Visualizing and Measuring the Geometry of BERT

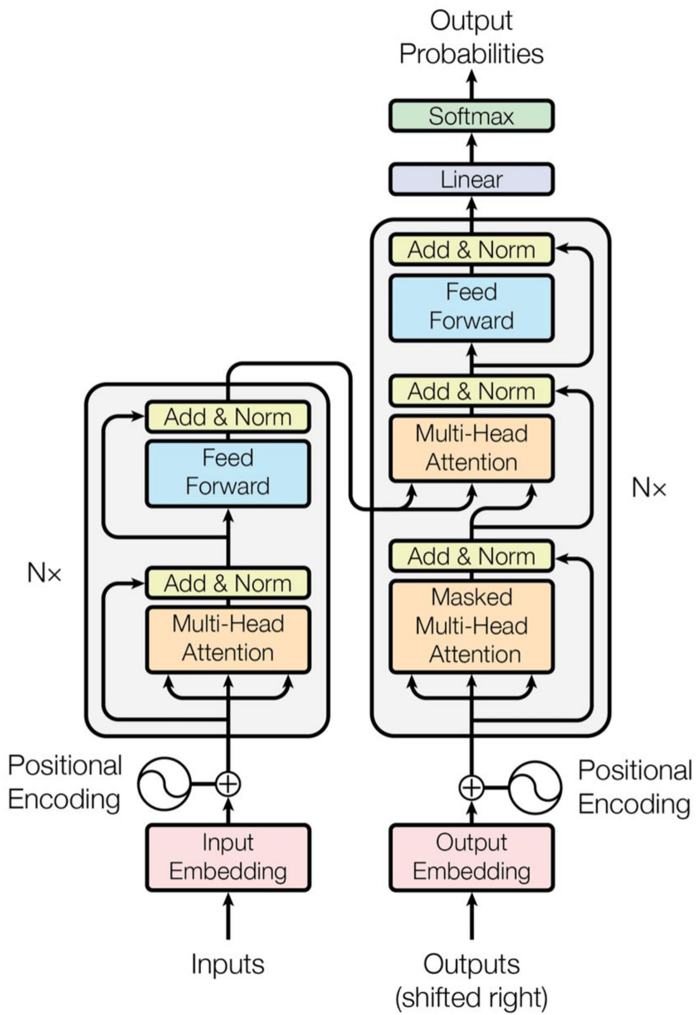
Andy Coenen,[*] Emily Reif,[*] Ann Yuan[*]
Been Kim, Adam Pearce, Fernanda Viégas, Martin Wattenberg
Google Brain
Cambridge, MA
{andycoenen,ereif,annyuan,beenkim,adampearce,viegas,wattenberg}@google.com

Motivation of the paper

- BERT is extracting a set of useful features from raw text - which features are extracted?
- How are these features represented internally?
- Especially:
    - Hewitt and Manning (2019) find evidence of geometric representation of entire parse trees in BERT's activation space
    - This work investigates how BERT represents syntax
    - **It shows evidence that attention matrices contain grammatical representations**
    - **It shows that BERT distinguishes word senses at a very fine level**
    - **Much of this semantic information appears to be encoded in a relatively low-dimensional subspace**

BERT - attention is all you need

BERT's model architecture is a multi-layer bidirectional **Transformer encoder** based on the original implementation described in Vaswani et al. (2017)



Syntatic information - are they encoded?

