·测量与评价·

基于多模态表征融合的试题相似性检测研究

史　磊　万宇哲　李　睿　张玲玲　罗敏楠　刘　欢

［摘　　要］随着信息技术与人工智能的快速发展，海量的在线试题涌现出来，试题查重检测成为教育考试数字化工程中维护命题安全与考试公平的关键环节。试题相似性检测技术是实现试题查重检测的核心任务之一。考试命题技术的不断演进，使得现有的试题呈现典型的多模态特性，包含题干、公式、示意图、知识点等多种模态信息。然而，现有试题相似性检测研究大多忽视了试题的多模态特性；此外，现有方法直接应用传统文本与视觉分析技术对公式、示意图和知识点进行处理，导致难以准确表征试题。鉴于此，本文提出了基于多模态表征融合的试题相似性检测框架，并结合向量数据库技术，实现高效的试题查重检测。首先，针对试题中公式、示意图和知识点的不同特点，本文提出了公式主干提取、示意图理解和层次化知识点预测等技术，并基于注意力机制进行多模态信息的融合，实现试题的准确表征。其次，为了满足海量试题查重检测的实时性需求，本文利用向量数据库进行试题向量表示的存储与检索，实现高效的查重

作者简介：史磊，西安交通大学计算机科学与技术学院，硕士研究生，主要研究方向为多模态学习；万宇哲，西安交通大学计算机科学与技术学院，硕士研究生，主要研究方向为试题难度预测；李睿，西安交通大学继续教育学院，主要研究方向为试题评价分析；张玲玲，西安交通大学计算机科学与技术学院，主要研究方向为机器学习、示意图理解；罗敏楠，西安交通大学计算机科学与技术学院，主要研究方向为机器学习、多模态学习；刘欢，西安交通大学计算机科学与技术学院，主要研究方向为机器学习、智慧教育。

说明：本文系“2023教育考试与评价”研讨会分论坛主题报告。

检测。

[关 键 词] 相似性检测;示意图理解;知识点预测;多模态融合

一、背景介绍

国家“十四五”规划和2035年远景目标纲要明确提出“加快数字化发展,建设数字中国”,旨在积极迎接数字时代的到来,推动网络强国的建设。随着数字化浪潮在全球范围内的蓬勃发展,教育信息化领域也随之蓬勃兴起。线下试题库和在线教学系统在不断涌现,导致大量多元化的试题数据积累。这种激增的数据量给试题设计带来了挑战,试题的重复设计和雷同现象呈逐渐上升趋势。这不仅会影响教育资源的有效利用,还在一定程度上损害考试的公平性,也会对考试的公信力产生负面影响。这一问题在高考领域尤为突出,社会培训机构频繁地从事试题押题行为,在市场上推出高考押题卷等现象屡见不鲜,这使得命题工作者在设计试题时需要特别注意,避免与已存在于市场上的模拟试题产生重复,从而增加了命题的难度。

试题查重检测是维护命题安全和考试公平的关键环节。试题相似性检测技术是实现试题查重检测的核心任务之一。考试命题技术的不断演进,使得现有的试题呈现典型的多模态特性,包含题干、公式、示意图、知识点等多种模态信息。试题相似性检测的前提是试题表征,试题表征又涉及不同的模态表征与融合。然而,现有试题相似性检测研究大多忽视了试题的多模态特性。如Huang等研究者[①]引入多任务学习机制,提出了中文教育预训练语言模型$BERT_{Edu}$,尽管其在多个数据集上优异、克服了练习中数学公式和术语的多样性,但仍然集中于文本分析,未能充分考虑如示意图、知识点等相关模态信息,从而限制了相似性检测的全面性。此外,现有方法直接应用传统文本与视觉分析技术对试题进行处理,难以准确表征

① Huang T, Li X. An empirical study of finding similar exercises [DB/OL]. (2021-11-16)[2023-10-15]. https://arxiv.org/pdf/2111.08322.pdf.

试题。例如，Tong 等人[①]提出考虑试题图文结合知识点相关信息的知识感知多模态网络 KnowNet，尽管结合了知识层次取得了一定进展，但是未充分考虑示意图的特殊性与题干公式的特殊性；类似的，Liu 等人[②]提出了一种基于注意机制的多模态神经网络模型 MANN，接受包含文本、图像和概念的多模态习题作为输入，使用传统卷积神经网络(Convolutional Neural Network，CNN)提取图像特征，但是没有进行示意图理解，知识点的层次化特性也没有考虑。

在实际应用中，试题相似性检测的需求通常涉及大规模的试题数据库，可能包含数百万或上千万道试题，这些试题多种多样，涵盖了不同知识点、不同数据类型和复杂的语义信息。传统的数据库以基础的数值或字符形式存储数据，往往无法有效地处理多模态试题数据的复杂性，也难以满足实时检索和分析的需求。同时，传统数据库的索引结构也难以有效地处理向量相似度搜索和邻近性查询任务。基于这些挑战，向量数据库应运而生。这是一种新兴的数据库形式，它以向量的形式来存储数据，并构建索引以加速相似度检索。此外，向量数据库还采用了近似搜索算法，加快向量的检索过程，以应对大规模、多模态、复杂的试题数据。

本文的研究旨在提供一种创新的试题查重检测方案，通过多模态表征与融合技术的运用，结合向量数据库的支持，来解决试题查重的复杂性。接下来，将在第二部分介绍相关工作，第三部分深入探讨试题的多模态表征融合，第四部分引入向量数据库，为高效准确的相似性检测奠定基础，第五部分对本文进行了总结与展望。

二、相关工作

随着教育信息化进程的深入，试题库和教育系统得到了飞速的发展，积累了大量多模态

① Tong W，Tong S，Hunag W，et al. Exploiting knowledge hierarchy for finding similar exercises in online education systems [C]//2020 IEEE International Conference on Data Mining (ICDM). IEEE，2020:1298－1303.

② Liu Q，Huang Z，Huang Z，et al. Finding similar exercises in online education systems [C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining，2018:1821－1830.

