Enhancing Reasoning Path Diversity in Commonsense Question Answering

Dataset Name: CommonsenseQA (CSQA)

Dataset: `https://paperswithcode.com/dataset/commonsenseqa`

Paper Title:
Self-Consistency Improves Chain of Thought Reasoning in Language Models

Paper: `https://arxiv.org/pdf/2203.11171v4`

Code

Code Repository: `https://github.com/codelion/optillm/blob/main/optillm/self_consistency.py`

Input and Output

Input: Data, pretrained language models (like BERT, T5).
Output: Comparison of the self-consistency method with greedy decoding and other baseline methods (e.g., beam search) to highlight its advantages.

Metrics

Provide metrics like:

- Number of unique reasoning paths.
- Path similarity measures (e.g., cosine similarity).
- Diversity scores (e.g., Shannon entropy) to assess the diversity of generated reasoning paths.

Current Progress

Literature Review

Studied the paper "Self-Consistency Improves Chain of Thought Reasoning in Language Models" to understand the self-consistency method and its advantages over greedy decoding and beam search.
Key insights:

- Self-consistency improves reasoning diversity by generating multiple reasoning paths and selecting the most consistent answer.
- It enhances performance on commonsense reasoning tasks like CommonsenseQA.

Dataset Preparation

- Downloaded and preprocessed the CommonsenseQA dataset.
- Verified the dataset's structure and ensured compatibility with the OptiLLM framework.

Codebase Exploration

- Explored the OptiLLM codebase, focusing on the `self_consistency.py` module.
- Identified key functions for reasoning path generation and evaluation.
- Reviewed the implementation of metrics like reasoning path diversity and similarity.

Initial Experiments

- Ran baseline experiments using greedy decoding and beam search.
- Observed limitations in reasoning diversity and accuracy.
- Implemented the self-consistency method and observed improvements in reasoning diversity.

Observations

- The self-consistency method generates more diverse reasoning paths compared to greedy decoding and beam search.
- Preliminary results indicate higher accuracy and reasoning diversity on the CommonsenseQA dataset.
- Metrics like Shannon entropy and cosine similarity effectively quantify reasoning path diversity.

Plan of Action

Short-Term Goals

- **Complete Implementation:**
  - Finalize the implementation of the self-consistency method in the OptiLLM framework.
  - Integrate additional metrics for reasoning path evaluation.
- **Run Experiments:**
  - Conduct experiments on the CommonsenseQA dataset using self-consistency, greedy decoding, and beam search.
  - Collect and analyze results for reasoning path diversity and accuracy.
- **Documentation:**
  - Document the implementation details and experimental results.
  - Prepare visualizations for metrics like reasoning path diversity and similarity.

Long-Term Goals

- **Optimize the Self-Consistency Method:**
  - Explore task-specific customizations to prompts for further improvements.
  - Experiment with different pretrained language models (e.g., GPT, T5).
- **Benchmarking:**
  - Compare the self-consistency method with state-of-the-art approaches on commonsense reasoning tasks.
- **Publish Results:**
  - Compile the findings into a comprehensive report or paper.
  - Share the results with the research community.

References

1. **Paper:** `https://arxiv.org/pdf/2203.11171v4`
2. **Dataset:** `https://paperswithcode.com/dataset/commonsenseqa`
3. **Codebase:** `https://github.com/codelion/optillm/blob/main/optillm/self_consistency.py`