520 EXAM - Q.2

- Saransh Sharma [ss4368]

ESSAY Q-1: WARGAMES

"*WAR, WHAT IS IT GOOD FOR ?...........ABSOLUTELY NOTHING!*" ~RUSH HOUR SONG

The scene starts off with the protagonists dealing with a relatively tense situation. The world's fate is in the hands of a sentient AI - Joshua, residing inside the "WOPR" military computer system - and even though there are people who are skeptical of the actual computer science that is portrayed in the movie, after having taken this course, I can say that a lot of the capabilities shown are actually pretty feasible, and in fact I'd go so far as to say that the movie was way ahead of its time in terms of the picture of AI that it paints for the world.
I noticed the following things relating our course-work to the movie:

- Joshua can be seen as possibly utilizing a Feed Forward Neural Network since it utilizes a dataset of a large number of games played against itself which can be analogous to training in a neural network. It uses this capability to predict optimal moves to take. This is feasible by modern standards and not a stretch in my opinion.
- Another thing of note throughout this clip and the movie as a whole was Joshua's ability to understand and respond to questions/commands given in English by the characters. This possibly utilizes concepts of NLP and is akin to modern day applications like Google Assistant or Siri (to name a few popular ones). It could be using concepts like semantic parsing (converting some input given in natural language into a logical representation that is useful for the computer) and grammatical tagging to tag the portions of the input as a particular part of speech(noun,verb,adj. etc), tokenisation. This might be what helps Joshua understand commands better. This is also quite feasible by modern standards.
- We can also relate Joshua's decision to stop playing Global Thermonuclear war, seeing no end in sight as possibly a reward-utility scenario in an MDP, wherein it can see no reward

in the future and so stops the game altogether(reaches stationary state). This also seems like the right thing to do and is feasible in my opinion.

- We can also notice that the initial game, tic-tac-toe, is an adversarial (and if so, futile) game that is deterministic and fully solved. Here we see only draws because both players are the same AI and that AI is playing adversarially from both sides to ‘minimize’ the other player’s score. We studied about min-max search (for adversarial games) and αβ-Pruning which reduces the search space and can be used to solve a game of tic-tac-toe. This is theoretically accurate according to me.
- I think value iteration for tic-tac-toe is also a good approach since we can define utilities and rewards(for terminal states - the state is 3 Xs/3 Os etc). The state space can be thought of as the location of the Xs & Os on the grid.
- But jumping from tic-tac-toe to Global Thermonuclear War seems far-fetched as even if it learned how to solve tic-tac-toe, that data will not necessarily be useful for Thermonuclear War as the state-spaces are completely different and the AI’s correlation regarding this seems unjustified in my opinion.
- Its shown in the movie, that Joshua can do a whole bunch of different things at a very high level, where in reality as we learnt, there is usually a tradeoff, even for state of the art systems - AIs can be very good at the few things they OR they can do a lot of basic things at a lower level of sophistication. This trends somewhat in the direction of Artificial General Intelligence [7] - which is essentially the ability of an AI to learn any “intellectual” task in the same manner that humans do- which is further off into the future than what is currently possible. This in my opinion is a bit of a stretch.

LESSONS !

- We can’t hold a computer program (even a possibly sentient one!) to standards higher than we hold ourselves to, having built them ourselves with our own biases and imperfections.

  Relating to this particular case for instance, there could’ve been a human child who got their hands on the ‘nuclear football’

& who single handedly, could put millions of people at risk and possibly not understand what they did wrong and what the consequences could've been for their decisions.

- **Sometimes**, 'playing around and having fun with it' can get you in deep, deep trouble
- The 80's were a prime time to be a sci-fi fan
-

ESSAY Q-2: PART-1

Why are Large Language Models prone to making things up in this way? Why don't they know better? Be thorough and draw on our discussions of machine learning

⇒ LLMs [1] sometimes tend to regurgitate false information when queried, this can be because they deal with very large datasets and have to digest a lot of information. Another reason is that LLMs don't really "understand" the meaning of what they output, they formulate/predict their answers on the basis of contextual information (context being derived from the previous words in the sentence/sequence) in order to respond as humans do and don't really have the ability to check the factual accuracy of the output they produce, therefore sometimes end up giving out incoherent or even false/misleading information that might not really exist such as citing non-existent research papers to support a claim or giving false and potentially even dangerous information such as benefits of eating glass etc [2].[in the case of Galactica] that might have existed in the dataset they were trained on. [Developers of such models generally issue warnings to users in a similar vein as well]

ESSAY Q-2: PART-2

To build a system that takes as input a natural language prompt from the user and generates a good Google search prompt, we could:

- We start with data collection - We generate a dataset containing search prompts and the respective natural language prompts as no such dataset already exists to the best of my knowledge - this can be done by first using a voice to speech dataset and then using those natural language prompts and generating search results from google. We then pre-process [3] the above dataset (as we did in Project-3 for model V, Vpart) by performing some or all of the following operations on the dataset - removing punctuations, tokenizing, lowercasing etc. to standardize, declutter and dimensionally reduce the data, this helps in improving performance and accuracy of the model down the line. We can also use a portion of the data we have collected for training and keep the remaining for testing the model down the line (as we did in Project-3 for model V, Vpart).

- We can then use GloVe Embedding [4] to convert words into numeric and more aptly, vector values which also capture the text's underlying semantic meaning. The embedded output can be given as input to the model in the next step.

- After this, we can use some variation of a neural network (eg. RNN) to train a model that predicts the prompts based on previous searches and the given prompt in numeric/vector format after embedding.

  We can also use LSTMs (Long-Short Term Memory) [5][6] as they are able to understand deep rooted dependencies in data and this is especially suitable in our case of language processing as the prediction being made and meaning of each word depends heavily on context.

The above steps can help us generate targeted google search prompts.

REFERENCES

[1] "How truthful is GPT-3? A benchmark for language models." LessWrong, 16 September 2021, https://www.lesswrong.com/posts/PF58wEdztZFX2dSue/how-truthful-is-gpt-3-a-benchmark-for-language-models. Accessed 16 December 2022.

[2] Heikkilä, Melissa. "Trust large language models at your own peril." MIT Technology Review, 22 November 2022, https://www.technologyreview.com/2022/11/22/1063618/trust-large-language-models-at-your-own-peril/. Accessed 16 December 2022.

[3] "Text Preprocessing NLP | Text Preprocessing in NLP with Python codes." Analytics Vidhya, 25 June 2021, https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/. Accessed 17 December 2022.

[4] "GloVe: Global Vectors for Word Representation." Stanford NLP Group, https://nlp.stanford.edu/projects/glove/. Accessed 18 December 2022.

[5] Graves, Alex. "Understanding LSTM Networks -- colah's blog." Colah's Blog, 27 August 2015, https://colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed 18 December 2022.

[6] "A Guide to Long Short Term Memory (LSTM) Networks." KnowledgeHut, 19 November 2022, https://www.knowledgehut.com/blog/web-development/long-short-term-memory. Accessed 18 December 2022.

[7] Simon, Charles. "The human touch: 'Artificial General Intelligence' is next phase of AI." C4ISRNET, 11 November 2022, https://www.c4isrnet.com/cyber/2022/11/11/the-human-touch-artificial-general-intelligence-is-next-phase-of-ai/. Accessed 16 December 2022.

My dog Alex - a return gift