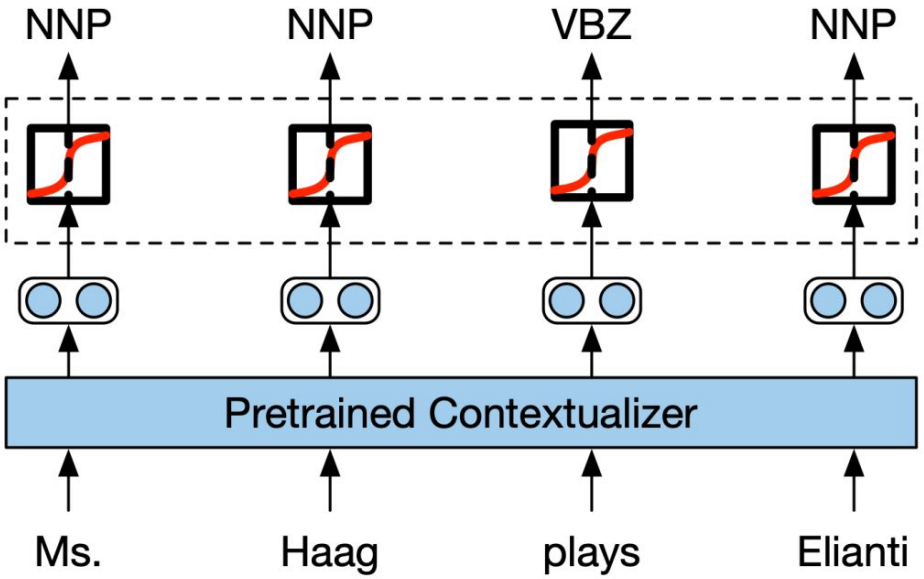
Контекстность эмбеддингов

Григорьев Петр, Сапожникова Дарья,
Хамдеева Дилара, Константин Матвеев

Контекстные эмбеддинги

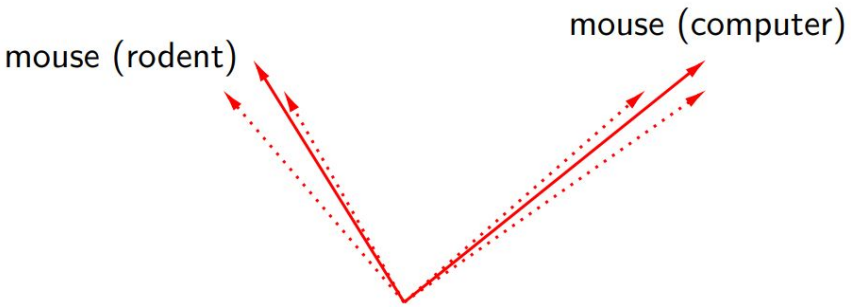
- **До 2018:** статические эмбеддинги (skipgram, GloVe, etc.)
- **После 2018:** контекстные эмбеддинги (ELMo, BERT, etc.)
- Контекстные **>>** статические, полезны даже вне модели



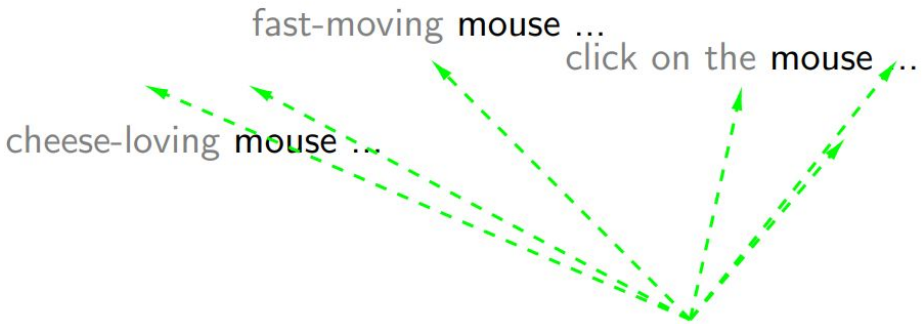
Интуиция вопроса

Насколько контекстные эмбеддинги зависят от контекста?

Слово - один из векторов набора, соответствующего смыслам слова



Бесконечно много представлений, зависящих от контекста



Интуиция вопроса

Насколько контекстные эмбеддинги зависят от контекста?

A panda dog is running on the road.

A dog is trying to get bacon off its back.

