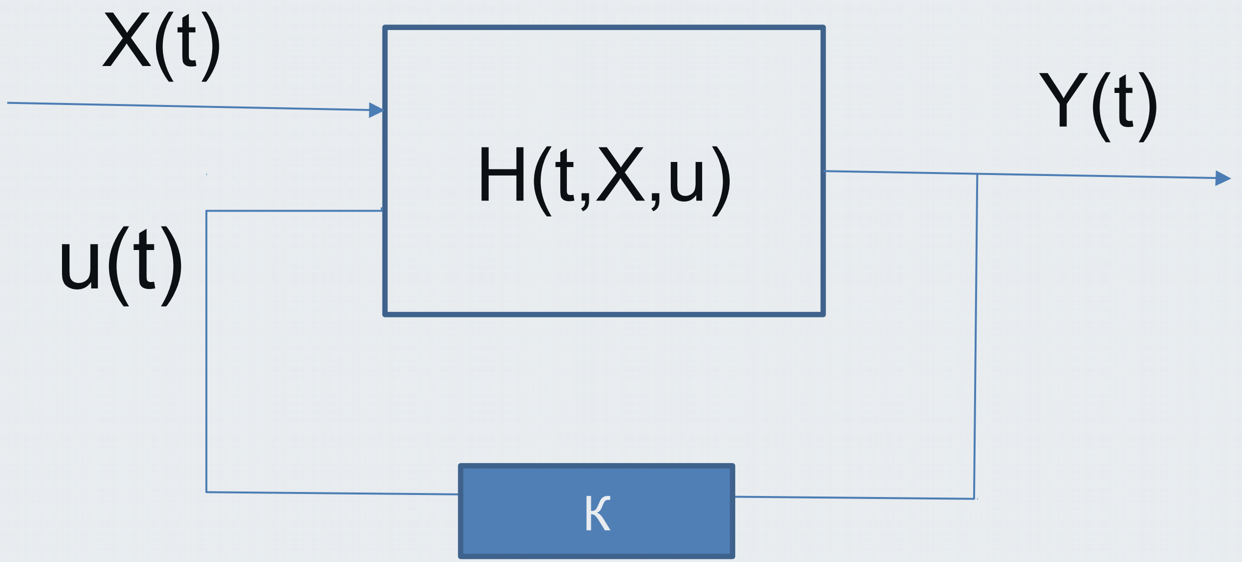
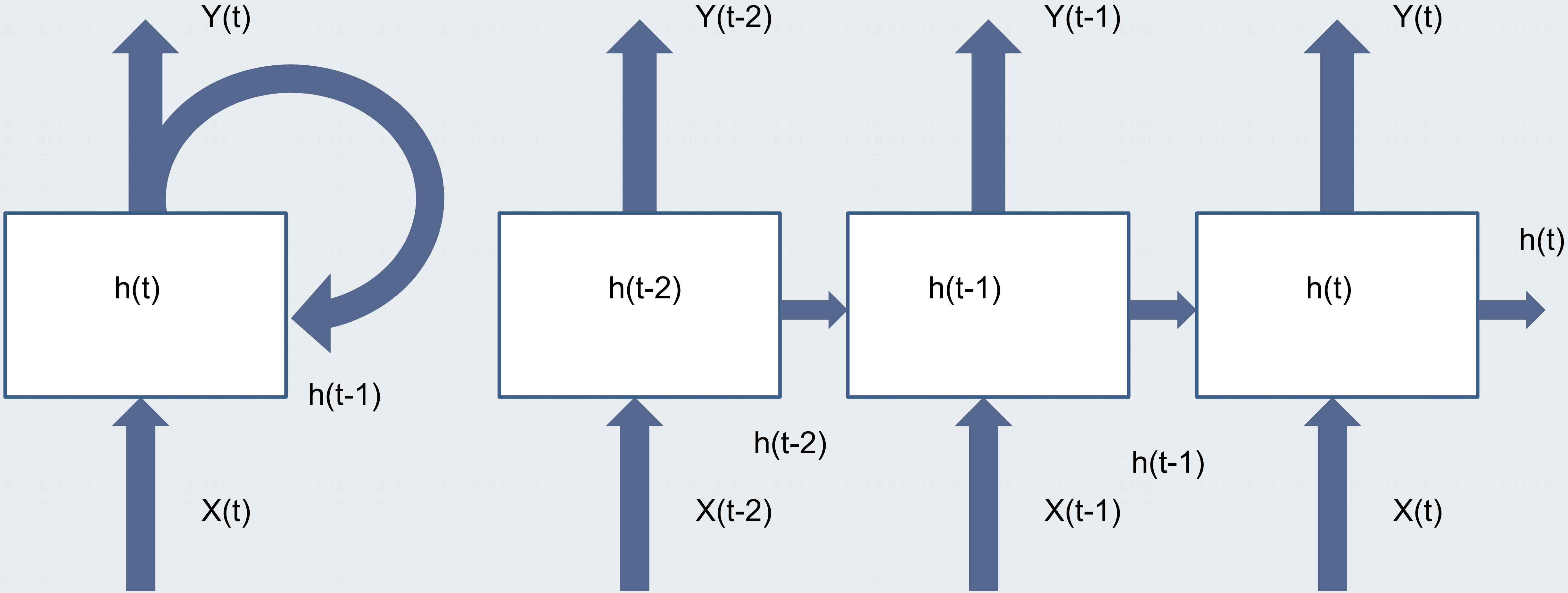
Рекуррентные модели и нейронные сети



