Week 9 Report

Thomas Deatherage
NFHM/BioCosmos
18 October 2024

Time Slot

1) What progress did you make in the last week?

- Weekly meetings with GaTech NFHM collaborators & UF team.
- Shared results from fine-tuning experiments on florence-2
- Adding Ollama and InternVL support to the codebase. Learned a few things I'll relate below

2) What are you planning on working on next?

- Working with Roman to Unicom Clip with one of the bio-focused datasets.
  - One of us will work on captioning images in dataset with an LLM
  - One of us will work on Unicom-style clustering

3) Is anything blocking you from getting work done?

- Nope.

Abstracts & Summaries

Lit review replaced by journal club:
https://humanaugmente-e7j6563.slack.com/archives/C07R7SK1VQU/p1729114900884069

Scripts and Code Blocks

To complement the work of Ben Yu and Bree Shi, this week I set about integrating an LLM directly into the codebase. However, I had a couple of false starts doing this.

My first step was using Ollama. Ollama is nice because it serves as a wrapper for multiple LLMs. This is nice because it allows developers to quickly experiment with different models using a consistent interface. Although, InternVL is not one of the default models included in

the repo, Ollama allows pulling in custom models, so long as they conform to GGUF format. InternVL is not in GGUF format. However, it's technically possible to convert LLMs to GGUF. Unfortunately, this does not readily seem to be the case for InternVL. Although I did try. In short, multi-modal models are *not* currently supported by Ollama. And trying to add that support is currently beyond my skillset. More details can be found in these two issues on the llama.ccp and internVL repos.

According to the second issue, Ollama support is being actively worked on and is currently "in progress". No timetables yet, but I've subscribed to both issues. In the meantime, I'm going to set aside trying to integrate InternVL into our repo. Although I could integrate it without Ollama – and will do that if Ollama-support takes too long – it would be much better to have it.

If you try the conversion yourself, you can see specifically where the error is:

```
$python llama.cpp/convert_hf_to_gguf.py internvl2-8B-hf --outfile
internvl2-8b.gguf --outtype q8_0 --verbose

INFO:hf-to-gguf:Loading model: internvl2-8B-hf
ERROR:hf-to-gguf:Model InternVLChatModel is not supported
Model conversion attempt completed. Output file: internvl2-8b.gguf
Conversion may have failed or produced an unusable output.
```

In short, vision (and likewise, multimodal) models are not supported by Ollama.

Flow Charts/Diagrams

Nothing new to show here.

Documentation

Nothing new this week.

Results Visualization + Proof of Work

Little to visualize. However, I have the code in place to support Ollama. Whenever InternVL is available (which I'm hoping *does* happen; there's high-demand for it apparently), we'll be ready to run it directly in our codebase. In short, we just need to rebuild with:

```
# ./devcontainer/compose.yml
  ollama:
    build:
      context: ../
      dockerfile: ./ollama/Dockerfile # Or directly from the ollama image
    volumes:
      - ollama-data:/root/.ollama
    ports:
      - 11434:11434
    networks:
      - default
```

Next Week's Proposal

- Unicom-ification of clip with one of the bio-related datasets, e.g., Arboretum, Tree of Life, VLM4Bio, etc.