RESEARCH TOPICS

2025-12-18

AUTOMATIC ANALYSIS AND GRADING OF UTML UML DIAGRAMS

Douwe Osinga
d.r.osinga@student.utwente.nl

Supervisor
dr. ir. Vadim Zaytsev
v.zaytsev@utwente.nl

Supervisor
dr. Nacir Bouali
n.bouali@utwente.nl

Faculty of Electrical Engineering,
Mathematics, and Computer Science

UNIVERSITY OF TWENTE.

1. INTRODUCTION

UML diagrams play a significant role in computer science, as they allow for communicating system designs in a standardised format. During technical studies, students are often required at some point to make a UML diagram for a graded assignment or exam.

However, the grading of these diagrams can often be a costly and lengthy process, involving multiple paid members of staff. Therefore, the automation of this task is an interesting topic.

In this Research Topics paper, I examine the current state of autograding diagrams and propose something - TODO proposal.

2. PROBLEM STATEMENT

The grading of (UML) diagram submissions by students can often be a costly and lengthy process, involving multiple paid members of staff, which can take multiple hours of active work[1] .

The automatisation of grading diagrams could reduce the cost and time required for universities and other institutions, providing financial benefit for universities and allowing for quicker grading times. Of course, these solutions must not be worse than human grading in terms of accuracy, consistency, and fairness.

Specifically, we are interested in the automatic grading of UTML UML diagrams, a recent in-house developed diagram format of the University of Twente [1] .

2.1. Research Questions

In order to examin the feasibility of automatically grading UML diagrams, we provide a main research question (**MRQ**), supported by research questions (**RQs**).

To what extent can (UTML) UML diagrams be graded automatically?

We aim to answer the main research question with the following sub-research questions:

RQ1: What existing work exist for automatically analysing and/or grading UML diagrams?
- **RQ1a**: What correction models are employed by existing works?

RQ2: To what extent can Intended Learning Objectives be translated into different types of autograder correction models?

RQ3: To what extent are existing solutions suitable for use in autograding UTML diagrams with regards to (1) UTML support, (2) availability of source code, (3) grading transparency, (4) grading consistency, (5) fairness in grading, (6) ease of linking ILOs to grading instructions, and (7) ease of integration into the grading process?

RQ4: To what extent can suitable autograders be adjusted, extended, and/or incorporated to be able to grade UTML UML diagrams?

RQ5: To what extent do suitable autograders compare to human grading in the context of grading first-year UML exam questions?

[1] From personal experience.

3. RELATED WORK

In order to answer research questions **RQ1** until **RQ4**, we conducted a small-scale literary study, collecting works from sources such as Google Scholar[2] and ResearchGate[3]

Work [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]

More focused on interactivity: [18]

Work on AI [19], [20]

Nondeterminism of AI [21], [22], [23] + counterarg: inherent lack of transparency, risks of nondeterminism in grading (see sources) == bad because same solution might not give same grade), lack of consistency (contexxt window, importance of reducing prompt length, …)

Further proof of unreliability of using Large Language Models (LLMs) for automatic grading: "In the evaluation based on UC4, GPT deducts points for missing relationships between specified actors and use cases, but theses relationships existed in the UML use case" [24, p.13] , and "While the models would provide a final score as requested in the prompt's response format, this core often did not match the actual sum of points awarded in their criterion-by-criterion assessment.[19, p.164] . Bouali et al. identify the problem perfectly, stating that "This discrepancy can be attributed to the autoregressive nature of LLMs, where they generate responses token by token".

I believe that the observation from N. Bouali, M. Gerhold, T. U. Rehman, and F. Ahmed [19] highlights the underlying problem of using LLMs for automatic grading. Because these models are in their very essence based on predicting tokens [25] , there is no formal guarantee that grades are produced with accuracy. The fact that LLMs produce grades that correlate with human grading does not mean that this grading is done in a fair, consistent, or reliable manner.

Experience on TAs [26]

Reliability of human marking/grading in general [27]

4. TOOLS AND TECHNIQUES

Adopt existing tool(s), make own tool, what frameworks/languages, …

5. PLANNING

TODO: Graduation planning. Phases, goals per phase.

[2] *https://scholar.google.com/*

[3] *https://www.researchgate.net*

BIBLIOGRAPHY

[1] "UTML." [Online]. Available: https://github.com/andrewjh9/UTML

[2] M. Hosseinibaghdadabadi, O. A. N. Almerge, and J. Kienzle, "Automated Grading of Use Cases," in *2023 ACM/IEEE 26th International Conference on Model Driven Engineering Languages and Systems (MODELS)*, IEEE, 2023. [Online]. Available: https://ieeexplore.ieee.org/iel7/10343461/10343549/10343598.pdf

[3] O. Anas, T. Mariam, and L. Abdelouahid, "New method for summative evaluation of UML class diagrams based on graph similarities," 2021. [Online]. Available: https://www.academia.edu/download/66135135/70_22270_EM_26aug_20feb_L.pdf

[4] F. Batmaz, "Semi-Automatic Assessment ofStudents' Graph-Based Diagrams," 2010. [Online]. Available: https://www.academia.edu/download/66135135/70_22270_EM_26aug_20feb_L.pdf

[5] W. Bian, O. Alam, and J. Kienzle, "Automated Grading of Class Diagrams," in *2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, IEEE, Sept. 2019, pp. 700–709. doi: 10.1109/models-c.2019.00106.

[6] W. Bian, O. Alam, and J. Kienzle, "Is automated grading of models effective?: assessing automated grading of class diagrams," in *Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems*, in MODELS '20. ACM, Oct. 2020, pp. 365–376. doi: 10.1145/3365438.3410944.

