Structure-Preserving Graph Contrastive Learning for Mathematical Information Retrieval

Chun-Hsi Ku[1], Hung-Hsuan Chen[1,*]

[1]*Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan, 320317*

Abstract

This paper introduces Variable Substitution as a domain-specific graph augmentation technique for graph contrastive learning (GCL) in the context of searching for mathematical formulas. Standard GCL augmentation techniques often distort the semantic meaning of mathematical formulas, particularly for small and highly structured graphs. Variable Substitution, on the other hand, preserves the core algebraic relationships and formula structure. To demonstrate the effectiveness of our technique, we apply it to a classic GCL-based retrieval model. Experiments show that this straightforward approach significantly improves retrieval performance compared to generic augmentation strategies. We release the code on GitHub.[1].

Keywords

Mathematical Information Retrieval, Graph Augmentation, Graph Contrastive Learning

1. Introduction

Mathematical Information Retrieval (MIR) is a critical subfield of information science, essential to advance scientific discovery by enabling the effective search and retrieval of mathematical content from vast digital corpora [1, 2]. Effective MIR systems are foundational to next-generation scholarly information access platforms, enabling researchers to navigate the enormous scientific corpus beyond simple keyword matching. Unlike traditional Information Retrieval (IR), which primarily focuses on text-based queries and documents, MIR must contend with the unique structural and semantic complexities inherent in mathematical formulas. Although established IR systems successfully leverage techniques such as TF-IDF and various measures of semantic similarity to retrieve relevant textual documents [1, 2, 3, 4], these methods often fall short when applied to mathematical content [5]. MIR necessitates models capable of interpreting the intricate syntactic structure of mathematical expressions, recognizing that formulas with different surface appearances can represent the same underlying concept. This inherent characteristic distinguishes MIR significantly from conventional text-centric IR paradigms.

Recent progress in the MIR domain has been driven substantially by the application of sophisticated machine learning techniques, particularly graph neural networks (GNNs) and related methods [6, 7, 8]. These approaches aim to generate meaningful and effective representations (embeddings) of mathematical formulas by meticulously capturing their structure and the complex relationships among constituent mathematical symbols. Previous research has investigated various graph-based and text-based strategies to navigate the specific challenges posed by MIR [5, 8, 7, 6]. Among these, graph contrast learning (GCL) has emerged as a particularly promising direction, especially for mitigating the common scarcity of labeled relevance data (i.e., explicit relevance scores) in formula retrieval tasks [6]. GCL frameworks learn robust formula embeddings by treating retrieval as a contrastive problem, aiming to maximize the similarity between different augmented views of the same formula while simultaneously minimizing similarity with unrelated formulas. Models such as MathBERT [9] and TangentCFT [5] have demonstrated the potential to integrate both formula structure and the surrounding textual context,

[1] https://github.com/lazywulf/formula_ret_aug

SCOLIA '26: Second International Workshop on Scholarly Information Access (SCOLIA), April 2, 2026, Delft, The Netherlands

[*] Corresponding author.

jimmyku123@gmail.com (C. Ku); hhchen1105@acm.org (H. Chen)
https://in.ncu.edu.tw/~hhchen/ (H. Chen)
0000-0001-5137-4449 (H. Chen)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-03-26

with TangentCFT frequently serving as a formidable baseline, especially for retrieval models that rely solely on formula structure.

However, a significant hurdle arises when applying standard GCL methodologies to MIR, specifically regarding the data augmentation step, which is crucial for contrastive learning. The widely used graph augmentation techniques prevalent in the GCL literature, including node drop, edge masking, and feature masking [10], often prove detrimental when applied to the typically small graph structures representing mathematical formulas. Within MIR, the compact nature of formula graphs means that even seemingly minor alterations introduced by these conventional augmentations can drastically distort the formula's fundamental meaning or structural integrity. Removing a single critical operator node or masking an edge signifying a key dependency can easily render the formula syntactically incorrect or semantically nonsensical. This sensitivity arises because nearly every node and edge in a formula graph carries substantial semantic weight. Consequently, employing inappropriate augmentation methods can severely disrupt the vital relationships among mathematical symbols, ultimately impeding the model's ability to learn effective representations and leading to suboptimal retrieval performance.

To address the inherent limitations of conventional augmentations within the MIR context, we introduce a straightforward yet highly effective graph augmentation method tailored explicitly for mathematical formulas, termed Variable Substitution. This technique is designed to introduce the necessary representational variance required for effective contrastive learning while rigorously preserving the core structural and semantic integrity of the original mathematical expression. By strategically focusing on the substitution of variablesâĂŤelements whose specific identity often matters less than their role within the structure, rather than altering the graph's fundamental topology or critical operator nodes, Variable Substitution effectively navigates the pitfalls of standard techniques. This approach aims to preserve the essential mathematical relationships encoded in the graph structure, thereby addressing the identified shortcomings of existing augmentations for formula graphs.

