Bay Hodge
6-26-23
NLP Project Proposal

Title: Evaluating Logical Coherence of Conversations between AIs
Team members: Bay Hodge

Description:

- I will evaluate the output resulting from two generative language models in conversation. The project will involve training GPTs on datasets of varying sources, quality, and size. Then, I'll evaluate how logically coherent the resulting conversation is using human judgment and a tool called "Coh-Metrix, which analyzes texts on over 200 measures of cohesion, language, and readability" (https://link.springer.com/article/10.3758/bf03195564).
- I'm interested in pursuing this project to better understand working with existing deep learning models and connecting them in a pipeline, training natural language models, and learning more about generative language models and how they are evaluated. The hypothesis I aim to test is the extent to which more data and data of a higher quality result in more logically coherent natural language outputs. Here, I want to understand why people say that OpenAI's edge over its competitors is its enormous quantity of data and high data quality. As a secondary purpose for the project, I want to simulate conversations between trained models from very different sources (see next paragraph).
- The training data will consist of a variety of text data sets from a variety of sources. For example, I'll train one model on a dataset of Star Wars movie scripts and the other on a dataset of medical papers. And perhaps, a dataset of ancient English texts and a dataset of modern ones.
- I plan to write code to create the architecture of and train the GPTs, connect them in a back-and-forth pipeline, and then implement a secondary evaluation process that tests the coherence of their conversation. The end product will look like this: the user will be prompted to choose which two AIs they want to converse with each other, and then the user can press a single button that generates a conversation between the desired AIs and outputs coherence scores from Coh-Metrix.

Literature summary:

Attention is All You Need (https://arxiv.org/abs/1706.03762)

- This seminal paper describes how attention mechanisms alone could effectively capture the relationships between words in a sequence, eliminating the need for recurrent or convolutional neural networks. Because of this, I plan to use the transformer architecture described in this paper to build and train the generative language models.

- In the paper, they train a machine translation model on millions of sentence pairs, optimized using Adam with a variable learning rate, and regularized with residual dropout and smoothing. It works in an encoder-decoder sequence in which the encoder maps an input sequence of symbol representations to a sequence of continuous representations. The decoder then takes this latter sequence to generate an output of symbols one element at a time.
- One particular source that might help with the implementation of this is Andrej Karpathy's video on building GPT from scratch (https://www.youtube.com/watch?v=kCc8FmEb1nY), which follows this paper closely, and implements the decoder portion of the transformer.

Language Models are Few-Shot Learners (https://arxiv.org/abs/2005.14165)

- OpenAI's GPT-3 and other GPT models are trained on a massive chunk of the internet to predict the next word in a document and then finetuned using reinforcement learning with human feedback. The big leap performed by GPT-3 was that it showed that a model trained on a massive amount of data improves task-agnostic, few-shot performance (i.e the model is given a few demonstrations of the task at inference time but without fine-tuning/updates to the weights) and can be preferred to a model trained with 10x less data that is then fine-tuned.
- 8 different sizes of models are trained in this paper, with larger models requiring larger batch sizes and smaller learning rates. The model and architecture are nearly the same as in GPT-2, except that they alternate dense and locally banded sparse attention patterns in the layers of the transformer.
- GPT-3 could be relevant for our project because it allows for fine-tuning, with instructions here: https://platform.openai.com/docs/guides/fine-tuning

LoRA: Low-Rank Adaptation of Large Language Models (https://arxiv.org/abs/2106.09685)

- An alternative option to training encoder-decoder models from scratch, which could possibly end up requiring more data and time than I can afford, is fine-tuning a pre-trained model using the aforementioned various datasets. LoRA describes a less expensive methodology for fine-tuning models by "freezing the pre-trained model weights and injecting trainable rank decomposition matrices into each layer of the Transformer architecture."

Coh-Metrix: Analysis of text on cohesion and language (https://link.springer.com/article/10.3758/bf03195564)

- Coh-Metrix analyzes texts "on over 200 measures of cohesion, language, and readability." For example, one measure that will be useful for this project is Referential Cohesion, which "measures how much words, word stems, or concepts overlap within a text; low referential cohesion can cause a reader to have difficulty connecting ideas between sentences” (https://soletlab.asu.edu/t-e-r-a/). This metric can help us measure coherence on a word-to-word and sentence-to-sentence level. Another metric, Deep Cohesion, “measures how events and ideas are related throughout the entire text; with greater overlap suggesting greater overall cohesion” (https://soletlab.asu.edu/t-e-r-a/). This metric can help us measure coherence with regard to the entire conversation.

Plan of action:

- The project consists of two parts: a system to process the data and train the models, and an end-product that deploys the train models so that a user can choose the two they want to generate a conversation. Finishing the first part is of the highest priority because I can test my desired hypotheses without the second part, which is the “packaging.”
- The plan is as follows: First, collect all relevant data sources.
- Then, follow papers and other relevant tutorials to create a working generative language model that accepts text inputs and outputs text responses. Then, I train these models on the data.
- Then, I need to write a software pipeline that allows these models to converse with each other, i.e. one’s input is the other’s output.
- I hope to have a pair of encoder-decoder models that can converse with each other by the midterm presentation.
- Then, evaluate the logical coherence of these conversations using (1) my own judgment and (2) Coh-Metrix. Supplying the Coh-Metrix tool with text that the tool can read as well as in a way that makes sense for measuring coherence between a prompt and a response will likely take some additional thought and engineering.
- Then, make improvements to the inference speed of the models, if necessary. Training speed is of less concern since when the model is trained, I will not need to retrain them, but inference speed affects the useability of the project.
- Finally, finalize the end product described above. The final paper I plan to write will have discussions and charts on the human judgment I performed on the resulting conversations, but the end product will also output conversations to the user in a fashion that they will easily be able to read the conversation and make their own intuitive judgment on whether the conversation made sense. The final presentation will be a combination of explaining the paper results and demonstrating the end product.

Description of the evaluation process:

- As a baseline for each trained model, I’ll measure its coherence from its initial output (as generated by a predetermined human-input prompt), i.e. without any conversation being had with another model. Then, I’ll measure how the coherence degrades with each subsequent back-and-forth of the conversation. As a baseline for measuring the conversation coherence, I’ll have each trained model converse with a model trained on noise - i.e., a set of sentences generated as a string of random words.

- As a first test if a conversation is logically coherent, we can simply use a human test (read the conversation, see if it makes sense, and grade it on certain metrics, like “Does it stay on topic?”, “Do the responses make sense?”, and “Do the individual sentences of one of the models make sense at all?”
- As a second test, I plan to use Coh-Metrix, an open-source software toolkit that provides a wide range of linguistic and discourse features to assess the coherence of the text. Coh-Metrix analyzes various aspects such as lexical diversity, cohesion, syntactic complexity, and semantic similarity.