试题数据,包括文本题干、公式、示意图和知识点等。本文将这些数据用于试题的表征学习,并进行相似度检测与分析。

具体的相关工作可分为以下两种,即多模态表征融合与相似性检测。

(一)多模态表征融合

单模态的深度学习模型在多个任务上取得了显著的成就,但是由于其只能获取特定模态的信息,在涉及多模态关联的情况下表现平平。近年来,随着多模态数据研究的不断深入,研究者开始积极探索如何将深度学习技术应用于涵盖多种模态的数据。如Cao等人①旨在通过将图像和文本的深度特征表示映射到一个共享的视觉—语义嵌入空间,以实现图像和文本之间的密切融合和内在的跨模态的语义映射。Ma等人②设计并使用一种多模态CNN来应对图像问答任务,并证明其可以有效学习图像问题的联合表示。Radford等人③提出创新性多模态预训练模型CLIP(Contrastive Language-Image Pre-Training),核心思想是将图像和文本信息映射到一个共享的语义空间中对比学习,使得模型学到文本—图像对的匹配关系,为模型赋予了深度的多模态理解能力。

但是学界在试题表征的多模态融合方面的研究相对较少,较少关注如何充分整合试题的不同模态信息。多模态整合试题数据的方式能够有效地处理和深刻理解多异构数据,挖掘多模态数据的相关性④。李洋洋等人⑤采用协同注意力机制分别获取试题文本引导的试题图片特征和试题图片引导的文本特征,通过门控机制动态进行表征融合,进一步增强了对

---

① Cao Y, Long M, Wang J, et al. Deep visual-semantichashing for cross-modal retrieval [C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016:1445 - 1454.

② Ma L, Lu Z, Li H. Learning to answer questions from image using convolutional neural network [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2016, 30(1).

③ Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning. PMLR, 2021:8748 - 8763.

④ Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2):423 - 443.

⑤ 李洋洋,谭曦,陈艳平,等.基于多模态学习的试题知识点分类方法[J].中文信息学报,2023,37(7):143 - 151.

上下文关键语义的捕捉能力。Mu 等人[①]提出了一种双向对比表示网络 BCRNet，使用文本—图像匹配策略捕捉文本和图像之间的语义关联，一定程度上解决了现有多模态方法在基于试题表示学习中的局限性，但是这些方法并没有充分结合示意图的特殊性与预测题目内隐含的知识点等信息的融合感知，来进一步提高相似性检测的准确性和高效性。充分利用示意图的深度表征，预测试题所隐含的知识点，并进行多模态融合，变得尤为重要。

（二）相似性检测

传统试题相似性检测通常可以分为四种。第一种是基于最长的公共子序列查找算法，相似性检测即为找到两个序列中最长的公共子序列[②]。第二种是通过将内容向量化，计算向量相似度如余弦相似度、Jaccard 相似度等，这种方法一定程度上忽略了文本的语义信息，同时速度较慢。第三种方法是利用倒排索引的方式进行查重，如吴嘉楠等人[③]使用开源搜索引擎 Lucene，采用倒排索引的方式来建立索引，对试卷分题后进行试题查重，虽然该方法具有较快的搜索速度，但无法支持对部分文档的修改，并限制了索引中包含的数据量以及索引的更新频率。最重要的是，它同样不支持试题的语义查重。第四种是使用 Simhash 算法进行重复性检测，它可以粗粒度地识别试题库中的冗余部分，李莉等[④]提出多特征权重 Simhash 算法，提高重复性检测的效率，对具有抄袭行为的恶意用户进行准确识别。但是该方法存在一定的不足，即 Simhash 算法更适合长文本内容[⑤]，针对较短的试题题目表现较差，影响查重检测的准确性。

---

① Mu J, Zhang X, Du Y, et al. BCRNet: bidirectional contrastive representation network for deep multimodal learning of exercise representations in online education systems [J]. Neurocomputing, 2023:126409.

② Hunt J W, MacIlroy M D. An algorithm for differential file comparison [M]. Murray Hill: Bell Laboratories, 1976.

③ 吴嘉楠，容振邦.基于 Lucene 的试卷查重系统设计与实现[J].信息技术与信息化，2016(5):32－35.

④ 李莉，杨春艳，朱江文，等.区块链下社交网络用户抄袭识别方案[J/OL].计算机应用:1－13[2023－10－15].http://kns.cnki.net/kcms/detail/51.1307.TP.20230516.1311.008.html.

⑤ Yang S K, Chou C. A hybrid methodology of effective text-similarity evaluation [C]//New trends in computer technologies and applications: 23rd International Computer Symposium, ICS 2018, December 20－22, 2018, Revised Selected Papers 23. Springer Singapore, 2019:227－237.

随着深度学习的迅速发展和广泛应用,崭新技术和深度神经网络的不断演进,为多模态数据处理提供了更高的性能和精度。为获取包含试题语义信息的表征,沈钢[①]采用特征提取技术对试题文本进行查重统计,可以提取到文本试题的语义信息。但是基于内容属性的语义相似度计算在算法迁移、人工成本和句法结构适应方面存在挑战,Wang等人[②]提出建模教育视频关键帧中嵌入的空间和时间信息,通过捕捉片段之间语义关联的细粒度相似性测量模型VENet,利用注意机制,选择适合给定练习的融合通道来增强相似性度量,但仍有未充分考虑类似知识点的多模态试题信息的语义融合,一定程度上限制了相似性度量在试题查重领域的准确性和全面性。

因此,有必要研究多模态试题中深度语义信息的相似性检测方案,以提供更精确全面的查重分析。

三、多模态试题表征与融合

首先将不同模态的信息向量化,提取不同模态的深度语义信息,然后进行有效融合,综合多模态信息,使其具有更丰富的语义表达,为后续向量数据库进行存储和查询提供数据基础。

(一)试题多模态特性与问题定义

分析试题的多模态特性是为了将试题中不同模态的信息整合到统一的特征表示中,从而支持试题相似性计算和查重检测的准确性和效率。试题的多模态特性是指试题由不同的模态信息组成,例如一道数学题目如图1所示,题干与公式构成了对题目的主干描述,示意图作为补充信息图示化问题背景或详细信息如左下角所示,预测抽取具有与题目内容相关

① 沈钢.基于文本相似性检索技术解决命题中重题检测问题的实践:以北京市自学考试命题为例[J].中国考试,2018(3):38-42.

