-作业五、多种大模型的下游任务测试

崔多
1479518308@qq.com

一 摘要

本文选用 GPT3，ERNIE，T5 等多种大模型，通过提示工程的方式，根据文本分类，文本生成，文本预测等来检验和对比不同模型下游任务上的性能，并对模型效果进行评估和比较。

二 研究内容简介

通过对比不同大型语言模型在多个下游 NLP 任务上的表现，可以更好地了解各个大模型的优缺点，为实际应用提供指导。

GPT-3、ERNIE 和 T5 都是目前非常流行的大型语言模型，它们在不同的任务上都有出色的表现。比如，GPT-3 在生成式任务上表现出色，ERNIE 在中文 NLP 任务上表现优异，T5 则在多语言翻译任务上表现出色。

三 实验方法

3.1 GPT-3 模型

GPT-3（Generative Pre-trained Transformer 3）是由 OpenAI 开发的一种自然语言处理模型，它是目前最先进的语言模型之一。GPT-3 是基于 Transformer 架构的深度神经网络模型，它使用了超过 1750 亿个参数进行训练，是目前最大的语言模型之一。

GPT-3 的训练数据来自于互联网上的大量文本，包括书籍、文章、博客、维基百科等。通过这些数据的训练，GPT-3 可以自动生成高质量的自然语言文本，包括文章、对话、摘要、翻译等。其主要特点是它可以在没有任何人工干预的情况下生成高质量的文本，同时还可以执行各种自然语言处理任务，如问答、文本分类、命名实体识别等。它还可以在不同的任务之间进行迁移学习，从而提高模型的性能。

GPT-3 的出现引起了广泛的关注和讨论，它被认为是自然语言处理领域的一次重大突破，为自然语言处理技术的发展带来了新的可能性。

3.2 ERNIE 模型

ERNIE（Enhanced Representation through kNowledge IntEgration）是百度公司开发的一种自然语言处理模型，它是基于 Transformer 架构的深度神经网络模型。ERNIE 的主要特点是它可以将外部知识与文本信息相结合，从而提高模型的表现。

ERNIE 的训练数据来自于互联网上的大量文本，包括新闻、百科、论坛等。与其他语言模型不同的是，ERNIE 还使用了一些外部知识库，如百度百科、互动百科等，从而提高模型的知识表示能力。

ERNIE 的主要应用包括文本分类、命名实体识别、关系抽取、机器阅读理解等。在这

些任务中，ERNIE 都取得了很好的表现，并且在一些竞赛中取得了优异的成绩。

除了基于中文的 ERNIE，百度还开发了 ERNIE-Gram 模型，它是基于英文的 ERNIE 模型，可以处理英文文本。ERNIE-Gram 的训练数据来自于互联网上的大量英文文本，包括新闻、维基百科、博客等。ERNIE-Gram 在英文文本分类、情感分析等任务中也取得了很好的表现。

ERNIE 是一种非常有前途的自然语言处理模型，它的知识表示能力和表现都非常优秀。随着 ERNIE 的不断发展和完善，相信它将在自然语言处理领域发挥越来越重要的作用。

3.3 T5 模型

T5（Text-to-Text Transfer Transformer）是由 Google Brain 开发的一种自然语言处理模型，它是基于 Transformer 架构的深度神经网络模型。T5 的主要特点是它可以将不同的自然语言处理任务转换为文本到文本的转换任务，从而实现多任务学习。其训练数据来自于互联网上的大量文本，包括书籍、文章、维基百科等。与其他语言模型不同的是，T5 将不同的自然语言处理任务转换为文本到文本的转换任务，如文本分类、命名实体识别、机器翻译等。这种方法可以使模型在不同的任务之间进行迁移学习，从而提高模型的性能。

T5 模型的主要应用包括文本分类、命名实体识别、机器翻译、问答系统等。在这些任务中，T5 都取得了很好的表现，并且在一些竞赛中取得了优异的成绩。除了基于英文的 T5 模型，Google 还开发了 mT5 模型，它是基于多语言的 T5 模型，可以处理多种语言的文本。mT5 的训练数据来自于互联网上的大量多语言文本，包括新闻、维基百科、博客等。mT5 在多语言文本分类、机器翻译等任务中也取得了很好的表现。

3.4 bert 模型

BERT（Bidirectional Encoder Representations from Transformers）是一种预训练的自然语言处理模型，由 Google 在 2018 年发布。BERT 使用 Transformer 架构，可以在大规模语料库上进行无监督的预训练，然后在各种下游 NLP 任务上进行微调，从而实现了在多个任务上的卓越表现。

BERT 的主要特点是双向性，即它可以同时考虑文本的上下文信息，而不是像传统的语言模型那样只考虑前面的文本。这使得 BERT 能够更好地理解语言的含义和语境，并在各种 NLP 任务中取得更好的结果。

