Skip-gram & negative sampling

Практикум на ЭВМ, весна 2018

Попов Артём Сергеевич

МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП

29 марта 2018 г.

Задача построения word embeddings

Дано: $D = \{w_1, w_2, \ldots, w_N\}$ — текстовая коллекция
$w_i \in W$ — словарь коллекции

Найти: векторное представление $v_w \in \mathbb{R}^m$ для каждого слова w , где $m \ll |W|$

Какие представления считать хорошими?
Семантически (синтаксически) близким словам соответствуют близкие вектора

Какие слова считать близкими?
Использовать *гипотезу дистрибутивности*: слова, которые встречаются рядом с одними и теми же словами, имеют схожее значение

Модель Skip-gram

Идея: по слову предсказать все слова, находящиеся рядом

. . . an efficient $\underbrace{\text{method for}}_{\text{context, k = 2}}$ learning $\underbrace{\text{high quality}}_{\text{context, k = 2}}$ vector . . .

Модель Skip-gram:

$$\mathcal{L}(U, V) = \sum_{i=1}^{N} \sum_{\substack{j=-k \\ j \neq 0}}^{k} \log p(w_{i+j}|w_i) \to \max_{V,U}$$

$$p(c|w) = \underset{c \in W}{\text{softmax}} \langle v_c, u_w \rangle = \frac{\exp(\langle v_c, u_w \rangle)}{\sum_{c' \in W} \exp(\langle v_{c'}, u_w \rangle)}$$

Одна итерация SGD для модели: $O(|W|m)$

Способы упрощения модели

1. Явная аппроксимация софтмакса
   - Hierarchical Softmax
   - Differentiated Softmax
   - ...
2. Методы, основанные на сэмплировании
   - Noise Contrastive Estimation
   - Negative sampling
   - Importance sampling
   - Self-normalization
   - ...

Множество элементарных исходов в модели Negative sampling

Множество элементарных исходов в модели Skip-gram:

$$\Omega = \{W \times W\} \qquad p(w, c) = \underset{c \in W}{\mathrm{softmax}} \langle v_c, u_w \rangle p(w)$$

Интерпретация: вероятность встретить пару (w, c) в коллекции

Пусть теперь у нас будет $|W| \times |W|$ множеств элементарных исходов, каждое состоит из двух элементов:

$$\Omega_{wc} = \{0, 1\} \qquad p(1|c, w) = \sigma(\langle v_c, u_w \rangle)$$

Интерпретация: вероятность того, что пара (w, c) может встретиться в коллекции

Наивный функционал модели Negative sampling

Наивный функционал, максимизирующий правдоподобие:

$$\sum_{i=1}^{N} \sum_{\substack{j=-k \\ j \neq 0}}^{k} \log p(1|w_{i+j}, w_i) \to \max_{V,U}$$

Тривиальное решение:

$$w_{i+j} = w_i, \ ||w_i||^2 = 1000, \text{ для всех пар } i, i+j \text{ из выборки}$$

Механизм «регуляризации» — добавление негативных примеров (negative samples)

negative sample — случайно сгенерированное слово из распределения $p(w)^{3/4}$ в пару к w_i

Функционал модели Negative sampling

Итоговый функционал:

$$\sum_{i=1}^{N}\left(\sum_{\substack{j=-k\\ j\neq 0}}^{k}\log p(1|w_{i+j},w_i)+\sum_{\substack{k=1\\ w'_k\sim p(w)^{3/4}}}^{K}\log p(0|w_i,w'_k)\right)\to\max_{V,U}$$

Можно записать так:

$$\sum_{i=1}^{N}\left(\sum_{\substack{j=-k\\ j\neq 0}}^{k}\log p(1|w_{i+j},w_i)+K\mathbb{E}_{w\sim p(w)^{3/4}}\log p(0|w_i,w)\right)\to\max_{V,U}$$

Функционал модели Negative sampling

Немного другой алгоритм, генерируются негативные пары слов:

$$\sum_{i=1}^{N} \sum_{\substack{j=-k \\ j \neq 0}}^{k} \log p(1|w_{i+j}, w_i) + \sum_{(w,c) \in D'} \log p(0|w, c) \to \max_{V,U}$$

D' — множество сгенерированных негативных примеров

$$D' : \left\{ (w, c) \sim p(w)^{3/4} p(c) \right\}$$

В skip-gram генерация пар усложняет жизнь, но в каких-то задачах может упрощать