This paper contributes the following advancements to the field of Mathematical Information Retrieval. First, we introduce Variable Substitution, a simple yet powerful graph augmentation method designed explicitly for MIR, which preserves the essential formula structure during the data augmentation phase of contrastive learning. Second, we present comprehensive experiments demonstrating that Variable Substitution yields significant improvements in formula retrieval performance compared to both existing standard graph augmentation techniques and the established state-of-the-art baseline, TangentCFT [5]. Third, we analyze the efficacy of Variable Substitution across distinct mathematical graph representations, namely Symbol Layout Trees (SLTs) and Operator Trees (OPTs), showcasing its robustness and adaptability by consistently outperforming baseline methods on both structures.

2. Related Work

Mathematical Information Retrieval (MIR) poses unique challenges because it requires understanding both the structure and semantics of mathematical expressions. Previous works have introduced various graph-based and text-based approaches [5, 9, 6]. Among these, GCL has demonstrated effectiveness in learning formula embeddings by treating formula retrieval as a contrastive learning problem. Models like TangentCFT [5] and MathBERT [9] have incorporated both formula structure and text, with TangentCFT serving as a strong baseline for formula-only retrieval models.

Several augmentation techniques have been proposed for GCL, including node dropping, edge masking, and feature masking. However, these approaches often struggle with the small size of formula graphs, where even minor augmentations can significantly alter the formula's meaning. To address this, we propose an augmentation method explicitly tailored to mathematical formulas.

3. Method

This section outlines our methodology, including graph structure generation, token embedding generation, graph contrastive learning with Variable Substitution, and the online query module. An overview

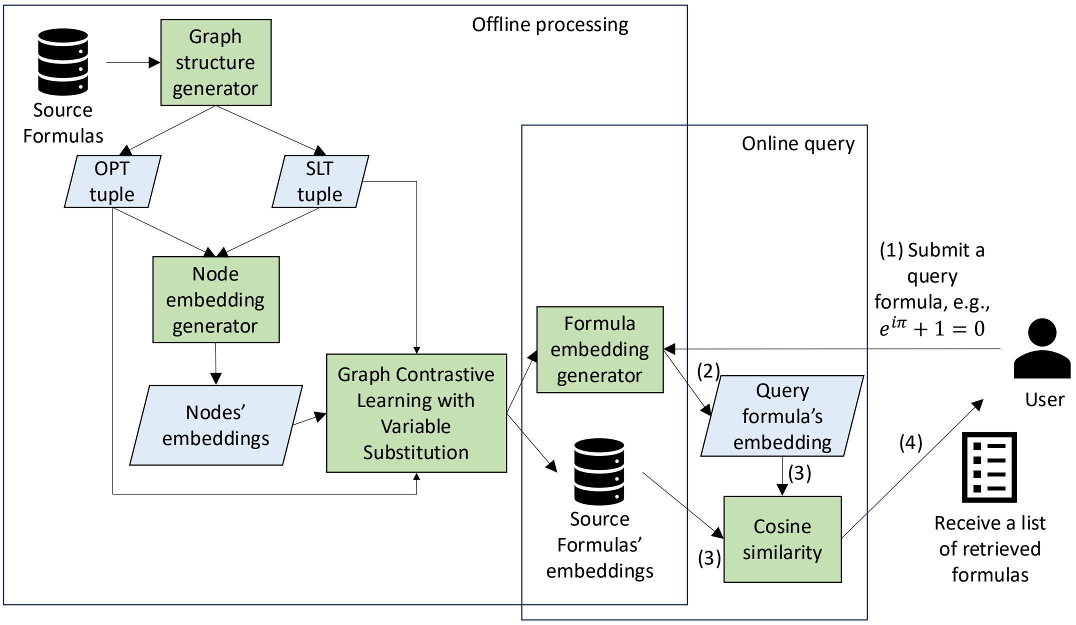

Offline processing
Source Formulas
Graph structure generator
OPT tuple
SLT tuple
Node embedding generator
Nodes' embeddings
Graph Contrastive Learning with Variable Substitution
Online query
Formula embedding generator
Source Formulas' embeddings
(1) Submit a query formula, e.g., $e^{i\pi}+1=0$
User
(2)
Query formula's embedding
(3)
Cosine similarity
(3)
(4)
Receive a list of retrieved formulas

Figure 1: The online and offline processing of the entire framework; we focus on Graph Contrastive Learning with Variable Substitution in this paper.

is given in Figure 1.

