Word Embeddings

Sergey Aksenov

Higher School of Economics

13 сентября 2021 г.

Векторные представления

- *Векторное представление* (embedding) — сопоставление произвольному объекту некоторого числового вектора в пространстве фиксированной размерности
- Наиболее известный вид – векторные представления слов (word embedding)
- Векторы могут обладать разнообразными полезными свойствами, отражать близость объектов в разных смыслах
- Для слов это может быть семантическая близость

Зачем нужны векторные представления

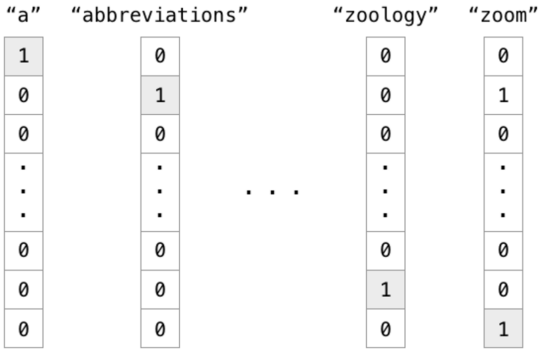
В современных подходах эмбеддинги используются в качестве признаков для решения почти любых задач машинного обучения

В текстовой аналитике это:

1. выделение именованных сущностей (NER)
2. выделение частей речи (POS-tagging)
3. машинный перевод
4. кластеризация документов
5. классификация документов, анализа тональности (sentiment)
6. ранжирование документов
7. генерация текста

One-hot encoding

Самый простой способ кодирования категориальных признаков:



Полученные векторы огромные и ортогональные

Ссылка на источник картинки

SVD для получения эмбеддингов слов

1. По корпусу текстов D со словарём T строим *матрицу со-встречаемостей* $X_{|T|\times|T|}$

   Возможны различные варианты учёта со-встречаемости слов:

   - сумма по всей коллекции числа попаданий пары слов в окно фиксированного размера
   - количество документов, хоть раз содержащих пару слов
   - количество документов, хоть раз содержащих пару слов в окне

   Понижаем размерность:

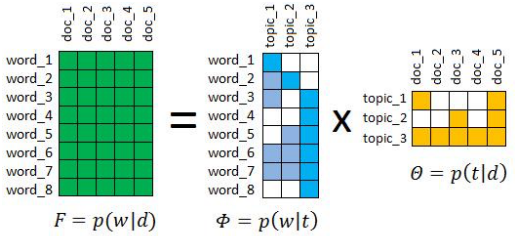
2. SVD-разложение: $X = USV^T$

3. Из столбцов матрицы U выбираются первые K компонент

Недостатки SVD

1. Относительно низкое качество получаемых представлений
2. Сложность работы с очень большой и разреженной матрицей
3. Сложность добавления новых слов/документов (решается инкрементальными методами построения)

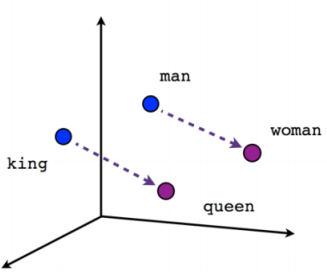
Тематическое моделирование

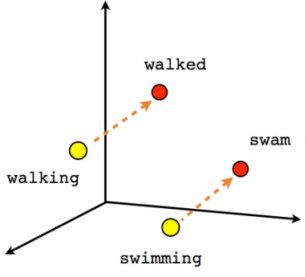


- Классические тематические модели получают на вход матрицу «мешка слов» или tf-idf и строят два типа распределений:
  - слов в кластерах-темах
  - тем в документах
- По факту получается стохастическое матричное разложение
- Строки матрицы «слова-темы» можно использовать в качестве эмбеддингов
- Современные реализации инкрементально обучаются на больших данных
- Как и SVD, простые TM не учитывают локальный контекст

word2vec

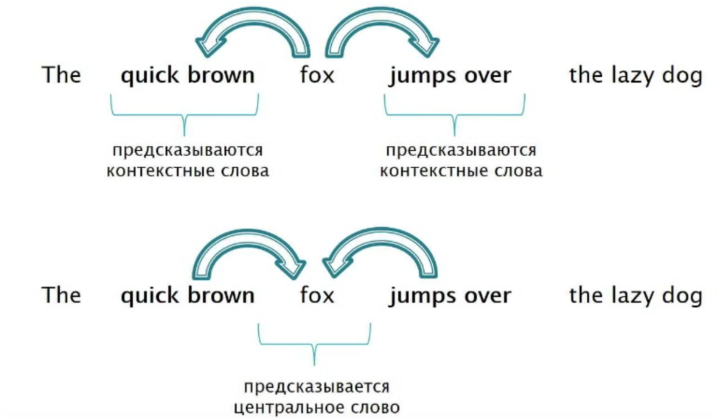
- **word2vec** — группа алгоритмов, предназначенных для получения вещественных векторных представлений слов
- **Идея:** «Слова со схожими значениями разделяют схожий контекст»
- Как правило, в векторном представлении семантически близкие слова оказываются рядом





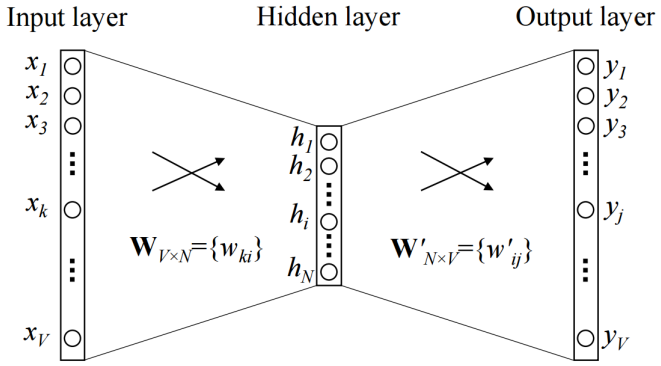
Don't count, predict!

Две модели: **Skip-gram** и **Continuous BOW**

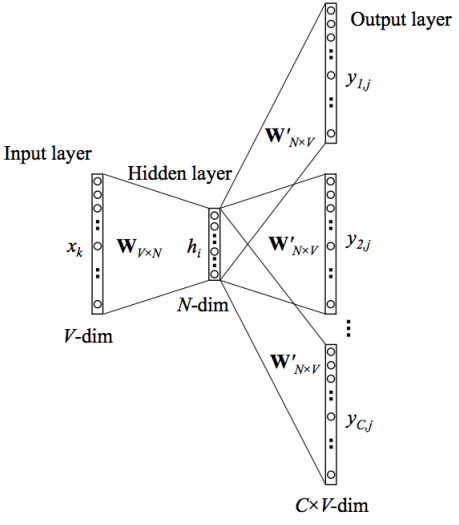


Ссылка на источник картинки

Модель CBOW (единичный контекст)



Модель Skip-gram



Производительность обучения Skip-gram

Подсчёт softmax — вычислительно дорогая операция

Применяются различные методы аппроксимации:

1. Softmax-based
   - **Иерархический softmax**
   - Дифференциальный softmax
   - CNN-softmax
   - ...
2. Sampling-based
   - **Negative Sampling**
   - Noise Contrastive Estimation
   - ...

---

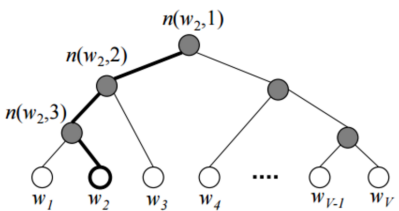
http://ruder.io/word-embeddings-softmax/index.html

Иерархический softmax

Вычисление softmax — дорогая операция, $O(V)$
Иерархический softmax — $O(log(V))$

Построим бинарное дерево, листьями которого будут уникальные слова коллекции (например дерево Хаффмана)

Выходы скрытого слоя связываются с внутренними узлами дерева (V-1 штук)



Ссылка на источник картинки

Иерархический softmax

Идея: в процессе обучения при фиксированном контексте нас интересует только предсказываемое слово

Вероятность того, что w будет выходным словом:

$$p(w = w_{out}) = \prod_{j=1}^{L(w)-1} \sigma(\{n(w, j+1) = lch(n(w, j))\} v_{n(w,j)}^T u)$$

- $L(w)$ — длина пути от корня до слова w
- $n(w, j)$ — j-я вершина на этом пути
- $\sigma(x)$ — сигмоида
- $\{\text{true}\} = +1$, $\{\text{false}\} = -1$
- $lch(n)$ — левый потомок вершины n
- $u = v_{w_{inp}}$ в случае skip-gram, $u = 1/h \sum_{k=1}^{h} v_{inp,k}$ (усреднённый вектор контекста) в случае CBOW

Иерархический softmax

Вероятность того, что w будет выходным словом:

$$p(w = w_{out}) = \prod_{j=1}^{L(w)-1} \sigma(\{n(w, j+1) = lch(n(w, j))\} v_{n(w,j)}^T u)$$

Считаем множество бинарных вероятностей, на каждом шаге можно пойти налево или направо с вероятностями

$$p(n, left) = \sigma(v_n^T u)$$
$$p(n, right) = 1 - p(n, left) = \sigma(-v_n^T u)$$

Затем на каждом шаге вероятности перемножаются и получается искомая формула

Negative Sampling

- Можно изменить постановку задачи и функционал качества.
- Решаем задачу бинарной классификации: $z = 1$ — пара $(w, s) \in D$, $z = 0$ — нет. $(s \in c(w))$

$$p(z = 1|(w, s)) = \frac{1}{1 + \exp(-v_w^T v_s)} = \sigma(v_w^T v_s)$$

- Запишем новый функционал правдоподобия:

$$\mathcal{L} = \sum_{(w,s)\in D_1} \log \sigma(v_w^T v_s) + \sum_{(w,s)\in D_2} \log \sigma(-v_w^T v_s),$$

$$D_1 = \{(w, s) : s \in c(w)\}, \quad D_2 = \{(w, c) : s \notin c(w)\}$$

Negative Sampling

$$\mathcal{L} = \sum_{(w,s)\in D_1} \log \sigma(v_w^T v_s) + \sum_{(w,s)\in D_2} \log \sigma(-v_w^T v_s),$$

- Но множество всех отрицательных примеров отсутствует
- Выход — для каждого рассматриваемого слова w генерировать в качестве отрицательных примеров случайные слова из T
- Функционал оптимизируется с помощью SGD

word2vec и PMI

Pointwise Mutual Information:

$$PMI(w, c) = \log \frac{\#(w, c)|D|}{\#(w)\,\#(c)}$$

$$PPMI(w, c) = \max\{PMI(w, c), 0\}$$

$$SPPMI(w, c) = PPMI(w, c) - \log k$$

k — число отрицательных примеров (negative samples)

Можно показать, что модель SGNS является разложением матрицы SPPMI (строки — слова, столбцы — контексты)

Важно: под контекстом в данном случае понимается *центральное слово*, на основании которого предсказываются стоящие рядом слова

Global Vectors

- На пальцах: GloVe = SVD + word2vec
- Строим матрицу $X \in \mathbb{R}^{V \times V}$, x_{ij} — количество раз, когда слово i встречается в контексте слова j (в окне $\leqslant 9$ слов)
- $P_{ij} = \dfrac{x_{ij}}{X_i}$ – вероятность j в контексте i (X_i – сумма $i-$й строки).
- Строим функцию $F(w_i, w_j, \hat{w}_k)$, показывающую, какое из слов i и j вероятнее увидеть в контексте k (> 1 если i чаще)

$$F(w_i, w_j, \hat{w}_k) = \frac{P_{ik}}{P_{jk}}$$

- Все w_x — векторные представления соответствующих слов

---

https://www.aclweb.org/anthology/D14-1162

GloVe

- Предлагается следующая функция

$$F((w_i - w_j)^T \hat{w}_k) = \frac{F(w_i^T \hat{w}_k)}{F(w_j^T \hat{w}_k)}, \quad F(w_i^T \hat{w}_k) = P_{ik}$$

