

# 自然语言模型性能对比

ZY2203402 冯士轩

## 一、作业题目

任意选取 3~4 个目前前沿语言大模型（不限中文或者英文），通过提示工程的方法来检验和对比不同模型下游任务上的性能，可选择 3~5 个不同的自然语言下游任务来进行测试。

## 二、实验原理

本次作业使用了 Bert、T5、GPT-2 三种语言模型，完成文本分类、摘要生成、问答、翻译四种下游任务。

### 2.1 Bert 模型

Bert 是一个预训练的神经网络模型，基于 transformer 架构，能学习通用的句子表示形式，适用于各种自然语言处理任务。它采用预训练-微调的方法，在大型语料库上进行预训练，然后通过微调到特定的语言任务上进行应用。BERT 的预训练任务包括掩蔽语言模型和下一句预测。在微调阶段，可以使用各种具体的自然语言处理任务训练 BERT 模型，不需要从头开始训练新的模型。BERT 在 NLP 领域应用广泛，适用于文本分类、问答、命名实体识别等任务。

### 2.2 T5 模型

T5（Text-to-Text Transfer Transformer）是一种预训练语言模型，由 Google 在 2019 年提出。T5 模型将各种自然语言处理任务统一为文本到文本转换任务（Text-to-Text Transfer），通过在大规模语料库上进行无监督预训练，学习到通用的文本表示形式，然后通过微调在各种具体 NLP 任务中进行使用，包括问答、文本分类、翻译等。T5 模型同样采用了 transformer 架构，其中包括了 encoder 和 decoder 两个模块，可以完成生成式任务（例如翻译）和分类任务（例如情感分析）。T5 模型的目标是让模型通过输入一个文本序列，输出一个目标文本序列的过程，因此其输入和输出的模式都是文本序列。T5 的优点是具有通用性，能够应对不同的 NLP 任务，同时其性能也已经在多项任务上超过了之前的 SOTA（state-of-the-art）模型。

### 2.3 GPT-2 模型

GPT-2（Generative Pretrained Transformer 2）是一种基于 transformer 架构的预训练语言模型，由 OpenAI 在 2019 年发布。这个模型在大规模语料库上进行了无监督的预训练，从

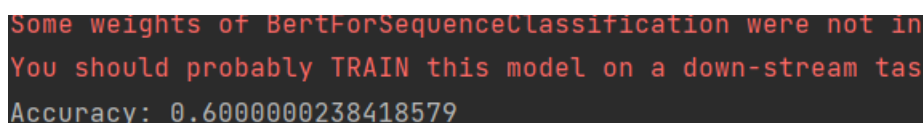
而学习到了通用的语言表示,可以在各种自然语言处理任务中进行微调,如问答、文本分类、生成文本等。GPT-2 模型使用了自回归的方法,即在输入一个初始文本片段后,按照模型所学到的语言规律依次生成下一个单词,即自动生成完整的文本。它的训练方式通过最大化训练数据的似然度完成,具有极强的语言生成能力,并能够跨越短文本到长文本的长度范围,能够生成几乎与人类水平相当的连贯和自然的文本。GPT-2 与其前身 GPT 相比,在模型规模、训练数据和训练任务等方面都有了极大的提高,成为了自然语言生成领域的重要研究成果。

## 三、实验过程

### 3.1 文本分类任务

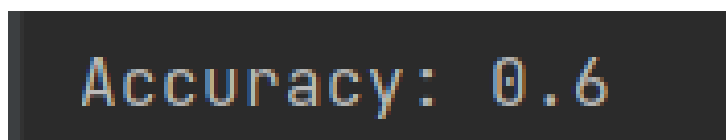
文本分类任务是给定一段文本,将其分类到事先定义好的若干类别之一。在这次实验当中我们选定五个文本:["这家餐馆的食物真是太好吃了","这部电影真棒","这辆车很难开","这本书太无聊了","这首歌非常好听"],使用三种模型将它们以积极和消极两种类型进行分类。将预测的标签和其真实标签进行对比,判断他们在文本分类任务的性能。

Bert 的分类结果如下图所示:



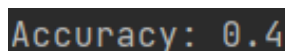
```
Some weights of BertForSequenceClassification were not in  
You should probably TRAIN this model on a down-stream tas  
Accuracy: 0.6000000238418579
```

T5 模型的分类结果如下图所示:



```
Accuracy: 0.6
```

GPT-2 模型的分类准确率如下图所示



```
Accuracy: 0.4
```

从上图可以看出 GPT-2 的任务性能要弱于 Bert 和 T5 模型,由于分类文本数量的原因,无法对 Bert 和 T5 模型进行进一步的对比。

### 3.2 问答任务

问答任务是给定一个问题和一篇文本,从文本中寻找答案并返回。在本次实验中我们选用的问题和文本分别是:

```
text = "紫禁城又称故宫，是中国北京市中心的一组古建筑群，是明清两代的皇宫。"
```

```
question = "紫禁城是什么？"
```

```
text1 = "France is a country in Europe. Its capital is Paris and its population is over 66 million."
```

```
question1 = "What is the capital of France?"
```

使用三种模型对这两个问题进行问答，通过答案判断模型在次任务的性能。

Bert 模型在此项任务的回答是：

```
城    ，  中国北京中心的一古    ，  明清代的皇    。  
. its capital is paris
```

T5 在此项任务的回答是：

```
Paris  
not_duplicate
```

GPT-2 在此项任务的回答是：

```
你 应该 可能 训练 这个 模型 在 一个 下游 任务 上 以 便 能够 回答 问题  
答案： Paris and its population is over  
答案1： 是中国北京市中心的一组古建筑群
```

从三个模型在问答任务的回答上来看，英文问题基本都能找到巴黎这个单词，但是 GPT-2 模型生成的答案中含有多余答案。而中文问题 GPT-2 回答效果最好，Bert 模型能大概明白意思，而 T5 模型可能是因为选用模型不当而导致问题和答案不匹配。

### 3.3 摘要生成任务

摘要生成任务是给定一段文字，从这段文字中提炼关键信息，生成新的文字。三个模型的性能指标使用 ROUGE 指标评估效果。ROUGE 指标有多种形式，例如 ROUGE-1（基于 unigram 的重叠率）、ROUGE-2（基于 bigram 的重叠率）、ROUGE-L（最长公共子序列的 F1 得分）等。其中，ROUGE-1 和 ROUGE-2 关注相邻字词的匹配，ROUGE-L 匹配最长公共子序列，并考虑了单词级别的语义表达差异。在 ROUGE 指标中，R、P、F 分别表示召回率、准确率和 F1 得分，分别表示包含多少原始文本中的信息、正确的信息数量占有所有生成的信息数量的比例、同时反映了召回率和准确率，是二者的调和平均值。

由于在实际操作时，输入中文文本表现不佳，故选用英文文本来测试三个模型的性能。此项任务我们输入的文本为 `input_text = "The method of machine learning effectively utilizes the powerful computational performance of computers, utilizing statistical knowledge theory to`

efficiently and reasonably model massive amounts of text information, and can discover hidden attributes hidden in massive amounts of text information". 设定的输出最大长度为 50.

Bert 在此项任务当中的表现:

生成的摘要为: the method of machine learning effectively utilizes the powerful computational performance of computers, utilizing statistical knowledge theory to efficiently and reasonably model massive amounts of text information, and can discover hidden attributes hidden in massive amounts of text information this method.

ROUGE 评分为: [{'rouge-1': {'r': 0.9090909090909091, 'p': 0.9375, 'f': 0.9230769180781065}, 'rouge-2': {'r': 0.918918918918919, 'p': 0.8947368421052632, 'f': 0.9066666616675557}, 'rouge-l': {'r': 0.9090909090909091, 'p': 0.9375, 'f': 0.9230769180781065}}]

```
[{"this IS NOT expected if you are initializing BertForMaskedLM from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification). The method of machine learning effectively utilizes the powerful computational performance of computers, utilizing statistical knowledge theory to efficiently and reasonably model massive amounts of text information, and can discover hidden attributes hidden in massive amounts of text information. {'rouge-1': {'r': 0.9090909090909091, 'p': 0.9375, 'f': 0.9230769180781065}, 'rouge-2': {'r': 0.918918918918919, 'p': 0.8947368421052632, 'f': 0.9066666616675557}, 'rouge-l': {'r': 0.9090909090909091, 'p': 0.9375, 'f': 0.9230769180781065}}]
```

T5 在此项任务当中的表现:

生成的摘要为: learning effectively utilizes the powerful computational performance of computers, utilizing statistical knowledge theory to efficiently and reasonably model massive amounts of text information, and can discover hidden attributes hidden in massive amounts of text information.