3.1. Graph Structure Generator

Given a mathematical formula, the graph structure generator converts it into graphs that capture the semantic and syntactic relationships among numbers, variables, and operators. We employ two graph structures: the Symbol Layout Tree (SLT) and the Operator Tree (OPT). SLT captures the spatial arrangement of symbols, whereas OPT focuses on operational semantics by representing operators as internal nodes and operands as child nodes [5, 6]. These graphs serve as input for the subsequent learning modules.

3.2. Token Embedding Generator

The token embedding generator (TEG) utilizes the fastText model to generate embeddings for each node in the graph [5, 6, 11, 12]. First, the TEG applies random walks to sample paths from the SLT or OPT graphs. These paths are then encoded using fastText, producing 100-dimensional embeddings for each node. Each embedding reflects the local neighborhood of the symbol within the graph structure, capturing both positional and contextual information. These embeddings serve as the basis for constructing formula-level graph representations.

3.3. Graph Contrastive Learning with Variable Substitution

Researchers have increasingly used GCL to generate formula embeddings without relying on labeled relevance scores [6, 10]. However, popular graph augmentation techniques, such as node/edge dropping or attribute masking, are ill-suited for mathematical graphs. Even minor modifications—like dropping an operator or variable—can fundamentally change a formula's interpretation. Such augmentations are likely to introduce destructive noise, hindering the model's ability to learn meaningful representations.

To address this, we propose a controlled augmentation method, Variable Substitution, which preserves the structural and semantic integrity of formulas. In our GCL setup, we first create an "augmented view" of a formula graph by applying Variable Substitution: nodes representing variables are randomly substituted with other variables, and nodes representing numbers are swapped with different numbers.

This process alters node identities while preserving the graph's topology. Positive pairs are then formed by the original formula graph and its augmented view. Negative pairs consist of the original graph and any other formula graph within the same training batch. The model is trained to minimize the distance between positive pairs and maximize the distance between negative pairs in the embedding space, thereby learning robust representations that capture the similarity among abstract formulas.

Once the model has learned to generate formula embeddings through contrastive learning, we store them in a database for efficient retrieval. Overfitting is unlikely to be a concern because contrastive learning emphasizes distinguishing between formulas based on inherent structural similarities rather than relying on human-labeled pairs. This enables our model to generalize more effectively to new, unseen formulas, without being constrained by the biases or limitations inherent in human labels.

3.4. Online Query Module

The online query module retrieves relevant formulas in response to user queries. When a user submits a query formula, the system generates an embedding for the query based on the trained formula embedding generator. The system then computes the cosine similarity between the query formula embedding and the embeddings of all formulas in the database. Based on these similarity scores, the system ranks the formulas in descending order and returns the most relevant results to the user.

4. Experiments

4.1. NTCIR-12 MathIR Dataset

We evaluate our method using the NTCIR-12 MathIR dataset [13], a benchmark commonly used for mathematical information retrieval (MIR) tasks. The dataset comprises an extensive collection of mathematical formulas extracted from Wikipedia and relevance judgments for a set of query formulas. The relevance scores are integers between 0 and 4, with higher scores indicating a closer match between the query and the retrieved formulas. This dataset is specifically designed to test both exact and approximate formula-matching capabilities.

4.2. Evaluation Metrics: Bpref and Full vs. Partial Match

For evaluation, we use the binary preference metric, bpref, which is particularly suitable for scenarios with incomplete relevance judgments. Bpref measures how often relevant documents are ranked higher than irrelevant ones without assuming that all relevant documents have been labeled in the dataset. This makes it an ideal metric for our experiments, where relevance judgments are limited to a subset of formula pairs.