$\vec{dog} = \vec{dog} \Longrightarrow$ *Нет зависимости от контекста*

$\vec{dog} \neq \vec{dog} \Longrightarrow$ *Зависимость от контекста*

Постановка вопроса

Насколько контекстные эмбеддинги зависят от контекста?

---

1) Насколько различны представления одного слова?
2) Схожи ли представления слов в одном контексте (предложении)?
3) Насколько можно заменить контекстные представления статичными?

Используемые модели

ELMo, 2018, 2 слоя

BERT, 2018, 12 слоев

GPT-2, 2019, 12 слоев

Используемые метрики

SelfSim, сходство с собой

IntraSim, сходство внутри предложения

MEV, максимально объяснимая дисперсия

OpenAI

Используемые метрики

Ниже SelfSim, сходство с собой

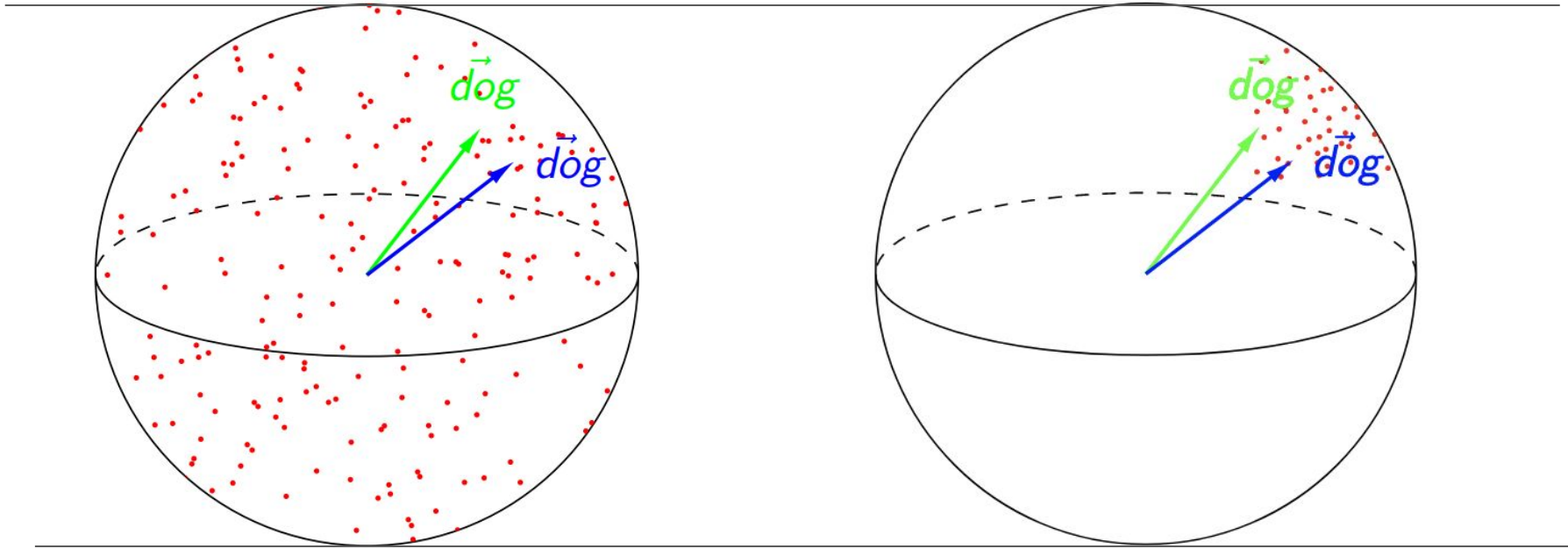
Выше IntraSim, сходство внутри предложения

Ниже MEV, максимально объяснимая дисперсия

Зависимость от контекста!

