# 基于前沿 NLP 大模型的不同下游任务性能对比报告

宋浩瑜 ZY2203207

## 目录

# 摘要

本报告旨在分析并对比当前前沿自然语言处理（NLP）大模型在不同下游任务上的性能。我们将重点关注几种预训练模型，如 GPT 系列以及其他相关模型，并探讨它们在情感分析、机器翻译、文本生成等任务中的表现。

# 1. 引言

随着深度学习技术的不断发展，自然语言处理领域涌现出了众多强大的预训练模型。这些模型通常采用 Transformer 架构，经过大量无标签文本数据的预训练，进而在各种下游任务中进行微调。本报告将重点比较这些模型在不同任务上的性能。

# 2. 下游任务概述

## 2.1 情感分析

情感分析任务旨在从文本数据中识别和提取作者的情感倾向。具体来说，这包括判断文本中表达的情感是积极的、消极的或中性的。

例子 1：这款手机的电池续航非常出色。屏幕显示效果也很清晰，运行速度非常快。总体来说，我非常满意这次购物。

例子 2：酒店房间设施陈旧，床垫很硬，让人难以入睡。前台服务态度冷淡，与客人沟通不耐烦。这次的住宿体验令人失望。

例子 3：这部电影的剧情设定很有创意，视觉效果震撼。虽然部分角色塑造略显薄弱，但总体而言还是值得一看的佳作。

## 2.2 机器翻译

机器翻译任务是将一种语言的文本自动翻译成另一种语言的过程。这类任务通常分为统计机器翻译（基于概率模型）和神经机器翻译（基于深度学习技术）。

例子：自然语言处理是人工智能领域的一个重要分支，它关注计算机理解、解释和生成人类自然语言的方法。NLP 旨在使计算机能够与人类有效地交流，处理大量文本数据，提取出有意义的信息，从而辅助决策、回答问题等。NLP 技术已广泛应用于众多领域，如搜索引擎、聊天机器人、机器翻译、文本摘要、情感分析等。

## 2.3 文本生成

文本生成是自然语言处理（NLP）领域的一个核心任务，指的是利用计算机和算法自动创建符合语法规则、具有意义且连贯的自然语言文本。文本生成可以应用于各种场景，如新闻撰写、社交媒体内容创建、自动邮件回复、聊天机器人对话等。

例子 1：帮我写一个请假条，我因为新冠不舒服，需要请假 3 天，请领导批准

例子 2：写一个诗歌，关于冬天

### 2.4 摘要生成

摘要生成任务是从原始文本中提取关键信息，生成简洁明了的摘要。摘要可以分为抽取式摘要（从原文中直接选取关键句子）和生成式摘要（生成全新的描述）。

例子：根据今天发布的官方报告，本月初中国东北部的洪涝灾害已造成至少50人死亡，23人失踪。受灾地区包括辽宁、吉林和黑龙江三个省份。数以千计的房屋被毁，约34万人被迫撤离家园。为抗击洪水，当局动用了大量的救援物资和人力资源，同时也呼吁民众提供支持。政府承诺将投入更多资金重建受灾地区，加快修复基础设施，恢复正常生产和生活秩序。

## 3. NLP 大模型性能对比

相比 BERT 类的“小模型”，大模型在文本生成，文本理解方面能展现出更多惊喜的效果。大模型的架构上可以分成 Encoder-Decoder 和 Decoder-only 两种类型。虽然随着 GPT-3 的大火，但这也不表明大模型只能是 Decoder-only 类型。

### Encoder-Decoder

Encoder-Decoder 大模型是一种在自然语言处理（NLP）以及其他领域广泛应用的深度学习架构。这种模型包括两个主要部分：编码器（encoder）和解码器（decoder）。编码器负责理解输入序列，将其转换为固定长度的向量表示；解码器则基于该表示生成目标序列。常见的 Encoder-Decoder 模型采用 Transformer 架构，以实现有效的长距离依赖关系捕获和并行计算。

输入表示：输入文本被分割成一系列标记（token），然后通过词嵌入（word embeddings）转换为向量表示。为了保留序列中的位置信息，会添加位置编码（positional encoding）到词嵌入中。

编码器：编码器通常由多层堆叠组成，每层都包含一个或多个子层，例如自注意力层、前馈神经网络层、层归一化（Layer Normalization）等。编码器通过处理输入序列，捕获不同词汇之间的关系，并将整个序列压缩成一个固定长度的向量表示。

解码器：解码器也有多层堆叠结构，与编码器类似地包含自注意力层、前馈神经网络层等。解码器的自注意力层关注目标序列中的词汇关系，而跨注意力层则关注输入序列和目标序列之间的关系。这样，解码器可以基于编码器的向量表示来生成目标序列。

掩蔽：在训练过程中，为了防止信息泄露（例如，在文本生成任务中提前预测下一个词汇），解码器会使用掩蔽机制。这可以确保模型在生成目标序列时仅关注当前位置及其之前的标记。

输出层：最后，模型通过线性层和激活函数（如 softmax）将解码器的隐藏状态转换为与预定义类别相对应的概率分布。然后可以通过选取具有最高概率的类别作为预测结果或者采用搜索策略（如贪心搜索、集束搜索等）来生成文本序列。

