Visualizing and Measuring the Geometry of BERT

Björn Bebensee (2019–21343)
Topics in Artificial Intelligence

September 26, 2019

This paper by Coenen et al. [1] deals with the question which linguistic features of text are being extracted by transformers and how the information is stored internally. Previous work by Hewitt and Manning [2] has found evidence of geometric representations of syntactic and linguistic features (such as parse trees). In this work the authors aim to extend this investigation into geometric representations in BERT. They find evidence that attention matrices encode grammatical representations, they show that BERT manages fine-grained distinctions of word senses and they find that this information is encoded in a low-dimensional space.

Coenen et al. investigate attention matrices to find possible representations of dependency grammar relations. They define an *attention probe* as follows: given a model-wide attention vector, which is simply the concatenation of all $\alpha_{i,j}$ from every attention matrix in every attention head in every layer, an attention probe is the task of finding the encoded relation of a pair of tokens $(token_i, token_j)$ within the model-wide attention vector. They find that using a corpus of parsed sentences and their model-wide attention vectors (for each pair of tokens in a sentence), they can successfully train a linear classifier to recognize whether a dependency relation exists between a pair of tokens and even the type of dependency relation. This suggests that the attention factors do encode this kind of information of dependency relations. They further explain why the parse tree distance corresponds to the square of Euclidean distance (as found in [2]) using *power-2 embeddings*. They compare the parse tree embeddings found in BERT to these power-2 embeddings (see figures 2 and 3 in the paper).

Lastly, Coenen et al. inquire into whether and how context-dependent embeddings capture different meanings of the same word and explore their geometric investigations. They find that it is possible to build a nearest-neighbor classifier which, given context embeddings, can find the closest sense of a word. This suggests that it is context embeddings may indeed encode word-sense information. They further find that most semantic information is contained in the earlier-layer embeddings which suggests that word-sense information may be encoded in a lower-dimensional space.

References

[1] Coenen, Andy, et al. "Visualizing and Measuring the Geometry of BERT." *arXiv preprint arXiv:1906.02715* (2019).

[2] Hewitt, John, and Christopher D. Manning. "A structural probe for finding syntax in word representations." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019.