arXiv:2603.08012v1 [cs.IR] 9 Mar 2026

保结构图对比学习的数学信息检索

Chun-Hsi Ku、Hung-Hsuan Chen计算
机科学与信息工程系 国立中央大学 台湾桃园
jimmyku123@gmail.com,
hhchen1105@acm.org

摘要

本文介绍了变量替换，这是一种针对数学公式检索领域的图对比学习（GCL）专用图增强技术。标准的图对比学习增强技术往往会扭曲数学公式的语义，尤其是对于较小且结构高度复杂的图而言。相比之下，变量替换能够保留核心代数关系和公式结构。为展示该技术的有效性，我们将其应用于一种经典的基于图对比学习的检索模型。实验表明，与通用增强策略相比，这种简单直接的方法显著提升了检索性能。我们在 GitHub 上发布了代码。[1]。

***K*eyw词**数学信息检索 · 图增强 · 图对比学习

1 引言

数学信息检索（MIR）是信息科学的一个重要分支，通过支持从海量数字语料库中有效搜索和检索数学内容，对推动科学发现至关重要 [1, 2]。有效的 MIR 系统是新一代学术信息访问平台的基础，使研究人员能够超越简单的关键词匹配，在浩瀚的科学语料库中进行探索。与主要关注基于文本的查询和文档的传统信息检索（IR）不同，MIR 必须应对数学公式固有的独特结构和语义复杂性。尽管成熟的 IR 系统成功地利用词频-逆文档频率等技术以及各种语义相似度度量来检索相关文本文档 [1, 2, 3, 4],，但这些方法应用于数学内容时往往效果不佳 [5]。MIR 需要能够解读数学表达式复杂句法结构的模型，并认识到外观不同的公式可能表达相同的底层概念。这一固有特性使 MIR 与传统的文本中心型 IR 范式有显著区别。

近期，数学信息检索（MIR）领域的进展很大程度上得益于先进机器学习技术的应用，尤其是图神经网络（GNN）及相关方法 [6, 7, 8]。这些方法旨在通过细致捕捉数学公式的结构及其组成数学符号之间的复杂关系，生成有意义且有效的表示（嵌入）。先前的研究探讨了各种基于图和基于文本的策略，以应对数学信息检索面临的特定挑战 [5, 8, 7, 6]。其中，图对比学习（GCL）已成为一个颇具前景的方向，尤其有助于缓解公式检索任务中带标签的相关性数据（即明确的相关性分数）普遍匮乏的问题 [6]。图对比学习框架将检索视为一个对比问题，通过最大化同一公式不同增强视图之间的相似度，同时最小化其与无关公式之间的相似度，来学习稳健的公式嵌入。MathBERT [9] 和 TangentCFT [5] 等模型展现了融合公式结构与周围文本上下文的潜力；TangentCFT 经常被用作强有力的基线，尤其适用于仅依赖公式结构的检索模型。

[1] https://github.com/lazywulf/formula_ret_a ug

然而，将标准图对比学习方法应用于数学信息检索时，数据增强环节会带来一项重大挑战，而这一环节对对比学习至关重要。图对比学习文献中广泛采用的图增强技术，包括节点删除、边掩码和特征掩码 [10], ，应用于通常规模较小的数学公式图时，往往会适得其反。在数学信息检索中，公式图结构紧凑，这意味着即便是这些传统增强方法带来的细微改动，也可能严重扭曲公式的基本含义或结构完整性。删除一个关键运算符节点，或掩蔽一条表示重要依赖关系的边，都可能轻易导致公式在语法上错误或语义上毫无意义。这种敏感性源于公式图中几乎每个节点和每条边都承载着重要的语义权重。因此，采用不恰当的增强方法会严重破坏数学符号之间至关重要的关系，最终妨碍模型学习有效表示的能力，并导致检索性能不佳。

为应对数学信息检索场景中传统增强方法的固有局限，我们提出一种简单而高效、专为数学公式设计的图增强方法，称为变量替换。该技术旨在引入有效对比学习所需的必要表示差异，同时严格保持原始数学表达式的核心结构与语义完整性。通过有策略地聚焦于变量替换——这些元素的具体身份往往不如其在结构中的作用重要——而非改变图的基本拓扑结构或关键运算符节点，变量替换有效规避了标准技术的缺陷。该方法旨在保留编码于图结构中的基本数学关系，从而解决现有公式图增强方法的上述不足。

