Why a neural network forgets

Description

Conversational AI, such as ChatGPT, has limited capacity to recall the content of earlier conversations. OpenAI does not disclose all the details of its operations, but scholars estimate that ChatGPT4 can process and recall up to 10,000 words in a single session or thread. Thatâ??s a substantial improvement on earlier models, but it doesnâ??t ensure continuity over a longer period of time.

Unlike an AI, a human consultant would obviously recall what was said some months earlier across numerous conversational threads. Most significantly, human consultants likely integrate the history of their conversations into their current interactions and advice. Long-term recollection for conversational AI hits limits. Itâ??s worth reviewing the operations of neural network models to see why.

Links and nodes

A natural language processing neural network consists of something like 1,028 input nodes depending on the design of the model. [See my revision to this in the comment section.] The number of output nodes corresponds to the number of words or tokens in the modelâ??s dictionary (about 50,000 in the ChatGPT model). There will be many layers of *hidden nodes* between the input and output nodes. Think of these layers of nodes as a stack with the inputs at the top and the outputs below. Input signals propagate from the top to the bottom.

The nodes in each layer of the network have input links from the nodes in the layer above and output links to the nodes in the layer below.

The input nodes at the top of the stack receive tokens in vector form. A single token vector is represented as a list of 1,028 floating point numbers. Iâ??ve shown how these vectors are derived in other posts. See post: Attention scores.

Conversational AI (e.g. ChatGPT) predicts what token or word may follow from another, like predictive text on a smartphone message editor. The neural network is trained on millions of existing texts available online and in official documents (a corpus). Training a neural network is a computationally intensive and time consuming process designed to facilitate contextual prediction, taking account of order and word meanings consistent with the way words are used in the training corpus. The details of

neural network training are for another discussion.

Once trained, the network will consist of a very large set of parameters, made up of the values attached to links between nodes, and bias values for each node. By some estimates thatâ??s about 1.5 trillion parameters for ChatGPT4. Once trained and tuned, these values remain fixed. They donâ??t change as people interact with the model.

The interaction stage is known as *inference*. Thatâ??s when someone asks a question, provides an input or otherwise feeds prompts into the model and expects a response.

Threads

Before it is submitted to the neural network during inference, the model modifies each input vector (pre-processing) through various matrix operations to take account of context and position in the token stream (the sentence). So, these pre-processing steps involve positional encoding. They modify the 1,028 vector values of the input token to reflect the position of each word in the input sequence and capture sensitivity to context.

During inference, each node in a hidden layer typically receives inputs from multiple nodes in the preceding layer. These inputs correspond to the output values of the nodes in the previous layer. If a node is deep in the network, then the inputs will have already undergone transformations as they propagate from the preceding layers.

When I asked ChatGTP for an account of the process it filled in some detail: The hidden node takes the transformed inputs from the previous layer, applies its own set of weights to each input value, and sums them up along with the bias term. An activation function generates the output of that particular hidden node. This process occurs across all nodes in the network as each calculates its output from its inputs. Values propagate through the network in this way eventually activating the output nodes.

The overall output from a given input is a probability distribution across all the tokens. The model uses that distribution to predict or select the next token in a sequence, e.g., If I type â??Graffiti is a social good,â?• ChatGPT might follow â??goodâ?• with â??thereforeâ?•, eventually producing the sequence, â??therefore we need to preserve its cultural valueâ?•. Thereâ??s a probabilistic aspect to the output, and it will likely give a different response for similar prompts at different times.

The succession of input tokens follows the same procedure. The propagation process starts afresh with each new token. Nothing is stored in the network from one token to the next. Each input vector has been already modified (pre-processed) through various matrix operations to take account of context and position in the token stream (the sentence). The input vectors take account of the most recent history of interactions for the current thread.

Trying to remember

Conversational AI platforms typically retain a history in text form of past threads, but these are independent conversations. The model does not use that textual history for recollection. Researchers are looking at ways to extend the memory capabilities of conversational AI models.

One method is for ChatGPT to provide a summary of the exchange and feed that into subsequent dialogues as input during inference, but that also hits severe limits as the model has no means of integrating text from an extended dialogue history.

I asked ChatGPT about attempts by researchers to give it a long-term memory. I summarise its response here.

> In so far as the conversational AI model reflects human linguistic cognitive processes it fails. As indicated above the values or weights attached to the links between nodes, along with the bias values for each node, are set during training and do not change during inference. On the contrary, human beings form new synaptic connections to integrate new information.

Temporal Memory Models

To improve long-term recollection in conversational AI, researchers are considering more dynamic models like differentiable neural computer (DNC) and Transformer-based architectures. In these models, a form of artificial memory is added to the network. This memory allows the network to write, read, and erase data over time, similar to the way the human cognitive system can modify its synaptic connections. These types of models can maintain a larger conversation history and provide more detailed responses based on prior context.

Memory-Augmented Networks

Memory-Augmented Neural Networks (MANNs) are networks that incorporate a large, addressable memory into the architecture, enabling the network to read from and write to the memory in a manner controlled by the input and output. This allows the network to â??learnâ?• which parts of its memory are important for particular tasks and how to use this information to produce more accurate outputs.

These models are complex and computationally intensive. The inability to recall the content of earlier interactions is a significant challenge for developers of current conversational AI models.

Note

- Featured image is the Victorian State Library, Melbourne (May 2023).

Category

1. Artificial Intelligence

Tags

1. AI
2. conversation

3. memory
4. neurons
5. recollection

Date Created
July 22, 2023
Author
rcoyne99