SkillFactory

Recurrent Neural Networks Named Entity Recognition

Sidorov Nikita

MLE (NLP) Sber

What we will learn today

- Problems of working with text;
- when we can use RNN;
- different tasks of sequence modelling task;
- how RNN process data;
- problems of simple RNN;
- evolution of RNNs;
- NER task;
- NER metrics;
- data labeling for NER.

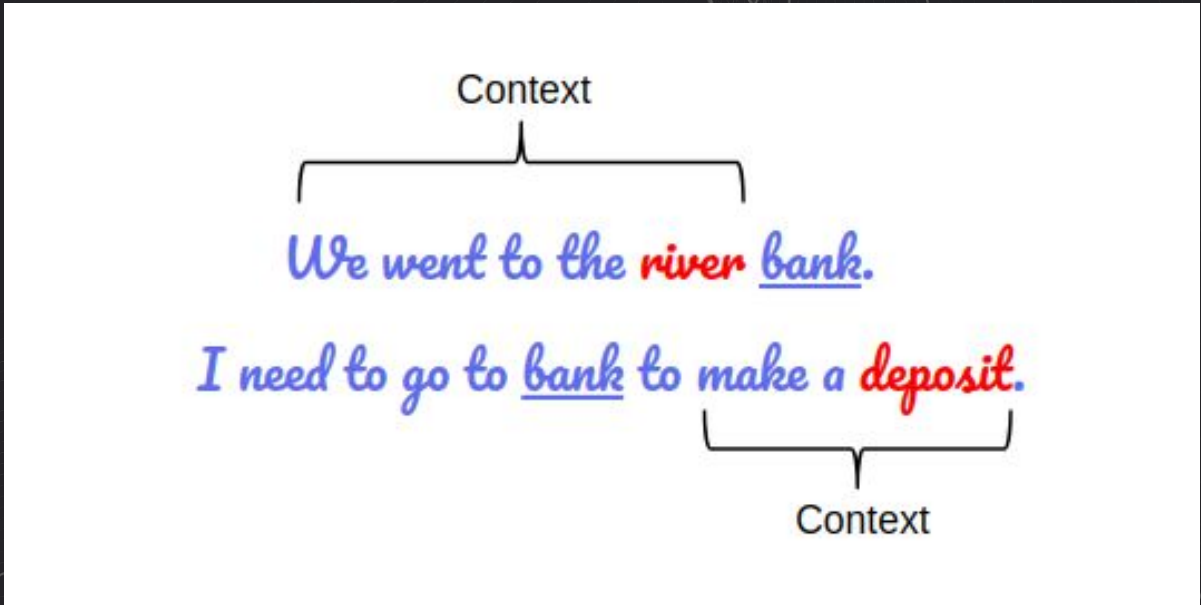
Problems of working with text

I have a cat. My cat is one year old. Her name is _____.

Many people love sports. The most popular sport is football. Perhaps that's why the most popular Instagram account has _____.

To develop memory, it is often advised to learn poetry. In this case, one of the best authors whose works come to mind is _____.

Problems of working with text



Types of sequence modelling tasks

- text classification;
- speech recognition;
- stock market predictions;
- language generation;
- ...

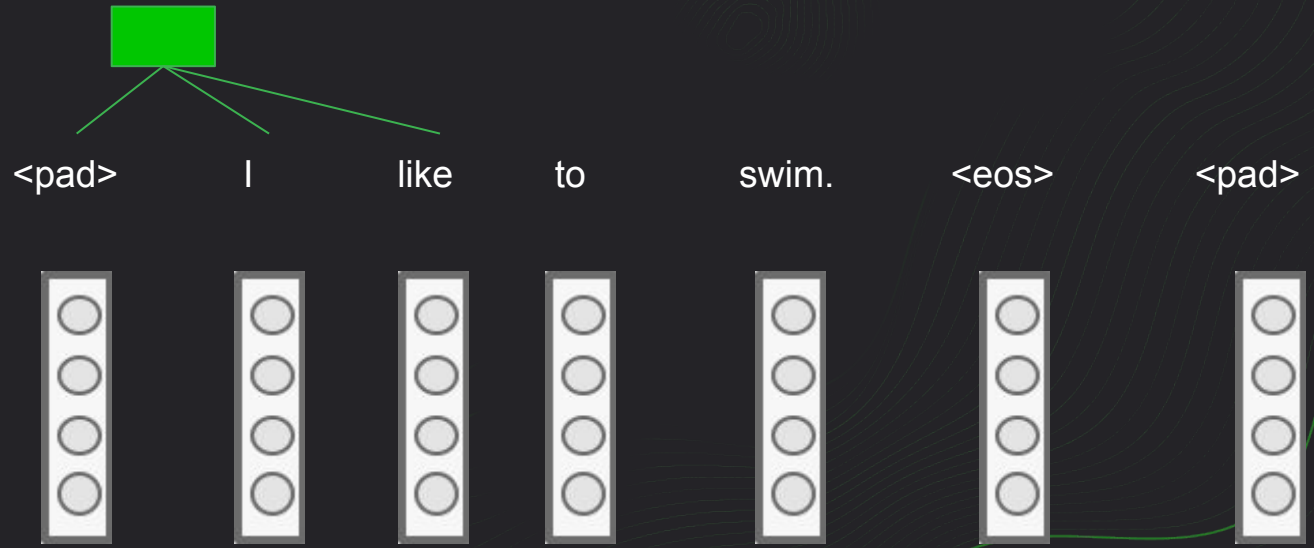
Recap - Bag of Words (BOW)

Cat can meow.

Dog can bark and can not meow.

