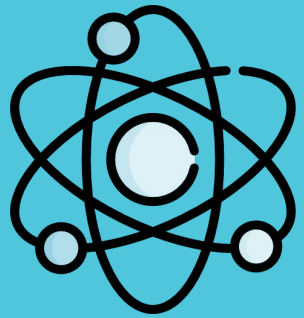
Harvard John A. Paulson
School of Engineering
and Applied Sciences

ScienceTutor



October 2023

AC215 Project Midterm Presentation

Sijia Li, Ziqing Luo, Yuqing Pan, Jiashu Xu, Xiaohan Zhao

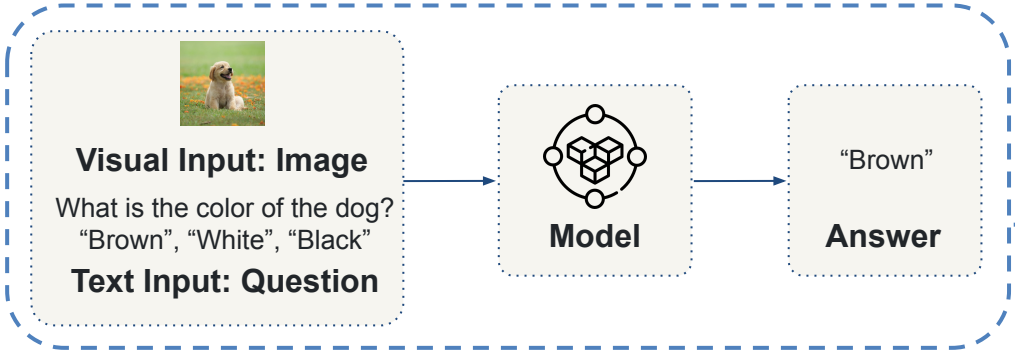
Project Overview

Background and Motivation:

- **Growing interest in multi-modal models and their applications**
- **Visual Question Answering (VQA)**
  - Provide answers to questions about input images in natural language

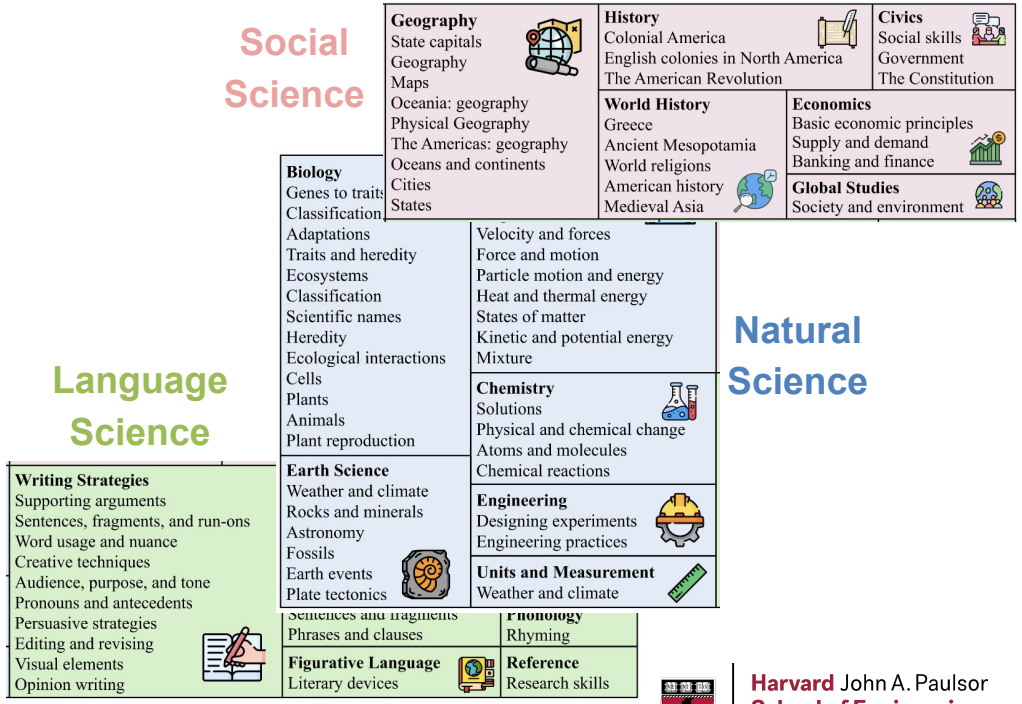
Project Goal:

