γGAMMAS: Improving Mathematical Reasoning in Vision Language Models Through Synthetic Data Generation

Ramgopal Venkateswaran, Shubhra Mishra
{ram1998, shubhra}@stanford.edu

Background

- **MATHVISTA:** Created a finegrained-benchmark for mathematical reasoning in VLMs
- **GSM8K/MATH:** Presented benchmarks for mathematical reasoning in LLMs for K-8 and high-school-level reasoning, respectvely
- **Textbooks Are All You Need:** demonstrated the importance of high-quality data by training Phi-1 using synthetic textbook-quality data

Problem Statement

Improving VLMs for math-problem solving tasks requires us to be able to generate synthetic high-quality training data *without* using VLMs in the loop, and then finetuning them with computationally efficient methods.

Contributions

- We propose GAMMAS: a pipeline to **Generate Advanced Multi-modal Mathematical And Synthetic data**. Overall, we generate 860 training samples and a validation set of size 200, comprising questions that test visual reasoning using bar charts and line plots.
- We **finetune InternLM-XComposer2-VL-1.8B**, using our generated training data
- We evaluate its performance both with the MATHVISTA "testmini" dataset and our own validation set. Within MATHVISTA, **we investigate the performance improvement on not just line plot and bar chart based questions, but also on other related categories** to understand the transfer learning capabilities of the model.

Evaluating Our Fine-Tuned Models

- Input: a mathematical question with a relevant image and multiple choice answers (letters and yes/no questions)
- Output: a multiple choice answer
- Evaluation metric: accuracy on various subsections of testmini (TestM) and our own validation set (GVal)
  - TestM-B: testmini bar chart questions
  - TestM-L: testmini line chart questions
  - TestM-B: testmini non-bar-and-line-chart questions
  - GVal-B: bar chart questions in the validation set we create
  - GVal-L: line chart questions in the validation set we create

Dataset & Methods

GAMMAS: Generating Advanced Multi-modal Mathematical and Synthetic Data

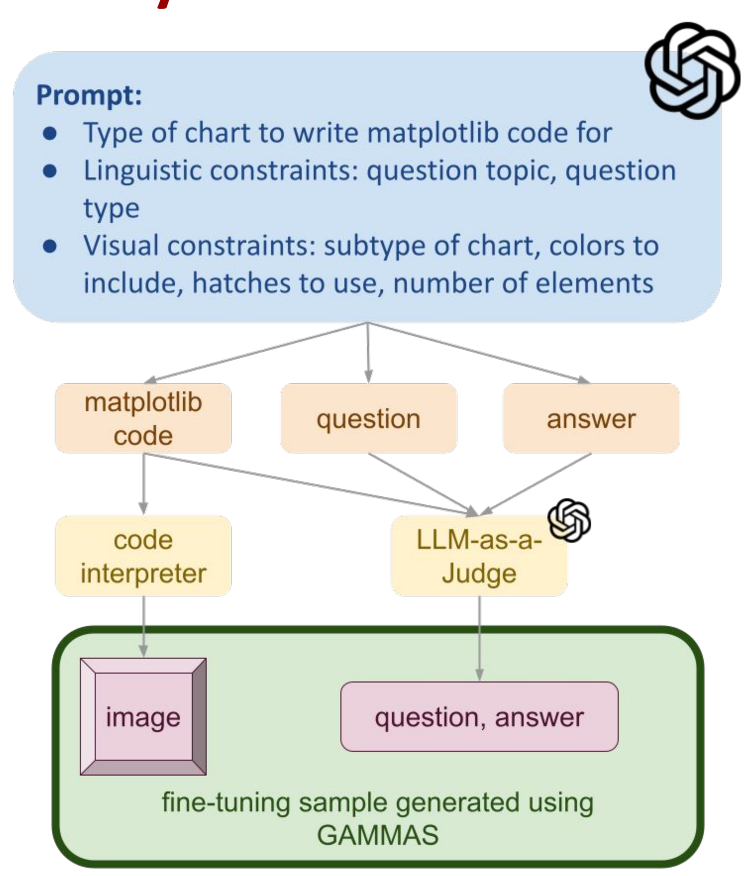

Figure 1: GAMMAS

Figure Type	Precision (%)	Recall (%)
Bar Chart	100.0	85.2
Line Plot	95.2	93.1

Table 1: Precision and recall analysis of LLM-as-a-judge on our generated data.

- Generate 1060 filtered samples (860 train, 200 test)

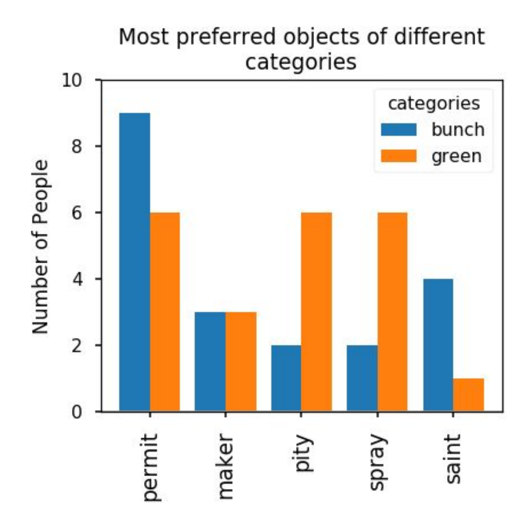

Figure 2: Example Bar Chart from MATHVISTA. Associated question: How many people like the least preferred object in the whole chart?

