基於圖神經網路的數學式檢索：符號/變數變換的圖增強技術探索

顧竣熙 ，陳弘軒
國立中央大學資訊工程系
110502007@cc.ncu.edu.tw, hhchen@ncu.edu.tw

摘要

數學式的結構複雜且具高度符號性，傳統的文本匹配方法（例如 tf-idf）難以有效應用於數學式檢索。因此，圖神經網路（Graph Neural Networks, GNNs）因其處理圖結構數據的能力，成為解決數學式檢索問題的有效工具。本研究基於對比學習架構，提出了一種新的增強方法——符號/變數交換，並探討該方法對檢索效能的影響。我們使用 NTCIR-12 資料集進行實驗，將符號/變數交換方法與其他增強技術進行比較。結果顯示，該方法在完全相關的檢索任務中表現優異，提升了檢索的精確度。這表明，合適的圖形增強技術在數學式檢索中能有效維持樣本對的語意一致性，並顯著提升檢索效能。

關鍵字：數學式檢索，圖神經網路，對比學習，圖增強技術

Abstract

The complex structure and symbolic nature of mathematical equations make traditional text matching methods (e.g., tf-idf) ineffective for formula retrieval. As a result, Graph Neural Networks (GNNs), which excel at processing graph-structured data, have emerged as a promising solution for this problem. This study introduces a novel augmentation method, Symbol/Variable Swapping, within a contrastive learning framework and investigate its impact on retrieval performance. Using the NTCIR-12 dataset, we compare our method with other augmentation techniques. The results demonstrate that Symbol/Variable Swapping significantly improves retrieval accuracy in fully relevant tasks, highlighting its effectiveness in maintaining semantic consistency between sample pairs. This indicates that suitable graph augmentation strategies are crucial for enhancing formula retrieval performance.

Keywords：Math Information Retrieval, GNN, Contrastive Learning, Graph Augmentation

1. 研究背景

數學式具有複雜的符號結構和階層式的表示方式，這使得傳統的文本匹配方法（如 tf-idf）在數學式檢索中的應用效果有限。這些方法無法有效捕捉數學式的結構和符號間的關聯，因此在檢索與相似度比較方面存在明顯的限制。

數學式可以轉換為樹狀結構，而圖神經網路（Graph Neural Networks, GNNs）正適合處理這類圖結構的資料表示。已有研究使用圖對比學習進行自我監督訓練，並初步展現了 GNNs 在數學式檢索領域的潛力[1][2][3][4]。這些研究常採用圖形增強技術來生成正樣本對，進而進行對比學習。然而，由於數學式的樹結構通常較小，若增強技術使用不當，可能會破壞原本的語意一致性，進而影響樣本對間的關聯性與檢索效能。

本研究的重點在於探索合適的增強方法以提升數學式檢索的效能。為此，我們提出了一種新的增強技術——符號/變數交換，此方法依據節點類型來替換節點值。我們依據前人的實驗設計進行重現，並將我們的方法與其他常見的增強技術進行比較，以評估其在檢索任務中的表現。

2. 研究方法

本研究專注於參考文獻[1]中所採用的其中一個模型——圖對比學習模型（Graph Contrastive Learning, GraphCL）[3]。該模型透過增強技術生成增強圖形，與原始圖形形成正樣本對，並與其他圖形及其增強圖形形成負樣本對，藉由對比學習來學

習圖形嵌入（如圖一所示）。

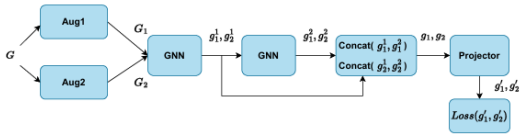

圖 1. GraphCL 的架構

實驗中使用的資料集為 NTCIR-12，包含 58 萬筆數學式，以及 20 筆檢索資料。我們使用 Tangent-CFT 工具取得數學式的符號嵌入，並轉換成符號布局樹（Symbol Layout Tree, SLT）及運算子樹（Operator Tree, OPT）的圖形表示。

在實驗中，我們應用了以下增強策略：特徵遮蔽（Feature Masking）、邊移除（Edge Removing）、節點丟棄（Node Dropping）、邊屬性遮蔽（Edge Attribute Masking）、隨機選擇（隨機應用前述四種增強方法，機率為 0.1）、符號/變數交換（Symbol/Variable Swapping）。除隨機選擇外，所有增強操作的機率均設為 30%。實驗涵蓋了六種不同的批次大小，並比較了兩種表示方式在部分相關與完全相關檢索任務中的表現。

最後，我們使用 bpref 指標來進行效能比較。相似度評估的部分，採用了原實驗的相關性評估方法[1]，K 值設為 1000。

3. 研究結果與討論

圖 2 和圖 3 的橫軸分別標示了各增強方法，從左至右依次為：特徵遮蔽、邊移除、節點丟棄、邊屬性遮蔽、隨機選擇、以及符號/變數交換。這些實驗結果是基於五筆資料的平均值而得。

在完全相關的檢索任務中，符號/變數交換方法的 bpref 指標相比其他增強技術有顯著提升，增幅約為 0.03 到 0.05 之間。這顯示該增強方法能有效維持樣本對之間的關聯性。特別是當使用符號布局樹（Symbol Layout Tree, SLT）表示方式時，符號/變數交換方法的表現尤為突出。推測其原因可能是 SLT 中的邊資訊在圖嵌入過程中扮演了更為重要的角色，而節點的資訊相對影響較小。這也驗證了論文 [1] 中對於「SLT 邊特徵在檢索任務中具有一定程度的重要性」的觀點，表明選擇合適的增強技術確實能有效提升模型的檢索效能。

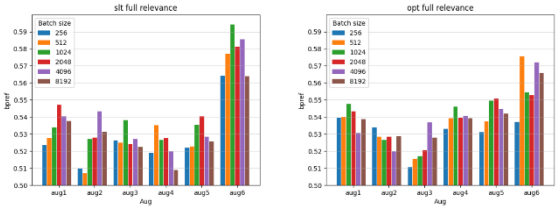

圖 2. 完全相關實驗 (左: SLT, 右: OPT)

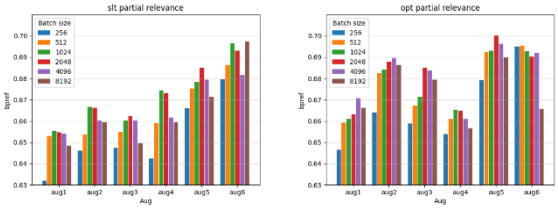

圖 3. 部分相關實驗 (左: SLT, 右: OPT)

4. 參考文獻

[1] Wang, P.-S., and H.-H. Chen. "The Effectiveness of Graph Contrastive Learning on Mathematical Information Retrieval." International Workshop on Graph-Based Approaches in Information Retrieval (2024).

[2] Sun, Fan-Yun, et al. "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization." arXiv preprint arXiv:1908.01000 (2019)

[3] You, Yuning, et al. "Graph contrastive learning with augmentations." Advances in neural information processing systems 33 (2020): 5812-5823.

[4] Thakoor, Shantanu, et al. "Large-scale representation learning on graphs via bootstrapping." arXiv preprint arXiv:2102.06514 (2021).