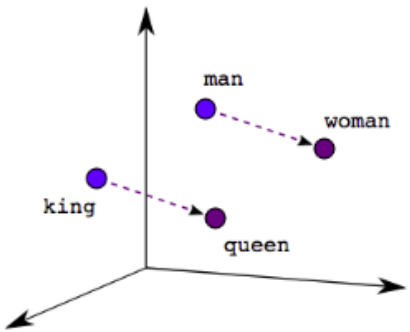
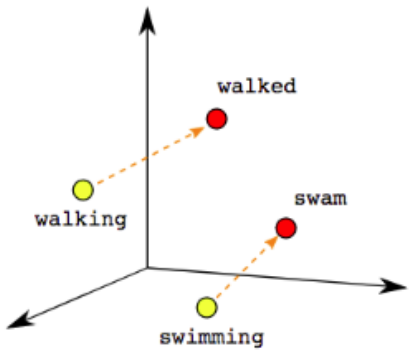
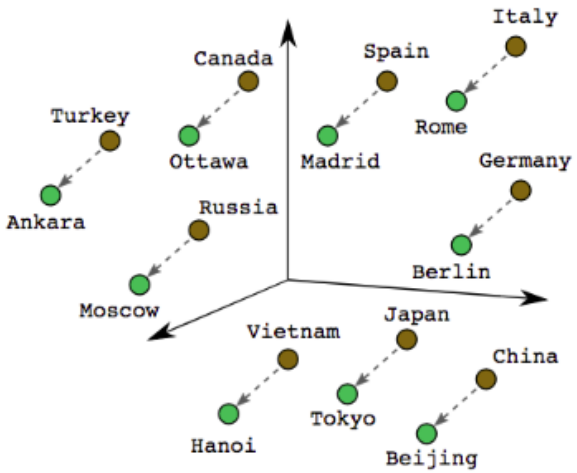
Векторные представления объектов

Виктор Китов
victorkitov.github.io







Содержание

Стандартное представление слов

- Обозначим D=размер словаря.
- Стандартные представления слов $x \in \mathbb{R}^D$:
  - $x_w = \mathbb{I}[w$ встретился в документе$]$
  - $x_w = TF_w = \#[w$ встретился в документе$]$
  - $x_w = TF_w IDF_w$, $IDF_w = \frac{N}{N_w}$
    - N - # документов
    - N_w - # документов, содержащих w хотя бы раз.
- V велико, поэтому нужно компактное представление (word embedding) $x \in \mathbb{R}^K$, $K << D$:
  - меньше входов=>меньше параметров=>ниже переобучение
  - возможность учитывать семантическое сходство/различие
    - например, синонимы "автомобиль" и "машина"

Интерпретируемые векторные представления слов

- Можно из слов извлекать интерпретируемые признаки:
  - x^1: часть речи
  - x^2: род (м/ж/ср - для существительных)
  - x^3: время (пр/наст/буд - для глаголов)
  - x^4: $\mathbb{I}$[начинается с заглавной буквы]
  - x^5: # букв
  - x^6: категория: машинное обучение, физика, биология, ...
  - x^7: подкатегория: обучение с учителем, без учителя, частичное обучение, ...
  - ...
- Необходимо придумывать признаки под задачу, производить разметку.
- Легче работать с неинтерпретируемыми признаками, но которые извлекаются автоматически.

Неинтерпретируемые представления слов

- Хотим, чтобы семантически близким словам соответствовали близкие представления.
- Дистрибутивная гипотеза (distributional hypothesis): слова близки по смыслу <=> они часто встречаются совместно
- "точность бустинга", "бустинг дал точность", "ниже точность, по сравнению с бустингом"
  - "точность" и "бустинг" связаны!
- Типичная размерность векторного представления $\in [300, 500]$.

Представления фраз

Можно обрабатывать фразы как отдельные "слова".

- Коллокации (неслучайно часто встречающиеся слова):

$$(w_i, w_j)\text{-коллокация} \iff \frac{p(w_i w_j) - \delta}{p(w_i)p(w_j)}$$

Представления фраз

Можно обрабатывать фразы как отдельные "слова".

- Коллокации (неслучайно часто встречающиеся слова):

$$(w_i, w_j)\text{-коллокация} \iff \frac{p(w_i w_j) - \delta}{p(w_i)p(w_j)}$$

$> threshold$. δ - параметр, снижающий значимость редко совстречающихся слов.