② Wang X, Huang W, Liu Q, et al. Fine-grained similarity measurement between educational videos and exercises [C]. Proceedings of the 28th ACM International Conference on Multimedia, 2020:331-339.

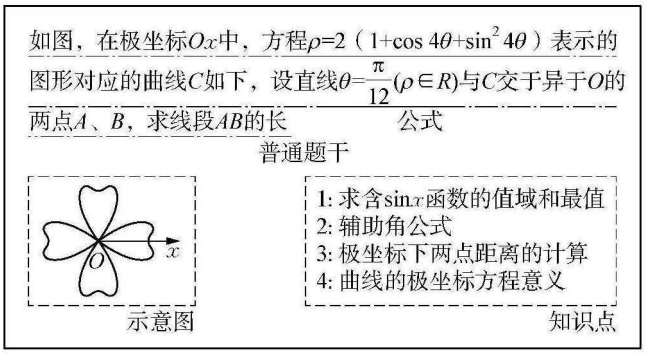


图 1　试题的多模态特性

的知识点信息如右下角所示，多重数据形式共同构成了一道试题的多模态特性。

试题相似性检测旨在对多模态试题的深度语义表征进行相似度计算，随后返回 topk 条相似度高的试题集合，此过程可以表示为以下定义：

给定一道试题 Q 作为输入，对试题进行预处理，删除 HTML 标签和标点符号等，获得题干 S，拆分题干获取题干文本 T 和对应的数学公式 LaTeX 格式 L，提取示意图 D，预测试题相关的知识点 N，目标是融合不同模态的信息来获取题目 Q 的深度语义表征，问题可以定义为：

$$split(S) \rightarrow T, L \tag{公式 1}$$

$$M(T, L, D, N, \theta) \rightarrow \mathbf{r}, \mathbb{R} \tag{公式 2}$$

$$\Gamma(\mathbf{r}, \sigma) \rightarrow \mathbb{Q}, \mathbb{S} \tag{公式 3}$$

其中 M 代表多模态表征与融合模型，θ 为其参数。r 代表试题深度语义表征，Γ 代表在多维向量空间中相似性检测的任务，σ 是 Γ 的参数，$\mathbb{Q}$, $\mathbb{S}$分别代表返回 topk 条相似性试题的试题内容和相似性评分的集合，$\mathbb{R}$= $\{Q^{Text}, Q^{Latex}, Q^{Diagram}, Q^{Notion}\}$ 依次为文本模态、公式、示意图以及概念知识点的表征集合。

（二）试题表示学习

本节将介绍试题不同模态信息的表示学习，分别为题干文本与公式（题干 S）的表示学

习、示意图理解以及知识点的预测,提取到每个不同模态的深度语义信息,为后文特征融合以及相似性检测提供可靠的模态语义表示。

1. 题干文本与公式表示

在试题表示学习的领域中,题干文本和公式的表征是至关重要的组成部分,通过对题干信息进行预处理,拆分提取其中的公式并转换为 LaTeX 格式,采用 ERNIE① 作为剩余题干文本的编码器,获得试题题干文本的语义表征 Q^{Text},众所周知,BERT 是一种基于 Transformer 模型的自然语言处理预训练模型,具备出色的特征表示能力。然而,随着研究的深入发现 BERT 在某些方面存在一定的潜在局限性,尤其是在处理中文文本时,虽然考虑了上下文语义但还缺少一定的知识信息②,因此百度提出了深度融合知识的 ERNIE 预训练语言模型,如图 2 所示,ERNIE 的编码器采用深度 Transformer,通过对词汇、实体等语义单元的掩码,实现了对知识的深度融合。采用多任务学习持续更新预训练模型,如图 2 右侧所示,在引入新任务时,使用上一阶段已训练好的模型参数进行初始化,在训练新任务时又同时与上个阶段的

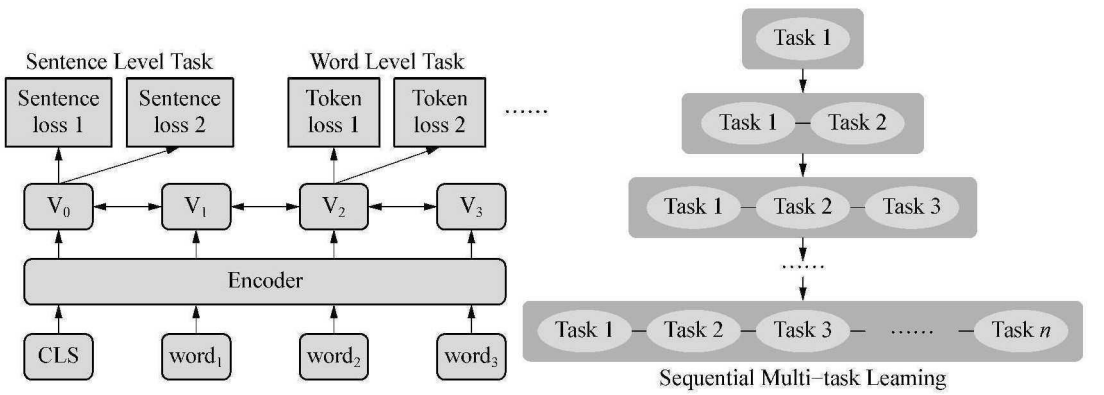


图 2 ERNIE 框架

来源:Sun Y, Wang S, Li Y, et al. ERNIE 2.0: a continual pre-training framework for language understanding [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5):8968 - 8975.

