CONCEPT UNDERSTANDING IN LARGE LANGUAGE MODELS: AN EMPIRICAL STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale Language Models (LLMs) like GPT have demonstrated their superior comprehension and expressiveness across a wide range of tasks, and exhibited remarkable capabilities in real-world applications such as Copilot and New Bing. Hence, it is crucial to investigate LLM's potential and limitations for trustworthy performance in both academia and industry. In this paper, we focus on exploring LLMs' ability in understanding concepts, especially for the abstract and the concrete. To this end, we construct a dataset D-Concept based on WordNet for a classic NLP task, hypernym discovery. The dataset consists of two subsets for abstract entities and concrete entities, respectively. We use pre-trained LLMs (from BERT to ChatGPT) to conduct hypernym discovery for these two kinds of concepts. The experimental results suggest that the large language model's understanding of abstract concepts is significantly weaker than that of concrete concepts.

1 INTRODUCTION

In the past few years, large-scale language models (LLMs) have become the frontiers of research in academics. The recently released ChatGPT has further exhibited the potential of large language models in various downstream tasks with advanced technologies and engineering efforts (Brown et al., 2020). On traditional natural language understanding tasks, state-of-the-art results are frequently being updated. On the other hand, the language understanding ability of LLMs also affects their deployments in real-world scenarios, making exploration of this research problem a new topic.

This paper focuses on the understanding ability of large language models from the perspective of abstract and concrete concepts. Such explorations are inspired by the fact that different real-world tasks usually require the understanding ability on different levels of abstraction. A better comprehension of concrete concepts may be desirable in tasks that are related to physical entities like classifying the type of animals, while other tasks ask the LLMs to know more about abstract entities such as distinguishing different human emotions or logical reasoning.

To achieve this goal, we first construct a new dataset D-Concept based on WordNet, a lexical database of semantic relations between words (Miller, 1995). The dataset follows the setting of a classic NLP task of hypernym discovery, and it can reflect the language model's ability in understanding abstract and concrete concepts. The noun branch of WordNet is explicitly divided into abstract and physical categories, which is an appropriate prior. Therefore, we follow this division to construct two sub-dataset on the abstract and physical branches, respectively, with each data record consisting of a pair of words from the corresponding branches. The task is to determine whether these two words are hypernyms or not. Note that WordNet is a tree-like structure, so different ancestors of a single hyponym are its hypernyms. The explored large language models range from BERT to the GPT series models (including GPT-3.5 and ChatGPT). Experimental results show that, in the hypernym discovery task, the language model's understanding ability of words from the abstract branch is significantly weaker than that of words from the physical branch.

Our contributions can be summarized as follows: 1. We construct a new dataset for the hypernym discovery task, which can be used to explore the large language model's ability between abstract and concrete entities. 2. We investigate the performance of LLMs of different scales on this task and find that the performance improves as the model size increases, but consistently performs worse on abstract tasks than on concrete ones.

2 EXPERIMENT

2.1 EXPERIMENT SETUP

WordNet, a large lexical database of English, groups nouns into sets of cognitive synonyms (synsets), each expressing a distinct concept and can be regarded as an entity. Hyperonymy relation ("bed"-"furniture") is one typical lexical relation on WordNet due to the nature that WordNet is a hierarchical graph with entities as nodes and hyperonymy relations as edges.

Abstract branch and physical branch are two main components in WordNet representing abstract concepts and concrete concepts, respectively. To explore how LLMs represent concepts, we visualize embeddings of concepts in WordNet generated by GPT-3.5 with T-SNE. As shown in Figure 1, embeddings of two types of concepts are roughly divided into two clusters, revealing that there indeed exist some differences between them. To dive deeper into such differences, we create a new concept understanding-related dataset based on the hypernym discovery task.

When constructing the dataset, on each branch, firstly we randomly sample a distance value d for positive examples (i.e., one entity and its corresponding hypernym). Secondly, we randomly sample the first entity and then the second entity is selected from an entity set where each entity is at a distance d from the first entity. Negative entity pairs are randomly and separately chosen. The dataset is split as Train:Valid:Test=2:4:4.

We compare 5 LLMs (BERT (Devlin et al., 2018), T5 (Raffel et al., 2020), CLIP (Radford et al., 2021), OpenClip (Cherti et al., 2022), GPT-3.5 (Brown et al., 2020)). We freeze their embeddings and add a linear layer to conduct binary classification. Cross-Entropy Loss is used as the loss function and we choose Accuracy (ACC), AUC and F1 score as evaluation metrics. For ChatGPT, we use the prompt-answering paradigm to obtain its answers (i.e., we use "Is Word1 the hypernym of Word2?" as a prompt to get "yes" or "no" from ChatGPT).

2.2 EXPERIMENTAL RESULT

Table 1 shows LLMs' performance on different types of concepts under the hypernymy task. We can find that LLMs consistently perform worse for abstract concepts compared with physical concepts. It implies that there is improvement room for LLMs in learning embeddings especially for abstract concepts. The poor results on abstract concepts may raise the risk of handling abstract concept-related tasks when applying LLMs. We can also discover that when the model scale increases, the performance on both tasks also improves except ChatGPT, where the embedding can not be obtained and is tested without fine-tuning or few-shot learning.

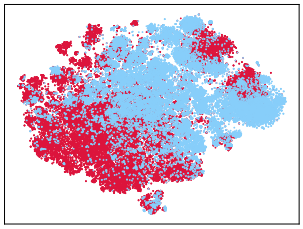

Figure 1: Embedding visualization for abstract (red) and concrete concepts (blue).

Table 1: Hypernym discovery result.

	Abstract			Concrete		
Model	ACC	AUC	F1	ACC	AUC	F1
BERT	85.28	85.28	84.35	89.65	89.61	89.12
T5	85.85	85.85	85.43	89.45	89.41	88.88
CLIP	87.93	87.93	87.62	91.53	91.49	91.13
OpenClip	88.85	88.85	88.47	91.60	91.56	91.22
GPT-3.5	89.45	89.45	89.04	91.98	91.94	91.60
ChatGPT	64.13	64.14	44.10	78.98	78.75	73.04

3 CONCLUSION

We construct a new dataset on the hypernym discovery task based on WordNet in order to explore the understanding ability of large language models on abstract and concrete entities. Experimental results show that LLMs are not good at handling abstract entities no matter the model size, which inspires researchers to mitigate this gap in the future.

URM STATEMENT

We acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.