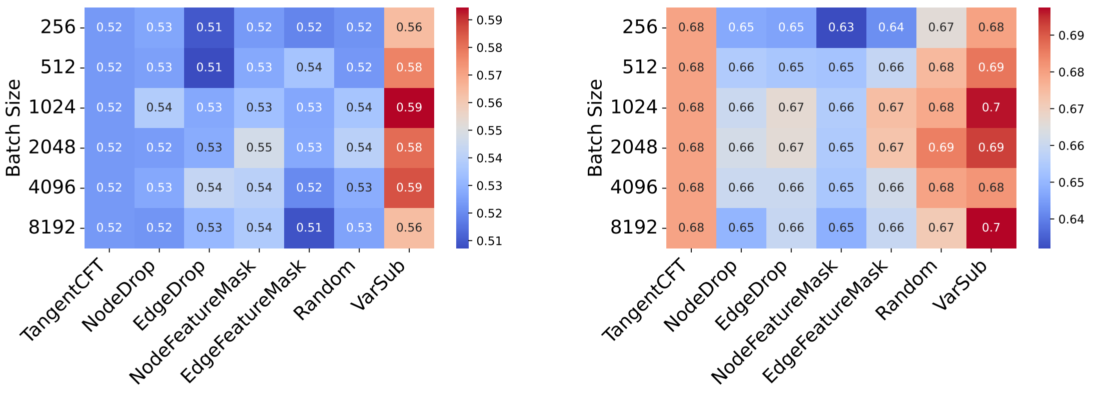
Since the bpref metric operates in a binary setting (relevant or irrelevant), but the dataset annotations range from 0 to 4, we must apply a threshold to convert the dataset labels to binary. We use two thresholds for this purpose. First, we consider only formulas with a score of 3 or higher to be relevant, and all others to be irrelevant; we refer to this approach as "full relevance". Second, we treat only formulas with a score of 0 as irrelevant, and all other scores are considered relevant; we refer to this approach as "partial relevance".

4.3. Compared methods

We compare Variable Substitution with several generic graph augmentation strategies. To ensure a fair comparison, we use TangentCFT [5] as the base model for all augmentation methods. The compared augmentation strategies include: Node Drop, which randomly removes nodes from the graph; Edge Drop, which randomly removes edges; Node Feature Mask, which masks the features of sampled nodes; and Edge Feature Mask, which masks the features of sampled edges. Finally, the Random strategy randomly selects one of the four aforementioned techniques for each graph.

Since contrastive learning is often sensitive to batch size, we evaluated different batch sizes across all graph augmentation strategies.

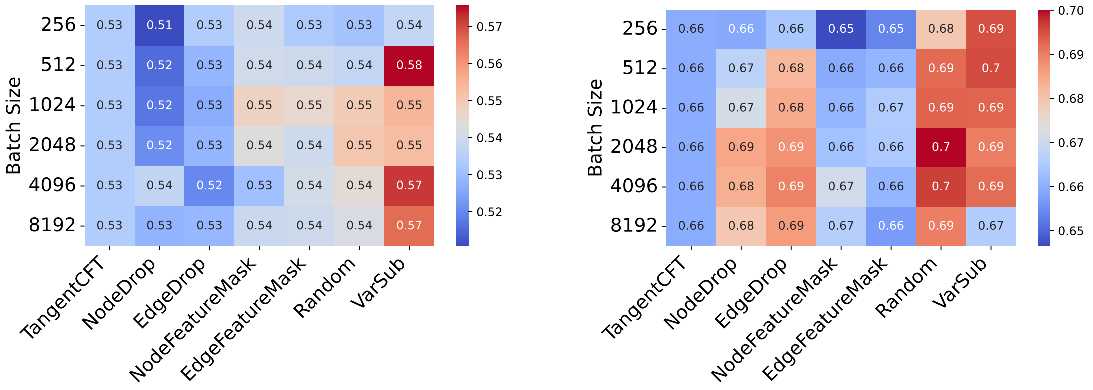
4.4. Results



(a) The full relevance setting

(b) The partial relevance setting

Figure 2: The bpref scores using the SLT layout



(a) The full relevance setting

(b) The partial relevance setting

Figure 3: The bpref scores using the OPT layout

The results, presented as heat maps in Figure 2 and Figure 3, show the performance of different augmentation methods in various batch sizes for the SLT and OPT layouts, respectively. Overall, Variable Substitution demonstrates superior performance, particularly in the SLT representation, which appears to be more sensitive to structural changes.

Figure 2 illustrates the results for the SLT structure, which captures the spatial layout of formula symbols. In this context, Variable Substitution shows a distinct advantage. Under the "full relevance" setting, it achieves a top bpref score of 0.59, yielding a significant margin over the next best methods, which score at most 0.54. This significant gap underscores the importance of preserving the topological structure. Generic augmentations, such as Node Drop or Edge Drop, can severely disrupt the spatial arrangement (e.g., removing a superscript), thereby corrupting the formula's meaning. In contrast, Variable Substitution maintains the complete layout, enabling the model to learn the formula's abstract structure more effectively. Under the "partial relevance" threshold, it again achieves the highest score of 0.70, reinforcing its superiority.

