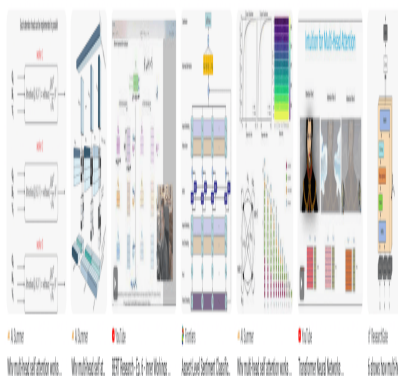




Attending to more than one thing at a time

Description

In an earlier post ([Attention Scores](#)) I considered how automated natural language processing (NLP) models attempt to simulate the way a listener or reader will focus on key words and groups of words in a sentence to decide how to respond to the sentence. I won't repeat the calculation here. But recall that the automated NLP method derives an attention score for each word in a sentence based on how close it is to each of the other words in the sentence in terms of semantic proximity. Words in these NLP models are represented as long sequences of numbers (vectors). The semantic proximity of two words is the distance between them in this [multidimensional features space](#). My example sentence "graffiti is a social good" calculated this (simplified) series of attention scores.

graffiti	is	a	social	good
0.59	0.40	0.37	0.58	0.69

An automated natural language model calculates and deploys these scores in both its training phase and also in the so-called *inference* or *prediction* phase. The latter is active when someone types comments or questions into a conversational AI system and expects a response. The model calculates the attention scores of the words in the input sentences. The model makes use of the attention scores, along with word sequences and word meanings, abetted by its functions as a fine-tuned neural network to generate (predict) the sequence of words that would make sense as a response – the words that would follow. See post [The next word](#).

Multiple attention heads

Human speakers are adept at placing emphasis on key words that they want listeners to pay attention to. *Attention* is under the influence of the person reading text out loud or delivering improvised speech. Irrespective of delivery, authors and readers might choose different words in a sentence as worthy of attention. In terms of automated natural language processing, a sentence might exhibit multiple patterns of potential attention scores as calculated in my earlier post ([Attention Scores](#)). NLP experts refer to the generation in their models of more than one pattern of attention for the same input sentence

as multi-head attention.

Itâ??s easiest to illustrate multi-head attention with sentences longer than the one I provided above. Iâ??ll use the sentence: â??Graffiti is both an art form and a means of disrupting the semiotic coding of the city.â?•

I recruited ChatGPT4 to help provide an illustration of the kind of multi-head attention scoring it uses. Hereâ??s one calculation where attention focuses according to the subject-action relationship: **graffiti** and **disrupting**. Those two words have higher scores calculated as in my earlier post about [attention scores](#). Note that the scores in this illustration are just to two decimal places. (They would have many more places in the actual model.)

Graffiti	is	both	an	art	form	and	a
0.50	0.05	0.05	0.05	0.05	0.05	0.05	0.05
means	of	disrupting	the	semiotic	coding	of	the
0.10	0.05	0.10	0.05	0.05	0.05	0.05	0.05

Focussing on the object being acted upon (the city) the model could produce a different distribution of scores, emphasising **semiotic**, **coding** and **city**.

Graffiti	is	both	an	art	form	and	a
0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
means	of	disrupting	the	semiotic	coding	of	the
0.05	0.05	0.05	0.05	0.40	0.40	0.05	0.05

Focussing on the method or means of action, the attention scores could appear something like the following, putting the emphasis on **art**, **form**, **means** and **disrupting**.

Graffiti	is	both	an	art	form	and	a
0.05	0.05	0.05	0.05	0.20	0.05	0.05	0.05
means	of	disrupting	the	semiotic	coding	of	the
0.30	0.05	0.20	0.05	0.05	0.05	0.05	0.05

I have shown the possible outcomes from just 3 attention heads. Natural language models such as ChatGPT4 will deploy a fixed number (8-16) of attention heads. Itâ??s worth noting a few points about this multi-head method of dealing with attention.