① Sun Y, Wang S, Li Y, et al. ERNIE 2.0: a continual pre-training framework for language understanding [C]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(5):8968 - 8975.

② Sun Y, Wang S, Li Y, et al. ERNIE: enhanced representation through knowledge integration [DB/OL]. (2029 - 04 - 19)[2023 - 10 - 15]. https://www.xueshufan.com/publication/2938830017.

旧任务进行同步训练，从而使得模型能够更全面地捕捉概念和语义信息。

从相似性检测的角度来看，公式中常量和变量的变化并不会引起公式本质的变化。首先，数学公式存在变量差异性，即同一数学公式可以包含不同的变量，但公式的本质语义结构是相同的。其次，常数的多样性也是一个显著特点，数学公式中的常数可以取各种不同的值，但是它们的存在并不改变数学公式的本质含义。举例来说，像 $\frac{2x^2+4y}{3}$ 和 $\frac{4a^2+2b}{6}$ 在相似性分析方面是等价的，因此首先对 LaTeX 公式进行预处理，消除表达式中的不同变量和常数，进行主干提取，如图 3 所示。

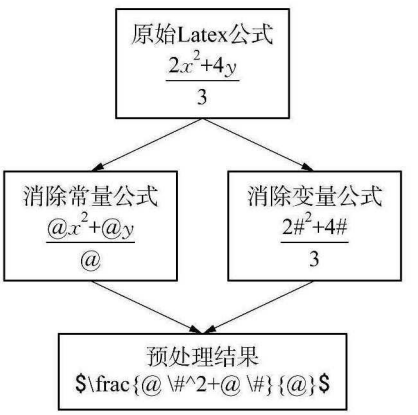


图 3　公式主干提取

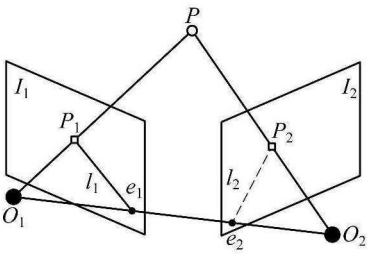
采用预训练模型 MathBERT[①] 提取公式主干特征的向量表示，充分学习 LaTeX 公式的结构和语义信息。MathBERT 是针对数学公式理解的预训练模型，使用运算符树（OPT）捕获数学公式的语义级结构信息。OPT 是一个由运算符节点和操作数节点组成的有向无环图，其中每个节点代表数学公式中的一个运算符或操作数。

2. 示意图理解

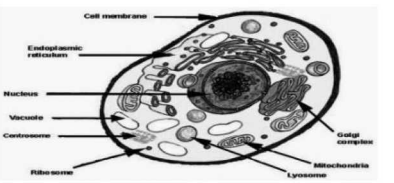
示意图是用抽象的图形符号表示系统、设备的工作原理及其组件相互关系的简图。如

① Peng S, Yuan K, Gao L, et al. MathBERT: a pre-trained model for mathematical formula understanding [DB/OL]. (2021-05-02)[2023-10-15]. https://arxiv.org/abs/2105.00377v1.

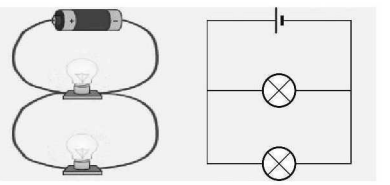
图 4(a)～(c)所示,试卷中包含几何类、生物类、物理类示意图。



(a) 几何类



(b) 生物类



(c) 物理类

图 4　不同学科示意图

多模态试题中示意图的准确理解是实现试题表征的重要组成部分。然而,与传统自然图像不同,示意图理解面临两大挑战。

(1) 小样本问题。示意图由领域专家绘制,获取代价高昂,样本有限导致现有数据驱动的模型过拟合现象严重。对自然图像数据集和示意图数据集的规模进行对比,如图 5 所示,常见的自然图像数据集规模远高于示意图数据集规模,其中最为突出的 ImageNet 自然图像数据集规模是 FoodWeb 示意图数据集的上万倍。

