Машинное обучение

Intro: что нас ждет?

Власов Кирилл Вячеславович

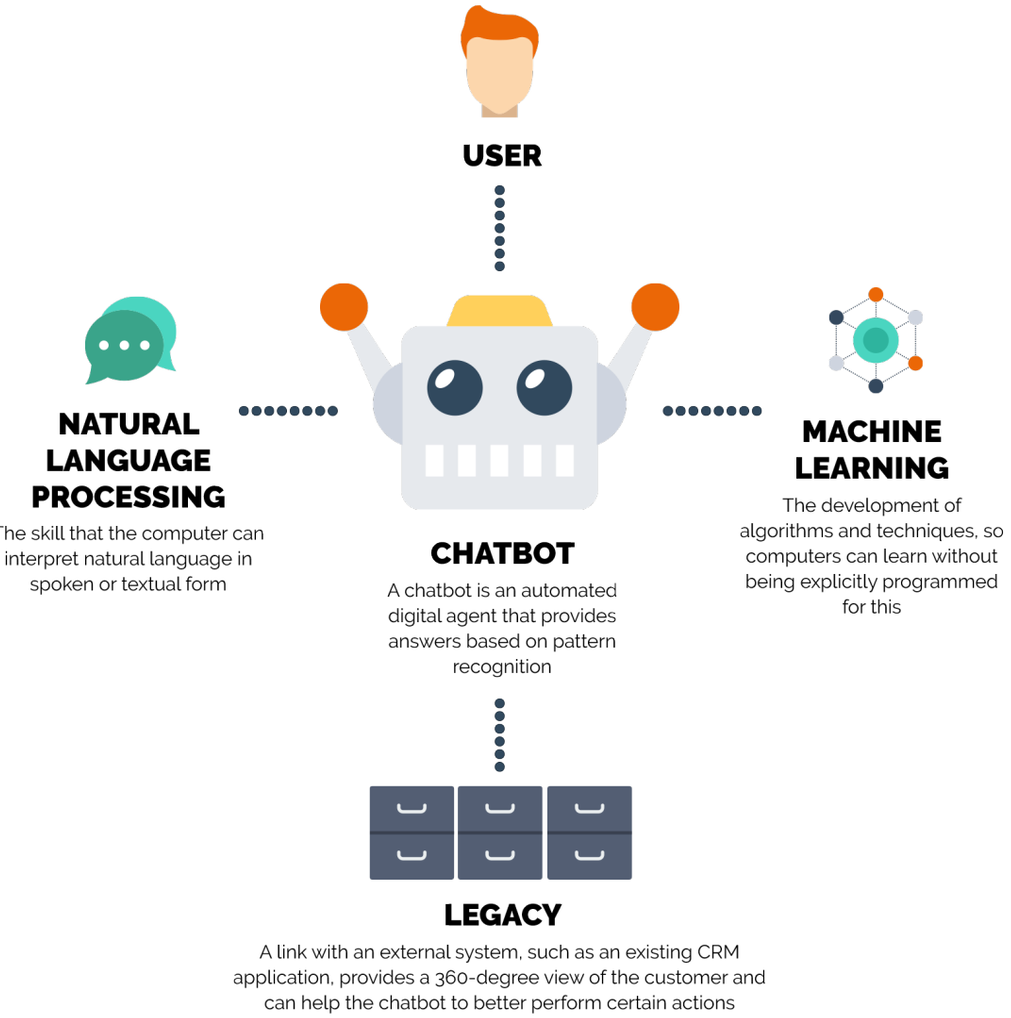
2019


WE NEED TO GO DEEPER

ФИВТ

https://github.com/ml-mipt/ml-mipt

Intro: что нас ждет?



Natural Language Processing

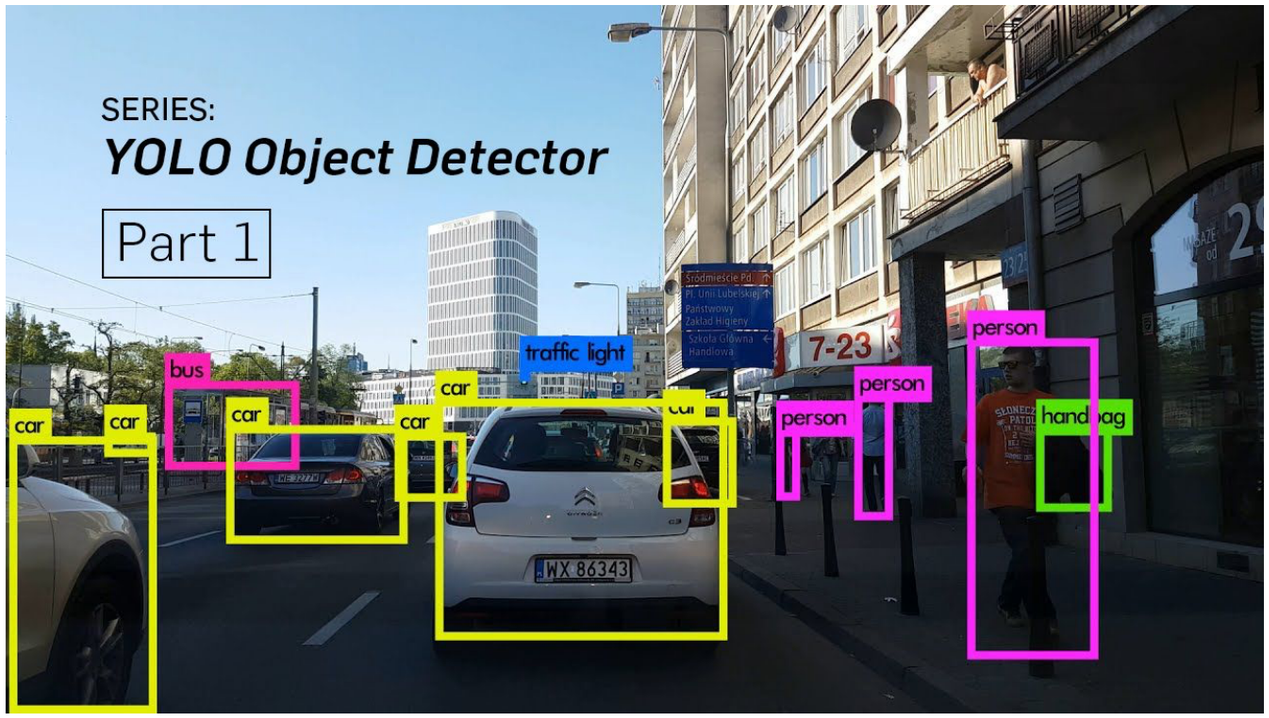
Attention
Encoder-decoder architecture.
BERT
GLOVE

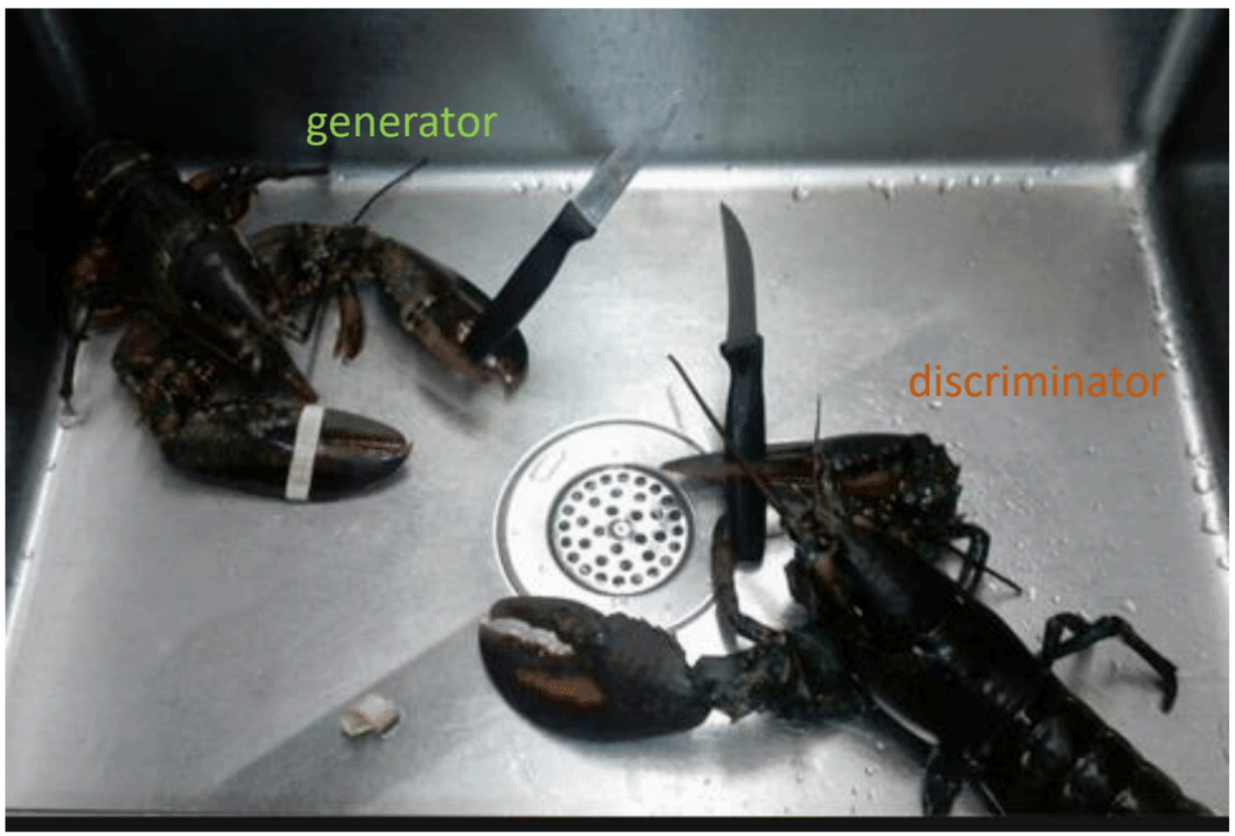
Intro: что нас ждет?

Reinforcement learning

Q-learning
Policy gradient methods
Sequence Learning

Intro: что нас ждет?

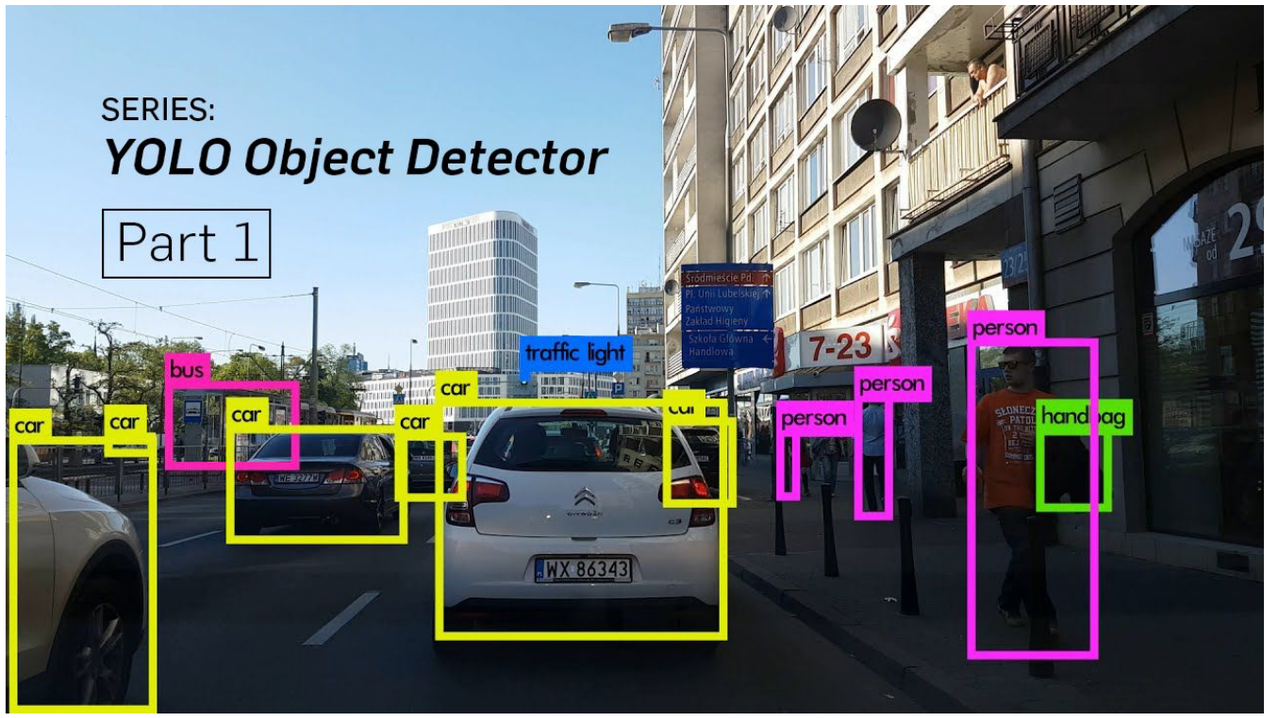




Deep CV

Object Detection (YOLO)
Segmentation (U-net)
GANs/VAE

Intro: что нас ждет?



