Deep Learning Blog Post
CS4240
Group - 27

April 14, 2024

Student Name and No.	Responsible for
Naga Subhash Malladi (6006388)	Reproduction + Ablation study of the model
Darryn Biervliet (4466624)	Reproduction + Ablation study of the model

Table 1: Individual Contributions

1 Introduction

In this post we are discussing the reproducing of the work done in the paper "Improved Baselines with Visual Instruction Tuning" to the LLaVA model through the visual instruction tuning, that enhanced the performance of the model, this is done by using a specific configuration of the CLIP (Contrastive Language–Image Pre-training) with ViT-L-336px (Vision Transformer Large with 336px resolution inputs) coupled with an MLP (Multilayer Perceptron) projection along with this training of the model is done on the specific academic oriented Visual Question Answering (VQA) data has lead to this performance improvement and made it effective across 11 benchmark tests (For the 13B model), along with this an ablations study is performed by training the model on the small language model like gemma.

2 Understanding of the original paper

The following are the reproducible aspects of the paper

1. Model Configuration and Data Efficiency: The modifications made to the LLaVA model, including the use of CLIP-ViT-L-336px with an MLP projection, are clearly defined and could be replicated if the exact configurations and parameters used are detailed in the paper.

2. Dataset: The use of 1.2M publicly available data for training is another reproducible aspect, especially since it emphasizes public data. The reproducibility here would depend on the availability of this dataset or the details provided to compile a similar dataset.

3. Training Process: Completing training in about a day on a specific hardware setup (single 8-A100 node) is notable. Reproducing this would require access to similar hardware and understanding the training protocol used, including any optimizations or specific settings.

4. Benchmarking: The achievement across 11 benchmarks suggests that the benchmarks themselves are known and likely publicly accessible. Reproducibility in this area would entail applying the model to these benchmarks following the same evaluation methods the authors used.

3 Architecture

The LLaVA model, as described in the information you provided, is essentially composed of three core components:

- *Large Language Model (LLM)*: This serves as the backbone for comprehending user instructions and generating responses. The LLM is responsible for linguistic understanding and output generation based on the inputs it receives, which have been contextualized through the vision-language connector.

- *Vision-Language Connector (MLP)*: The connector, specifically an MLP (Multilayer Perceptron) in the improved LLaVA architecture, bridges the gap between the visual and textual domains. This MLP replaces the simpler linear projection originally used, enhancing the model's ability to understand and translate between visual inputs and their linguistic representations. It's a key component for aligning the vision encoder outputs to the language model's understanding.

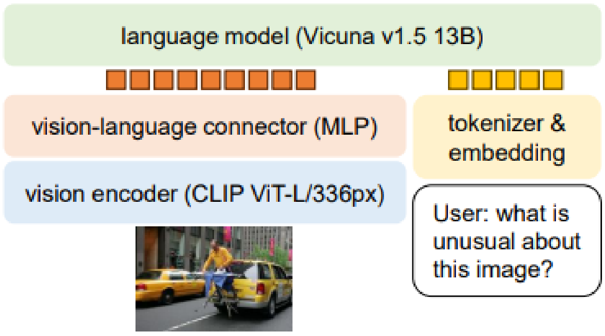

Figure 1: Model Architecture

- *Vision Encoder (like CLIP)*: A pre-trained visual backbone, such as CLIP (Contrastive Language-Image Pre-training), encodes visual features from input images. CLIP, known for its robust performance in understanding a wide range of visual concepts in a way that's alignable with textual descriptions, provides the visual understanding necessary for the model to process and reason about images.

This architecture enables LLaVA to handle complex multimodal tasks, including detailed descriptions, conversational-style question answering, and complex reasoning involving visual inputs, with remarkable data efficiency and performance.

4 Reproduction Methodology

This section discusses the systematic approach employed to replicate the findings presented in the paper, focusing on the LLaVA 1.5 model. Our replication process is structured around a two-stage instruction-tuning procedure, augmented by an ablation study to assess the impact of different language models on performance.

4.1 Procedure:

1. **Stage 1: Feature Alignment Pre-training**
   - Utilized the `pretrain.sh` script to update only the projection matrix based on a subset of the CC3M dataset.
2. **Stage 2: Fine-tuning for Specific Use Cases**
   - Employed `finetune.sh` for end-to-end fine-tuning, adapting the model to distinct scenarios:
     - **Visual Chat**: Enhancing daily user interaction capabilities.
3. **Stage 3: Ablation Study**
   - Substituted the original Vicuna 7B LLM with smaller models like Gemma and using benchmark scripts for comparative analysis.

4.2 Resources and Adjustments:

Scripts from the LLaVA 1.5 GitHub repository (`pretrain.sh` and `pretrain_xformers.sh`) were foundational tools in our replication effort. Adjustments to hardware limitations were made via `zero3_offload.json` configurations for CPU offloading, ensuring model fine-tuning despite GPU constraints.

5 Challenges in Reproduction Process

The replication of LLaVA 1.5 was met with several challenges, primarily hardware-related, impacting our ability to fully train the model as intended.