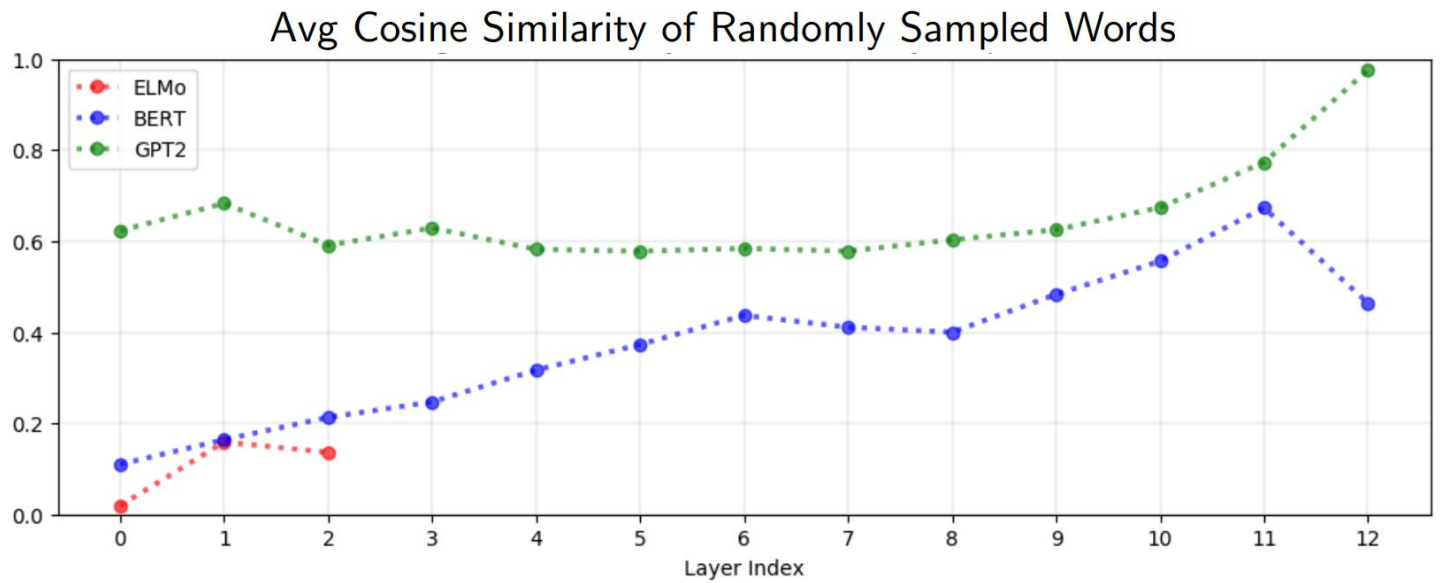
Анизотропия

Векторные представления не равномерно распределены в пространстве



Анизотропия

Векторные представления не равномерно распределены в пространстве

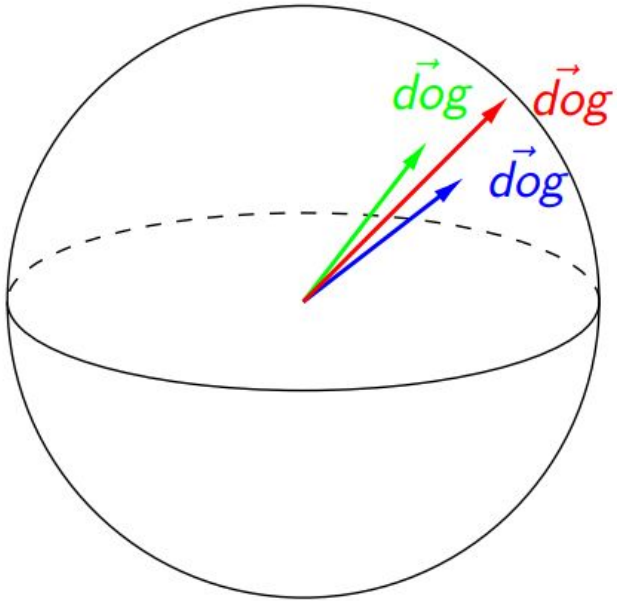


SelfSim

Средняя косинусная близость представлений одного слова

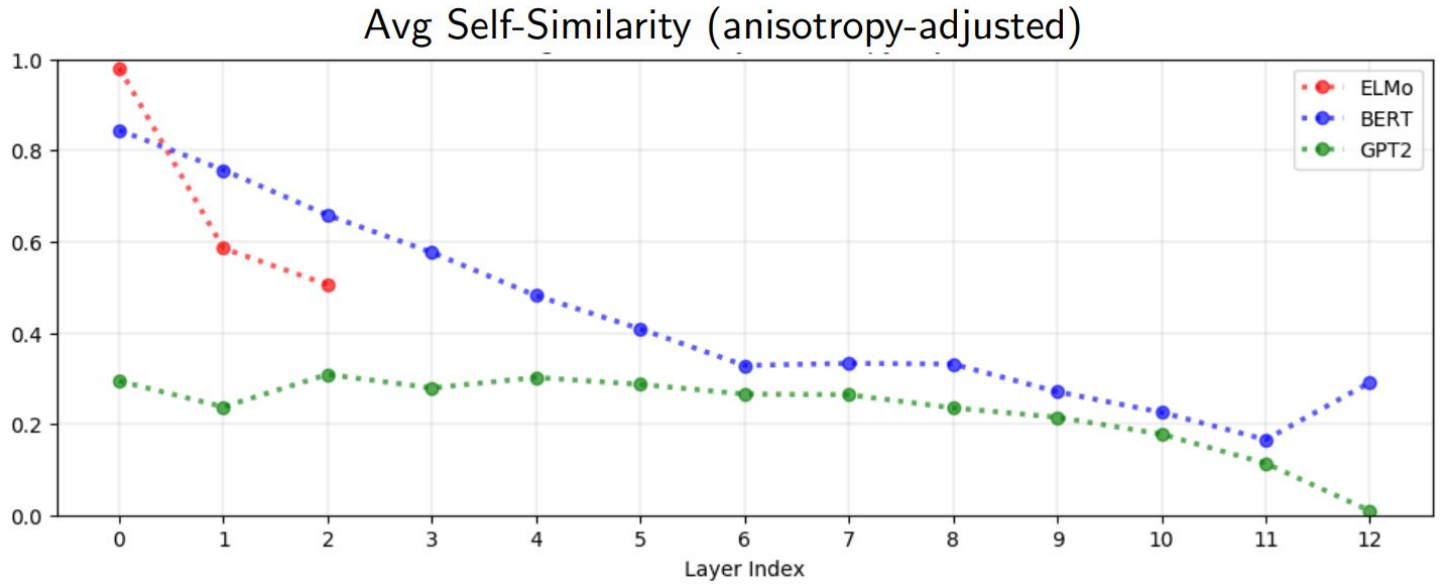
- Представления берутся из одного слоя
- W - слово, *s[i] = s[j] = w*
- $f_l(s, i)$ - представление слова *s[i]* на слое *l*
- *Вычитаем сходство пар случайных слов*

