Reasoning Vision Language Model on Toaster

Backprop Battalion

Aniket Suhas Borkar
Anuj
Apoorva Gupta
Divyansh
Rajeev Kumar
Sandeep Nitharwal

Problem Statement

Vision-Language Models (VLMs) combine visual and textual data to perform multi-modal tasks like image captioning, visual question answering (VQA), and text-to-image generation.

Our project aims to explore the possibility of running VLMs on edge devices like mobile phones and personal laptops by combining various model compression techniques used in VLMs and LLMs.

Low Compute Device

+

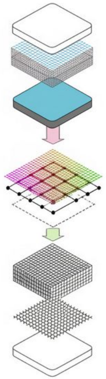
Our work

Advanced Vision capabilities on low compute Devices

We also aim to benchmark the reasoning capabilities of VLMs on tasks like VQA on charts which requires reasoning over both visual and textual information.

Existing Techniques & Challenges

Generally following techniques are employed:
1. **Distillation** – Reducing size while maintaining accuracy
2. **Compression** – Via quantization or cache optimization.
3. **Fine-Tuning & More Data** – Improving model understanding



Compressing Models

Challenges
1. **Limited Resources** – Edge devices have low compute power and memory.
2. **High Latency** – Complex VLM has processing time.
3. **Complexity** – It is hard to get better performance with low parameter Vision Models.

Low compute devices

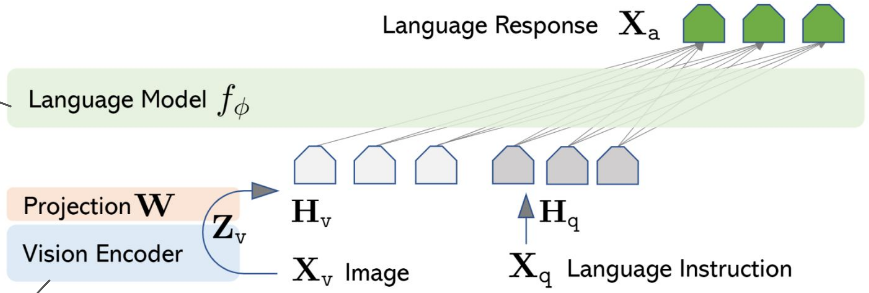
Related Work: Vision-Language Model Architecture

Basic Architecture

> Comprises a vision encoder and a language model
> Vision encoder extracts image features; language model processes text

Language Models

> Takes both visual and textual inputs to generate meaningful responses
> Uses transformers for text processing which captures long-range dependencies and contextual relationships



Extracts meaningful features from an image
Uses deep learning models like Vision Transformers (ViTs) or CNNs (Convulation Neural Network)

Training Objectives

Encourages alignment between vision and text using contrastive loss or masked language modeling

Related Work: SmolVLMs for Edge Devices

We have selected some of the small vision language models which we want to work with. While searching we looked for best capabilities at the smallest of the size.

Small VLMs developed by **Huggingface**

SmolVLM-256M: Ultra-efficient, runs on low-power devices.
SmolVLM-500M: Better performance with minimal GPU usage.
SmolVLM-2.2B: A more capable model

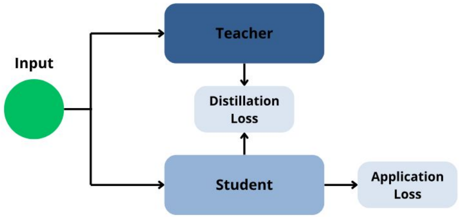
InternVL

Small VLMs developed by **OpenGVLab**(China)

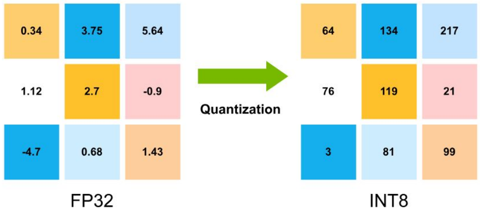
InternVL2_5-1B-MPO: an advanced multimodal large language model (MLLM) series that demonstrates superior overall performance

Key Challenges with these models: Struggles with math reasoning and planning, heavy reliance on text, limited visual reasoning

