Title	グラフニューラルネットワークを用いた知識グラフ表現学習に関する研究
Author(s)	曹, 逸銘
Citation	
Issue Date	2026-03
Type	Thesis or Dissertation
Text version	ETD
URL	https://hdl.handle.net/10119/20595
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 博士

氏　　　　　名	Cao Yiming		
学 位 の 種 類	博士（情報科学）		
学 位 記 番 号	博情第 582 号		
学位授与年月日	令和 8 年 3 月 25 日		
論 文 題 目	STUDY ON KNOWLEDGE GRAPH REPRESENTATION LEARNING WITH GRAPH NEURAL NETWORK		
論 文 審 査 委 員	Nguyen Le Minh	JAIST	Professor
	SHIRAI Kiyoaki	JAIST	Professor
	HASEGAWA Shinobu	JAIST	Professor
	Ashwin Ittoo	HEC, Liege University	Professor
	Satoshi Tojo	Asia University	Professor

論文の内容の要旨

Graph Neural Networks have emerged as the standard paradigm for representation learning on graph date. However, the practical application to real-world, large-scale graphs faces three main challenges: (1) noisy or uninformative graph topologies, (2) the prohibitive cost of generating high-quality node features, and (3) the difficulty of scaling to massive, heterogeneous graphs. This dissertation attempts to investigates structured models that are able to systematically addresses these challenges.

First, we address the challenge of learning on noisy topologies by proposing SSDA Framework, a differentiable, self-supervised data augmentation framework. SSDA jointly learns optimal policies for both structural and attribute augmentation in an end-to-end manner, enhancing the robustness of GNN models.

Second, we propose the LME framework to achieve a balance in the quality of LLM generated features and the computational cost. We conduct an analysis of feature engineering strategies for Text-Attributed Graphs. Our proposed LME pipeline, which utilizes an LLM as an intelligent keyword extractor to generate sparse features for the nodes, achieves over 90% of SOTA performance while being 2.4× smaller in storage and enabling 3.7× faster downstream GNN training.

Finally, to apply on massive heterogeneous graphs, we introduce the LME-Prop model. This scalable framework propagates the efficient LME sparse features from a 10% anchor set to the entire graph. The core innovation is a Dual-Channel Prop-GNN that fuses signals from the original topology (Structural Channel) and a new, efficiently-constructed *k*-NN graph (Semantic Channel). We demonstrate on the large-scale ogbn-mag benchmark that LME-Prop recovers over 80% of the full-graph SOTA performance while reducing total featurization time by 4.3×.

In conclusion, this dissertation establishes a complete, efficient, and scalable methodology for deploying semantically-aware GNNs on real-world, industrial scale graphs.

Keywords: Graph neural network, Graph representation learning, Data augmentation, Large language models, Heterogeneous graph, Computational efficiency.

論文審査の結果の要旨

This dissertation addresses fundamental challenges that limit the practical deployment of Graph Neural Networks on real-world, large-scale graphs, namely noisy graph structures, the high cost of node feature construction, and scalability to massive heterogeneous graphs. The problem formulation is clear, well-motivated, and highly relevant to both academic research and industrial applications. The dissertation makes three coherent and complementary contributions.

First, the proposed SSDA framework introduces a differentiable, self-supervised data augmentation strategy that jointly learns structural and attribute augmentations, improving robustness to noisy or uninformative graph topologies. This approach is principled and strengthens GNN performance without requiring manual augmentation design.

Second, the LME framework provides a thoughtful analysis of LLM-based feature engineering for text-attributed graphs and proposes an efficient pipeline that uses LLMs as intelligent keyword extractors to generate sparse node features.

Finally, the LME-Prop model extends these ideas to massive heterogeneous graphs through a scalable propagation mechanism based on a dual-channel architecture that integrates structural and semantic neighborhoods. The results on the benchmark data convincingly show that the method recovers a large fraction of full-graph performance with substantial reductions in featurization time.

Overall, the dissertation is well-structured, experimentally rigorous, and demonstrates a strong balance between methodological innovation and system-level efficiency. The proposed frameworks collectively establish a scalable and semantically-aware approach to graph representation learning. This is an excellent dissertation and we approve awarding a doctoral degree to Cao Yiming.