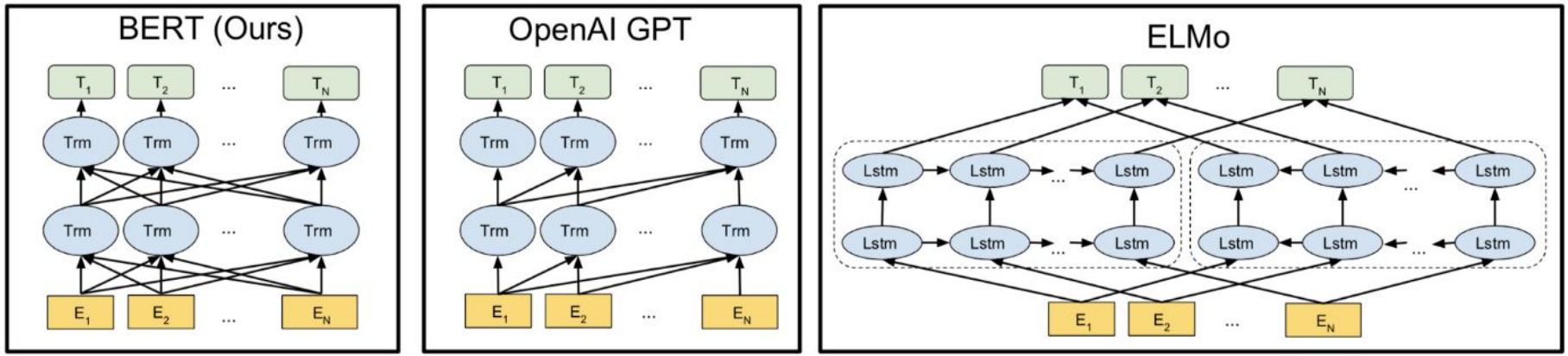
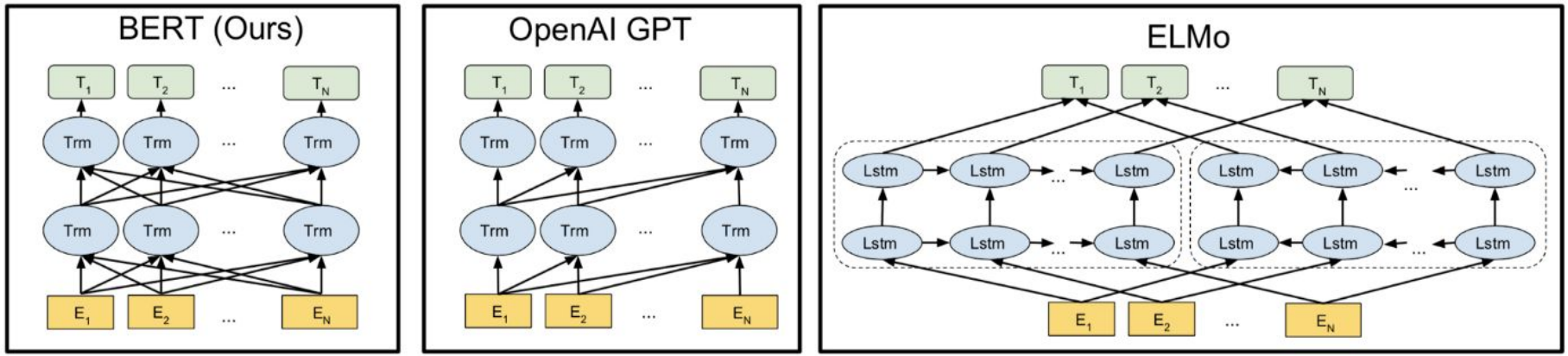
How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings (2019)

Данг Куинь Ньы, Шошин Борис,
Стрельцов Артем, Коган Александра
БПМИ182

Контекстные модели



Контекстные модели



1. Насколько контекстны представления?
2. Конечно ли число векторных представлений для одного слова?

Модели и данные

Модели:

- ELMo (2 скрытых слоя)
- BERT base (12 скрытых слоёв)
- GPT-2 (12 скрытых слоёв)

Данные:

- Из задач SemEval Semantic Textual Similarity (2012 - 2016)
- Слова встречаются хотя бы в пяти разных контекстах

Метрики контекстности

- *w* - слово
- *s_1,…, s_n* - предложения, в которых встречается слово *w* на позициях *i_1,…, i_n* соответственно
- *f_l(s, i)* - функция, возвращающая представление *s[i]* на слое *l* модели *f*

Метрики контекстности

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

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

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

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

- *w* - слово
- *s_1,…, s_n* - предложения, в которых встречается слово *w* на позициях *i_1,…, i_n* соответственно
- *f_l(s, i)* - функция, возвращающая представление *s[i]* на слое/модели *f*

Метрики контекстности

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

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

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

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

- *w* - слово
- *s_1,…, s_n* - предложения, в которых встречается слово *w* на позициях *i_1,…, i_n* соответственно
- *f_l(s, i)* - функция, возвращающая представление *s[i]* на слое/модели *f*

Метрики контекстности

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

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

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

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

- *w* - слово
- *s_1,…, s_n* - предложения, в которых встречается слово *w* на позициях *i_1,…, i_n* соответственно
- *f_l(s, i)* - функция, возвращающая представление *s[i]* на слое/модели *f*
- *σ_1...σ_m* - первые *m* сингулярных чисел матрицы *[f_l(s_1, i_1)... f_l(s_n, i_n)]*

Анизотропия

Зависимость метрик от распределения векторов:

1. векторы равномерно распределены

   --> *Self-Sim* = 0.95 - плохая контекстуализация

2. векторы анизотропны (например, косинусная близость в среднем равна 0.99)

   --> *Self-Sim* = 0.95 - хорошая контекстуализация

Анизотропия

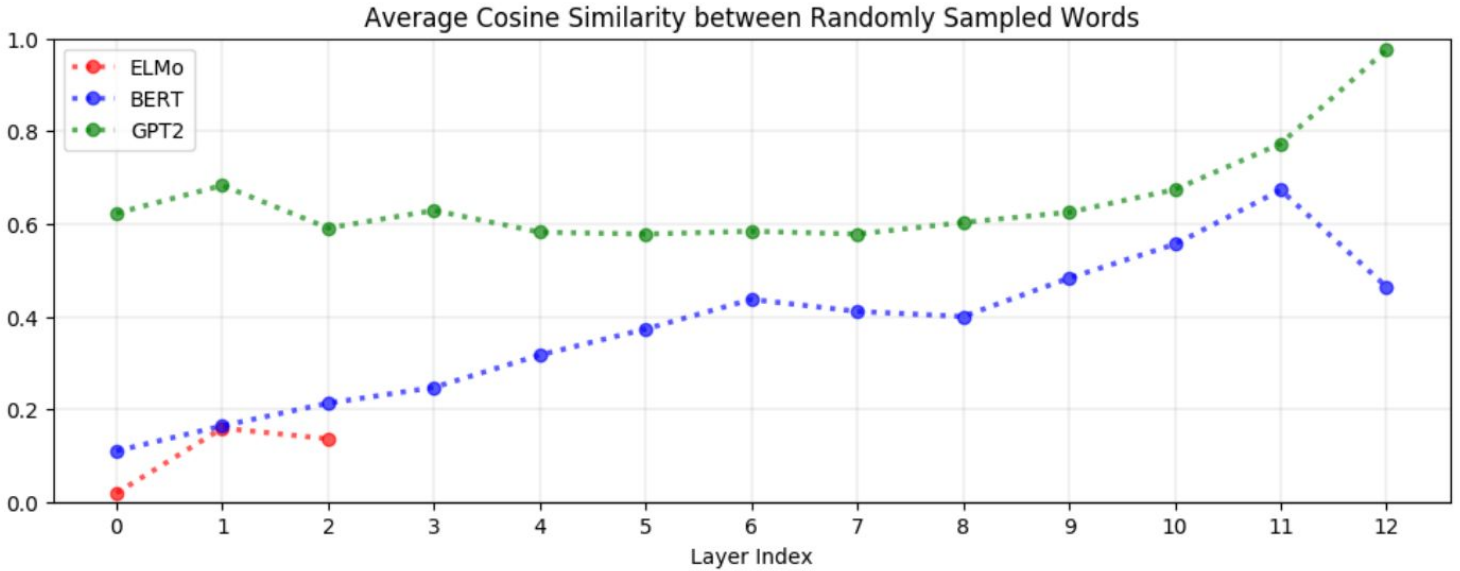
Введём *“анизотропные бейзлайны”*:

- отдельный для каждой из трёх метрик, для каждого слоя модели
- представления выбираются из равномерного распределения
- чем анизотропнее представления в слое, тем ближе бейзлайн к 1

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

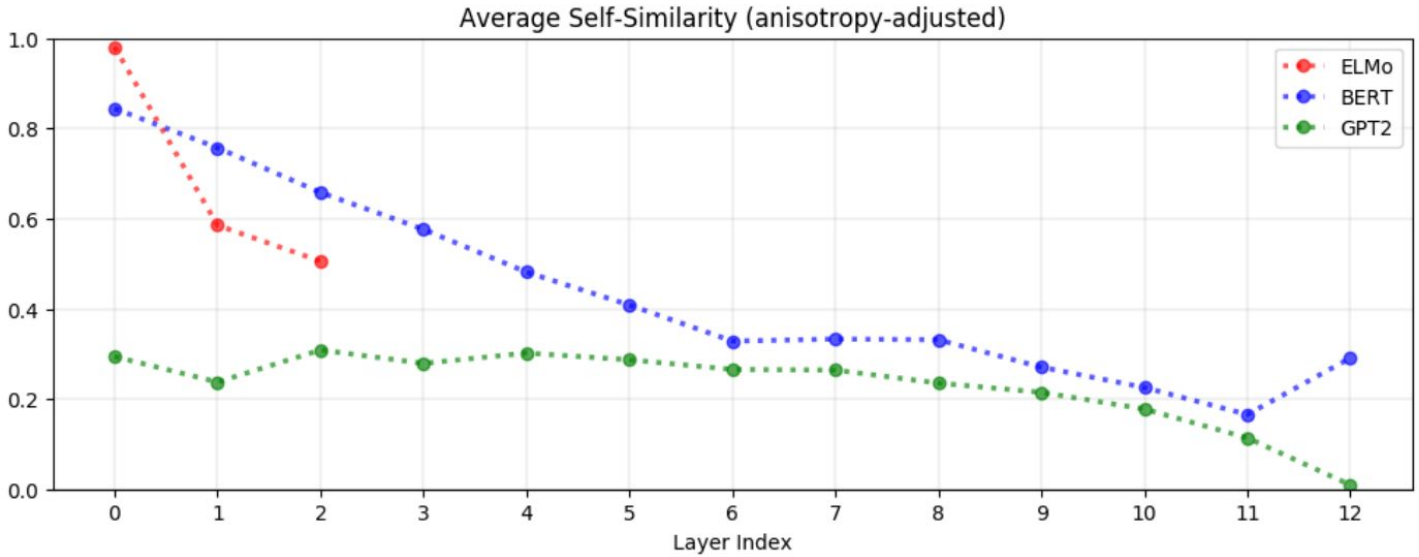
$$SelfSim^*_\ell(w) = SelfSim_\ell(w) - Baseline(f_\ell)$$

Результаты. Анизотропия



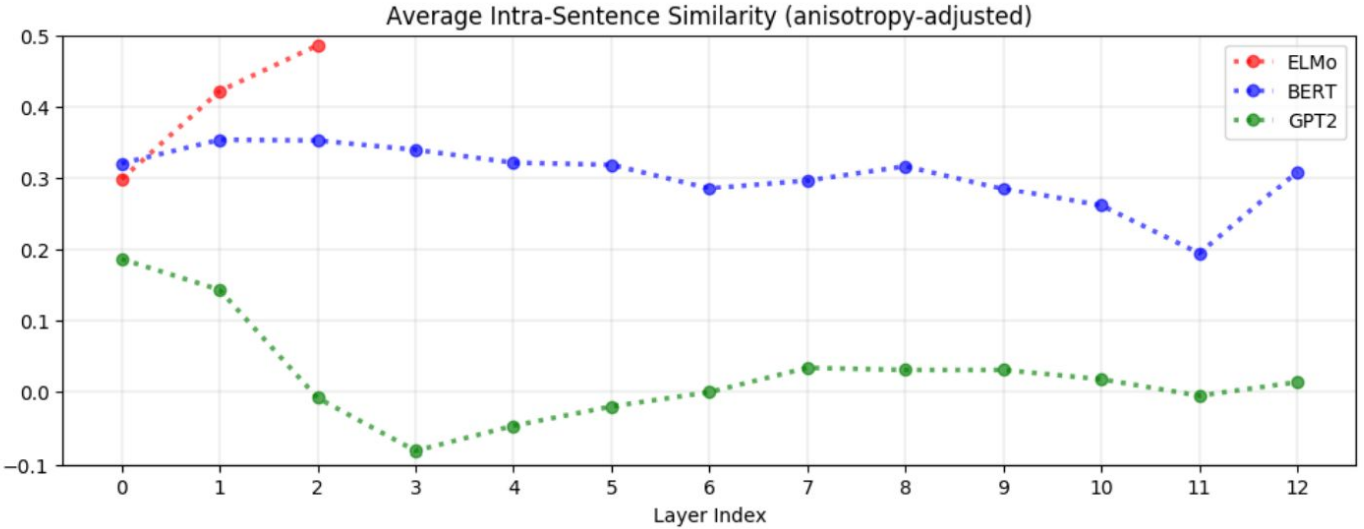
- во всех (кроме вводных) слоях контекстные представления анизотропны
- представления более анизотропны на последних слоях

Результаты. Контекст-специфичность



- чем глубже слой, тем более контекстно-специфичны представления
- наиболее контекстно-специфичные представления у стоп-слов!

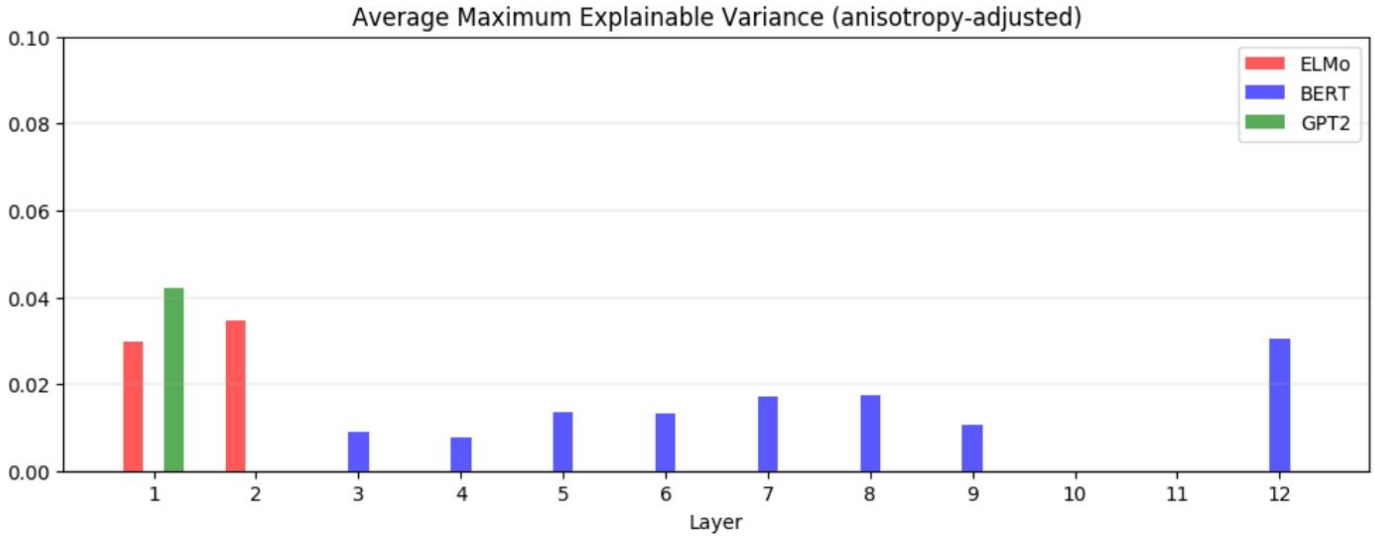
Результаты. Контекст-специфичность



Представления слов из одного предложения:

- ELMo: *более* похожи друг на друга в последних слоях
- BERT: *менее* похожи друг на друга в последних слоях
- GPT-2: похожи друг на друга *не более чем случайно выбранные слова*

Результаты. Контекстные и статичные представления



- не более 5% дисперсии контекстных представлений объясняется статичными представлениями