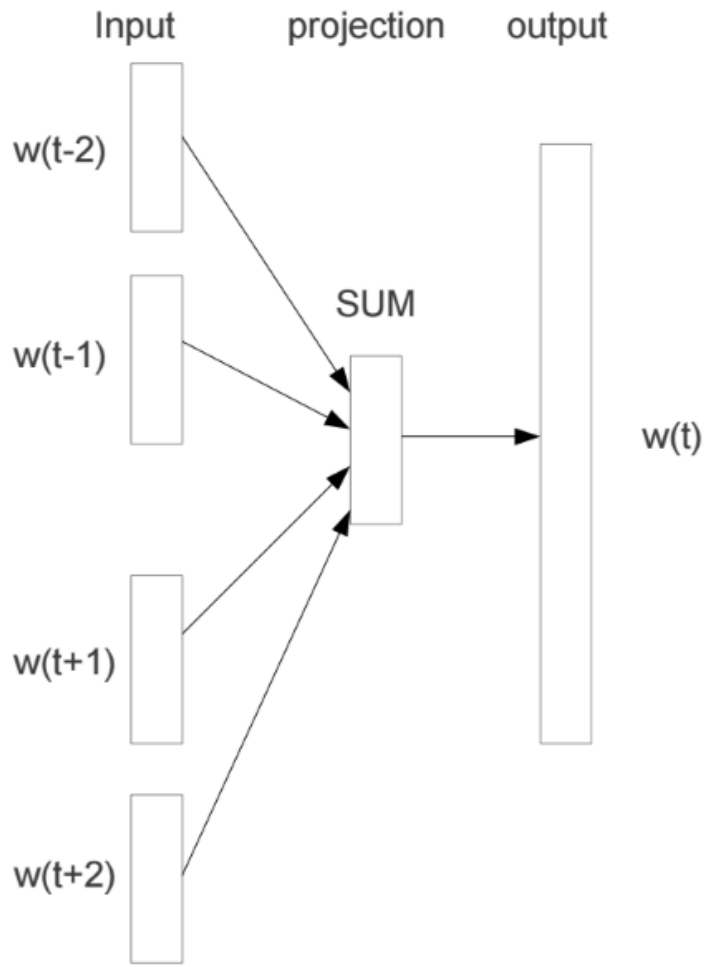
Содержание

Word2vec

- Для каждого w оценим:
  - целевое представление слова v_w
  - контекстное представление слова u_w
    - впоследствии можно не использовать, усреднить или конкатенировать с целевым представлением

CBOW: идея

Continuous bag of words (CBOW): предсказываем центральное слово по контексту.



CBOW: модель

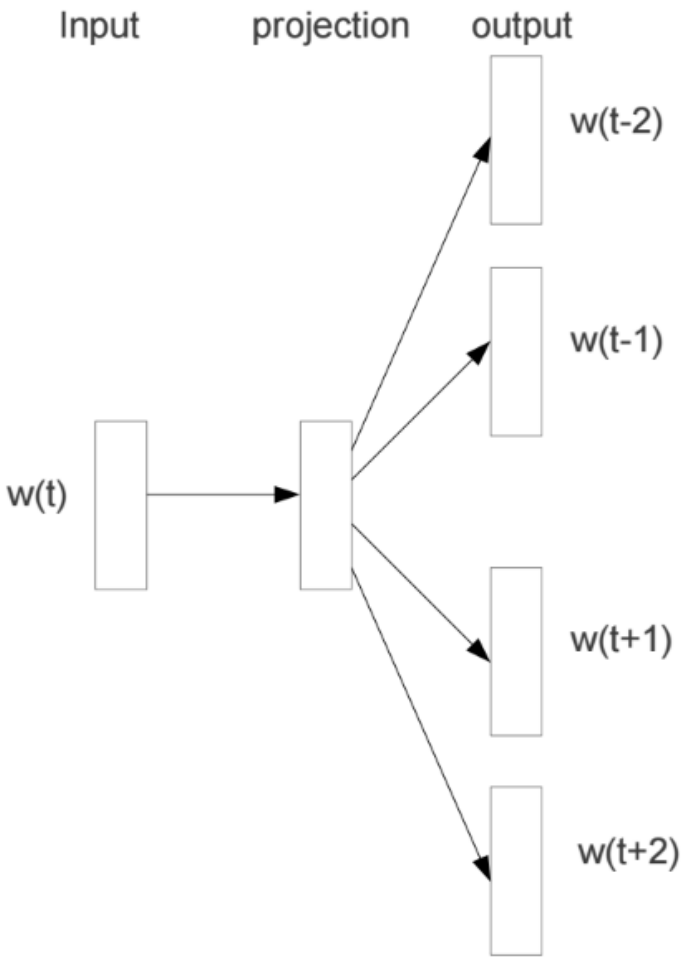
$$\frac{1}{T}\sum_{t=1}^{T}\ln p(w_t|w_{t-K},..w_{t-1},w_{t+1},...w_{t+K})\to\max_{\theta}$$

где $u_c=\sum_{-K\le i\le K,\, i\neq 0} u_{w_{t+i}}$ и

$$p(w_t|w_{t-c},..w_{t-1},w_{t+1},...w_{t+c})=\frac{\exp\left(u_c^T v_{w_t}\right)}{\sum_{w=1}^{D}\exp\left(u_c^T v_w\right)}$$

Skip-gram: идея

Skip-gram: предсказываем контекст по центральному слову:



Skip-gram: модель

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-K\le i\le K,\, i\neq 0} \ln p(w_{t+i}|w_t) \to \max_{\theta}$$

$$p(w_{t+i}|w_t) = \frac{\exp\left(u_{w_t}^T v_{w_{t+i}}\right)}{\sum_{w=1}^{D}\exp\left(u_{w_t}^T v_w\right)}$$

Комментарии

- Можем извлекать представления для др. объектов из последовательностей.
  - символы, биграммы, триграммы символов (см. *FastText*), предложения
  - нуклеотиды в ДНК последовательности
  - сервисы, заказанные клиентом компании
- Можем использовать ансамбли представлений
  - сумма, среднее, конкатенация

Содержание

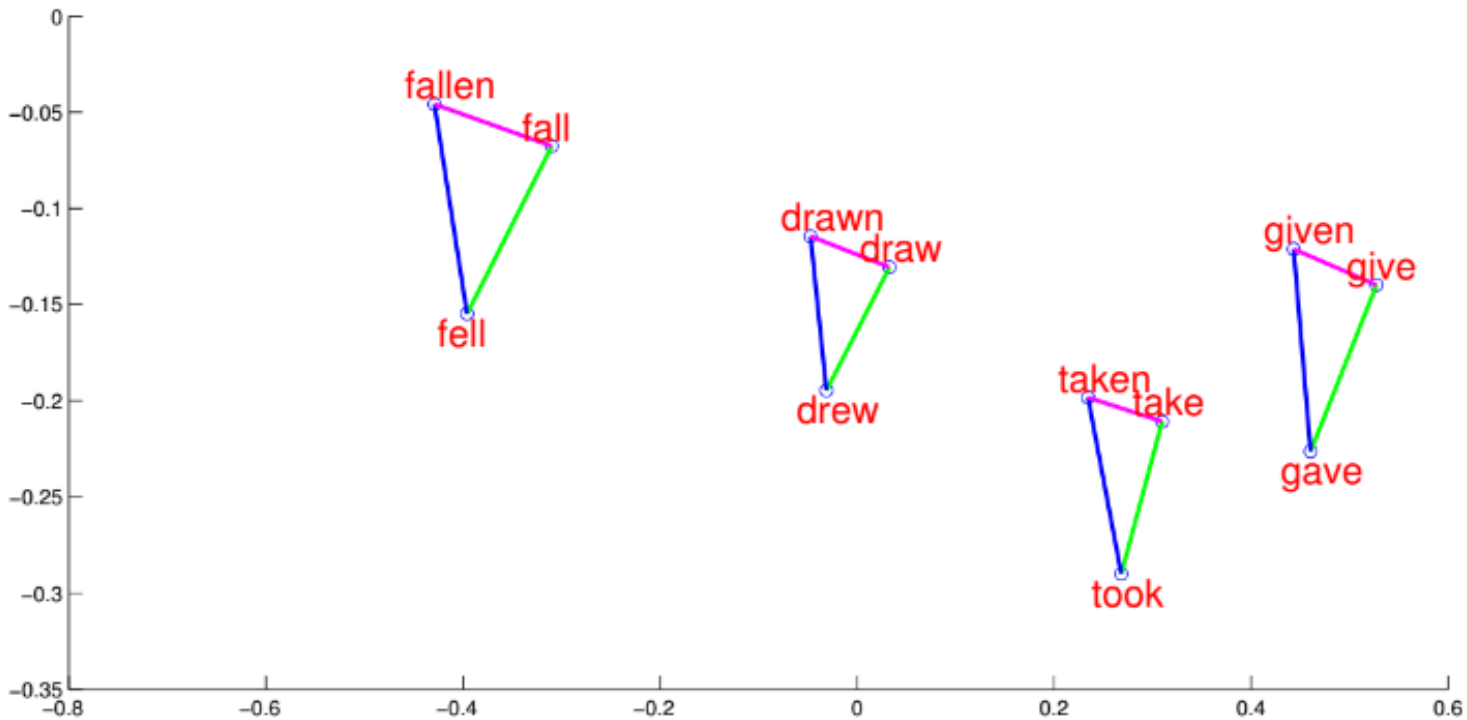
Похожие слова по представлению[1]

- Ближайшие соседи слова в пространстве эмбеддингов - слова, похожие по смыслу
(корпус GoogleNews, cosine-sim):
  - student -> teacher, faculty, school, university
  - car -> truck, jeep, vehicle
  - country -> nation, continent, region

[1] http://epsilon-it.utu.fi/wv_demo/

Формы слов

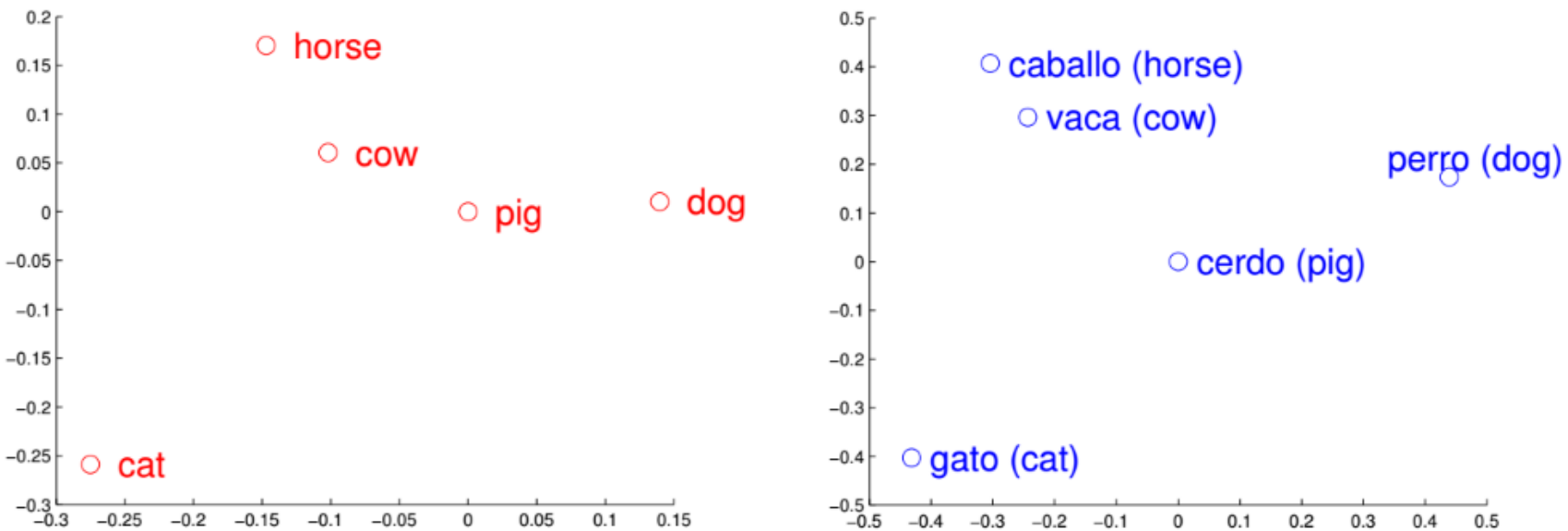
Одинаковые слова в разных формах образуют похожие структуры:



Представления могут помочь строить др. формы новых и редких слов.

