BERT Visualization

Xueren Ge

Haibei Zhu

Goal

- Tackle “Anisotropy” problem within BERT
  - How to visualize anisotropy problem
  - How to solve anisotropy problem
- Understand why BERT work
  - Does BERT really learn contextual information?
  - The importance of each layer in BERT
- How to build better BERT
  - Structure
  - Dataset

Metrics

- Self-similarity

$$SelfSim_{\ell}(w) = \frac{1}{n^2 - n} \sum_{j} \sum_{k \neq j} \cos(f_{\ell}(s_j, i_j), f_{\ell}(s_k, i_k))$$

  - The word w in layer l is the average cosine similarity between its contextualized representations across its n unique contexts.

  - E.g.
    - S1: [w, a, b, c, d]
    - S2: [g, h, j, k, w]

  - IF $SelfSim_l(w)$ is high
    - model doesn’t contextualize the representations at all
  - IF $SelfSim_l(w)$ is low
    - model contextualize the representations totally

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

Metrics

- Intra-sentence similarity

$$IntraSim_{\ell}(s) = \frac{1}{n} \sum_{i} \cos(\vec{s}_{\ell}, f_{\ell}(s,i))$$

$$\text{where } \vec{s}_{\ell} = \frac{1}{n} \sum_{i} f_{\ell}(s,i)$$

  - the average cosine similarity between its word representations and the sentence vector
  - E.g.
    - S1: [w, a, b]
    - $\vec{s}_l = \frac{1}{3}(embedding(w) + embedding(a) + embedding(b))$
    - $intraSim(s) = \frac{1}{3}[\cos(embedding(w), \vec{s}_l) + \ \dots]$
  - IF $intraSim(s)$ low
    - 2 words in the same sentence don’t have a similar meaning simply because they share the same context

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

Metrics

- Maximum explainable variance

$$MEV_\ell(w) = \frac{\sigma_1^2}{\sum_i \sigma_i^2}$$

  - the proportion of variance in *w*'s contextualized representations for a given layer that can be explained by their first principal component

  - E.g.
    - S1: [w, a, b], S2: [c, w, d]
    - $[embedding(w_1), embedding(w_2)]$ ---> SVD

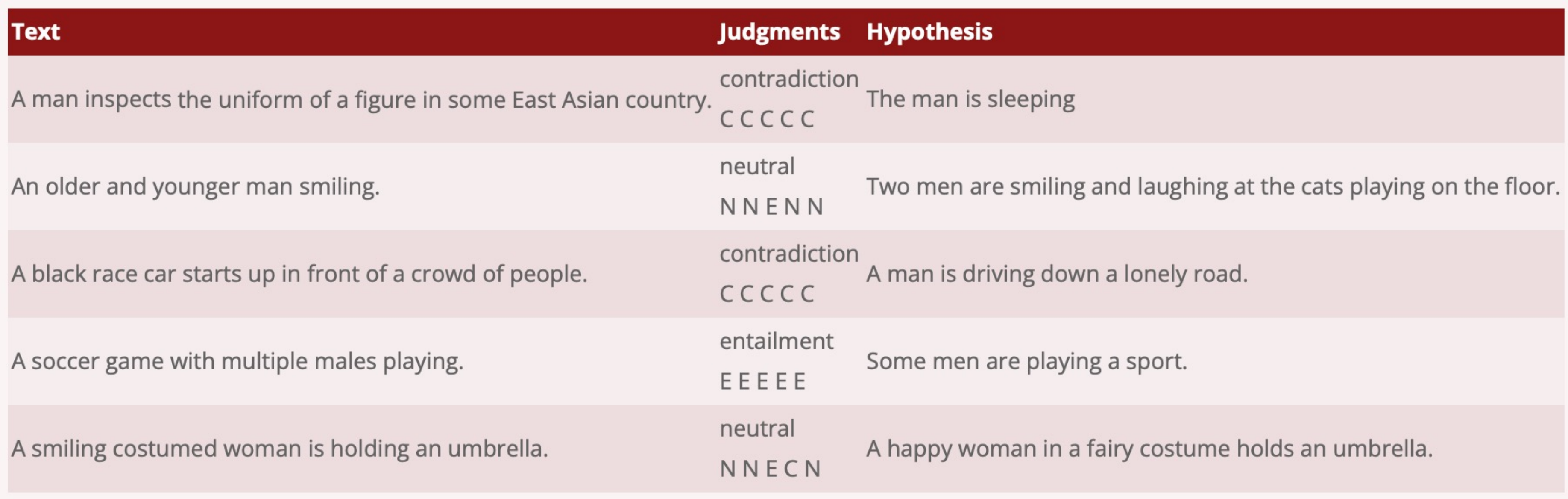
  - IF $MEV_l \rightarrow 0$
    - Static embedding will be poor