Encoder-Decoder 模型在各种 NLP 任务中都取得了显著的成功，例如机器翻译、文本摘要、问答系统等。一个典型的例子是谷歌的 BERT（Bidirectional Encoder Representations from Transformers），它在大规模语料库上进行双向预训练，可有效改进各种下游任务的性能。

## Decoder-only

Decoder-only 大模型是一种主要应用于自然语言处理（NLP）任务的深度学习架构。这种模型仅包括解码器（decoder）部分，而不像其他常见架构（如编码器-解码器（encoder-decoder）结构）那样具有编码器（encoder）部分。Decoder-only 大模型通常采用 Transformer 架构，以有效处理长距离依赖关系和并行计算。

输入表示：输入文本首先被拆分成一系列标记（token），然后通过词嵌入（word embeddings）转换为向量表示。为了保留序列中的位置信息，会添加位置编码（positional encoding）到词嵌入中。

自注意力机制：Decoder-only 模型使用多头自注意力机制（multi-head self-attention）来捕获不同词汇之间的关系。这使得模型可以同时关注序列中所有位置的信息，而不受距离的限制。

层次化堆叠：模型通常具有多层（layer），每层都包含一个或多个子层，例如自注意力层、前馈神经网络层、层归一化（Layer Normalization）等。通过堆叠这些层，模型可以学习到更复杂和抽象的特征。

掩蔽：在训练过程中，为了避免信息泄露（例如，在文本生成任务中提前预测下一个词汇），解码器使用掩蔽机制。这样可以确保模型只关注当前位置及其之前的标记。

输出层：最后，模型通过线性层和激活函数（如 softmax）将隐藏状态转换为与预定义类别相对应的概率分布。然后可以通过取最高概率的类别作为预测结果或者采用搜索策略（如贪心搜索、集束搜索等）来生成文本序列。

Decoder-only 大模型的一个经典代表是 OpenAI 的 GPT（Generative Pre-trained Transformer）系列。通过在大规模语料库上进行无监督预训练，GPT 模型能够学习到丰富的语言知识，进而在各种下游 NLP 任务中取得卓越性能。

下面我们分别展示一下不同模型之间的效果和差异

## 3.1 ChatYuan-large-v1

ChatYuan 模型为 Encoder-Decoder 大模型，使用了中英文的语料进行训练，拥有 783M 个参数，生成的效果如下所示。

### 3.1.1 情感分析

褒义；中性；褒义

### 3.1.2 机器翻译

*Natural language processing is an important branch of artificial intelligence, which aim to understand, interpret and produce methods of computer interacting with and help computers interest in humans, procure a lot of text data, taking meaningful information to assist with decision making and answer questions. NLP technology has been applied in many fields such as search engine, chat robot, machine translation, text summary, affective analysis and others.*

### 3.1.3 文本生成

（1）请假条

*尊敬的领导：*

*您好!*

*因最近新冠肺炎疫情，我感到不舒服，所以请求请假 3 天，时间定于 2 月 13 日至 2 月 16 日。*

*我将尽快恢复健康，争取早日重返工作岗位，以更好地为公司服务。*

*谢谢您的批准!*

*_此致*

*敬礼!*

*XXX*

（2）诗歌

*寒冬漫天雪花飘，*

*凝成一捧冰霜。*

*冰晶映照着晨曦，*

*融化在大地间，*

*让爱充满每个角落。*

*清晨雪白的空气，*

*唤醒沉睡的记忆，*

*雪花漫天飘飞，*

*一片一片将大地包裹，*

*把爱装进心里，*

*温暖每个角落，*

*让我们在寒冷中感受爱与快乐。*

#### 3.1.4 文本摘要

*官方报告：中国东北部洪灾已致 50 人死亡*

## 3.2 Langboat/bloom-1b4-zh

该模型为 Decoder-only 结构，基于 bigscience/bloom-1b7 对词表进行了裁剪，保留中文部分。参数量 1.4B。

#### 3.2.1 情感分析

褒义的、消极的、褒义的

#### 3.2.2 机器翻译

翻译失败

#### 3.2.3 文本生成

（1）请假条

*因本人身体不适，需要请假 3 天，请领导批准。*

（2）诗歌

*雪，是冬天里最美的花朵，是冬天里最美的风景。*

#### 3.2.4 文本摘要

*洪涝灾害是自然灾害中的一种，它给人们的生活带来了很大的影响，所以我们要做好预防工作，减少洪涝灾害的发生。下面是小编为大家整理的预防洪涝灾害的方法，欢迎大家阅读。预防洪涝灾害的方法 1、加强防洪排涝工作，做好防洪排涝工作*

