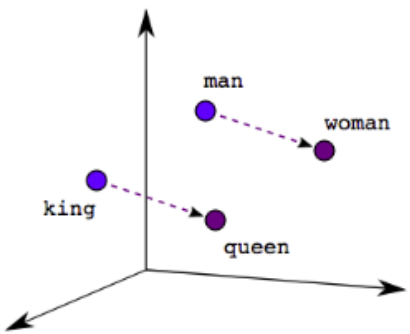
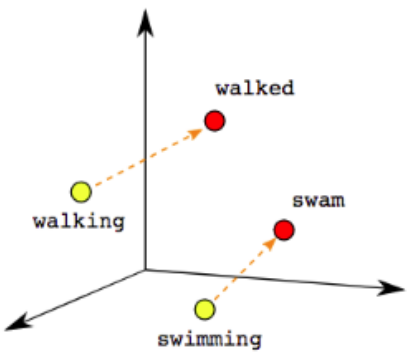
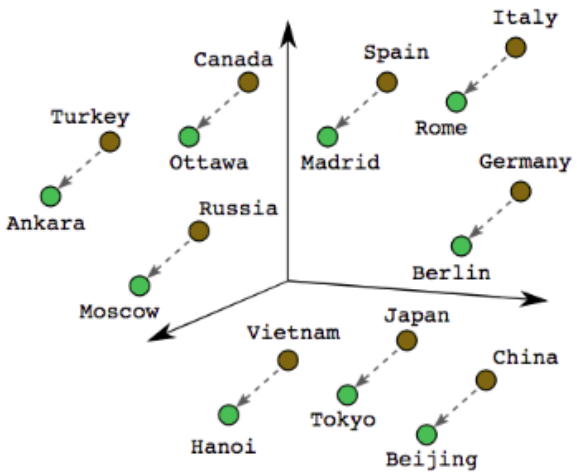
Векторные представления объектов

Виктор Китов

v.v.kitov@yandex.ru







Содержание

Стандартное представление слов

- Обозначим V=размер словаря.
- Стандартные представления слов $x \in \mathbb{R}^V$:
  - $x_w = \mathbb{I}[w$ 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) $x \in \mathbb{R}^K$, $K << V$:
  - 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):

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

δ - параметр, снижающий значимость редко совстречающихся слов.

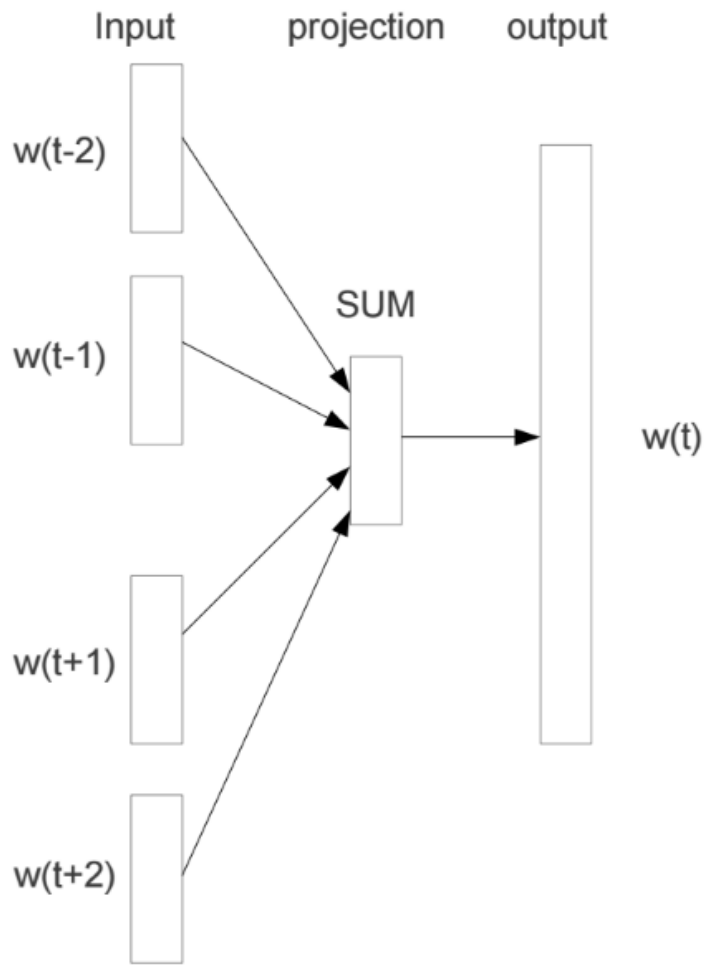
Содержание

Word2vec

- Для каждого w оценим:
  - целевое представление слова v_w
  - контекстное представление слова $\tilde{v}_w$
    - впоследствии можно не использовать, усреднить или конкатенировать с целевым представлением

