AIM825 Course Project

Multimodal Visual Question Answering with Amazon Berkeley Objects Dataset

Deadline: 15th May 2025, 11:59PM

Overview

This assignment involves creating a multiple-choice Visual Question Answering (VQA) dataset using the Amazon Berkeley Objects (ABO) dataset, evaluating baseline models, fine-tuning using Low-Rank Adaptation (LoRA), and assessing performance using standard metrics. Your model would finally be evaluated on a hidden dataset and hidden metric.

Constraints/ Restrictions

The following constraints are to be followed -

1. Training Compute - Students must train the models on free cloud GPUs - Colab or Kaggle. Kaggle is preferred as it allows for multi-GPU training with 2x16GB GPUs leading to a total of 32GB GPU memory. Do keep in mind you need the memory to load the model as well as do the required gradient updates.
2. Model Size - The final model submitted should not be more than 7 billion in parameter size.
3. Team Size - Students can form inter-section/intra-section teams of size at most 3. Please follow the same team structure with members which you formed for the first mini-project.

Dataset

Amazon Berkeley Objects (ABO) Dataset

- **Contents**: 147,702 product listings with multilingual metadata and 398,212 unique catalog images.
- Link to dataset - Here. In this please use the small variant in the downloads section so that you don't have size issues (The original variant is 100GB in size vs the small

variant which has 3GB). The small variant has metadata as csv and images in 256x256 format.

Objectives

1. **Data Processing:** Understand how to leverage the metadata and linked images using which you will be able to continue to Point 2.
2. **Data Curation**: Generate a single word answer VQA dataset using multimodal APIs or on-device models. For eg: For a red bag, you can have a question curated like - What is the color of the bag? Ans: Red.
3. **Baseline Evaluation**: Assess off-the-shelf VQA models without fine-tuning.
4. **Fine-Tuning with LoRA**: Apply LoRA to fine-tune selected models.
5. **Evaluation**: Measure performance using Accuracy, F1 Score, and propose additional metrics.

Detailed Instructions

1. Data Curation (7 Marks)

- **Tools**:
    - **Gemini 2.0 API**: Utilize for multimodal data processing. Refer to Google AI studio to get started (check out how to use the python API).
    - **On-Device Models**: Consider using Ollama for local model deployment. Google AI for DevelopersOllama+3Medium+3Introduction | 🔗 LangChain+3. You can use the latest version of LLaMa, Qwen and other Open Source Multimodal models on devices with faster inference to quickly curate dataset. The choice is up to you to use API based models or on-device servings.

- **Process**:
    - Select a subset (if you want) of the ABO dataset suitable for your project scope.
    - Design prompts to generate questions and a single word answers per image. For a given image, you are allowed to generate multiple types of questions based on the dataset covering questions which one can answer just by looking at the image.
    - Ensure diversity in question types and difficulty levels.

- Document the data curation process, including tools, prompts, and preprocessing steps.

2. Baseline Evaluation (4 Marks)

- **Models**:
  - Use pre-trained VQA models (eg: BLIP-2 / CLIP / VilBERT etc.). Hugging Face Forums
- **Evaluation**:
  - Run inference on the curated dataset without any fine-tuning.
  - Record performance metrics as a baseline for comparison.
  - You can also experiment with limited training data or reduced training iterations to create variant baselines.

3. Fine-Tuning with LoRA (7 Marks)

- **Approach**:
  - Apply LoRA for parameter-efficient fine-tuning of selected models. Medium
  - Focus on models with fewer parameters to optimize resource usage. Submissions with smaller models used may receive additional points.
  - You can also explore additional model compression techniques like Quantization or model speedup/optimisation techniques like Flash Attention, KV Cache for faster inference to reduce iteration time. Do note if you do anything interesting in your report for bonus marks.
- **Documentation**:
  - Detail the fine-tuning process, including model selection, training parameters, and any challenges encountered.

4. Evaluation Metrics (2 Marks)

- **Standard Metrics**:

  If you perform direct token comparison then you can convert your task into a correct (True) / in-correct (False) then you can define a few standard metrics like -
  Accuracy: Proportion of correctly answered questions. Directly compare tokens
  F1 Score: Harmonic mean of precision and recall.

- **Additional Metrics**:

- Check literature and discover some other metrics like - BERTScore, BARTScore etc. Please use this too to gauge the performance along with token-level analysis
- Propose and justify any additional metrics that could provide deeper insights into model performance.

5. Iterative Improvement

- Repeat steps 1 to 4, refining your dataset and models based on evaluation results to achieve optimal performance before the deadline.

Deliverables

1. **Git Repository**:
   - All scripts and notebooks used for data curation, model training, and evaluation.
2. **Comprehensive Report**:
   - **Data Curation**: Detailed explanation of the dataset creation process, including tools and prompts used.
     Model Choices: Rationale for selected models and any alternatives considered.
     Fine-Tuning Approaches: Description of the fine-tuning process, including LoRA.
   - **Evaluation Metrics**: Analysis of model performance using chosen metrics.
   - **Any additional contribution/ novelty**
3. **Inference Script**:
   - A script capable of loading the trained model and performing inference on new data. A draft of the script structure expected will be released on LMS as the deadline approaches.
4. **Curated Dataset:** Provide the QA dataset that you created using the multimodal models in a CSV format. We request you not to attach the image again in the final submission just the metadata of Question, Options, and the correct Option is fine.

Evaluation Rubric

Component	Marks
Data Curation	6
Baseline Evaluation	3
Fine-Tuning with LoRA	7
Evaluation Metrics	2
Functioning Inference Script	2
Hidden Set Performance	5
Total	**25**

Note: The hidden set performance will be assessed using a separate dataset and metric not disclosed to students.

Additional Resources

- **Gemini 2.0 API Documentation**: Google AI for Developers
  Ollama Model Library: Ollama
- **LoRA Fine-Tuning Guide**: Medium
- **BLIP-2 Fine-Tuning Example**: Hugging Face Forums
- **vLLM Guide:** https://docs.vllm.ai/en/latest/
- **Model Quantization and PEFT:** https://huggingface.co/docs/peft/en/developer_guides/quantization