### 3.3 IDEA-CCNL/Wenzhong2.0-GPT2-3.5B-chinese

该模型为 Decoder-only 结构，基于悟道数据集预训练的中文版的 GPT2，参数量 3.5B

#### 3.3.1 情感分析

褒义的、消极的、褒义的

#### 3.3.2 机器翻译

失败

#### 3.3.3 文本生成

（1）请假条

*尊敬的领导： 我因为新冠不舒服，需要请假 3 天，请假时间为：2018 年 1 月 1 日-2018 年 1 月 31 日。*

（2）诗歌

*冬天，是一个美丽的季节，是一个温暖的季节。*

#### 3.3.4 文本摘要

模型太大生成失败

### 3.4 ChatGPT-4

#### 3.4.1 情感分析

褒义的、消极的、褒义的

#### 3.4.2 机器翻译

*Natural Language Processing (NLP) is an important branch of artificial*

*intelligence that focuses on methods for computers to understand, interpret, and generate human natural language. The goal of NLP is to enable computers to effectively communicate with humans, process large amounts of text data, extract meaningful information, and assist in decision-making, answering questions, and more. NLP techniques have been widely applied in various fields, such as search engines, chatbots, machine translation, text summarization, sentiment analysis, etc.*

### 3.4.3 文本生成

（1）请假条

*尊敬的领导：*

*您好！很抱歉地通知您，我最近出现了一些新冠病毒相关的不适症状，为了确保团队及公司整体的安全和健康，遵从医生建议，我需要休息并进行隔离观察。因此，我申请请假 3 天（从即日起算）以便养病、恢复，并根据情况接受进一步诊断与治疗。*

*在休假期间，我会尽量保持手机畅通，如有紧急工作需求，请随时联系我。同时，我已将待处理的工作事项交接给我的同事，确保工作顺利进行。*

*感谢您对我的理解与支持。我会尽快康复并回到岗位，继续为团队贡献力量。希望您能批准我的请假申请。*

*再次感谢！*

*顺祝商祺，*

*[您的名字]*

*[日期]*

（2）诗歌

*雪花舞动的冬日晨，*

*白茫茫的世界显得宁静。*

*寒风轻拂过窗棂，*

*岁月如歌，冬天温馨。*

*银装素裹的大地闹，*

*孩童们欢笑在雪堆旁。*

*滑冰者翩翩起舞，*

*构筑梦想，抵挡严寒。*

*鸟儿啼鸣离别南方，*

*枯枝留待春暖花开。*

*炉火照耀着家庭团圆，*

*分享快乐，温暖人心。*

*冬日寒夜星光璀璨，*

*镶嵌在这银色幕布上。*

*我们依偎在爱的怀抱里，*

*期盼春暖，又迎冬天。*

#### 3.4.4 文本摘要

*本月初中国东北部洪涝灾害已导致至少 50 人死亡，23 人失踪。受灾地区包括辽宁、吉林和黑龙江三个省份，数千房屋被毁，约 34 万人被迫撤离。政府承诺投入更多资金重建受灾地区，并加速修复基础设施。*

## 4. 性能对比分析

根据生成的结果，我们从 1-5 分进行打分，以最高分 5 分表示真实情景下做到的效果。

| | 情感分析 | 机器翻译 | 文本生成 | 摘要生成 |
|---|---|---|---|---|
| ChatYuan-large-v1 | 1 | 4 | 5 | 2 |
| bloom-1b4-zh | 4 | 翻译失败 | 1 | 2 |
| GPT2-3.5B-chinese | 4 | 翻译失败 | 2 | 生成失败 |
| ChatGPT-4 | 5 | 5 | 5 | 5 |

在情感分析中 Decoder-only 的三个模型都很好的分辨了情感，而 Encoder-Decoder 的 ChatYuan 模型在进行消极语句的分析时出现了一些问题。

由于 bloom-1b4-zh 以及 GPT2-3.5B-chinese 两个模型可能未加入英文预料进行训练，所以并未成功翻译；相比 ChatYuan 模型，ChatGPT-4 模型翻译的结果显

得更加流畅自然。

在文本生成方面，ChatYuan 以及 ChatGPT-4 模型生成的结果远远胜过其他两个模型的结果。通过结果对比，ChatYuan 以及 ChatGPT-4 模型能很好地理解并扩展收集到的信息，从语料库中生成符合逻辑的文本；而 bloom-1b4-zh 以及 GPT2-3.5B-chinese 虽然在模型的参数量上远超 ChatYuan，但是就结果来看，这两个模型虽然能较好的理解文本需求，却不能很好的生成结果，这可能是由于模型内部对于 prompt 的设置比较简单直接，使得生成的结果虽然满足基本要求，却无法达到满意的水平。

## 5. 结论

本报告对比了前沿 NLP 大模型在各种下游任务的性能。由于参数量的不同很难直接评判某种方法是否更好，不过相对而言，在同等参数量的情况下，Encoder-Decoder 模型相比 Decoder-Only 模型能取得更好的效果，不过对于参数判断的精确度有较高的要求。而 Decoder-Only 模型对于语言风格的理解上更加准确，可能相比于编码解码的复杂结构，单次解码更不容易导致内部信息的丢失。

总体而言，ChatYuan 以及 ChatGPT-4 模型在大多数任务上具有较高的准确率。然而，在实际应用时，还需考虑因素如计算资源、微调数据量等，并根据项目需求选择合适的模型。