- **Develop an educational application**
  - Provides answers to science questions to children via a chatbot and potentially other features
- **Technical objectives:**
  - Collect VQA data, finetune LLaVA on the science domain, design and deploy UI, build ML workflow on GCP



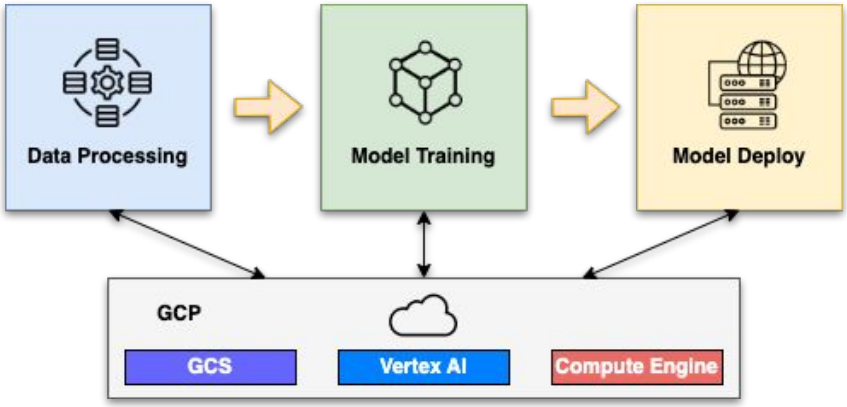
Data Source: ScienceQA

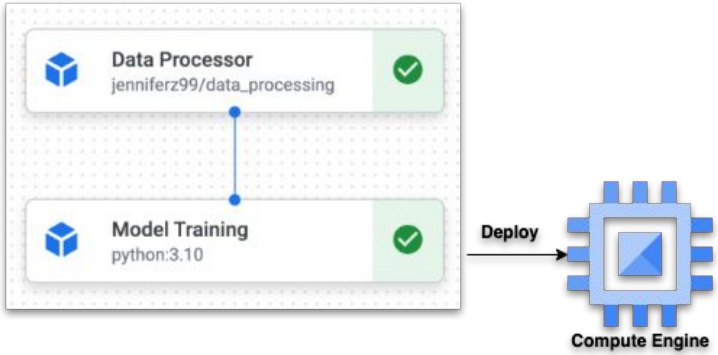
- **~21k multimodal multiple choice questions from elementary and high school science curricula**
- **Relevant attributes:**
  - Image, question, choices, answer, subject, topic, category
- **Highly relevant to our project**
- **High quality, mature VQA dataset**



[Link to Dataset]