Результаты. Контекстные и статичные представления

Static Embedding	SimLex999	MEN	WS353	RW	Google	MSR	SemEval2012(2)	BLESS	AP
GloVe	0.194	0.216	0.339	0.127	0.189	0.312	0.097	0.390	0.308
FastText	0.239	**0.239**	**0.432**	0.176	0.203	0.289	0.104	0.375	0.291
ELMo, Layer 1	0.276	0.167	0.317	0.148	0.170	0.326	0.114	**0.410**	0.308
ELMo, Layer 2	0.215	0.151	0.272	0.133	0.130	0.268	0.132	0.395	0.318
BERT, Layer 1	0.315	0.200	0.394	**0.208**	**0.236**	**0.389**	**0.166**	0.365	**0.321**
BERT, Layer 2	**0.320**	0.166	0.383	0.188	0.230	0.385	0.149	0.365	**0.321**
BERT, Layer 11	0.221	0.076	0.319	0.135	0.175	0.290	0.149	0.370	0.289
BERT, Layer 12	0.233	0.082	0.325	0.144	0.184	0.307	0.144	0.360	0.294
GPT-2, Layer 1	0.174	0.012	0.176	0.183	0.052	0.081	0.033	0.220	0.184
GPT-2, Layer 2	0.135	0.036	0.171	0.180	0.045	0.062	0.021	0.245	0.184
GPT-2, Layer 11	0.126	0.034	0.165	0.182	0.031	0.038	0.045	0.270	0.189
GPT-2, Layer 12	0.140	-0.009	0.113	0.163	0.020	0.021	0.014	0.225	0.172

- использование главных компонент контекстных представлений на первых слоях лучше FastText и GloVe на многих бенчмарках

Выводы

- у фиксированного слова, скорее всего, неограниченное количество контекстных представлений
- контекстные представления высоко анизотропны
  - предлагается ввести в задачах штраф на анизотропию
- из контекстных представлений можно получить хорошие статичные представления слов

Содержание и вклад работы

Данная работа показывает насколько различны контекстуальные эмбеддинги для одного и того же слова в разных предложениях. Автор сравнивает их на примере таких популярных моделей, как BERT, ELMo и GPT-2.

Было обнаружено, что верхнии слои этих моделей создают более контекст зависимые представления, чем нижнии. Также было обнаружено, что эмбеддинги одного и того же слова в разных контекстах наиболее похожи для ELMo и наиболее различны для GPT-2

Плюсы работы

- Обоснованность логических выводов и высокое качество написанного текста
- Полнота раскрытия темы. Автор сравнил эмбединги на разных моделях, слоях моделей и разных типах слов
- Научная новизна статьи. До этого не было статей на эту тему

Минусы работы

Малая практическая ценность. Результаты представляют интерес в основном в исследовательских целях

Узкая область проводимого исследования: задача исследования различности контекстуальных эмбедингов является довольно специфичной

Недостаточное описание используемых метрик. Автор не указал являются эти метрики широко используемыми или придуманными специально для этой работы

Трудность воспроизведения результатов, полученных автором. При использовании другого корпуса текстов результаты получились сильно отличные от результатов статьи

Заключение

Автор проделал хорошее и детальное исследование выбранной им темы. Но несмотря на это в работе есть некоторые недостатки. Исходя из этого, я считаю, что автор заслуживает оценки 8/10. В своей оценке я уверен на 4/5.

Информация о статье

- Выпущена в 2019 году, представлена в виде устного доклада на конференции EMNLP 2019
- Автор -- PhD из Стэнфорда, работает в Facebook, специализируется на representation learning и оценке моделей
- Много работ по исследованию моделей, в том числе их геометрии, например, “Rotate *King* to get *Queen*: Word Relationships as Orthogonal Transformations in Embedding Space”.
- Скорее всего, в данной статье есть вдохновение от работ “The strange geometry of skip-gram with negative sampling” (2017) и “What do you learn from context? Probing for sentence structure in contextualized word representations” (2019)

Информация о статье

- У статьи 169 цитирований (немало)
- В частности, большинство цитирований связаны с исследованием геометрии представлений и их контекстуальности, например, в статье “Circles are like Ellipses, or Ellipses are like Circles? Measuring the Degree of Asymmetry of Static and Contextual Embeddings and the Implications to Representation Learning”.
- Прямых конкурентов у статьи нет, есть разве что некоторое пересечение со статьями примерно того же времени, например, про замечание о косинусной близости на различных слоях

Практическое применение

- Ввести штраф за анизотропию (получится что-то вроде регуляризации для того, чтобы модель была более task-specific)
- В целом статья более исследовательская, поэтому какого-то прямо целенаправленного применения в продакшне нет, но зато она позволяет лучше понять устройство описанных в статье моделей

Источники

- Основная статья: https://arxiv.org/pdf/1909.00512.pdf
- Ноутбук с примерами: https://colab.research.google.com/drive/17aQRrHIOMC-doCTbzsWEdIuzgfcm3f-H?usp=sharing