0 | - |
| 倒数第二隐藏 | 95.6 | - |
| 上次隐藏 | 94.9 | - |
| 加权总和最后四个隐藏 | 95.9 | - |
| 康卡特 最后四个隐藏 | 96.1 | - |
| 加权和所有 12 层 | 95.5 | - |

表 7：公共的负对数似然估计-2003 命名实体识别结果。超参数是使用开发 集选择的。报告的开发和测试分数是使用这些超参数在 5 次随机重启中取平均值。

图层。我们使用第一个子标记的表示形式作为 命名实体识别 标签集上的文字级分类器的输入。

为了消融微调方法，我们通过从一个或多个层中提取激活来应用基于特征的方法，*而无需*微调外变双向编码器表示法的任何参数。这些上下文嵌入用作分类层之前随机初始化的双层 768 维 双向长短期记忆模型 的输入。

结果如表7所示。外变双向编码器表示法大型以最先进的方法具有竞争力。性能最佳的方法是连接来自预训练变形金刚的前四个隐藏层的令牌表示，这仅比微调整个模型落后 0.3 F1。这表明外变双向编码器表示法对于微调和基于特征的方法都是有效的。

# 6 结论

最近由于语言模型迁移学习而进行的实证改进表明，丰富的、无监督的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 结果使即使是低资源任务也能 从深度单向架构中受益。我们的主要贡献是进一步将这些发现推广到深度 *双向* 架构中，允许相同的预训练模型成功处理广泛的自然语言处理任务。

# 引用

艾伦·阿克比克、邓肯·布莱斯和罗兰·沃尔格拉夫。2018. 用于序列标记的上下文字符串嵌入.第 *27 届计算语言学国际会议论文集*，第 1638-1649 页。

Rami Al-Rfou、Dokook Choe、Noah Constant、Mandy Guo和Llion Jones。2018. 具有更深层次自我关注的字符级语言建模. arXiv  *预印本arXiv：1808.04444*。

久保田理惠和张彤。2005. 从多个任务和未标记数据中学习预测结构的框架.*机器学习研究杂志*，6（11月）：1817–1853。

路易莎·本蒂沃利、贝尔纳多·马格尼尼、伊多·达甘、华庄当和达尼洛·詹皮科洛。2009. 第五届PASCAL承认文本蕴涵挑战。在*TAC*中。美国国家情报和安全总局。

约翰·布利策、瑞安·麦克唐纳和费尔南多·佩雷拉。2006. 结构对应学习的领域适应.在 *2006年自然语言处理经验方法会议论文集*，第120-128页。计算语言学协会。

Samuel R. Bowman、Gabor Angeli、Christopher Potts 和 Christopher D. Manning。2015. 用于学习自然语言推理的大型注释语料库.在*EM自然语言处理*中。计算语言学协会。

彼得·布朗、彼得·德苏扎、罗伯特·默瑟、文森特·德拉·彼得拉和珍妮弗·赖。1992. 基于类的自然语言 n-gram 模型.*计算语言学*， 18（4）：467–479.

Daniel Cer、Mona Diab、Eneko Agirre、Inigo LopezGazpio 和 Lucia Specia。2017. [Semeval-2017 任务 1：语义文本相似性](https://doi.org/10.18653/v1/S17-2001)多语言和[跨语言重点评估。](https://doi.org/10.18653/v1/S17-2001) 第*11届语义评估国际研讨会（SemEval-2017）论文集*，第1-14页，加拿大温哥华。计算语言学协会。

西普里安·切尔巴、托马斯·米科洛夫、迈克·舒斯特、齐戈、托斯滕·布兰茨、菲利普·科恩和托尼·罗宾逊。2013. 衡量统计语言建模进展的十亿字基准.arXiv  *预印本arXiv：1312.3005*。

陈志、张海、张晓、赵林.2018.