- Тогда можно взять $F(x) = \exp(x)$, а w_i таким, что

$$w_i^T \hat{w}_k = \log(P_{ik}) = \log(x_{ik}) - \log(X_i)$$

- С учётом фиксированности данных о коллекции, перепишем задачу

$$w_i^T \hat{w}_k + b_i + \hat{b}_k = \log(x_{ik}), \quad b_i + \hat{b}_k = \log(X_i), \quad b_x - \text{bias}$$

- Оптимизация с помощью AdaGrad

GloVe

- Оптимизируемый функционал:

$$J = \sum_{i,j=1}^{V} f(x_{ij})(w_i^T \hat{w}_j + b_i + \hat{b}_j - \log(x_{ij}))^2$$

- $f(x)$ должна
  - $f(0) = 0$
  - $f(x)$ не убывает
  - $f(x)$ относительно мала для больших x
- $$f(x) = \begin{cases} (x/x_{max})^\alpha, & x < x_{max} \\ 1, & \text{else} \end{cases}$$
- $\alpha = 0.75$, $x_{max} = 100$

N-символьные эмбеддинги: FastText

- N-символьные эмбеддинги показывают более высокое качество, чем эмбеддинги символов
- При этом они тоже хорошо справляются с редкими и OOV-словами

  Пример: N-граммы для слова `apple` (min=3, max=6):

  `<ap`, `<app`, `<appl`, `<apple`, `app`, `appl`, `apple`, `apple>`, `ppl`, `pple`, `pple>`, `ple`, `ple>`, `le>`

FastText – библиотека (и алгоритм) для получения векторных представлений слов и классификации текстов

- Архитектурно такая же, как у word2vec (CBOW и skip-gram)
- Строит эмбеддинги символьных N-грамм
- Эмбеддинг слова получается усреднением эмбеддингов N-грамм

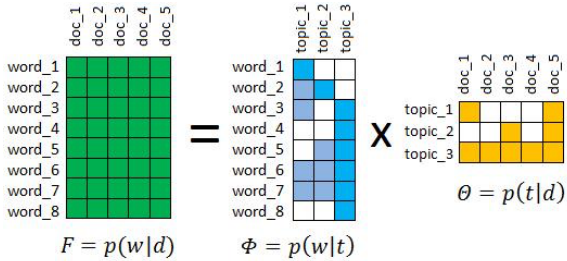
FastText

- Для борьбы с ростом размерности признакового пространства используется *hashing trick*
- В FastText есть встроенный классификатор на основе двуслойной полносвязной сети
- Вход – эмбеддинг документа как сумма эмбеддингов слов
- На выходе стоит иерархический софтмакс с деревом Хаффмана
- За счёт использования символьных N-грамм может из коробки хорошо работать с сырыми текстами
- Ссылки на статьи: https://arxiv.org/pdf/1607.01759.pdf, https://arxiv.org/pdf/1607.04606.pdf

Эмбеддинги предложений и документов

- Часто в задачах текстовой аналитики нужные эмбеддинги не слов в документах, а самих документов или их частей
- Самый простой способ получения – взвешенная сумма эмбеддингов слов (например, с tf-idf в качестве весов)
- Такой подход показывает хорошие результаты и часто используется на практике, но есть и более интересные методы
- Если есть хороший способ найти эмбеддинг предложения, вектор документа можно опять-таки получить усреднением по предложениям
- Качество оценивается по метрике задачи (внешний критерий) или, например, качеству поиска размеченных аналогий (внутренний)

Снова тематическое моделирование



- Классические тематические модели получают на вход матрицу «мешка слов» или tf-idf и строят два типа распределений:
  - слов в кластерах-темах
  - тем в документах
- Столбцы матрицы «темы-документы» можно использовать в качестве эмбеддингов документов
- Они отражают близость документов со схожей тематикой