ROUGE 评分为: [{'rouge-1': {'r': 0.9032258064516129, 'p': 1.0, 'f': 0.9491525373858086}, 'rouge-2': {'r': 0.8857142857142857, 'p': 1.0, 'f': 0.9393939344123049}, 'rouge-l': {'r': 0.9032258064516129, 'p': 1.0, 'f': 0.9491525373858086}}]

```
摘要: learning effectively utilizes the powerful computational performance of computers, utilizing statistical knowledge theory to efficiently and reasonably model massive amounts of text information. {'rouge-1': {'r': 0.9032258064516129, 'p': 1.0, 'f': 0.9491525373858086}, 'rouge-2': {'r': 0.8857142857142857, 'p': 1.0, 'f': 0.9393939344123049}, 'rouge-l': {'r': 0.9032258064516129, 'p': 1.0, 'f': 0.9491525373858086}}]
```

GPT-2 在此项任务当中的表现:

生成的摘要为: The method of machine learning effectively utilizes the powerful computational performance of computers, utilizing statistical knowledge theory to efficiently and reasonably model massive amounts of text information, and can discover hidden attributes hidden in massive amounts of text information.

ROUGE 评分为: [{'rouge-1': {'r': 1.0, 'p': 1.0, 'f': 0.999999995}, 'rouge-2': {'r': 1.0, 'p': 0.9722222222222222, 'f': 0.9859154879587383}, 'rouge-l': {'r': 1.0, 'p': 1.0, 'f': 0.999999995}}]

```
摘要: The method of machine learning effectively utilizes the powerful computational performance of computers, utilizing statistical knowledge theory to efficiently and reasonably model massive amounts of text information. The method of machine learning effectively utilizes the powerful computational performance of computers, utilizing statistical knowledge theory to efficiently and reasonably model massive amounts of text information. {'rouge-1': {'r': 1.0, 'p': 1.0, 'f': 0.999999995}, 'rouge-2': {'r': 1.0, 'p': 0.9722222222222222, 'f': 0.9859154879587383}, 'rouge-l': {'r': 1.0, 'p': 1.0, 'f': 0.999999995}}]
```

三个模型生成的摘要字数都在 40 左右。从 ROUGE 性能指标来看, GPT-2 产生的摘要

质量要高于 Bert、T5 模型的，而 T5 模型生成质量要高于 Bert 模型。

### 3.4、翻译任务

此次实验翻译任务是将中文翻译成英文。

这里我们使用将"你好，你今天怎么样"翻译为"Hello, how are you today"。

Bert 模型的翻译结果：

```
Source text: 你好，你今天怎么样
Translated text: hello, how are you today as as as as
```

T5 模型的翻译结果是：

```
Source text: 你好，你今天怎么样?
Translated text: Hallo, wie sind Sie heute?
```

GPT-2 的翻译结果是

```
Source text: 你好，你今天怎么样
Translated text: Hello. How are you today?
```

从上面结果来看 GPT-2 完美的将中文翻译为英文，而 Bert 在翻译最后控制不住长度导致生成了无效信息，而 T5 模型在翻译效果不好。

## 四、总结

这次作业由于我对各个模型不够了解，各个模型无论是调参方面还是最佳模型使用方面都做得不是很少。在实际操作时，由于选用的是预训练模型，在使用时我训练很少或者没有训练，导致对模型在任务上性能对比影响很大。

本学期经过大作业的训练，我已经掌握了一些自然语言处理的方法，但这些方法还比较简单，针对更加复杂的自然语言处理的问题，还需要以后继续深入地学习。