Semantic Code Search using Hypergraph Neural Networks

Pranjali Jadhav (112103052), Dr. Y. V. Haribhakta (College Guide), Samip Sharma (Industry Guide)

Aim

Exploring Semantic Code Search using Hypergraph Neural Networks to capture higher-order multi-way relationships between functions and concepts.

Objectives

➢Capture **higher-order relationships** between functions and concepts using Hypergraph Neural Networks.
➢Retrieve **top-k relevant code snippets** for a given natural language query.
➢Learn meaningful function and concept embeddings through **contrastive learning** over hyperpaths.

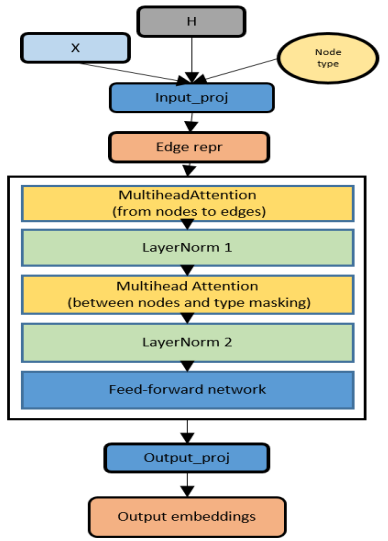
Introduction

- Engineers often struggle to navigate complex tools using traditional documentation.
- Newcomers find it hard to locate code examples that match their intent.
- This project uses Hypergraph Neural Networks for semantic code search to bridge that gap via natural language queries.

Dataset used

- Dataset used for training is the Python subset of CodeSearchNet
- The corpus is a dataset of 2 million (comment, code) pairs from opensource libraries hosted on GitHub
- Out of these, 30000 function-docstring pairs were used for the hypergraph construction.

Model Architecture



Example

Q. How to find synonyms of a word?

```
def findSynonyms(self, word, num):
        """
        Find synonyms of a word

        :param word: a word or a vector representation of word
        :param num: number of synonyms to find
        :return: array of (word, cosineSimilarity)

        .. note:: Local use only
        """
        if not isinstance(word, basestring):
            word = _convert_to_vector(word)
        words, similarity = self.call("findSynonyms", word, num)
        return zip(words, similarity)
```

Results

Metric	HGNN Model	CodeBERT	SBERT	E5-small-v2
NDCG@3	**0.905346**	0.897424	0.894881	0.996800

Table 6.1: NDCG@3 comparison across models

Related work

Title	Publish date	Summary
CodeBERT: A Pre-Trained Model for Programming and Natural Languages	18 September 2020	CodeBERT is a bimodal pre-trained model that uses a masked language modeling (MLM) and replaced token detection (RTD) objective to align code with comments/docstrings.
You are AllSet: A Multiset Learning Framework for Hypergraph Neural Networks	28 March 2022	A framework that models hypergraph propagation using multiset functions rather than traditional adjacency-based methods.

Implementation

•A hypergraph links function nodes (from docstrings) to concept nodes (from TF-IDF keywords) via hyperedges.
•The AllSet Transformer performs attention-based message passing and node aggregation.
•A contrastive loss trains the model to align each function with its concept-based hyperpath.
•At inference, a query is encoded and matched to concept paths to retrieve top-k relevant functions.

Conclusion

•The proposed HGNN-based semantic code search models complex links between functions and conceptual keywords.
•It improves code retrieval, boosts developer productivity, and streamlines API exploration.

References

[1] A. Chien, C. Qian, and C. Shah, ”AllSet: Hypergraph Transformer for Set Representation Learning,” in International Conference on Machine Learning (ICML), 2022.
[2] X. Wang, H. Huang, J. Ye, and E. Xu, ”Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[3] CodeSearchNetChallenge Dataset, GitHub Repository, Available: https://github.com/github/CodeSearchNet.