6.7960: Final Project Proposition

Due on Friday, November 14, 2024

Isola and Beery

Gatlen Culp and Adriano Hernandez
Collaborators: ChatGPT o1-preview

Resources Used: None

Introduction

There is debate about whether Large Language Models (LLMs) learn similar semantic representations despite differences in architecture and training data. There has been previous research into similarity between representations on other models, but so far as I can find, nothing of the sorts have been conducted on the embeddings of Large Language Models. Evidence suggests that similar circuits can be found across different models. Certain

In the broader context of deep learning, questions arise regarding the extent to which LLMs learn approximately the same semantic concepts and whether their representations could be disentangled to an underlying "platonic model."

This project aims to investigate the degree to which different LLMs share similar semantic representations by training a translation model capable of transforming the embedding (produced from an identical string) from one model into one of another and measuring the "translation loss" between the real and translated embeddings. This project was directly inspired by an issue I had encountered during my summer project building a data visualization and analysis tool that used LLM embeddings and would have benefitted by interoperability between cheap/old and expensive/new models.

Motivation

- Determine the extent to which different LLMs share semantic representations.
- Identify domains where translation loss is minimized or maximized.
- Assess the potential for creating a standardized embedding space for interoperability.
- Provide insights contributing to the development of interoperable AI systems and efficient embedding reuse.

Project Plan and Timeline

- **Week 1 - Setup**:
  - Conduct a literature review on semantic representations in LLMs.
  - Finalize the selection of models and datasets.
  - Set up computational environment and resources.
  - Begin assembling diverse dataset covering various domains.
  - Plan data cleaning and normalization approach.
- **Week 2 - Data and Training**:
  - Complete data collection and preprocessing pipelines
  - Generate embeddings using GPT-4, Claude 3, and LLaMa 3.
  - Develop and train translation networks (feed-forward neural networks).
- **Week 3 - Analysis**:
  - Evaluate translation loss using MSE.
  - Compute cosine similarity and other statistical measures.
  - Apply dimensionality reduction (t-SNE/UMAP) for visualization.
  - Analyze translation loss patterns across different domains.
  - Start blog rough draft.
- **Week 4 - Presentation**:

- Perform statistical significance testing.
- If time permits:
  * Investigate smaller/fine-tuned models
  * Explore translation between NLP and non-text modalities
- Draft the final research blog post with visualizations.

Conclusion

This project will shed light on the semantic representations learned by different LLMs and assess the feasibility of translating embeddings across models. The findings could promote more efficient use of computational resources through embedding reuse and contribute to the understanding of universal representations in LLMs.