Обработка естественного языка, лекция 6: механизм внимания

Пожидаев М. С.

15 декабря 2021 г.

Механизм внимания

Проблема прежняя

Влияние состояний кодера ближе к последним элементам входной последовательности сильнее, чем влияние элементов в начале.

Attention!

Позволим в вектор контекста декодера делать вклад всем состояниям кодера, но с учётом коэффициентов внимания.

Общая формула состояния декодера

$$s_t = f(s_{t-1}, y_{t-1}, c_t)$$

- s_{t-1} — состояние декодера на предыдущем шаге;
- y_t — текущий элемент выходной последовательности;
- c_t — текущий вектор контекста.;

Вычисление вектора контекста

$$c_t = \sum_{i=1}^{n} \alpha_{ti} h_i$$

Коэфициенты соответствия $\alpha_{tt'}$ позволяют определить значимость элементов входной последовательности, представленных соответствующими скрытыми состояниями кодера $h_{t'}$ на элемент выходной последовательности y_t через влияние на соответствующее скрытое состояние декодера s_t. Необходимо обратить внимание, что влияние на элемент выходной последовательности y_t может оказывать любой элемент входной последовательности.

Вычисление коэффициентов соответствия

$$\alpha_{ti} = \frac{\exp(score(s_{t-1}, h_i))}{\sum_{i'=1}^{n} \exp(score(s_{t-1}, h_{i'}))}$$

α_{ti} — это применение $softmax(x)$ к оценке соответствия пары элементов, первый из которых принадлежит выходной последовательности в положении t, а второй — входной в положении i, обозначая, насколько хорошо они соответствуют друг другу.

Подход Бахданау

Кодер представляет собой двунаправленную рекуррентную сеть.

$$score(s_t, h_i) = W_1 \tanh(W_2 h_i + W_3 s_t)$$

Подход Луонга

Кодер представляет собой однонаправленную рекуррентную сеть.

Варианты подсчёта соответствия:

- $score(s_t, h_i) = h_i \cdot s_t$
- $score(s_t, h_i) = W(h_i \cdot s_t)$
- $score(s_t, h_i) = W_1 \cdot tanh(W_2(h_i + s_t))$

Self-attention

Как и ячейка рекуррентной ИНС, Self-attention кодирует некоторую входную последовательность с представлением результата в виде вектора, но при этом рекуррентной ИНС не является.

Преимущество!

В основу способа кодирования ложится информация о взаимосвязях внутри входной последовательности.

Особенности Self-attention

1. Длина входной последовательности должна быть фиксированной.
2. Данные в каждом x_i должны быть равномерно распределены по всем элементам (one-hot нельзя!).
3. Может применяться для иерархического кодирования.

Query, key и value

Каждый элемент входной последовательности x_t дополнительно кодируется в три новых вектора: *query* (q_t), *key* (k_t) и *value* (v_t).
Кодирование происходит путём добавления трёх линейных ячеек:

$$q_t = W_q x_t$$
$$k_t = W_k x_t$$
$$v_t = W_v x_t$$

Вычисление внимания

Вычисление значения внимания для x_t производится по следующей формуле:

$$\alpha(x_t) = softmax([a_{t1}, a_{t2}, \ldots, a_{tn}])v_t$$

$$ati = \frac{q_t k_i}{\sqrt{n}}$$

В этой формуле вектор k принимает по очереди все значения векторов k_i.

Спасибо за внимание!

Веб-сайт: https://marigostra.ru/
E-mail: msp@luwrain.org