	Cat	can	meow	Dog	bark	and	not
Sentence 1	1	1	1	0	0	0	0
Sentence 2	0	2	0	1	1	1	1

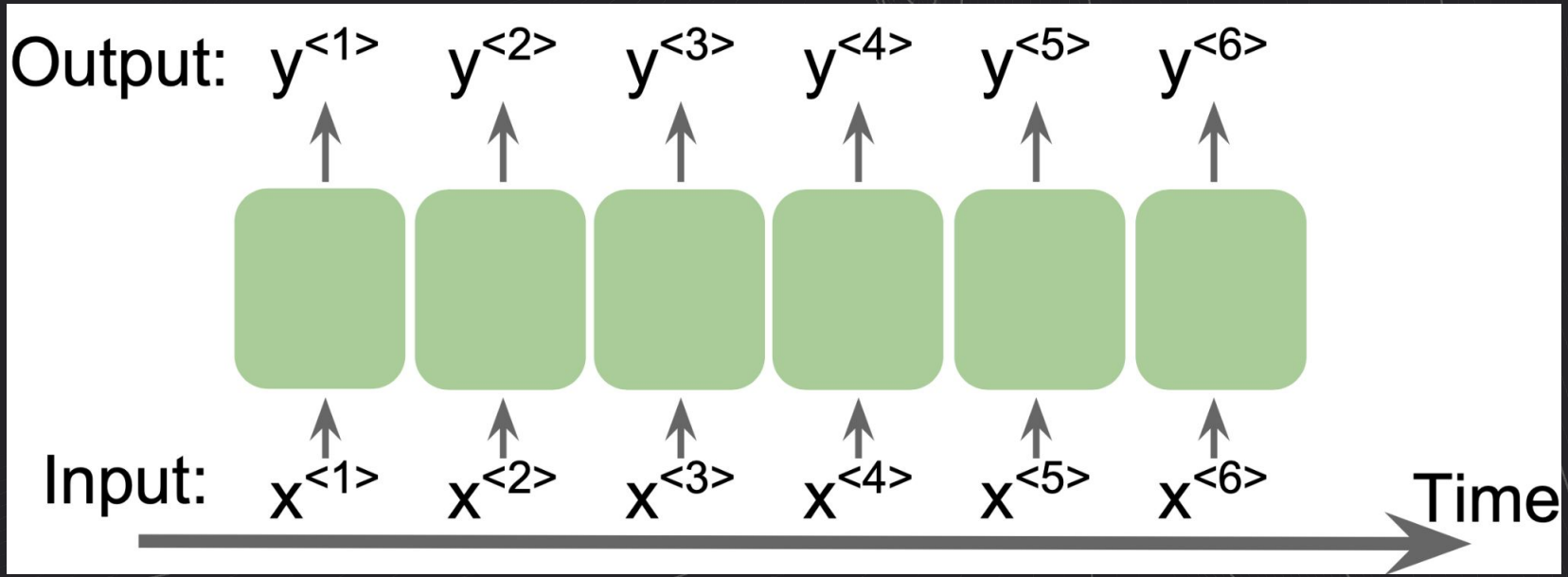
Recap - Convolution for texts



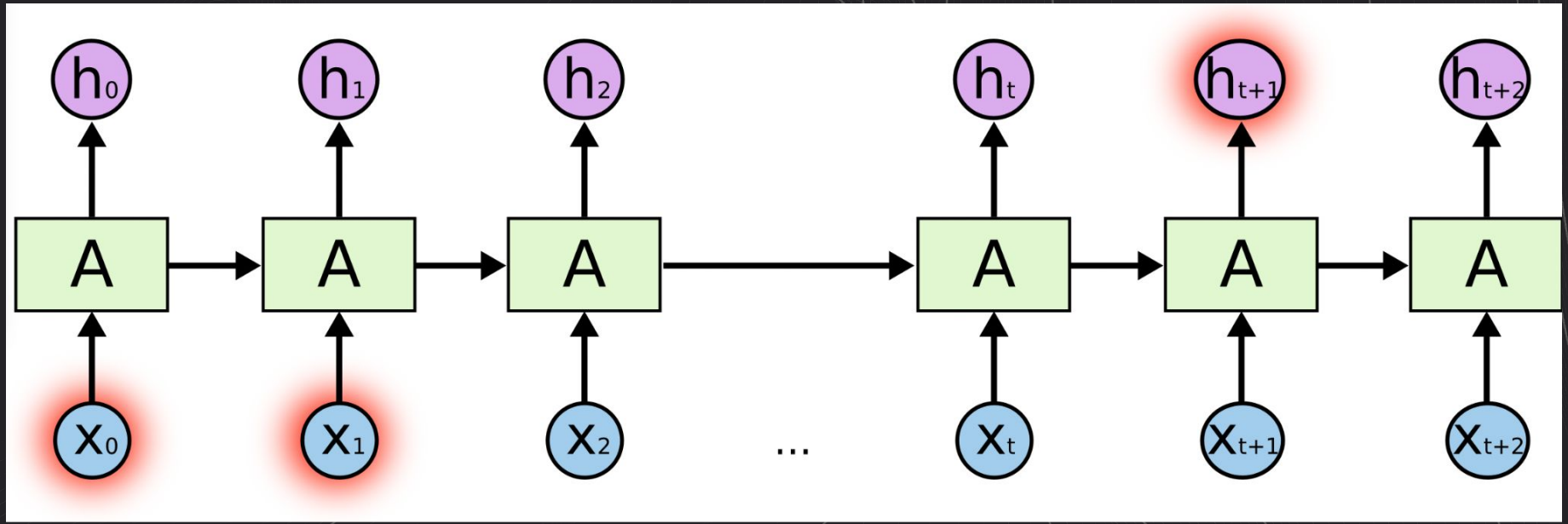
Have they any problems?

Look at the data

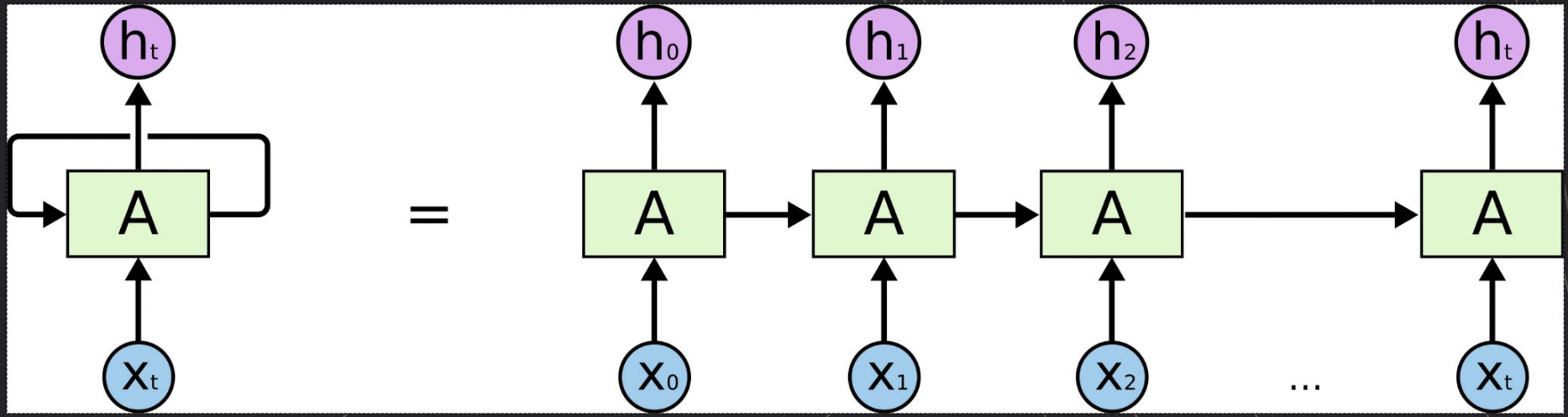
Sequential data is not independent and identically distributed.