Слова на разных языках

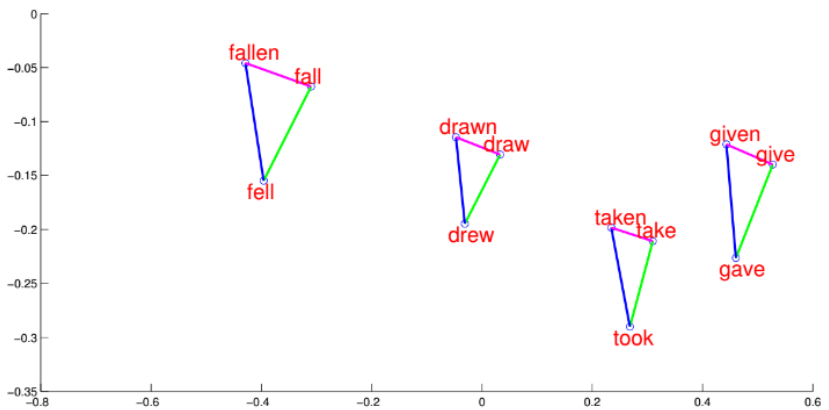
Слова на разных языках группируются похожим образом:



Представления слов могут помочь в переводе на др. язык.

Семантическая регулярность

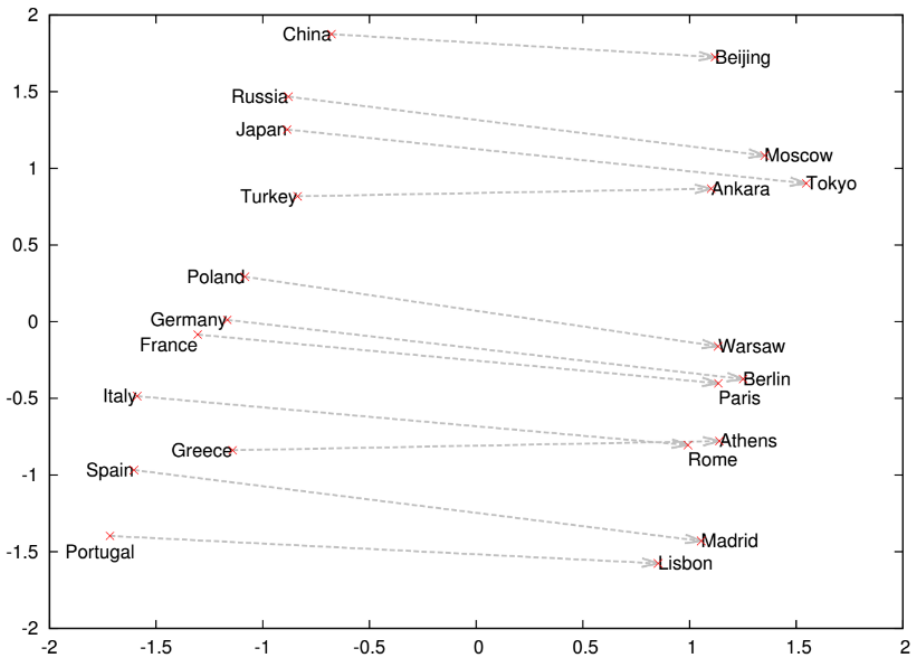
Слова, связанные семантически определенным образом группируются единообразно:



(prince-princess)+queen≈king. Может помочь в системе автоматических ответов на вопросы.

Семантическая регулярность

Слова, связанные семантически определенным образом группируются единообразно:



(Beijing-China)+Russia≈Moscow! Может помочь в системе автоматических ответов на вопросы.

Содержание

Вычислительная сложность Word2vec

CBOW: $$\frac{1}{T}\sum_{t=1}^{T}\ln p(w_t|w_{t-c},..w_{t-1},w_{t+1},...w_{t+c})\to\max_{\theta}$$

где $u_c=\sum_{-K\le i\le K,\,i\neq 0}u_{w_{t+i}}$ и

$$p(w_t|w_{t-K},..w_{t-1},w_{t+1},...w_{t+K})=\frac{\exp\left(u_c^T v_{w_t}\right)}{\sum_{w=1}^{D}\exp\left(u_c^T v_w\right)}$$

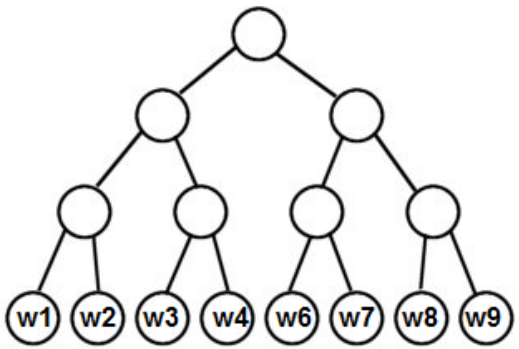
SkipGram: $$\frac{1}{T}\sum_{t=1}^{T}\sum_{-K\le i\le K,\,i\neq 0}\ln p(w_{t+i}|w_t)\to\max_{\theta}$$

$$p(w_{t+i}|w_t)=\frac{\exp\left(u_{w_t}^T v_{w_{t+i}}\right)}{\sum_{w=1}^{D}\exp\left(u_{w_t}^T v_w\right)}$$

Проблема: знаменатель вычисляется за $O(D)$.