- Question: what is encoded in attention matrices?
- We are using an **attention probe**, which is checking dependency relation between two tokens using **model-wide attention vector**
- Model-wide attention vector is formed by concatenating the entries in every attention matrix from every attention head in every layer

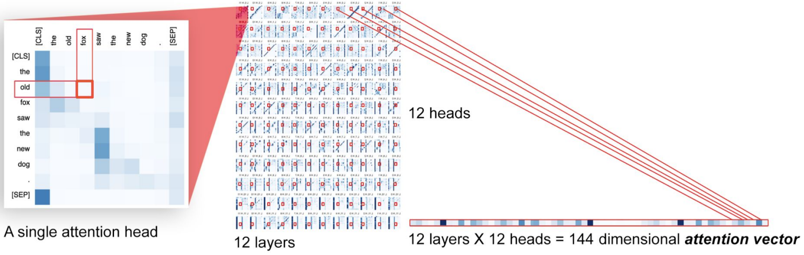


Figure 1: A *model-wide attention vector* for an ordered pair of tokens contains the scalar attention values for that pair in all attention heads and layers. Shown: BERT-base.

Probing method

- Dataset is based on Penn Treebank
- 30 relations checked with more than 5000 examples in the data set
- Each sentence was run through BERT-base to obtain the model-wide attention vector
- Models used:
  - Goal 1: to predict whether there is a dependency relation between two tokens
  - Goal 2: which type of dependency relation exists between two tokens, given the dependency relation's existence

Probing method - results

- Classifier 1: 85.8% of accuracy
- Classifier 2: 71.9% of accuracy

The aim is to gauge whether model-wide attention vectors contain a relatively simple representation of syntactic features. The success of this simple linear probe suggests **that syntactic information is in fact encoded in the attention vectors.**

Word senses - semantics

- The embeddings produced by BERT (& transformer models) depend on context
- Thus they should capture the particular shade of meaning of a word as used in a sentence

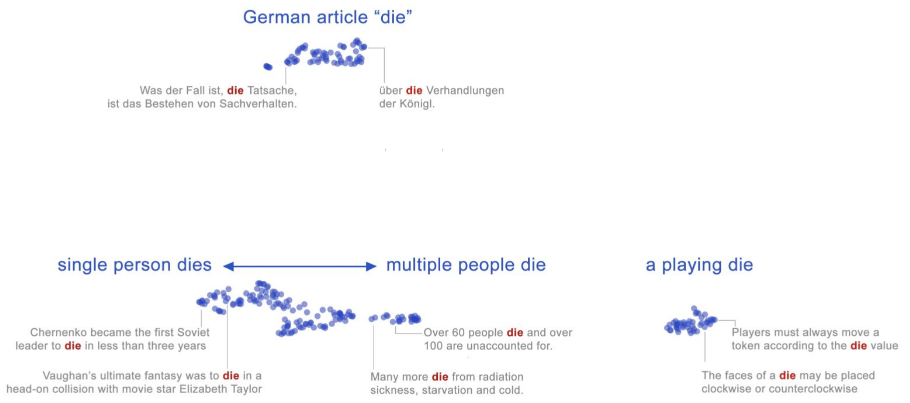


Figure 4: Embeddings for the word "die" in different contexts, visualized with UMAP. Sample points are annotated with corresponding sentences. Overall annotations (blue text) are added as a guide.

Visualization of semantics

- We observe clear clusters relating to word senses
- Different senses of a word are typically spatially separated
- Within the clusters there is often further structure related to fine shades of meaning

Is it possible to find quantitative corroboration that word senses are well-represented?

Figure 4: Embeddings for the word "die" in different contexts, visualized with UMAP. Sample points are annotated with corresponding sentences. Overall annotations (blue text) are added as a guide.

Semantics - quantitative validation

- Already prepared datasets
- To back up visualization results, a simple word-sense disambiguation classifier is trained.
- For a given word with *n* senses, a nearest-neighbor classifier is trained, where each neighbor is the centroid of a given word sense's BERT embedding
- To classify a new word, we find the closed of these centroids

Method	F1 score
Baseline (most frequent sense)	64.8
ELMo [20]	70.1
BERT	**71.1**
BERT (w/ probe)	**71.5**

F1 scores for WSD

Semantics - quantitative validation

- Already prepared datasets
- To back up visualization results, a simple word-sense disambiguation classifier is trained.
- For a given word with *n* senses, a nearest-neighbor classifier is trained, where each neighbor is the centroid of a given word sense's BERT embedding
- To classify a new word, we find the closed of these centroids