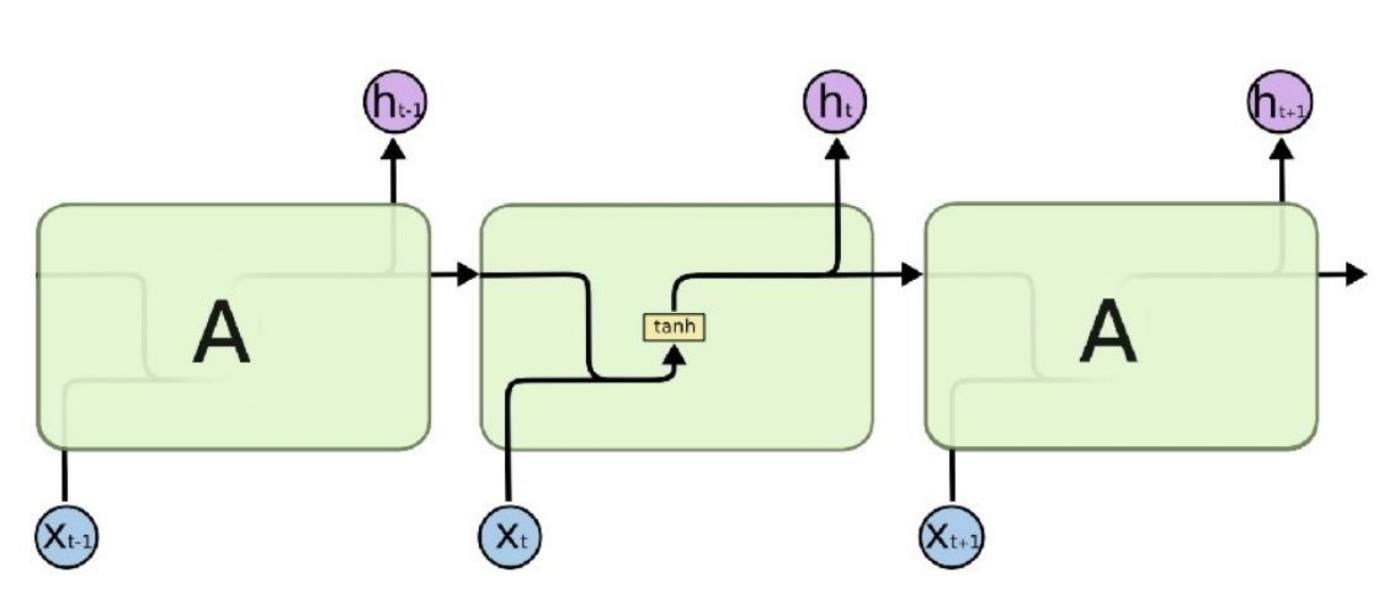
How RNN deal with data



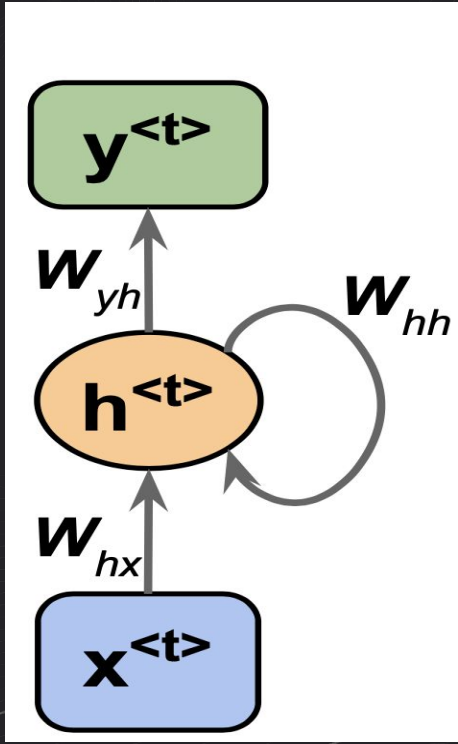
How RNN deal with data



How RNN works



How RNN works



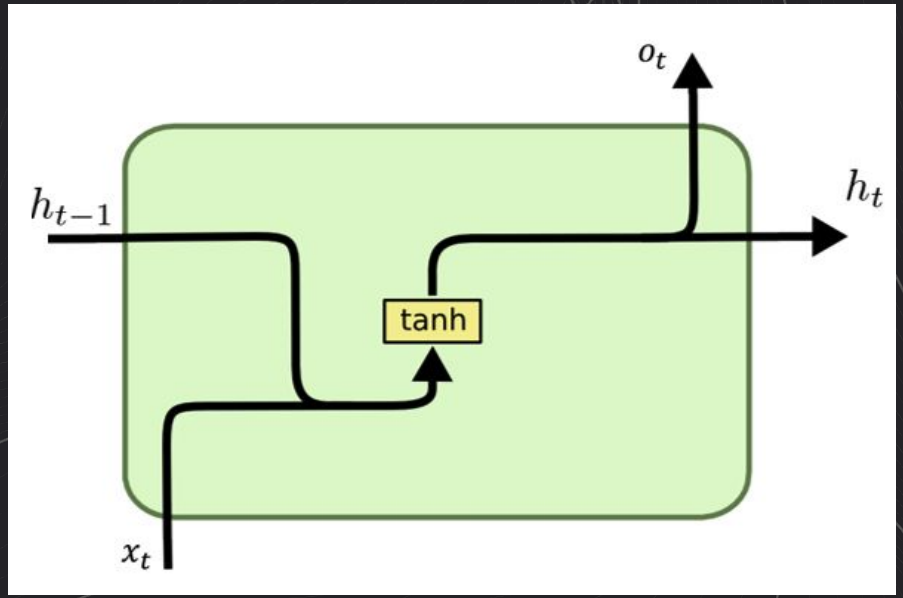
RNN has 2 inputs and 2 outputs!

It means we need 2 weight matrices!