Иерархический SoftMax

- Бинарное дерево:
  - предсказываемые слова языка - листья.
  - u_c-эмбеддинг центрального слова (контекста)



- Пусть w_I - входное слово, а w_O - выходное (предсказываемое) слово в Skip-Gram.
- Тогда $w_I \to u_{w_I}$, $w_O \to \{v_{w_O,j}\}_j$ для каждого узла дерева j.
- Для каждого узла j:

$$p(\text{left}|j) = \sigma\left(v_j^T u_c\right),$$
$$p(\text{right}|j) = 1 - \sigma\left(v_j^T u_c\right) = \sigma\left(-v_j^T u_c\right)$$

- $p(w|\text{context})$ =произведение вероятностей дойти до него.

Иерархический SoftMax

- Сложность вычисления $p(w_O|w_I)$ существенно снижается:

$$O(D) \to O(\log_2 D)$$

- Целевым представлением для w теперь уже является входной эмбеддинг u_w
  - а не набор выходных, связанных с переходами в дереве
- Эффективнее работает не сбалансированное дерево, а дерево Хафмана
  - более частотным словам - более короткие пути.

Негативное сэмплирование

- Негативное сэмплирование (negative sampling)[2] - аппроксимация максимизация правдоподобия.
- Для каждой реальной (позитивной) пары (w_t, w_{t+i}) сэмплируем S негативных случайно $\left(w_t, w_{j(1)}\right), \ldots \left(w_t, w_{j(D)}\right)$.

$$\underbrace{\ln\left(\frac{1}{1+e^{-u_{w_t}^T v_{w_{t+i}}}}\right)}_{\sigma\left(+u_{w_t}^T v_{w_{t+i}}\right)} + \sum_{k=1}^{S} \underbrace{\ln\left(\frac{1}{1+e^{+u_{w_t}^T v_{w_{t+i}}}}\right)}_{\sigma\left(-u_{w_t}^T v_{w_{j(k)}}\right)} \to \max_{u_{w_t}, v_{w_{t+i}}}$$

- S ~2-5. $p\left(w_{j(k)}\right) \propto p(w)^{3/4}$-чаще сэмлируем редкие слова.

[2] Distributed Representations of Words and Phrases

fastText[3]

- Работает как skip-gram (предсказываем слова контекста по центральному слову)
- Раньше совместимость была $u_t^T v_{t+i}$
- В fastText $w \to \left[u_w, \{p_j\}_{p\in\text{n-grams}(w)}, v_w\right]$
  - $u_w, \{p_j\}_{p\in\text{n-grams}(w)}$- входные эмбеддинги для известного слова
  - v_w - выходной эмбеддинг для предсказываемого
- Новая совместимость

$$u_{w_t}^T v_{w_{t+i}} + \sum_{j\in\text{n-grams}(w_t)} p_j^T v_{w_{t+i}}$$

- Пример 3-грамм person:
"<person>","<pe","per","ers","rso","son","on>"
  - предлагается использовать все n-граммы, $3 \le n \le 6$.
- Для слов вне словаря работает, используем только n-граммы $\sum_{j\in\text{n-grams}(w)} p_j$.

[3] https://arxiv.org/pdf/1607.04606.pdf код и данные:fasttext.cc

Содержание

Матрица совстречаемости слов

- $X \in \mathbb{R}^{VxV}$ - матрица со-встречаемости слов (word co-occurrence matrix)
- X_{ij} =#{слово j встретилось в контексте слова i}.
- Пример для контекста ± 1 слово:

I like deep learning. I like NLP. I enjoy flying.

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

Разложение матрицы совстречаемости

- Hyperspace Analogue to Language (HAL)[4]: эмбеддинги из низкорангового разложения
  - напр. строки U либо столбы V^T из SVD.
  - эмбеддинги доминируются частыми словами!

[4] Lund and Burgess, 1996.
[5] Bullinaria and Levy, 2007
[6] https://aclanthology.org/D14-1162.pdf

Разложение матрицы совстречаемости

- Hyperspace Analogue to Language (HAL)[4]: эмбеддинги из низкорангового разложения
  - напр. строки U либо столбы V^T из SVD.
  - эмбеддинги доминируются частыми словами!
- Модификация[5]: счётчик совстречаемости→PPMI

$$PPMI(w_1, w_2) = \max\{0, PMI(w_1, w_2)\} = \max\{0, \ln \frac{P(w_1, w_2)}{P(w_1)P(w_2)}\}$$

[4] Lund and Burgess, 1996.
[5] Bullinaria and Levy, 2007
[6] https://aclanthology.org/D14-1162.pdf

Разложение матрицы совстречаемости

- Hyperspace Analogue to Language (HAL)[4]: эмбеддинги из низкорангового разложения
  - напр. строки U либо столбы V^T из SVD.
  - эмбеддинги доминируются частыми словами!
- Модификация[5]: счётчик совстречаемости→PPMI

$$PPMI(w_1,w_2)=\max\{0,PMI(w_1,w_2)\}=\max\{0,\ln\frac{P(w_1,w_2)}{P(w_1)P(w_2)}\}$$

- GloVe[6]: матр. факторизация $\log X\approx W^T\tilde{W}+B+\tilde{B}$

$$\sum_{i,j=1}^{D} f\left(X_{ij}\right)\left(w_i^T\tilde{w}_j+b_i+\tilde{b}_j-\log X_{ij}\right)^2\to\min_{w,\tilde{w},b,\tilde{b}}$$

$f(X_{ij})$ некоторая ↑ функция весов, $f(0)=0$.