Динамическая система

Рекуррентные модели



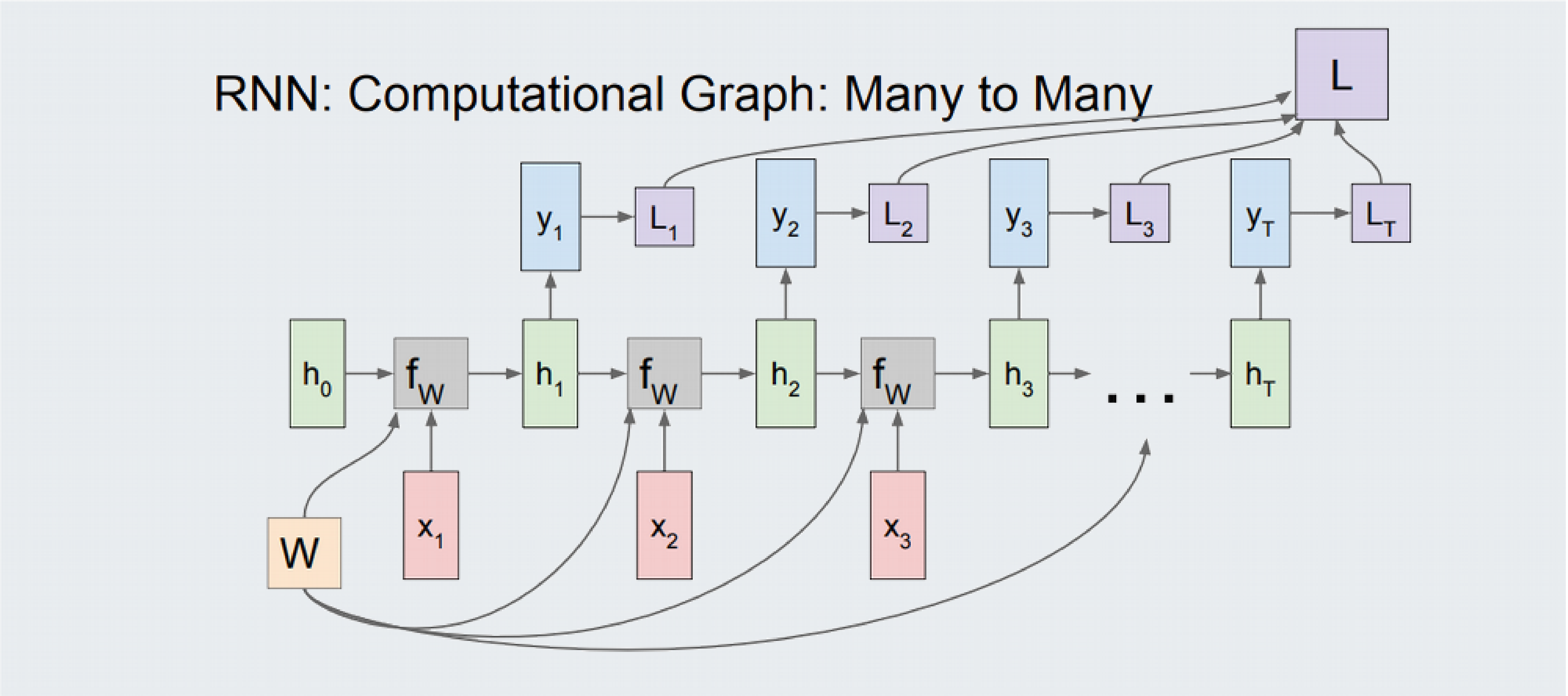
$$h(t) = f(W, h(t-1), X(t))$$

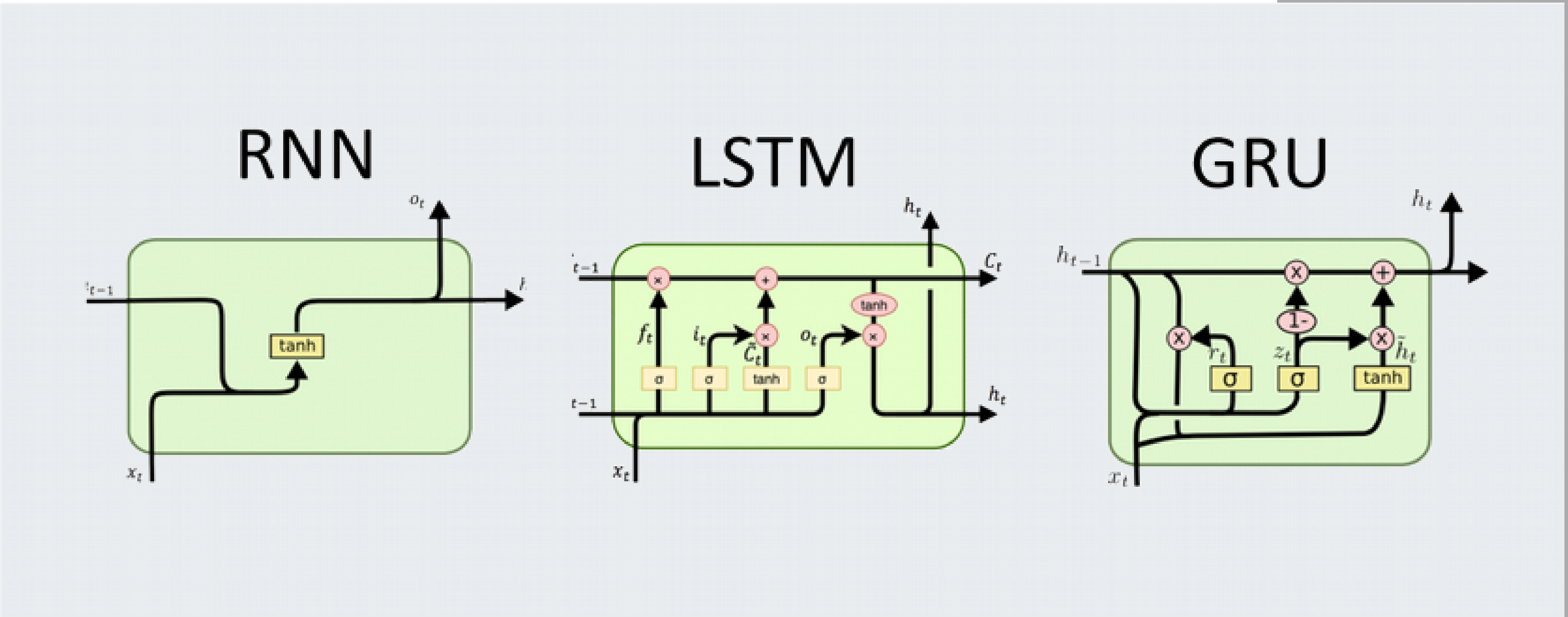
$$h(t) = \tanh\big(Wh \cdot h(t-1) + Wx \cdot X(t)\big)$$

$$Y(t) = Wy \cdot h(t)$$

Рекуррентные модели

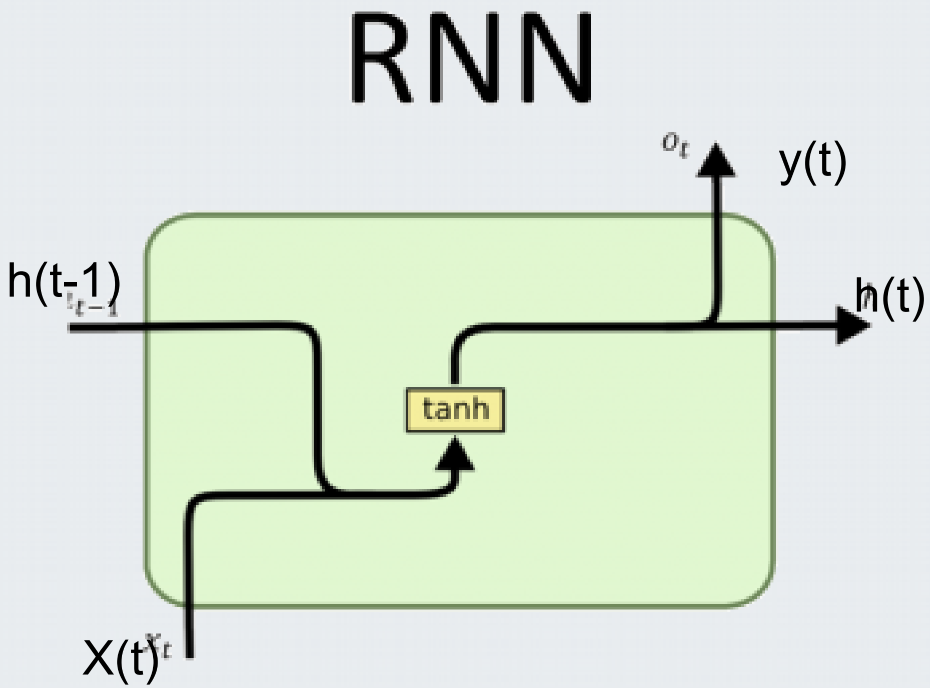
Рекуррентные модели





Виды рекуррентных узлов

Обратное Распространение Во Времени



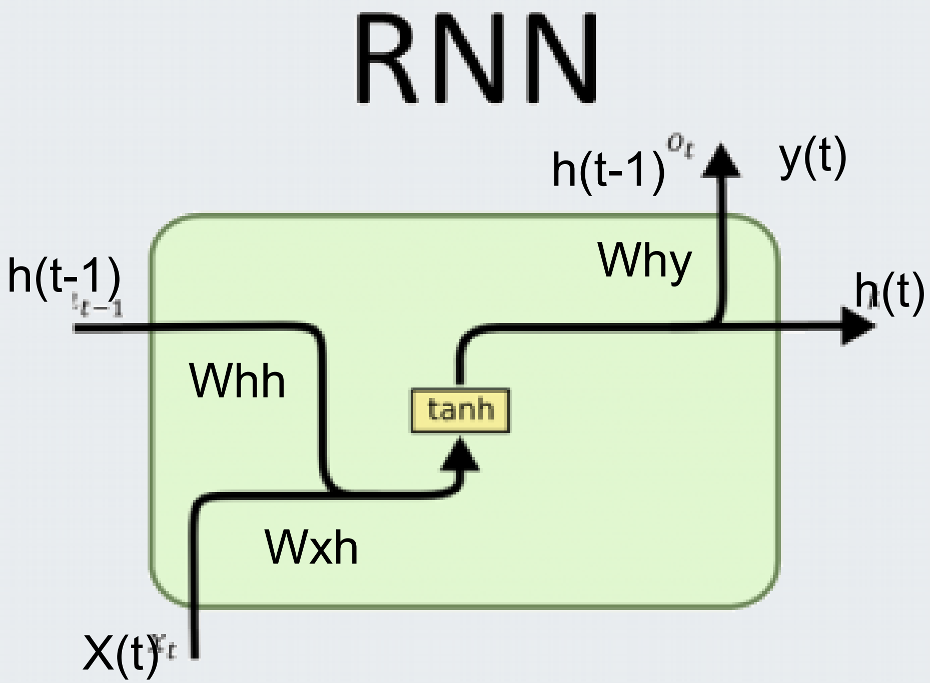
$$h(t) = \tanh\left(W_{xh} \cdot X(t) + W_{hh} \cdot h(t-1) + b_h\right)$$

$$y(t) = W_{hy} \cdot h(t) + b_y$$

$$Pc = soft\max(y(t))$$

$$L = -\log\left(Pc\right)$$

Обратное Распространение Во Времени



Инициализация весовых матриц :

