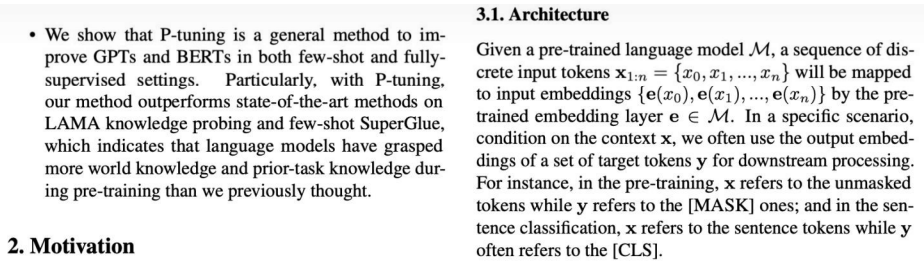


- We show that P-tuning is a general method to improve GPTs and BERTs in both few-shot and fully-supervised settings. Particularly, with P-tuning, our method outperforms state-of-the-art methods on LAMA knowledge probing and few-shot SuperGlue, which indicates that language models have grasped more world knowledge and prior-task knowledge during pre-training than we previously thought.

2. Motivation

3.1. Architecture

Given a pre-trained language model $\mathcal{M}$, a sequence of discrete input tokens $\mathbf{x}_{1:n} = \{x_0, x_1, ..., x_n\}$ will be mapped to input embeddings $\{\mathbf{e}(x_0), \mathbf{e}(x_1), ..., \mathbf{e}(x_n)\}$ by the pre-trained embedding layer $\mathbf{e} \in \mathcal{M}$. In a specific scenario, condition on the context $\mathbf{x}$, we often use the output embeddings of a set of target tokens $\mathbf{y}$ for downstream processing. For instance, in the pre-training, $\mathbf{x}$ refers to the unmasked tokens while $\mathbf{y}$ refers to the [MASK] ones; and in the sentence classification, $\mathbf{x}$ refers to the sentence tokens while $\mathbf{y}$ often refers to the [CLS].

【P-Tuning】 一种自动学习 prompt pattern 的方法（附源码）

何枝
欣赏每一个用逻辑阐述观点的人，不喜欢无论据的情绪输出。

prompt 是当前 NLP 中研究小样本学习方向上非常重要的一个方向。

如果你对 prompt 还不太了解的话，不妨先看看这个视频：

简单了解prompt learning是什么?
4.6 万播放 · 48 赞同 视频

举例来讲，今天如果有这样两句评论：

> 1. 什么苹果啊，都没有苹果味，怪怪的味道，而且一点都不甜，超级难吃！
> 2. 这破笔记本速度太慢了，卡的不要不要的。

现在我们需要根据他们描述的商品类型进行一个分类任务，

即，第一句需要被分类到「水果」类别中；第二句则需要分类到「电脑」类别中。

一种直觉的方式是将该问题建模成一个传统文本分类的任务，通过人工标注，为每一个类别设置一个id，例如：

```
{
    '电脑': 0,
    '水果': 1,
    ....
}
```

这样一来，标注数据集就长这样：

```
什么苹果啊，都没有苹果味，怪怪的味道，而且一点都不甜，超级难吃！    1
这破笔记本速度太慢了，卡的不要不要的。      0
...
```

这种方法是可行的，但是需要「较多的标注数据」才能取得不错的效果。

由于大多数预训练模型（如BRET）在 pretrain 的时候都使用了 [MASK] token 做 MLM 任务，而我们在真实下游任务中往往是不会使用到 [MASK] 这个 token，这就意味着今天我们在训练下游任务时需要较多的数据集去抹平上下游任务不一致的 gap。

那，如果我们没有足够多的训练数据怎么办呢？

prompt learning 的出现就是为了解决这一问题，它将 [MASK] 的 token 引入到了下游任务中，将下游任务构造成和 MLM 类似的任务。

举例来讲，我们可以将上述评论改写为：

```
这是一条[MASK][MASK]评论：这破笔记本速度太慢了，卡的不要不要的。
```

然后让模型去预测两个 [MASK] token 的真实值是什么，那模型根据上下文能推测出被掩码住的词应该为「电脑」。

由于下游任务中也使用了和预训练任务中同样的 MLM 任务，这样我们就可以使用更少的训练数据来进行微调了。

但，这还不是 P-tuning。

通过上面的例子我们可以观察到，构建句子最关键的部分是在于 prompt的生成，即：

```
「这是一条[MASK][MASK]评论：」(prompt) + 这破笔记本速度太慢了，卡的不要不要的。(content)
```

被括号括起来的前缀（prompt）的生成是非常重要的，不同 prompt 会极大影响模型对 [MASK] 预测的正确率。

那么这个 prompt 怎么生成呢?

我们当然可以通过人工去设计很多不同类型的前缀 prompt，我们把他们称为 prompt pattern，例如：

```
这是一条[MASK][MASK]评论：
下面是一条描述[MASK][MASK]的评论：
[MASK][MASK]：
...
```

但是人工列这种 prompt pattern 非常的麻烦，不同的数据集所需要的 prompt pattern 也不同，可复用性很低。

那么，我们能不能通过机器自己去学习 prompt pattern 呢?

这，就是 P-Tuning。

1. P-Tuning

人工构建的模板对人类来讲是合理的，但是在机器眼中，prompt pattern 长成什么样真的关键吗?

机器对自然语言的理解和人类对自然语言的理解很有可能不尽相同，我们曾经有做一个 model attention 和人类对语言重要性的理解的对比实验，发现机器对语言的理解和人类是存在一定的偏差的。

Attention可以用来解释模型吗?
380 播放 · 1 赞同 视频

那么，我们是不是也不用特意为模型去设定一堆我们觉得「合理」的 prompt pattern，而是让模型自己去找它们认为「合理」的prompt pattern 就可以了呢?

