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

Kawin Ethayarajh
Stanford University

Выполнил Карлов В.А.
БПМИ193

Статичные vs Контекстуальные эмбеддинги

- Статичные представления слов – не зависят от контекста. Каждому слову всегда ставим в соответствие один и тот же вектор.
- Контекстуальные представления – одно и то же слово кодируем разными векторами в зависимости от контекста.

Анизоторпия

- Анизотропия векторных представлений – это их свойство быть неравномерно распределенными по координатам в векторном пространстве.
- Иными словами, анизотропные векторы лежат в узком конусе.

Ключевые находки

1. Во всех 3-х моделях контекстуальные эмбеддинги анизотропны;
2. В разном контексте представления одного и тоже слова отличаются, причем их различие растет в более глубоких слоях;
3. После поправки на эффект анизотропии, в среднем, менее 5% различий в контекстуальных эмбеддингах слова быть объяснено их первым основным компонентом.

Используемые данные

- В исследовании использовались данные из «SemEval Semantic Textual Similarity tasks 2012 - 2016 » Agirre et al., 2012, 2013, 2014, 2015)
- Они содержат предложения, в которых одно и то же слово встречается в разных контекстах:
  - *A panda dog is running on the road.”*
  - *A dog is trying to get bacon off his back.”*

Метрики контекстуальности

- Пусть w – слово, встречающееся в предложениях $\{s_1, ..., s_n\}$ на позициях $\{i_1, ..., i_n\}$. То есть $w = s_1[i_1] = ... = s_n[i_n]$.
- $f_l(s, i)$ – это представление слова $s[i]$ в слое l.

1) Self-similarity:

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

Интуитивно: среднее косинусное расстояние представлений одного и того же слова (в рамках фиксированного слоя)

Метрики контекстуальности

- Пусть $s = (w_1, \ldots, w_n)$ – предложение, состоящее из n слов.
- $f_l(s, i)$ – это представление слова $s[i]$ в слое l.

2) Intra-sentence similarity :

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

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

Интуитивно: среднее косинусное расстояние представлений слов предложения с усредненным вектором представлений

Метрики контекстуальности

- Пусть w – слово, встречающееся в предложениях $\{s_1, ..., s_n\}$ на позициях $\{i_1, ..., i_n\}$. То есть $w = s_1[i_1] = ... = s_n[i_n]$.
- $f_l(s, i)$ – это представление слова $s[i]$ в слое l.
- Пусть $[f_l(s_1, i_1) ... f_l(s_n, i_n)]$ – матрица представлений. $\sigma_1 ... \sigma_m$ – m первых сингулярных значений этой матрицы.

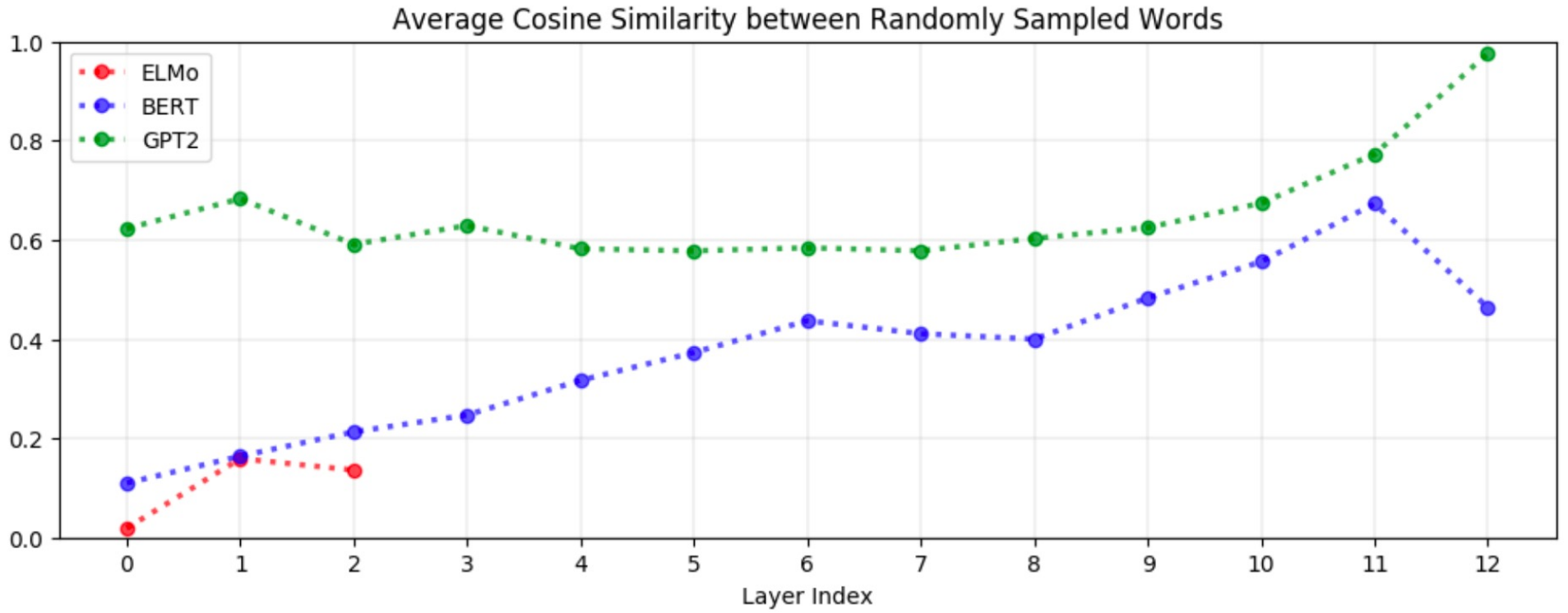
3) Maximum explainable variance:

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

Интуитивно: часть дисперсии представлений слова *w,* которая объясняется первой принципиальной компонентой

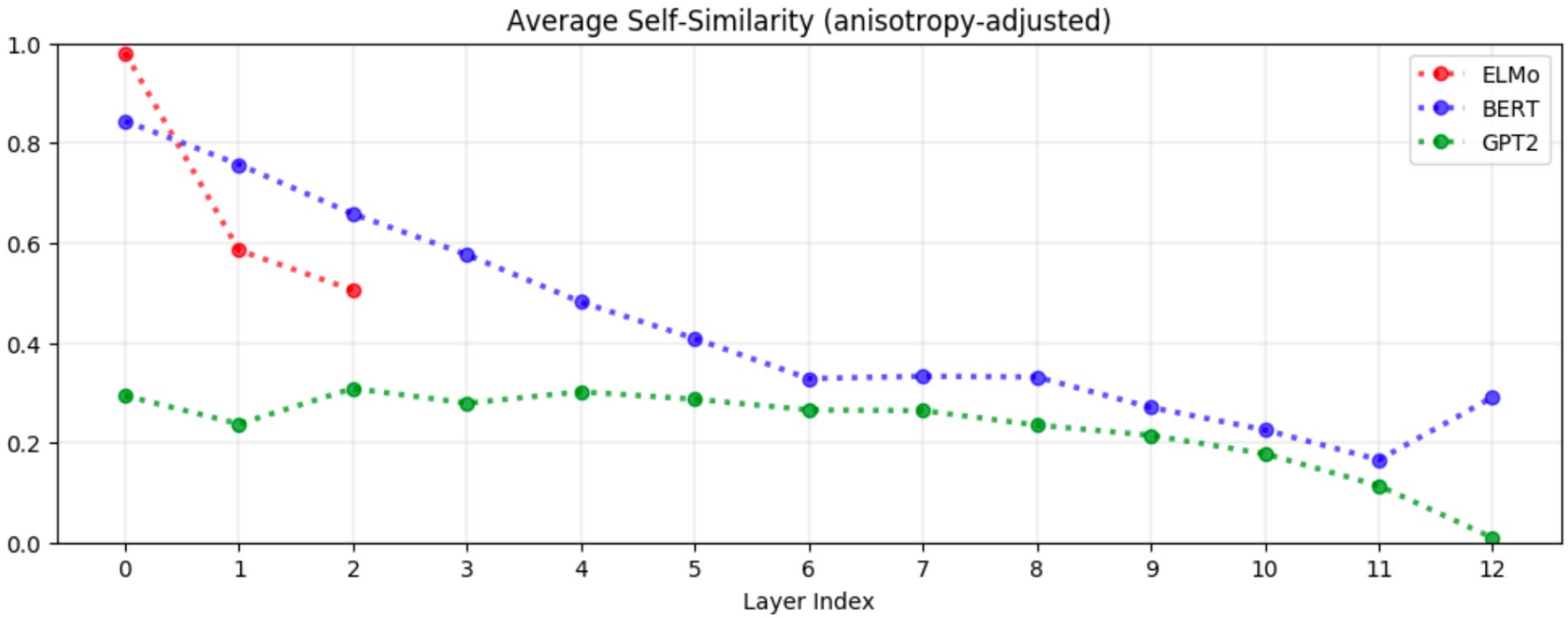
Результаты (про анизотропию)

- Контекстуальные представления анизотропны во всех слоях, кроме начального. Анизотропия растет с глубиной слоя.



Результаты (про контекстуальность)

- Контекстуальность представлений растет с глубиной слоя.



Результаты (про контекстуальность)

- Представления «стоп-слов» имеют наибольшую контекстульность среди всех других слов (наименьшее значение *self-similarity*)
  - “the”
  - “of”
  - ”to”

Результаты (про контекстуальность)

- После поправки на эффект анизотропии, в среднем, менее 5% различий в контекстуальных эмбеддингах слова быть объяснено статичным эмбеддингом.