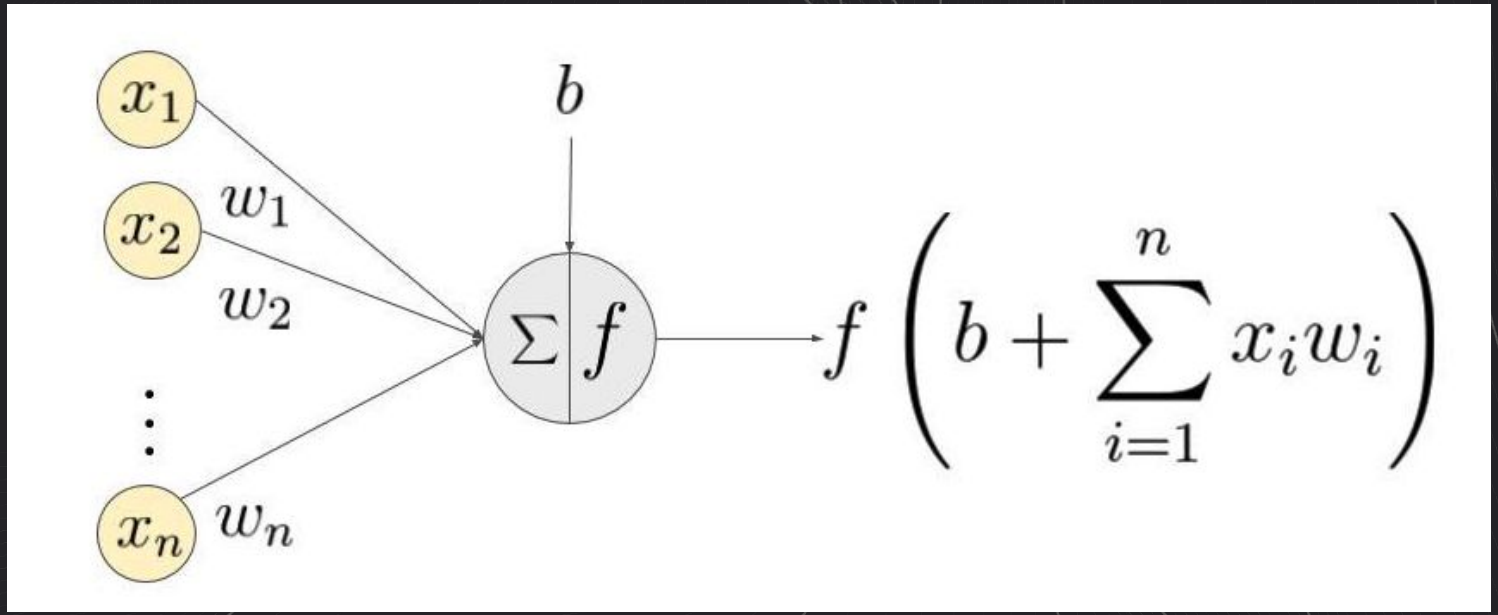
$$W_h = \left[W_{hh} ; W_{hx} \right]$$

Intuition of RNN

Intuition: Let`s add some functionality to classify, what context we need and what context is more appropriate to remember. Tanh function can do it!



Calculations in simple FCNN



Calculations in RNN

$$h_0 = \bar{0}$$

$$h_1 = \sigma\big(\langle W_{\text{hid}}[h_0, x_0]\rangle + b\big)$$

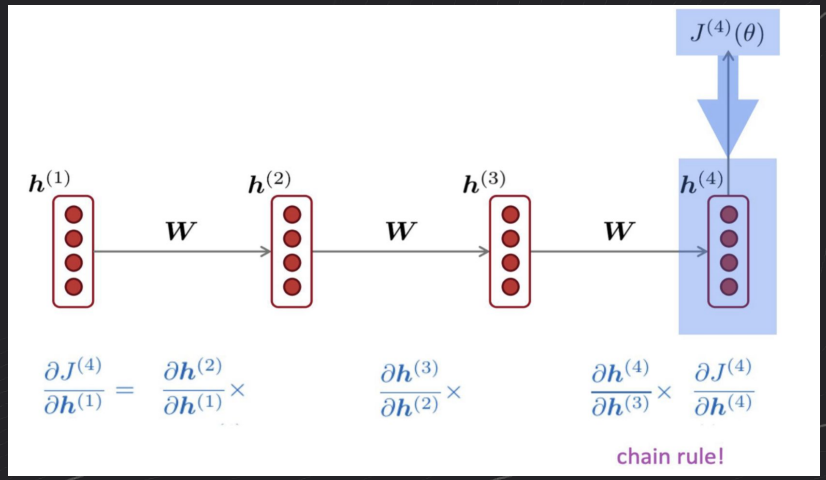
$$h_2 = \sigma\big(\langle W_{\text{hid}}[h_1, x_1]\rangle + b\big) = \sigma\big(\langle W_{\text{hid}}[\sigma\big(\langle W_{\text{hid}}[h_0, x_0]\rangle + b, x_1]\rangle + b\big)$$

$$h_{i+1} = \sigma\big(\langle W_{\text{hid}}[h_i, x_i]\rangle + b\big)$$

$$P(x_{i+1}) = \text{softmax}\big(\langle W_{\text{out}}, h_i\rangle + b_{\text{out}}\big)$$

Problems of RNN

When the derivatives are small, the gradient signals get smaller and smaller as it backpropagates further.(Vanishing gradient problem)



If gradient becomes to big, then optimizer update step becomes too big. (Exploading gradient problem)

learning rate

$$\theta^{new} = \theta^{old} - \alpha \nabla_\theta J(\theta)$$

gradient

Vanishing gradient problem

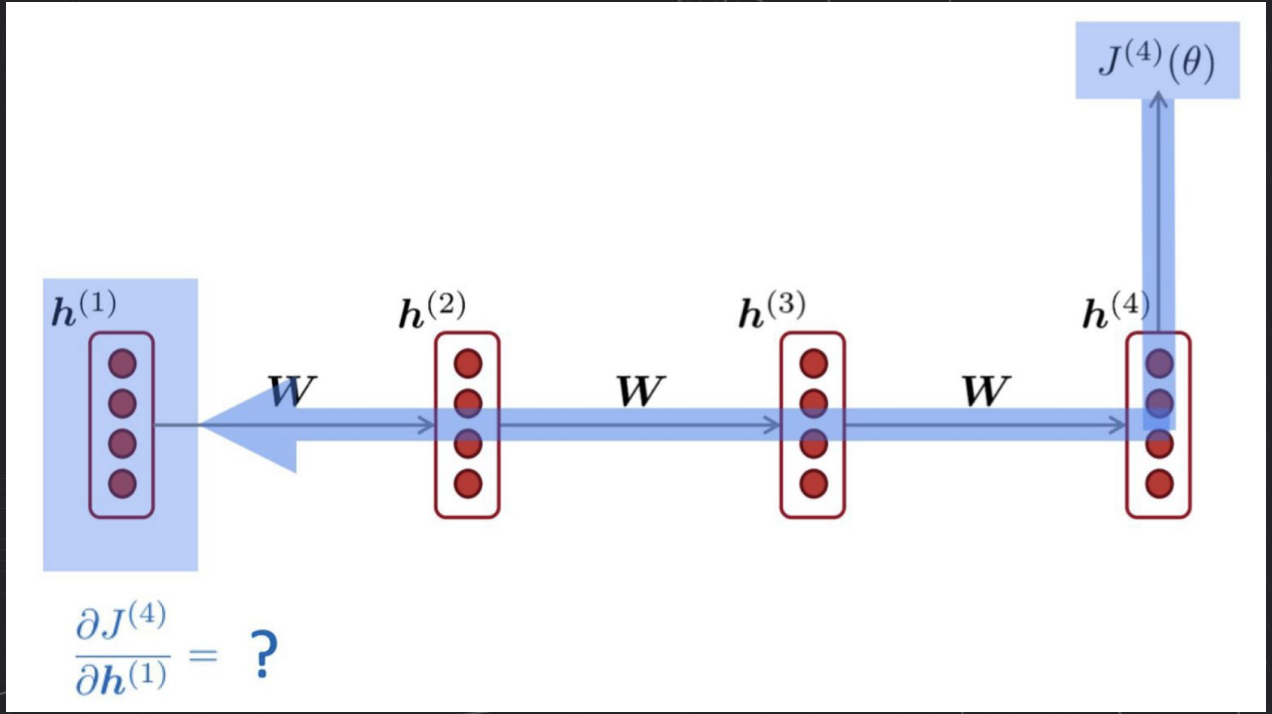
Vanishing problem is present in **all** deep neural networks architectures.