  - IF $MEV_l \rightarrow 1$
    - Static embedding can replace for contextualized representations

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

Data

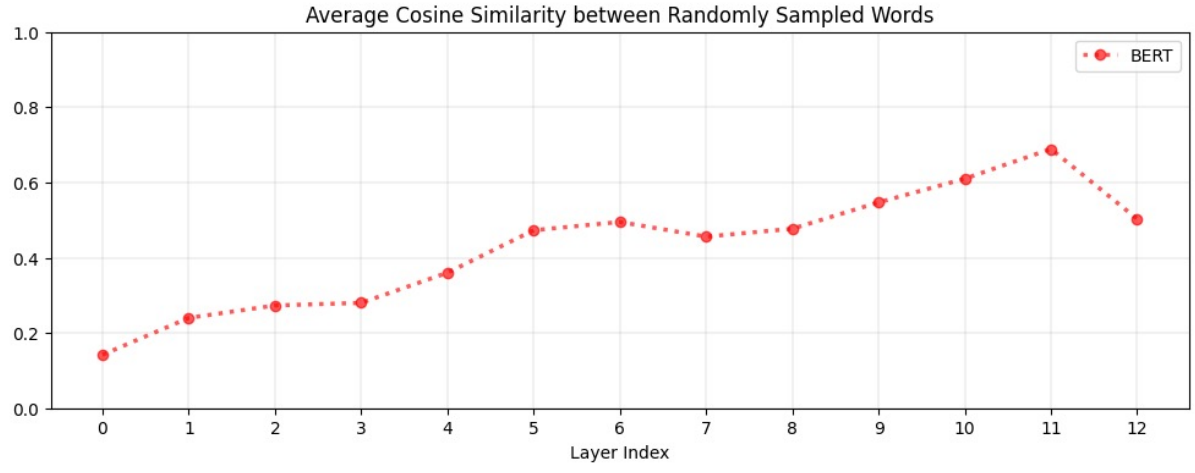
- Stanford Natural Language Inference (SNLI) Corpus
  - https://nlp.stanford.edu/projects/snli/



Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

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

Anisotropy



We randomly select 2 words and calculate the cosine similarity, and find that as layer goes deeper, any words cosine similarity grows →
Any 2 random words have similar representation, which means, the whole vector space tend to be a cone instead of ball.

- How to adjust for it?
  - Use anisotropy baseline

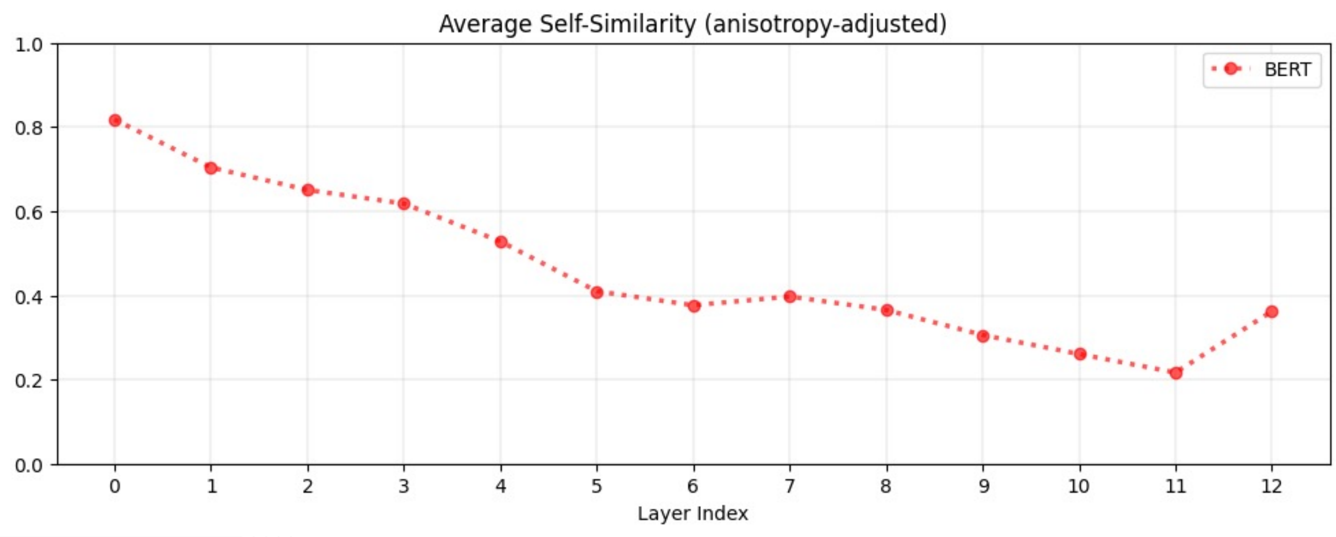
$$Baseline(f_\ell) = \mathbb{E}_{x,y \sim U(\mathcal{O})}\left[\cos(f_\ell(x), f_\ell(y))\right]$$