Related Work - Existing ways to compress models



Student-Teacher distillation pipeline



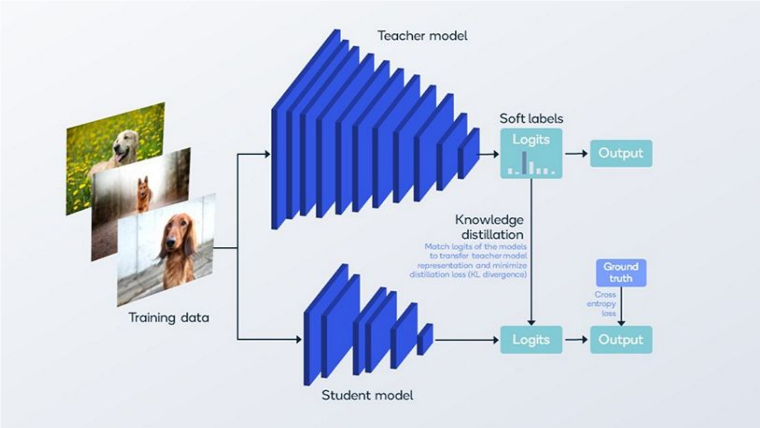
Parameter quantization

Major existing methods of model compression used in vision tasks are:

1. **Knowledge Distillation**: Involves minimizing the differences between the outputs of a larger teacher model and a smaller student model. The student model can then be used on edge devices, and has improved performance
2. **Pruning**: Removes low-relevance parameters leading to faster inference. Multiple criterion can be used to determine parameters to prune
3. **Quantization**: Conversion of compute-intensive float32 weights into computationally lighter representations such as int8. Uses QAT to train
4. **Low Rank Factorization**: This involves replacing the model weights' matrices with their low-rank approximations using methods such as SVD

Related Work - Distillation

- Teacher is pre-trained and its weights are fixed and used to train student
- While training the student, the **KL divergence** between the o/p distributions of teacher and student is added to the loss function of the student



- Penalizes the student if it goes too far from the teacher
- Techniques exist which transfer **response-based knowledge** by matching the output distributions, as well as to transfer **feature-based knowledge** by transforming features into a common feature space and minimizing the difference between the features of the two models
- Other techniques such as self-distillation (one part of the model trains other) and online distillation (where teacher is not pre-trained) exist

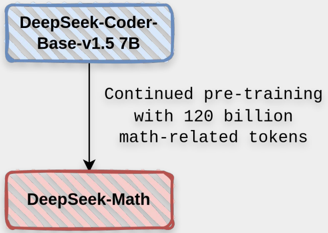
Related Work - Capabilities Improvement in Text-based Model

Improve/Fine-Tune

Idea

- Improve architecture
- Scale up parameters
- Fine-tune on specialized datasets

DeepSeek-Math



51.7 % accuracy on MATH benchmark

Tulu 3

Fine-tuned models like LLaMA or GPT variants on curated instruction-following datasets

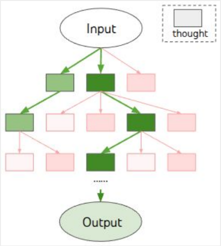
Spend More Time Reasoning

Idea

- Decompose problems into intermediate steps or “thoughts”

Tree of Thoughts (ToT)

- Structure reasoning process into a tree



Improved GPT-4’s accuracy on Game of 24 **4 ➡74 %**

DeepSeek-R1

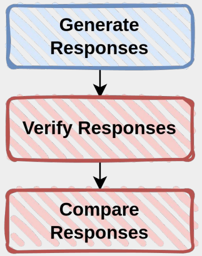
Employs GRPO - iterative refinement of outputs based on human or model-generated evaluations

Search Over More Solutions

Idea

- Generate multiple candidate responses and select the best

Inference-Time Search



Gemini v1.5 Pro surpasses o1-Preview

Method	AIME	MATH
Pass@1	1 / 15	426 / 500
Consistency@200	4 / 15	460 / 500
Consistency@1,000	3 / 15	460 / 500
Verification@200	**8** / 15	**467** / 500
o1-Preview@1	7 / 15	428 / 500

Related Work - Bringing reasoning to VLMs

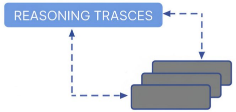
Improve VLM chain-of-thought reasoning [Oct’24]