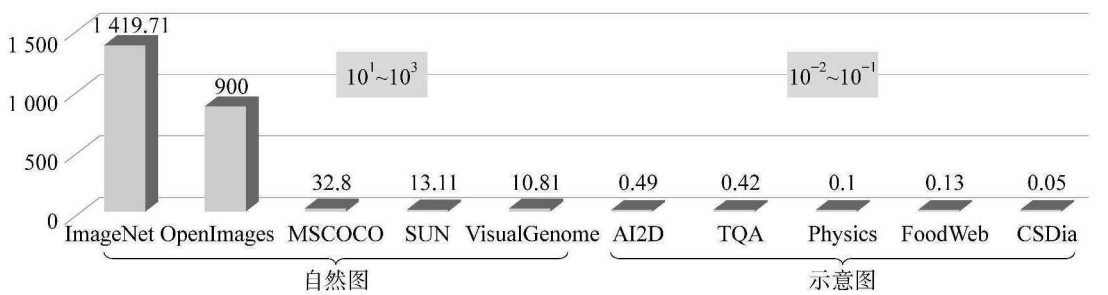


图 5　自然图像与示意图数据集规模对比

此外，对自然图像与示意图数据集中的类别样本数进行对比，如图 6 所示，自然图中不同类别样本数量较为均衡，示意图中的低频类别分布更为广泛，长尾现象显著。

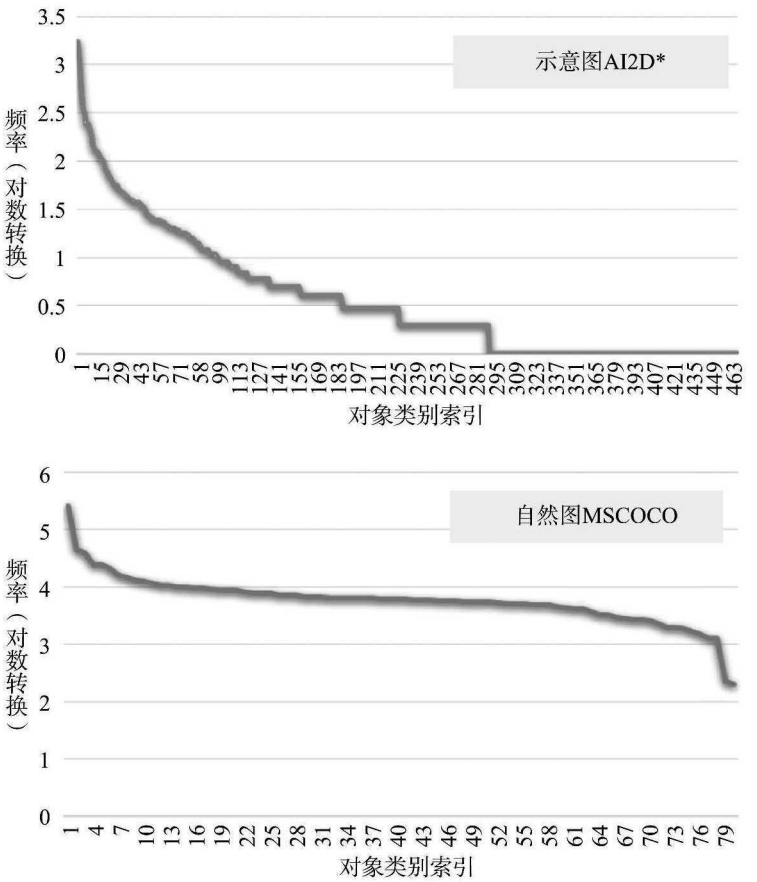


图 6　自然图像与示意图数据集类别样本数对比

（2）底层视觉稀疏现象。自然图像侧重于视觉冲击，充分展示物体的颜色及细致的纹理；示意图强调物体具体构造等学科相关的知识。从底层视觉特征进行分析，图 7 展示了自然图样本和示意图样本的颜色分布直方图。从图中可以看出，自然图的 RGB 分布较为均匀，色彩信息非常丰富；示意图的 RGB 分布在值 255 的附近较为频繁，而其他颜色值的频率非常低。这表明了示意图中存在大量的白色背景，而有用的前景信息的视觉色彩却极为稀疏。

针对示意图理解的上述挑战，本文设计了基于格式塔感知的小样本示意图表征模型，借

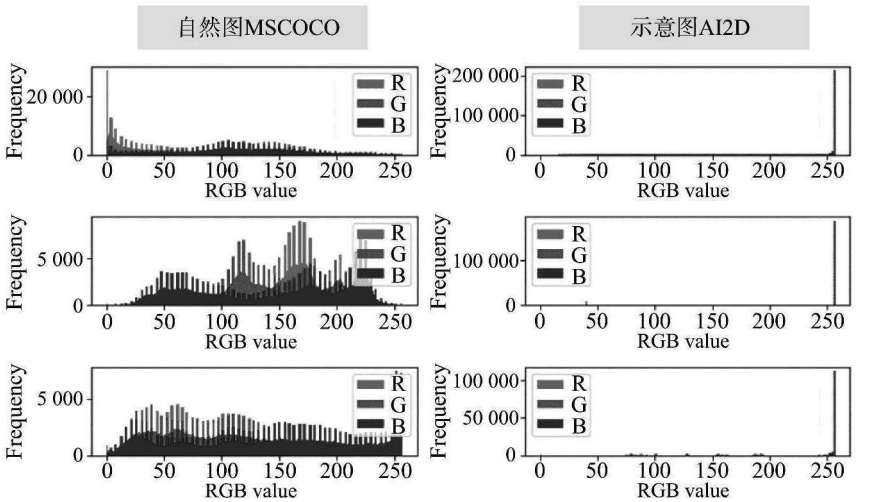


图 7　自然图像与示意图底层视觉特征对比

助示意图中对象检测任务对示意图进行理解①。人类视觉感知符合格式塔准则②,即人眼视觉系统趋于将图像中相似的、接近且具有平滑轮廓的区域感知为一个更有意义的整体。该模型能够增强示意图局部表征,实现小样本场景下的示意图对象检测。

示意图对象检测任务定义:给定示意图 D,通过特征编码器 $\phi(D)$和检测器 $\psi(\phi(D))$检测出 D 中的对象集合$\mathbb{Y}$,$\mathbb{Y}$中包含对象类别标签$\mathbb{O}$和位置坐标$\mathbb{b}$,其形式化定义如下:

$$\mathbb{Y}=\psi(\phi(D)),\text{其中}(\mathbb{O},\mathbb{b})\in\mathbb{Y}$$

模型一共包含三个部分。首先是预处理模块,从三个视角对输入的示意图进行分块处理,分别是颜色分支、位置分支以及边缘分支,提取相应分支的视觉特征。其次是格式塔编码器模块,将初始化的图块表征经过格式塔感知聚合输出局部对象的表征。具体来说,分别采用相似性、接近性和平滑性格式塔准则将三个分支的图块构建为三种子图,如图 8 所示。接下来,采用图神经网络对不同感知准则下的示意图图块进行特征更新。

最后是示意图对象解码模块,采用基于 Transformer 的目标检测框架输出示意图中对象

① Hu X, Zhang L, Liu J, et al. GPTR: gestalt-perception transformer for diagram object detection [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

② Koffka K. Principles of gestalt psychology [M]. London: Routledge, 2013.