Deep CV

Object Detection (YOLO)
Segmentation (U-net)
GANs/VAE

ML: Unsupervised Learning?


Why is it, when something happens, It is always you three?


Loss=NaN
Vanishing gradients
Exploding gradients


ME, TRYING TO RUN
SOME SIMPLE NEURAL NETWORK
PAIN
CUDA ERROR
IMPORT FAILED
PROBLEMS
STRESS
PAIN
10^10-LINES-LONG
TRACEBACK
TENSORFLOW ERROR
GOOGLE COLAB
imgflip.com

Intro: что нас ждет?

Quizes

Homeworks

InClass competitions


Каждый пятый подросток хотя бы раз
пробовал участвовать на Kaggle

Intro: что нас ждет?

Quizes

Homeworks

~~InClass competitions~~ → kaggle

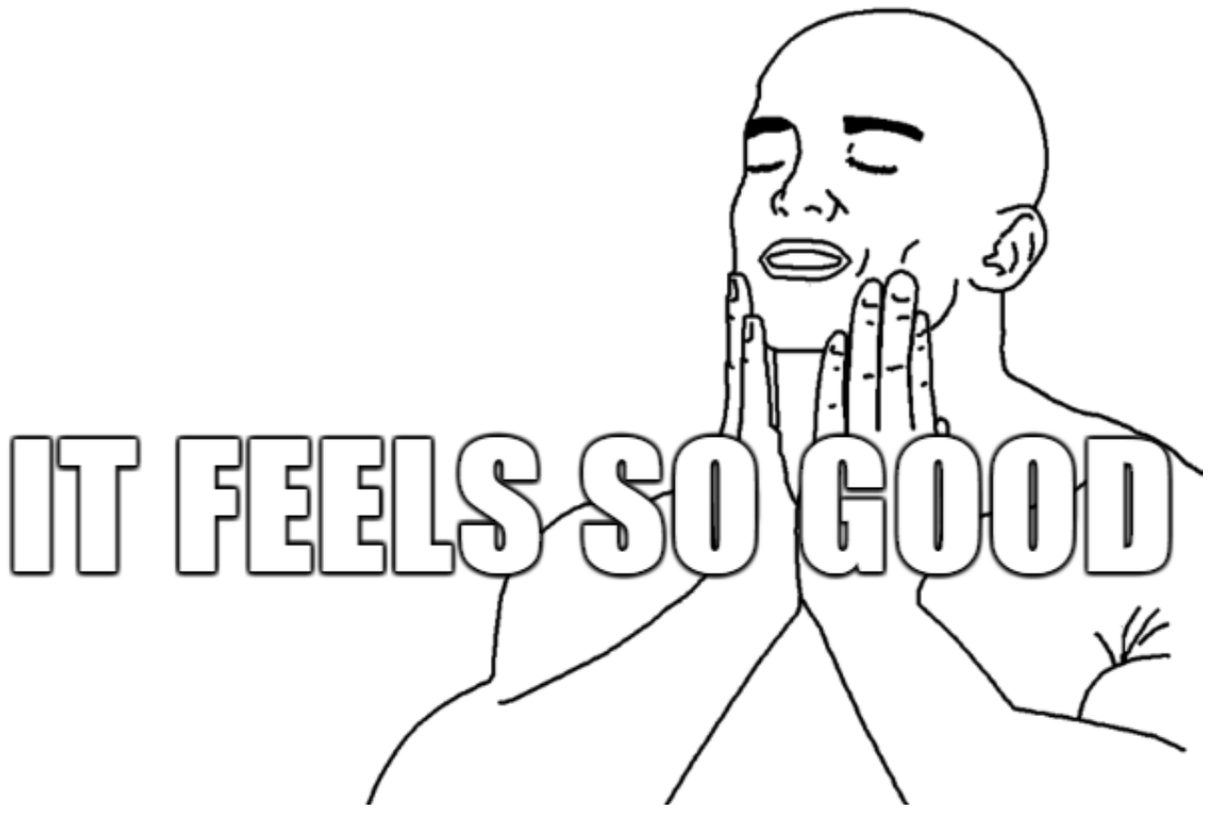
Intro: что нас ждет?

Quizes

Homeworks

~~InClass competitions~~ → kaggle

+ Individual project



ML engineers at the beginning of the project vs. **ML engineers at the end of the project**

Intro: что нас ждет?

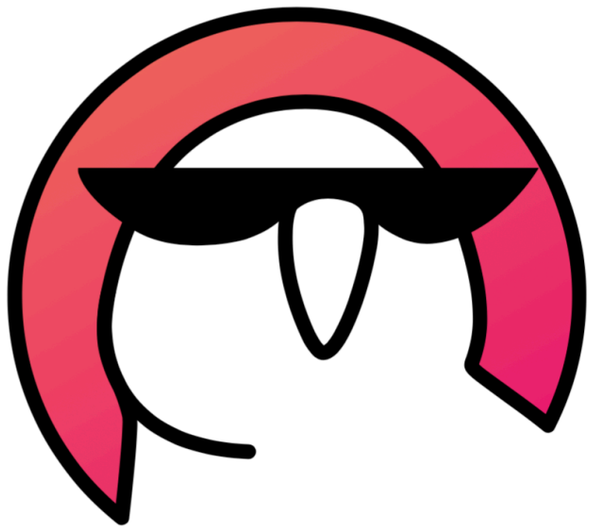
Quizes

Homeworks

~~InClass competitions~~ → kaggle

+ Individual project

- 💡 Идея
- 👨‍🎓 Формализация задачи
- 📦 Поиск данных
- 🔬 Исследование и выбор решения
- 🛠️ Реализация
- 🎉 Внедрение (API, web-service, bots, whatever)



Open Data Science


НЕ ПРОСТО КАЙФ
МТС
А КАЙФИЩЕЕЕЕ
risovach.ru

Машинное обучение

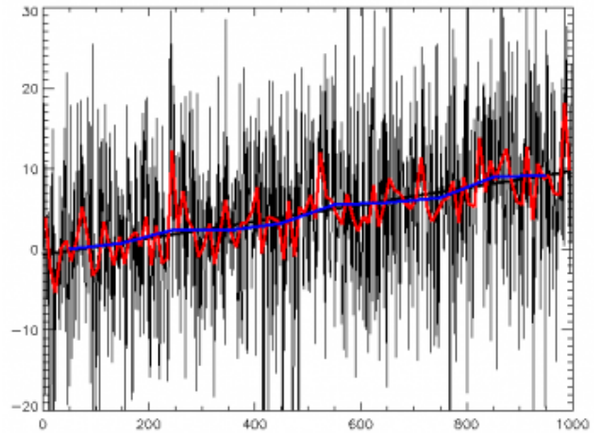
Лекция 9

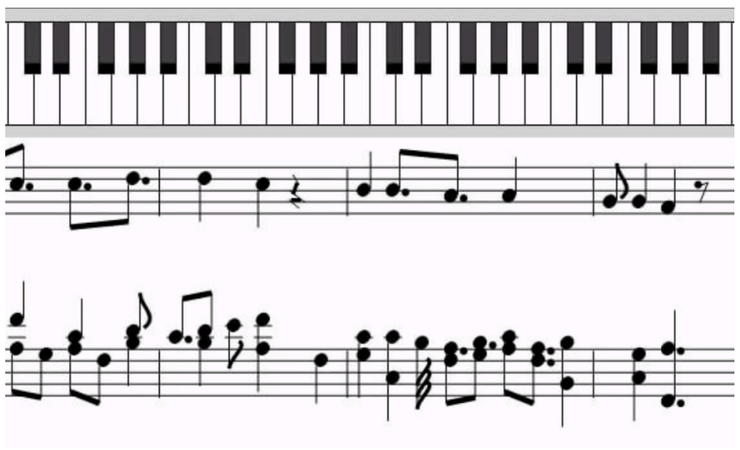
Recurrent Neural Networks

Власов Кирилл Вячеславович

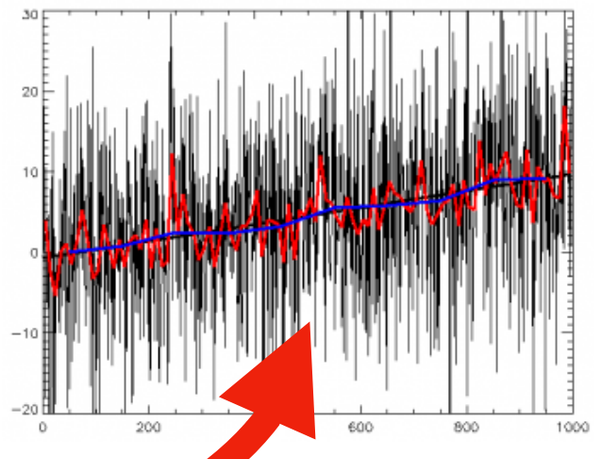
2019

Мотивация обработки последовательностей

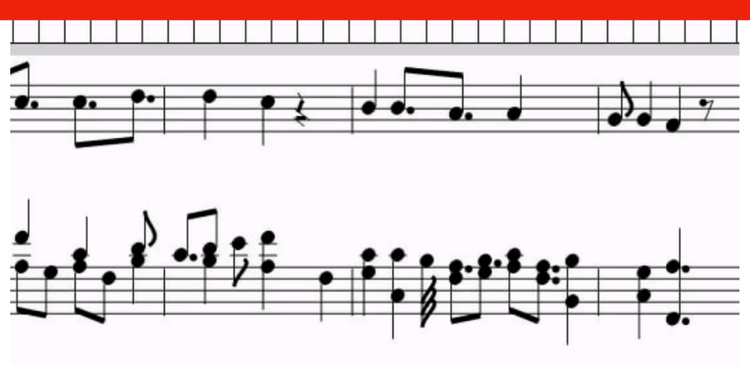




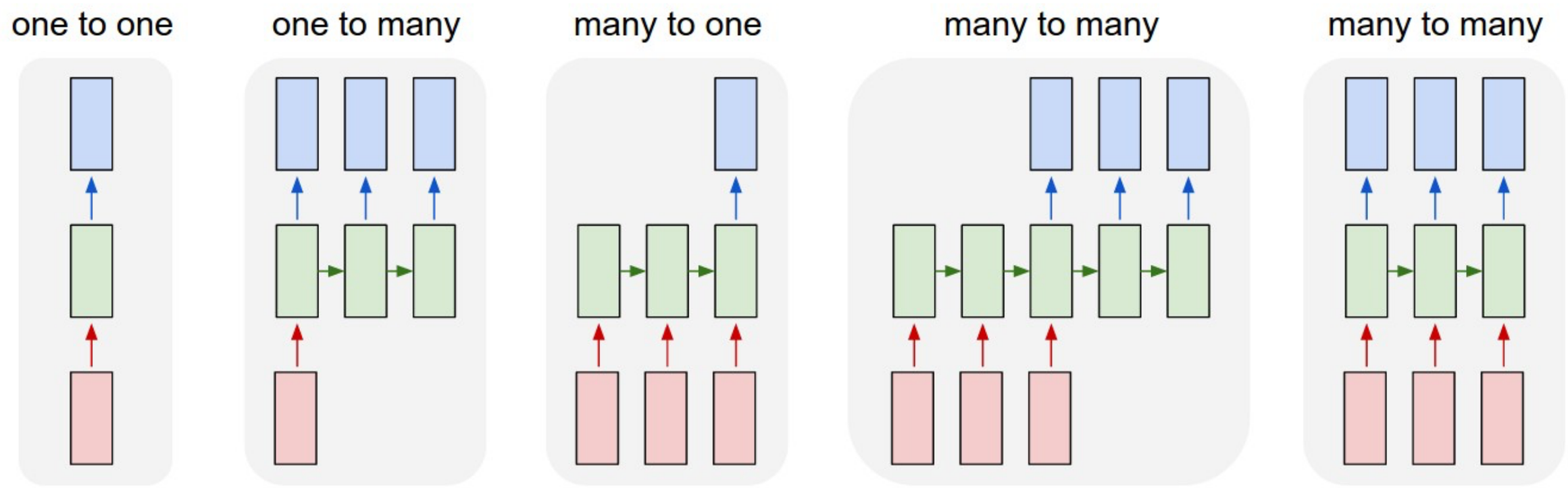
Мотивация обработки последовательностей



stats: ARIMA/SARIMA/etc.

