Обработка естественного языка. Лекция 4 Механизм внимания

Михаил Пожидаев

26 сентября 2023 г.

Механизм внимания

Слабость связи между декодером и кодером

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.

Что влияет на декодер?

Общая формула состояния декодера

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

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

Усилим декодер

Порядок 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 c_t . Необходимо обратить внимание, что влияние на элемент выходной последовательности y_t может оказывать любой элемент входной последовательности.

Получение α_{ti}

Вычисление коэффициентов соответствия

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

 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 Богданова

Кодер — двунаправленная RNN

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

Подход Луонга

Кодер — однонаправленная RNN

Варианты подсчёта соответствия:

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

Self-attention

Внутреннее внимание

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. Применяется только для кодирования последовательности в вектор.

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:

$$lpha(x_t) = softmax([a_{t1}, a_{t2}, \dots, a_{tn}])v_t$$
 $ati = rac{q_t k_i}{\sqrt{n}}$

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

Positional encoding

Кодирование положения токена

Непонятно, где это было!

Self-attention не сохраняет информацию о положении исходного элемента, поэтому добавим дополнительное кодирование (position encoding).

$$PE(pos, 2i) = \sin(pos/10000^{2i/d})$$

 $PE(pos, 2i + 1) = \cos(pos/10000^{2i/d})$

- ▶ pos индекс входного токена;
- ightharpoonup 2i и 2i + 1 индекс компоненты вектора;
- ightharpoonup d размерность векторного пространства.

Multi-head attention

Разобьём исходное векторное пространство входной последовательности:

- 1. Векторы query, key и value заменяются k проекциями (пусть k = 8).
- 2. Каждая проекция обрабатывается отдельно своим механизмом Self-attention.
- 3. Результат склеивается обратно.

Плюсы:

- обработка становится параллельной;
- повышается общее качество обработки.

Вычисление проекций

Ключевой параметр — количество проекций

Таким образом, функция Multi-head attention может быть представлена следующим образом:

$$multihead(Q, K, V) = concat(head_1, \dots, head_n)W'_o$$

В приведённом выражении n — количество проекций, а каждый $head_i$ представляет отдельный вариант использования механизма внимания.

Спасибо за внимание!

Всё о курсе: https://marigostra.ru/materials/nlp.html

E-mail: msp@luwrain.org

Канал в Телеграм: https://t.me/MarigostraRu