CBOW: идея

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



CBOW: модель

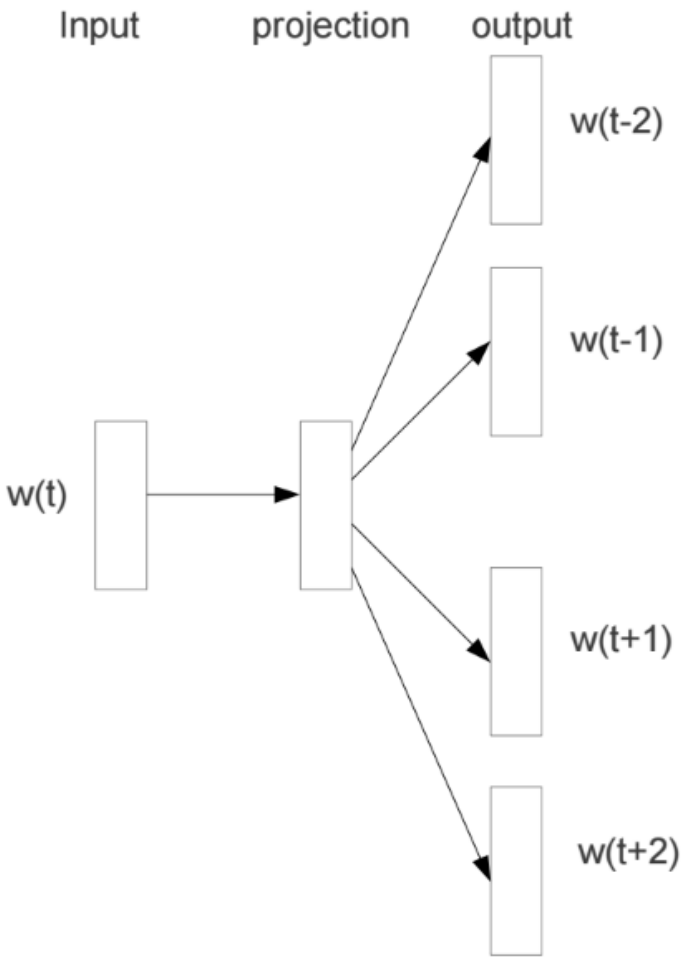
$$\frac{1}{T}\sum_{t=1}^{T}\ln p(w_t|w_{t-c},..w_{t-1},w_{t+1},...w_{t+c}) \to \max_{\theta}$$

где $\tilde{v}_{context} = \sum_{-c\le i\le c,\, i\neq 0}\tilde{v}_{w_{t+i}}$ и

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

Skip-gram: идея

Skip-gram: предсказываем контекст по центральному слову:



Skip-gram: модель

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

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

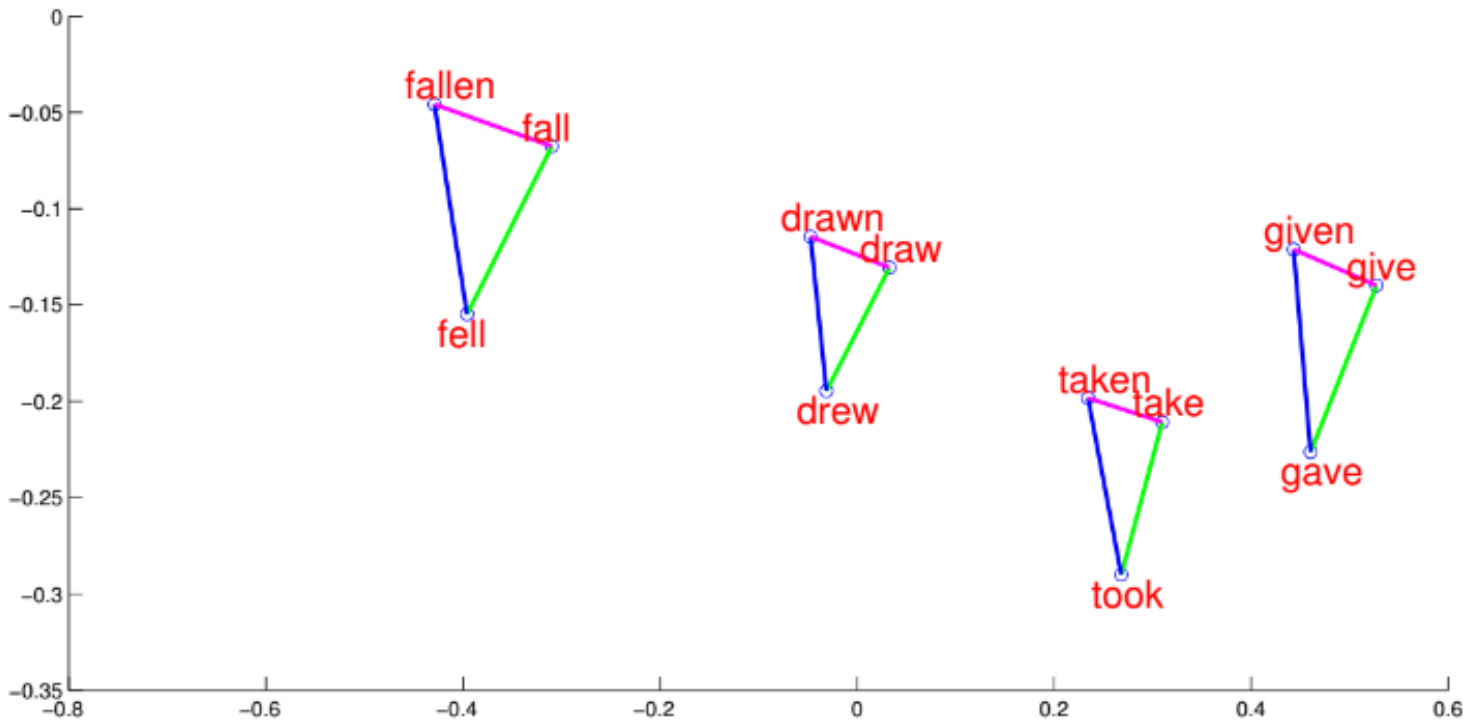
Комментарии

- 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

Формы слов

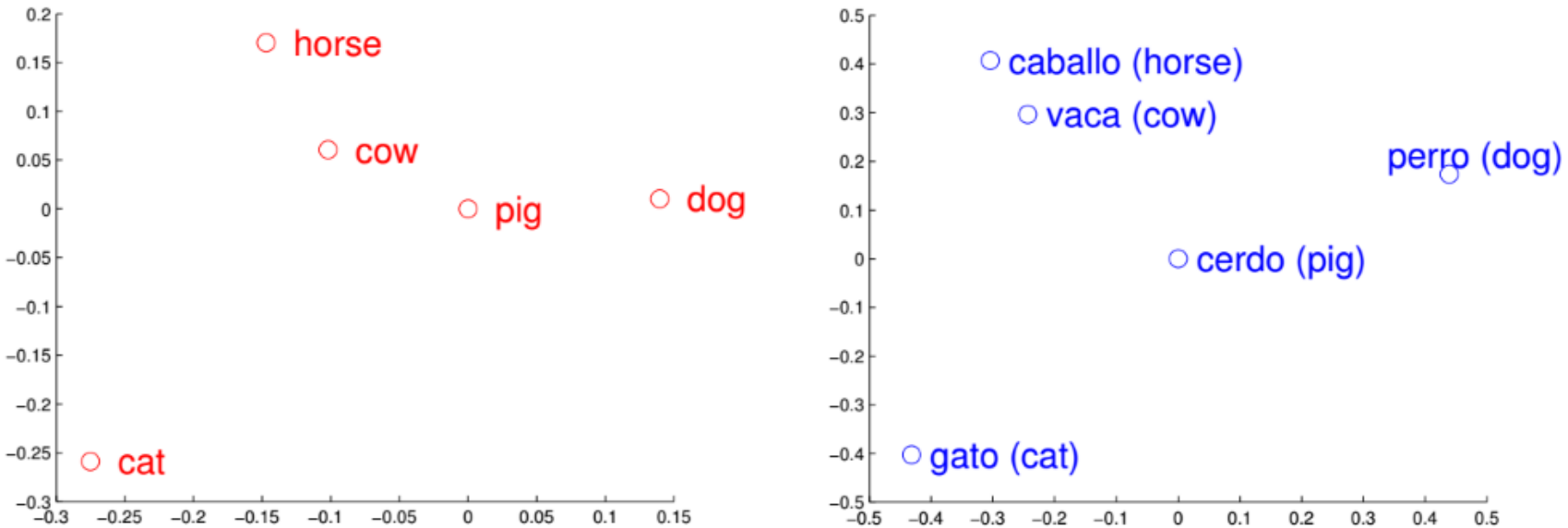
Одинаковые слова в разных формах образуют похожие структуры:



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

Слова на разных языках

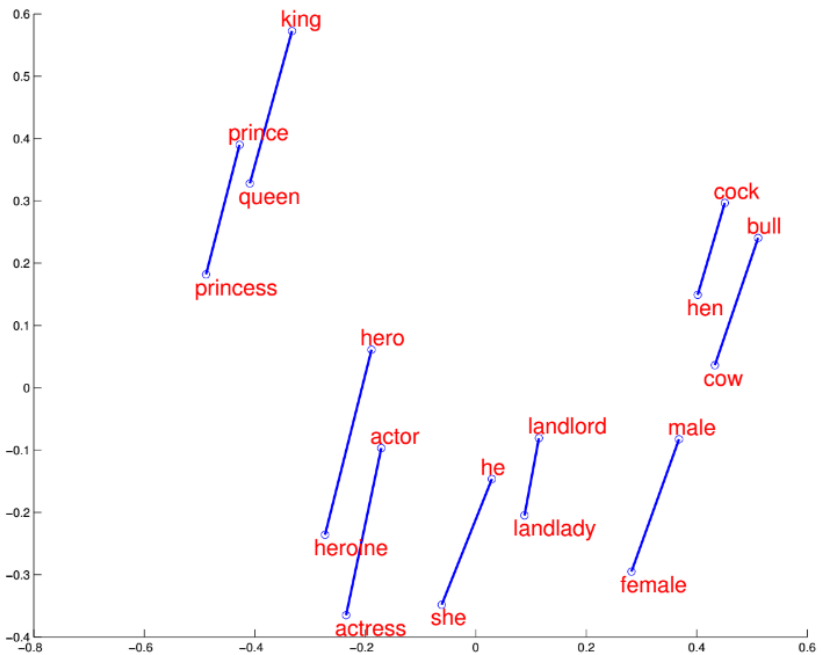
Слова на разных языках группируются похожим образом:



Представления слов могут помочь в переводе на др. язык.

Семантическая регулярность

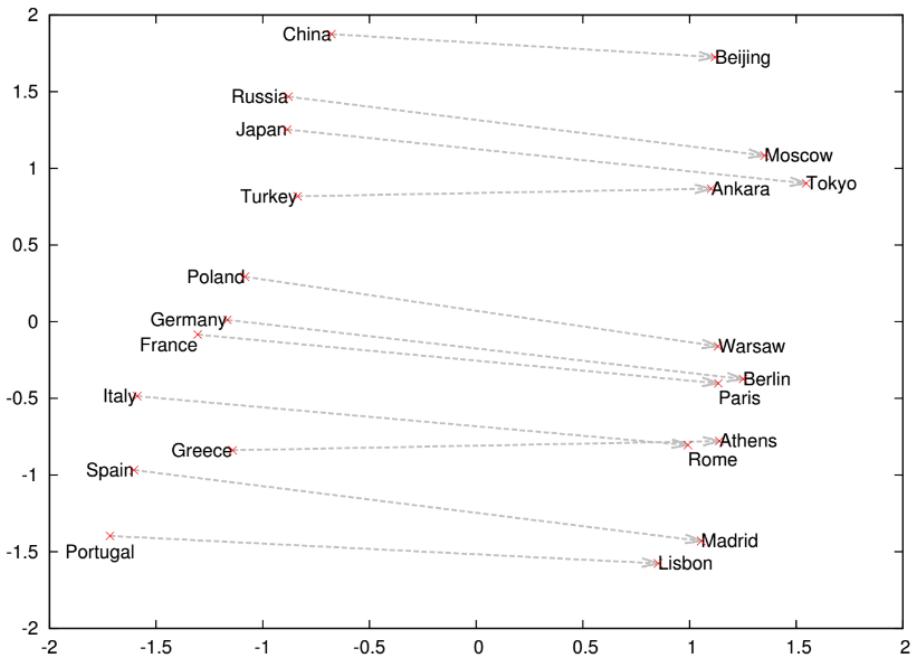
Слова, связанные семантически определенным образом группируются единообразно:



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

Семантическая регулярность

Слова, связанные семантически определенным образом группируются единообразно:



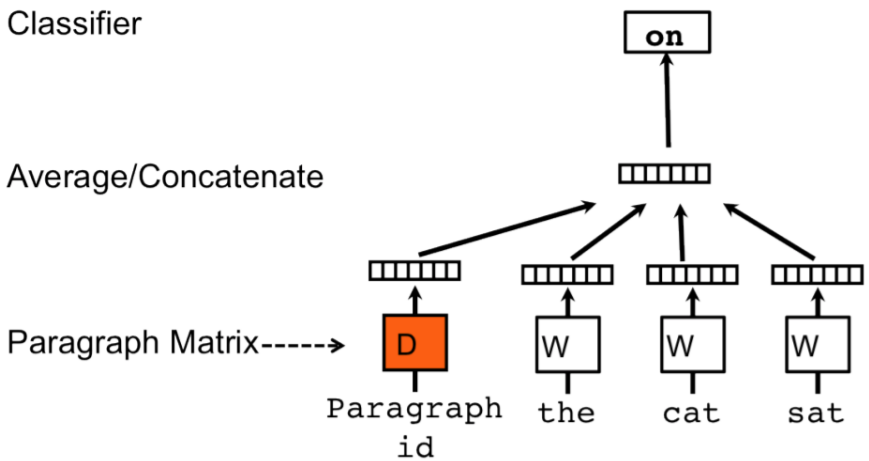
(Beijing-China)+Russia≈Moscow! Может помочь в системе автоматических ответов на вопросы.

Содержание

Представления параграфов - мотивация

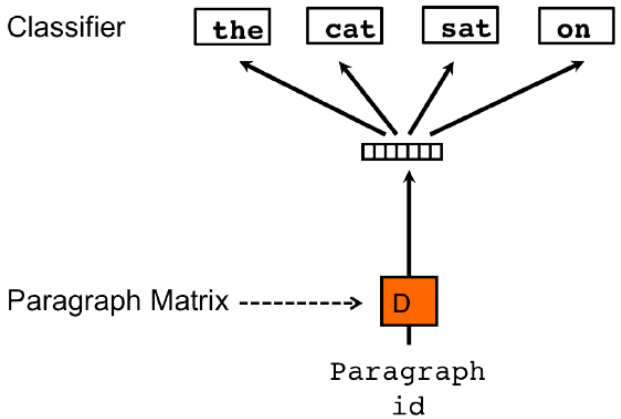
- Необходимо получить векторные представления параграфов (документов, предложений,...).
- Простой подход: усреднить слова, входящие в параграф.
  - или взвешенно усреднить, учитывая частоту встречаемости слов и их тематику.
- Точнее работает непосредственное представление самих параграфов.