Due to chain rule/choice of nonlinearity function gradient can become vanishingly small during backpropogation.

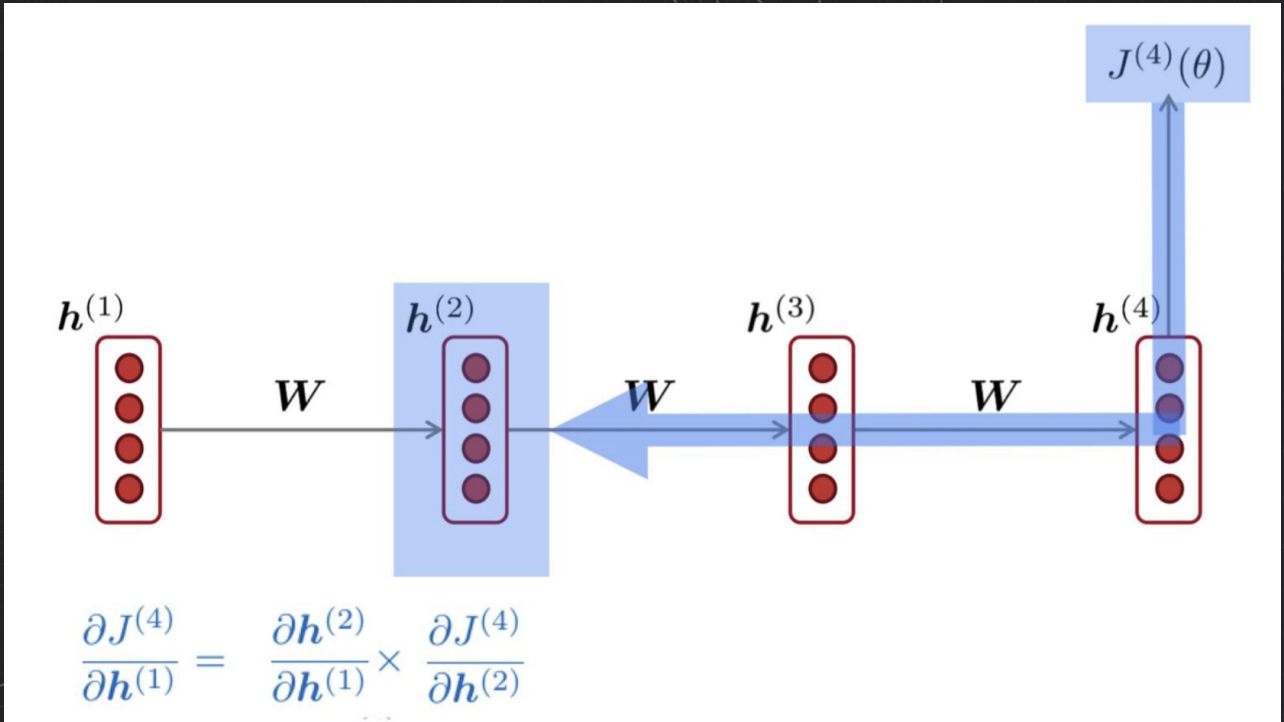
Lower level of NN are hard to train and are trained slower

Though vanishing/exploding gradients are a general problem, RNN are particularly unstable due to repeated multiplication by the same weight matrix[Bengio et al, 1994]. Gradients magnitude drops exponentially with connection length.