- Uses GPT-4o reasoning data to finetune VLMs, followed by RL methods

Dataset

Reasoning Data Distillation

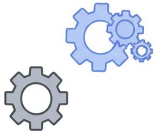
Utilize GPT-4o to augment existing VQA datasets to generate CoT instances. Includes examples from ChartQA, DocVQA, etc



SFT

Supervised Fine Tuning for CoT

Using the dataset generated in previous step, the base architecture VLM is fine-tuned to improve CoT predictions



DPO-RL

Reinforcement Learning

Using the fine-tuned model from the SFT phase, multiple responses are generated and paired up to be used for the DPO algorithm

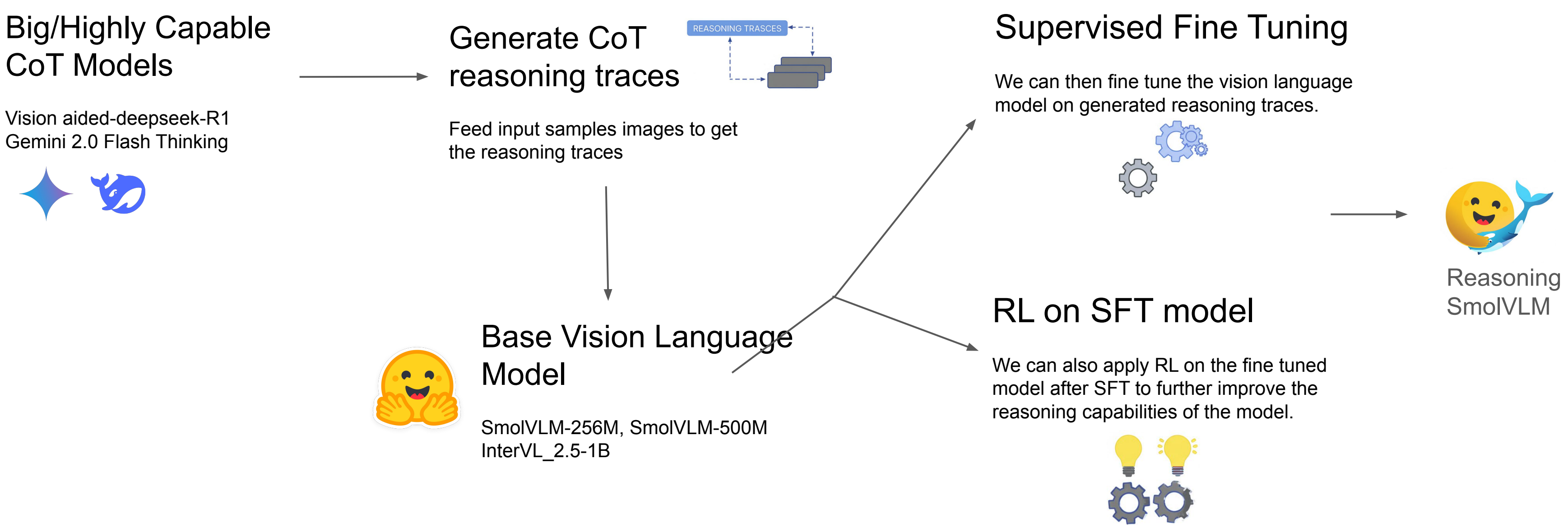
Related Work - Bringing reasoning to VLMs

R1-VL [March’25]

- In this paper, Zhang et al. design a novel RL framework called **StepGRPO** to allow multimodal models to improve reasoning capabilities
- Previous works used **outcome-based** strategies to model the rewards for RL training - this is not optimal, in particular for VLMs
- StepGRPO introduces rule-based reasoning rewards which provide **step-wise feedback** to the model, rewarding logical consistency and key-steps
- Claims improved performance on benchmarks related to Math based vision tasks

Our Pipeline

Closely building on top of our reference paper, we plan to apply reasoning CoT traces of current models on small Vision Language Models. Additionally we also want to incorporate the modern RL techniques to replace the DPO method in the paper.

