BERT Pre-training of Deep Bidirectional Transformers for Language Understanding

用于语言理解的深度双向变压器的BERT预训练

论文笔记

1. Abstract

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT representations can be fine-tuned with just one additional output layer to create state-of-theart models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE benchmark to 80.4% (7.6% absolute improvement), MultiNLI accuracy to 86.7% (5.6% absolute improvement) and the SQuAD v1.1 question answering Test F1 to 93.2 (1.5 absolute improvement), outperforming human performance by 2.0.

我们引入了一种名为BERT的新语言表示模型，它代表变形金刚的双向编码器表示。与最近的语言表示模型（Peters et al。，2018; Radford et al。，2018）不同，BERT旨在通过联合调节所有层中的左右上下文来预训练深度双向表示。因此，只需一个额外的输出层即可对预先训练的BERT表示进行微调，以便为各种任务创建最先进的模型，例如问答和语言推断，而无需基本的任务特定架构修改。 BERT在概念上简单且经验丰富。它在11项自然语言处理任务中获得了最新的最新成果，包括将GLUE基准推至80.4％（绝对改进率7.6％），MultiNLI精度达到86.7％（绝对改进5.6％）和SQuAD v1.1问题回答测试F1到93.2（1.5绝对改进），超过人类表现2.0。

1. Problem

Language model pre-training has shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2017, 2018; Radford et al., 2018; Howard and Ruder, 2018). These tasks include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition (Tjong Kim Sang and De Meulder, 2003) and SQuAD question answering (Rajpurkar et al., 2016), where models are required to produce fine-grained output at the token-level.

There are two existing strategies for applying pre-trained language representations to downstream tasks: feature-based and fine-tuning. The feature-based approach, such as ELMo (Peters et al., 2018), uses tasks-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning the pretrained parameters. In previous work, both approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques severely restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a leftto-right architecture, where every token can only attended to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentencelevel tasks, and could be devastating when applying fine-tuning based approaches to token-level tasks such as SQuAD question answering (Rajpurkar et al., 2016), where it is crucial to incorporate context from both directions.

语言模型预训练已经证明对改进许多自然语言处理任务是有效的（Dai和Le，2015; Peters等，2017,2018; Radford等，2018; Howard和Ruder，2018）。这些任务包括句子级任务，如自然语言推理（Bowman et al。，2015; Williams et al。，2018）和释义（Dolan和Brockett，2005），旨在通过整体分析来预测句子之间的关系，以及令牌级任务，如命名实体识别（Tjong Kim Sang和De Meulder，2003）和SQuAD问题回答（Rajpurkar等，2016），其中模型需要在令牌级别生成细粒度输出。

将预训练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，2018），使用特定于任务的体系结构，其包括预先训练的表示作为附加特征。微调方法，例如Generative Pre-trained Transformer（OpenAI GPT）（Radford等，2018），引入了最小的任务特定参数，并通过简单地微调预训练参数来训练下游任务。在以前的工作中，两种方法在预训练期间共享相同的目标函数，在这些方法中，他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严重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OpenAI GPT中，作者使用左右架构，其中每个令牌只能处理Transformer的自我关注层中的先前令牌（Vaswani等，2017）。这些限制对于句子级别任务来说是次优的，并且在将基于微调的方法应用于令牌级别任务（例如SQuAD问答）时可能是毁灭性的（Rajpurkar等，2016），其中从两个方向合并上下文至关重要。

1. Idea

In this paper, we improve the fine-tuning based approaches by proposing BERT: Bidirectional Encoder Representations from Transformers. BERT addresses the previously mentioned unidirectional constraints by proposing a new pre-training objective: the “masked language model” (MLM), inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. Unlike left-to-right language model pre-training, the MLM objective allows the representation to fuse the left and the right context, which allows us to pre-train a deep bidirectional Transformer. In addition to the masked language model, we also introduce a “next sentence prediction” task that jointly pre-trains text-pair representations. The contributions of our paper are as follows:

• We demonstrate the importance of bidirectional pre-training for language representations. Unlike Radford et al. (2018), which uses unidirectional language models for pretraining, BERT uses masked language models to enable pre-trained deep bidirectional representations. This is also in contrast to Peters et al. (2018), which uses a shallow concatenation of independently trained leftto-right and right-to-left LMs.