本文对数学信息检索领域做出如下贡献。首先，我们提出变量替换，这是一种专为数学信息检索设计的简单而强大的图增强方法，可在对比学习的数据增强阶段保持基本公式结构。其次，我们开展了全面实验，表明与现有标准图增强技术以及既有的先进基线TangentCFT [5]相比，变量替换能显著提升公式检索性能。第三，我们分析了变量替换在不同数学图表示上的有效性，即符号布局树（SLTs）和运算符树（OPTs），结果表明该方法在两种结构上均持续优于基线方法，展现出稳健性与适应性。

2 相关工作

数学信息检索（MIR）面临独特的挑战，因为它需要同时理解数学表达式的结构和语义。以往研究提出了各种基于图和基于文本的方法 [5, 9, 6]。其中，图对比学习（GCL）将公式检索视为对比学习问题，在学习公式嵌入方面展现出良好效果。TangentCFT [5] 和 MathBERT [9] 等模型结合了公式结构与文本，其中 TangentCFT 是纯公式检索模型的强大基线。
研究者为图对比学习提出了多种增强技术，包括节点丢弃、边掩码和特征掩码。然而，这些方法往往难以适用于规模较小的公式图，因为即使是轻微的增强也可能显著改变公式的含义。为此，我们提出了一种专为数学公式设计的增强方法。

3 方法

本节介绍我们的方法，包括图结构生成、词元嵌入生成、结合变量替换的图对比学习，以及在线查询模块。总体概览见图1。

3.1 图结构生成器

给定一个数学公式，图结构生成器将其转换为图，以捕捉数字、变量和运算符之间的语义与句法关系。我们采用两种图结构：符号布局树（SLT）和运算符树（OPT）。SLT捕捉符号的空间排列，而OPT通过将运算符表示为内部节点、将操作数表示为子节点来关注操作语义 [5, 6]。这些图作为后续学习模块的输入。

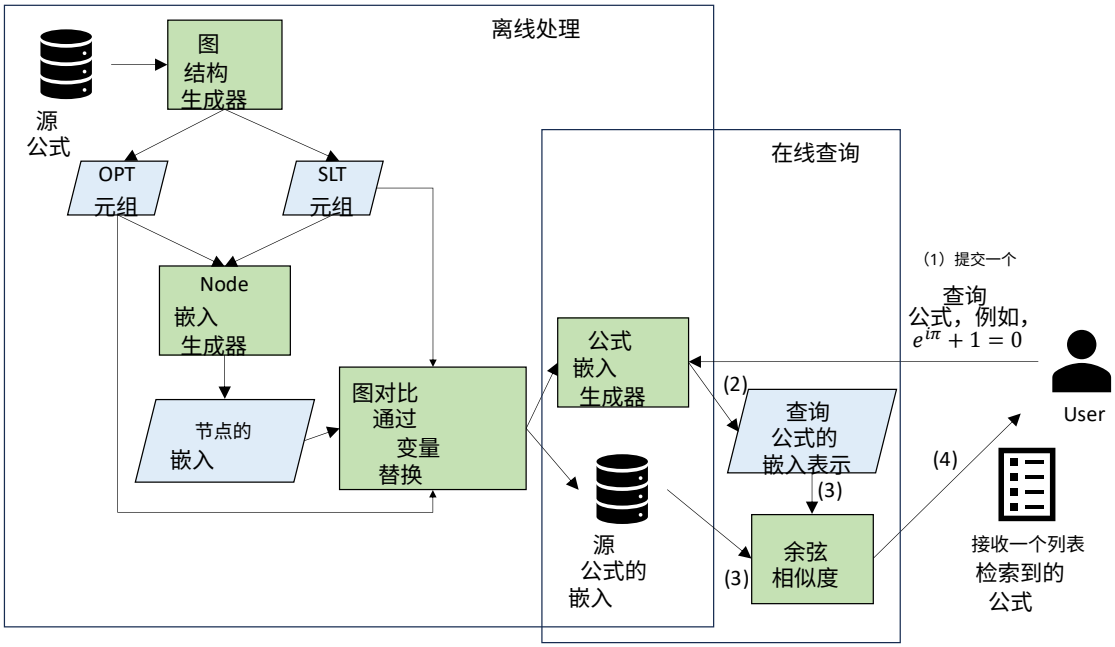


图1：整个框架的在线和离线处理；本文聚焦于结合变量替换的图对比学习。

3.2 词元嵌入生成器

词元嵌入生成器（TEG）利用fastText模型为图中的每个节点生成嵌入 [5, 6, 11, 12]。首先，TEG应用随机游走从SLT或OPT图中采样路径。随后使用fastText对这些路径进行编码，为每个节点生成100维嵌入。每个嵌入反映符号在图结构中的局部邻域，同时捕捉位置信息与上下文信息。这些嵌入是构建公式级图表示的基础。