成、mlm loss 计算 三

1.1 prompt token(s) 生成

既然现在我们不用人工去构建 prompt 模板，我们也不清楚机器究竟喜欢什么样的模板……

那不如我们就随便凑一个模板丢给模型吧。

听起来很草率，但确实就是这么做的。

我们选用中文 BERT 作为 backbon 模型，选用 vocab.txt 中的 [unused] token 作为构成 prompt 模板的元素。

> [unused] 是 BERT 词表里预留出来的未使用的 token，其本身没有什么含义，随意组合也不会产生很大的语义影响，这也是我们使用它来构建 prompt 模板的原因。

那么，构建出来的 prompt pattern 就长这样：

```
[unused1][unused2][unused3][unused4][unused5][unused6]
```

1.2 mask label 生成

完成 prompt 模板的构建后，我们还需要把 mask label 给加到句子中，好让模型帮我们完成标签预测任务。

我们设定 label 的长度为2（'水果'、'电脑'，都是 2 个字的长度），并将 label 塞到句子的开头位置：

```
[CLS][MASK][MASK]这破笔记本速度太慢了，卡的不要不要的。[SEP]
```

其中 [MASK] token 就是我们需要模型帮我们预测的标签 token，现在我们把两个部分拼起来：

```
[unused1][unused2][unused3][unused4][unused5][unused6][CLS][MASK][MASK]这破笔记本速度太
```

这就是我们最终输入给模型的样本。

1.3 mlm loss 计算

下面就要开始进行模型微调了，我们喂给模型这样的数据：

```
[unused1][unused2][unused3][unused4][unused5][unused6][CLS][MASK][MASK]这破笔记本速度太
```

并获得模型预测 [MASK] token 的预测结果，并计算和真实标签之间的 CrossEntropy Loss。

P-Tuning 中标签数据长这样：

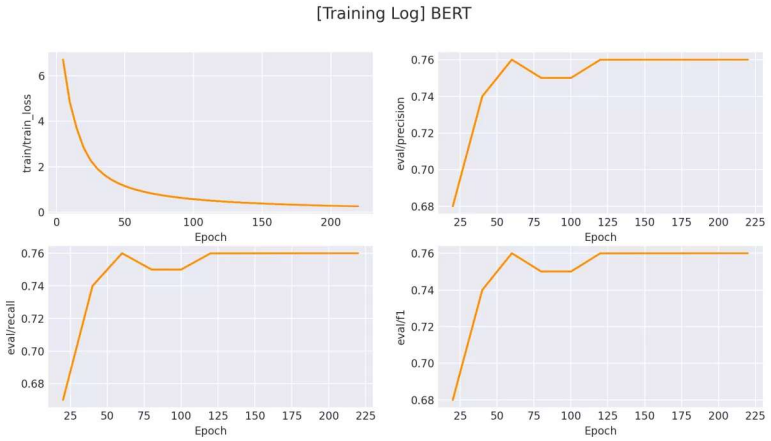
```
水果    什么苹果啊，都没有苹果味，怪怪的味道，而且一点都不甜，超级难吃！
电脑    这破笔记本速度太慢了，卡的不要不要的。
...
```

也就是说，我们需要计算的是模型对 [MASK] token 的输出与「电脑」这两个标签 token 之间的 CrossEntropy Loss，以教会模型在这样的上下文中，被 [MASK] 住的标签应该被还原成「物品类别」。

我们选用 63 条评论（8 个类别）的评论作为训练数据，在 417 条评论上作分类测试，模型 F1 能收敛在 76%。

通过实验结果我们可以看到，基于 prompt 的方式即使在训练样本数较小的情况下模型也能取得较为不错的效果。

相比于传统的分类方式，P-Tuning 能够更好的缓解模型在小样本数据下的过拟合，从而拥有更好的鲁棒性。



P-tuning 训练曲线图

论文链接：arxiv.org/pdf/2103.1038...

源码链接：

https://github.com/HarderThenHarder/transformers_tasks/tree/main/prompt_tasks/p-tuning
github.com/HarderThenHarder/transformers_tasks/tre...

编辑于 2023-01-07 19:53 · IP 属地河北

自然语言处理　小样本学习 (Few-Shot Learning)　深度学习（Deep Learning）

写下你的评论...

7 条评论　默认　最新

sqlplus
请问前辈 只有两张卡能走这个方向吗😭
01-26 · IP 属地北京　赞

Justeagles ‣ **何枝**
小样本指的是下游任务训练数据要求少，又不是指模型小啊。prompt要想效果好，预训练模型肯定不能小的。
02-02 · IP 属地中国香港　赞

何枝 作者 ‣ **sqlplus**
嗯嗯是的，可以尝试一下，prompt- tuning会比pretrain甚者是fine tune用到更少的训练数据，这意味着实验耗时更少，对算力要求更低（至于效果怎么样就不能保证了）
01-27 · IP 属地四川　赞

展开其他 2 条回复 >

play
厉害呀看了好多p-tuning的文章都没弄懂，现在终于明白了
01-14 · IP 属地新加坡 赞

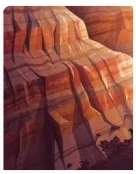
茶一白
好牛 膜拜大佬
2022-11-15 · IP 属地江苏 赞

文章被以下专栏收录

何小枝与NLP的快乐日常
纸上得来终觉浅。

推荐阅读

集成学习三大法宝-bagging、boosting、stacking
周宁



深度学习力0的第一步
张轩铭