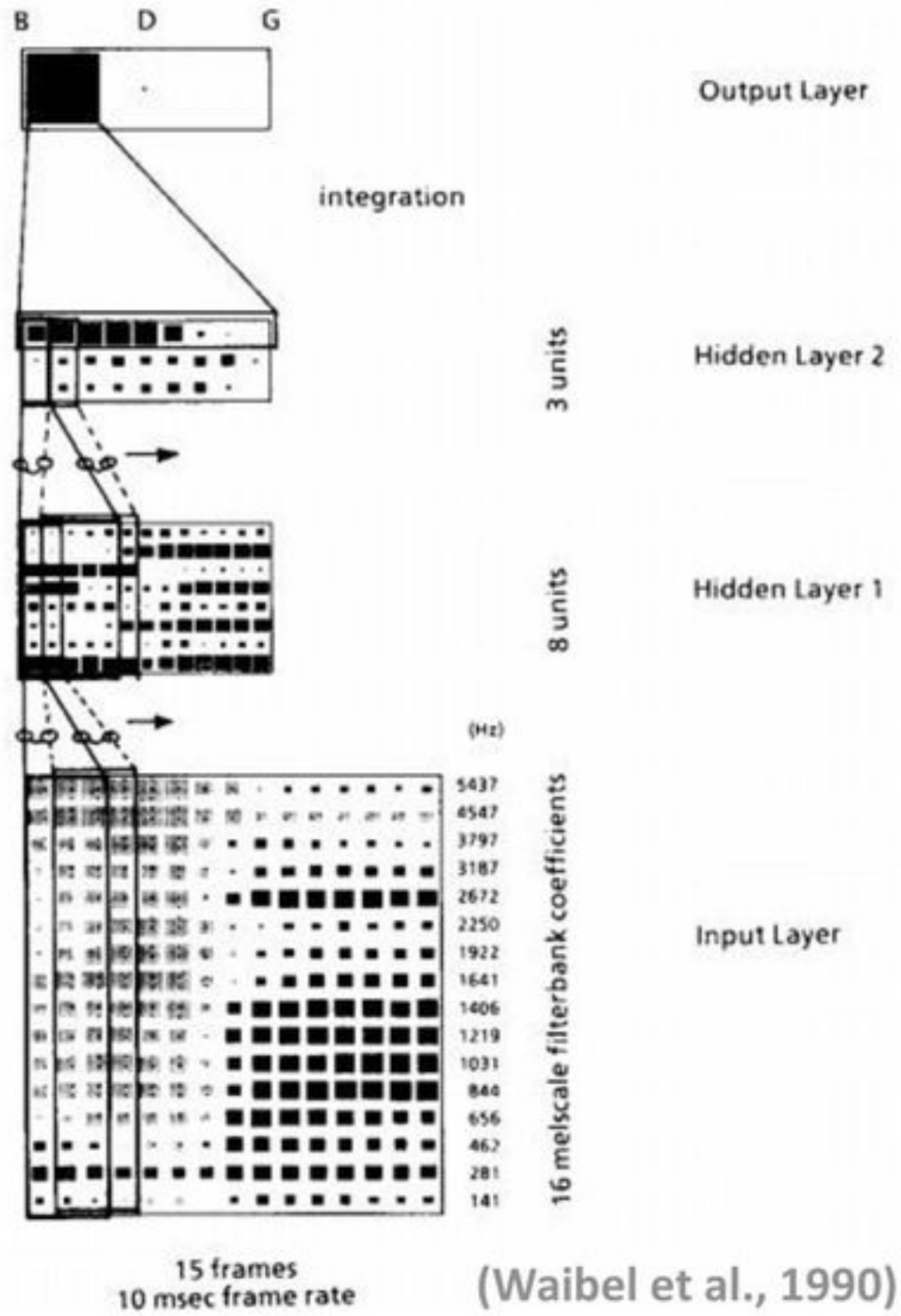


Мотивация обработки последовательностей



Andrej Karpathy: The Unreasonable Effectiveness of Recurrent Neural Networks

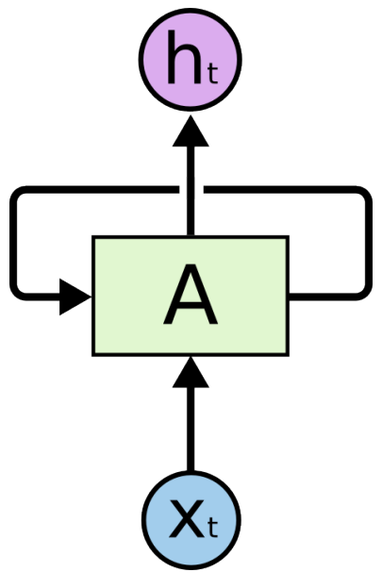
Time delay neural network



Waibel, Hinton, 1989

Phoneme Recognition Using Time-Delay Neural Networks

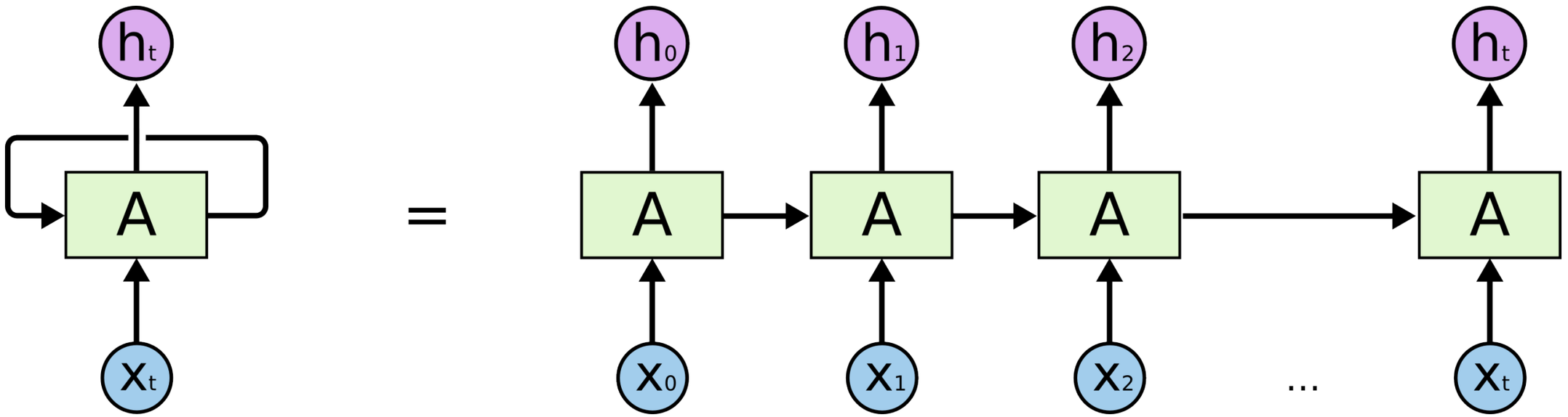
Рекуррентный нейрон



$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t),$$

Christopher Olah: Understanding LSTM Networks

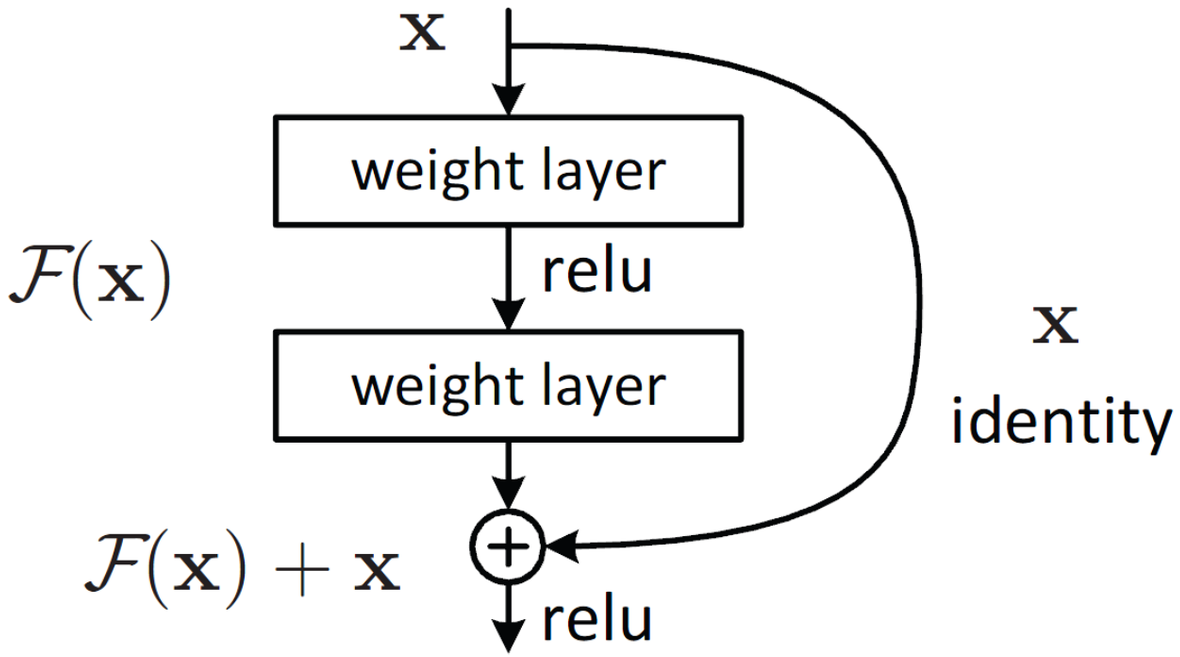
Рекуррентный нейрон



$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t),$$

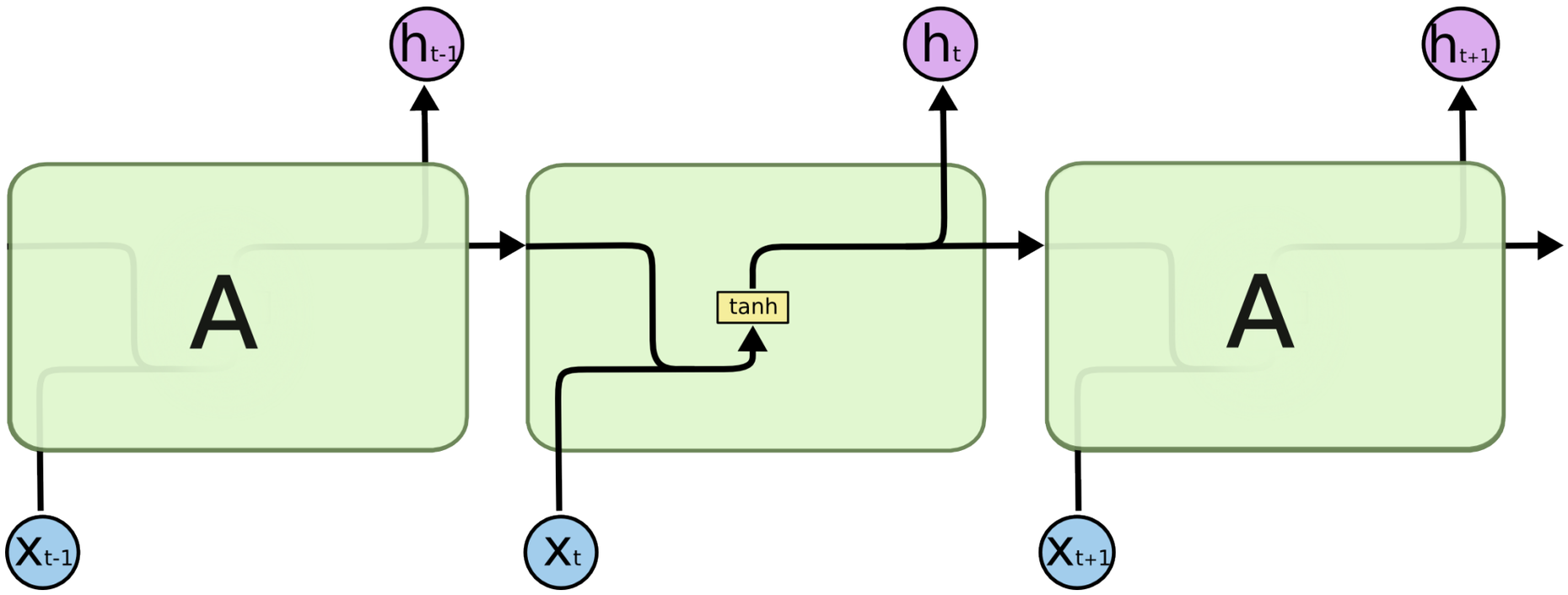
Christopher Olah: Understanding LSTM Networks