[Quora问题对。](https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs)

克里斯托弗·克拉克和马特·加德纳。2018. 简单有效的多段落阅读理解.在 *ACL* 中。

Kevin Clark、Minh-Thang Luong、Christopher D Manning和Quoc Le。2018. 具有交叉视图训练的半监督序列建模.*在2018年自然语言处理经验方法会议论文集*，第1914-1925页。

罗南· 科洛伯特 和杰森·韦斯顿。2008. 自然语言处理的统一架构：具有多任务学习的深度神经网络。在第 *25届机器学习国际会议论文集*，第160-167页。阿克姆。

亚历克西斯·康诺、杜维·基拉、霍尔格·施温克、洛·巴罗和安托万·博德斯。2017. 来自[自然语言推理数据的通用句子表示的监督](https://www.aclweb.org/anthology/D17-1070)学习[.](https://www.aclweb.org/anthology/D17-1070) 在*2017年自然语言处理经验方法会议论文集*，第670-680页，丹麦哥本哈根。计算协会

语言学。

戴和郭五乐。2015. 半监督序列学习.在 *神经信息处理系统的进展*中，第3079-3087页。

邓俊杰， 董伟， 索彻， L.-J.李、李克和L. 飞飞.2009. 图像网：一个大规模的分层图像数据库。在*CVPR09*中。

威廉·B·多兰和克里斯·布罗克特。2005. 自动构建量刑释义语料库.*第三届释义国际研讨会论文集（IWP2005）。*

威廉·费杜斯、伊恩·古德费罗和安德鲁·戴。2018. Maskgan：通过填写 .arXiv  *预印本arXiv：1801.07736*。

丹·亨德利克斯和凯文·金佩尔。2016. [桥接非线性和随机正则子与高斯误差线性单元.](http://arxiv.org/abs/1606.08415) *CoRR*， abs/1606.08415.

Felix Hill，Kyunghyun Cho和Anna Korhonen。2016. 从未标记数据中学习句子的分布式表示. 在*北美分会2016年会议记录*中

*计算语言学协会：人类语言技术*。计算语言学协会。

杰里米·霍华德和塞巴斯蒂安·鲁德。2018.  [文本分类的通用语言模型微调.](http://arxiv.org/abs/1801.06146) 在 *ACL* 中。计算语言学协会。

胡明昊， 彭玉兴， 黄震， 邱希鹏， 魏辅苣， 周明.2018. 用于机器阅读理解的增强助记符阅读器.在*IJCAI*。

Yacine Jernite，Samuel R. Bowman和David Sontag。2017. [基于话语的快速联合国监督句子表征学习](http://arxiv.org/abs/1705.00557)目标[.](http://arxiv.org/abs/1705.00557) *CoRR*， abs/1705.00557.

Mandar Joshi，Eunsol Choi，Daniel S Weld和Luke Zettlemoyer。2017. Trivia问答：用于阅读理解的大规模远程监督挑战数据集。在 *ACL* 中。

瑞安·基罗斯、朱玉坤、鲁斯兰·萨拉胡迪诺夫、理查德·泽梅尔、拉奎尔·乌尔塔松、安东尼奥·托拉尔巴和桑贾·菲德勒。2015. 跳过思想向量.在*神经信息处理系统的进展*中，第3294-3302页。

郭乐和托马斯·米科洛夫。2014. 句子和文档的分布式表示.在*机器学习国际会议*上，第1188-1196页。

Hector J Levesque，Ernest Davis和Leora Morgenstern。2011. 威诺格勒模式挑战.在*Aaai春季研讨会：常识推理的逻辑形式化*，第46卷，第47页。

Lajanugen Logeswaran和Honglak Lee。2018. [学习句子重述的有效框架.](https://openreview.net/forum?id=rJvJXZb0W) 在*学习表征国际会议*上。

布莱恩·麦肯、詹姆斯·布拉德伯里、熊 彩明 和理查德· 索彻。2017. 在翻译中学习：语境化的词向量。在 *NIPS*。

Oren Melamud，Jacob Goldberger和Ido Dagan。2016. context2vec：使用双向 长短期记忆模型 学习通用上下文嵌入。在 *CoNLL* 中。

托马斯·米科洛夫、伊利亚·苏茨克弗、陈凯、格雷格·科拉多和杰夫·迪恩。2013. 单词和短语的分布式表示及其组合性.神经*信息处理系统进展26*，第3111-3119页。柯伦协会，

公司。

Andriy Mnih和Geoffrey E Hinton。2009. [一种可扩展](http://papers.nips.cc/paper/3583-a-scalable-hierarchical-distributed-language-model.pdf)的[分层分布式语言模型.](http://papers.nips.cc/paper/3583-a-scalable-hierarchical-distributed-language-model.pdf)在D.科勒，D.舒尔曼斯，Y. 本吉奥和L. Bottou，编辑，*神经信息处理系统进展21*，第1081-1088页。柯伦联合公司

