FROM HANDWRITTEN MATH TO LATEX: A DEEP LEARNING APPROACH WITH ERROR CORRECTION

Anthony Brogni
Department of Computer Science & Engineering
University of Minnesota - Twin Cities
Minneapolis, MN, USA
brogn002@umn.edu

Evan Pochtar
Department of Computer Science & Engineering
University of Minnesota - Twin Cities
Minneapolis, MN, USA
pocht004@umn.edu

Victor Hofstetter
Department of Computer Science & Engineering
University of Minnesota - Twin Cities
Minneapolis, MN, USA
hofst127@umn.edu

Ajitesh Parthasarathy
Department of Computer Science & Engineering
University of Minnesota - Twin Cities
Minneapolis, MN, USA
parth057@umn.edu

ABSTRACT

This document outlines our proposal for the CSCI 5527 Final Project. It includes an introduction to the goal of our project and the corresponding motivation, the related work used to develop our methods, a proposed approach to the implementation of the solution, and a project plan with key milestones and target dates.

1 INTRODUCTION

This project aims to address the problem of converting handwritten mathematical expressions into accurate LaTeX code, a crucial task for digitizing mathematical content. The primary objective is to design a system that can efficiently recognize handwritten math expressions and generate LaTeX code while ensuring minimal syntax errors. Our project combines computer vision, sequence modeling, and error correction techniques to improve accuracy beyond existing methods.

2 MOTIVATION

Converting handwritten mathematical expressions into machine-readable formats, such as LaTeX, presents challenges due to the complexity of symbols, multi-line expressions, and varying handwriting styles. While several models exist for handwriting recognition, they often generate incorrect or incomplete LaTeX syntax. This problem is interesting because solving it would greatly benefit fields that require digitizing mathematical notes, such as education, scientific research, and online learning platforms.

3 RELATED WORK

There is significant research on handwritten math expression recognition, particularly using convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Papers such as IM2LaTeX Kanervisto (2016) and others have achieved reasonable accuracy in handwritten-to-LaTeX conversion. Recent advances like transformers and attention mechanisms have also been used to improve sequence generation tasks. However, these models often suffer from LaTeX syntax errors in their output, which require manual post-processing.

One promising direction is MathWriting Gervais et al. (2024), a dataset for handwritten mathematical expression recognition, which provides a larger and more diverse set of handwritten math expressions than previously available datasets. We plan to leverage this comprehensive dataset for our project.

4 PROPOSED APPROACH

We propose a hybrid approach to this problem involving two major components:

1. Handwritten Image to LaTeX Conversion: A convolutional neural network (CNN) for image feature extraction, followed by a transformer-based sequence generation model for LaTeX output. We will incorporate attention mechanisms to better capture spatial relationships between symbols.
2. LaTeX Error Correction using a Fine-Tuned Language Model: We will employ a fine-tuned large language model (LLM) to post-process the generated LaTeX and correct any syntax or semantic errors. The model will refine the output iteratively, ensuring valid LaTeX code that adheres to grammar and syntax rules.

5 PROJECT PLAN AND MILESTONES

To ensure the successful completion of this project, we have outlined the following key milestones:

- **Milestone 1 (Target date 4/1/2025):** Initial setup and data preprocessing. Visualize and analyze the MathWriting dataset for suitability.
- **Milestone 2 (Target date 4/14/2025):** Implement CNN and transformer-based models for image-to-LaTeX conversion. Train models on MathWriting data.
- **Milestone 3 (Target date 4/20/2025):** Create and submit the project progress report and lightning talk assignments.
- **Milestone 4 (Target date 5/1/2025):** Develop and integrate the fine-tuned LLM for LaTeX syntax correction. Conduct end-to-end testing.
- **Final Deliverables (Deadline: 5/14/2025):** Fully functional system that converts handwritten mathematical expressions into LaTeX with error correction and final project report.

REFERENCES

Philippe Gervais, Asya Fadeeva, and Andrii Maksai. Mathwriting: A dataset for handwritten mathematical expression recognition, 2024. URL https://arxiv.org/abs/2404.10690.

Anssi Kanervisto. im2latex-100k , arxiv:1609.04938, June 2016. URL https://doi.org/10.5281/zenodo.56198.