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



Представления одного слова

1) Больше зависят от контекста в более высоких слоях
2) Самая высокая зависимость у шумовых слов (stopwords, ‘the’, ‘of’)
3) Различие в контекстах, а не полисемия, влияет на различия



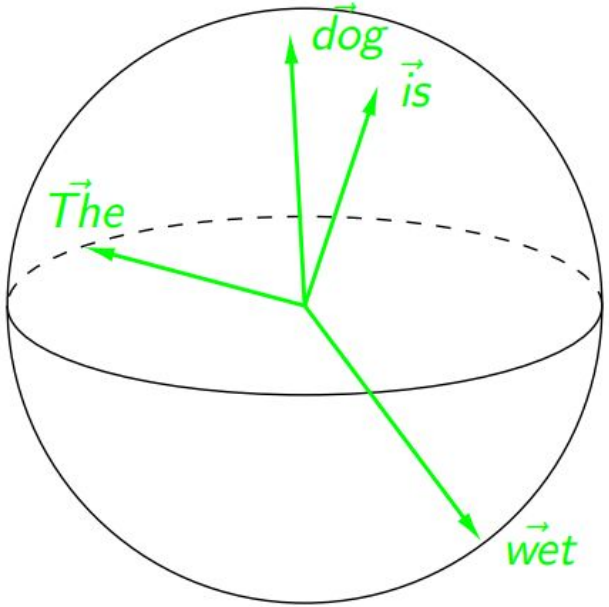
IntraSim

Средняя косинусная близость между словом и контекстом

- Представления берутся из одного слоя
- W - слово, *s[i] = s[j] = w*
- $f_l(s, i)$ - представление слова *s[i]* на слое *l*
- *Вычитаем сходство пар случайных слов*