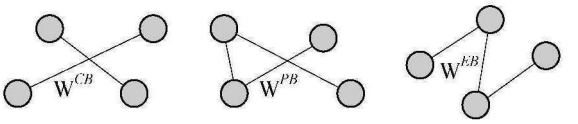


图 8　颜色、位置和边缘子图

的位置坐标和类别标签。其中，编码器的输出 $Q^{Diagram} = \phi(D)$，即为示意图的表征向量，作为后续多模态特征融合模块的输入。模型详细信息请参考论文[①]。

3. 知识点预测

知识点是试题的一个重要特征，是考生理解并解答试题时所需知识的集合，知识点的重合程度也是判断两道题目是否相似的重要维度。与题干、公式、示意图等其他特征不同，试题的知识点往往不会直接显示在试题文本和图形信息中，而是需要利用试题的文本和图形特征进行推断。这种给定一个试题(文本、图片、符号、公式等)，预测该试题涉及的知识点集合 $Q^{Notion} = \{n_1, n_2, \cdots, n_j\}$ 的工作即为知识点标注任务。

在传统方法中，知识点识别和标注工作大多由相关学科专家或教学经验丰富的教师凭借知识和经验来完成。人工标注的方法虽然准确率较高，但这种方法费时费力，且人工标注对标注工作者的先验知识依赖程度较高、主观性较强，因此其使用往往受到限制。在机器学习领域，知识点标注任务可视为多标签文本分类问题。现有对知识点进行自动标注工作的方法大多先利用 BERT 模型提取试题的文本特征，之后利用 MLP 多个二分类器根据提取到的文本特征对试题所涉及的知识点进行标注。现有方法虽然明显提高了试题知识点标注的效率，但仍面临以下三个问题：一是在多模态信息融合的试题中，模型如何自适应融合试题中的文本和示意图信息。例如，图 9(a)和图 9(b)均是初中学段的数学试题，图 9(a)试题的知识点信息全部来自题目文本信息，而图 9(b)试题的知识点主要由示意图中线条的位置关系预测得到；二是由于知识点标签具有层次化的特点，模型如何利用知识点标签的层次结构来约束多分类过程，以确保最终分类结果的准确性和完备性；三是在部分知识点标注信息缺乏或

① Hu X, Zhang L, Liu J, et al. GPTR: gestalt-perception transformer for diagram object detection [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

冗余的情况下,模型如何保证鲁棒性,缓解标签缺失及模糊的影响。

图 9　知识点的不同来源

为解决现有方法对知识点标注时面临的难点,本文构建了知识点预测模型,如图 10 所示的模型。

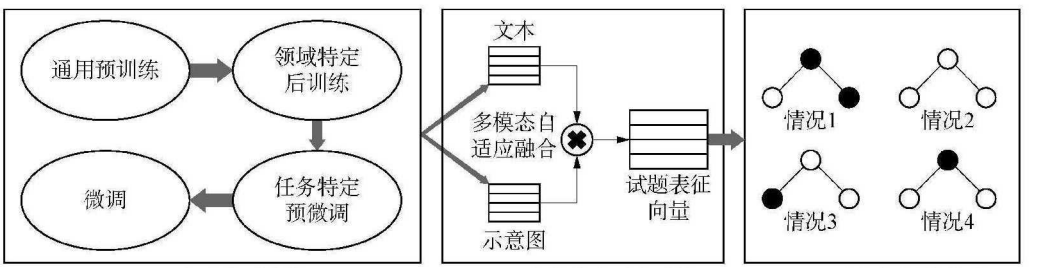


图 10　知识点预测模型

该模型包括以下三个步骤:第一步,多阶段预训练(Multi-Stage Pretraining)。多阶段预训练可以提高模型的泛化能力和表示能力,通常包括通用预训练(General Pretrain)、领域特定后训练(Domain-Specific Post-training)、任务特定预微调(Task-Specific Pre-finetune)和微调(Finetune),每个阶段可使用不同语料进行训练以提高模型各个方面的性能,使得模型逐步学习通用语言特征并适应特定领域和最终任务。第二步,多模态自适应融合(Multi-Modal Adaptive Fusion)。使用多模态自适应融合方法可以提高模型整合不同模态的信息的能力,并根据任务的需要自适应地调整权重或融合的方式,从而增强模型对试题文本、示意图信息的理解能力。第三步,多标签层次分类(Multi-Label Hierarchical Classification)。与传统的多个二分类器方法相比,多标签层次分类方法可以考虑标签结果之间的层次关系,以提高模型

的分类性能，可对试题中涉及的各层次所有知识点均进行标注。

（三）多模态融合

试题中不同模态之间存在一定的关联，为避免引入更多噪声，深度理解试题语义信息，在获取到试题异构数据的表征后，采用注意力机制进行多模态表征的融合。这种多模态融合方法的有效性已在教育领域的应用中得到验证，如陈恩红等[①]提出的 QuesNet，将异构信息通过不同的嵌入表征映射向量表示，使用 BiLSTM 通过多头自注意力机制生成统一的向量表示，深度提取多模态试题的语义信息，从试题语义理解和试题结构理解上进行深度表征，有效解决了多源异构试题表征难的问题。

如图 11 所示，在提取到不同模态信息的语义表征后，采用基于注意力机制的融合层(Attention-based Fusion Layer，ABFL)整合多模态试题信息，为输入试题 Q 学习一个统一的

