Neural Similarity Search on Supergraph Containment (Extended abstract)

Hanchen Wang[†‡], Jianke Yu[‡], Xiaoyang Wang[§], Chen Chen[♭✉], Wenjie Zhang[§], Xuemin Lin[♮]
[†]Zhejiang Gongshang University, China, [‡]University of Technology Sydney, Australia,
[§]University of New South Wales, Australia, [♭]University of Wollongong, Australia, [♮]Shanghai Jiao Tong University, China
Hanchen.Wang@uts.edu.au, jianke.yu@student.uts.edu.au,
{xiaoyang.wang1, wenjie.zhang}@unsw.edu.au, chenc@uow.edu.au, xuemin.lin@sjtu.edu.cn

***Abstract*—Supergraph search is a fundamental graph query processing problem. Supergraph search aims to find all data graphs contained in a given query graph based on the subgraph isomorphism. In other words, the goal is to determine if part of the query graph is the same as a smaller data graph. Existing algorithms construct the indices and adopt the *filtering-and-verification* framework, which is usually computationally expensive and can cause redundant computations. Recently, various learning-based methods have been proposed for a good trade-off between accuracy and efficiency for query processing tasks. However, to our knowledge, no learning-based method is proposed for the supergraph search task. In this paper, we propose the first learning-based method for similarity search on supergraph containment, named Neural Supergraph similarity Search (NSS). NSS first learns the representations for query and data graphs and then efficiently conducts the supergraph search on the representation space, the complexity of which is linear to the number of data graphs. The carefully designed Wasserstein discriminator and reconstruction network enable NSS to capture better the interrelation, structural and label information between and within the query and data graphs. Experiments demonstrate that the NSS is up to 6 orders of magnitude faster than the state-of-the-art exact supergraph search algorithm in query processing and is more accurate than the other learning-based solutions.**

***Index Terms*—Supergraph Matching, Graph Neural Network, Graph Reconstruction, Wasserstein Distance**

I. Introduction

In this work, we focus on the supergraph search problem, which aims to determine all the data graphs that are contained in the query graphs as subgraphs. Due to the hard tractability of the supergraph search, the existing algorithms generally construct the indices and conduct the supergraph search based on the indices to obtain the solution [1]. As a result, existing algorithms for this problem have two main limitations: The construction of the indices incurs significant overhead, and the query processing of the existing methods is also costly.

Motivated by the importance of the problem and the expensive time cost of existing exact solutions, we design a learning-based method for supergraph search which could significantly reduce the preprocessing and query processing time for this problem. To be more specific, we propose the first learning-based model for similarity search on supergraph containment, namely Neural Supergraph similarity Search (NSS). NSS generates the representations for query and data graphs utilizing graph neural networks. With the given query, NSS performs the graph containment search in the vector representation space, whose time complexity is linear to the number of data graphs in the dataset. Instead of simply applying the graph neural networks to obtain the approximation results, NSS is carefully designed to improve the qualities of learned representations by preserving the graph properties and the interrelations between the query and data vertices. Specifically, we design a reconstruction network that generates vectors that are as similar as possible to the initial features based on the properties (*e.g.,* labels) of vertices. The inputs of the reconstruction network are the representations obtained by GNNs. Therefore, NSS has a better property-preserving ability in representation learning. The Wasserstein discriminator [2] is also incorporated in our model. The discriminator minimizes the Wasserstein distance between query and data graphs, which implicitly induces an alignment of vertices from two graphs. So NSS enjoys high accuracy without additional time cost.

Contributions. The contributions of this paper are summarized as follows: (1) To the best of our knowledge, our proposed NSS is the first learning-based model that solves the supergraph containment problem with similarity search. (2) With the careful design and adaption of machine learning techniques and graph neural networks, NSS could significantly reduce the time cost of the supergraph search while achieving high approximate accuracy. (3) Extensive experiments are conducted on five real-life datasets. The results indicate that NSS is 6 orders of magnitude faster compared to the state-of-the-art method IDAR in terms of query processing time.

II. Preliminaries

The data graph is denoted as $g = (V(g), E(g))$, where $V(g)$ is the set of vertices and $E(g)$ is the edges in the data graph. The set of data graphs is denoted as D. Each vertex is mapped to the label by the label mapping function L. The query graph, denoted by Q, is a potential supergraph of data graph g.

Definition II.1 (Subgraph Isomorphism)**.** Given a query graph $Q = (V, E)$ and a data graph $g = (V', E')$, a subgraph isomorphism is an injective function f_{iso} from V to V' such that (1) $\forall v \in V, L(v) = L(f_{iso}(v))$; and (2) $\forall e_{(u,v)} \in E, e_{(f_{iso}(u), f_{iso}(v))} \in E'$.

Problem Statement. Given a query graph Q and a set D of data graphs, the supergraph search problem is to find all data graphs in D that are subgraphs of Q based on the subgraph

