Anurag Asthana, Yosen Lin, Rohan Sachdeva, Hanzen Shou

Dr. Cottrell

Deep Learning Semantle Project Proposal

The goal of our project will be to train a machine learning model that can solve Semantle puzzles in as few guesses as possible. For context, the Semantle website (https://semantle.com/) provides some very good information on what the game is and how it is played [1]. Essentially, Semantle is a variant of Wordle in which the goal is to guess a target word (not necessarily five letters) with the feedback for each guess given in terms of a similarity score; this similarity score measures how semantically similar or how similar in meaning the guess is to the target word. A score of 100 is the maximum score indicating a winning guess, while -100 is, in theory, the lowest similarity score that a guess can achieve.

A key motivation for the project is for enjoyment, as some of our games have gone into hundreds of guesses, while a few times, we have managed to beat the game in around twenty guesses, so we want to assess how well Deep Learning models could perform against a human player on average. Another motivation is that our model would be learning how similar certain words are to others and gaining an understanding of relationships between words, which has many applications in natural language processing (NLP). Furthermore, an extension to consider is to be able to train the model to solve Semantle puzzles in different languages, and the options offered through the website are Swedish, Hebrew, Spanish, Portuguese, French, German, Turkish, Russian, Dutch and Korean.

There are several plausible methods of data collection that we can employ here. As of now, we have access to a total of 785 Semantle puzzles which includes today's puzzle and past puzzles (https://www.thewordfinder.com/semantle-archives/), so we could have the model query guesses to these websites and fetch the similarity scores directly [2]. We can also use the following open-source project on Github which applies Word2Vec in generating unlimited Semantle puzzles (https://github.com/qwertyasdef/Semantle-Unlimited) [3]. As mentioned in the description of this project, it is possible to build a local instance to play these Semantle games. By setting this up, we may be able to query our guesses into the local instance and fetch the similarity scores accordingly.

For our model architecture, one idea that we had is to use a Recurrent Neural Network with LSTM cells as a policy network for reinforcement learning to generate sequences of words. Within this framework, we can use similarity scores as a form of reward signal. During the training process, the RNN's policy will be iteratively updated to maximize the expected cumulative reward. Using this optimization, we aim to enhance the quality and relevance of the generated word sequences.

For related work, we found an interesting PhD thesis *Semantic Representation and Inference for NLP* by Dongsheng Wang (https://arxiv.org/pdf/2106.08117.pdf), which contains essential, relevant background ideas in natural language processing for our project [4].

Links and References:

[1] Semantle Website - https://semantle.com/

[2] Semantle Archives - https://www.thewordfinder.com/semantle-archives/

[3] Semantle Unlimited Project - https://github.com/qwertyasdef/Semantle-Unlimited

[4] *Semantic Representation and Inference for NLP* - https://arxiv.org/pdf/2106.08117.pdf