Training your AI

### Description



The training data for a GPT model such as ChatGPT4 consists of many (hundreds of billions!) tokens harvested in sequence from various online sources, such as Wikipedia amongst many others. This source data is processed as a continuous stream independent of document boundaries.

## Tokenize

The initial training task is to tokenize the entire corpus, identifying recurring fragments of words, syllables, punctuations, capitalisations, and text symbols, as well as complete words. That involves analysis of the data stream to create a lexicon of identifiable subword units or word fragments. Then the model divides the incoming stream into discrete tokens.

Hereâ??s an example of a tokenized sentence that includes a mixture of complete words and word fragments. Note that tokens here incorporate spaces and punctuation marks:

â??Unâ?•, â??conventionalâ?•, â?• graffitiâ?•, â?• oftenâ?•, â?• reâ?•, â??contextâ?•, â??ualâ?•, â??izesâ?•, â?• theâ?•, â?• cityâ?•, â??scapeâ?•, â??.â?•

The training method treats the source text data as a continuous stream, albeit chunked in ways convenient for its method of processing. It doesnâ??t take account of the concept of documents or passages and does not consider document boundaries.

## Chunk

This chunking operates as a kind of sliding window into the data stream. The window constitutes the context across which the model considers each token in turn, predicts the token that follows and adjusts the network iteratively to minimise error in its ability to correctly predict that next token in the sequence. The context size for GPT-3 is 2,048 tokens, but is larger for later versions.

Each time the model processes its chunk of 2,048 tokens, and before it inputs them to the training neural network, it must calculate the attention scores and positional embedding for each token in that chunk to predict the token that follows.

## Predict

The input to the neural network during training as it encounters each token in the stream is a vector, a list of 1,024 floating point numbers. The output from the network is a prediction for the next token. The network optimisation algorithm adjusts the network parameters so that the prediction corresponds to the actual next token.

Having adjusted the neural network parameters for that window, the model repeats the process for the next context, advancing by one token. At strategic stages, the model will iterate through the entire corpus to minimise the error for the entirety of its token predictions.

## Infer



Once it is trained, the model can be deployed for inference, i.e., recruited for conversation with a human being or other NLP application. The neural network will generate output predictions for new input sequences, even if the network has not encountered those exact input sequences in its training data.

Thatâ??s the â??superpowerâ?• of a neural network. If it has been trained appropriately and with well-coded input methodologies, it can generate new, original, and seemingly creative outputs. The outputs are statistically probable responses based on patterns in the training data.

As the model processes tokens, it can synthesize new terms and infer word usages (meanings) for out-of-vocabulary terms not seen in the input data, but based on patterns learned from the training data.

Itâ??s worth noting that output predictions during inference are positioned according to a probability distribution. The degree of randomness in the output is a hyper-parameter for the output of the model. Randomness in the output obviates unwanted repetition and can result in output more consistent with the way a human being would respond in a conversation.

## Note

- This blog is my own composition. In conversational mode I drew on ChatGPT4 to help with explanations of GPT models. It does that far better than any other sources I have used to date. I also asked ChatGPT4 to evaluate my explanation. It generally agreed with what I wrote and offered a few suggestions incorporated here.

**Category**

1. Artificial Intelligence

**Tags**

1. natural language processing

2. neural network
3. tokenization
4. training

**Date Created**
August 12, 2023
**Author**
rcoyne99