Exploring the Viability of Utilising Multi-Modal GPT Models with Local Hardware for Image Text Detection

Richard Finlay Tweed MSci
Department of Futile Research, No Associated University

2024-03-14

Abstract

This paper evaluates the viability of using multi-modal GPT models with local hardware acceleration for text detection in images. Despite initial optimism, our results indicate that, like our attempts at Ceilidh dancing, this approach was doomed from the beginning. We conclude that current GPT models do not effectively understand images, a finding that should surprise few.

1 Introduction

Remember the first time you tried to explain Generative Pre-trained Transformers (GPTs)[1] to your parents? That's how we felt trying to make GPT models understand images on a local android phone—sorry, computer. Our journey into the abyss of multi-modal learning was fuelled by a mix of naive hope and a profound misunderstanding of our own research capabilities.

2 Methodology

We utilised the latest in multi-modal GPT models, specifically OpenAI's ChatGPT GPT-4. Our local hardware setup consisted of a Pixel 6 Pro running a custom developed image recognition software [2]. We had a serverless Golang binary [3] running in front as a load balancer because only HTTPS with valid certificates on port 443 were supported.

2.1 Architecture

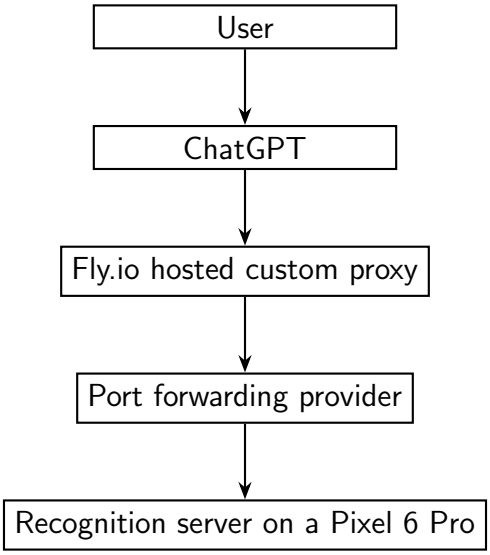

Figure 1: Communication Flow from ChatGPT to the Recognition Server.

2.2 Image Text Detection: Successful Meme detection

The chosen library worked well on our screenshots of Mastodon memes, completing the detection in under a second. This detection consumed no water (unlike commercial operators [4]) and was powered by our personal solar panel so caused no scope 1 CO2 emissions[5]. It also fulfilled our data residency requirements, as the text detection was performed in our residence's living room. The fact the images got sent to OpenAI first, and via an intermediary, should be ignored.

3 Results: Missing Data

While we convinced the GPT that our server exists, and that it has an API worthy of its use, we entirely failed to get it to send any image provided. It preferred to send empty request bodies regardless of how many riches offered or penalties threatened. From this we inferred that this multimodal GPT doesn't actually understand images, and there's some supporting middleware that displays the images when the model wants to present them, rather than them being in the context in full. Some others had some luck tricking the GPT in order to send their files, mostly via base64 encoding but we were not so fortunate[6].

4 Discussion: Were We Mad To Try This?

Yes. Yes, we were. It became abundantly clear that expecting current multi-modal GPT models to understand and interpret images on local hardware was like expecting a tractor to be able to drive through Camden market.

5 Conclusions: A Reflection on Our Overambitious Dreams

This investigation into the application of multi-modal GPT models for image text detection on local hardware was a fun failed experiment. Our conclusion — that current models lack a true understanding of images — is a testament to our excessive imagination and underwhelming execution. We await the day when more advanced models emerge, models that do not just see images but understand them, embracing their complexity with the grace of a thousand bees. Until then, we recommend doing text recognition the old fashioned way, with our eyes.

Acknowledgements

The author would like to express their gratitude to several individuals and machines who made this research possible:

- Diana Licheva, for her patience and dry humour throughout the research process.
- The reviewer(s) for skimming through this nonsense.
- ChatGPT, for making it far faster to create a LaTeX document, for generating a template and for giving this research a reason to exist.
- The kettle, for providing hot water for tea every time it was required.

References

[1] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[2] Richard Finlay Tweed. *TextRecogServer*. URL: https://github.com/RichardoC/TextRecogServer.

[3] Richard Finlay Tweed. *ImagePassthroughServer*. URL: https://github.com/RichardoC/image-passthrough-sigbovik-2024.

[4] Pengfei Li et al. *Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models*. 2023. arXiv: 2304.03271 [cs.LG].

[5] World Business Council for Sustainable Development and World Resources Institute. *The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard*. World Business Council for Sustainable Development, 2004. ISBN: 9781569735688. URL: http://pdf.wri.org/ghg_protocol_2004.pdf.

[6] Marc Breaux. *Unable to upload files from a custom ChatGPT session via an API action*. https://community.openai.com/t/unable-to-upload-files-from-a-custom-chatgpt-session-via-an-api-action/508636/5. Accessed: 2024-03-21. 2024.