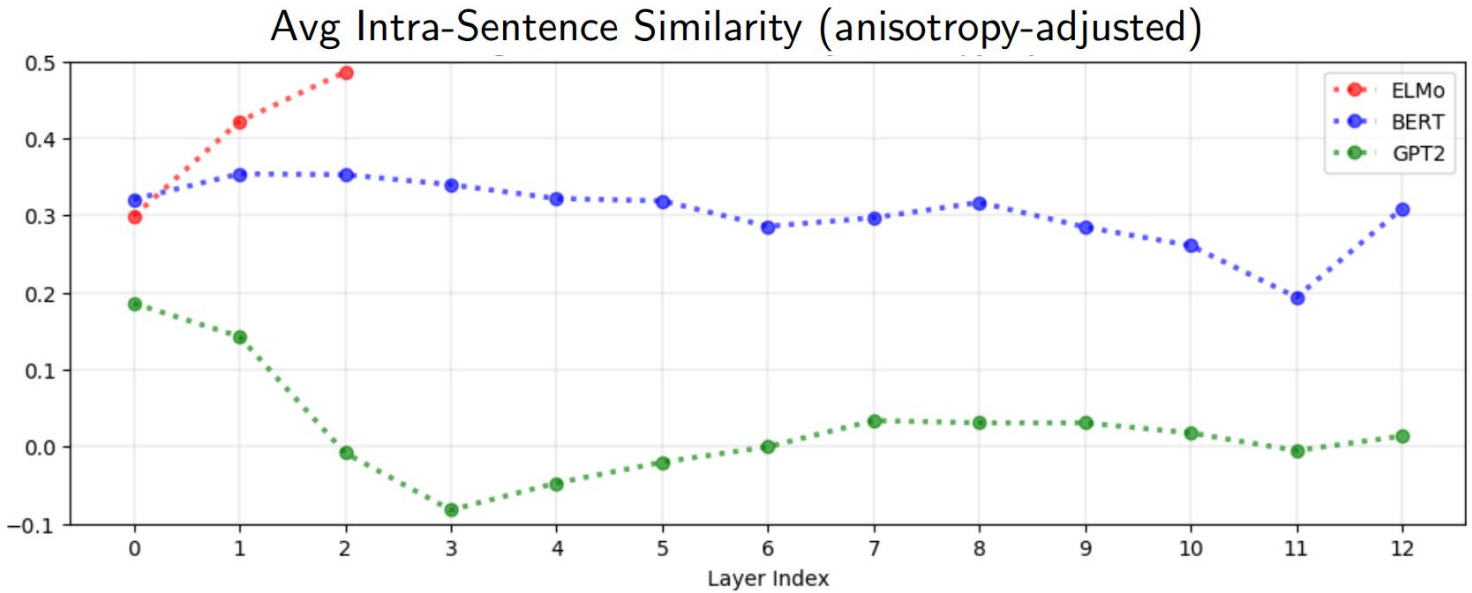
$$IntraSim_\ell(s) = \frac{1}{n}\sum_i \cos(\vec{s}_\ell, f_\ell(s,i))$$

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



Слова в одном контексте

1) ELMo - высокая схожесть на высоких слоях
2) BERT - схожесть падает на более высоких слоях
3) GPT-2 - схожесть, сравнимая с парой случайных слов