[7] S. Foss, T. Urazova, and R. Lawrence, "Learning UML database design and modeling with AutoER," in *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*, in MODELS '22. ACM, Oct. 2022, pp. 42–45. doi: 10.1145/3550356.3559091.

[8] R. Jebli, J. E. Bouhdidi, and M. Y. Chkouri, "Assessing Students' UML Class Diagrams: a NewAutomated Solution," in *2023 7th IEEE Congress on Information Science and Technology (CiSt)* |, IEEE, 2023. [Online]. Available: https://ieeexplore.ieee.org/iel7/10409867/10409868/10409936.pdf

[9] S. Modi, H. A. Taher, and H. Mahmud, "A Tool to Automate Student UML diagram Evaluation," 2021. [Online]. Available: https://www.academia.edu/download/72488756/575.pdf

[10] N. H. Ali, Z. Shukur, and S. Idris, "Assessment System For UML Class DiagramUsing Notations Extraction," 2007. [Online]. Available: https://www.researchgate.net/profile/Zarina-Shukur/publication/253243639_Assessment_System_For_UML_Class_Diagram_Using_Notations_Extraction/links/55487af30cf2b0cf7acec2e4/Assessment-System-For-UML-Class-Diagram-Using-Notations-Extraction.pdf

[11] N. H. Ali, Z. Shukur, and S. Idris, "A Design of an Assessment System for UML Class Diagram," in *2007 International Conference on Computational Science and its Applications (ICCSA 2007)*, IEEE, Aug. 2007, pp. 539–546. doi: 10.1109/iccsa.2007.2.

[12] P. Thomas, N. Smith, and K. Waugh, "An approach to the automatic grading of imprecise diagrams," technical report, 2006. doi: .org/10.21954/ou.ro.00016046.

[13] P. Thomas, “Grading Diagrams Automatically,” 2004. [Online]. Available: https://oro.open.ac.uk/90155/1/2004_01.pdf

[14] P. Thomas, N. Smith, and K. Waugh, “Automatically Assessing Diagrams,” in *Proceedings of the IADIS International Conference on e-Learning*, 2009. [Online]. Available: https://www.researchgate.net/profile/Pete-Thomas/publication/42799920_Automatically_assessing_diagrams/links/0fcfd5060076dd8ba2000000/Automatically-assessing-diagrams.pdf

[15] P. Thomas, N. Smith, and K. Waugh, “Automatically assessing graph-based diagrams,” *Learning, Media and Technology*, vol. 33, no. 3, pp. 249–267, 2008, doi: 10.1080/17439880802324251.

[16] M. Striewe and M. Goedicke, “Automated Checks on UML Diagrams,” in *ITiCSE’11*, in ITiCSE ’11. ACM, June 2011, pp. 38–42. doi: 10.1145/1999747.1999761.

[17] N. Smith, P. Thomas, and K. Waugh, “Automatic Grading of Free-Form Diagrams with Label Hypernymy,” in *2013 Learning and Teaching in Computing and Engineering*, IEEE, Mar. 2013, pp. 136–142. doi: 10.1109/latice.2013.33.

[18] S. Foss, “AutoER: A System for the Automatic Generation and Evaluation of UML Database Design Diagrams,” 2022. [Online]. Available: https://open.library.ubc.ca/media/download/pdf/24/1.0421624/4

[19] N. Bouali, M. Gerhold, T. U. Rehman, and F. Ahmed, “Toward Automated UML Diagram Assessment: Comparing LLM-Generated Scores with Teaching Assistants,” 5220. [Online]. Available: https://research.utwente.nl/files/496461589/134819.pdf

[20] D. R. Stikkolorum, P. van der Putten, C. Sperandio, and M. R. Chaudron, “Towards Automated Grading of UML Class Diagrams with Machine Learning,” 2019. [Online]. Available: https://ceur-ws.org/Vol-2491/paper80.pdf

[21] H. He and T. Machines, “Defeating Nondeterminism in LLM Inference,” 2025, [Online]. Available: https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/

[22] M. Brenndoerfer, “Why Temperature=0 Doesn't Guarantee Determinism in LLMs,” 2025, [Online]. Available: https://mbrenndoerfer.com/writing/why-llms-are-not-deterministic

[23] B. Atil *et al.*, “Non-Determinism of “Deterministic” LLM Settings,” 2025. [Online]. Available: https://arxiv.org/pdf/2408.04667

[24] C. Wang, B. Wang, P. Liang, and J. Liang;, “Assessing UML Diagrams by GPT: Implications for Education,” technical report, 2025. [Online]. Available: https://www.researchgate.net/publication/397720325_Assessing_UML_Diagrams_by_GPT_Implications_for_Education

[25] A. F. Ferraris, D. Audrito, L. D. Caro, and C. Poncibò, “The architecture of language: Understanding the mechanics behind LLMs,” *Cambridge Forum on AI: Law and Governance*, vol. 1, pp. 1–19, 2025, doi: 10.1017/cfl.2024.16.

[26] F. Ahmed, N. Bouali, and M. Gerhold, “Teaching Assistants as Assessors: An Experience Based Narrative,” 2024. [Online]. Available: https://research.utwente.nl/files/457355611/126242.pdf

[27] M. Meadows and L. Billington, “A Review Of THe Literature On Marking Reliability.” [Online]. Available: https://assets.publishing.service.gov.uk/media/5a820a57e5274a2e87dc0d5a/0505_Meadows_and_Billington_CERP_RP.pdf