Evaluating your AI

Description

Iâ??ve been implementing small scale trials of automated natural language processing routines that deploy the same methods as ChatGPT, i.e., implementations of the so-called Transformer architecture. That requires training on sequences of words in an original source document, equating each word to a semantic encoding (i.e., a long vector of 30 floating point numbers sourced from a publicly available dictionary of words and tokens), positional encodings which are vectors derived from the positions of words in the text, and finally an attention mechanism that factors in a calculation about the importance of words in sentences.

These factors combine to provide input to a neural network model. I had to write the code in Python to prepare the data for the model, but the neural network is implemented and managed as a library routine (Tensorflow.Keras) accessed via high-level parameters called from my Python program. The attention mechanism is folded into the operations of the neural network. Itâ??s a complicated process and involves layers in the network. So it is accessed from Python via high-level specifications. I didnâ??t have to design the network or the attention mechanism.

My coding expert was ChatGPT4. As I have said in previous posts (e.g., AI makes AI), it has leveraged my own programming skills to a more professional level. Manuals, websites, and Youtube videos would be insufficient to achieve this. ChatGPT4 acts as an expert programmer, tutor, and troubleshooter.

AI for kids

My goal is to produce an NLP (natural language processing) model that generates some coherent (grammatical) text consistent with the training corpus, and will continue the thread of text (a prompt) supplied by a user.

Iâ??ve learned so far that to replicate the performance of a fully-fledged NLP is beyond my desktop capability and expertise. However, it is possible to produce a trained model that effectively reproduces the language patterns of a constrained training text.

Writing for children operates with limited vocabularies. Hereâ??s an infantilised version of some of my own text, courtesy of ChatGPT4:

> First, I want to talk about the scary part of smart computers before I talk about the good things they can do. Here, â??end of the worldâ?? really means â??big news.â?? When we talk about smart computers and why they make us worry, we learn a lot about people. More and more, we use computers to talk to each other, especially in cities. These smart talking computers show us what we say and how we talk in many places.

I tried to train on a text so formed but it fared no better than the grown up version. Such texts are still a challenge for my desktop AI.

A templated source text

The strategy I adopted to test my training model was to write a program to automatically generate text that closely follows a template, and see if a model trained on that text could reproduce the performance of the template. The template I adopted followed the pattern

Preamble adjective noun adverb verb adjective noun.

The terms in the template are selected at random

preamble: â??many people sayâ??, â??the council maintains thatâ??, â??local residents assertâ??, â??the consultants thinkâ??, â??retailers sayâ??, â??the developers thinkâ??
adjective: â??the localâ??, â??the historicâ??, â??the modernâ??, â??the publicâ??, â??theâ??
noun: â??neighbourhoodâ??, â??districtâ??, â??viewâ??, â??developmentâ??, â??amenityâ??, â??infrastructureâ??
verb: â??exploitsâ??, â??complementsâ??, â??enhancesâ??, â??threatensâ??, â??makes the most ofâ??, â??benefits fromâ??
adverb: â??possiblyâ??, â??certainlyâ??, â??probablyâ??, â??arguablyâ??

Hereâ??s a segment of the text so produced.

> â?¦ Local residents assert the view threatens the amenity. The council maintains that the view threatens the amenity. Retailers say the historic amenity complements the historic development. The consultants think the historic amenity complements the historic development. The consultants think the historic development certainly complements the historic development. Retailers say the modern neighbourhood certainly complements the historic development. The developers think the modern view certainly complements the historic development. The developers think the public infrastructure complements the historic development. Local residents assert the modern view probably complements the historic development. â?¦

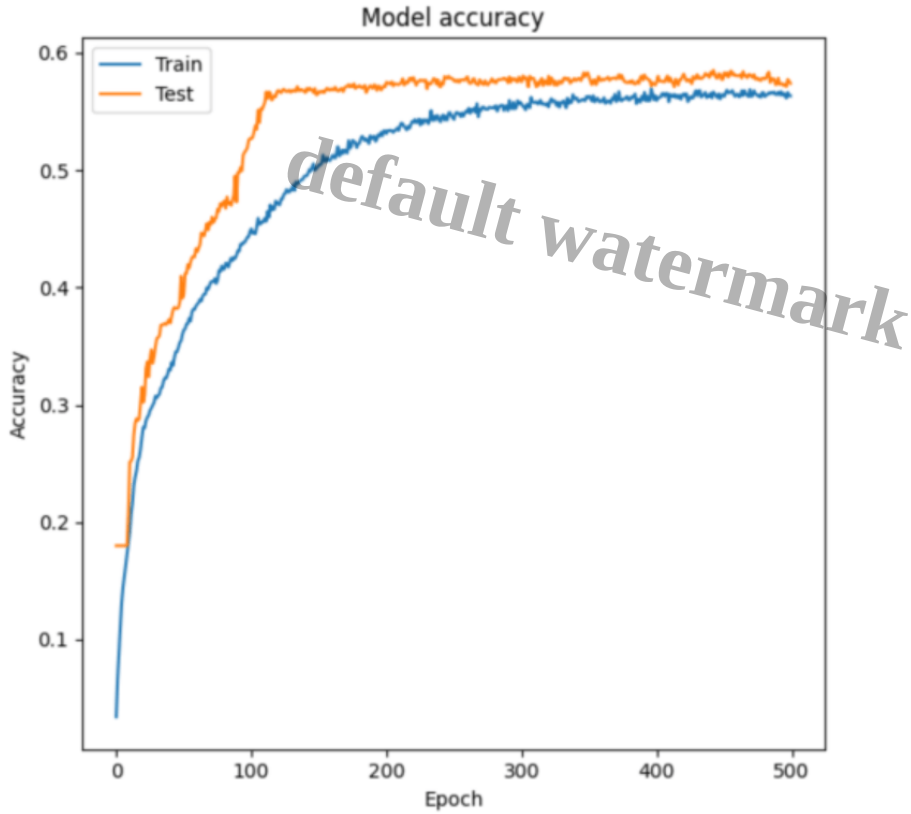
This uninspiring 21,314 word document is made up of just 39 words and contains 2,000 of the 103, 680 possible sentences conforming to the template. I put the sentences in reverse alphabetical word order to simulate some sense of context among sentences. It effectively clusters together sentences with