ResNet


x
weight layer
relu
$\mathcal{F}(\mathbf{x})$
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu


VGG-19
34-layer plain
34-layer residual

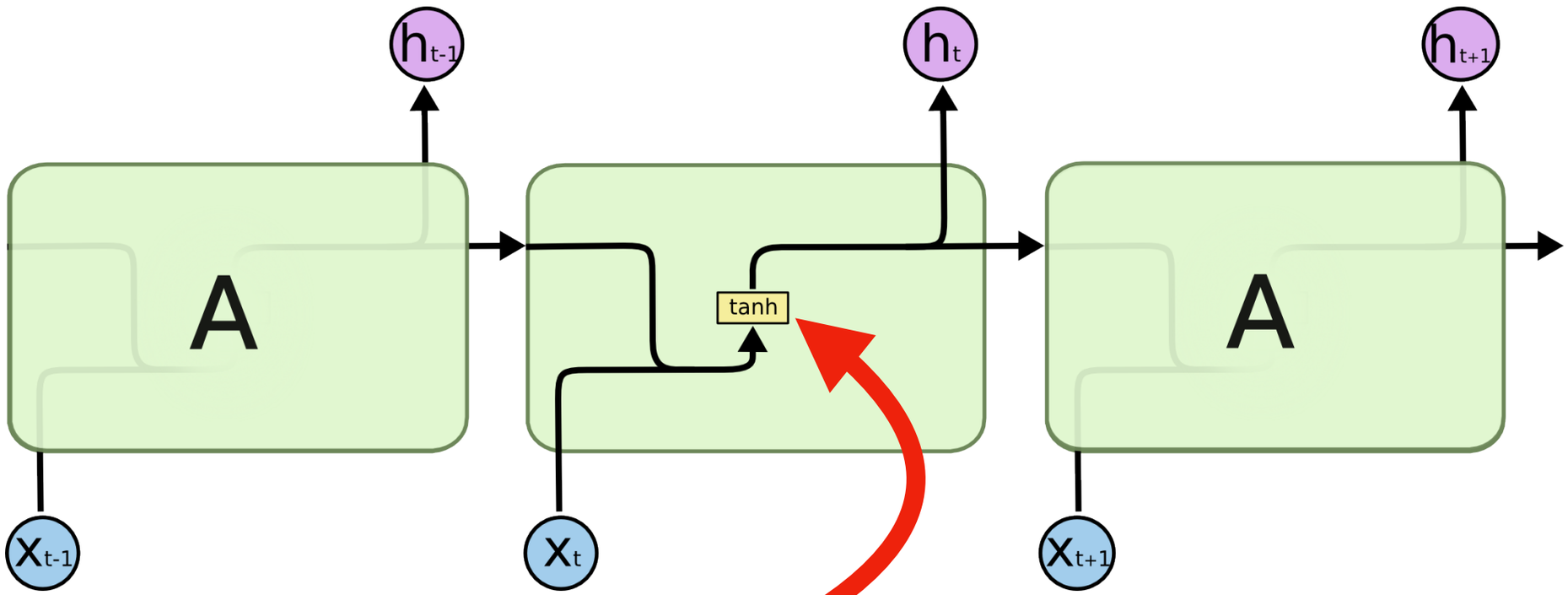
Vanila RNN



$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t),$$

Christopher Olah: Understanding LSTM Networks

Vanila RNN

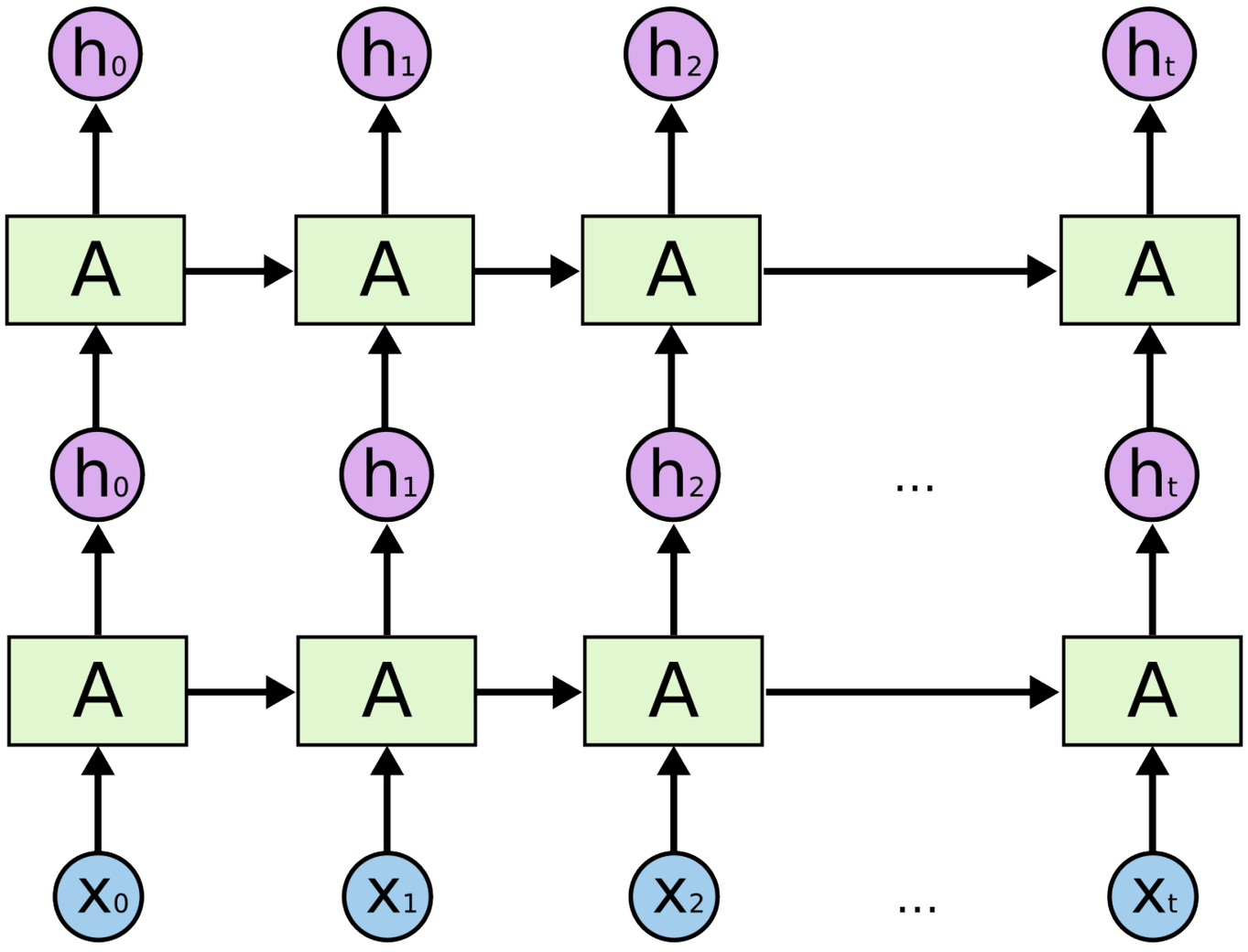


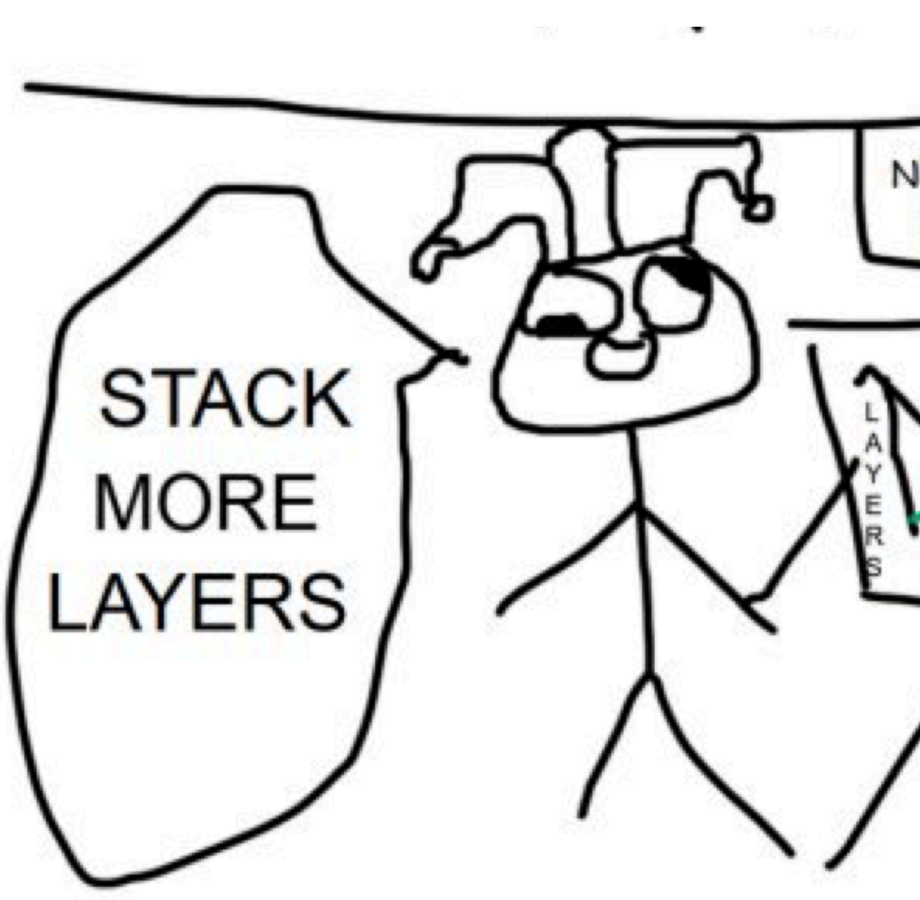
Почему tanh(), а не sigmoid() или ReLU?

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t),$$

Christopher Olah: Understanding LSTM Networks

Больше слоев





Примеры

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

Lemma 0.1. *Assume (3) and (3) by the construction in the description.*

Suppose $X = \lim |X|$ *(by the formal open covering* X *and a single map* $\underline{Proj}_X(\mathcal{A}) = \mathrm{Spec}(B)$ *over* U *compatible with the complex*

$$Set(\mathcal{A}) = \Gamma(X, \mathcal{O}_{X,\mathcal{O}_X}).$$

When in this case of to show that $\mathcal{Q} \to \mathcal{C}_{Z/X}$ *is stable under the following result in the second conditions of (1), and (3). This finishes the proof. By Definition ?? (without element is when the closed subschemes are catenary. If* T *is surjective we may assume that* T *is connected with residue fields of* S*. Moreover there exists a closed subspace* $Z \subset X$ *of* X *where* U *in* X' *is proper (some defining as a closed subset of the uniqueness it suffices to check the fact that the following theorem*

(1) f *is locally of finite type. Since* $S = \mathrm{Spec}(R)$ *and* $Y = \mathrm{Spec}(R)$.

Proof. This is form all sheaves of sheaves on X. But given a scheme U and a surjective étale morphism $U \to X$. Let $U \cap U = \coprod_{i=1,\ldots,n} U_i$ be the scheme X over S at the schemes $X_i \to X$ and $U = \lim_i X_i$. □

The following lemma surjective restrocomposes of this implies that $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} = \mathcal{F}_{\mathcal{X},\ldots,0}$.