$$SelfSim^*_\ell(w) = SelfSim_\ell(w) - Baseline(f_\ell)$$

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

Results

- Self-Similarity

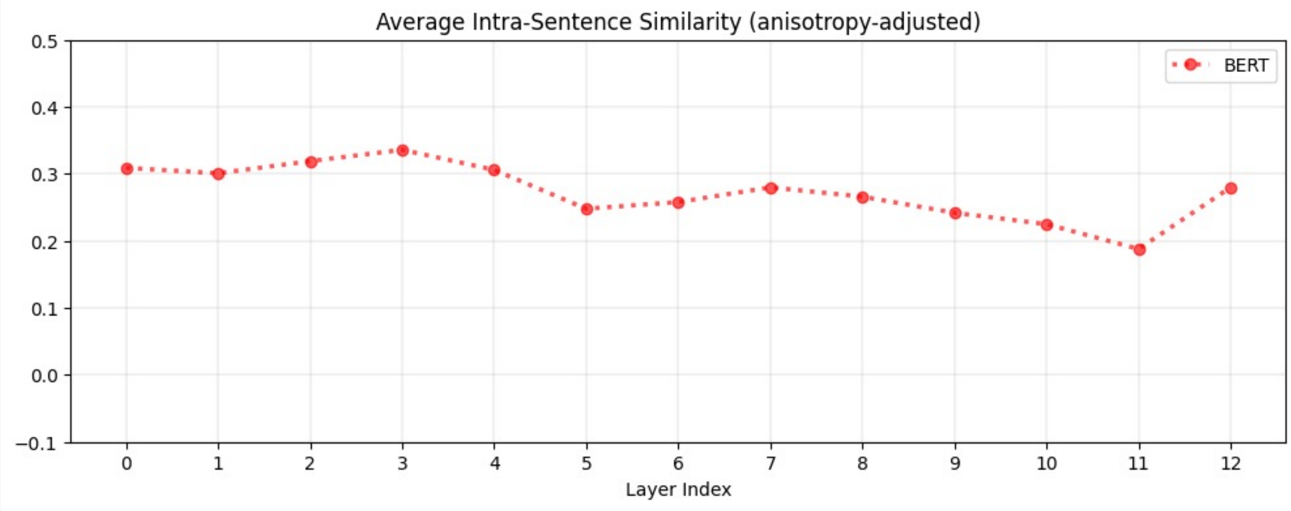


- The higher the layer, the lower the self-similarity, suggesting that contextualized word representations are more context-specific in higher layers.

