Shane Aung
Joshua Larson
Samuel Alaniz
Project Proposal
ECE 382V, Spring 2025

Title

Human Activity Recognition using Small Vision Large Language Models

Description

Sensing humans in their homes is a tricky issue that balances privacy and possibilities of improving health monitoring, security monitoring, and smart home automatons. Traditionally HAR has relied on a variety of sensors to detect human activity. Those including vibration, EMI, sound, light intensity, and WiFi. Camera's have always been an option, but privacy issues surrounding sending home video data to the internet to be processed was too intrusive. Luckily with the advancements of Large Language Models, their compute requirements have significantly reduced. With quantized version of models now running on phones and raspberry pi's. In essence, keeping your data local. This proposal is intended to explore HAR with quantized version of Multi-modal LLM's.

Motivation

Recent innovations in Large Language Model quantization and distillation have brought the possibility to run LLM's on the edge. This can bring a world where you can simply ask the LLM "what is the user doing in the image" and you'll get back a response that is context aware and specific to the user. This brings about a world where an on edge LLM can help care for our elderly, IE know if they get hurt or aren't taking their medicine. It can ensure security of your home by monitoring the exterior for suspicious behavior. Even building patterns of life, and recognizing anomalous behavior of suspicious people around your home. It can recognize when a party is over and can turn off the lights. We hope to shed light on current capabilities and short comings into this space for LLM's.

Related Work

LLM's that can process visual data are also known as VLM's. Similar work has been done with HAR and VLM's, although there was not an emphasis for running on the edge. [1] Also, in that paper, multiple VLM's were used and coordinated with another text only LLM's. They are models that can move pixels into text, with some general reasoning behind them. They are trained on large corpses of images and text, and are able to answer questions about a photo or describe a photo for you. [2] We plan on testing on similar datasets that other research papers have referenced, like NTU RGB+D. [3]

Resources Needed

- **Laptop:** Apple MacBook Air, M2, 16GB RAM, 10-Core GPU
- **Laptop:** HP Envy x360 (Windows 10), 12GB RAM, GPU: UHD Graphics 620, i7-8550U CPU
- **Laptop:** Apple MacBook Pro, M1 Pro, 16GB RAM, 10-Core
- **Smartphone:** iPhone 14 Pro, A16 Bionic, 6GB RAM, 6-Core GPU
- **Smartphone:** iPhone 15 Pro, A17 Bionic, 6GB RAM, 6-Core GPU
- **Raspberry Pi:** 16GB RAM, GPU: VideoCore VII, Architecture: Quad-core 64-bit Arm Cortex-A76

Expected Timeline

There are 8 weeks between the due date of the project proposal and the due date of the project report. I will break the timeline down into 2 week intervals.

- **Weeks 1 - 2** Setup Dev Environment, Model Selection, Begin Data Preparation
- **Weeks 3 - 4** Finish Data Preparation, Start Implementing System, Begin Testing, Project Update
- **Weeks 5 - 6** Refine Data Pipeline, Get More Data, Final Testing
- **Weeks 7 - 8** Final analyses, Write Report and Presentation

References

[1] Harsh Lunia. *Can VLMs Be Used on Videos for Action Recognition? LLMs Are Visual Reasoning Coordinators*. 20 July 2024, `https://www.researchgate.net/publication/382459094_Can_VLMs_be_used_on_videos_for_action_recognition_LLMs_are_Visual_Reasoning_Coordinators`.

[2] *A Dive into Vision-Language Models*. (n.d.). Available at: `https://huggingface.co/blog/vision_language_pretraining`. Accessed 2 March 2025.

[3] A Review on Human Activity Recognition using Vision-Based Method. (2017). *Journal of Healthcare Engineering*, Vol. 2017, pp. 1–31. Available at: `https://doi.org/10.1155/2017/3090343`.