Ankur P Parikh，Oscar Tackstr ̈ om，Dipanjan Das和̈ Jakob Uszkoreit。2016. 自然语言推理的可分解注意力模型.在*EM自然语言处理*中。

Jeffrey Pennington，Richard Socher和Christopher D. Manning。2014. 手套：[用于单词表示的全局向量。](http://www.aclweb.org/anthology/D14-1162) 在*自然语言处理的经验方法（EM自然语言处理）*中，第1532-1543页。

马修·彼得斯、瓦利德·阿马尔、钱德拉·博伽瓦图拉和罗素·鲍尔。2017. 具有双向语言模型的半监督序列标记.在 *ACL* 中。

马修·彼得斯、马克·诺依曼、莫希特·艾耶尔、马特·

加德纳、克里斯托弗·克拉克、肯顿·李和卢克·泽特尔莫耶。 2018a. 深度语境化单词表示。在*全国艾滋病协会*。

马修·彼得斯、马克·诺依曼、卢克·泽特尔莫耶和易文涛。 2018b. 剖析上下文词嵌入：架构和表示。在*2018年自然语言处理经验方法会议论文集*，第1499-1509页。

亚历克·拉德福德、卡西克·纳拉西姆汉、蒂姆·萨利曼斯和伊利亚·苏茨克弗。2018. 通过无监督学习提高语言理解.技术报告，OpenAI。

Pranav Rajpurkar，Jian Zhang，Konstantin Lopyrev和Percy Liang。2016. 小队：100，000+ 个文本机器理解问题。在*2016年自然语言处理经验方法会议论文集*，第2383-2392页。

Minjoon Seo，Aniruddha Kembhavi，Ali Farhadi和Hannaneh Hajishirzi。2017. 机器理解的双向注意力流.在*ICLR*中。

理查德·索彻、亚历克斯·佩雷利金、让·吴、杰森

Chuang，Christopher D Manning，Andrew Ng和Christopher Potts。2013. 情感树库语义组合的递归深度模型.在 *2013年自然语言处理经验方法会议论文集*，第1631-1642页。

孙傅， 李林阳， 邱希鹏， 刘洋.2018. U-net：机器阅读理解与无法回答的问题。 arXiv  *预印本arXiv：1810.06638*。

威尔逊·泰勒。1953. 完形填空程序：测量可读性的新工具。 *新闻通报*， 30（4）：415–433.

Erik F Tjong Kim Sang和 Fien De Meulder。

2003. conll-2003 共享任务简介：与语言无关的命名实体识别。在 *CoNLL* 中。

约瑟夫·图里安、列夫·拉蒂诺夫和约书亚·本吉奥。2010. 单词表示：一种简单而通用的半监督学习方法.在*计算语言学协会第48届年会论文集*，ACL '10，第384-394页。

阿希什·瓦斯瓦尼、诺姆·沙泽尔、尼基·帕尔马、雅各布

乌什科雷特、莱昂·琼斯、艾丹·戈麦斯、卢卡斯·凯撒和伊利亚·波洛苏欣。2017. 注意力就是你所需要的一切。在*神经信息处理系统的进展*中，第6000-6010页。

帕斯卡·文森特、雨果·拉罗切尔、约书亚·本吉奥和皮埃尔-安托万·曼扎戈尔。2008. 使用去噪自动编码器提取和组合鲁棒特征.在第 *25 届机器学习国际会议论文集*，第 1096-1103 页。阿克姆。

亚历克斯·王、阿曼普雷特·辛格、朱利安·迈克尔、菲利克斯·希尔、奥马尔·利维和塞缪尔·鲍曼。 2018a. 自然语言理解能力评估：用于自然语言理解的多任务基准和分析平台。在 *2018 年 EM自然语言处理 研讨会 Blackbox自然语言处理：分析和解释 自然语言处理 的神经网络的*论文集中，第 353-355 页。