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

Results

- IntraSimilarity

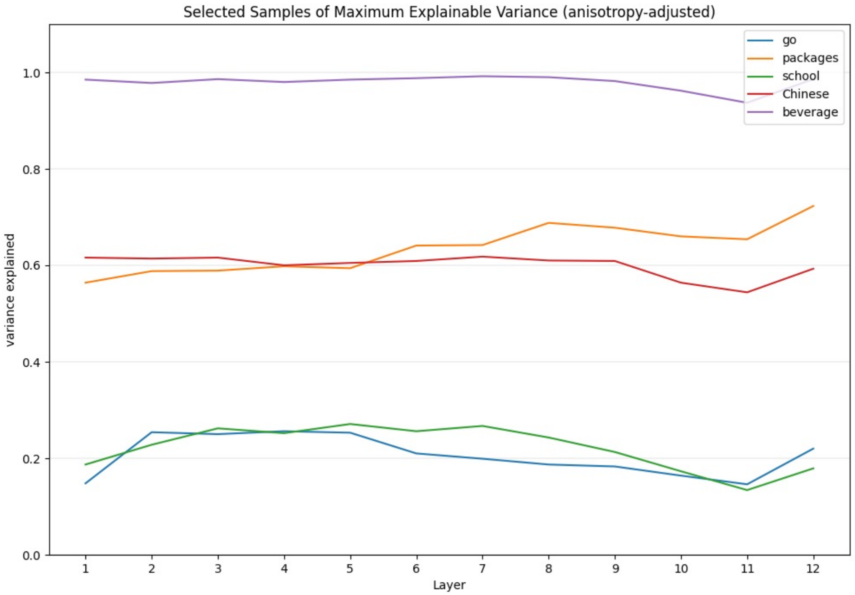


  - As layer goes deeper, 2 words in the same sentence have lower Intra-sentence similarity → 2 words don't have the same meaning even when they share the same context.

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

Results

- Maximum explainable variance

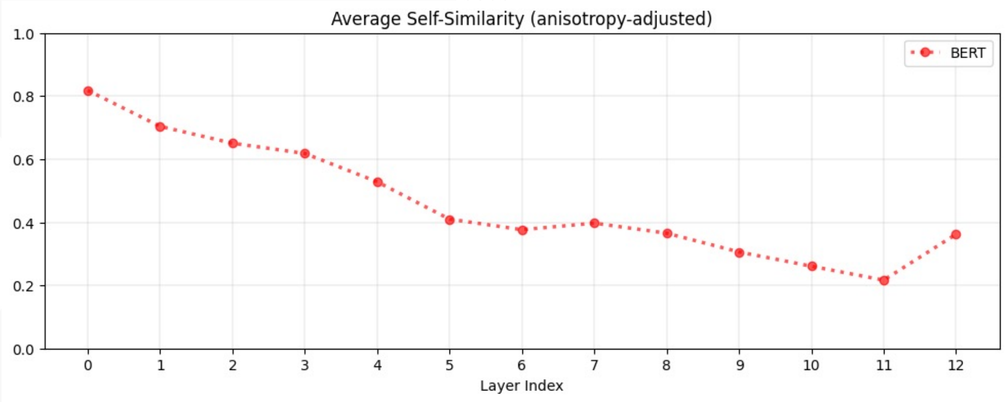


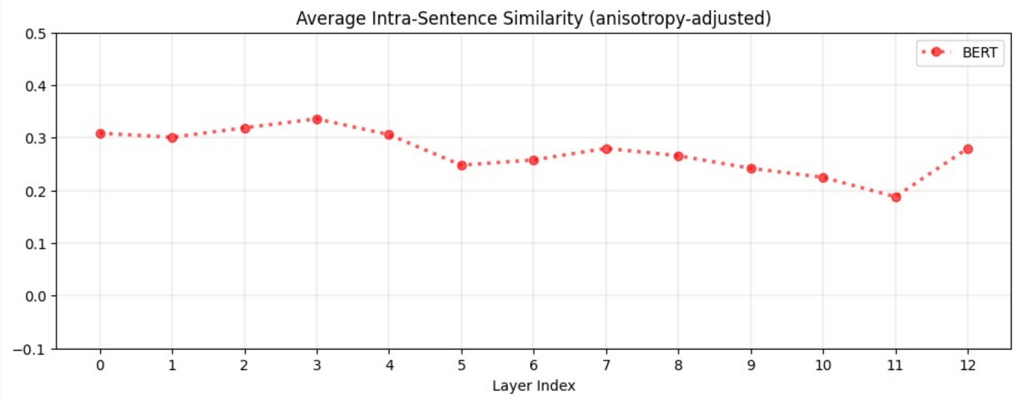
  - We randomly select samples and plot its MEV, for example, ’beverage’ is almost 1, which means in all context , in all layers its meaning is almost the same. And can be replaced by static embedding.
  - While ’go’, ‘school’ may have low multiple meaning in different context.

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

Results

- IntraSimilarity





Interesting Findings:

In layer 12, both SelfSim and intraSim increases, which means

1. Word gets less contexualized, the same word tend to have similar contexlization embedding. (bad)
2. Two words in the same sentence have a similar meaning because they share the same context. (bad)

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

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