A similar, though less pronounced, trend is observed for the OPT structure shown in Figure 3, which represents the formula's operational hierarchy. Variable Substitution consistently outperforms other techniques across all batch sizes, achieving a bpref score of 0.58 in the "full relevance" setting, compared to 0.55 for the random augmentation. In the "partial relevance" setting, Variable Substitution and random strategy lead with a score of 0.70. It suggests that the operational semantics of OPTs may be slightly more resilient to random alterations than the strict spatial rules of SLTs. Nevertheless, the consistent lead of Variable Substitution confirms that preserving the integrity of the operator-operand tree is still the most effective strategy.

Across both graph representations, we note two general trends. First, larger batch sizes, which are typically expected to improve contrastive learning by providing more negative examples, only yield marginal performance gains. Second, the results are highly stable; we repeated each experiment 5 times, and the standard deviations were minimal across all settings (typically 0.001 to 0.009). These findings collectively underscore the effectiveness and robustness of Variable Substitution as a structure-preserving augmentation technique for math formula search.

5. Discussion

This paper introduces a simple yet effective augmentation technique, Variable Substitution, for graph contrastive learning in the context of math formula search. Through extensive experiments, we demonstrate that this domain-specific method outperforms generic augmentation strategies, particularly in identifying structurally similar formulas. Our results suggest that preserving the core structural relationships between symbols and variables is critical to improving formula retrieval performance.

Future research could explore more sophisticated or targeted augmentation techniques that preserve mathematical semantics while increasing the diversity of the training data. Additionally, we are interested in applying this structure-preserving augmentation approach to other IR tasks involving structured data, such as chemical formula retrieval.

Acknowledgments

We acknowledge the support of Taiwan's National Science and Technology Council under grant number 113-2221-E-008-100-MY3.

Declaration on Generative AI

The authors used LLMs to improve readability. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of documentation 28 (1972) 11–21. doi:10.1108/eb026526.

[2] J. Wu, K. M. Williams, H.-H. Chen, M. Khabsa, C. Caragea, S. Tuarob, A. G. Ororbia, D. Jordan, P. Mitra, C. L. Giles, Citeseerx: Ai in a digital library search engine, AI Magazine 36 (2015) 35–48. doi:10.1609/aimag.v36i3.2601.

[3] C. Caragea, J. Wu, A. Ciobanu, K. Williams, J. Fernández-Ramírez, H.-H. Chen, Z. Wu, L. Giles, Citeseer x: A scholarly big dataset, in: Advances in Information Retrieval: 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings 36, Springer, 2014, pp. 311–322. doi:10.1007/978-3-319-06028-6_26.

[4] H.-H. Chen, P. Treeratpituk, P. Mitra, C. L. Giles, Csseer: an expert recommendation system based on citeseerx, in: Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, 2013, pp. 381–382. doi:10.1145/2467696.2467750.

[5] B. Mansouri, S. Rohatgi, D. W. Oard, J. Wu, C. L. Giles, R. Zanibbi, Tangent-cft: An embedding model for mathematical formulas, in: Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval, 2019, pp. 11–18. doi:10.1145/3341981.3344235.

[6] P.-S. Wang, H.-H. Chen, The effectiveness of graph contrastive learning on mathematical information retrieval, in: International Workshop on Graph-Based Approaches in Information Retrieval, Springer, 2024, pp. 60–72. doi:10.1007/978-3-031-71382-8_5.

[7] L. Pfahler, K. Morik, Self-supervised pretraining of graph neural network for the retrieval of related mathematical expressions in scientific articles, arXiv preprint arXiv:2209.00446 (2022). doi:10.48550/arXiv.2209.00446.

[8] S. Peng, L. Gao, K. Yuan, Z. Tang, Image to latex with graph neural network for mathematical formula recognition, in: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16, Springer, 2021, pp. 648–663. doi:10.1007/978-3-030-86331-9_42.

[9] S. Peng, K. Yuan, L. Gao, Z. Tang, Mathbert: A pre-trained model for mathematical formula understanding, arXiv preprint arXiv:2105.00377 (2021). doi:10.48550/arXiv.2105.00377.

[10] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, Y. Shen, Graph contrastive learning with augmentations, Advances in neural information processing systems 33 (2020) 5812–5823. URL: https://proceedings.nips.cc/paper/2020/file/3fe230348e9a12c13120749e3f9fa4cd-Paper.pdf.

[11] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146. doi:10.1162/tacl_a_00051.

[12] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651 (2016). doi:10.48550/arXiv.1612.03651.

[13] M. P. Kato, K. Kishida, N. Kando, T. Saka, M. Sanderson, Report on ntcir-12: The twelfth round of nii testbeds and community for information access research, SIGIR Forum 50 (2017) 18âĂŞ27. doi:10.1145/3053408.3053413.