王伟、明燕、陈武.2018年b.用于阅读理解和问答的多粒度分层注意力融合网络。计算 *语言学协会第56届年会论文集（第1卷：长论文）。*计算语言学协会。

亚历克斯·沃施塔特、阿曼普雷特·辛格和塞缪尔·鲍曼。2018. 神经网络可接受性判断. arXiv  *预印本arXiv：1805.12471*。

阿迪娜·威廉姆斯、尼基塔· 南吉亚和塞缪尔·鲍曼。2018. 通过推理理解句子的广泛覆盖挑战语料库.在 *全国艾滋病协会*。

吴永辉， 迈克·舒斯特， 陈志峰， 郭五

穆罕默德·诺鲁齐，沃尔夫冈 ·马切里，

马克西姆·克里昆， 曹元， 高秦， 克劳斯·马切里， 等. 2016.谷歌的神经机器翻译系统：弥合人工翻译和机器翻译之间的差距。 arXiv  *预印本arXiv：1609.08144*。

Jason Yosinski，Jeff Clune，Yoshua Bengio和Hod Lipson。2014. 深度神经网络中的特征可转移性如何？在*神经信息处理系统的进展*中，第3320-3328页。

亚当斯·魏宇、大卫·多哈、明汤隆、瑞

赵、陈凯、穆罕默德·诺鲁兹和郭五乐。2018. 问答Net：将局部卷积与阅读理解的全局自我关注相结合。在*ICLR*中。

罗文·泽勒斯、约纳坦·比斯克、罗伊·施瓦茨和崔 叶金 。2018. Swag：用于扎根常识推理的大规模对抗数据集。在 *2018年自然语言处理经验方法会议（EM自然语言处理）的论文集*。

朱玉坤、瑞安·基罗斯、里奇·泽梅尔、鲁斯兰·萨拉胡迪诺夫、拉奎尔·乌尔塔松、安东尼奥·托拉尔巴和桑贾·菲德勒。2015. 对齐书籍和电影：通过看电影和读书来实现类似故事的视觉解释.在*IEEE计算机视觉国际会议论文集*，第19-27页。

附录“外变双向编码器表示法：预培训

用于语言理解的深度双向变压器”

我们将附录分为三个部分：

* 附录A中提供了外变双向编码器表示法的其他实施细节;
* 我们实验的其他细节见附录 B;和
* 附录C中提供了其他消融研究。

我们提出了外变双向编码器表示法的其他消融研究，包括：

* + 训练步骤数的影响;和
  + 不同掩蔽程序的消融。

# 外变双向编码器表示法的其他详细信息

A.1 预训练任务图解

我们在下面提供了预训练任务的示例。

有掩码的语言模型和掩码程序 假设未标记的句子是我的狗是毛茸茸的，并且在随机掩码过程中我们选择了第 4 个文字（对应于毛茸茸），我们的屏蔽过程可以进一步说明

* 80%的时侯：用[MASK]标记替换文字，例如，我的狗是 毛茸茸的→我的狗是 [MASK]
* 10%的时侯：用随机文字替换文字，例如，我的狗多毛→ 我的狗是苹果
* 10%的时侯： 保持 文字不-

改变，例如， 我的狗多毛 → 我的狗多毛。这样做的目的是使表示偏向于实际观察到的单词。

此过程的优点是，外变模型编码器不知道将要求它预测哪些单词或哪些单词已被随机单词替换，因此它被迫保留每个输入标记的分布上下文表示 形式。此外，由于随机替换仅发生在所有文字的 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，有掩码的语言模型仅对每批15%的文字进行预测，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。外变双向编码器表示法使用双向变压器。开源人工智能 生成式预训练的外变网络模型 使用从左到右的外变网络模型。 嵌入式语言模型 使用独立训练的从左到右和从右到左的 长短期记忆模型 的串联来生成下游任务的特征。在这三者中，只有外变双向编码器表示法表示在所有层中同时以左上下文和右上下文为条件。除了架构差异之外，外变双向编码器表示法和开源人工智能 生成式预训练的外变网络模型是微调方法，而嵌入式语言模型是基于功能的方法。 |

来收敛。在C.1节中，我们证明了有掩码的语言模型确实比从左到右的模型（预测每个文字）略慢，但有掩码的语言模型模型的经验改进远远超过了增加的训练成本。

下一句预测 下一句预测任务可以用以下示例来说明。

输入 = [CLS] 该男子去了 [掩码] 商店 [SEP]