5.1 Major Challenges:

1. **Hardware Limitations:**
   - The `pretrain.sh` script aims to train the 13 Billion parameter LLaVA model, encountering memory constraint errors due to our hardware setup. This necessitated adjustments to the training protocol to accommodate resource limitations.
   - Inference of the 7B parameter model is still quite slow. Doing the standard Vqa v2 benchmark, which is one of the benchmarks that the authors used to evaluate their model, would take 13 hours to run on our hardware. This made it difficult for us to compare the results of both our fine tuned and reproduce model to the original paper. For bench marking, we decided to do a quick inference

test, to get a feel of how the model would perform. This would give us a quick indication of whether the model is capable of understanding images, and if it could perform question answering over the images.

2. **Memory Constraints:**

   - The `pretrain_xformers.sh` script, intended for a smaller 7 Billion parameter model, also presented challenges. While more manageable, it still required strategic resource allocation to proceed without interruption.

5.2 Solutions and Workarounds:

Resource Management: Implementation of `zero3_offload.json` for deepspeed configurations allowed for CPU offloading. Although this approach slowed the training process, it enabled the completion of fine-tuning stages within our hardware constraints. For the pretrainig and finetuning of Gemma, some changes to the original (pre)training script had to be made. These were minor chances as a result to the the mandatory version update to use the Gemma model. Besides these minor fixes we did not deviate from the original implementation.

5.3 Implications:

These hardware and memory challenges underscore the importance of resource scalability and efficiency in reproducing large-scale multimodal models. They highlight the need for adaptable methodologies that can accommodate varying hardware capabilities without compromising the integrity of the reproduction process.

6 Observations and Results

6.1 Observation of the Ablation Study

In the ablation study, we opted to replace the Vicuna 7B language model with the smaller Gemma model. This decision was primarily driven by the desire to explore the performance capabilities of smaller language models and to assess their training efficiency. Training and fine-tuning Vicuna 7B required approximately five days, whereas Gemma completed both processes within a single day. This significant reduction in training time suggests that smaller language models like Gemma could offer substantial efficiencies, particularly in applications where rapid deployment or frequent retraining is necessary.

6.2 Results

A comprehensive report of our training runs can be found here. Notably, we see that during pretraining, the gemma model performed worse than the baseline Vicuna 7B model. The model seems to be less capable in interpreting the image captions pairs. During fine tuning however, we see a strong reduction in the loss during training, indicating that we are improving on the question answering capabilities of the model. We could not afford to do extensive parameter experimentation is the scope of this project, nor could we afford to iteratively experiment with different architectures of the projection layer.

7 Inference

The inferences are done for both the models using the CLI inference script.

Figure 2: Input image for the CLI inference test

1. Conversation with Vicuna v1.5 7B

```
USER: what can be seen in the image?
ASSISTANT: In the image, we can see a large building with a sign on it, a tree, and a train station.
USER: Which letter is colored blue?
ASSISTANT: The letter "U" in the image is colored blue.
```

Figure 3: Benchmark results comparing the Vicuna and Gemma language models.

Our very limited finetuned model shows great initial capability of QA over images. Not only did it recognize we were in a public transport area but it was also able to accurately answer the follow up question. These results are (weak) evidence of a successfully finetuned model.

2. With Language mode (Gemma 2B)

```
Human: What can be seen in the image?
Assistant:  In the image, there are two street signs and a red building in the image.</s> is visible in the image, with a
 red and white color scheme.</s>ing the building is a large, green, and blue.</s>, along with the tall building, suggests
 that the area might be a busy area or a public event. The presence of the building and the presence of the building sugg
ests that the area is likely a public transportation or a public space.</s>-way, the presence of a building and a buildin
g suggests that the area might be a part of a public transportation or a public space.</s>-shaped clock can be seen
Human: which letter is colored blue?
Assistant:  A large building with a large building, a building with a clock tower, and a building is located near a train
 station.</s>, on the other hand, are likely to be a part of a train station or a public area.</s>, the presence of a clo
ck tower and a clock tower suggests that the area is likely a public or a public area.</s>, the presence of the clock tow
er and the presence of the building suggests that the area is likely a public or a public area where the building is bein
g used.</s>'s presence and the presence of a clock tower in the background suggests that the area might be a part of
Human:
```

Figure 4: Benchmark results comparing the Vicuna and Gemma language models.

Our Gemma model performs significantly worse. Although it seems to have some vague understanding of the situation, it hallucinates a lot of details. Furthermore, it is not capable of answering or understanding follow up questions. Despite our fining tuning efforts, the finetuned model seems to poorly on unseen data.

8 Conclusion

From the inference results we can conclude that our finetuned Vicuna-7B is capable of performing basic QA over unseen images. This hints towards a successful reproduction of the original paper. For our ablation study, we tried to substitute the Vicuna-7B model with the more lightweight Gemma-2B. During pre-training we found that the training loss plateaued significantly higher than Vicuna. The cause of this remains unclear, although speculatively, it could be due to the lack of expressively of the 'smaller' model, a sub optimal choice of the projector layer, or a bad hyper parameter choice. During the finetuning stage we did observe a significant reduction in loss. Our hypothesis is that, although we observed some improvement during finetuning, the base capabilities of image understanding of the pretrained Gemma-2B severely limited its final answering capability. Regardless of our unsuccessfully attempt we would not discourage any future efforts of similar substitutions since our results can only be considered as preliminary due to insufficient time and computational resources.