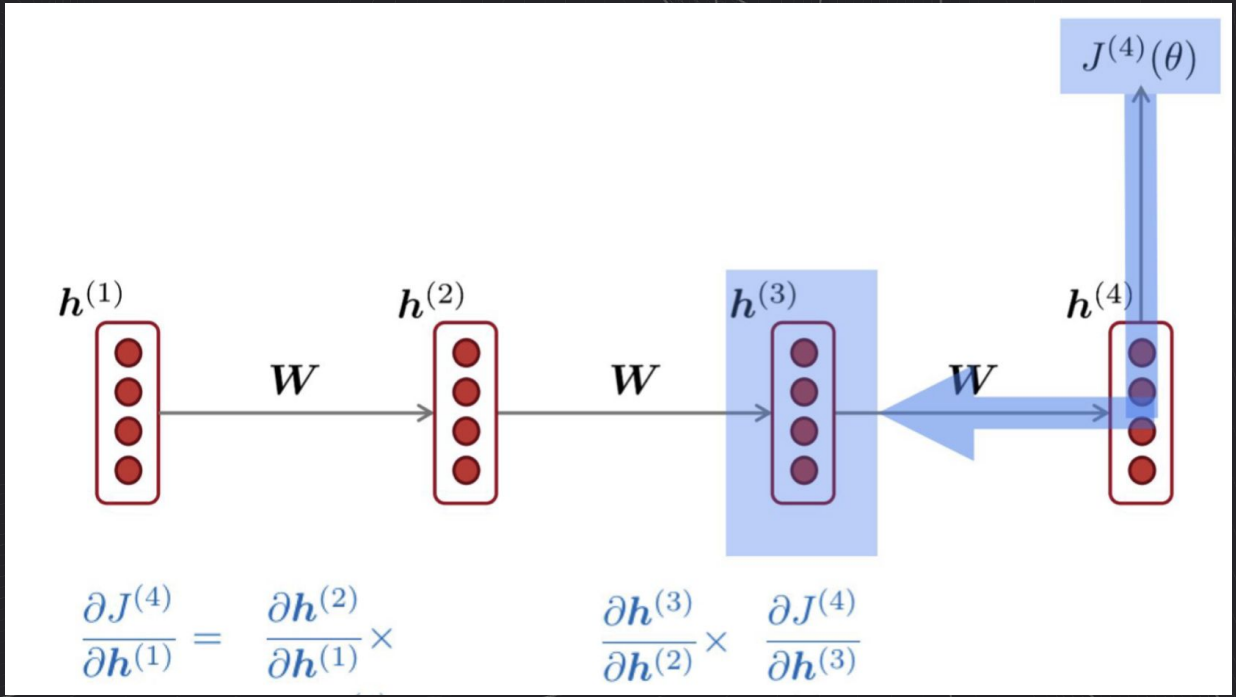
Vanishing gradient problem



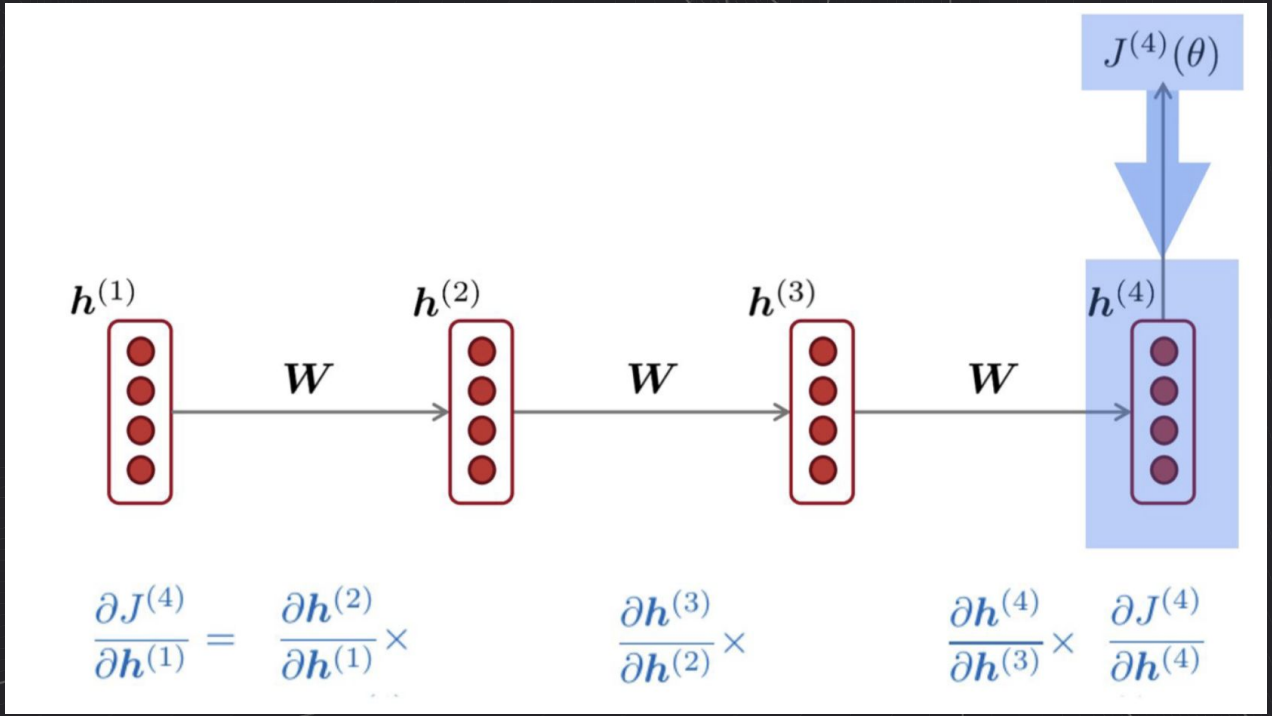
Vanishing gradient problem



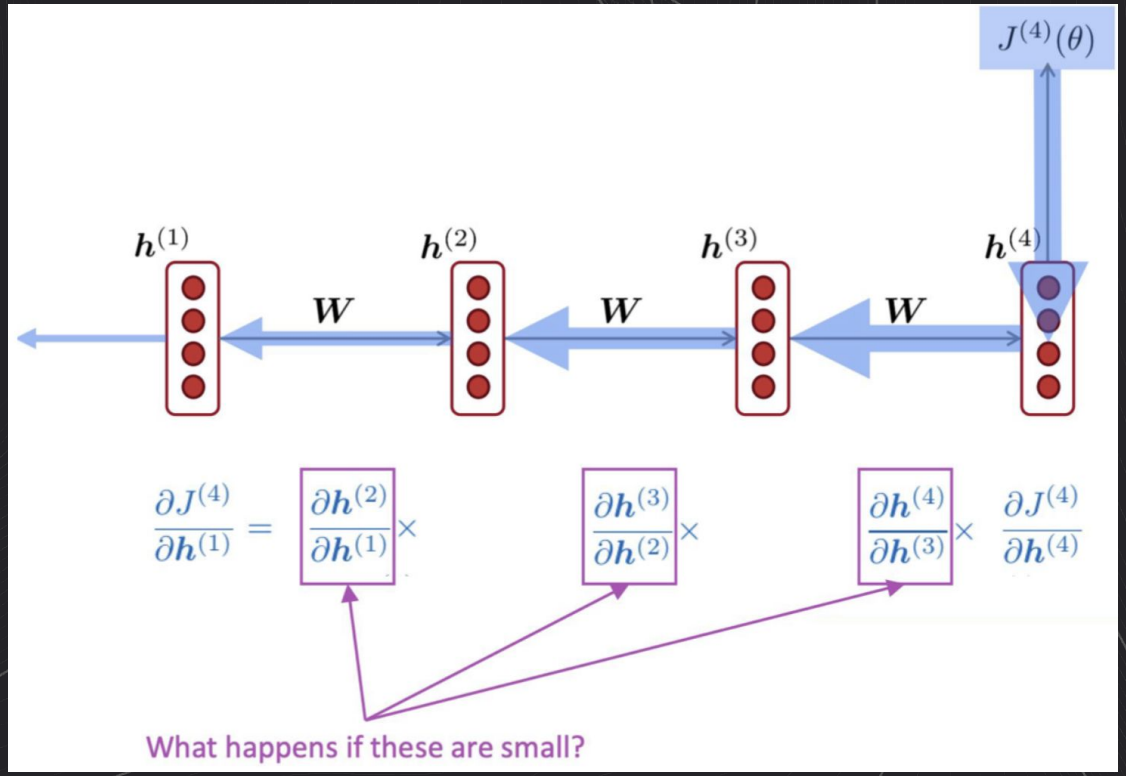
Vanishing gradient problem



Vanishing gradient problem



Vanishing gradient problem



Exploding gradient problem

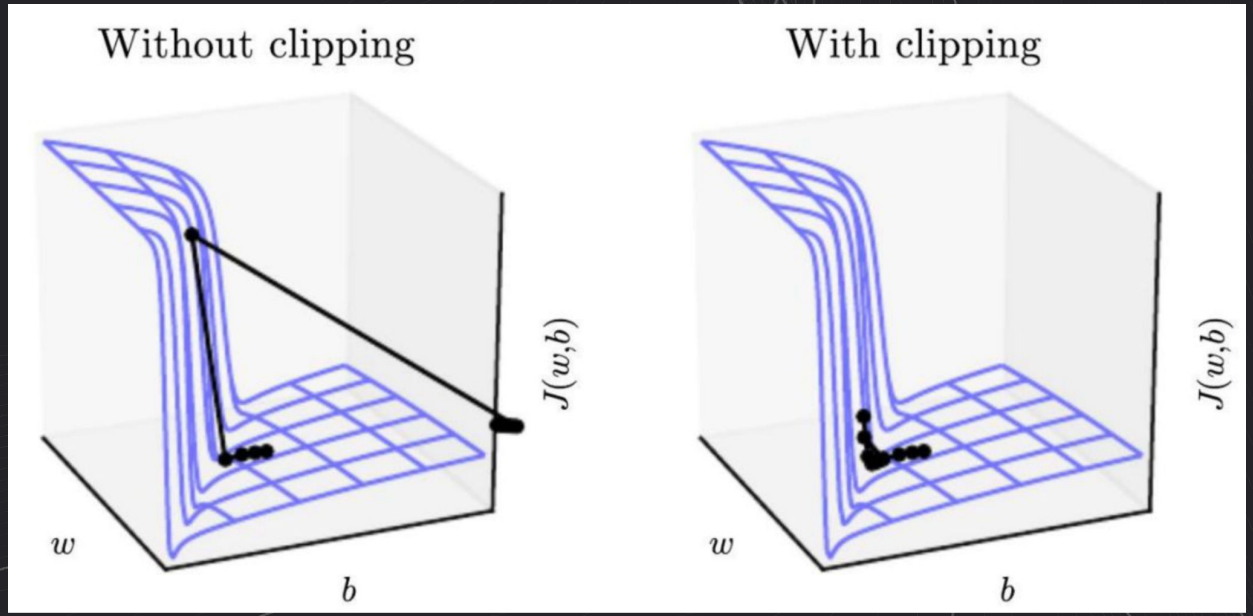
If gradient becomes to big, then optimizer update step becomes too big.