Lemma 0.2. *Let* X *be a locally Noetherian scheme over* S*,* $E = \mathcal{F}_{X/S}$*. Set* $\mathcal{I} = \mathcal{J}_1 \subset \mathcal{I}'_n$*. Since* $\mathcal{I}^n \subset \mathcal{I}^n$ *are nonzero over* $i_0 \leq \mathfrak{p}$ *is a subset of* $\mathcal{J}_{n,0} \circ \overline{A}_2$ *works.*

Lemma 0.3. *In Situation ??. Hence we may assume* $\mathfrak{q}' = 0$.

Proof. We will use the property we see that $\mathfrak{p}$ is the mext functor (??). On the other hand, by Lemma ?? we see that

$$D(\mathcal{O}_{X'}) = \mathcal{O}_X(D)$$

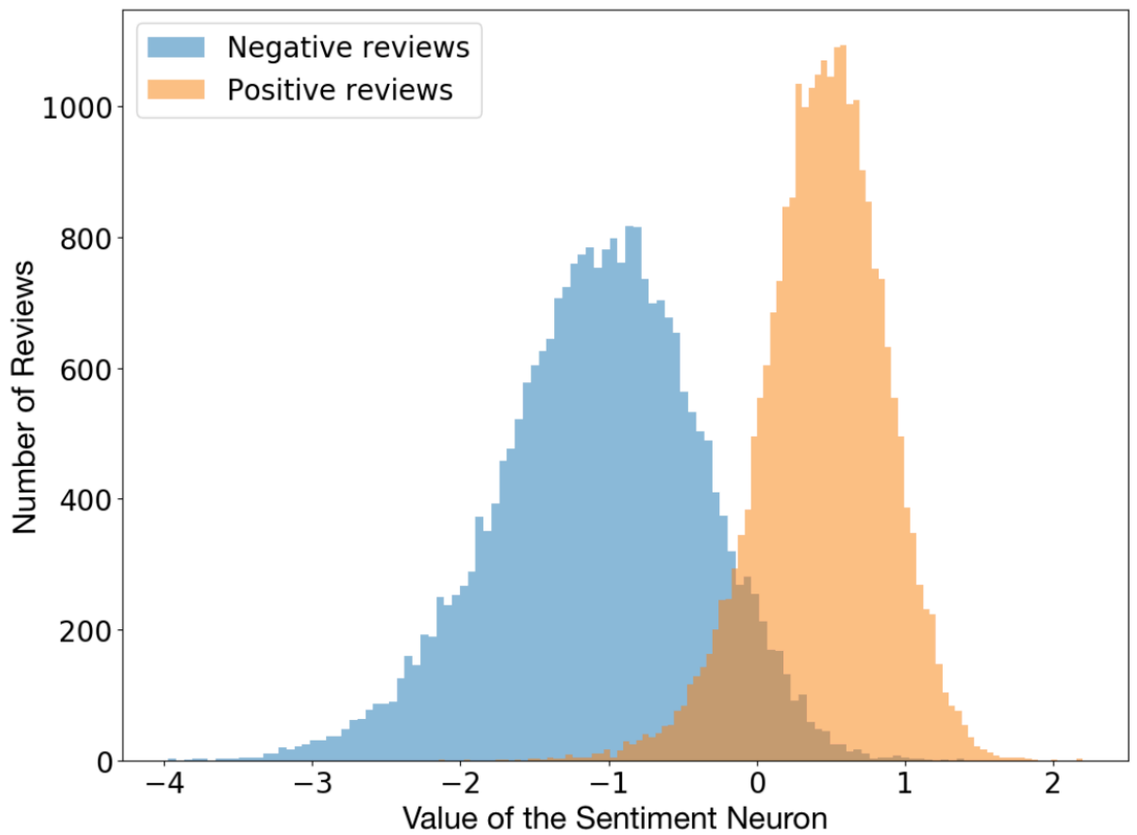
where K is an F-algebra where δ_{n+1} is a scheme over S. □

Andrej Karpathy: The Unreasonable Effectiveness of Recurrent Neural Networks

Примеры

SENTIMENT FIXED TO POSITIVE	SENTIMENT FIXED TO NEGATIVE
Just what I was looking for. Nice fitted pants, exactly matched seam to color contrast with other pants I own. Highly recommended and also very happy!	The package received was blank and has no barcode. A waste of time and money.
This product does what it is supposed to. I always keep three of these in my kitchen just in case ever I need a replacement cord.	Great little item. Hard to put on the crib without some kind of embellishment. My guess is just like the screw kind of attachment I had.
Best hammock ever! Stays in place and holds it's shape. Comfy (I love the deep neon pictures on it), and looks so cute.	They didn't fit either. Straight high sticks at the end. On par with other buds I have. Lesson learned to avoid.
Dixie is getting her Doolittle newsletter we'll see another new one coming out next year. Great stuff. And, here's the contents - information that we hardly know about or forget.	great product but no seller. couldn't ascertain a cause. Broken product. I am a prolific consumer of this company all the time.
I love this weapons look . Like I said beautiful !!! I recommend it to all. Would suggest this to many roleplayers, And I stronge to get them for every one I know. A must watch for any man who love Chess!	

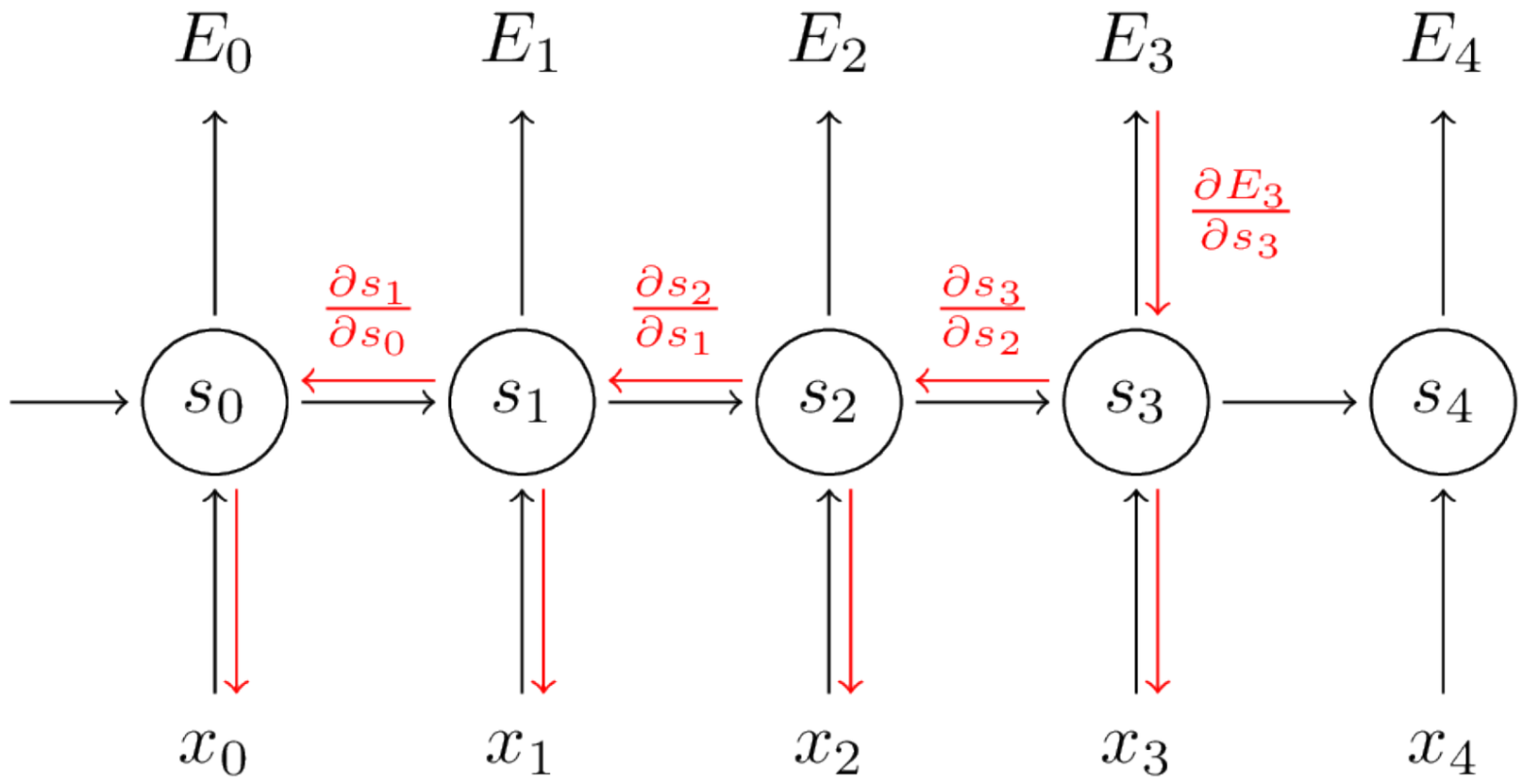
OpenAI: Unsupervised Sentiment Neuron



Обучение RNN

$$s_t = \tanh(Ux_t + Ws_{t-1})$$
$$\hat{y}_t = \mathrm{softmax}(Vs_t)$$

$$E_t(y_t, \hat{y}_t) = -y_t \log \hat{y}_t$$
$$E(y, \hat{y}) = \sum_t E_t(y_t, \hat{y}_t) = -\sum_t y_t \log \hat{y}_t$$



$$\frac{\partial E_3}{\partial V} = \frac{\partial E_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial V} = \frac{\partial E_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_3}\frac{\partial z_3}{\partial V} = (\hat{y}_3 - y_3) \otimes s_3$$

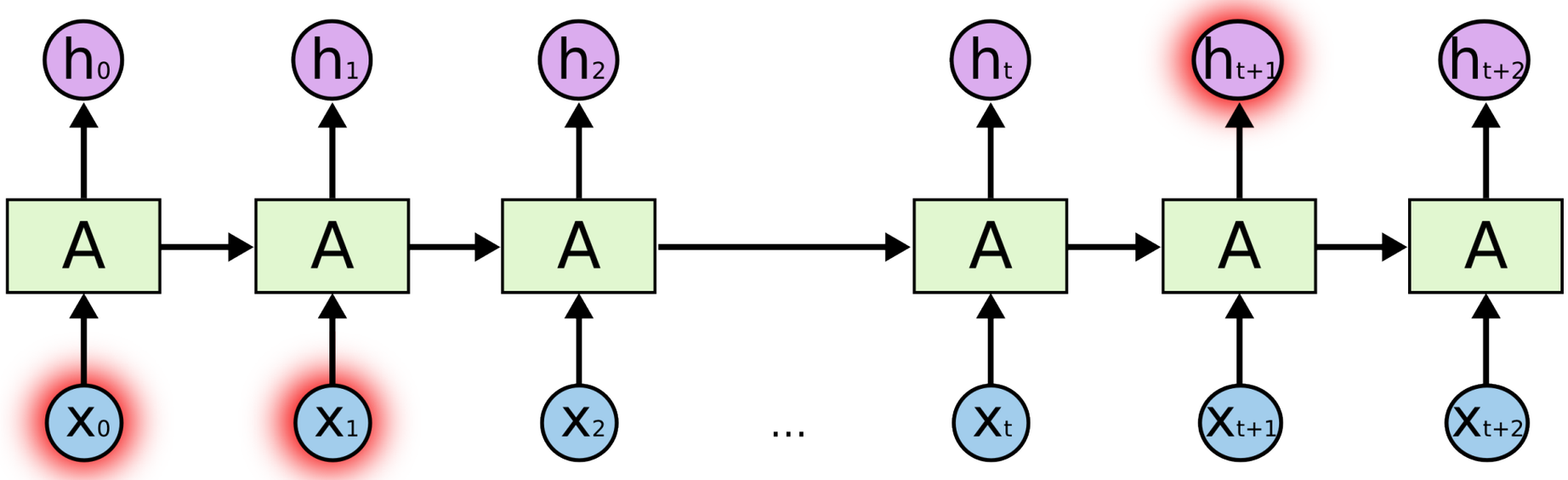
$$\frac{\partial E_3}{\partial W} = \frac{\partial E_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial s_3}\frac{\partial s_3}{\partial W}$$

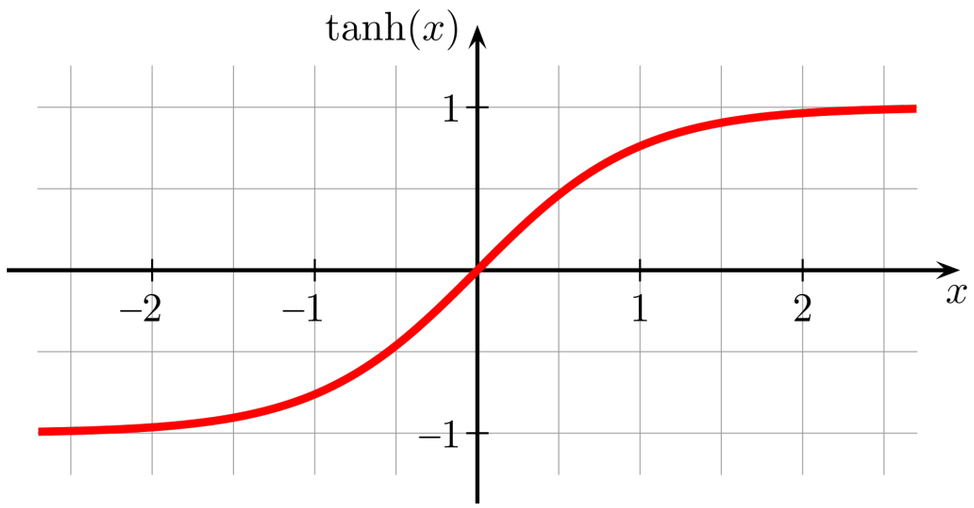
Ho: $s_3 \longleftarrow s_2 \longleftarrow s_1$

$$\frac{\partial E_3}{\partial W} = \sum_{k=0}^{3} \frac{\partial E_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial s_3}\frac{\partial s_3}{\partial s_k}\frac{\partial s_k}{\partial W} \Longrightarrow \frac{\partial E_3}{\partial W} = \sum_{k=0}^{3} \frac{\partial E_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial s_3}\left(\prod_{j=k+1}^{3} \frac{\partial s_j}{\partial s_{j-1}}\right)\frac{\partial s_k}{\partial W}$$

Denny Britz: Backpropagation Through Time and Vanishing Gradients

Проблема долгих зависимостей

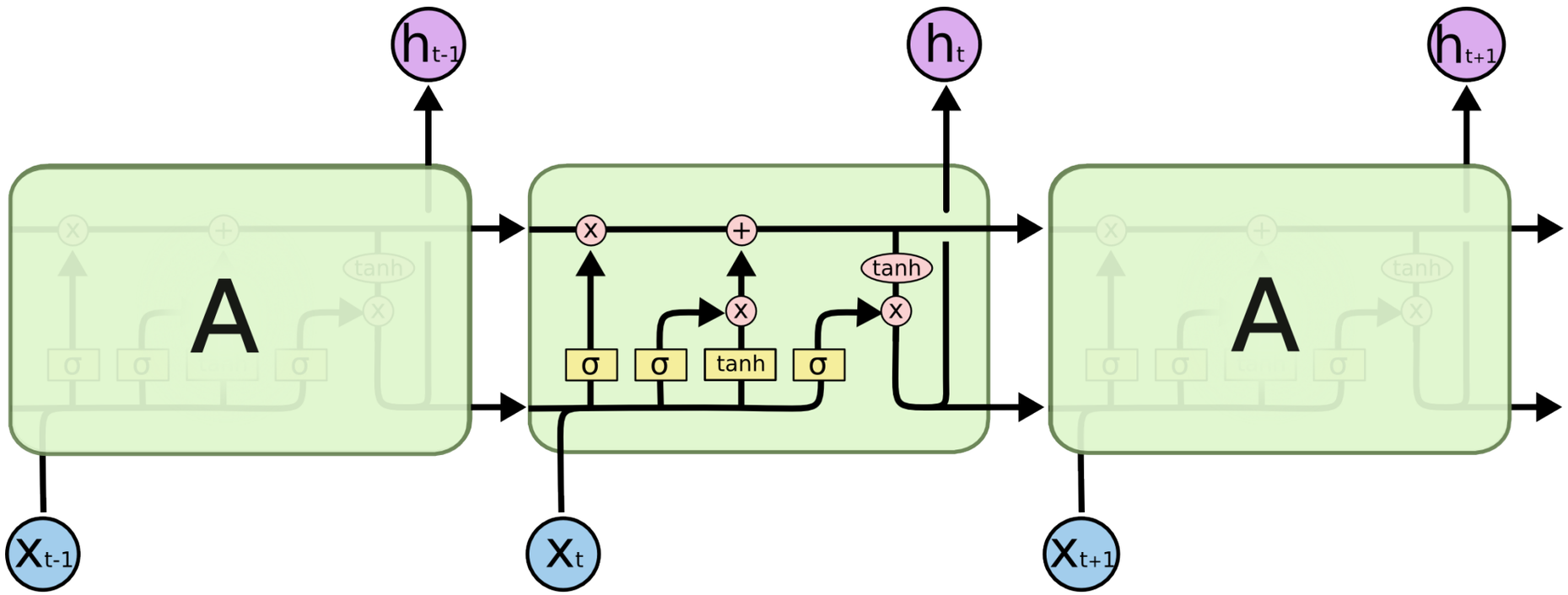




$$s_t = \tanh(Ux_t + Ws_{t-1})$$
$$\hat{y}_t = \mathrm{softmax}(Vs_t)$$

$$\frac{\partial E_3}{\partial W} = \sum_{k=0}^{3} \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial s_3} \left(\prod_{j=k+1}^{3} \frac{\partial s_j}{\partial s_{j-1}} \right) \frac{\partial s_k}{\partial W}$$

Long Short Term Memory networks

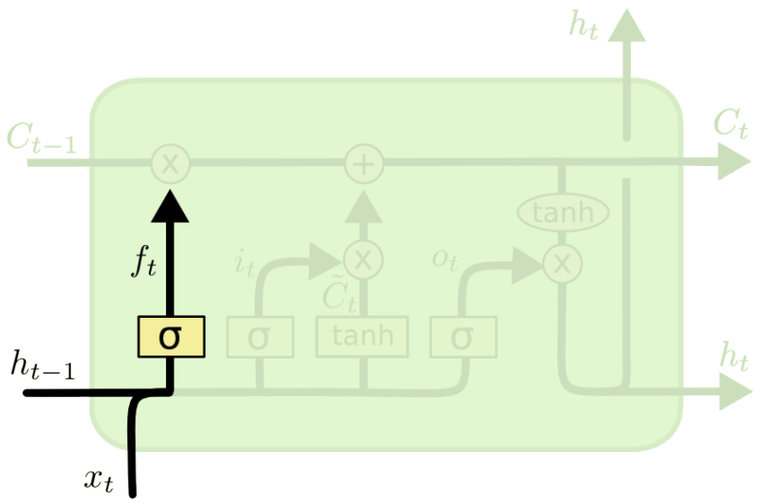


Hochreiter, Sepp & Schmidhuber: Long Short-term Memory

Christopher Olah: Understanding LSTM Networks

Long Short Term Memory networks

Forget gate

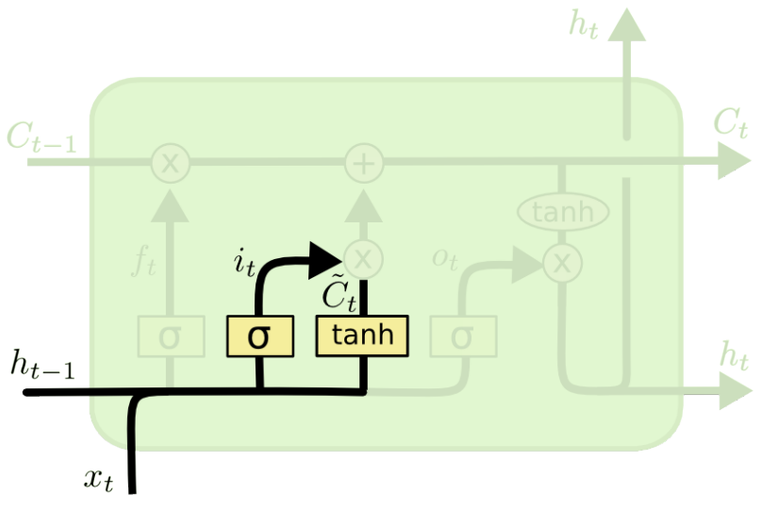


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Christopher Olah: Understanding LSTM Networks

Long Short Term Memory networks

Input gate



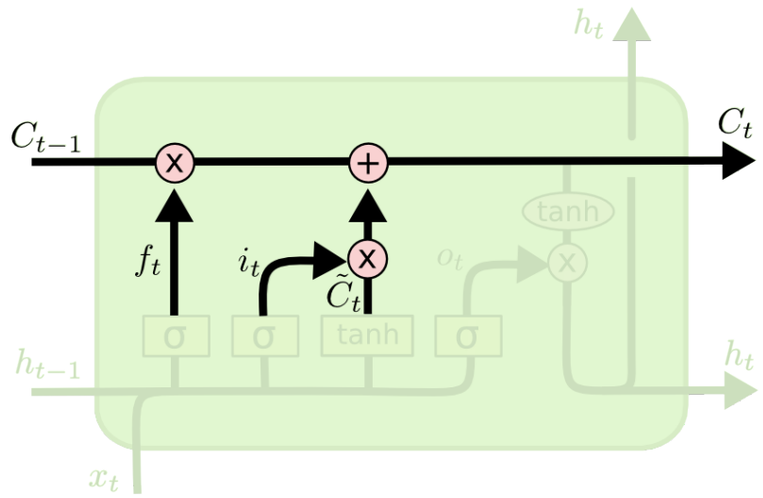
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Christopher Olah: Understanding LSTM Networks

Long Short Term Memory networks

Cell update

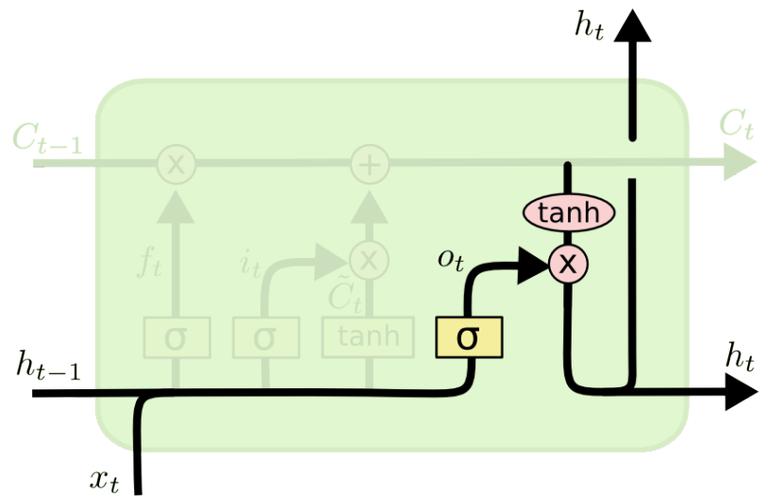


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Christopher Olah: Understanding LSTM Networks

Long Short Term Memory networks

Output gate



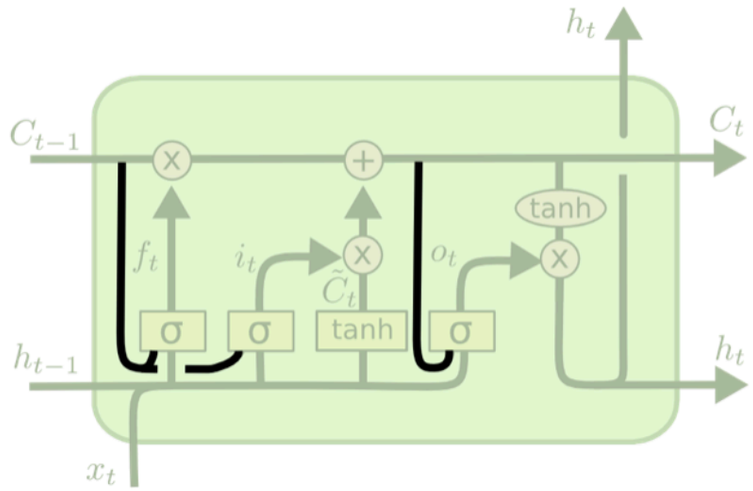
$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

Christopher Olah: Understanding LSTM Networks

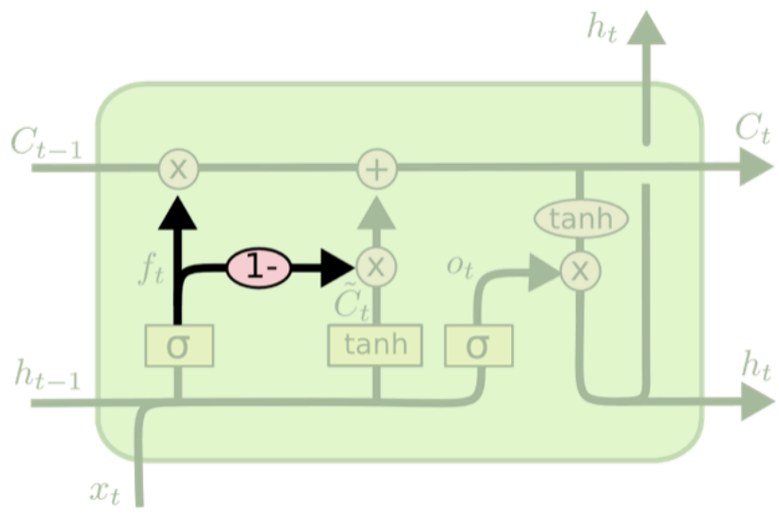
Варианты LSTM

Peephole connections



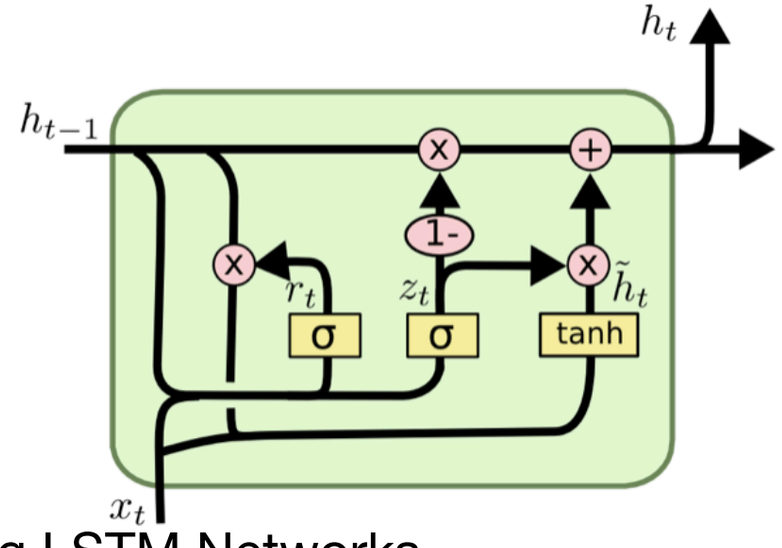
$$f_t = \sigma\left(W_f \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [\boldsymbol{C_t}, h_{t-1}, x_t] \; + \; b_o\right)$$

Coupled forget and input gates



$$C_t = f_t * C_{t-1} + (1 - \boldsymbol{f_t}) * \tilde{C}_t$$

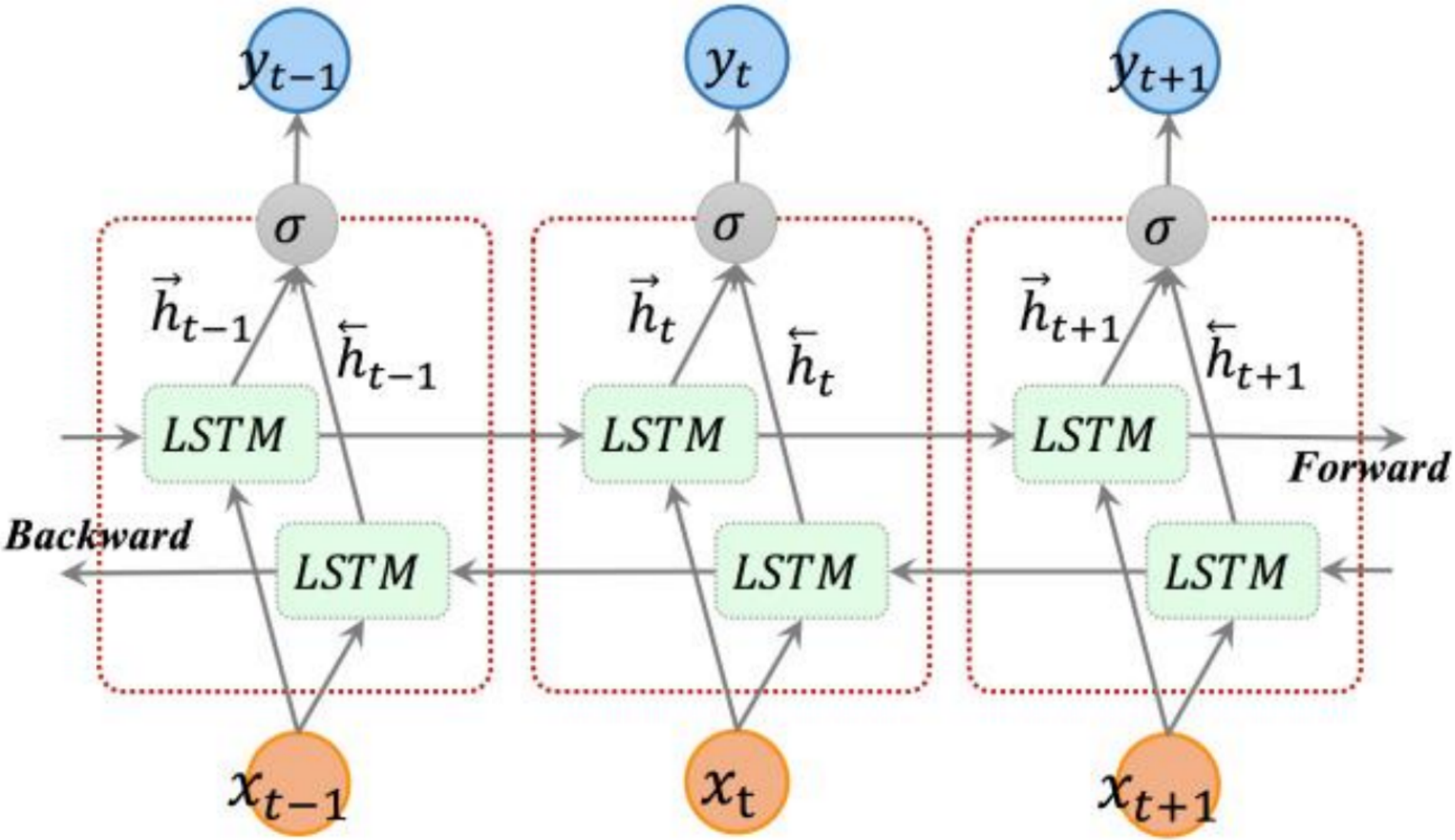
Gated Recurrent Unit (GRU)



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Christopher Olah: Understanding LSTM Networks

Bidirectional RNN



Ссылки

Курс: Машиное обучение от ФИВТ

Курс: Deep Learning на пальцах

Andrej Karpathy: The Unreasonable Effectiveness of Recurrent Neural Networks

Christopher Olah: Understanding LSTM Networks

Denny Britz: Backpropagation Through Time and Vanishing Gradients

OpenAI: Unsupervised Sentiment Neuron

Hochreiter, Sepp & Schmidhuber: Long Short-term Memory

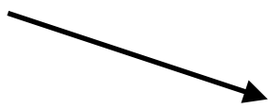
Recap: Natural Language processing

Feature engineering

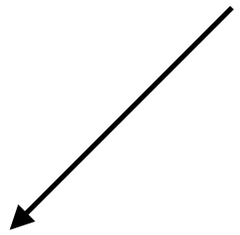
Bag of words

Три документа:

- Артур любит смотреть фильмы
- Ольга тоже любит фильмы
- Денис любит смотреть футбол



Артур любит смотреть фильмы
Ольга тоже Денис футбол



id	Артур	Ольга	Денис	Любит	смотреть	Фильмы	тоже	Футбол
1	1	0	0	1	1	1	0	0
2	0	1	0	1	0	1	1	0
3	0	0	1	1	1	0	0	1

Feature engineering

TF (term frequency — частота слова) — отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова в пределах отдельного документа.

$$\mathrm{tf}(t,d) = \frac{n_t}{\sum_k n_k},$$

где n_t есть число вхождений слова t в документ, а в знаменателе — общее число слов в данном документе.

IDF (inverse document frequency — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции.

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{\, d_i \in D \mid t \in d_i \,\}|},$$ [2]

где

- |D| — число документов в коллекции;
- $|\{\, d_i \in D \mid t \in d_i \,\}|$ — число документов из коллекции D, в которых встречается t (когда $n_t \neq 0$).

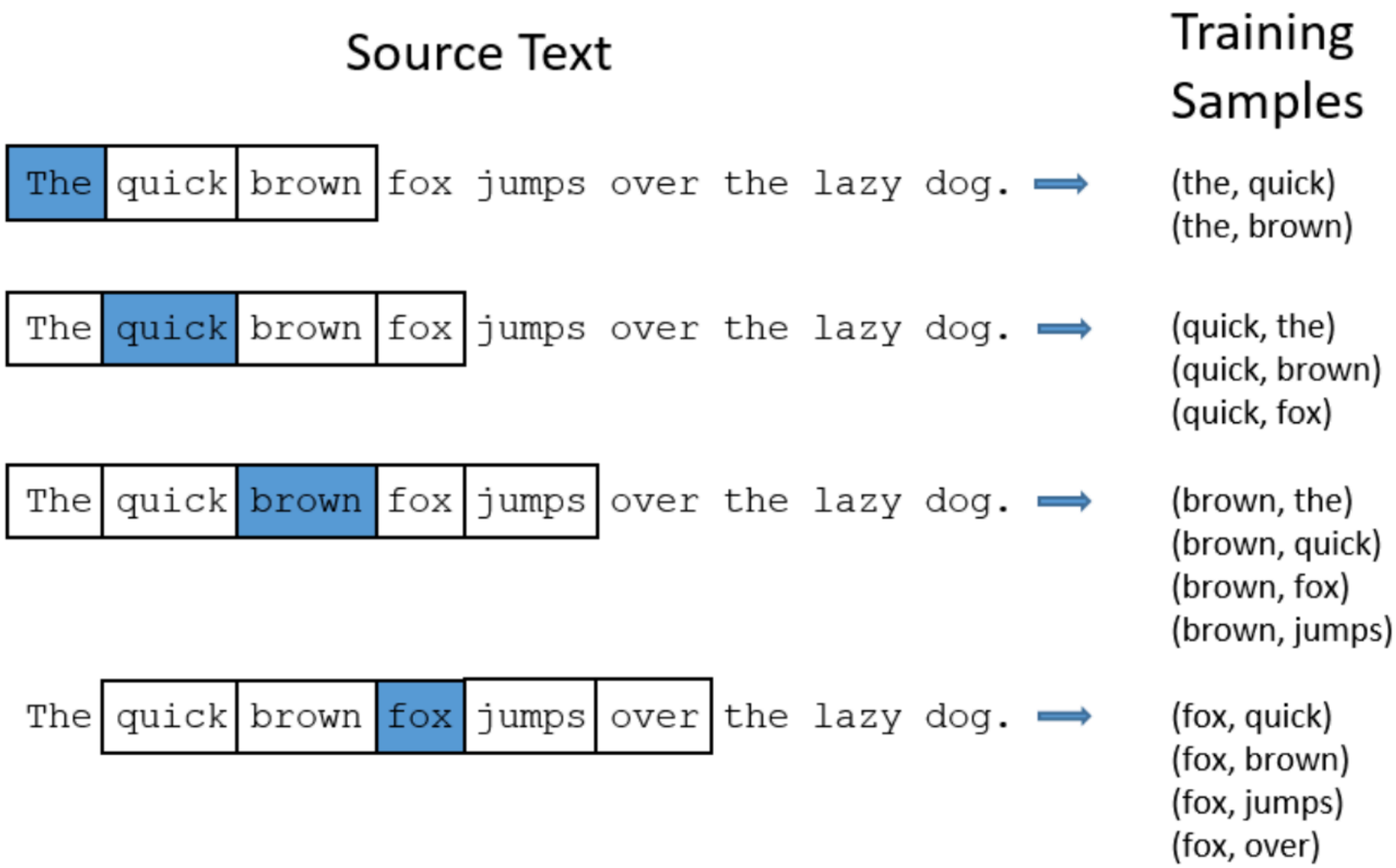
$$\text{tf-idf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

Большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

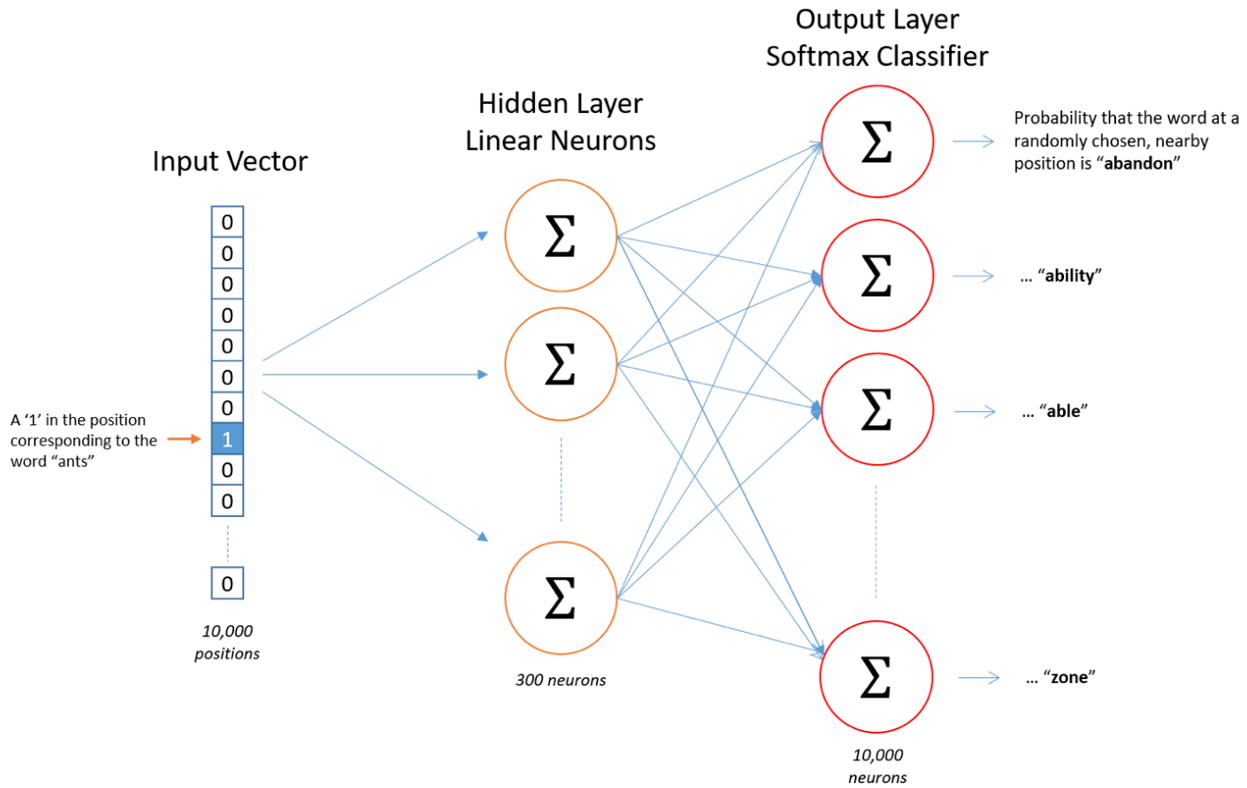
Проблемы

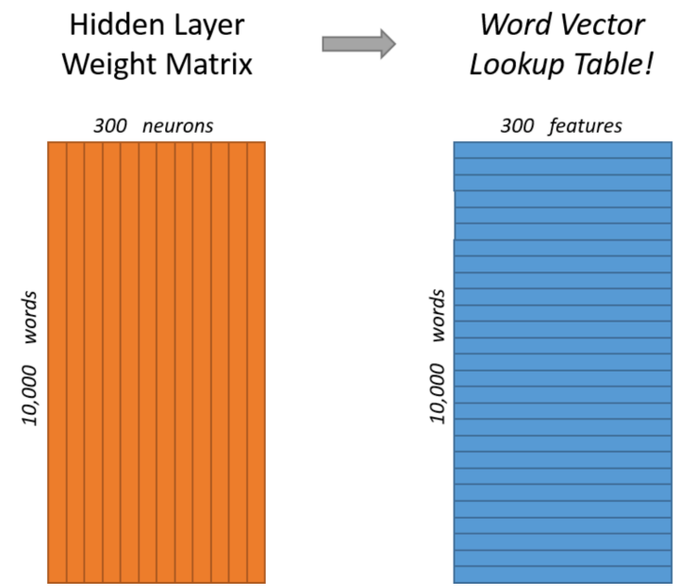
Огромные и сильно разряженные матрицы

Word2Vec

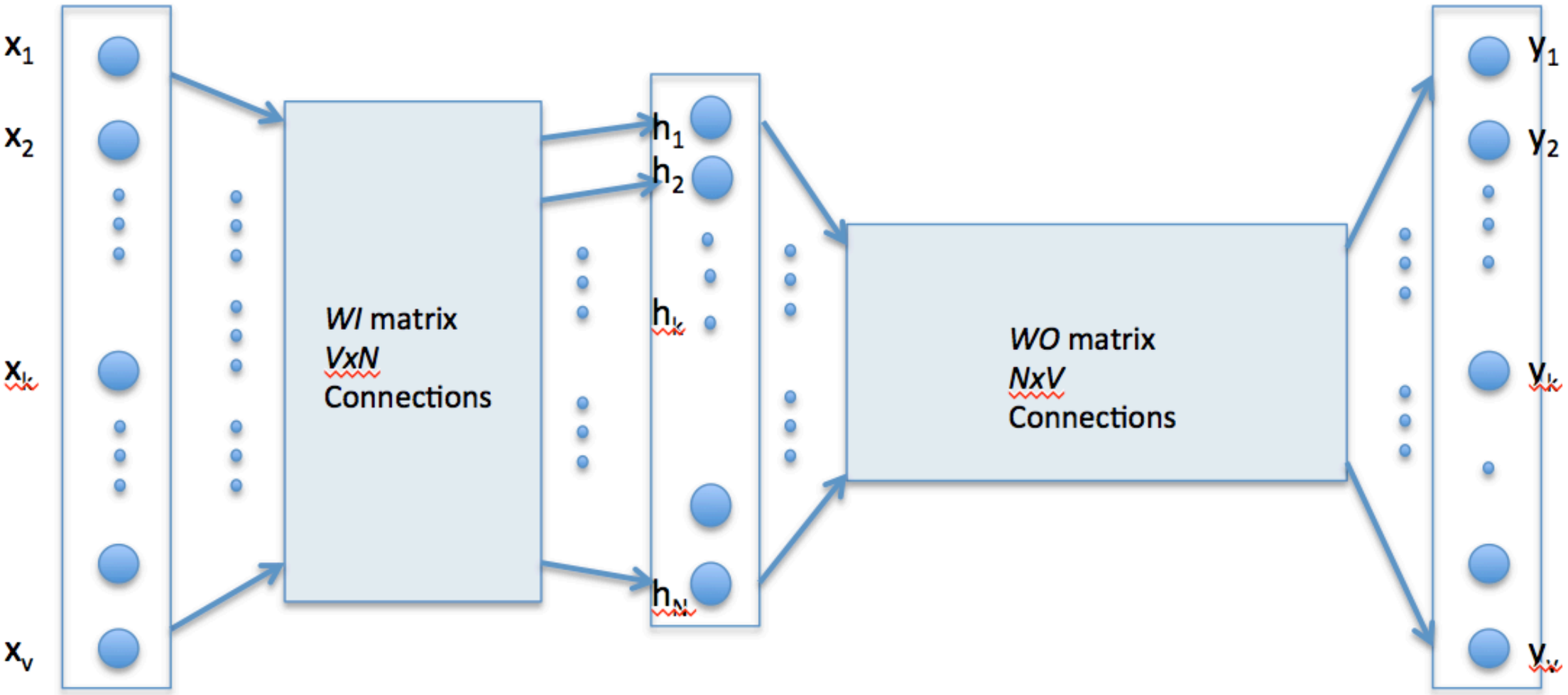

Source Text
Training Samples
The quick brown fox jumps over the lazy dog.
(the, quick)
(the, brown)
The quick brown fox jumps over the lazy dog.
(quick, the)
(quick, brown)
(quick, fox)
The quick brown fox jumps over the lazy dog.
(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)
The quick brown fox jumps over the lazy dog.
(fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)

Word2Vec

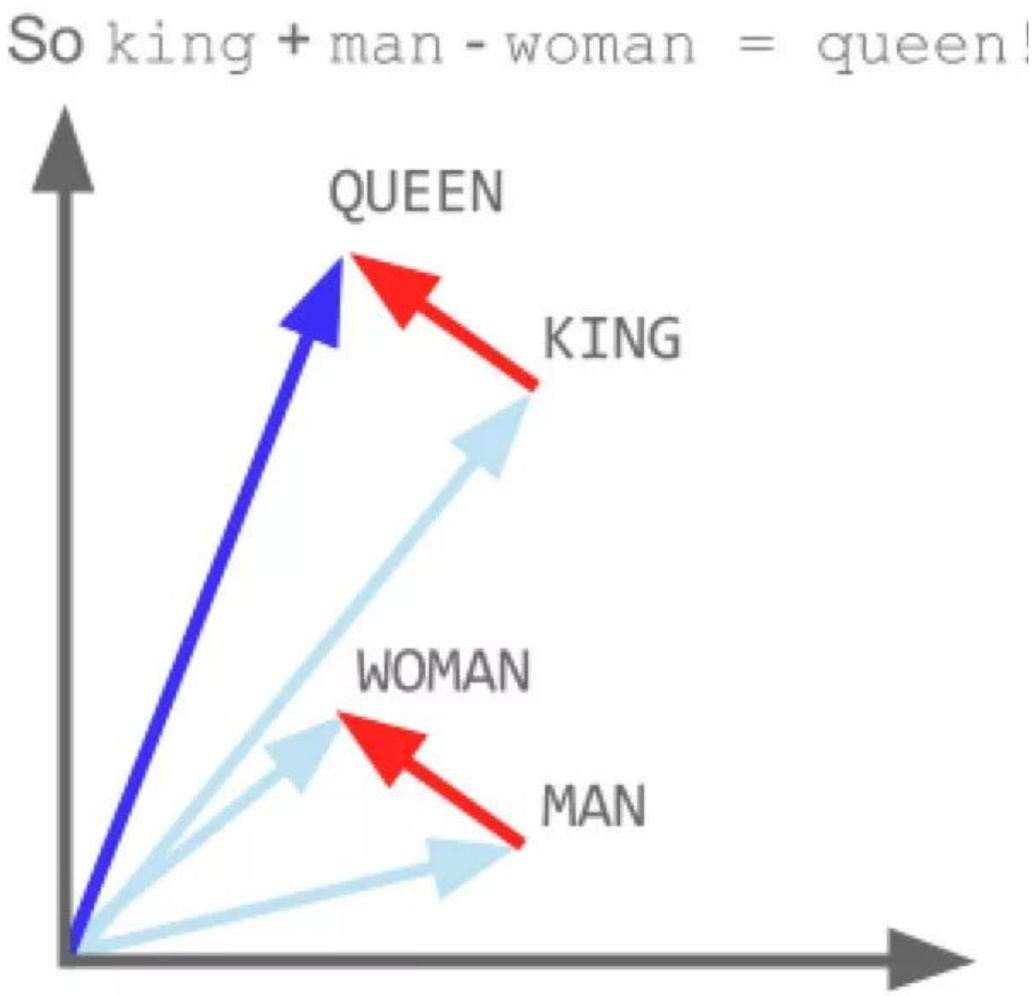

Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"
A '1' in the position corresponding to the word "ants"
0
0
0
0
0
0
1
0
0
0
10,000 positions
300 neurons
10,000 neurons


Hidden Layer
Weight Matrix
Word Vector
Lookup Table!
300 neurons
10,000 words
300 features
10,000 words

Word2Vec


x1
x2
xk
xV
WI matrix
VxN
Connections
h1
h2
hk
hN
WO matrix
NxV
Connections
y1
y2
yk
yV
Input Layer
Hidden Layer
Output Layer

Word2Vec



- Geopolitics: Iraq - Violence = Jordan
- Distinction: Human - Animal = Ethics
- President - Power = Prime Minister
- Library - Books = Hall
- (Moscow - Russia) + France = ?

И это все?

FastText
Glove
BERT

word2vec + LSTM = ❤️