Paragraph vector: модель PV-DM



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

Paragraph vector: модель PV-DBOW

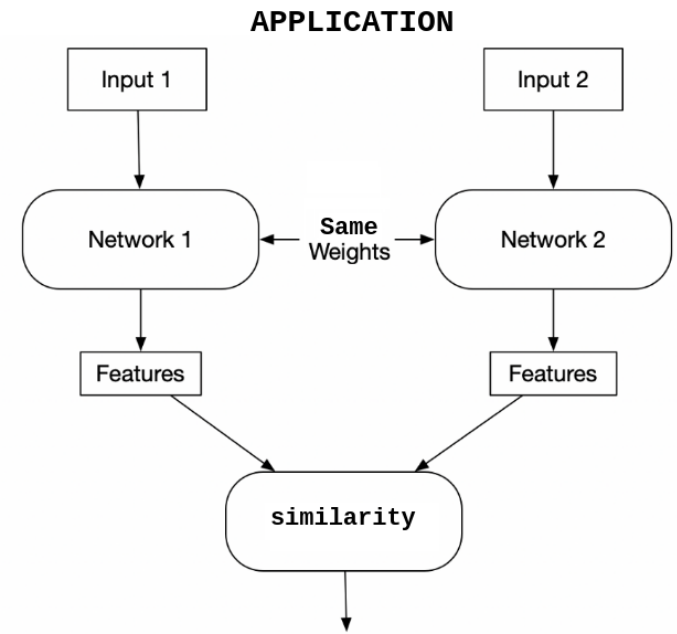


- Во время обучения делим документы на параграфы. Каждому параграфу -> векторное представление.
- Оценивается skip-gram: предсказываются случайные слова параграфа по представлению параграфа.
  - проще PV-DM, нужно хранить только представления параграфов
- Называется *Distributed Bag of Words version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

Содержание

Сиамская сеть

- Сиамская сеть использует 2 представления произвольных объектов.
- Прогноз - результат сравнения представлений.

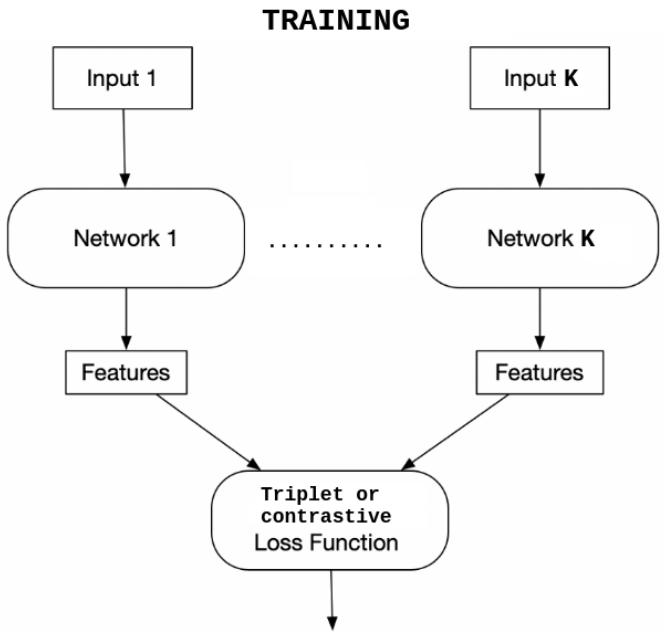


Мотивация: находим связь между сравниваемыми объектами.

Примеры приложений

- Классификация:
  - вход: 2 объекта или тестовый объект и центроид класса
  - выход: похожесть объектов или близость к определенному классу
- Поисковая система
  - вход: документ и поисковый запрос (м. быть поиск по картинке)
  - выход: степень релевантности документа запросу
- Обнаружение перефразирования:
  - вход: 2 предложения
  - выход: насколько они близки по смыслу
- Проверка подписи
  - вход: сканы 2х подписей
  - выход: их степень принадлежности одному человеку

Обучение



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

Функции потерь

Контрастные потери (contrastive loss):

- обучение на случайных парах объектов x_i, x_j

$$\mathbb{I}[y_i = y_j] \left\| f_\theta(x_i) - f_\theta(x_j)) \right\|^2 + \mathbb{I}[y_i \neq y_j] \max\left\{0, \alpha - \left\| f_\theta(x_i) - f_\theta(x_j)) \right\|\right\}^2$$

Тройные потери (triplet loss):

- обучение на случайных тройках x, x^+, x^-.
- x - опорный объект (anchor)
- x^+ - похожий на x (например, того же класса)
- x^- - не похожий на x (например, др. класса)
- $\alpha > 0$ - гиперпараметр

