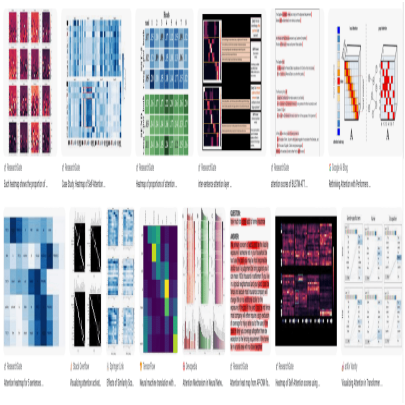

Attention scores

Description

One of the important techniques in the new wave of highly successful generative natural language (NLP) models is the use of *attention scores* in neural network (NN) training. Here I continue the investigation I started in previous posts into how NLP works.

To recap: a word in NLP models is represented as a point in a multidimensional feature space. The axes defining this space are unlabelled, that is, they are opaque to human scrutiny, but correspond to values attached to nodes in the hidden layers of a neural network. We can assume the NN procedure indicate clustering of words in the context of the training corpus on the basis of similar features. All words in an NLP system are defined with unique coordinates within this common â??semantic space.â?• The number of dimension defining this space are typically large, e.g. 2,048 in ChatGPT. A word will be defined by a string of coordinates 2,048 numbers long, each to 17 decimal places.

The coordinates for each word emerge from processing the very large corpus of texts encountered during NN training. The coordinates mostly reflect the proximities between words as they occur in sentences and in the context of each other. As I stated in a previous post (Architecture in multidimensional feature space) there are simple algorithms that calculate the proximity of words to one another in such a *semantic space.* So, â??graffitiâ?• is closer to â??artâ?• than â??bicycle.â?• Such relationships are general and derived from the entire training corpus.

Ambiguity

â??Graffiti,â?• â??artâ?• and â??bicycleâ?• have dominant meanings and well-established contexts of use, but many words are overtly ambiguous. A common example is the word â??bank.â?• Do I mean the building or the side of a river? Human beings have little difficulty in deciding between these meanings based on context: â??I deposited my money in the bankâ?•; â??The duck swam to the bank.â?• In these cases an automated NLP system needs to calculate the relationships between words in specific sentences in the training set as well as meanings within a generalised semantic space. The local context is needed.

Word usage is nuanced in any case. Though an obvious example, the challenge posed by the ambiguity of the word â??bankâ?• constitutes a perennial challenge for language use: capturing the relationships between word usage that follows general definitions and usage in the local context of a particular sentence, paragraph or larger segment of text.

Advanced NLP platforms (such as ChatGPT) deploy methods known as â??attention scoringâ?• to address the relationship between the general usage of words and their usage in the local context. The use of pronouns provides a similar challenge: â??She gave it to her friend.â?• Is â??sheâ?• referring to the same entity as â??herâ?•?

Embedding vectors

The NLP literature refers to coordinates of a word in a multidimensional feature space as *embedding vectors*. In the advanced methods I am referring to here, *attention* algorithms modify the generalised word embedding parameters to account for the local contexts of words. Hereâ??s how one helpful post by an NLP blogger (Kotamraju) explains the role of *attention* algorithms.

> â??A self-attention module works by comparing every word in the sentence to every other word in the sentence, including itself, and reweighing the word embeddings of each word to include contextual relevance.â?•

Itâ??s worth noting that this attentional manipulation of word embeddings occurs prior to NN training on the entire corpus of texts. Itâ??s a pre-training process. The full neural network training routine operates on modified word embeddings that capture both the positional embedding of words (word sequences in sentences as described in the previous post) and the attentional information I am about to describe.

An example of attention scoring

Word embeddings of 2,048 and 17 decimal places are too unwieldy to use in a demonstration. I asked ChatGPT to provide some plausible embeddings for a list of several words each just 4 dimensions, and to 2 decimal places. I canâ??t assume these vectors are calculated, but they seem to work, at least for illustration. I supplied the words. My ChatGPT â??tutorâ?• provided the vectors.

- â??cityâ?•: [0.8, -0.2, 0.5, 0.4]
- â??goodâ?•: [-0.1, 0.5, 0.9, 0.3]
- â??graffitiâ?•: [-0.2, 0.8, 0.3, -0.5]
- â??socialâ?•: [0.7, -0.4, -0.5, 0.6]
- â??treeâ?•: [0.2, 0.9, 0.1, -0.3]
- â??aâ?•: [-0.3, 0.6, 0.2, -0.1]
- â??isâ?•: [0.5, -0.1, 0.7, 0.2]
- â??notâ?•: [-0.5, 0.4, -0.2, -0.1]

I created the sentence â??graffiti is a social goodâ?• and positioned the embedding vectors for each word on an excel spreadsheet.

graffiti	-0.2	0.8	0.3	-0.5
is	0.5	-0.1	0.7	0.2
a	-0.3	0.6	0.2	-0.1
social	0.7	-0.4	-0.5	0.6
good	-0.1	0.5	0.9	0.3

I have ignored capitalisation and punctuation. Imagine this sentence is one of millions that appears in a neural network training corpus. The attention module of a neural network system will process the words in this sentence to take account of these generalised embeddings and modify them to account for their relationships with other words in the sentence.