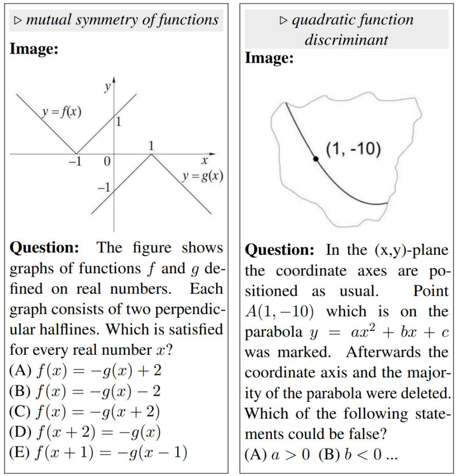


Deliverables

We plan to first do a proof of concept pilot drive of our methodologies while being extremely cost effective. Hence, we are sampling sub-dataset which requires **medium level thinking**.

How do we define **medium-level thinking** task?

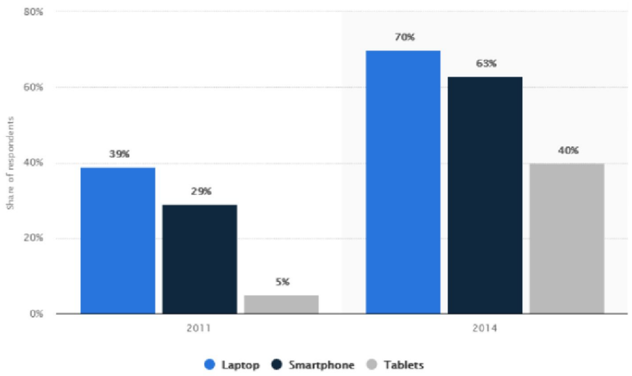
VQA dataset [1]



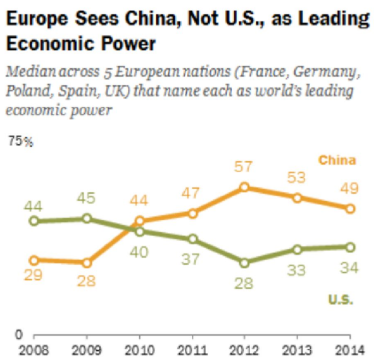
MathVision dataset [2]

Deliverables (continued)

We have selected **ChartVQA** as the benchmark consisting of **medium-level thinking** tasks.



Q6: Which digital device has most explosive increase in ownership from 2011 to 2014?
A: Tablets **Output:** Laptop



Q9: Which year shows the tiniest difference in values between China and US being seen as leading economic power across all the years?
A: 2010 **Output:** 2012

Finalizing

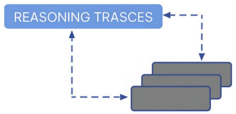
Dataset Collection

We plan to collect ~500 and 200 samples from train and test set of ChartVQA dataset

Dataset Collection

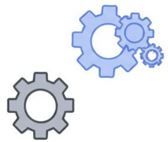
Generating Reaso-ning Traces

We will then generate the chain of thought reasoning of solving these tasks



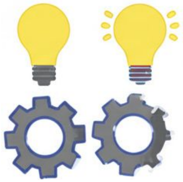
Supervised Fine Tuning

We can then fine tune the vision language model on generated reasoning traces.



RL on SFT model

We can also apply RL on the fine tuned model after SFT to further improve the reasoning capabilities of the model.



Timeline

30th March
Data Curation & Pre-processing

→

5th April
Implementing basic compression techniques and fine-tuning

→

7th April
Performance Benchmarking

→

12th April
Explore advanced techniques like RL on top of fine-tuned model

→

15th April
Finalizing deliverables and presentation

Group Contribution

All the members are equally contributing in all the project work and discussions.

Member Name	Roll No.	Papers Read
Aniket Suhas Borkar	210135	- R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization - Computer Vision Model Compression Techniques for Embedded Systems:A Survey - ScienceDirect
Anuj	210166	- Mathematics Visual Instruct Tuning: Mavis - MiniCPM-V: A GPT-4V Level MLLM on Your Phone
Apoorva Gupta	210179	- Visual Instruction Tuning - Analysis of Knowledge Distillation on Image Captioning Models
Divyansh	210355	- DeepSeek-Math - Tree of Thoughts
Rajeev Kumar	210815	- Sample, Scrutinize and Scale: Effective Inference-Time Search by Scaling Verification - ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning
Sandeep Nitharwal	210921	- DeepSeek-R1 - Tulu-3

Thank You!