3.3 Graph 采用变量替换的对比学习 tution

研究人员越来越多地使用图对比学习生成公式嵌入，而不依赖标注的相关性分数 [6, 10]。然而，常见的图增强技术（例如节点/边删除或属性掩码）并不适用于数学图。即使是细微的修改——例如删除运算符或变量——也可能从根本上改变公式的含义。这类增强很可能引入破坏性噪声，妨碍模型学习有意义的表示。

为了解决这一问题，我们提出了一种受控增强方法——变量替换，它能够保留公式的结构和语义完整性。在我们的图对比学习设置中，我们首先通过应用变量替换，为公式图创建一个“增强视图”：随机将表示变量的节点替换为其他变量，并将表示数字的节点替换为其他数字。此过程在保留图拓扑结构的同时改变节点标识。随后，以原始公式图及其增强视图构成正样本对。负样本对则由原始图与同一训练批次中的任意其他公式图组成。模型经过训练，以缩小嵌入空间中正样本对之间的距离，并扩大负样本对之间的距离，从而学习能够捕捉抽象公式间相似性的稳健表示。

模型通过对比学习掌握公式嵌入的生成方法后，我们会将其存储在数据库中，以便高效检索。过拟合不太可能成为问题，因为对比学习注重依据公式固有的结构相似性来区分公式，而非依赖人工标注对。这使我们的模型能够更有效地泛化到新的、未见过的公式，而不受人工标签固有偏差或局限性的制约。

3.4 在线查询模块

在线查询模块会根据用户的查询检索相关公式。用户提交查询公式后，系统会根据经过训练的公式嵌入生成器为该查询生成嵌入。随后，系统计算查询公式嵌入与数据库中所有公式嵌入之间的余弦相似度。系统根据这些相似度分数按降序排列公式，并向用户返回最相关的结果。

4 实验

4.1 NTCIR-12 MathIR 数据集

我们使用 NTCIR-12 MathIR 数据集 [13], 这一数学信息检索（MIR）任务中常用的基准数据集来评估我们的方法。该数据集包含从维基百科中提取的大量数学公式，以及针对一组查询公式的相关性判断。相关性分数为 0 到 4 之间的整数，分数越高表示查询公式与检索到的公式之间的匹配程度越高。该数据集专门用于测试公式的精确匹配和近似匹配能力。

4.2 评估指标：bpref 以及完全匹配与部分匹配

在评估中，我们使用二元偏好指标 bpref。该指标特别适用于相关性判断不完整的场景。bpref 衡量相关文档的排名高于不相关文档的频率，且不假设数据集已标注所有相关文档。因此，对于我们的实验而言，bpref 是一种理想的指标，因为相关性判断仅针对部分公式对。

由于 bpref 指标采用二元设置（相关或不相关），而数据集的标注范围为 0 到 4，因此我们必须设置阈值，将数据集标签转换为二元标签。我们采用两种阈值。首先，我们仅将得分为 3 或更高的公式视为相关，其他公式均视为不相关；我们将这种方法称为“完全相关”。其次，我们仅将得分为 0 的公式视为不相关，其他得分的公式均视为相关；我们将这种方法称为“部分相关”。

4.3 对比方法

我们将变量替换与几种通用图增强策略进行比较。为确保比较公平，所有增强方法均使用 TangentCFT [5] 作为基础模型。参与比较的增强策略包括：节点删除，即随机移除图中的节点；边删除，即随机移除边；节点特征掩码，即对采样节点的特征进行掩码；以及边特征掩码，即对采样边的特征进行掩码。最后，随机策略会为每个图随机选取上述四种方法之一。

