Kaggle - LLM Science Exam竞赛银牌算法概览

竞赛概览：

竞赛受OpenBookQA数据集的启发，要求参与者使用大型语言模型（LLM）回答一些科学上的困难问题，从ABCDE五个选项中选出正确答案。通过这项工作，研究者们希望更好地理解LLM测试自身的能力，以及在资源受限环境中运行LLM的潜力。在本次竞赛中，我们的算法主要使用基于RAG检索增强生成架构的自然语言问答模型。

所用算法：

1. 数据提取：分别对样本提示语prompt和外部大型知识语料库进行清洗和整理，prompt重复三次与问题选项拼接构成检索文本（query），将语料库进行整理和区分后形成三种：完整的所有文本（full text）、STEM相关文本片段1、STEM相关文本片段2。
2. 向量化Embedding：分别对检索文本与三种文本语料库使用embedding模型gte-base生成embedding（特征向量）。
3. 创建索引（Index）：使用Faiss对三种文本语料库的embedding创建Index，以便后续查询。
4. 检索（Retrieval）：使用向量内积（Inner Product, IP）相似度进行语义相似搜索，使用query文本embedding逐一在三种文本语料库Index中查询，获取top10文本作为最终检索结果。
5. 生成（Generation）：将基于完整文本（full text）语料检索结果作为训练集文本外部知识库（context），由context（知识库）+ prompt（问题）+ option（选项）构成完整样本。
6. 模型微调（Finetune）：将所得完整样本输入deberta-v3-large模型进行五分类训练，基于下述形式进行RAG和风格行为词汇适配：

五分类形式：外部知识文本{context}，请回答题目{question prompt }，选项{ option [1-5]}中最正确的是ABCDE中的哪一个？

1. 推理答案（Inference）：将在三种语料知识库得到检索文本分别输入微调后的deberta-v3-large模型预测每个选项概率，将预测结果按照加权平均得到最终的前三答案选项。

数据与模型链接：

STEM文本语料库1：https://www.kaggle.com/datasets/mbanaei/all-paraphs-parsed-expanded

STEM 文本语料库2：https://www.kaggle.com/datasets/mbanaei/stem-wiki-cohere-no-emb

Wikipedia完整语料库：https://www.kaggle.com/datasets/jjinho/wikipedia-20230701

Gte-base: https://huggingface.co/thenlper/gte-base

Faiss: https://github.com/facebookresearch/faiss

deberta-v3-large: https://huggingface.co/microsoft/deberta-v3-large

样本：https://www.kaggle.com/datasets/cdeotte/60k-data-with-context-v2

https://www.kaggle.com/datasets/cdeotte/40k-data-with-context-v2

https://www.kaggle.com/datasets/cdeotte/99k-data-with-context-v2

Silver Medal Algorithm Overview for

**Kaggle - LLM Science Exam** Competition

**Competition Overview:**

The competition, inspired by the OpenBookQA dataset, challenges participants to utilize Large Language Models (LLMs) to answer complex scientific questions, selecting the correct answer from options A, B, C, D, and E. Through this endeavor, researchers aim to gain a deeper understanding of the LLMs' capability to test themselves and the potential of deploying LLMs in resource-constrained environments. In this competition, our algorithm primarily employs a natural language question-answering model based on the Retrieval-Augmented Generation (RAG) architecture.

**Algorithm Descriptions:**

1. **Data Extraction**: Clean and organize the sample prompt and an external large knowledge corpus separately. The prompt, repeated three times, is concatenated with question options to form the retrieval text (query). The corpus is organized and divided into three types: full text, STEM-related text snippet 1, and STEM-related text snippet 2.
2. **Vectorization (Embedding)**: Embedding models (gte-base) are used to generate embeddings (feature vectors) for the retrieval text and the three text corpora separately.
3. **Index Creation:** Faiss is used to create an Index for the embeddings of the three text corpora to facilitate subsequent queries.
4. **Retrieval**: Semantic similarity searches are conducted using Inner Product (IP) similarity, querying each text corpus Index with the query text embedding and obtaining the top 10 texts as the final retrieval results.
5. **Generation**: The retrieval results based on the full text corpus serve as the training set text external knowledge base (context). The complete sample is formed by context (knowledge base) + prompt (question) + option.
6. **Model Fine-tuning**: The complete sample is input into the deberta-v3-large model for five-category training, with RAG and style-behavior vocabulary adaptation based on the following format:

Five-category format: External knowledge text {context}, please answer the question {question prompt}, which one is the most correct in the options {option [1-5]} among A, B, C, D, E?

1. **Answer Inference**: The retrieved texts from the three knowledge corpus are input into the fine-tuned deberta-v3-large model to predict the probability of each option. The prediction results are weighted averaged to obtain the final top three answer options.

**Data & Model Links:**

STEM Text Corpus 1：https://www.kaggle.com/datasets/mbanaei/all-paraphs-parsed-expanded

STEM Text Corpus 2：https://www.kaggle.com/datasets/mbanaei/stem-wiki-cohere-no-emb

Wikipedia Complete Corpus：https://www.kaggle.com/datasets/jjinho/wikipedia-20230701

Gte-base: https://huggingface.co/thenlper/gte-base

Faiss: https://github.com/facebookresearch/faiss

deberta-v3-large: https://huggingface.co/microsoft/deberta-v3-large

Samples：https://www.kaggle.com/datasets/cdeotte/60k-data-with-context-v2

https://www.kaggle.com/datasets/cdeotte/40k-data-with-context-v2

https://www.kaggle.com/datasets/cdeotte/99k-data-with-context-v2