Мультиязычные эмбеддинги

- Обычно семантические эмбеддинги строятся внутри в рамках одного языка
- Хочется иметь пространство, в котором векторы слов на разных языках, но с общей семантикой, были близки
- Тогда можно обучать модель на одном языке, а результаты использовать для другого
- Схожая задача явным образом решается в моделях машинного перевода путём оптимизации функционала глубокой нейросети
- Но строить эмбеддинги можно неявно (даже без чёткой разметки) и значительно проще с вычислительной точки зрения

Виды подходов

- **Моноязычный:** обучаем эмбеддинги для каждого языка отдельно и обучаем линейные преобразования между пространствами
- **Псевдо-мультиязычный:** обучаем эмбеддинги одновременно на синтетическом корпусе, в котором перемешаны слова на разных языках
- **Мультиязычный:** обучаемся на параллельном корпусе так, чтобы эмбеддинги похожих слов на разных языках были близки

Рассмотрим ниже несколько примеров.

Подробный структурированный обзор есть доступен по адресу http://ruder.io/cross-lingual-embeddings/

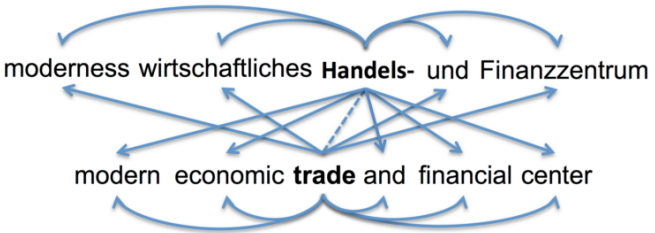
Linear projection

- Замечено, что схожие сущности (числа, имена животных) образуют в своих языковых пространствах на английском и испанском схожие по форме кластеры
- Отсюда вывод, что можно обучить независимо два векторных пространства и далее построить линейное преобразование между ними
- Для обучения преобразования отбирается 5000 наиболее популярных слов в исходном языке, для них подбираются переводы
- Дальше на этих 5000 парах с помощью SGD обучается матрица преобразования векторов
- `Mikolov, T., Le, Q. V., Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. Retrieved from http://arxiv.org/abs/1309.4168`

Random translation replacement

- Берётся исходный язык, все слова из него переводятся на целевой с помощью Google Translate
- На основании корпуса на исходном языке формируется корпус, в котором каждое слово с вероятностью 50% заменяется на свой перевод
- Дальше запускается обучение модели CBOW (как в w2v)
- `Gouws, S., Søgaard, A. (2015). Simple task-specific bilingual word embeddings. NAACL, 1302-1306.`

Bilingual skip-gram



- Очевидно необходима разметка выравнивания между словами
- Luong, M.-T., Pham, H., Manning, C. D. (2015). Bilingual Word Representations with Monolingual Quality in Mind. Workshop on Vector Modeling for NLP, 151-159.

Multilingual unsupervised and supervised embeddings

- Есть предобученные векторы FastText для большого числа языков
- Как и в других моноязычных подходах, здесь обучается линейное преобразование из исходного пространства в целевое
- Но здесь используются никакие параллельные данные
- Вместо этого обучается GAN
  - параметрами генератора является матрица преобразования
  - дискриминатор учится отличать преобразованные векторы исходного пространства от векторов целевого
- `https://github.com/facebookresearch/MUSE`

Выводы: чем пользоваться в жизни

- В общем случае самым простым, легковесным и эффективным инструментом является FastText
- В ситуации, когда нужно работать с большими текстами и нужна интепретируемость векторов – подойдут тематические модели
- Также при наличии разметки (пусть и синтетической) бывает полезным обучать модели наподобие DSSM
- Глубокие нейросетевые эмбеддинги в индустрии используются нечасто:
  - они требуют много железа, данных и времени (даже на дообучение)
  - скорость вывода может быть недостаточной, нужно думать, как ускорять (например, делать дистилляцию сети)
  - в жизни улучшение качества по сравнению с более простыми подходами может оказаться слишком небольшим
- Но если есть время на решение технических вопросов и дообучение, и точно понятно, что работать будет сильно круче, то они могут стать лучшим решением

Что дальше?

- context2vec
- USE
- ULMFIT
- ELMo
- BERT
- ERNIE
- GPT-2
- ...

А на сегодня всё!