A Journey of OpenAI GPT Models

The talk begins by giving an overview of the abilities and functions of Natural Language Processing (NLP) AI models. In particular, NLP models possess the ability to create probability distributions to identify sentences that are grammatically and contextually correct, as well as filling in blank words in a sentence. The talk then explored the progression of NLP technology, going from Google's Transformer model to the present ChatGPT model, based off GPT 3. The GPT series of models was specifically focused upon, and in particular their learning methodology evolution across the GPT models.

Even with so many parameters in the ChatGPT model, there are still many limitations. One of the more well known being its tendency to make up facts when answering questions sometimes, likely due to it only being concerned with generating text to conform to the pattern of an appropriate response. There are also biases related to some responses, due to the data it was trained upon. Furthermore, it is also possible to circumvent some of the policies set by OpenAI by rephrasing banned questions in certain ways. Overall, this was a good overview talk into the current state of NLP technology, and the history that led to today.

Transformer, GPT and ChatGPT

In the last few years, AI has exponentially outperformed human performance in text comprehension. Every year, NLP model sizes and parameter count grows by 10 times on average. This talk explained and offered an insight to the improvement of AI learning methodologies throughout history.

The GPT series of models all uses an encoder/decoder transformer. The transformer tried to tackle the problem of representing the semantic information of each word in a sentence. It does this by combining text and positional embeddings. The transformer uses multi-headed self attention to analyse many aspects of the input text in parallel to provide better contextual understanding between the concepts in the input. The GPT models then introduced feedforward and masked multihead attention in GPT 1, and added layer norm in GPT 2 to improve efficiency. Finally, GPT 3 introduced sparse attention amongst other improvements. Newer models such as the IntructGPT and ChatGPT added a human feedback system to further improve their capabilities.

Overall, this is an excellent overview into the evolution of the GPT series models, though the content was difficult to follow along at times for someone relatively new to the artificial intelligence field.

Fooling Deep Learning in Computer Vision and Addressing it with XAI

In this talk, we are shown ways of how a convolutional neural network (CNN) model can be manipulated and exploited by a malicious actor, and the potential solutions to address these vulnerabilities with explainable artificial intelligence (XAI). Currently, we refer to a neural network as a "black box", because it is impossible to determine how the model functions on a holistic level. Because of this, it is possible to cause unwanted behaviour by introducing perturbation such noise into an image, thereby causing incorrect classifications. Further, it is also possible to introduce a "backdoor" into the model by training on a poisoned dataset where a specific symbol is used as a trigger for a malicious behaviour.

One method of addressing these vulnerabilities is appending a few perturbation rectifying layers to the model to remove the artefacts. Apart from that, it is also possible to use XAI to detect backdoors in a model with explanations.

This talk was very interesting as it elaborates on content taught in CITS1003, by for instance explaining why a Label Universal Adversarial attack works on a whole class of image classification at a time. The research of solutions to resolving and hardening the security of AI models will prove to be critically important in the near future once AI technologies become more widespread in our lives.

From images to text & text to images

This talk described the possibility and applications of connecting models with differing modalities together. There are many potential applications in this technology, such as the ability to improve search engine queries, and the possibility for visually impaired people to have live captions in the real world. There are also medical applications of allowing diagnostic images to be analysed and described in text to a diagnostician.

An example of this technology can already be found in Microsoft Office's suggested caption feature for images, where a caption can be automatically generated from an image. This works by using a transformer to encode text embeddings from a convolutional neural network to image embeddings, which can then be decoded with a recurrent neural network.

This presentation was engaging and explained the concepts in a way suitable for beginners to the top. I am excited by the prospects that this technology can bring, but am also weary of the potential dangers if not implemented properly and securely. The medical applications in particular can be a high value and critical target that will be of great interest to potential malicious actors.

Fine-tuning GPT-3 on textbooks to build a tutorial bot for a level 2 class at UWA

This talk discussed how a Microsoft Teams chat bot was built to simulate a tutor for a level 2 unit at UWA. The chat bot in question leverages OpenAI's ChatGPT API to integrate an instance of the model with further targeted training of the unit content, which is possible due to its ability to persist data and remember conversations. The ChatGPT model was fed with PDFs of the unit textbooks to further its ability to answer technical questions, and the UWA policy documents to answer administrative questions. The model is made to be able to respond to questions across 3 domains; general, technical, and administrative.

Some of the motivations behind this project are the amount of similar and repetitive questions asked by students, and the high cost of tutors who answer these questions constantly. I was quite surprised to learn that a trial would take place this semester for a cohort of around 70 students for a chemical engineering unit. I did not expect that the ChatGPT technology would be used in a formal educational context in such a short time after it was introduced, albeit in early trial stages. There are reasons to be both concerned about the exploitation of this technology to cheat, and hopeful of the opportunities of ChatGPT technologies can bring to an educational context.