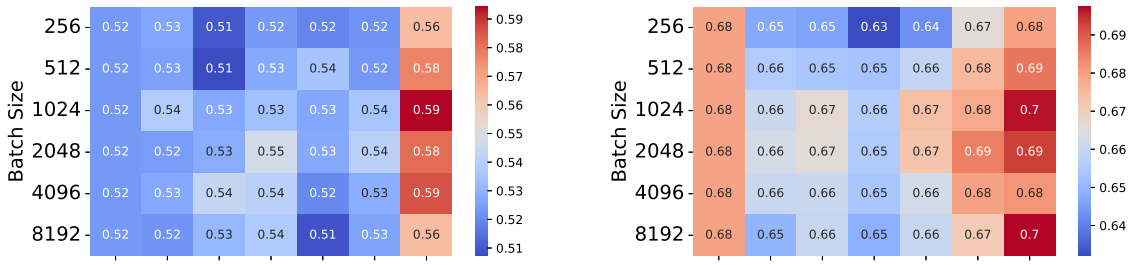
由于对比学习通常对批次大小较为敏感，我们在所有图增强策略上评估了不同的批次大小。

4.4 结果

结果以热力图形式呈现在图2和图3中，分别展示了不同增强方法在不同批次大小下对SLT和OPT布局的性能。总体而言，变量替换表现出更优的性能，尤其是在SLT表示中，该表示似乎对结构变化更为敏感。

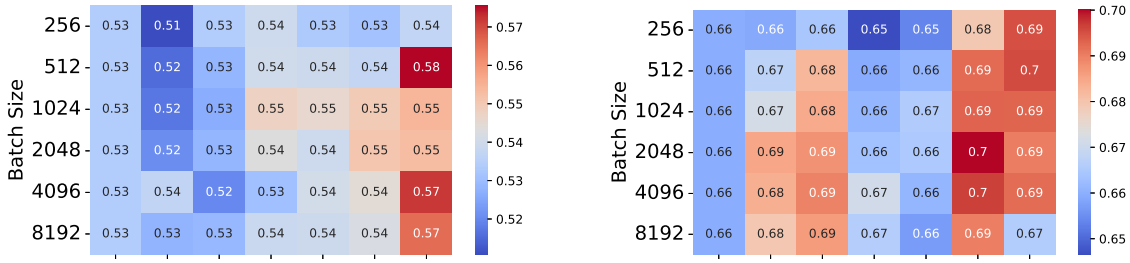
图2展示了SLT结构的结果，该结构捕捉公式符号的空间布局。在此背景下，变量替换展现出明显优势。在“完全相关”设定下，它取得了最高的bpref分数0.59，显著领先于次优方法（最高仅为0.54）。这一显著差距凸显了保持拓扑结构的重要性。通用增强（如节点删除或边删除）可能严重破坏空间排列（例如移除上标），从而扭曲公式的含义。相比之下，变量替换保持了完整布局，使模型能够更有效地学习公式的抽象结构。在“部分相关”阈值下，它再次取得最高分数0.70，进一步印证了其优越性。

图3所示的OPT结构也呈现出类似但不太明显的趋势，该结构表示公式的运算层次结构。在所有批量大小下，变量替换始终优于其他技术；在“完全相关”设置下，其bpref分数达到0.58，而随机增强为0.55。在



(a) 完全相关性设置 (b) 部分相关性设置

图2：使用 SLT 布局的 bpref 分数



(a) 完全相关性设置 (b) 部分相关性设置

图3：使用 OPT 布局时的 bpref 分数

“部分相关”设置下，变量替换和随机策略以0.70的分数领先。这表明，OPT的操作语义可能比SLT的严格空间规则更能抵御随机改动。尽管如此，变量替换始终保持领先，这进一步证实，保持运算符-操作数树的完整性仍是最有效的策略。

在两种图表示中，我们观察到两个普遍趋势。首先，较大的批次大小通常被认为能够通过提供更多负样本来提升对比学习，但实际带来的性能提升微乎其微。其次，结果非常稳定；我们将每项实验重复了 5 次，在所有设置下标准差都很小（通常为 0.001 到 0.009）。这些发现共同凸显了变量替换作为一种用于数学公式搜索的结构保持增强技术的有效性和稳健性。