Current Progress



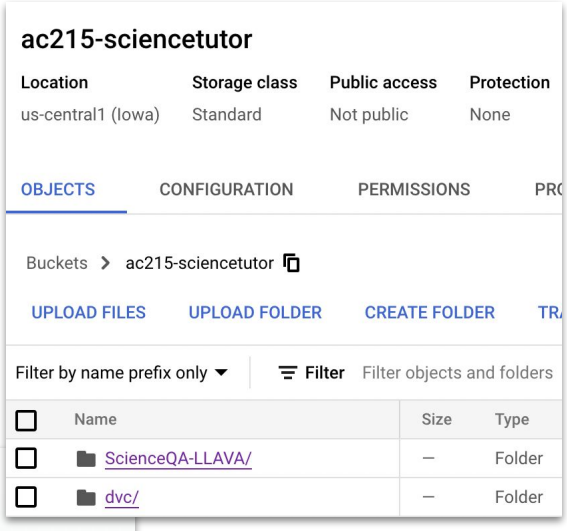


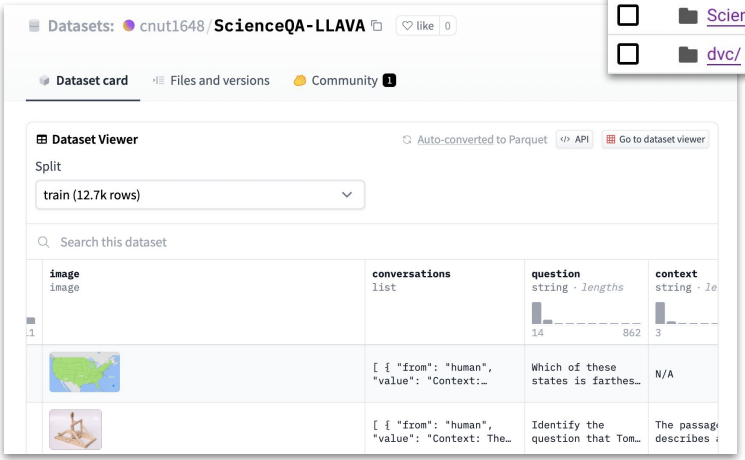
Implement containerized data management and model training pipeline

Build ML workflow with Vertex AI and Compute Engine

Application Pipeline - Data Processing

- **Prepare Data in LLaVA Format**
  - Extract information like questions, context, and choices
  - Create conversations between human and AI
- **Upload to GCS for data version control**
- **Upload to Hugging Face**



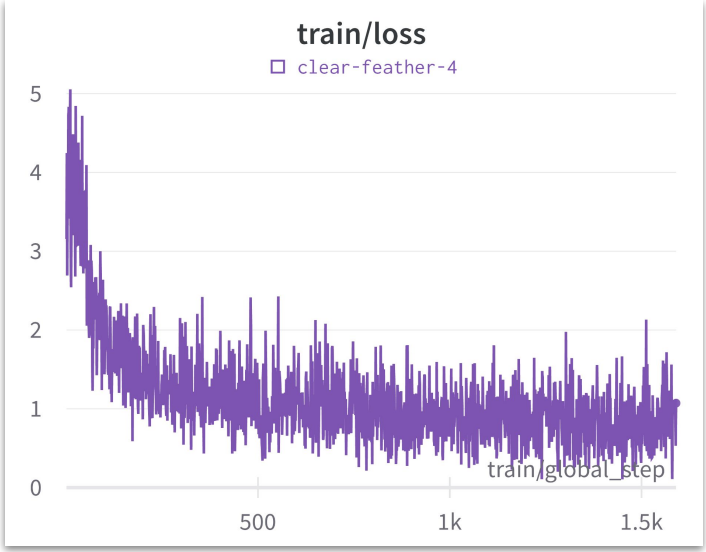


Harvard John A. Paulsor
School of Engineering
and Applied Sciences

Application Pipeline - Model Training

Training Steps:

- **Fork** the **LLaVA** repository to cnut1648 / LLaVA
- **Customize** the model to get compatible with preprocessed **LLaVA-ScienceQA** dataset
- **Train** 1 epoch and report to **wandb** by running **LLaVA/llava/train/train_mem.py**
- **Upload** model checkpoints to **GCS** / Hugging Face