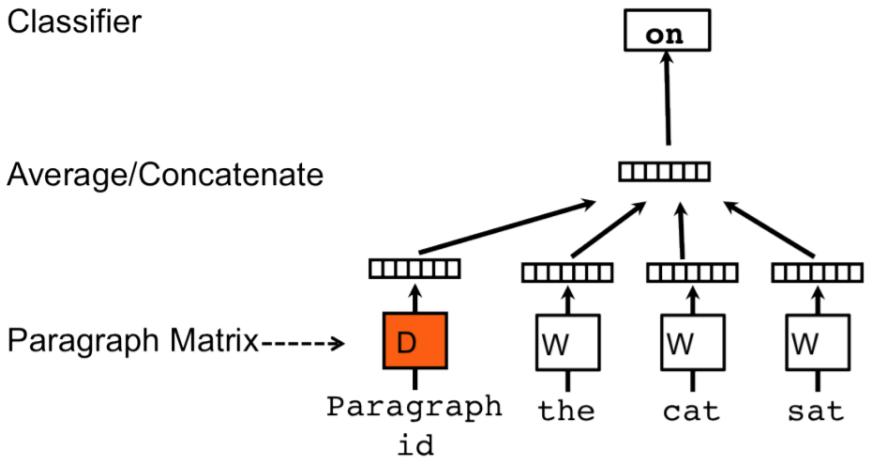
[4] Lund and Burgess, 1996.
[5] Bullinaria and Levy, 2007
[6] https://aclanthology.org/D14-1162.pdf

Содержание

Представления параграфов - мотивация

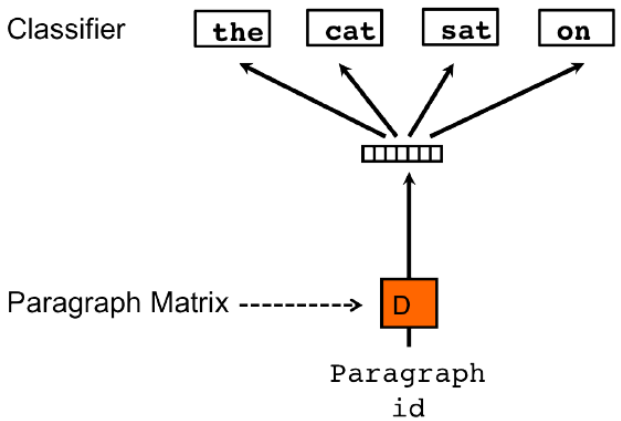
- Необходимо получить векторные представления параграфов (документов, предложений,...).
- Простой подход: усреднить слова, входящие в параграф.
  - или взвешенно усреднить, учитывая частоту встречаемости слов и их тематику.
- Точнее работает непосредственное представление самих параграфов.

Paragraph vector: модель PV-DM



- Во время обучения делим документы на параграфы. Каждому параграфу -> векторное представление.
- Оценивается CBOW, контекст: представление слов и парграфов.
- Можно усреднять или конкатенировать контексты слов и параграфа.
- Называется *Distributed Memory Model of Paragraph Vectors (PV-DM)*.

Paragraph vector: модель PV-DBOW

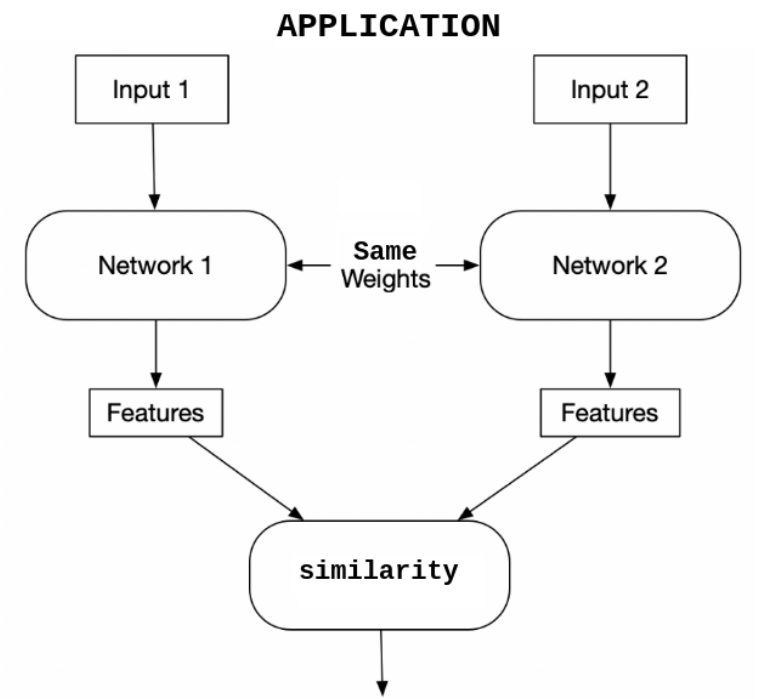


- Во время обучения делим документы на параграфы. Каждому параграфу -> векторное представление.
- Оценивается skip-gram: предсказываются случайные слова параграфа по представлению параграфа.
  - контекст: только представления параграфов
- Называется *Distributed Bag of Words version of Paragraph Vector (PV-DBOW)*

Содержание

Сиамская сеть

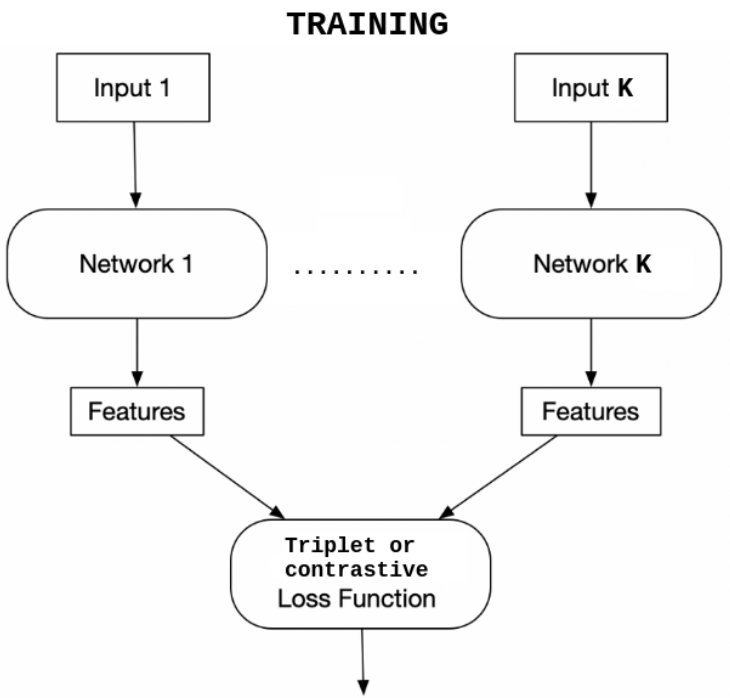
- Сиамская сеть (siamese network): объект $x_n \to$эмбеддинг e_n.
  - объекты могут быть разных доменов.
- Обучение: похожие объекты $\Rightarrow$ похожие представления.
  - похожесть: $\|\cdot\|_2^2$, $\langle \cdot, \cdot \rangle$, cos-sim



Примеры приложений

- Классификация:
  - вход: 2 объекта (обучение) или тестовый объект (применение)
  - выход: класс на основе близости к центроиду класса или по K-NN
- Поисковая система
  - вход: документ и поисковый запрос
    - возможен и поиск по изображению
  - выход: степень релевантности документа запросу
- Обнаружение перефразирования:
  - вход: 2 предложения
  - выход: насколько они близки по смыслу
- Проверка подписи
  - вход: сканы 2х подписей
  - выход: степень их принадлежности одному человеку