5 讨论

本文介绍了一种简单而有效的增强技术——变量替换，用于数学公式搜索领域的图对比学习。通过大量实验，我们证明，这种领域特定方法的表现优于通用增强策略，尤其是在识别结构相似的公式方面。结果表明，保留符号与变量之间的核心结构关系，对于提升公式检索性能至关重要。

未来的研究可以探索更复杂或更具针对性的增强技术，在提高训练数据多样性的同时保留数学语义。此外，我们也有兴趣将这种保留结构的增强方法应用于其他涉及结构化数据的信息检索任务，例如化学式检索。

致谢与生成式人工智能使用声明

我们感谢台湾国家科学及技术委员会通过项目编号113-2221-E-008-100-MY3提供的支持。作者使用Gemini改进语言表达和可读性。作者根据需要审阅并编辑了内容，并对内容承担全部责任。

参考文献

[1] Karen Sparck Jones。词项专指性的统计解释及其在检索中的应用。《文献学杂志》，28(1)：11–21，1972年。[2] Jian Wu、Kyle Mark Williams、Hung-Hsuan Chen、Madian Khabsa、Cornelia Caragea、Suppawong Tuarob、Alexander G Ororbia、Douglas Jordan、Prasenjit Mitra和C Lee Giles。Citeseerx：数字图书馆搜索引擎中的人工智能。《人工智能杂志》，36(3)：35–48，2015年。[3] Cornelia Caragea、Jian Wu、Alina Ciobanu、Kyle Williams、Juan Fernández-Ramírez、Hung-Hsuan Chen、Zhao- hui Wu和Lee Giles。Citeseer x：一个学术大数据集。载于《信息检索进展：第*36*届欧洲信息检索研究会议，*ECIR 2014*，荷兰阿姆斯特丹，*2014*年*4*月*13*日至*16*日。会议论文集*36*》，第311–322页。施普林格，2014年。[4] Hung-Hsuan Chen、Pucktada Treeratpituk、Prasenjit Mitra和C Lee Giles。Csseer：一种基于Citeseerx的专家推荐系统。载于《第*13*届*ACM/IEEE-CS*数字图书馆联合会议论文集》，第381–382页，2013年。[5] Behrooz Mansouri、Shaurya Rohatgi、Douglas W Oard、Jian Wu、C Lee Giles和Richard Zanibbi。Tangent- cft：数学公式的嵌入模型。载于《*2019*年*ACM SIGIR*信息检索理论国际会议论文集》，第11–18页，2019年。[6] Pei-Syuan Wang和Hung-Hsuan Chen。图对比学习在数学信息检索中的有效性。载于《基于图的信息检索方法国际研讨会》，第60–72页。施普林格，2024年。[7] Lukas Pfahler和Katharina Morik。科学文章中相关数学表达式检索的图神经网络自监督预训练。*arXiv*预印本*arXiv:2209.00446*，2022年。[8] Shuai Peng、Liangcai Gao、Ke Yuan和Zhi Tang。利用图神经网络将图像转换为LaTeX以进行数学公式识别。载于《文档分析与识别–*ICDAR 2021*：第*16*届国际会议，瑞士洛桑，*2021*年*9*月*5*日至*10*日，会议论文集，第二部分*16*》，第648–663页。施普林格，2021年。[9] Shuai Peng、Ke Yuan、Liangcai Gao和Zhi Tang。MathBERT：用于数学公式理解的预训练模型。*arXiv*预印本*arXiv:2105.00377*，2021年。[10] Yuning You、Tianlong Chen、Yongduo Sui、Ting Chen、Zhangyang Wang和Yang Shen。带增强的图对比学习。《神经信息处理系统进展》，33：5812–5823，2020年。[11] Piotr Bojanowski、Edouard Grave、Armand Joulin和Tomas Mikolov。利用子词信息丰富词向量。《计算语言学协会汇刊》，5：135–146，2017年。[12] Armand Joulin、Edouard Grave、Piotr Bojanowski、Matthijs Douze、Hérve Jégou和Tomas Mikolov。Fast- text.zip：压缩文本分类模型。*arXiv*预印本*arXiv:1612.03651*，2016年。[13] Makoto P. Kato、Kazuaki Kishida、Noriko Kando、Tetsuya Saka和Mark Sanderson。NTCIR-12报告：NII信息获取研究测试平台与社区的第十二轮评测。《*SIGIR*论坛》，50(2)：18–27，2017年2月。