- **инициализации Wxh (вход-скрытый),**
- **Whh (скрытый-скрытый),**
- **Why (скрытый-выход).**

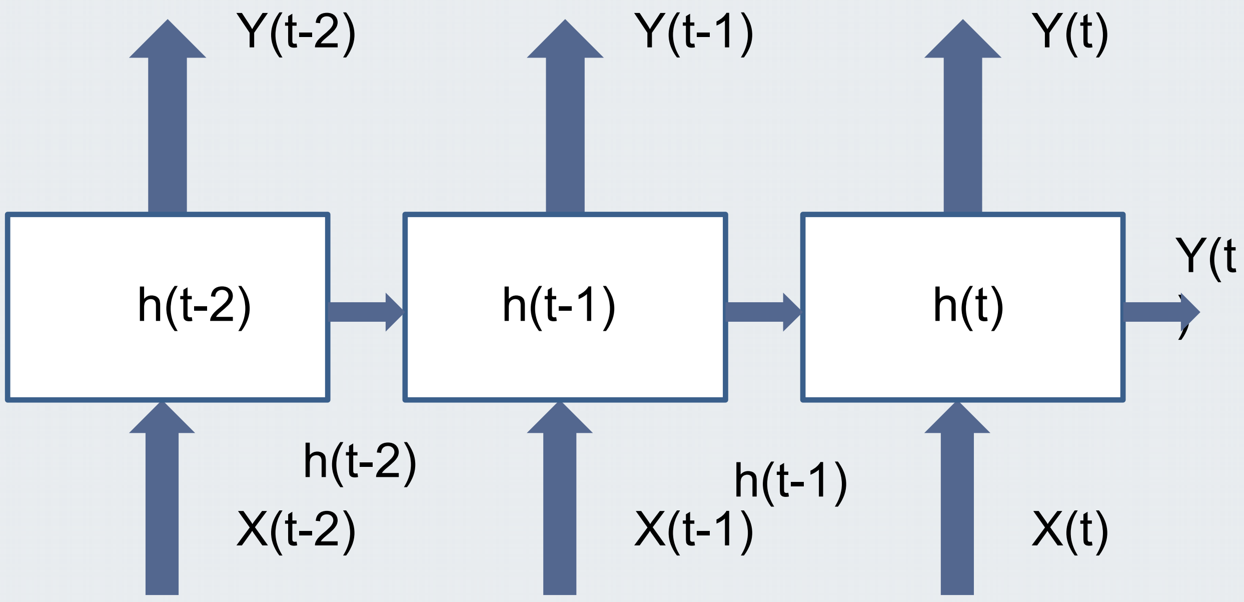
$$прямой \quad проход:$$

$$y(n), h(h), (x(t), h(t), y(t), t = 1, n)$$

$$обратный \quad проход \quad ищем:$$

$$\frac{\partial L(n)}{\partial W_{xh}}, \frac{\partial L(n)}{\partial W_{hh}}, \frac{\partial L(n)}{\partial W_{hy}}, \frac{\partial L(n)}{\partial b_h}, \frac{\partial L(n)}{\partial b_y}$$

Обратное Распространение Во Времени



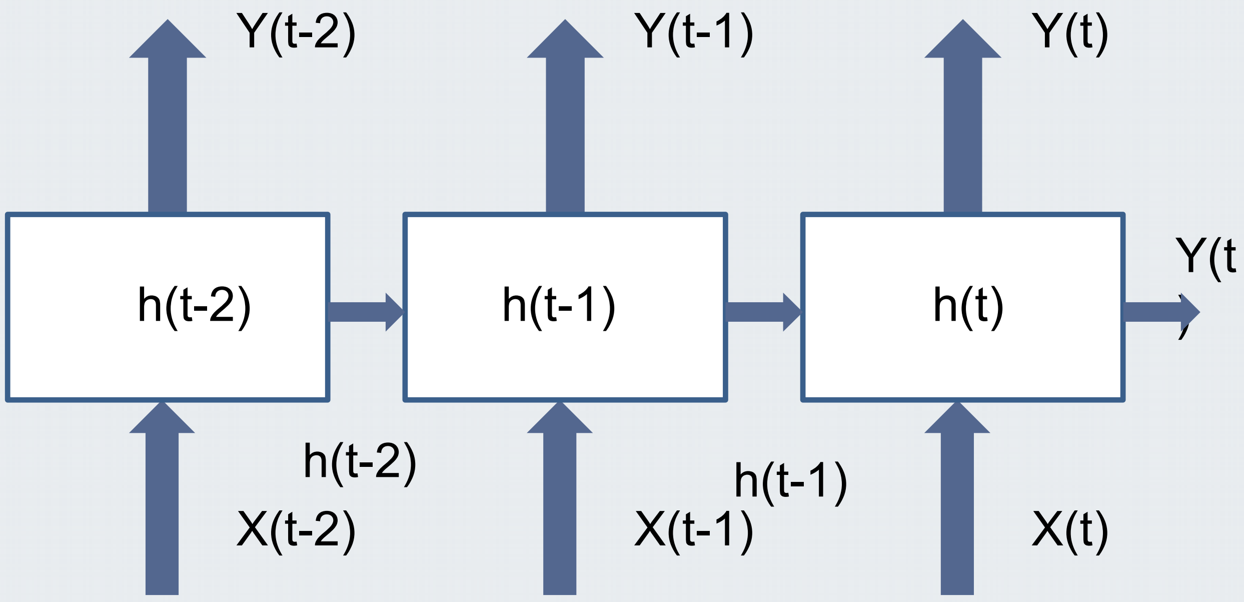
$$\text{прямой} \quad \text{проход}:$$
$$y(n), h(h), (x(t), h(t), y(t), t = 1, n)$$
$$\text{обратный} \quad \text{проход} \quad \text{ищем}:$$
$$\frac{\partial L(n)}{\partial W_{xh}}, \frac{\partial L(n)}{\partial W_{hh}}, \frac{\partial L(n)}{\partial W_{hy}}, \frac{\partial L(n)}{\partial b_h}, \frac{\partial L(n)}{\partial b_y}$$

$$\frac{\partial L(n)}{\partial W_{hy}} = \frac{\partial L(n)}{\partial y(n)} \frac{\partial y(n)}{\partial W_{hy}}, \quad y(n) = W_{hy} h(n) + b_y$$
$$\frac{\partial y(n)}{\partial W_{hy}} = h(n), \quad \frac{\partial y(n)}{\partial b_y} = 1$$
$$\frac{\partial L(n)}{\partial W_{hy}} = \frac{\partial L(n)}{\partial y(n)} h(n), \quad \frac{\partial L(n)}{\partial b_y} = \frac{\partial L(n)}{\partial y(n)}$$

$$h(t) = \tanh\left(W_{xh} \cdot X(t) + W_{hh} \cdot h(t-1) + b_h\right)$$
$$y(t) = W_{hy} \cdot h(t) + b_y$$

Backpropagation Through Time (BPTT):

Обратное Распространение Во Времени



$$\frac{\partial L(n)}{\partial W_{hy}} = \frac{\partial L(n)}{\partial y(n)} \frac{\partial y(n)}{\partial W_{hy}}, \quad y(n) = W_{hy} h(n) + b_y$$

$$\frac{\partial y(n)}{\partial W_{hy}} = h(n), \quad \frac{\partial y(n)}{\partial b_y} = 1$$

$$\boxed{\frac{\partial L(n)}{\partial W_{hy}} = \frac{\partial L(n)}{\partial y(n)} h(n),} \quad \boxed{\frac{\partial L(n)}{\partial b_y} = \frac{\partial L(n)}{\partial y(n)}}$$

прямой проход:

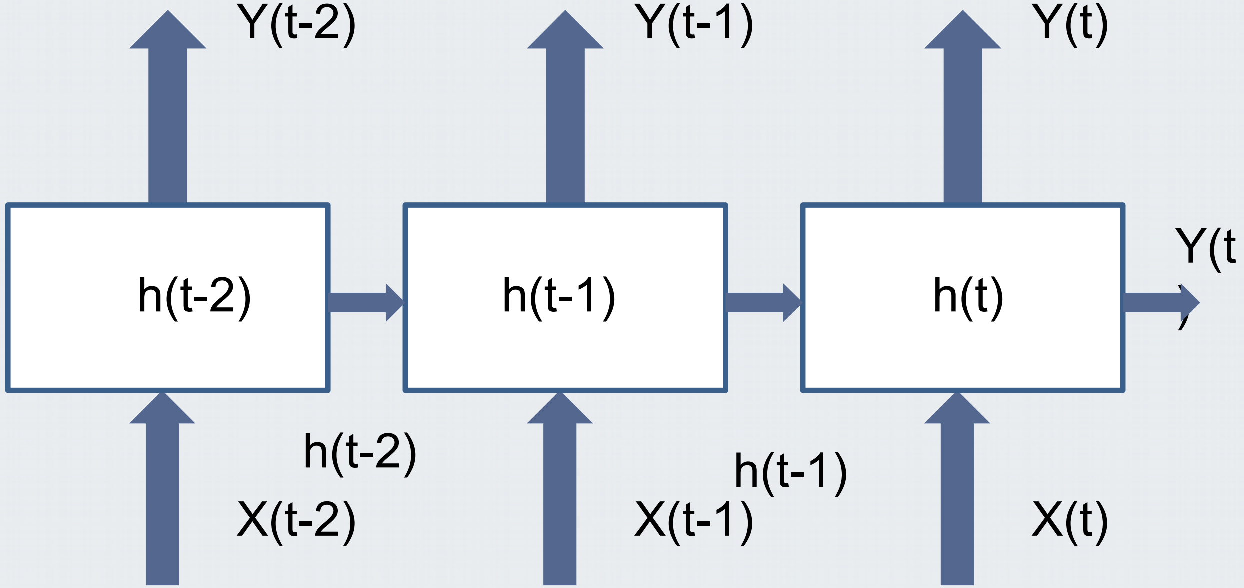
$y(n), h(h), (x(t), h(t), y(t), t = 1, n)$

обратный проход ищем:

$$\frac{\partial L(n)}{\partial W_{xh}}, \frac{\partial L(n)}{\partial W_{hh}}, \boxed{\frac{\partial L(n)}{\partial W_{hy}}}, \frac{\partial L(n)}{\partial b_h}, \boxed{\frac{\partial L(n)}{\partial b_y}}$$

Backpropagation Through Time (BPTT):

Обратное Распространение Во Времени



$$\frac{\partial L(n)}{\partial W_{xh}} = \frac{\partial L(n)}{\partial y(n)} \sum_t \frac{\partial y(t)}{\partial h(t)} \frac{\partial h(t)}{\partial W_{xh}}$$

$$\frac{\partial h(t)}{\partial W_{xh}} = (1 - h(t)^2) X(t)$$

$$\boxed{\frac{\partial L(n)}{\partial W_{xh}} = \frac{\partial L(n)}{\partial y(n)} \sum_t \frac{\partial y(t)}{\partial h(t)} (1 - h(t)^2) X(t)}$$

$\textit{прямой} \quad \textit{проход}:$

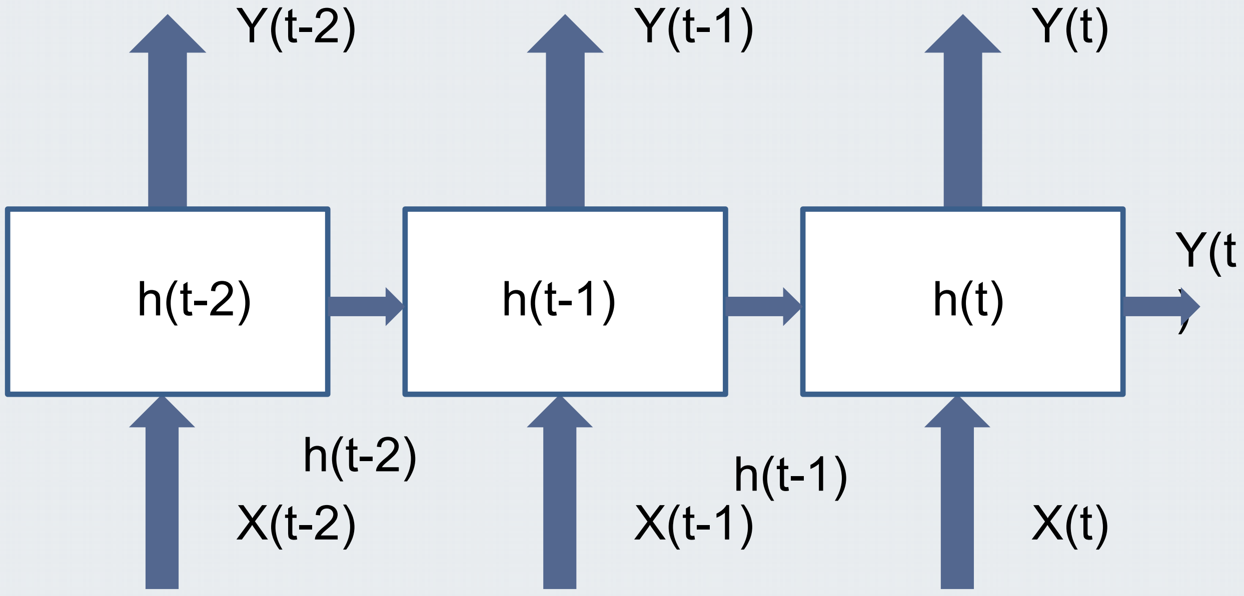
$y(n), h(h), (x(t), h(t), y(t), t = 1, n)$

$\textit{обратный} \quad \textit{проход} \quad \textit{ищем}:$

$$\boxed{\frac{\partial L(n)}{\partial W_{xh}}}, \frac{\partial L(n)}{\partial W_{hh}}, \frac{\partial L(n)}{\partial W_{hy}}, \frac{\partial L(n)}{\partial b_h}, \frac{\partial L(n)}{\partial b_y}$$

Backpropagation Through Time (BPTT):

Обратное Распространение Во Времени



$$\frac{\partial L(n)}{\partial W_{hh}} = \frac{\partial L(n)}{\partial y(n)} \sum_t \frac{\partial y(t)}{\partial h(t)} \frac{\partial h(t)}{\partial W_{hh}}$$

$$\frac{\partial h(t)}{\partial W_{hh}} = (1 - h(t)^2) h(t-1)$$

$$\boxed{\frac{\partial L(n)}{\partial W_{hh}} = \frac{\partial L(n)}{\partial y(n)} \sum_t \frac{\partial y(t)}{\partial h(t)} (1 - h(t)^2) h(t-1)}$$

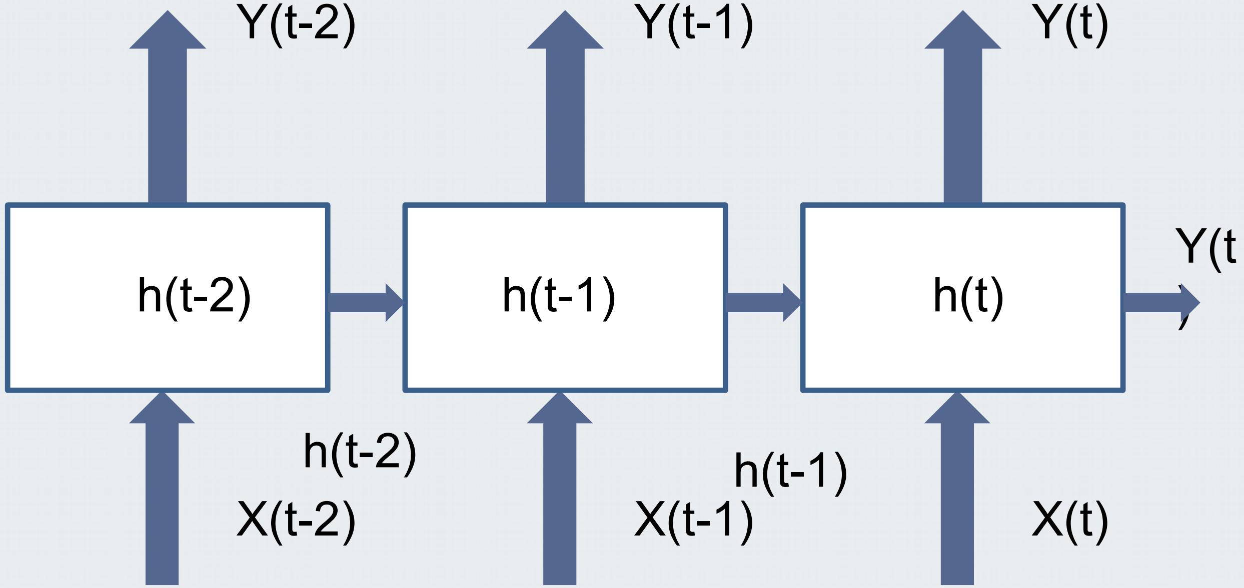
прямой проход:

$y(n), h(h), (x(t), h(t), y(t), t = 1, n)$

обратный проход ищем:

$$\frac{\partial L(n)}{\partial W_{xh}}, \boxed{\frac{\partial L(n)}{\partial W_{hh}}}, \frac{\partial L(n)}{\partial W_{hy}}, \frac{\partial L(n)}{\partial b_h}, \frac{\partial L(n)}{\partial b_y}$$

Обратное Распространение Во Времени



$$\frac{\partial L(n)}{\partial b_h} = \frac{\partial L(n)}{\partial y(n)} \sum_t \frac{\partial y(t)}{\partial h(t)} \frac{\partial h(t)}{\partial b_h}$$

$$\frac{\partial h(t)}{\partial b_h} = (1 - h(t)^2)$$

$$\boxed{\frac{\partial L(n)}{\partial b_h} = \frac{\partial L(n)}{\partial y(n)} \sum_t \frac{\partial y(t)}{\partial h(t)} (1 - h(t)^2)}$$

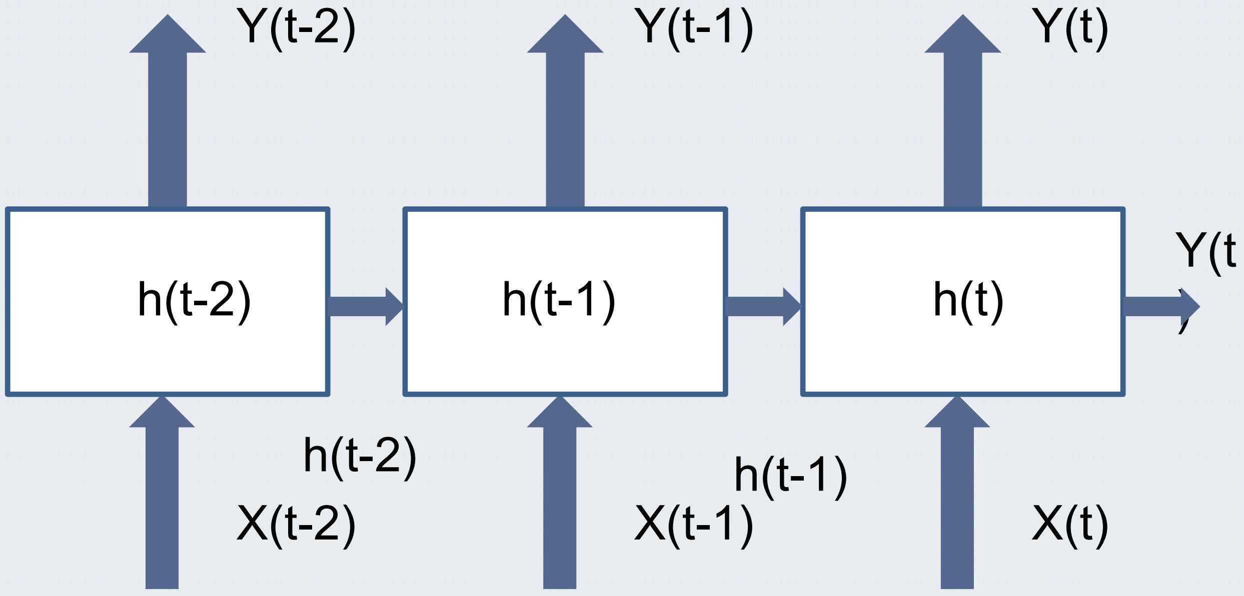
$\text{прямой} \quad \text{проход}:$

$y(n), h(h), (x(t), h(t), y(t), t = 1, n)$

$\text{обратный} \quad \text{проход} \quad \text{ищем}:$

$$\frac{\partial L(n)}{\partial W_{xh}}, \frac{\partial L(n)}{\partial W_{hh}}, \frac{\partial L(n)}{\partial W_{hy}}, \boxed{\frac{\partial L(n)}{\partial b_h}}, \frac{\partial L(n)}{\partial b_y}$$

Обратное Распространение Во Времени



$$\frac{\partial y(t)}{\partial h(t)} = \frac{\partial y(t)}{\partial h(t+1)} \frac{\partial h(t+1)}{\partial h(t)} = \frac{\partial y(t)}{\partial h(t+1)} (1 - h(t)^2) W_{hh}$$

$$\frac{\partial y(n)}{\partial h(n)} = W_{hy}$$

$$\frac{\partial L(n)}{\partial y(n)}, \qquad L(n) = -\ln(Pc), \quad Pc = soft\max(y(n))$$

$$\frac{\partial L(n)}{\partial y(n)} = \begin{cases} Pc, y\_true \neq C \\ Pc - 1, y\_true \neq C \end{cases}$$

прямой проход:

$$y(n), h(h), (x(t), h(t), y(t), t = 1, n)$$

обратный проход ищем:

$$\boxed{\frac{\partial L(n)}{\partial W_{xh}}, \frac{\partial L(n)}{\partial W_{hh}}, \frac{\partial L(n)}{\partial W_{hy}}, \frac{\partial L(n)}{\partial b_h}, \frac{\partial L(n)}{\partial b_y}}$$

Обратное Распространение Во Времени

Повторем с каждым примером

1. Установите входные данные
2. Установите начальное скрытое состояние (вектор нулей).
3. Сделайте прямой проход
4. Оцените потери
5. Определите производную потерь на выходе
6. Продвигайте ошибку вглубь сети
7. Накапливайте поправку для параметров сети
8. Пришли в первый узел -> делаем поправку параметров сети

Обратное Распространение Во Времени

Упрощения:

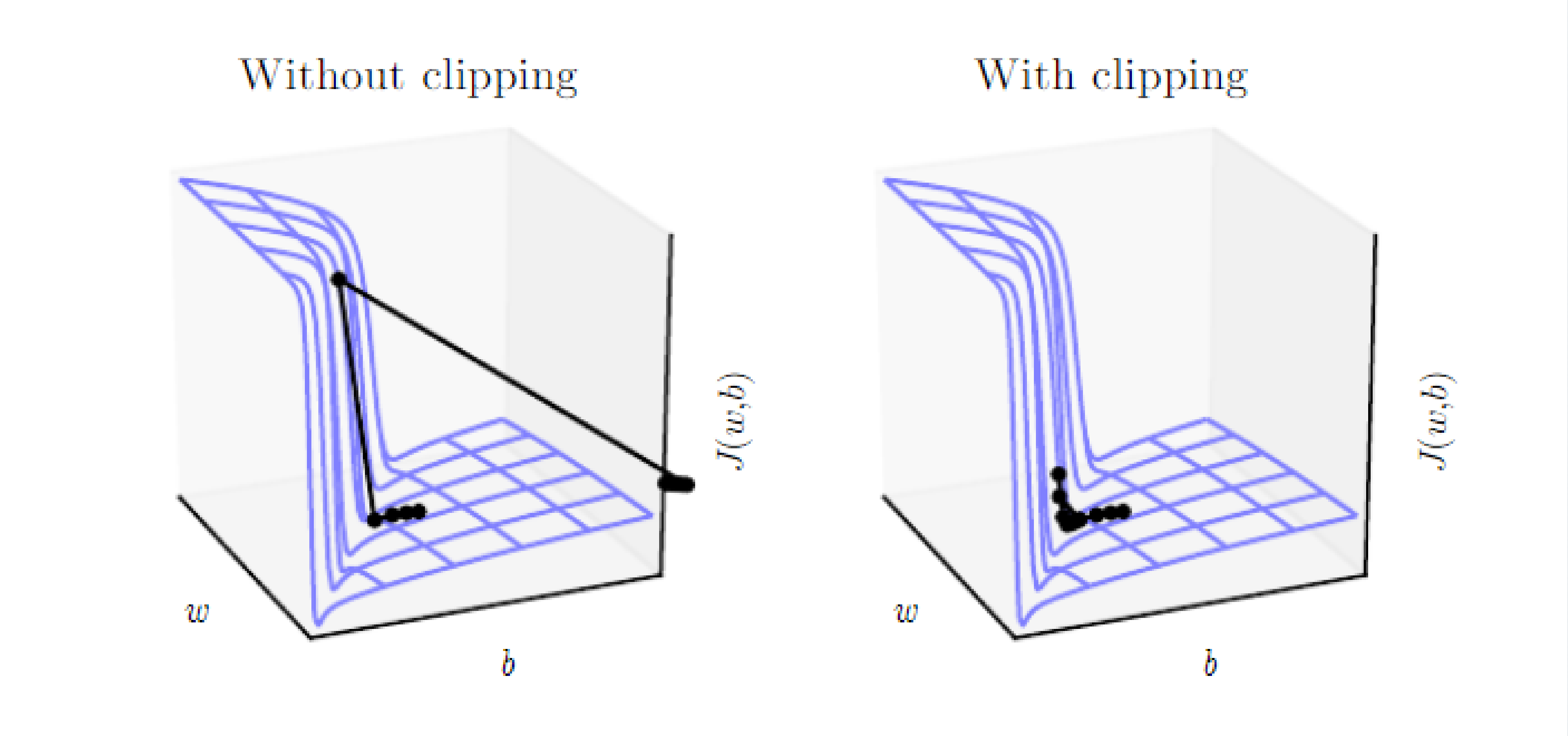
1. Объединяем $$\frac{\partial L}{\partial y}\frac{\partial y}{\partial h}=\frac{\partial L}{\partial h}$$

2. Закончив с обратным распространением во времени BPTT, используем обрезку на значениях градиента ниже -1 или выше 1. Это поможет избавиться от проблемы со **взрывными градиентами**.

3. Когда все градиенты подсчитаны, обновляем параметры веса и смещения, используя градиентный спуск

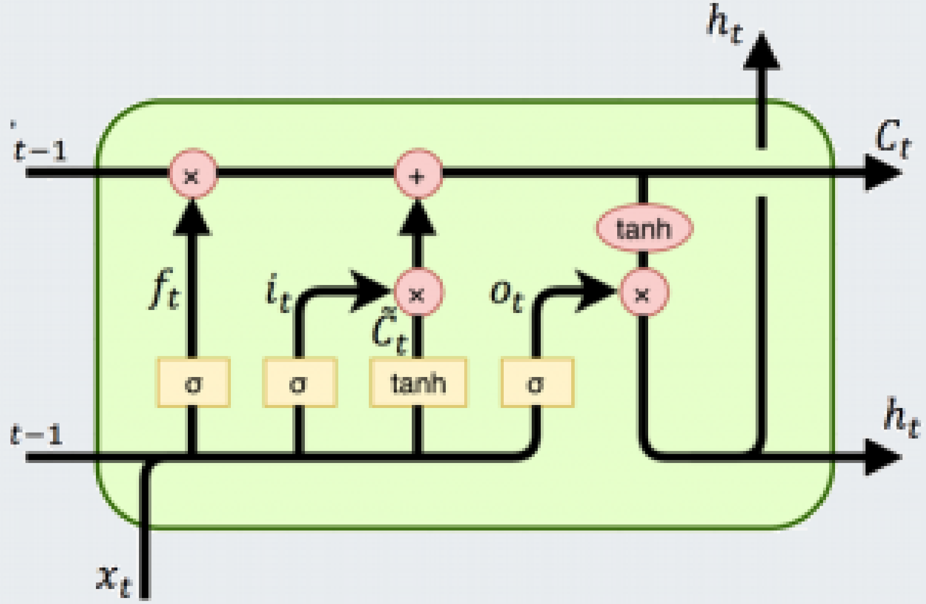
Градиентный взрыв

https://www.deeplearningbook.org/contents/rnn.html



Модель узла

LSTM



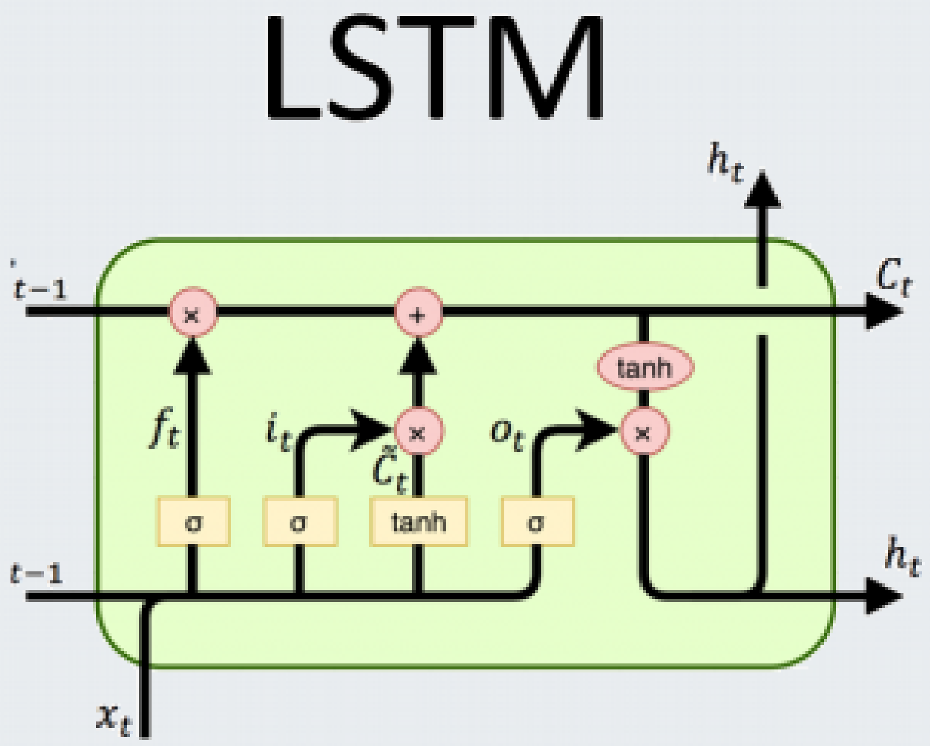
$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$

Переменные:

- x_t — входной вектор,
- h_t — выходной вектор,
- c_t — вектор состояний,
- W, U и b — матрицы параметров и вектор,
- f_t, i_t и o_t — векторы вентилей,
  - f_t — вектор вентиля забывания, вес запоминания старой информации,
  - i_t — вектор входного вентиля, вес получения новой информации,
  - o_t — вектор выходного вентиля, кандидат на выход.

- σ_g: на основе сигмоиды.
- σ_c: на основе гиперболического тангенса.
- σ_h: на основе гиперболического тангенса

BPTT: LSTM



https://nicodjimenez.github.io/2014/08/08/lstm.html

$$X_{ct} = [X_t, h_{t-1}]$$

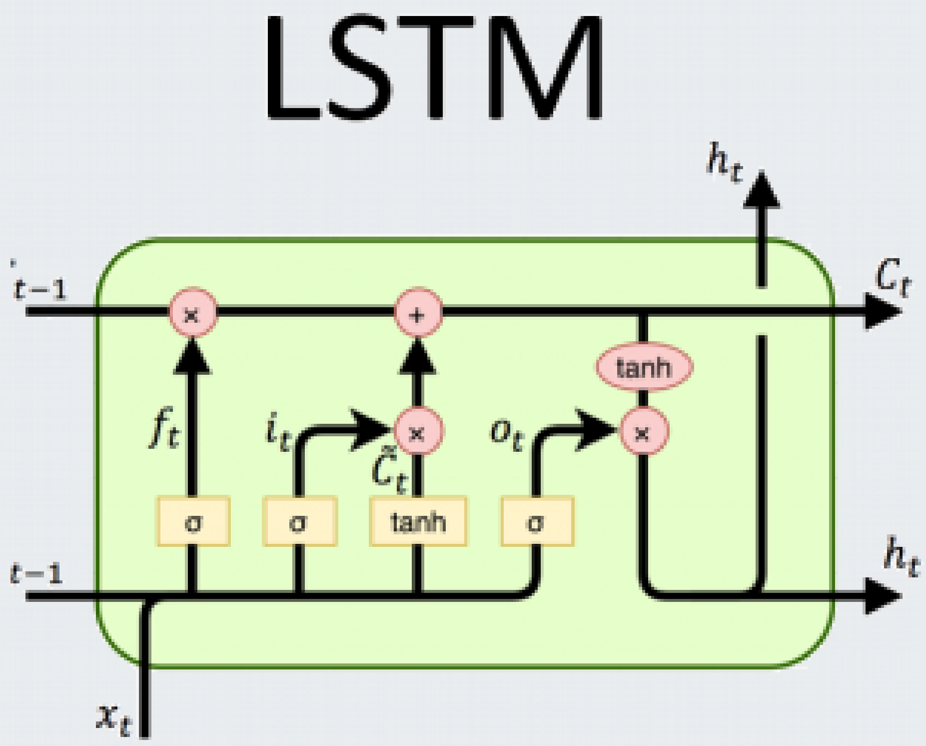
$$C_t = \phi(W_c X_{ct} + b_c)$$

$$i_t = \sigma(W_i X_{ct} + b_i)$$

$$f_t = \sigma(W_f X_{ct} + b_f)$$

$$o_t = \sigma(W_o X_{ct} + b_o)$$

BPTT: LSTM



https://nicodjimenez.github.io/2014/08/08/lstm.html

$$L_t = \|h_t - y_t\|^2, \qquad L = \sum_t L_t$$

$$\frac{\partial L}{\partial W} = \sum_t^T \sum_i^M \frac{\partial L}{\partial h_{it}} \frac{\partial h_{it}}{\partial W},$$

$M -$ *число ячеек*

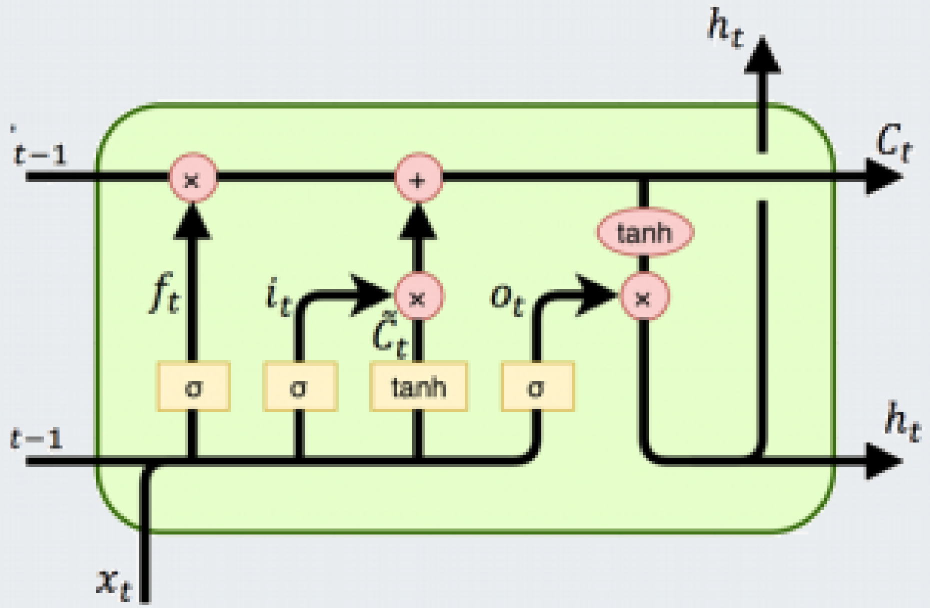
$T -$ *длина последовательности*

$$\frac{\partial L}{\partial h_{it}} = \sum_{s=t,T} \frac{\partial Ls}{\partial h_{it}}, \qquad Lt = \begin{cases} L_t + L_{t+1}, t < T \\ L_t, t = T \end{cases}$$

$$\frac{\partial L}{\partial h_t} = \frac{\partial L_t}{\partial h_t} + \frac{\partial L_{t+1}}{\partial h_t}, \qquad \frac{\partial L}{\partial h_T} = \frac{\partial L_T}{\partial h_t}$$

Модель узла

LSTM



«карусель константной ошибки»
(constant error carousel)

$$f_t = 1$$

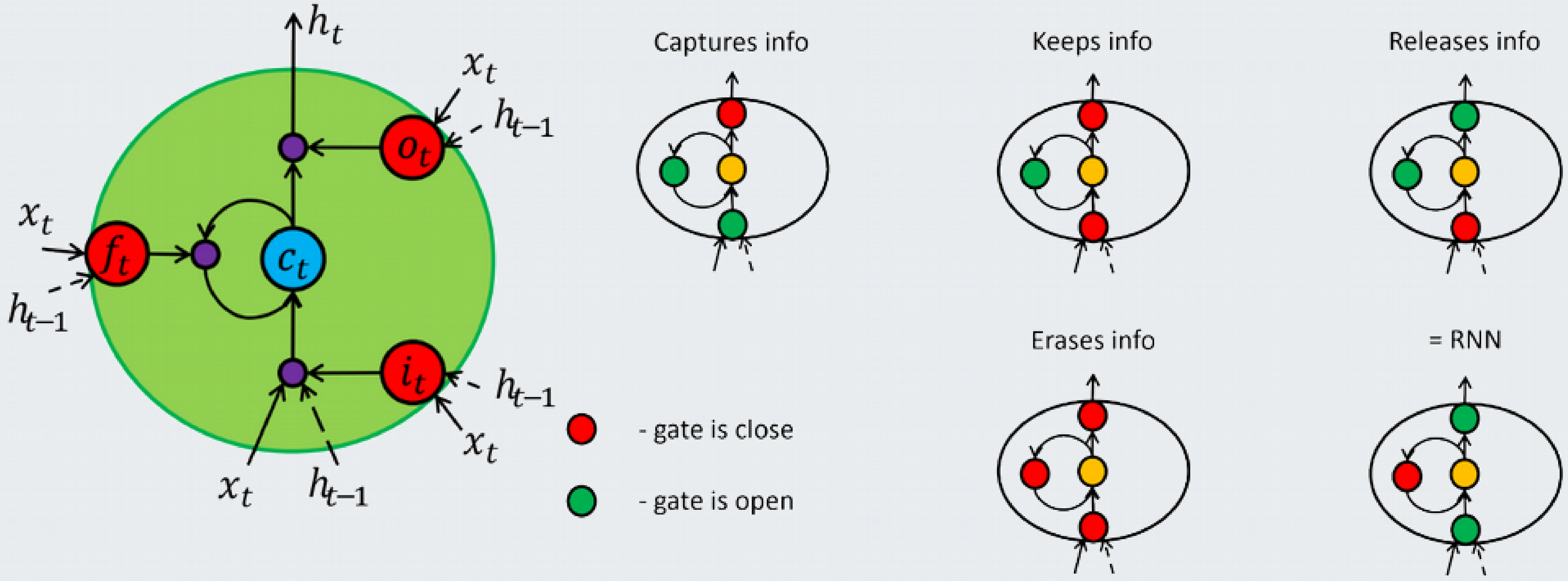
$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} X_t + W_{hc} h_t - 1 + b_c)$$

$$C_t = C_{t-1} + i_t \circ \tanh(W_{xc} X_t + W_{hc} h_t - 1 + b_c)$$

$$\frac{\partial C_t}{\partial C_{t-1}} = 1$$

$$b_f > 1$$

Модель узла: гейты

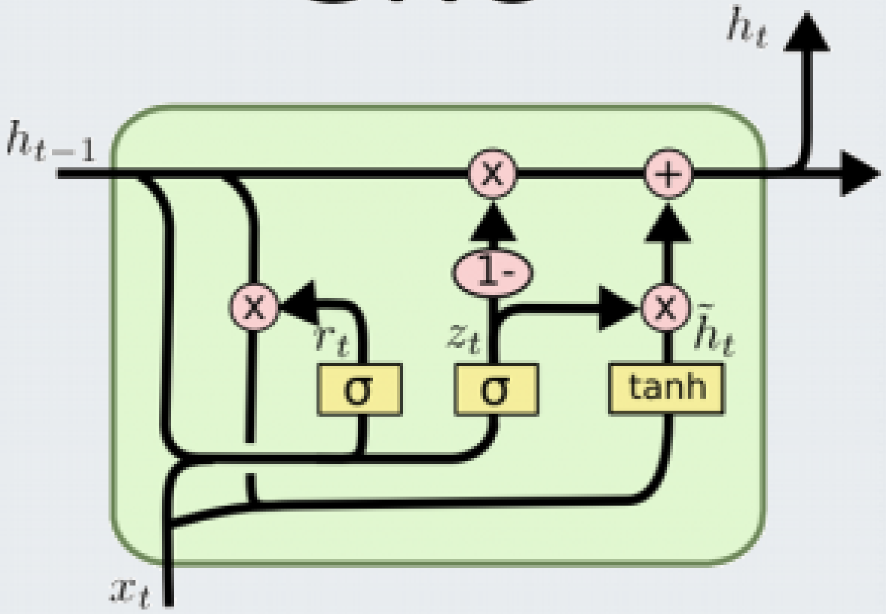


LSTM: Особенности

1. Много вариантов организации:
   1. «Замочные скважины» (11 матриц параметров)
   2. Дополнительные связи
   3. Без некоторых гейтов
   4. Двунаправленые
2. Некоторые простые архитектуры (без одного из гейтов!) не хуже «ванильноого» LSTM.
3. Градиент взрывается (но хотя бы не затухает) – делаем обоезку
4. Инициализация не всегда простая

Модель узла

GRU



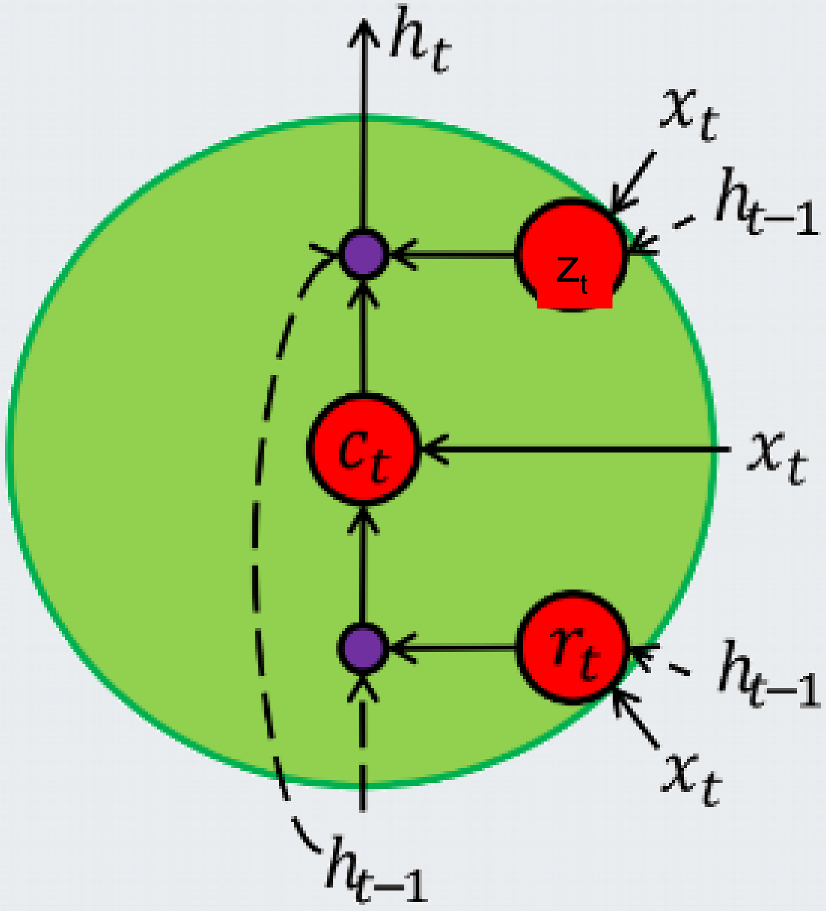
$\circ$ обозначает произведение Адамара. $h_0 = 0$.

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r)$$
$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \sigma_h(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h)$$

Переменные

- x_t: входной вектор
- h_t: выходной вектор
- z_t: вектор вентиля обновления
- r_t: вектор вентиля сброса
- W, U и b: матрицы параметров и вектор

Модель узла: Гейты



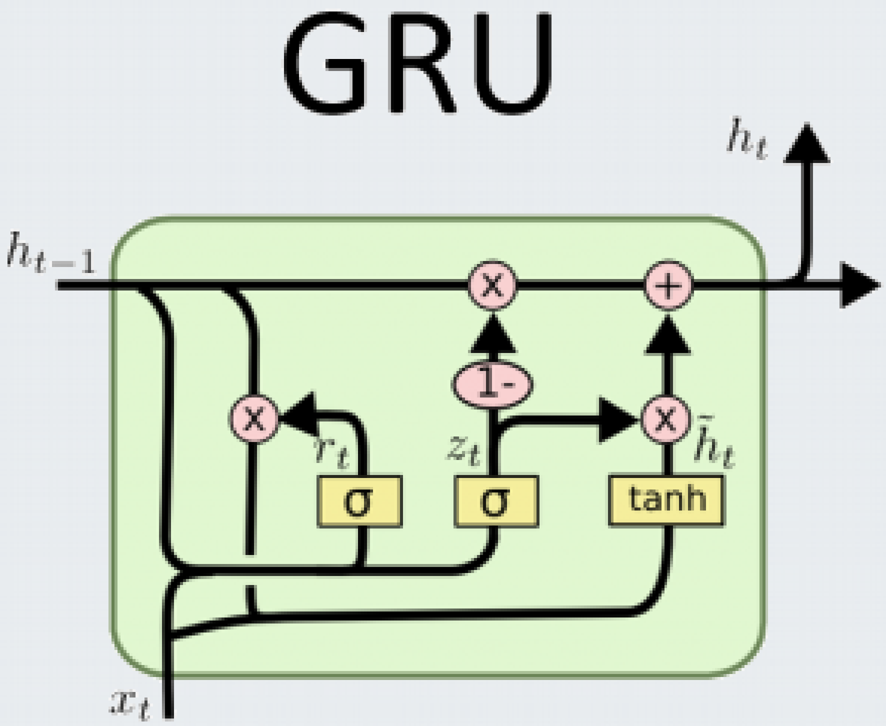
Протекание градиента: Карусель

$$c_t = \sigma(\mathrm{W_h x_t} + \mathrm{U_h}(\mathrm{h_{t\text{-}1}} \circ \mathrm{r_t}) + \mathrm{b_h})$$

$$h_t = (1 - z_t) \circ c_t + z_t \circ h_{t-1}$$

$$\frac{\partial h_{t+1}}{\partial h_t} = z_{t+1} \circ \frac{\partial c_{t+1}}{\partial h_t} + z_{t+1}, \quad b_z > 1$$

GRU: Особенности



1. Меньше матриц (6)
2. Меньше весов
3. Чуть хуже LSTM
4. Внутренний слой для многослойной модели.

RNN: Особенности

1. RNN имеют довольно простую общую структуру: уровни LSTM или GRU.
2. Все они выдают последовательность выходов, кроме верхнего.
3. DropOut и Batchnorm между слоями, и осторожно на рекуррентных связях
4. Слоев немного.
5. Skip-layer connections, как ResNet.
6. Состояния остаются, если нужно (в Keras : stateful=True)

RNN: Приложения

1. Временные ряды
2. Языковые модели
3. Обработка видеопотока
4. Модели с вниманием