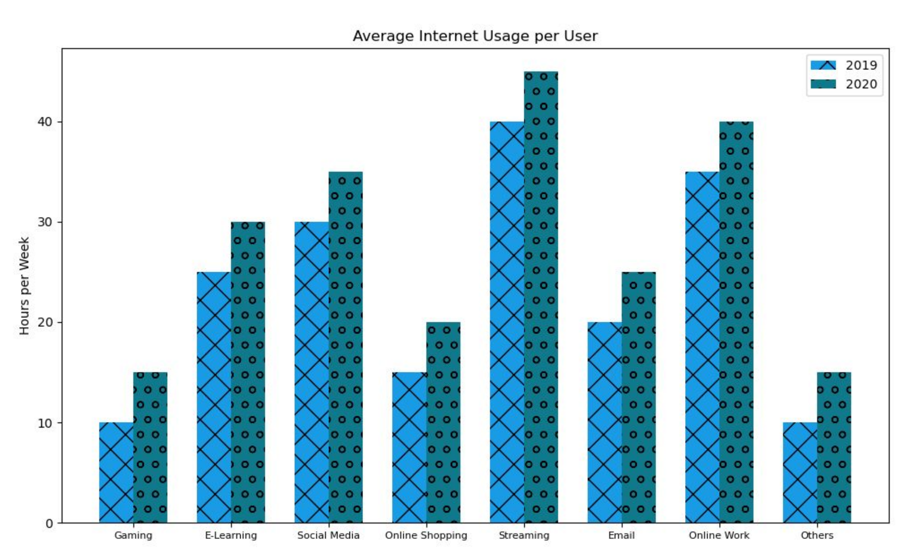

Figure 3: Example Bar Chart Generated Using GAMMAS. Associated question: According to the bar chart, what was the percentage increase in hours spent by users on Online Shopping from 2019 to 2020?

Fine-tuning

- We use LoRA and DoRA to finetune the language model (LM) component of InternLM-XComposer2-VL-1.8B
- We jointly fully fine-tune the vision encoder (ViT)
- We do not finetune the sampler which connects ViT and LM.
- LoRA
  - Fixes existing weights and learns an additional low-rank update matrix
  - Reduces GPU memory by only propagating gradients through matrices with a rank r << the ranks of the dens weight matrix
- DoRA
  - Adds an extra parameter *m* to tune the magnitude of the weight matrices
  - Increases memory consumption during training, but does not affect inference cost

Experiments & Analysis

- Training for all experiments was done on a single L4 GPU. We tuned the following hyperparameters:
  - r, the rank of the low-rank matrix in LoRA and DoRA
  - Learning rate: values ranging from 2e-4 to 8e-6
  - Adam optimizer, cosine learning rate schedule with weight decay of 0.1
  - Categories of figures that we train on
  - Whether or not we randomize options for MCQs
  - Whether or not we use filtered data from LLM-as-a-judge
  - Whether or not we modify the vision encoder

Model	GVal-B	TestM-B	TestM-L	TestM-O
Baseline	50.0	64.5	58.3	50.7
LoRA (r = 2)	64.1	-	-	-
LoRA (r = 8)	59.4	-	-	-
LoRA (r = 32)	67.2	61.3	58.3	54.4
Ablate ViT	60.9	61.3	58.3	55.5

Table2: Only Training on Synthetic Bar Charts

Model	GVal-L	TestM-B	TestM-L	TestM-O
Baseline	49.4	64.5	58.3	50.7
LoRA (32)	57.1	64.9	62.5	53.8

Table 3: Only Training on Synthetic Line Charts

Model	GVal-B/L	TestM-B	TestM-L	TestM-O
Baseline	50.0/49.4	64.5	58.3	50.7
LoRA (32)	68.8/63.6	67.7	70.8	54.2
DoRA (32)	71.9/58.4	71.0	70.8	54.6

Table 4: Training on Both Synthetic Bar and Line Charts

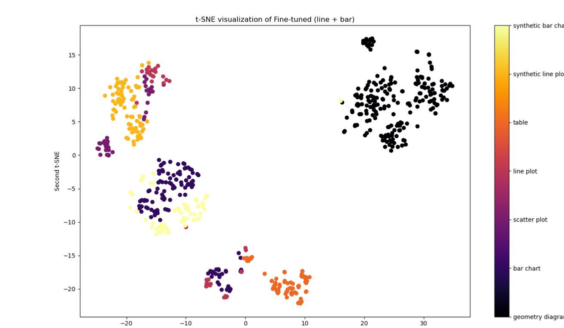

Figure 2: t-SNE

GAMMAS generates more challenging and diverse data than the current leading benchmark

- Baseline models consistently perform worse on GVal than they do on TestM
- This trend also generally holds for fine-tuned models, despite finetuning having also been done on data generated using GAMMAS

Data diversity improves model performance

- Models finetuned using only line- or bar-charts tend to either degrade in performance on TestM, or not improve at the level that models finetuned with both charts do

Impact of Vision vs. Language

- Tuning ViT together with LM improves performance on validation sets compared to tuning just LM; performance on TestM datasets is comparable for both. Two reasons could be: 1) overfitting 2) difficulty of GAMMAS

Conclusions & Future Work

- We present GAMMAS: a novel pipeline to generate synthetic multi-modal data for math problem-solving, which we use to finetune a small VLM using PeFT techniques, showing up to 17.2% improvement on math problem solving tasks
- In the future, not only do we wish to expand this work to additional tasks, but also explore techniques like ReFT, and measure the impact of fine-tuning the sampler as well