他买了一加仑 [掩码] 牛奶 [分割] 标签 = 是下一个

输入 = [CLS] 男人 [掩码] 到商店 [SEP]

企鹅 [掩码] 是飞行 ##less 鸟 [[分割] 标签 = 不是下一个

A.2 预训练程序

为了生成每个训练输入序列，我们从语料库中采样两个文本跨度，我们将其称为“句子”，即使它们通常比单个句子长得多（但也可以更短）。第一个句子接收 A 嵌入，第二个句子接收 B 嵌入。50%的时侯B是A之后的实际下一个句子，50%的时侯它是一个随机句子，这是为“下一个句子预测”任务完成的。对它们进行采样，使组合长度≤ 512 个文字。语言模型的 掩码在 字片 标记化后应用，统一屏蔽率为 15%，并且没有特别考虑部分词条。

我们使用 256 个序列（256 个序列 \* 512 个文字 = 128，000 个文字/批次）的批量大小进行 1，000，000 个步骤的训练，这大约是 33 亿字语料库的 40 个时期。我们使用学习率为 1e-4 *的 自适应估计矩 β*1 = 0*。*9， *β*2 = 0*.*999、L2重量衰减0*.*01、学习率预热超过前10000步，学习率呈线性衰减。我们在所有层上使用 0.1 的辍学概率。我们使用高斯误差线性单元激活（Hendrycks and Gimpel， 2016）而不是标准纠正的线性单元，遵循 开源人工智能 生成式预训练的外变网络模型。训练损失是平均有掩码的语言模型可能性和平均下一句预测可能性的总和。

外变双向编码器表示法基础的训练是在4个云张量处理单元的Pod配置上进行的（总共16个TPU芯片）。[[10]](#footnote-10) 外变双向编码器表示法 大型的训练是在16个云TPU（总共64个TPU芯片）上进行的。每次预训练需要 4 天才能完成。

较长的序列不成比例地昂贵，因为注意力与序列长度是二次的。为了加快实验中的预 处理 速度，我们对90%的步骤的序列长度为128的模型进行了预训练。然后，我们训练 512 序列的其余 10% 的步骤来学习位置嵌入。

A.3 微调程序

对于微调，大多数模型超参数与预训练中的参数相同， 但 批量大小、学习率和训练周期数除外。辍学概率始终保持在 0.1。最佳超参数值是特定于任务的，但我们发现以下可能值范围适用于所有任务：

* 批量大小： 16， 32
* 学习率（亚当）： 5e-5， 3e-5， 2e-5
* 周期数：2、3、4

我们还观察到，大型数据集（例如，100k+标记的训练示例）对超参数选择的敏感度远低于小数据集。微调通常非常快，因此只需对上述参数进行详尽搜索并选择在开发集上表现最佳的模型是合理的。

A.4 外变双向编码器表示法、嵌入式语言模型 和 开源人工智能 生成式预训练的外变网络模型 的比较

在这里，我们研究了最近流行的表示学习模型的差异，包括嵌入式语言模型，开源人工智能 生成式预训练的外变网络模型和外变双向编码器表示法。模型架构之间的比较如图 3 所示。请注意，除了架构差异之外，外变双向编码器表示法和开源人工智能 生成式预训练的外变网络模型是微调方法，而嵌入式语言模型是基于功能的方法。

与外变双向编码器表示法最相似的现有预训练方法是开源人工智能 生成式预训练的外变网络模型，它在大型文本语料库上训练从左到右的外变语言模型。事实上，外变双向编码器表示法中的许多设计决策都是有意做出的，以使其尽可能接近生成式预训练的外变网络，以便可以对这两种方法进行最低限度的比较。这项工作的核心论点是，第 3.1 节中提出的双向性和两个预训练任务 占了大部分实证改进，但我们确实注意到 外变双向编码器表示法 和 生成式预训练的外变网络 的训练方式之间还有其他几个差异：