similar predicates.

Evaluation

Itâ??s possible to monitor the training, as shown in the diagram below. Over 500 epochs, the training accuracy converged towards a plateau near 60%. The Keras neural network model sets aside some of the training data for testing the accuracy of the modelâ??s ability to predict sequences of tokens.

In my case, the test accuracy closely follows the training accuracy. Thatâ??s good as it suggests the model is generalizing to unseen data and is not overfitting. Overfitting is where a model seems capable of only predicting word sequences that are in the training data, and not unseen sequences.



Here are some tests while using the model to complete sentences. Thatâ??s where we require the model to predict what could come next after a few prompt words. Hereâ??s an illustration of that generalizing capability.

Prompt: **the council enhances** â?¦

the council enhances the modern amenity.
the council enhances the modern infrastructure.
the council enhances the neighbourhood.
the council enhances the historic development.
the council enhances the amenity.

Though the individual words exist in the corpus text there is no combination â??the council enhances.â?• It helps the credibility of these outputs that â??councilâ?• can go with either a singular or plural verb. As a further prompt, there are 7 occurrences in the corpus of the phrase â??the development makes.â?• Here are some outputs.

the development makes the most of the public amenity. [not in corpus]
the development makes the most of the infrastructure. [not in corpus]
the development makes the most of the modern amenity. [not in corpus]
the development makes the most assert the modern neighbourhood. [not in corpus]
the development makes the most of the local neighbourhood. [occurs once]
the development makes the most of the historic amenity. [occurs twice]

Here is a further prompt. There is no phrase in the corpus â??**the council makes**.â?• The model predicts:

the council makes the most of the district.
the council makes the development from the infrastructure.
the council makes the most of the local neighbourhood.
the council makes the most of the public development.
the council makes the most of the district.
the council makes the most of the development.
the council makes the most of the infrastructure.
the council makes the most of the district.
the council makes the most of the modern view.
the council makes the most of the public neighbourhood.

Here is a further prompt: â??**people say**.â?•

people say the neighbourhood enhances the modern amenity. [not in corpus]
people say the view. [13 times in corpus]
people say the district threatens the public district. [not in corpus]
people say the historic view. [7 times in corpus]
people say the modern amenity. [8 times in corpus]
people say the district neighbourhood complements the public historic infrastructure. [not in corpus]
people say the development most of the view. [not in corpus]
people say the modern district. [7 times in corpus]

Not all outputs are sensible but it is clear that the predictions follow the template, with some deviations. A lot depends on the coherence of the training corpus, and this randomly generated template-based demonstration shows little evidence that it has access to context or meaning.

Human filtering

I created a variant of the text generator with a view that a human censor might approve or disapprove of each generated sentence. The criteria would be whether a sentence makes sense and/or is relevant to a particular context. Though I programmed that feature, it proved tedious to filter more than a few sentences manually.

7 Many people say the public development probably complements the local neighbourhood.
Approve this sentence? (yes/no/stop): yes

8 Many people say the view arguably enhances the modern view.
Approve this sentence? (yes/no/stop): no
Sentence not approved, discarded.

9 Retailers say the local neighbourhood benefits from the historic neighbourhood.
Approve this sentence? (yes/no/stop): yes

10 Retailers say the modern amenity possibly benefits from the modern infrastructure.
Approve this sentence? (yes/no/stop): yes

Scaling up

This experiment was to show at least that my NLP model is training as it should and with a reasonable accuracy. The main shortcoming is that this performance is achieved with only a small vocabulary. The vocab to document ratio is about 39:23,314. To scale this up suggests that the training corpus should be over 500 times the size of the vocabulary.

My target document has a vocabulary of 3,440 words and is 14,000 words long. It would need to be 1.7 million words long to achieve a similar performance, and with the same vocabulary, though ChatGPT4 tells me I would need to scale up the model as well, ensure diversity within the training data, amongst several other critical performance enhancements.

Note

- The featured image is from Dall-E3 now accessed via ChatGPT4, and in response to the prompts: photocell banner graphic for a blog about AI evaluation; something a bit more grungy and apocalyptic. I meant to type â??photorealâ?• instead of â??photocell.â?•

Category

1. Artificial Intelligence

Tags

1. evaluation
2. Keras
3. neural network model
4. training

Date Created
November 11, 2023
Author
rcoyne99