This can cause bad updates: we took too large step and reach bad parameter configuration (with large loss).

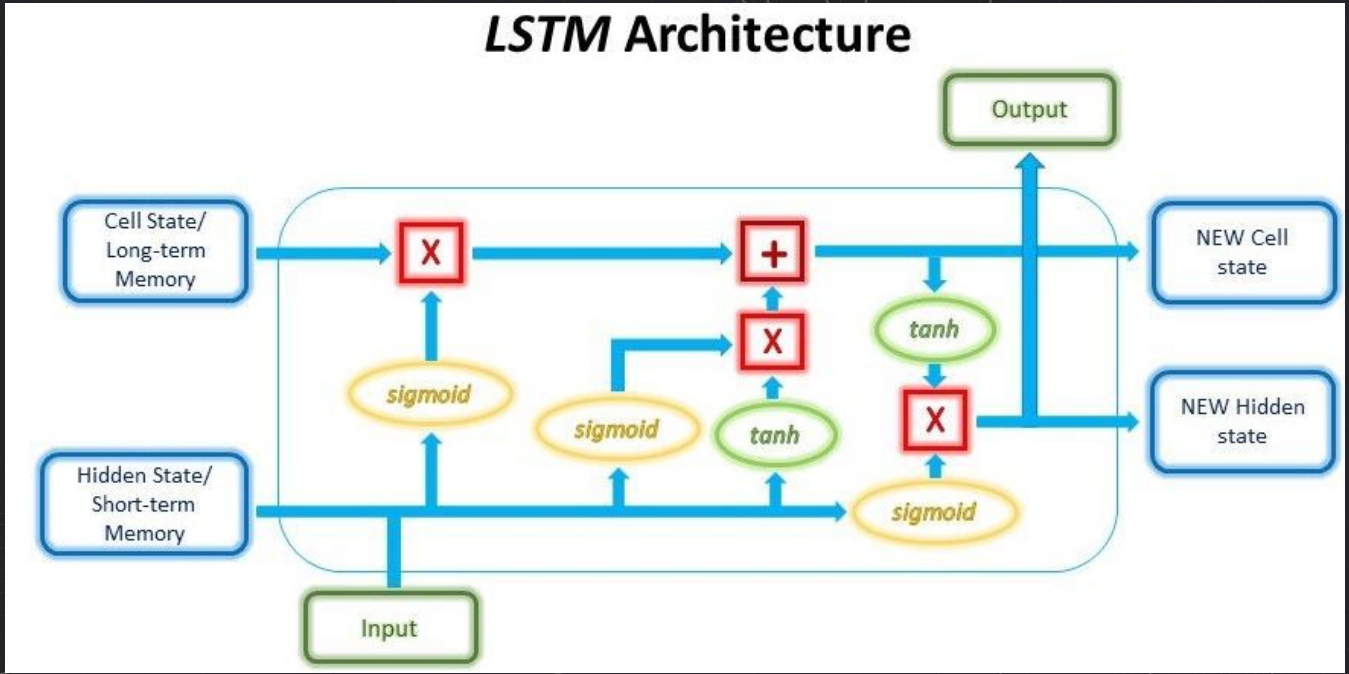
In the worst case, this will result in Inf or NaN in your network (than you need to restart training from the earlier checkpoint).

Exploding gradient problem

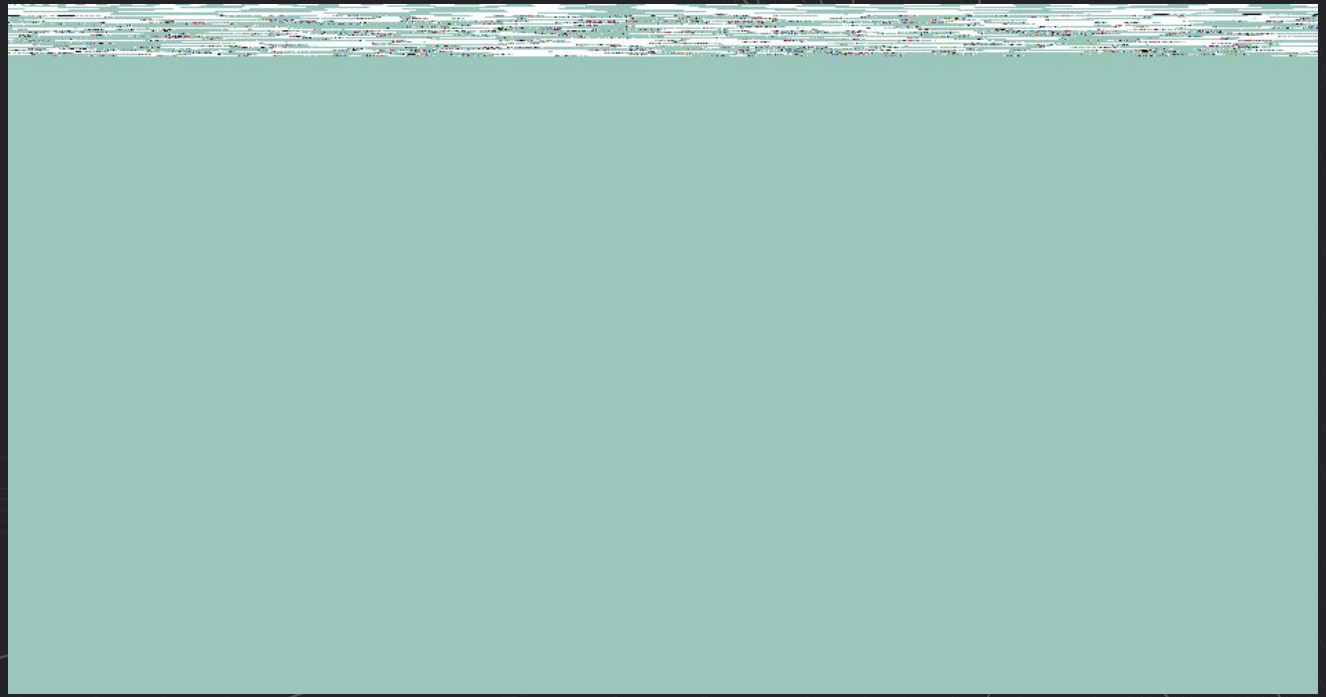
Possible solution -> clipping gradient by some value if norm of the gradient become larger then some threshold.