* 生成式预训练的外变网络在书本语料库（8亿字）上接受培训;外变双向编码器表示法在书本语料库800M字）和维基百科（2，500M字）上接受培训。
* 生成式预训练的外变网络 使用句子分隔符 （[SEP]） 和分类符 （[CLS]），它们仅在微调时引入;外变双向编码器表示法学习[SEP]，[CLS]和预训练期间的句子A / B嵌入。
* 生成式预训练的外变网络 接受了 1M 步的培训，批量大小为 32，000 字;外变双向编码器表示法接受了1M步骤的培训，批量大小为128，000字。
* 生成式预训练的外变网络 在所有微调实验中使用相同的 5e-5 学习率;外变双向编码器表示法选择特定于任务的微调学习速率，该速率在开发集上表现最佳。

为了隔离这些差异的影响，我们在 Section 5.1 中进行消融实验，证明大多数改进实际上来自两个预训练任务及其实现的双向性。

A.5 对不同任务进行微调的图解

在不同任务上微调外变双向编码器表示法的图示如图4所示。我们的任务特定模型是通过将外变双向编码器表示法与一个额外的输出层相结合而形成的，因此需要从头开始学习最少数量的参数。在这些任务中，（a） 和 （b） 是序列级任务，而 （c） 和 （d） 是文字等级任务。在图中，*E*  表示输入嵌入，*T* i 表示文字*i* 的上下文表示，[CLS] 是分类输出的特殊符号，[SEP] 是分隔非连续标记序列的特殊符号。

# B 详细实验设置

B.1 自然语言理解能力评估基准实验的详细说明。

表1中的自然语言理解能力评估结果是从[排行榜](https://gluebenchmark.com/leaderboard)和 [https://blog 中获得 https://自然语言理解能力评估benchmark.com/。](https://blog.openai.com/language-unsupervised) [openai.com/language-unsupervised。](https://blog.openai.com/language-unsupervised)自然语言理解能力评估基准包括以下数据集，其描述最初由Wang等人总结。（2018年a）：

M自然语言推理多流派自然语言推理是一项大规模的众包蕴涵分类任务（Williams等人。，2018）。给定一对句子，目标是预测第二个句子相对于第一个句子是*蕴涵*、*矛盾*还是*中性*。

QQP Quora Question Pair是一个二元分类任务，其目标是确定在Quora上提出的两个问题在语义上是否等效（Chen等人。， 2018）。

Q自然语言推理 问题自然语言推理是斯坦福问答数据集的一个版本（Rajpurkar et al.，2016），已转换为二元分类任务（Wang等人。，2018a）。正面示例是包含正确答案的（问题，句子）对，而否定示例是来自不包含答案的同一段落的（问题，句子）。

|  |
| --- |
| 类 类  标签 标签  图 4：在不同任务上微调 外变双向编码器表示法 的图示。 |

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，并带有人类对其情感的注释（Socher et al.， 2013）。

CoLA 语言可接受 性语料库是一个二元单句分类任务，其目标是预测英语句子在语言上是否“可接受”（Warstadt et al.， 2018）。

STS-B 语义文本相似性基准是从新闻标题和其他来源中提取的句子对的集合（Cer 等人。， 2017）。它们被注释为1到5的分数，表示这两个句子在语义上的相似程度。

微软研究院释义语料库 Microsoft Research Paraphrase Corpus由从在线新闻来源自动提取的句子对组成，并带有人类注释，用于识别对中的句子在语义上是否等效（Dolan和Brockett， 2005）。

RTE识别文本蕴涵是一个类似于M自然语言推理的二元蕴涵任务，但训练数据要少得多（Bentivogli等人）。，2009 年）。[[11]](#footnote-11)

W自然语言推理 Winograd 自然语言推理是一个小型的自然语言推理数据集（Levesque et al.， 2011）.自然语言理解能力评估网页指出，这个数据集的构建存在问题， [[12]](#footnote-12) 提交给 自然语言理解能力评估的每个经过训练的系统的表现都比预测多数类的65.1基线准确性差。因此，我们排除了这个集合，以便对 开源人工智能 生成式预训练的外变网络模型公平。对于我们的自然语言理解能力评估提交，我们总是预测多数类。

