Exploring Semantic Similarity: A Comparative Study of Synonym Vector Embeddings in GPT-2 and Human Cognition

Catherine Zhang (yz5609) Tianyu Wu (tw1802) Xintong Xie (xx964) Zhiyu Guo (zg915)

April 11, 2024

1 Research Question

We aim to investigate vector embeddings and representations of synonyms in Large Language Models (LLMs), particularly exploring whether GPT-2 understands word meanings similarly to humans. Our study will focus on comparing the similarity between vector embeddings of synonym pairs in GPT-2, contrasting these with human perceptions of the same pairs.

2 Methodology

Our approach involves inputting pairs of synonymous nouns into a pre-trained GPT-2 model and extracting their vector representations across various layers. We will use cosine similarity to assess the relationships between these embeddings at different layers. The study will particularly compare synonyms with nearly identical meanings and synonyms where one word might have multiple meanings.

We also plan to collect human ratings on the similarity of these synonyms (probably a smaller dataset) and compare these with the GPT-2 outputs. As well as feeding the human questionnaire to GPT-3.5 and GPT-4 for further comparison. Time permitting, we will also explore established word embedding libraries like FastText and GloVe for further comparisons.

3 Data Source

The primary resources for this study will be the pre-trained GPT-2 model from OpenAI and the WordNet database, which provides a robust list of synonym pairs for analysis.

4 Concerns & Doubts

1. For human results on synonym similarity, we could potentially get only very few results (less than 10) due to the time and scope constrain, would such a small sample be a problem?
2. Is there a way to compare GPT-2 with GPT-3.5 and GPT-4 to see the vector embedding differences?

References

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*. `https://arxiv.org/abs/1909.00512`

Princeton University. (2010). About WordNet. *WordNet. Princeton University*. `https://wordnet.princeton.edu/`

Radford, A., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1.8*, 9. `https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf`