Deep Learning. Recurrent neural networks. GRU. LSTM

Урок 8

Егор Конягин

24 июля 2019 г.

МФТИ & АО ”ЦОСиВТ”

Содержание

1. О временных последовательностях

2. RNN

3. GRU - gated recurrent unit

4. LSTM

О временных последовательностях

Временные последовательности

Когда мы рассматривали полносвязные и сверточные нейросети, мы имели дело со статическим набором данных, то есть не было в явном виде зависимости от времени t.

Примеры временных последовательностей:

- звук,
- речь (перевод, распознавание и тд.),
- видеопоток.

Обозначения

Поскольку входные данные - это некая временная зависимость, то мы будем использовать следующие индексы:

$$x^{<1>}, x^{<2>} \dots x^{<p>}.$$

Если вывод нейосети - это тоже последовательность, то используется аналогичная нотация:

$$y^{<1>}, y^{<2>} \dots y^{<t>}.$$

Соответственно:

$$X^{(i)<t>}$$

- t-ый элемент последовательности i-ого объекта выборки.

Кодирование текста

Предположим, что мы рассматриваем задачу перевода. Перед тем как непосредственно ее решить, требуется слова закодировать, т.е. перевести в векторы. Рассмотрим самый наивный подход:

$$\begin{bmatrix} a \\ Aaron \\ and \\ \dots \\ Zulu \end{bmatrix} \rightarrow One - hot - encoding \rightarrow a = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \\ 0 \end{bmatrix} \quad Aaron = \begin{bmatrix} 0 \\ 1 \\ \dots \\ 0 \\ 0 \end{bmatrix} \quad (1)$$

Почему не полносвязная нейросеть?

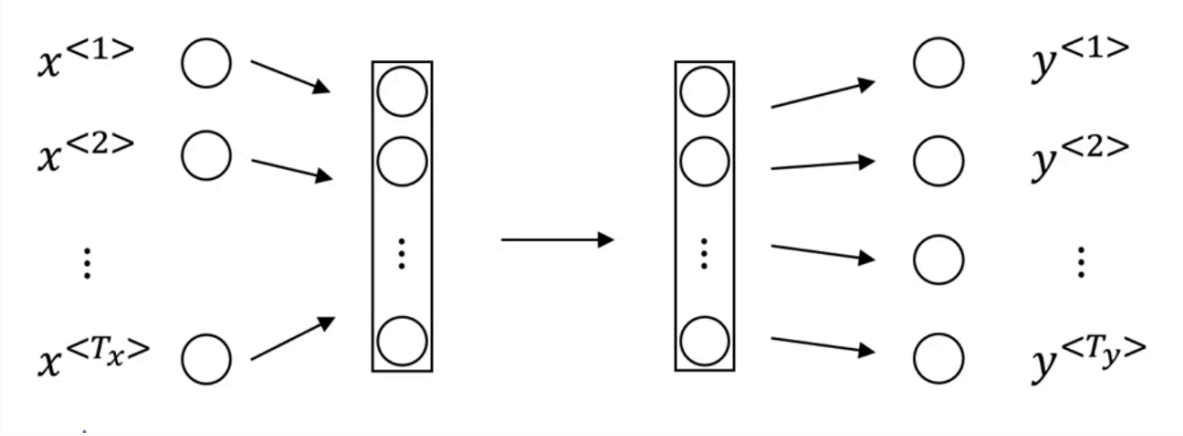

Рис. 1: Полносвязная модель

Проблема 1 Данные могут быть разной длины. **Проблема 2** Инвариантность относительно позиции. **Проблема 3** Отсутсвие упорядоченности. **Проблема 4** Очень большое кол-во параметров.

RNN

RNN. Forward propagation

В модели RNN ответ, полученный на t-ом элементе x, передается далее

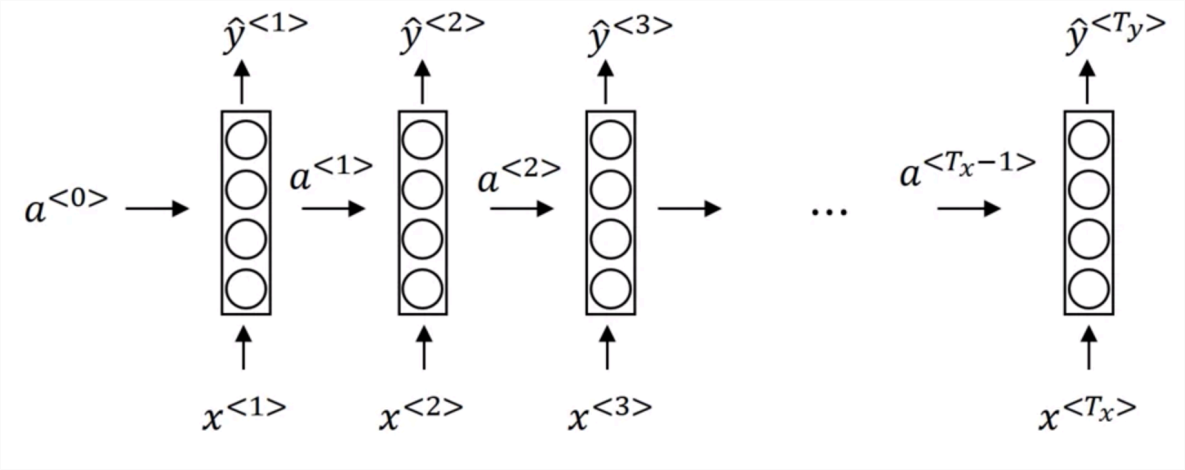

Рис. 2: Простейшая модель RNN. Источник: Andrew Ng

$$a^{<t>} = \sigma_1(w_{aa} \cdot a^{<t-1>} + w_{ax} \cdot x^{<t>} + b_a) \quad a^{<0>} = \overrightarrow{0} \tag{2}$$

$$y^{<t>} = \sigma_1(w_{ya} \cdot a^{<t-1>} + b_a) \tag{3}$$

RNN. Backward propagation

В качестве функции ошибки можно взять известную нам бинарную кросс-энтропию:

$$\mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -y^{<t>} \log \hat{y}^{<t>} - (1 - y^{<t>}) \log(1 - \hat{y}^{<t>}) \tag{4}$$

$$\mathcal{L}(\hat{y}, y) = -\sum_{t=0}^{T_x} y^{<t>} \log \hat{y}^{<t>} + (1 - y^{<t>}) \log(1 - \hat{y}^{<t>}) \tag{5}$$

Далее последовательно считаются $\frac{\partial \mathcal{L}}{\partial a^{<T\_x>}}, \frac{\partial \mathcal{L}}{\partial a^{<T\_x-1>}} \cdots$. После подсчета этих величин (backpropagation through time) можно рассчитать градиенты по параметрам и обновить параметры.

Разные типы RNN

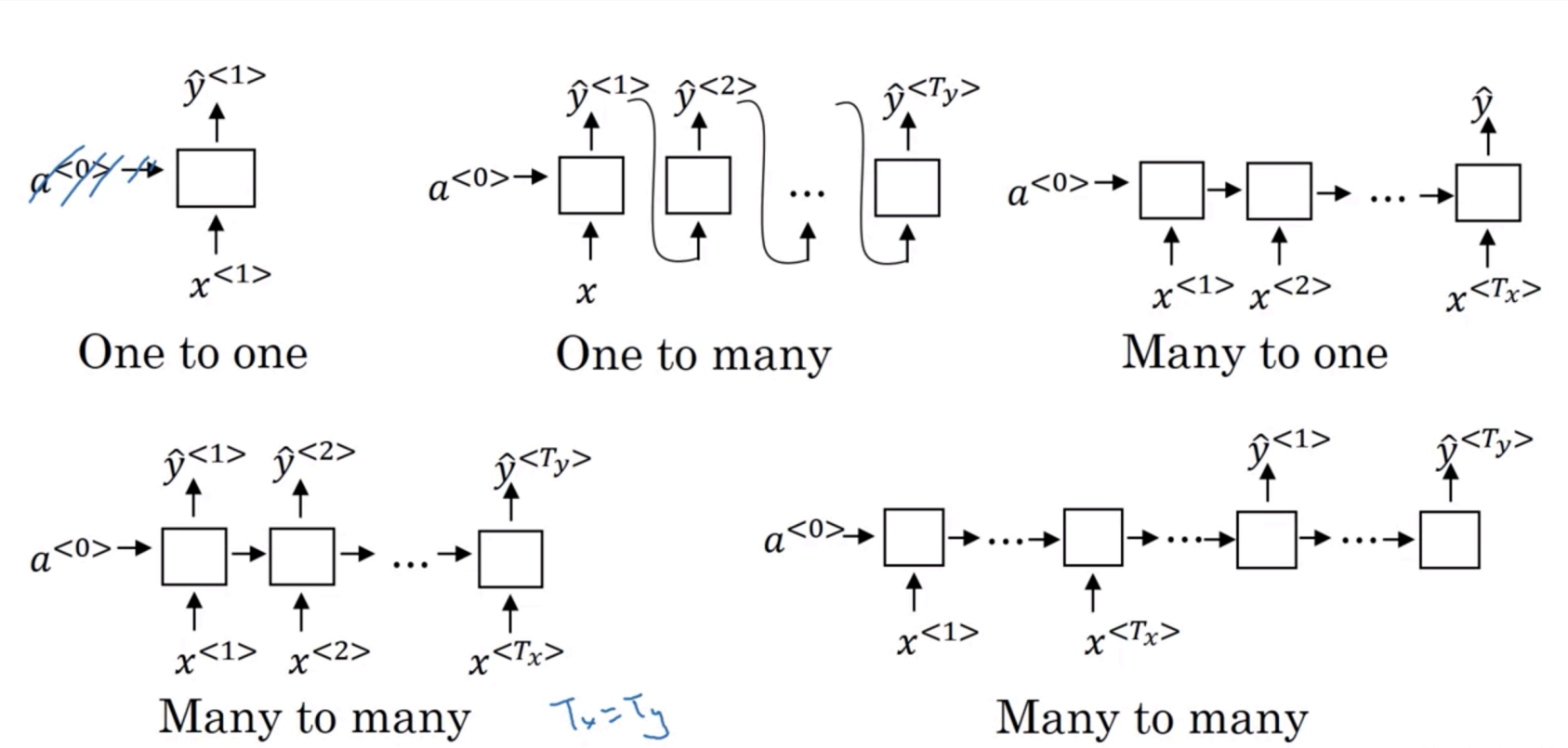

Рис. 3: Типы RNN. Источник: Andrew Ng

Проблема затухающих градиентов

Рассматривая модель текста, заметим следующее:

The cat which already ate was funny...

The cats which... were funny... Грамматическое число слова cat влияет на слова, которые могут стоять сильно дальше (т.н. долгосрочные зависимости).

Вследствие такого явления рассмотренная модель RNN ”принимает во внимание”только близкие к текущему объекты. Поэтому мы сталкнемся с проблемой затухания градиентов.

Техника gradient clipping: если $|(\frac{\partial \mathcal{L}}{\partial w})_i| > T \rightarrow (\frac{\partial \mathcal{L}}{\partial w})_i = T$.

GRU - gated recurrent unit

GRU. Начало

Вспомним forward-prop для RNN:

$$a^{<t>} = \sigma(w_a \cdot a^{<t-1>} + w_x \cdot x^{<t>} + b_a) \tag{6}$$

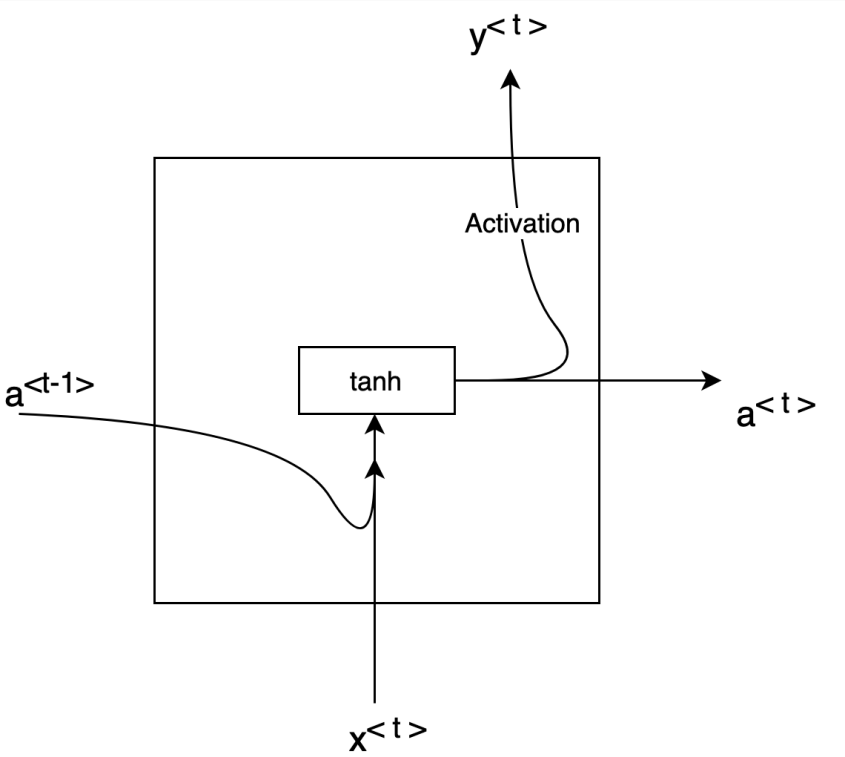

Рис. 4: RNN-ячейка

GRU

GRU имеет так называемый ”затвор”:

$$c^{<t>} = a^{<t>} \tag{7}$$

$$\hat{c}^{<t>} = \tanh(w_c c \cdot c^{<t-1>} + w_c x \cdot x^{<t>} + b_c) \tag{8}$$

$$\Gamma_u = \sigma_\Gamma(w_{uc} \cdot c^{<t-1>} + w_{ux} \cdot x^{<t>} + b_u) \tag{9}$$

ВАЖНО! σ_Γ должна иметь множество значений (0,1).

$$c^{<t>} = \Gamma_u \cdot \hat{c}^{<t>} + (1 - \Gamma_u) \cdot c^{<t-1>} \tag{10}$$

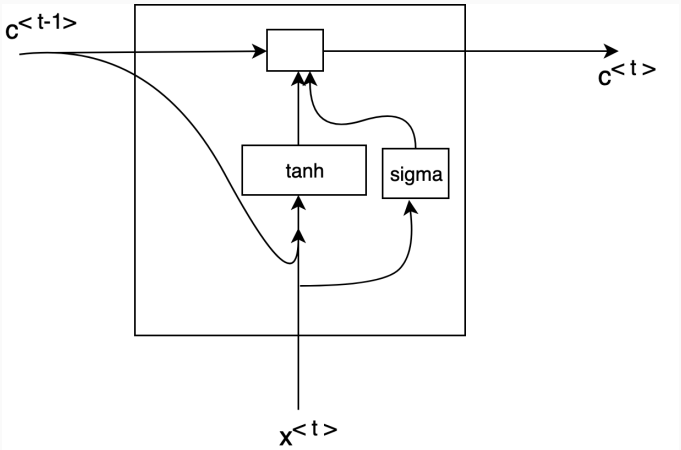

Рис. 5: GRU-ячейка

LSTM

LSTM - long short-term memory

Forward propagation:

$$\hat{c}^{<t>} = \tanh(w_{ca} \cdot a^{<t-1>} + w_{cx} \cdot x^{<t>} + b_c) \tag{11}$$

$$\Gamma_u = \sigma_\Gamma(w_{ua} \cdot a^{<t-1>} + w_{ux} \cdot x^{<t>} + b_u) \tag{12}$$

$$\Gamma_f = \sigma_\Gamma(w_{fa} \cdot a^{<t-1>} + w_{fx} \cdot x^{<t>} + b_u) \tag{13}$$

$$\Gamma_o = \sigma_\Gamma(w_{oa} \cdot a^{<t-1>} + w_{ox} \cdot x^{<t>} + b_o) \tag{14}$$

$$c^{<t>} = \Gamma_u \cdot \hat{c}^{<t>} + \Gamma_f \cdot c^{<t-1>} \tag{15}$$

$$a^{<t>} = \Gamma_o \cdot \tanh c^{<t>} \tag{16}$$

LSTM. Схема

Схематично описать LSTM можно следующим образом:

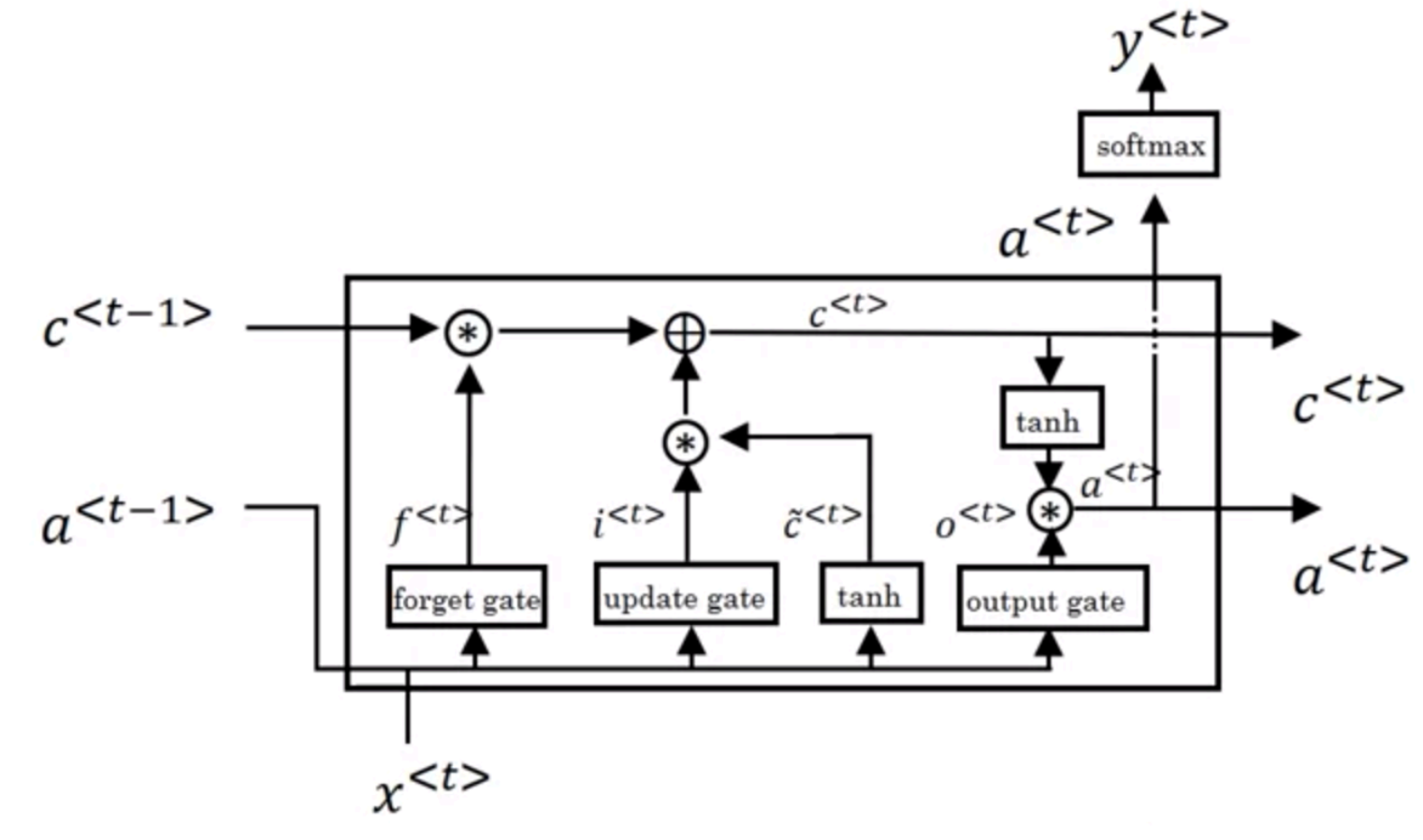

Рис. 6: LSTM-ячейка. Источник: Andrew Ng's classes

Deep RNN

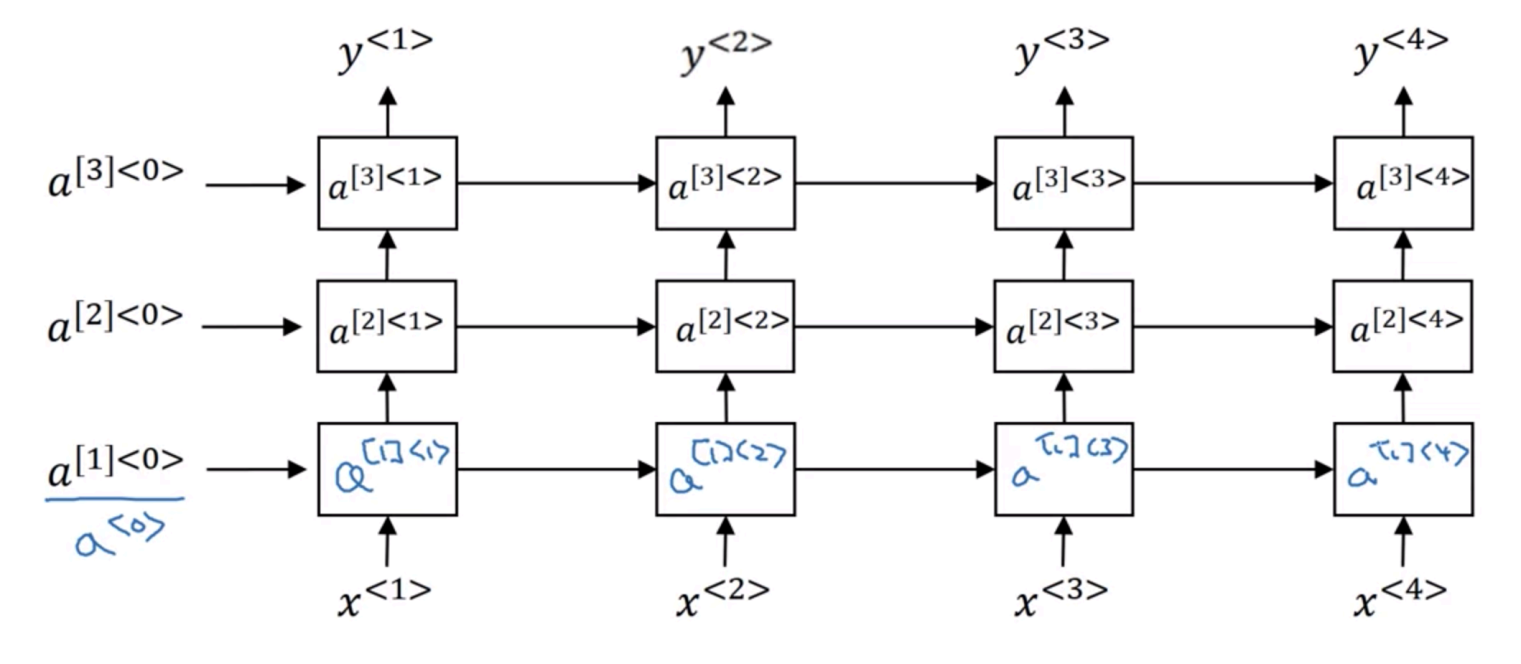

Рис. 7: Deep RNN. Источник: Andrew Ng's classes

RNN редко строятся глубже, чем в три слоя.

Summary

Мы рассмотрели сегодня

- простейшую сеть RNN;
- элемент GRU;
- элемент LSTM.