Fine-Tuning Report: Mathematical Expression OCR Model

Introduction

This report documents the fine-tuning process of Qwen2.5-VL 3B for mathematical expression optical character recognition (OCR), focusing on converting images of mathematical expressions to LaTeX format.

Dataset Processing Challenges

LaTeX Format Inconsistencies

When processing the M^{2}E dataset, we encountered non-standard LaTeX formatting in the annotations. For example:

```
{"name": "0.jpg", "tex": "1 5 . 4 \\times 4 = 6 1 . 6 \\t 1 5 . 4 \\n \\times 4 \\n 6 1 . 6"}
```

This representation, when visualized, produces incorrect formatting:

```
1 5 . 4 \times 4 = 6 1 . 6 \t 1 5 . 4 \n \times 4 \n 6 1 . 6
```

The correct LaTeX format should be:

```
\begin{tabular}{l}
$15.4 \times 4=61.6$ \\
15.4 \\
$\times \quad 4$ \\
\hline 61.6
\end{tabular}
```

Due to these inconsistencies, we selected only three high-quality datasets for training: im2latex, MLHME38K, and HME100K.

Base Model Requirements

Image Input Specifications

Since we selected Qwen2.5-VL 3B as our base model for fine-tuning, we needed to accommodate its image input requirements. The model requires a minimum dimension of 28 pixels for both width and height. During preprocessing, we resized all images while maintaining their aspect ratios to ensure all dimensions exceeded this minimum threshold.

Fine-Tuning Implementation Notes

Quantization Settings

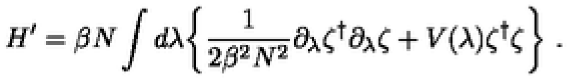
When using the Unsloth framework to fine-tune Qwen2.5-VL 3B, it's essential to set `load_in_4bit=True` during model loading. This parameter enables quantization during model export, ensuring the fine-tuned model can run efficiently on most computers with GPUs, making it more accessible to users with standard hardware.

Results Comparison

Performance Evaluation

Below is a comparison of results before and after fine-tuning:

Original Image:



$$H' = \beta N \int d\lambda \left\{ \frac{1}{2\beta^2 N^2} \partial_\lambda \zeta^\dagger \partial_\lambda \zeta + V(\lambda) \zeta^\dagger \zeta \right\} .$$

LaTeX Format Before Fine-Tuning:

H ^ { \prime } = B N \int d \lambda \left{ \frac { 1 } { 2B ^ { 2 } N ^ { 2 } } \partial _ { \lambda } \epsilon ^ { \dagger } \partial _ { \lambda } \epsilon + V (\lambda) \epsilon ^ { \dagger } \epsilon \right}

$$H' = BN \int d\lambda \left\{ \frac{1}{2B^2 N^2} \partial_\lambda \epsilon^\dagger \partial_\lambda \epsilon + V(\lambda) \epsilon^\dagger \epsilon \right\}$$

LaTeX Format After Fine-Tuning:

H ^ { \prime } = \beta N \int d \lambda \left{ \frac { 1 } { 2 \beta ^ { 2 } N ^ { 2 } } \partial _ { \lambda } \zeta ^ { \dagger } \partial _ { \lambda } \zeta + V (\lambda) \zeta ^ { \dagger } \zeta \right} \ .

$$H' = \beta N \int d\lambda \left\{ \frac{1}{2\beta^2 N^2} \partial_\lambda \zeta^\dagger \partial_\lambda \zeta + V(\lambda) \zeta^\dagger \zeta \right\} \ .$$

The fine-tuned model demonstrates significant improvement in mathematical OCR capabilities. When visualizing the LaTeX format generated after fine-tuning, it matches the original image perfectly, showing that the model can now accurately recognize single-line complex mathematical symbols and maintain proper formatting.

Future Improvements

Prompt Engineering

To further enhance the model's performance, we plan to optimize prompts by leveraging Qwen2.5-VL's visual grounding capabilities and powerful document parsing features. This approach should significantly improve the model's ability to accurately convert mathematical images to properly formatted LaTeX.