BERT Visualization

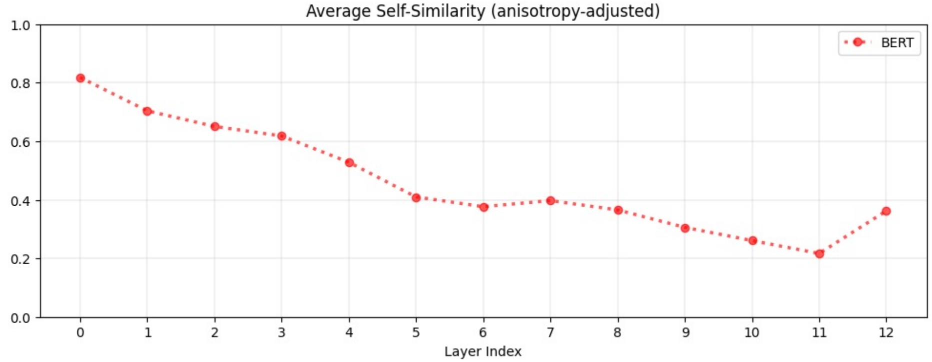
Xueren Ge

Haibei Zhu

Metrics & Results

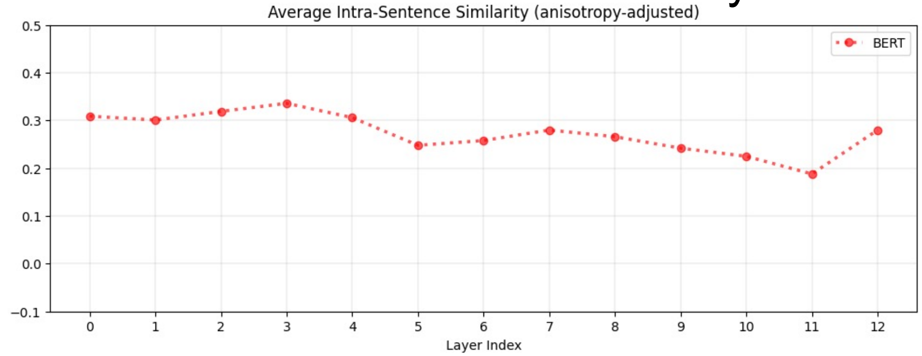
Dataset: Stanford Natural Language Inference (SNLI) Corpus

Self-similarity



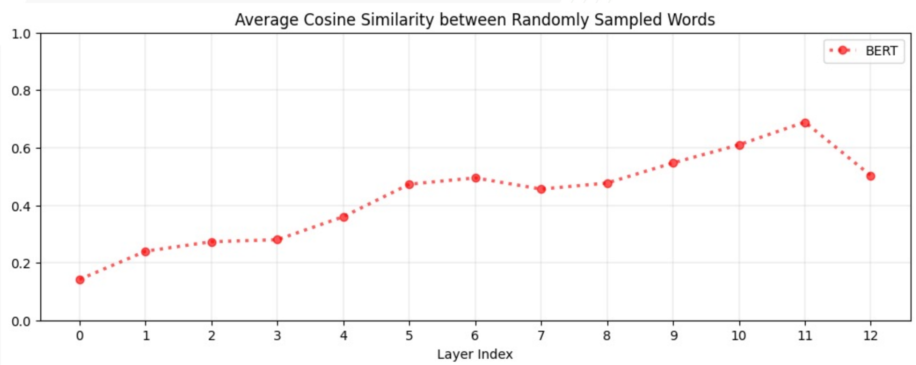
The higher the layer, the lower the self-similarity, suggesting contextualized word representations are more context-specific in higher layers.

Intra-Sentence similarity



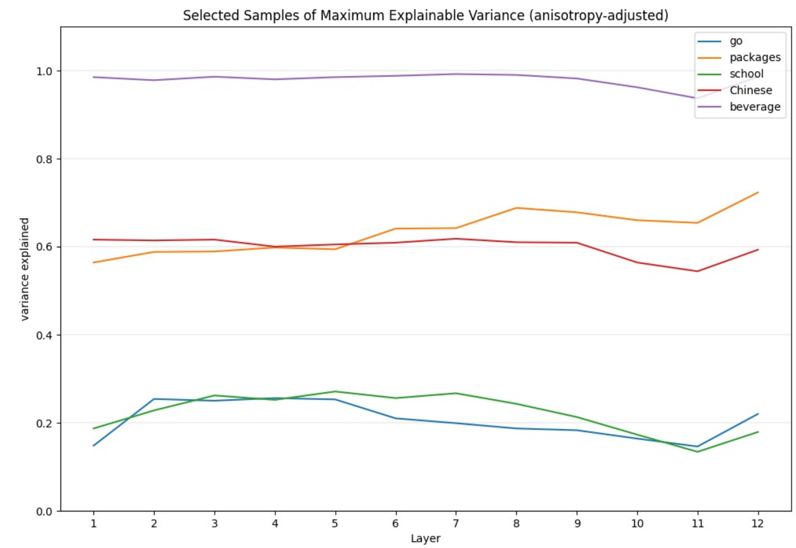
As layer goes deeper, the lower Intra-sentence similarity → 2 words don't have same meaning even when they share the same context.

Anisotropy



Any 2 random words have similar representation, which means, the whole vector space tend to be a cone instead of ball.

Maximum explainable variance



'beverage' is almost 1, which means in all context , in all layers its meaning is almost the same. And can be replaced by static embedding.

How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings

Future Goal

- Tackle “Anisotropy” problem within BERT
  - How to visualize anisotropy problem (done)
  - How to solve anisotropy problem (contrastive learning)
- Understand why BERT work
  - Does BERT really learn contextual information? (yes!)
  - The importance of each layer in BERT (Hugging Face)
- How to build better BERT
  - Structure
  - Dataset

Future Work

- Data-biased
  - Inspired by Last class ethics, we planned to change the data distribution, let it be biased-dataset, and see what Bert will learns and what’s going on in each layer
- Structure
  - We have seen layer 12 may have bad performances, we planned to use Hugging Face to froze some layers to see what Bert will react.
- Anisotropy
  - Maybe use contrastive learning to see how to solve this problem

How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings
https://aclanthology.org/D19-1006.pdf