Hereâ??s a potted summary of how that works. The attention algorithm takes each embedding vector in turn and calculates its proximity in the multidimensional feature space to every other word in the sentence. That is achieved by a simple multiplication and addition process (dot product). So each parameter in the â??graffitiâ?• vector is multiplied by each element in the â??goodâ?• vector and then summed to give an *alignment score* for semantic proximity to â??graffiti.â?• The scores for each word as it relates to â??graffitiâ?• are then turned into weightings (probabilities). Each weighting for the words in the sentence will add up to 1.0. The formula to do this last manipulation is known as *softmax*.

These weightings are then used in a calculation that modifies the embedding vectors for each word. Summing those modified vectors provides scores that serve to indicate the importance of each word to the meaning of the sentence as a whole.

Hereâ??s my spreadsheet that calculated the overall score for the word â??graffitiâ?• in the sentence. The final score for that word is 0.59.

	embedding vectors				alignment score	softmax weighting	weighted embedding vectors			
graffiti	**-0.2**	**0.8**	**0.3**	**-0.5**	1.02	0.36	0.01	0.23	0.03	0.0
is	0.5	-0.1	0.7	0.2	-0.07	0.12	-0.01	-0.01	0.03	-0.0
a	-0.3	0.6	0.2	-0.1	0.65	0.25	0.01	0.12	0.01	0.0
social	0.7	-0.4	-0.5	0.6	-0.91	0.05	-0.01	-0.02	-0.01	-0.0
good	-0.1	0.5	0.9	0.3	0.54	0.22	0.00	0.09	0.06	-0.0
							0.01	0.41	0.12	0.0

Here are the attention scores for all the words in the sentence, shaded for clarity.

graffiti	is	a	social	good
0.59	0.40	0.37	0.58	0.69

I was pleased to see that the attention scores look as one might expect for such a sentence. A linguistically capable human being reading or interpreting the sentence is indeed likely to focus their attention on â??graffitiâ?• and â??good.â?• That suggests that the method is correct, but also that the original embeddings provided in my conversation with ChatGPT were reasonable. I applied the same method to the sentence: â??A city is not a tree.â?•

a	city	is	not	a	tree
0.34	0.54	0.41	0.16	0.23	0.49

The question of what the NLP does with those attention scores is for another blog post. Itâ??s worth remembering at this stage that word embeddings are much larger than indicated here, that NLP systems typically process blocks of text that are longer than single sentences, and that NLP systems operate with tokens that include whole words, parts of words and punctuation. The overall objective of generative NLP is prediction: to predict the next word or token in a generative text sequence. If you think of the size of embedding vectors, it involves a huge amount of calculation to replicate what human language users take for granted as functions of basic language competence.

Bibliography

- Anon. â??Attention is all you need || Transformers Explained || Quick Explained.â?• *Developers Hutt*, 2022. Accessed 2 April 2023. https://www.youtube.com/watch?v=66selToeguE
- Galassi, Andrea, Marco Lippi, and Paolo Torroni. â??Attention in Natural Language Processing.â?• *IEEE Transactions on Neural Networks and Learning Systems* 32, no. 10 (2021): 4291-4308.
- Kotamraju, Saketh. â??An intuitive explanation of Self Attention: A step-by-step explanation of the multi-headed self-attention block.â?• *Towards Data Science*, 8 October, 2020. Accessed 4 April 2023. https://towardsdatascience.com/an-intuitive-explanation-of-self-attention-4f72709638e1
- See, Abigail. â??NLP with Deep Learning | Winter 2019 | Lecture 8 â?? Translation, Seq2Seq, Attention.â?• *Stanford Online*, Winter, 2019. Accessed January 10, 2023. https://www.youtube.com/watch?v=XXtpJxZBa2c&t=4337s
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. â??Attention Is All You Need.â?• In *31st Conference on Neural Information Processing Systems*, 1-15. Long Beach, CA, USA2017.
- Vig, Jesse. â??Deconstructing BERT, Part 2: Visualizing the Inner Workings of Attention.â?• *Towards Data Science*, 8 January, 2019. Accessed 3 April 2023. https://towardsdatascience.com/deconstructing-bert-part-2-visualizing-the-inner-workings-of-attention-60a16d86b5c1
- Walker, Richard. â??Attention â?? the beating heart of ChatGPT: Transformers & NLP 4.â?• *Luci Date*, 26 February, 2023. Accessed 12 March 2023. https://www.youtube.com/watch?v=sznZ78HquPc

Category

1. Artificial Intelligence

Tags

1. attention
2. NLP

Date Created
May 20, 2023
Author

rcoyne99