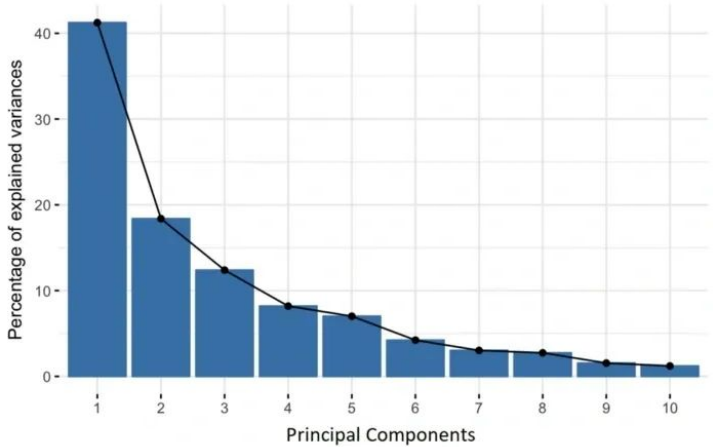


MEV

Доля дисперсии, объяснимая наибольшим собственным значением

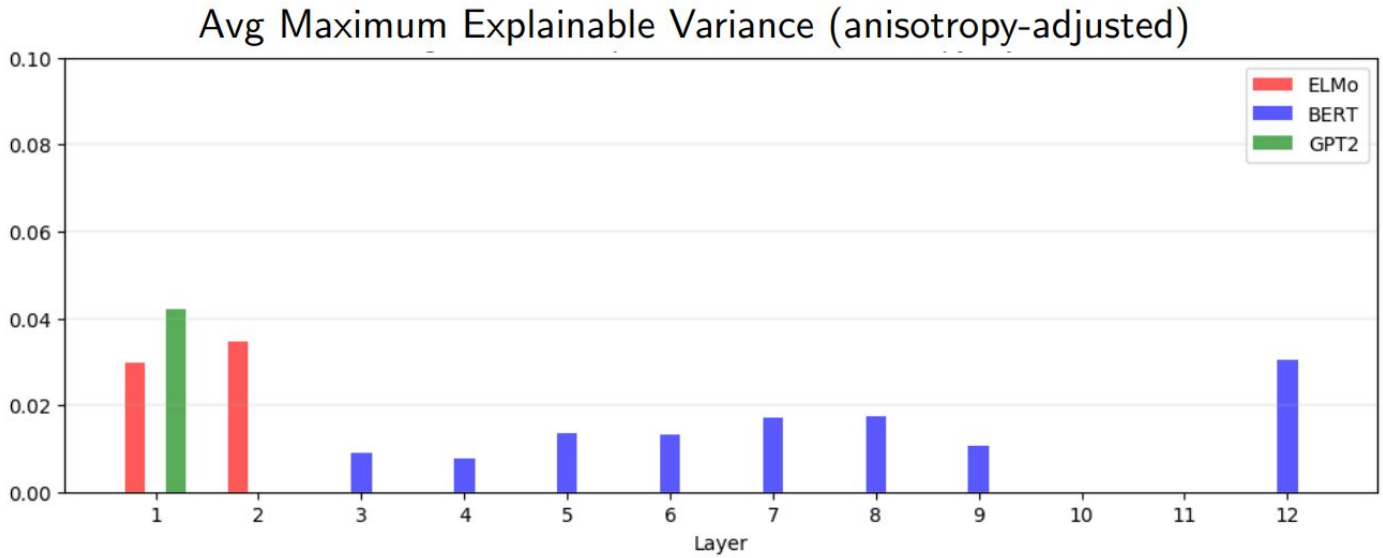
- Представления берутся из одного слоя
- Столбцы матрицы - представления слова
- $\sigma_1 \dots \sigma_m$ - собственные значения матрицы
- Вычитаем MEV набора случайных слов

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



Статические представления

1) Менее 5% дисперсии объяснимо статическими представлениями
2) Дисперсия всех слов лучше объяснима, чем дисперсия одного слова
3) Наблюдение опровергает гипотезу, что эмбеддинги соответсвуют смыслам слова



Статические представления

Возьмем главные компонент матриц всех векторов одного слова на слое

Такие представления обгоняют классические эмбеддинги на многих бенчмарках

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

Рецензия

Сильные стороны

- Все выводы,сделанные в статье, обоснованы и явно вытекают из экспериментов, подробное описание которых присутствует в тексте. Необходимые для интерпретации величины средних значений для введенных метрик также выведены конкретно под эту задачу.
- Рассматриваемые контекстные представления достаточно молоды, и на момент написания статьи проведенные исследования, изучающие их поведение фокусировались на изучении наличия влияния контекста на модель.
- Данный метод впервые был применен по отношению к контекстным векторных представлений (до этого подобные исследования проводились только со статичными представлениями).

Слабые стороны

- Хотелось бы понять логику выбора именно этих статических эмбеддингов для сравнения. Нам вкратце описывается что это - одни из лучших существующих моделей. Однако, лучше было бы также указать, как они ведут себя на выбранных наборах данных относительно других статических векторных представлений. Может быть под конкретно эту задачу кто-то из прямых конкурентов работает лучше.
- В работе вводятся метрики, для понимания численных значений которых требуется сравнение с некоторым бейзлайном. Однако, для метрики MEV данная величина опущена в тексте,, однако на нее ссылаются

> “ Though not visible in Figure 4, the raw MEV of many words is actually below the anisotropy baseline”

Текст статьи

В целом, статья написана понятным языком. Логические связи внутри текста не нарушены.

Единственным крупным замечанием можно выделить, что в обзоре литературы упоминается статья, которая использует подобный подход, однако ни здесь ни далее не уточняется в чем заключается степень похожести.

> “It is more similar to Mimno and Thompson (2017), which studied the geometry of static word embedding spaces.”

Воспроизводимость экспериментов

- Эксперименты описаны четко, нет опущенных шагов.
- Указаны версии предобученных моделей, которые были использованы:

  > “We use the pretrained models provided in an earlier version of the PyTorch-Transformers library”

- Также указаны, какие наборы данных были использованы:

  > “Our input data come from the SemEval Semantic Textual Similarity tasks from years 2012 - 2016 (Agirre et al., 2012, 2013, 2014, 2015)”

- Необходимые для экспериментов метрики, введенные в статье, подробно расписаны.

Список источников

- Kawin Ethayarajh, How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. EMNLP 2019 (oral).
- Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. arXiv preprint arXiv:1903.08855, 2019.
- https://github.com/kawine/contextual
- https://github.com/sonsus/albert_paraphras