四 实验结果

4.1 文本情感分类

1）数据集：

为了验证在数据集上不同大模型的文本情感分类问题，本文选用公开数据集 ChnSentiCorp 数据集作为测试对象。该数据集是一个中文情感分析数据集，由哈尔滨工业大学社会计算与信息检索研究中心发布。本研究采用酒店评论子集，其数据量为 9600 条。每个评论都被标记为正面、负面或中性情感。该数据集被广泛用于中文情感分析的研究和应用中，是中文情感分析领域中的重要数据集之一。其数据集实例如下：

'选择珠江花园的原因就是方便，有电动扶梯直接到达海边，周围餐馆、食廊、商场、超市、摊位一应俱全。酒店装修一般，但还算整洁。 泳池在大堂的屋顶，因此很小，不过女儿倒是喜欢。 包的早餐是西式的，还算丰富。 服务吗，一般', 1

数据集中，正向情感为 1，负向情感为 0。

2）情感分类结果：

本研究采用三种大数据模型，分别为：bert 模型，albert 模型以及 ernie 模型，权重数据来源于网站 https://huggingface.co/models.

根据大模型的输出矩阵，本文建立全连接神经网络，将输出数据分为两类，即正向情感与负向情感，。为了获取全连接权重，本文以 16 条评论为一组，训练 30 轮。将训练结果作用于随机抽取的 200 条评论，所获得的准确率如表 4.1 所示。

表 4. 1 不同模型文本分类准确率

模型名称	bert 模型	albert 模型	ernie 模型
准确率	0.75	0.78	0.71

在训练轮次较少的情况下，模型表现出来较好的文本分类能力，由于硬件限制，本文没有增加训练轮次，以探索模型在本测试集上的最优效果，但已经说明了大模型较好的泛化能力。

4.2 文本预测

1）文本内容：

本文选用 5 条文本进行内容理解与预测，文本内容如表 4.2 所示，mask 为等待预测的内容

表 2 预测文本

文本内容	**预测内容**
太阳从[MASK][MASK]升起，西方落下。	**东方**
[MASK][MASK]是中国的首都。	**北京**
[MASK][MASK][MASK][MASK]是一个成语，最早出自于战国·楚·宋玉《风赋》。空穴来风（穴：洞）指有孔洞便会进风。比喻消息和传闻的产生都是有原因和根据的；也比喻消息和传闻毫无根据。	**空穴来风**
[MASK][MASK][MASK]与红楼梦，三国演义，水浒传并称为中国四大名著。	**西游记**
[MASK][MASK][MASK][MASK]together with Dream of the Red Mansion, Romance of the Three Kingdoms, and Water Margin, it is known as China's four masterpieces.	**Journey to the West**

2）预测结果：

本文选用 gpt3.5 和 ernie 模型进行预测，对于前四条中文问题，两个模型都表现出来较好的预测效果，预测结果完全准确。

而对于最后一个英文问题，ernie 模型作为一个基于中文进行训练的模型，预测效果并不理想，生成为乱码，而 gpt3.5 给出了准确的答案。

4.3 文本生成模型

本文针对不同模型提出问题与获取到的答案如下表所示：

表 3 不同模型问题解答

Q	**At my age you will probably have learnt one lesson. Hypothesis: It's not certain how many lessons you'll learn by your thirties. Does the premise entail the hypothesis?**
T5	**it is not possible to tell**
GPT3	**No, the premise does not entail the hypothesis.**
bert	**at my age you will probably havent one lesson hass**

4.2 问答模型

以较为简单的任务为例：问答模型的结构如下所示，通过已有内容，根据问题得到答案，

展现了模型的文本提取能力。

表 4 问答模型实例

Q	What's my name?
C	My name is Clara and I live in Berkeley.
A	Clara

本文针对同一内容，对不同模型进行测设，所获得结果为：

表 5 不同温大模型回答结果

	roberta	bert	big_bird roberta
Q	Why is model conversion important?		
C	The option to convert models between FARM and transformers gives freedom to the user and let people easily switch between frameworks.		
A	gives freedom to the user	gives freedom	gives freedom to the user

五 结论

通过对比多种模型在多个下游任务上的表现，可以更好地了解它们的优劣势。比如，如果需要处理中文 NLP 任务，ERNIE 可能是更好的选择；如果需要进行多语言翻译，T5 可能更适合。同时，这种对比也可以为模型的改进提供指导，例如，可以通过对比不同模型在命名实体识别任务上的表现，来改进模型对实体识别的支持。

通过对比不同大型语言模型在多个下游 NLP 任务上的表现，可以更好地了解它们的性能和适用范围，为实际应用提供指导 ，且这种对比也可以为研究人员提供有关如何选择和使用不同模型的建议，以及如何进一步改进这些模型的方向和思路。因此，对比不同大型语言模型在多个下游 NLP 任务上的表现是非常有价值的，可以为 NLP 领域的研究和应用带来更多的启示和进展。