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

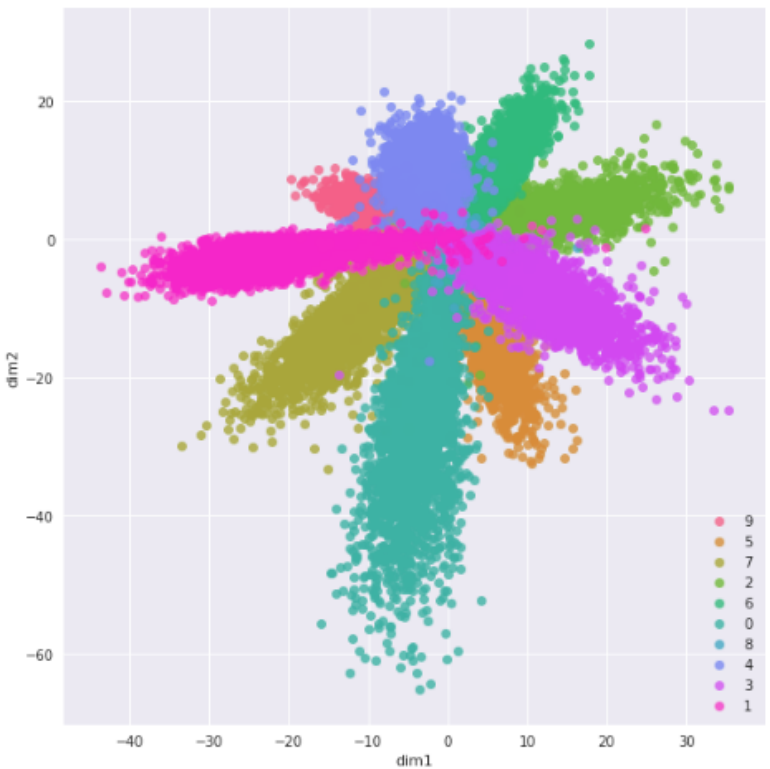
- Могут использоваться для metric learning $\rho_\theta(x, x')$.
- Обзор более продвинутых ф-ций потерь.

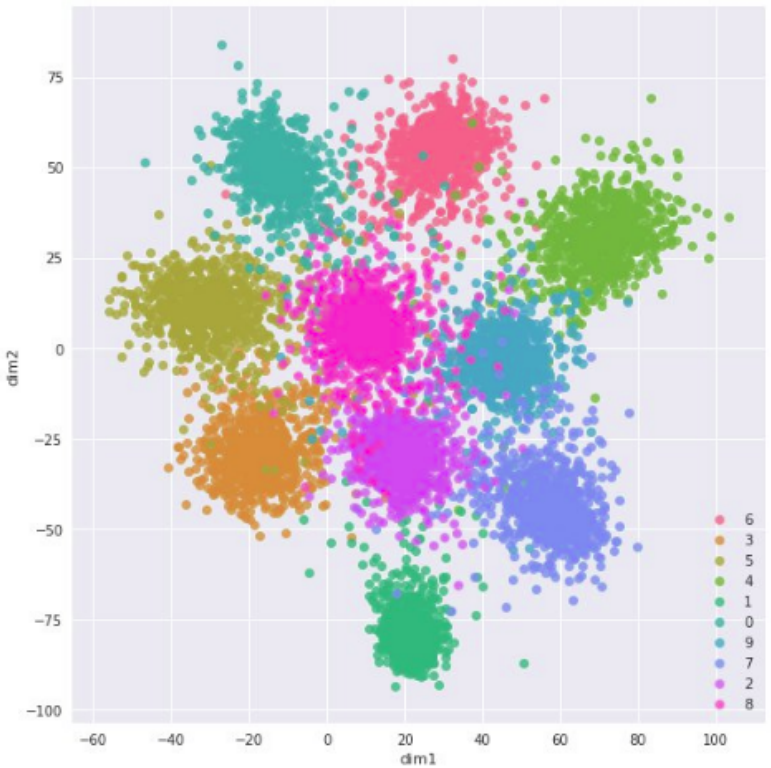
Сиамская сеть и классификация

- 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.
  - хорошо работает в ансамбе с классификатором
    - diversity↑, т.к. совсем др. принцип работы
  - требует больше обучения
    - обучение не на объектах, а на парах (contrastive loss) и тройках (triplet loss).

Представления объектов: классификация и сиамская сеть

Представления объектов: классификация и сиамская сеть для MNIST:





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 последовательностей (например, нуклеотиды в ДНК)
- **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, ...