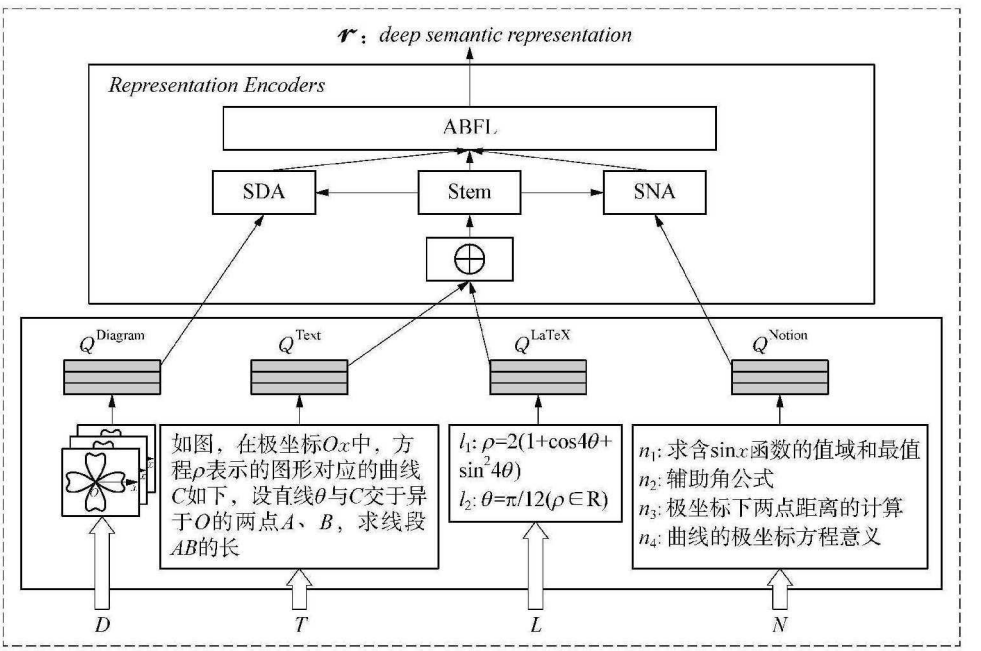


图 11　多模态融合结构

① 陈恩红，刘淇，王士进，等.面向智能教育的自适应学习关键技术与应用[J].智能系统学报，2021，16(5):886－898.

语义表示,公式表征与题干文本表征进行拼接得到融合题干表征,通过题干—示意图注意力机制(Stem-Diagram-Attention, SDA)捕捉文本与示意图之间的关联性,结合题干—知识点注意力机制(Stem-Notion-Attention, SNA)[①]来捕捉文本与知识点之间的关联性,随后通过 ABFL,得到统一语义表征 r。

其中 SDA 描述试题题干与示意图之间的注意力权值,给定试题题干表征 $S^r=(s_1, s_2, \cdots)$, s_i 为融合公式特征后每个 token 的语义表征,结合示意图特征图 $Q^{Diagram}=(d_1, d_2, \cdots)$, d_i 代表一题多图每个图的表征,LSTM 隐含层状态 h_{t-1} 拥有之前 $t-1$ 个词的信息,将其也一同参与线性变换,进行归一化后通过注意力分布得到 SDA 相关联的特征表示,类似的 SNA 协同知识点信息 $Q^{Notion}=(n_1, n_2, \cdots)$,学习题干与知识点之间的关系,具体计算如公式(4)到公式(10)所示,ABFL 层代表基于注意力机制的融合层,由 LSTM 结合和 Attention 构成,如公式(11)到公式(13)所示:

$$z_j = \boldsymbol{V}_{ad}\tanh(\boldsymbol{W}_{ad}(h_{t-1} \oplus s_t \oplus d_j)) \quad \text{(公式 4)}$$

$$\alpha_j = \frac{z_j}{\sum_{j=1}^{M} z_j} \quad \text{(公式 5)}$$

$$\boldsymbol{sda}_j = \sum_i \alpha_{j,i} d_j \quad \text{(公式 6)}$$

$$z'_j = \boldsymbol{V}_{an}\tanh(\boldsymbol{W}_{an}(h_{t-1} \oplus s_t \oplus n_j)) \quad \text{(公式 7)}$$

$$\beta_j = \frac{z'_j}{\sum_{j=1}^{L} z'_j} \quad \text{(公式 8)}$$

$$\boldsymbol{sna}_j = \sum_i \beta_{j,i} n_j \quad \text{(公式 9)}$$

$$h_j = LSTM(\sum_j (\boldsymbol{sda}_j \oplus \boldsymbol{sna}_j \oplus s_j)) \quad \text{(公式 10)}$$

① Liu Q, Huang Z, Huang Z, et al. Finding similar exercises in online education systems [C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018:1821 - 1830.

$$l_j = \tanh(W_{af} h_j) \quad \text{（公式 11）}$$

$$\gamma_j = \mathrm{softmax}(l_j) \quad \text{（公式 12）}$$

$$r = \sum_j \gamma_j h_j \quad \text{（公式 13）}$$

其中，⊕代表张量拼接，W_{ad}，V_{ad} 是 SIA 需要学习的参数，W_{an}，V_{an} 是 SKA 需要学习的参数，LSTM 输入序列每个元素均为 *sia*， *ska*， *stem* 拼接得到，W_{af} 为 ABFL 层中的参数，h_j 是 LSTM 中最后一层的输出，模型输出试题的统一表征 r。

四、向量数据库检索

向量数据库和传统数据库的差异如表 1 所示，传统数据库通过点查和范围查等方式执行精确匹配，只返回是否满足查询条件。

表 1　向量数据库和传统数据库对比

特点	向量数据库	传统数据库
数据类型	高效处理向量、嵌入式数据	处理结构化数据、文本等
检索速度	基于向量相似度快速，高效检索	通常较慢，特别是对高维数据
低延迟，高并发	是	否
查询方式	近似查找：查询结果是与输入条件最相似的	精确查找：点查/范围查 查询结果只能是否符合条件
适用领域	以图搜图、推荐系统 (大规模数据存储检索)	企业管理、数据分析等 (中小规模数据存储与检索)

向量数据库通过近似查找实现模糊匹配，输出的是概率上最符合条件的答案。向量数据库能够存储和处理海量数据，支持准确的相似性检索，完全满足试题查重检测任务的需求。

试题相似性检测即为多模态特征向量检索，向量检索是在一个给定向量数据集中，按照某种度量方式，检索出与查询向量相近的 K 个向量，即 K-Nearest Neighbor(KNN)，但由于

KNN计算量过大,通常只关注近似近邻问题。Milvus作为一种高性能的向量检索引擎崭露头角①,它是一款开源的分布式向量数据库,具备海量的向量数据存储、更新和查询能力,具有延迟低、吞吐量高的特点,能够为万亿级向量数据建立索引,用于高效管理大规模矢量数据,它支持FLAT、IVF_FLAT、IVF_SQ8、HNSW等向量索引算法,支持在向量相似度检索过程中进行标量字段过滤,以及多种相似性指标来实现高效近似最近邻搜索,从而快速查重相似向量。