Method	F1 score
Baseline (most frequent sense)	64.8
ELMo [20]	70.1
BERT	**71.1**
BERT (w/ probe)	**71.5**

F1 scores for WSD

Changing semantics of a word

- Already prepared datasets
- To back up visualization results, a simple word-sense disambiguation classifier is trained.
- For a given word with *n* senses, a nearest-neighbor classifier is trained, where each neighbor is the centroid of a given word sense's BERT embedding
- To classify a new word, we find the closed of these centroids

Embedding distance and context: a concatenation experiment

- If word sense is affected by context, and encoded by location in space, then we should be able to influence context embedding positions by systematically varying their context.
- To test this hypothesis, we performed an experiment based on a simple and controllable context change: **concatenating sentences where the same word is used in different senses.**

A: "He thereupon *went* to London and spent the winter talking to men of wealth." *went*: to move from one place to another.
B: "He *went* prone on his stomach, the better to pursue his examination." *went*: to enter into a specified state.

Embedding distance and context: a concatenation experiment

- **individual similarity ratio:** ratio of cosine similarity between the keyword embeddings and their matching sense centroids and the keyword embeddings and their opposing sense centroids.
- **concatenated similarity ratio**

- Hypothesis -> the keyword embeddings in the concatenated sentence would move towards their opposing sense centroids.
- Result -> We found that the average individual similarity ratio was higher than the average concatenated similarity ratio at every layer

Embedding distance and context: a concatenation experiment

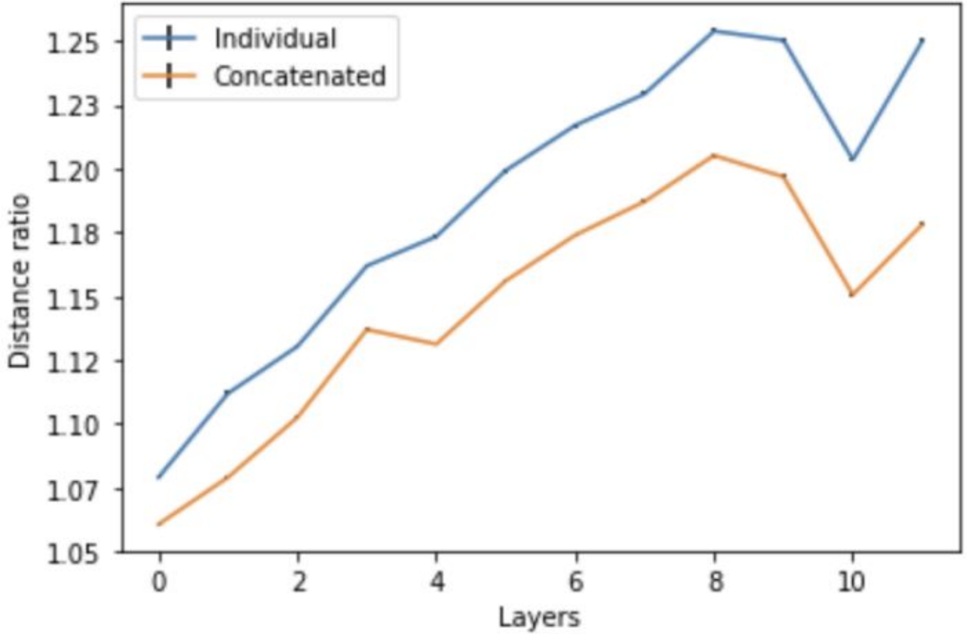


Figure 5: Average similarity ratio: senses A vs. B.

Further research questions

- What other meaningful subspaces exist? After all, there are many types of linguistic information that we have not looked for.
- What the internal geometry can tell us about the specifics of the transformer architecture.
- Can an understanding of the geometry of internal representations help us find areas for improvement, or refine BERT’s architecture?