# C 其他消融研究

C.1 训练步骤数的影响

图 5 显示了从已针对 k 步预先训练的检查点进行微调后的 M自然语言推理 开发阶段数据集准确性 。这使我们能够回答以下问题：

1. 问题：外变双向编码器表示法真的需要如此大量的预训练（128，000字/批\* 1，000，000步）才能达到很高的微调精度吗？

答：是的，与 1M 步相比，外变双向编码器表示法基础 在 1M 步上训练时在 M自然语言推理 上实现了近 500% 的精度。

1. 问题：有掩码的语言模型预训练的收敛速度是否比LTR预训练慢，因为每批中只预测了15%的单词，而不是每个单词？

答：有掩码的语言模型模型的收敛速度确实比LTR模型稍慢。然而，就绝对准确性而言，有掩码的语言模型模型几乎立即开始优于LTR模型。

C.2 不同掩码程序的消融术

在第 3.1 节中，我们提到 外变双向编码器表示法 在使用屏蔽语言模型 （有掩码的语言模型） 目标进行预训练时使用混合策略来屏蔽目标令牌。以下是评估不同掩蔽策略效果的消融研究。

200

400

600

800

1

*,*

000

76

78

80

82

84

预训练 步骤 （千）

MNLIDevAccuracy

伯特

基础

(

蒙面LM

)

伯特

基础

(

)

从左到右

图 5：训练步骤数的消融。这显示了微调后的 M自然语言推理 精度，从已针对 *k*  步预先训练的模型参数开始。x 轴是 *k* 的值。

请注意，掩码策略的目的是减少预训练和微调之间的不匹配，因为在微调阶段永远不会出现 [MASK] 符号。我们报告M自然语言推理和命名实体识别的Dev结果。对于 命名实体识别，我们报告了微调和基于特征的方法，因为我们预计基于特征的方法的不匹配将被放大，因为模型将没有机会调整表示。

屏蔽率 开发集结果

MASK SAME RND M自然语言推理 命名实体识别

微调 基于功能的微调

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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：不同掩蔽策略下的消融。

结果如表8所示。在表中，

MASK 意味着我们将目标文字替换为 有掩码的语言模型 的 [MASK] 符号;SAME 意味着我们保持目标文字不变 ; RND 意味着我们将目标文字替换为另一个随机文字。

表格左侧的数字表示有掩码的语言模型预训练期间使用的特定策略的概率（外变双向编码器表示法使用80%，10%，10%）。论文的右侧部分表示 Dev 集结果。对于基于特征的方法，我们将外变双向编码器表示法的最后4层连接为特征，这在第5.3节中被证明是最好的方法。

从表中可以看出，微调对不同的遮罩策略非常稳健。然而，正如预期的那样，在将基于特征的方法应用于命名实体识别时，仅使用M ASK策略是有问题的 。有趣的是，仅使用 RND 策略的性能也比我们的策略差得多。

1. https://github.com/tensorflow/tensor2tensor [↑](#footnote-ref-1)
2. http://nlp.seas.harvard.edu/2018/04/03/attention.html [↑](#footnote-ref-2)
3. 在所有情况下，我们将前馈/过滤器大小设置为4*H*,

   即，3072 表示*H*=768和 4096 为*H*=1024. [↑](#footnote-ref-3)
4. 我们注意到，在文献中，双向反式- [↑](#footnote-ref-4)
5. 最终模型在NSP上实现了97%-98%的准确率。 [↑](#footnote-ref-5)
6. 向量*C*不是没有微调的有意义的句子表示，因为它是用 NSP 训练的。 [↑](#footnote-ref-6)
7. GLUE数据集分布不包括测试标签，我们只为每个BERT进行了一次GLUE评估服务器提交基础和伯特大.10https://gluebenchmark.com/leaderboard [↑](#footnote-ref-7)
8. 问答网络 在 中描述俞等.(2018），但该系统在发布后有了很大的改进。 [↑](#footnote-ref-8)
9. 这我们使用的琐事问答数据由来自 琐事问答-Wiki 的段落组成，这些段落由文档中的前 400 个标记组成，其中包含至少一个提供的可能答案。 [↑](#footnote-ref-9)
10. https://cloudplatform.googleblog.com/2018/06/CloudTPU-now-offers-preemptible-pricing-and-globalavailability.html [↑](#footnote-ref-10)
11. 请注意，我们在本文中仅报告单任务微调结果。多任务微调方法可能会进一步提高性能。例如，我们确实观察到使用MNLI进行多任务训练的RTE有了实质性的改进。 [↑](#footnote-ref-11)
12. <https://gluebenchmark.com/faq> [↑](#footnote-ref-12)