1. The way the attention heads detect these attention patterns is determined by a set of weights in the layer of a trained neural network. They are not pre-set by a human designer of the model. Nor is their performance as easily interpreted by a human operator as suggested above. Developers

of these NLP systems have designed them to operate in a certain way, but cannot necessarily interpret the parameters so calculated.

2. The performance of the attention heads and their varied interpretations of blocks of text are learned during the training process: the diversity among attention heads arises naturally from the random initialization of model parameters and the optimization process during training (according to a conversation with ChatGPT4). Once trained, the performance of the training heads is fixed, it won't change throughout the life of the model.
3. The multi-attention process lends itself to parallel computer processing, which is an advantage for NLP training of extremely large corpora of texts.
4. These models do not only assign attention scores to words in sentences but to [tokens](#), i.e. fragments of words. Nor do the methods apply only to sentences, but any arbitrary block of text. When the model generates output during user interaction it will calculate attention scores for the whole block. The grammatical structure of the user input is not relevant in assigning attention scores. The content of lists, paragraphs, sentences, tables or random word orderings are treated in the same way.

Richer NLP

NLP researchers show that calculating these multiple attention profiles leads to a rich neural network representation of textual information. Notwithstanding the caveats above, these representations correspond loosely to human language capabilities. Competent human speakers and listeners are adept at adjusting their focus depending on context. At one moment the emotional cues in a sentence come to the fore, at other times it's the social relationships implied, or the practical actions invoked.

Focussing on particular word combinations in a sentence models something of this capability, especially if we recall that words and tokens in NLP models are coded as vectors in multidimensional features spaces. *Scribble, doodle, defacement, art* and *graffiti* are strongly related and in different ways, whether or not they appear in the same sentence. That contributes to the ability of conversational AI to give the appearance of choosing its words carefully according to context as it predicts the next word in a sequence of words in its responses to inputs.

The multi-head attention mechanism provides a means of dealing with ambiguity and context in language. For human speakers and listeners, various interpretations of a sentence are on hold until further information comes to light, or perhaps different interpretations persist as part of our understanding held in parallel.

There are also computational efficiencies in NLP multi-head attention, resulting in the capacity to deal with and store the many and varied contextual features within large blocks of text.

The various attention weights also lead to some transparency in an NLP model, improving its capability to provide a rationale for its word predictions, responses or actions: e.g. It's as if the model might say, 'I explained the creative potential of graffiti as I focussed on the proposition that it is a social good.'

A multi-head attention mechanism enables an NLP model to respond to my sentences ('Graffiti is both an art form and a means of disrupting the semiotic coding of the city') with statements such as 'Graffiti is a creative aspect of the urban fabric,' 'Graffiti is contentious,' 'Graffiti requires

management, etc. These follow-on sentences depend on the context established by other prior statements in the conversation, including those generated by the model itself.

An application utilizing the ChatGPT-4 neural network language model can save the history of its text-based conversations with a particular user. It can then use these previous interactions as input to the model for subsequent responses. This enables the model to generate responses that are contextually consistent with the ongoing conversation. The multi-head attention mechanism within the model, amongst other architectural elements, contributes to the diversity of potential responses to any given input. (This paragraph was edited by ChatGPT4.)

Notes

- ChatGPT4 assisted me in exploring this explanation and the examples. I like to think this was a conversation. I took what I could from the explanations, as well as web searches. At this stage there seems to be little online that explains the techniques without getting into the reeds of specialist NLP terminology, vectors, matrices and diagrams, some of which appear in the featured image for this post.
- I stumbled across a Youtube clip of actor Ian McKellen explaining one of Shakespeare's soliloquies. This excerpt from an actors' workshop makes clear how important it is for actors to pay attention to particular words in their interpretation and delivery. See Joseph, K. (2022). "Tomorrow, and tomorrow" Ian McKellen analyzes Macbeth speech (1979). Youtube. from <https://youtu.be/zGbZCgHQ9m8>.

Category

1. Artificial Intelligence

Date Created

July 1, 2023

Author

rcoyne99