# 大模型的知识分析与增强

## 1. 知识分析概述

预训练语言模型学到大量知识，被认为是“神经网络知识库”（LM as KB）。然而，这些知识的学习效果和原理不透明，因此需要对预训练语言模型进行知识分析。

### 研究背景

- **预训练语言模型**：指通过大规模语料进行自监督学习的语言模型，如BERT、GPT-3。这些模型在训练过程中从语料中学习了大量语言和世界知识。
- **神经网络知识库**：由于预训练语言模型掌握了大量知识，研究人员将其比作知识库。
- **透明性问题**：尽管模型表现出很强的知识检索和推理能力，但其内部知识表示和存储机制对用户来说是一个黑箱。

### 研究内容

1. **知识探测**：探测预训练语言模型掌握的知识，了解模型内存储的知识范围和准确性。
2. 知识定位：分析预训练语言模型中的知识存储机制，确定知识存储在模型的哪些层和神经元中。
    - **层粒度**：分析不同层次的知识存储情况。
    - **神经元粒度**：进一步细化到具体的神经元层面。
3. **知识学习机理分析**：分析影响预训练语言模型学习效果的因素，探讨模型如何从数据中学习和存储知识。

## 2. 知识探测

通过自然语言句子和提示，探测模型掌握的世界知识。LAMA项目通过将三元组或问答对形式的知识转化为填空形式进行探测。

### 世界知识探测

- **自然语言提示**：使用自然语言提示进行探测，无需重新训练模型。例如，使用BERT进行填空任务：“Dnate was born in ”，模型填补“Florence”。
- **LAMA项目**：LAMA（Language Model Analysis）项目将三元组或问答对形式的知识转化为填空形式，对模型进行探测和评估。

### 数据来源

1. **GoogleRE语料**：从Wikipedia抽取的三元组，关注“出生地”、“出生日期”、“死亡地”关系。
2. **T-Rex**：wikidata的一个子集，关注41种关系，每种关系最多保留1000个事实。
3. **ConceptNet**：常识知识库，实验关注16种单token关系，从OMCS中直接取句子。
4. **SQuAD**：常用QA数据集，选取305个上下文无关且答案为单token的问答对。

### 主要发现

- **BERT-large**：在一对一关系上表现相对较好，证明了预训练语言模型中蕴含着世界知识。
- **衍生研究**：研究prompt选择对探测结果的影响，例如AutoPrompt通过梯度搜索自动生成提示语，提高探测效果。

**严谨性分析**

- **探针实验条件**：实验发现，模型表现受各种偏差影响，包括提示选择、示例性样例和上下文推理。
- **质疑结论**：模型处理世界知识的机制和人们预期可能存在很大差别，实验设定需要更严谨。

## 3. 知识定位

研究如何确定语言模型中"知识"的存储位置和存取机制。

**知识存储机制**

- **Transformer前馈网络（FFN）**：被认为是存储大量具体知识的Key-Value存储器。
- **Logit Lens**：通过查看模型的隐藏状态，发现FFN模块存储了大量知识。

**主要方法**

1. **基于梯度归因**：提出知识归因的方法，将一个事实三元组定位到特定神经元中，如Dai等人的研究发现特定神经元与知识存在关联。
2. **基于因果分析**：通过因果分析，发现中间层的MLP对预测结果有决定性影响。

**实验发现**

- **FFN作为键值存储器**：FFN的输入层、宽隐层和窄隐层分别存储key向量和value向量，形成知识存储结构。
- **层级结构**：中间层的MLP存储了事实性知识，模型输出结果是各层结果的迭代精炼。

## 4. 长尾知识学习

探究预训练数据如何影响大模型中的知识，发现LLM回答问题的能力与预训练数据频率有较强相关性。

**主要结论**

- **频率相关性**：LLM回答问题的能力与预训练语料中相关信息文档的数目有较强的相关性。
- **知识移除实验**：移除预训练语料中相关信息会导致模型在相关问题上回答准确率下降。
- **解决方法**：扩大模型规模和检索增强可以提高模型对长尾知识的学习能力。

## 5. 知识萃取

从预训练语言模型中诱导出有用的显式符号化知识。

**传统小模型的知识萃取**

- **COMET**：将常识知识图谱ATOMIC作为训练数据，微调GPT，通过常识知识注入和激发，得到常识模型COMET。该模型能够扩展常识知识图谱，并生成新的常识知识。

**大模型萃取常识知识**

- **ATOMIC10X**：提出符号知识蒸馏框架，将常识知识从大模型（如GPT-3）转移到小模型，通过自动蒸馏生成的知识图谱在数量、质量和多样性方面超过了人类编写的知识图谱。

**非英语语言的知识萃取**

- **CN-AutoMIC**：从中等规模多语大模型（MT5）中提取中文常识知识，通过优化生成和过滤方法，在生成质量较差的条件下获取大规模高质量常识知识。

## 6. 幻觉现象与缓解

语言模型存在生成错误内容的幻觉问题。

**幻觉现象**

- **事实性幻觉**：生成的内容不忠实于既定的事实知识，可能误导用户。
- **忠实性幻觉**：生成的内容与之前生成的信息不一致或与用户提供的输入相冲突，影响用户体验和信任度。

**幻觉的来源**

- **知识缺乏或错误**：大模型可能缺乏相关知识或内化错误知识。
- **相关性误解**：模型将相关性误解为事实知识。
- **训练数据问题**：存在模仿性、重复性、社会偏见等问题。
- **自信过高**：大模型在生成错误答案时与生成正确答案时同样自信，缺乏说“不知道”的能力。

**缓解方法**

1. **训练数据清洗**：清洗预训练数据，避免内化错误知识，但需要重新预训练模型。
2. **优化解码方式**：减少解码过程的随机性，可能损害模型的创造力。
3. **改进指令数据**：提高模型拒绝错误回答的能力，可能导致模型过度拒绝。
4. **外部知识增强**：利用外部知识库减少模型幻觉，可能引入知识冲突问题。

## 7. 工具增强

为了弥补大模型的缺陷，如数学计算能力、超长上下文处理能力和获取新知识的能力，通过工具增强模型性能。

**Toolformer模型**

- **研究背景**：大模型难以解决即时信息更新、生成错误幻觉和数学计算等问题。通过调用外部工具可以补充模型相关知识。
- **训练方法**：通过in-context learning生成包含工具调用语句的数据，然后通过过滤得到微调模型需要的训练数据。
- **工具使用**：包括问答系统、计算器、维基百科搜索和翻译系统等。实验表明Toolformer在多任务上性能接近或超过GPT-3水平。

**未来展望**

- **ToolLLM**：构建高质量指令微调数据集（ToolBench），在ToolBench上微调LLaMA模型，解决开源LLM在理解人类指令并与工具（API）交互方面的不足。