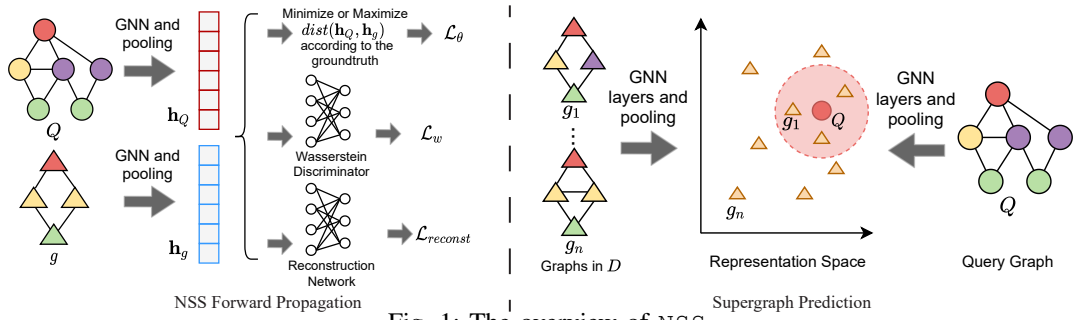


Fig. 1: The overview of NSS

TABLE I: Accuracy Evaluation Results

	NCI				PubChem				FDA				Wiki-Vote			
	Acc	**F1**	**Pre**	**Recall**	**Acc**	**F1**	**Pre**	**Recall**	**Acc**	**F1**	**Pre**	**Recall**	**Acc**	**F1**	**Pre**	**Recall**
NN-Baseline	77.70%	64.70%	91.80%	49.95%	72.63%	61.93%	85.80%	48.45%	78.30%	66.41%	77.76%	57.95%	79.48%	57.14%	74.34%	46.41%
GNN-Baseline	84.56%	78.98%	89.15%	70.90%	83.63%	80.79%	87.67%	74.91%	84.11%	76.37%	84.98%	69.35%	82.43%	65.72%	77.35%	57.13%
NSS w/o WD	86.51%	81.76%	91.48%	73.91%	84.42%	81.73%	88.62%	75.83%	86.88%	81.44%	85.47%	77.78%	85.21%	78.61%	68.50%	92.23%
NSS w/o reconst	86.39%	81.63%	91.15%	73.91%	84.89%	82.12%	90.03%	75.49%	86.84%	81.48%	85.09%	78.16%	85.14%	77.78%	69.55%	88.22%
NSS	**90.71%**	**87.83%**	**94.73%**	**81.86%**	**87.68%**	**85.78%**	**91.33%**	**80.87%**	**88.51%**	**84.23%**	**85.64%**	**82.85%**	**91.35%**	**86.28%**	**81.06%**	**92.23%**

isomorphism introduced in Definition II.1. That is, supergraph search is to compute the answer set $A_Q = \{g_i \in D | g_i \subseteq Q\}$.

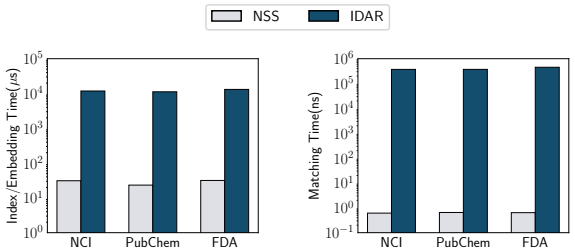
III. Our Approach

The framework of our proposed supergraph search model NSS is illustrated in Fig. 1. The left side of the Figure shows the forward propagation of NSS during the training process. NSS first trains a graph neural network model (GCN [3] is used as the default GNN) using three carefully designed objective functions, then uses the model to compute graph representations for query and date graphs, and finally predicts supergraph containment relationships based on the distances between these representations. Specifically, NSS has two main components: (1) the embedding network in which the vector representations are computed for the graphs and (2) the modules that produce the trainable objectives that enable NSS to preserve the structural and property information in and between the graphs and thus enhance the accuracy of the model. Besides, NSS utilizes the Wasserstein discriminator to preserve the interrelationship between graphs and the reconstruction model to preserve the label information. The prediction phase is illustrated on the right side of Fig. 1. In general, NSS embeds data graphs in set D into the vector space with a representation for each data graph. With the given query graph Q, the distances between Q and the data graphs on the representation space are utilized for the prediction , *i.e.,* the data graphs in D whose distances with Q in the embedding space are smaller than the threshold (within the pink circle on the right side) are selected as the subgraphs and vice versa.

IV. Experiment

To verify the accuracy of NSS, two naive learning-based baselines and two ablation baselines (NSS without reconstruction network or Wasserstein discriminator) are used for accuracy comparison. We also compare the efficiency of our method with the state-of-the-art exact approach IDAR [1]. Four real-world datasets, i.e., NCI, PubChem, FDA and Wiki-Vote.

The accuracy of experimental results is reported in Table I. The results in the table suggest that our model generally outperforms the other baselines in terms of accuracy. From the ablation study comparison, we can conclude that the reconstruction network and Wasserstein discriminator also play a critical role in the model. The efficiency comparison is reported in Fig. 2. Specifically, the preprocessing time comparison is reported in Fig. 2 (a), and the query processing time comparison is reported in Fig. 2 (b). We can find that the time costs are similar in the experiments on the three datasets. Regarding the indexing and embedding time comparison, NSS is nearly 3 orders of magnitude faster than IDAR. Additionally, on the LiveJournal dataset, NSS demonstrates its scalability. The average number of edges in its query graphs and data graph are $35,110,790$ and $4,925.25$, respectively. NSS completes embedding and prediction within an average time of 32.88 seconds and 149.13 microseconds, respectively. In conclusion, NSS innovates in supergraph search with GNNs framework, achieving unprecedented accuracy and speed.



(a) Indexing/Embedding Efficiency Comparison

(b) Query Processing Efficiency Comparison

Fig. 2: Efficiency Comparison Results

Acknowledgments This work was supported by ZJNSF LY21F020012.

References

[1] H. Kim, S. Min, K. Park, X. Lin, S.-H. Hong, and W.-S. Han, "Idar: Fast supergraph search using dag integration," *Proceedings of the VLDB Endowment*, vol. 13, no. 9, pp. 1456–1468, 2020.

[2] J. Gao, X. Huang, and J. Li, "Unsupervised graph alignment with wasserstein distance discriminator," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 426–435.

[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016.