设计采用Milvus作为多模态表征向量的存储检索引擎,整体流程如图12所示,多模态融合表征应用于向量相似度检测与存储,Milvus采用了基于向量的索引策略,可以更为准确地反映数据的内在结构。

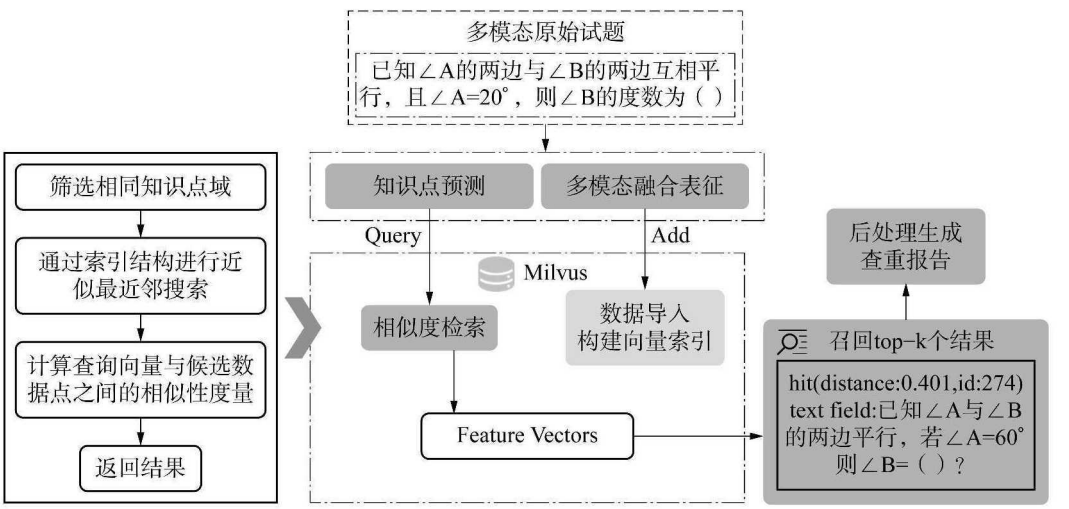


图12 整体流程

首先获取到多模态的试题表征 r 与预测的知识点信息 Q^{Notion},连接Milvus,初始化阶段批处理大量试题深度语义表征信息存储于Milvus中,Milvus会自动为批处理数据构建索引,提高搜索效率。在检索阶段中,随后应用抽取到的知识点来缩小检索空间,提供更精准的检索结果,Milvus可以作为相似度计算引擎,选用如欧氏距离(L2)、内积(IP)、余弦相似度、

① Wang J, Yi X, Guo R, et al. Milvus: A purpose-built vector data management system [C]. Proceedings of the 2021 International Conference, 2011.

Jaccard 距离等多种相似性指标来进行相似度计算，由于 Milvus 查询速度非常快，可以融合多种相似性度量方法，获取高效可靠的相似度计算方法。最后返回与查询条件最匹配的 top-k 条命中试题，并且生成查重报告，详细描述试题内容信息以及相似性信息，以满足多模态试题查重的需求。

五、未来与展望

本文提供了一种高效的试题相似性检测研究方法，通过多模态试题的表征与向量数据库的应用来满足设计需求，这有助于确保试题的独立性，避免试题的重复设计和雷同问题，从而维护了国家大型考试和整个教育领域的高质量和安全性。

方法采用深度预训练模型进行多模态的深度语义表征，可更好地理解试题中包含的文本、示意图以及其他多种信息，并且预测题目相关的知识点，结合注意力机制融合多模态特征，并且通过预测到的知识点对试题检索范围进行精确的缩放，结合向量数据库中多种试题相似度计算，在庞大的试题库中迅速找到与查询试题相似的题目，从而实现高效的试题检索。这不仅有助于查重检测，规避命题重复，还能够帮助命题工作者参考相关问题，生成更具时代背景和创新性的试题，提高考试的命题质量。

未来拟积极探索多模态表征方法的进一步优化，提高多模态表征的准确性，从而更全面、精细地捕捉试题中的语义信息；同时将加强对知识点标记的细粒度处理。准确的知识点预测对于相似性查重检测、构建知识图谱乃至提供个性化学习资源等都至关重要。未来将不断改进知识点预测算法，以提高其准确性和一致性；充分利用知识图谱，使得模型学习到具有知识信息的更深刻语义表征；开发更加智能化的可视化试题查重工具，使用户能够更轻松地理解和分析试题查重结果。

A Study on the Detection of Test Item Similarities Based on Multimodal Representation and Fusion

Lei SHI　Yuzhe WAN　Rui LI　Lingling ZHANG　Minnan LUO　Huan LIU

Abstract: With the rapid development of information technology and artificial intelligence, a vast number of online test items have been created. Similarity detection of test items has become a crucial component of the digitalization of educational assessments, ensuring the security of test item development and the fairness of examination. Test item similarity detection technology is one of the core tasks for implementing test item similarity detection. The evolving nature of test item generation technology has resulted in test items of typical multimodal characteristics, encompassing various modalities such as the question stem, formulas, diagrams and notions. However, most existing research on test items similarity detection overlooks the multimodal characteristics of test items. In addition, existing methods directly apply traditional text and visual analysis techniques to process formulas, diagrams, and notions, leading to difficulties in accurately representing test items. In view of this, this paper proposes a framework for test item similarity detection based on the fusion of multimodal representations, coupled with vector database technology, to achieve efficient similarity detection. Firstly, in addressing the unique attributes of formulas, diagrams, and notions in test items, this paper introduces techniques including formula core extraction, diagram comprehension, and hierarchical notion prediction. These methods, integrated with an attention mechanism, enable the amalgamation of multimodal information, thereby ensuring precise test item representation. Secondly, to meet the real-time demands of extensive test item similarity detection, this paper utilizes a vector database for

storing and retrieving test item vector representations，thereby achieving efficient similarity detection.

Keywords：similarity detection; diagram understanding; notion prediction; multimodal fusion