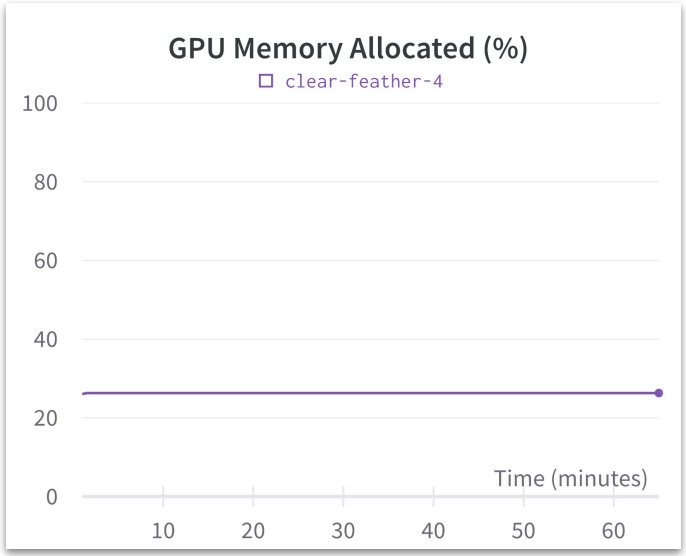


W&B Logging of Training Loss

Application Pipeline - Model Training

Optimization techniques to reduce memory usage:

- Choose smaller model (LLaVA 7B instead of 13B)
- Deepspeed ZERO-2 for multi-GPU
- LoRA (Low-Rank Adaptation)
- bf16 (bfloat16 floating-point format)
- Gradient checkpointing
- tf32 (TensorFloat-32 precision format)



W&B Logging of Memory Allocation

Application Pipeline - Model Deployment & Inference

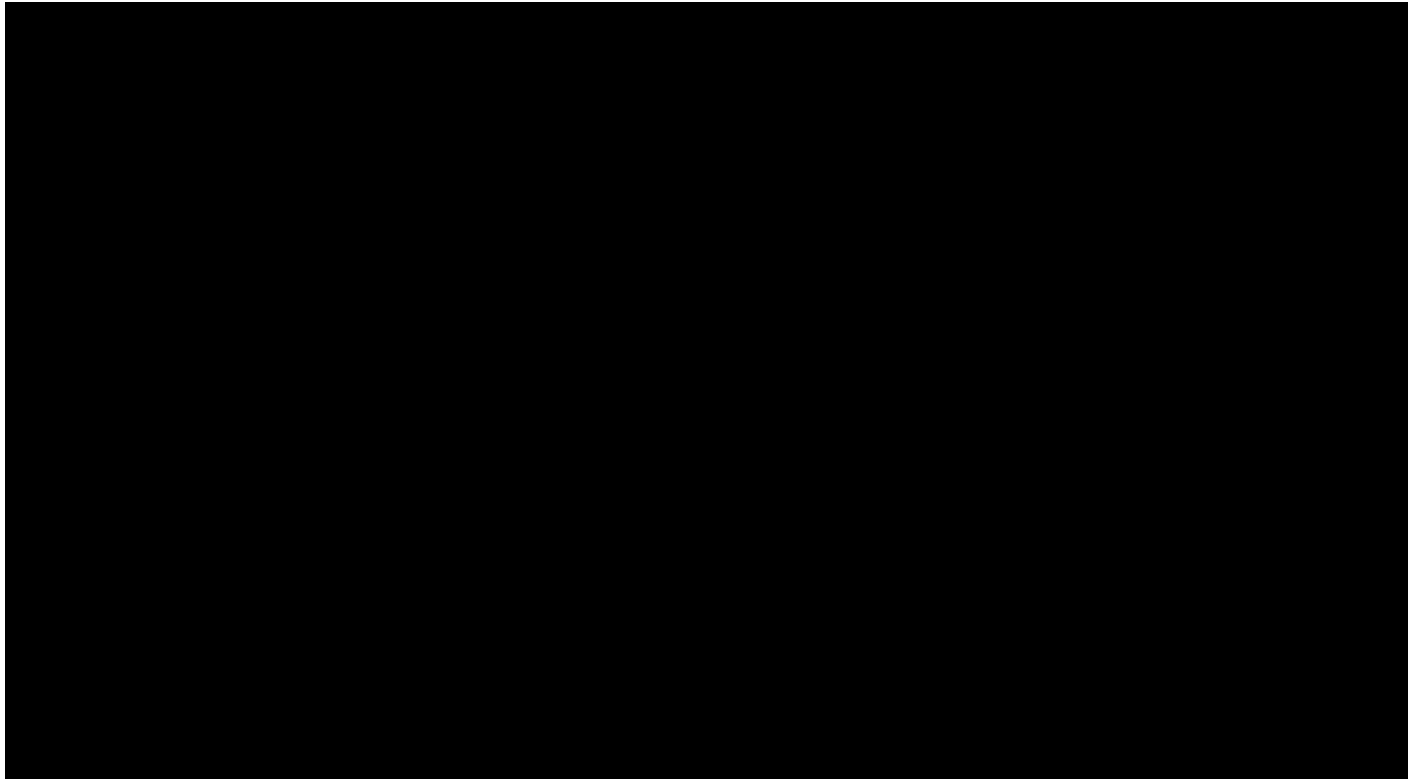
Deployment:

- **Vertex AI is not suitable**: Vertex AI builds an API endpoint for a model but we have our **own service and API** (a web server)
- Customize **web_server** Docker container and push image to **DockerHub**
- Deploy the container in an instance of **Google Compute Engine**

Inference:

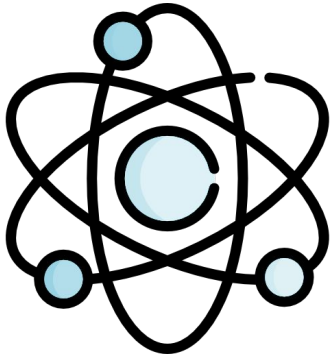
- Utilize **multi-GPU** inference and load the **4-bit quantized** model
- Achieve **accuracy 65.53%** on the testset of **LLaVA-ScienceQA**

Live Demo



Next Steps

- Customize UI with gradio.app
- Integrate potential **extensions**:
  - Quiz me 
  - Recommend Resources 
- Set up **CI/CD pipelines** to facilitate users' access to the app
- Scale with Kubernetes and deploy with Ansible

Science Tutor 

Harvard John A. Paulson
School of Engineering
and Applied Sciences

Thank You!

Agenda

1. **Project Background and Overview (50s)**
2. **Walk-Through of Current Progress**
   - **Overview: Application Pipeline Flow; Vertex AI Pipeline Flow (25s)**
     - Data Processing (25s)
     - Model Training (30s)
     - Model Deployment & Inference (30s)
   - **Live Demo (40s)**
3. **Next Steps (30s)**
   - **Chatbot Web App**
   - **Potential Extensions**

Project

Background an

- **Growing in**
  models an
- **Visual Que**
  - Provi
    input

nal application

s to science
dren via a chatbot
enhance
gs

Visual Input: Image

What is the color of the do
"Brown", "White", "Black

Text Input: Question

VQA in
omain!

+ New Chat

Does this passage...

What is the capital of...

Is the following trait...

What is the probability...

Select the organism in...

In this food chain, the...

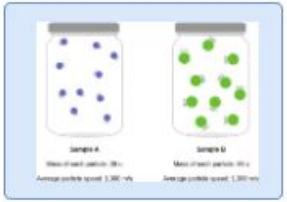
Compare the average...

Which organ controls...

Quiz Me!

Recommend Resources!

Science Tutor



Compare the average kinetic energies of the particles in each sample. Which sample has the higher temperature?

"neither; the samples have the same temperature", "sample A", "sample B"

"sample B"

Can you provide some explanation?

The particles in both samples have the same average speed, but each particle in sample B has more mass than each particle in sample A. So, the particles in sample B have a higher average kinetic energy than the particles in sample A.

Add Image

Type your question...

Application Mockup Design

(Subject to Change)

Harvard John A. Paulsor
School of Engineering
and Applied Sciences