• We show that pre-trained representations eliminate the needs of many heavilyengineered task-specific architectures. BERT is the first fine-tuning based representation model that achieves state-of-the-art performance on a large suite of sentence-level and token-level tasks, outperforming many systems with task-specific architectures.

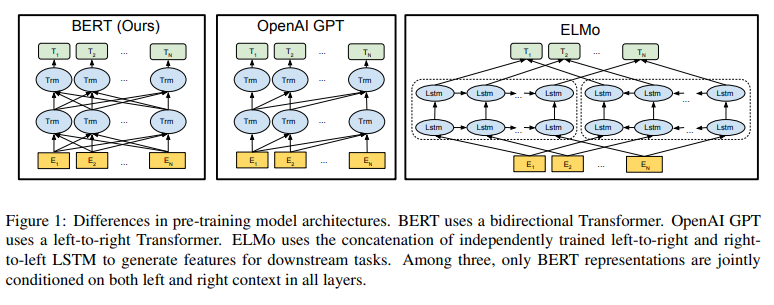
• BERT advances the state-of-the-art for eleven NLP tasks. We also report extensive ablations of BERT, demonstrating that the bidirectional nature of our model is the single most important new contribution. The code and pre-trained model will be available at goo.gl/language/bert.

在本文中，我们通过提出BERT：变换器的双向编码器表示来改进基于微调的方法。 BERT通过提出一个新的预训练目标来解决前面提到的单向约束：“掩盖语言模型”（MLM），受到完形任务的启发（Taylor，1953）。被掩盖的语言模型从输入中随机地掩盖一些标记，并且目标是仅基于其上下文来预测被掩盖的单词的原始词汇id。与从左到右的语言模型预训练不同，MLM目标允许表示融合左右上下文，这允许我们预训练深度双向变换器。除了蒙面语言模型，我们还引入了一个“下一句预测”任务，它共同预先训练文本对表示。本文的贡献如下：

•我们证明了双向预训练对语言表达的重要性。与Radford等人不同。 （2018），其使用单向语言模型进行预训练，BERT使用掩蔽语言模型来实现预训练的深度双向表示。这也与Peters等人形成对比。 （2018），其使用由独立训练的左右和右到左LM的浅层连接。

•我们展示了预先训练的表示消除了许多重型工程任务特定体系结构的需求。 BERT是第一个基于微调的表示模型，它在大量的句子级和令牌级任务上实现了最先进的性能，优于许多具有任务特定体系结构的系统。

•BERT推进了11项NLP任务的最新技术。我们还报告了对BERT的广泛消融，证明了我们模型的双向性质是最重要的新贡献。代码和预先训练的模型将在goo.gl/language/bert上提供。



1. Conclusion

Recent empirical improvements due to transfer learning with language models have demonstrated that rich, unsupervised pre-training is an integral part of many language understanding systems. In particular, these results enable even low-resource tasks to benefit from very deep unidirectional architectures. Our major contribution is further generalizing these findings to deep bidirectional architectures, allowing the same pre-trained model to successfully tackle a broad set of NLP tasks. While the empirical results are strong, in some cases surpassing human performance, important future work is to investigate the linguistic phenomena that may or may not be captured by BERT.

由于使用语言模型进行迁移学习，最近的经验改进表明，丰富的，无监督的预训练是许多语言理解系统的组成部分。 特别是，这些结果使得即使是低资源任务也能从非常深的单向体系结构中受益。 我们的主要贡献是将这些发现进一步推广到深度双向架构，允许相同的预训练模型成功解决一系列广泛的NLP任务。 虽然实证结果很强，在某些情况下超过人类表现，但未来重要的工作是调查BERT可能会或可能不会捕获的语言现象。