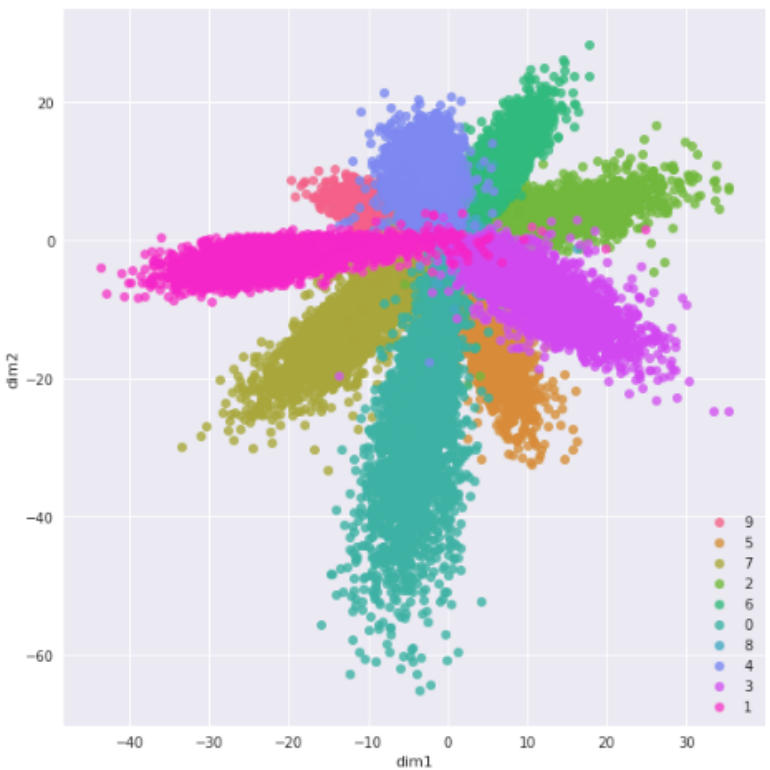
Обучение

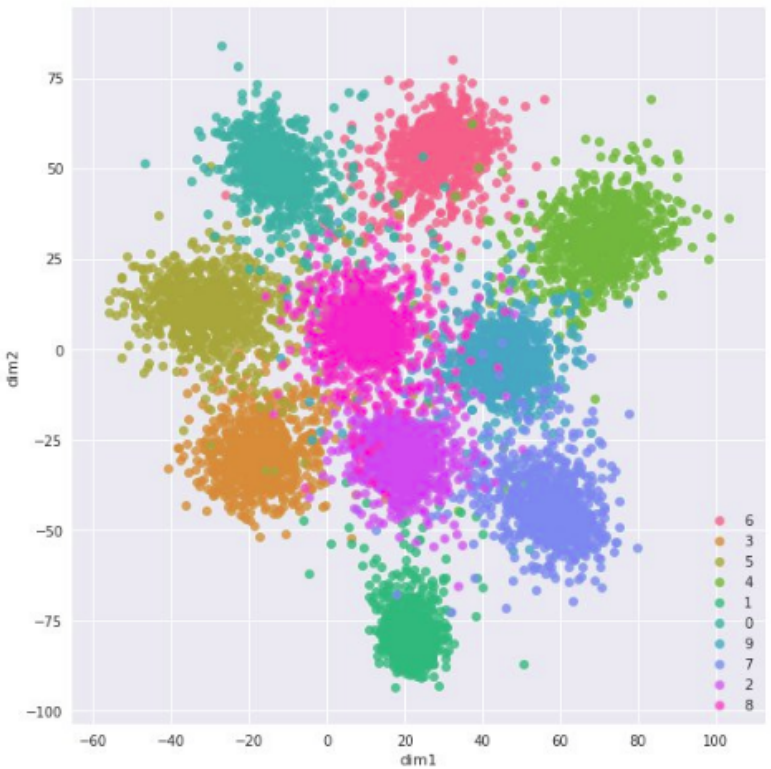


- Идея функции потерь:
  - представления похожих объектов должны быть близки
  - представления различных объектов должны быть далеки

Представления объектов: классификация и сиамская сеть

Представления объектов: классификация и сиамская сеть для MNIST:





Применение

После того, как сиамская сеть настроена, можно решать конечную задачу. Если
классификация, то можно

- инициализировать классификационную сеть первыми слоями сиамской сети (особенно для instance discrimination)
- решать классификацию в пространстве эмбеддингов.
  - метод ближайших центроидов
  - метод K ближайших соседей

Попарные потери

Попарные потери (pairwise contrastive loss, spring loss)[7]:

- обучение на случайных парах объектов x_i, x_j

$$\mathcal{L}(x_i, x_j) = \begin{cases} \rho(e_i, e_j)^2, & \text{если } x_i, x_j \text{ похожи} \\ \max\{0, \alpha - \rho(e_i, e_j)\}^2, & \text{если } x_i, x_j \text{ похожи} \end{cases}$$

- $\alpha > 0$ - гиперпараметр (мин. расстояние для непохожих объектов, когда не будет штрафа)
- $\rho(e, e')$ - обычно Евклидово
- число уникальных пар - $O(N^2)$.

[7] Выгоднее позволять похожим объектам небольшую вариацию в эмбеддингах. Предложите соответствующее изменение.

Тройные потери

Тройные потери (triplet loss):

- обучение на случайных тройках x, x^+, x^-.
- x - опорный объект (anchor)
- x^+ - похожий на x (positive)
- x^- - не похожий на x (negative)
- $\alpha > 0$ - гиперпараметр (мин. разница расстояний без штрафа)
- $\rho(e, e')$ - обычно Евклидово

$$\mathcal{L}(x, x^+, x^-) = \max\left\{\rho(e, e^+)^2 - \rho(e, e^-)^2 + \alpha; 0\right\}$$

- число уникальных пар - $O(N^3)$.

Вероятностные потери

- Вероятностные потери (InfoNCE loss, NCE=noise constrastive estimation)
- x - опорный объект (anchor)
- x^+ - похожий на x (positive)
- $x_1, ...x_S$ - набор непохожих на x объектов

$$\mathcal{L}(x, x^+, x_1^-, ...x_M^-) = -\ln \frac{e^{sim(e,e^+)}}{e^{sim(e,e^+)} + \sum_{m=1}^{S} e^{sim(e,e_m^-)}}$$

$$sim(e, e') = \frac{e^T e'}{||e|| \cdot ||e'||}$$

- $> O(N^3)$ уникальных сэмплов.

Комментарии

- Сэмплировать можно равномерно
  - по объектам (максимизируем микро-усредненные метрики per object)
  - по классам (максимизируем макро-усредненные метрики per class)
- Контрастное обучение можно использовать для metric learning $\rho_\theta(x, x')$.

Сиамская сеть и классификация

- Классификация
  - выучивает "что представляет каждый класс".
  - выдает степени соответствия x каждому классу.
- Сиамская сеть
  - выучивает "что отличает классы друг от друга".
  - выдает расстояния от x до каждого класса.
  - более устойчива к дисбалансу классов и редким классам (*one shot learning*)
    - при обучении каждый класс учитывается поровну
    - модель выучивает признаки, по которым можно судить о сходстве классов на частотных классах, потом сразу подхватывает их для редких.
  - извлекает больше информации из выборки
    - обучение не на объектах, а на парах и тройках объектов.
  - хороша в ансамбле с классификатором (↑разнообразие)

Применения

Заключение

- **Представления слов** отображают слова в компактные векторные представления.
  - может применяться
    - к биграммам, триграммам, коллокациям.
    - к символам - удобно для новых слов
    - к любым объектам из посл–тей (нуклеотиды в ДНК и др.)
- **Представления параграфов** отображают параграфы в векторные представления.
  - работают лучше, чем усреднение слов параграфа
- Представления можно находить для целевой или связанной задачи (language modeling, transfer learning)
- **Сиамская сеть** оценивает похожесть пар объектов.
  - применения: классификация (особенно one shot learning), детекция перефразирования, нахождение похожих изображений, ...