LSTM

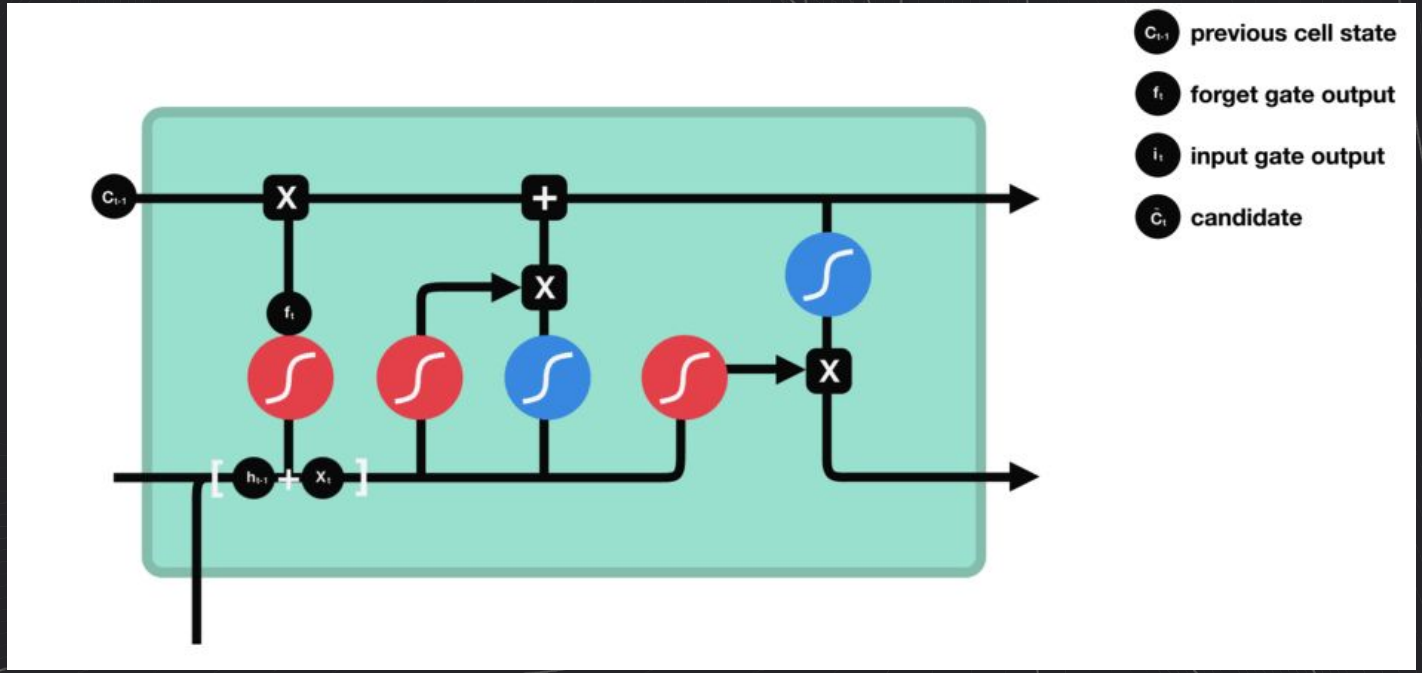

LSTM Architecture
Output
Cell State/
Long-term
Memory
X
+
NEW Cell
state
tanh
X
sigmoid
sigmoid
tanh
X
NEW Hidden
state
Hidden State/
Short-term
Memory
sigmoid
Input

LSTM
Forget gate

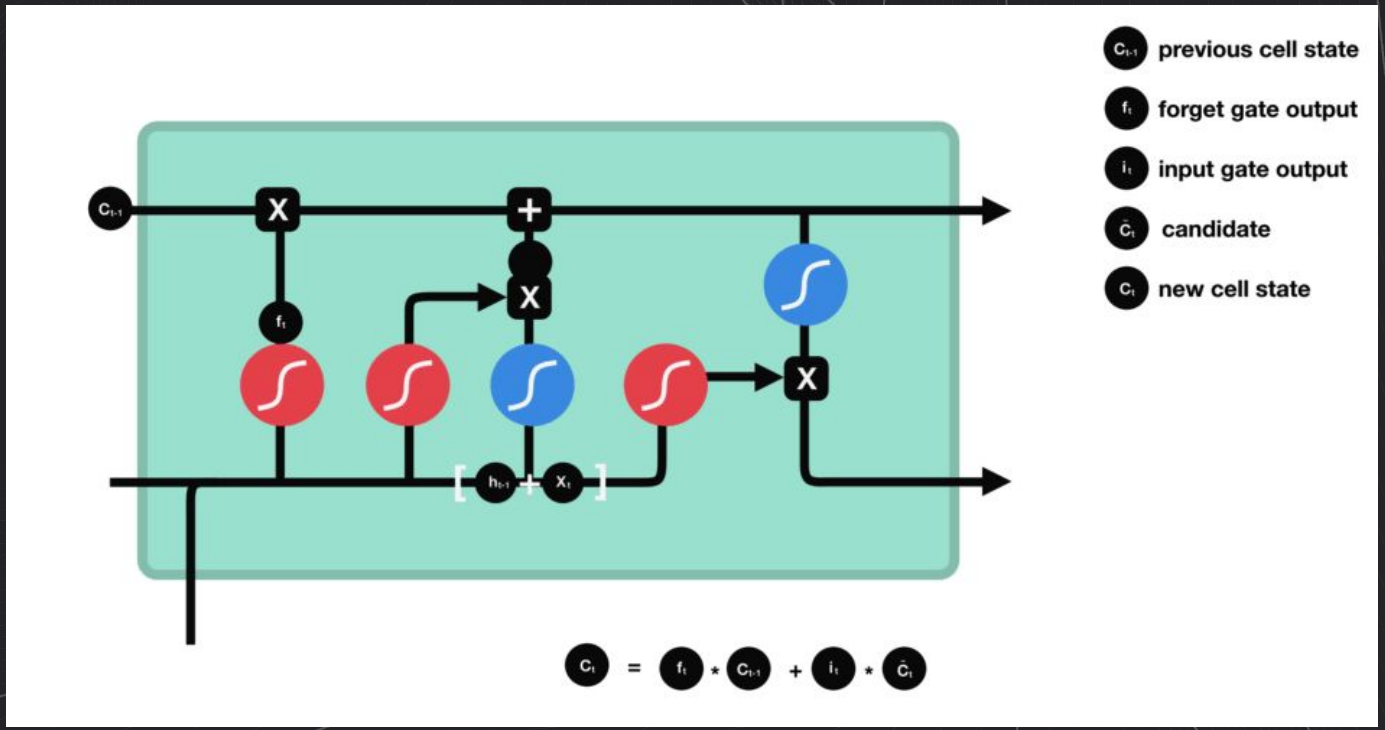


images from this article

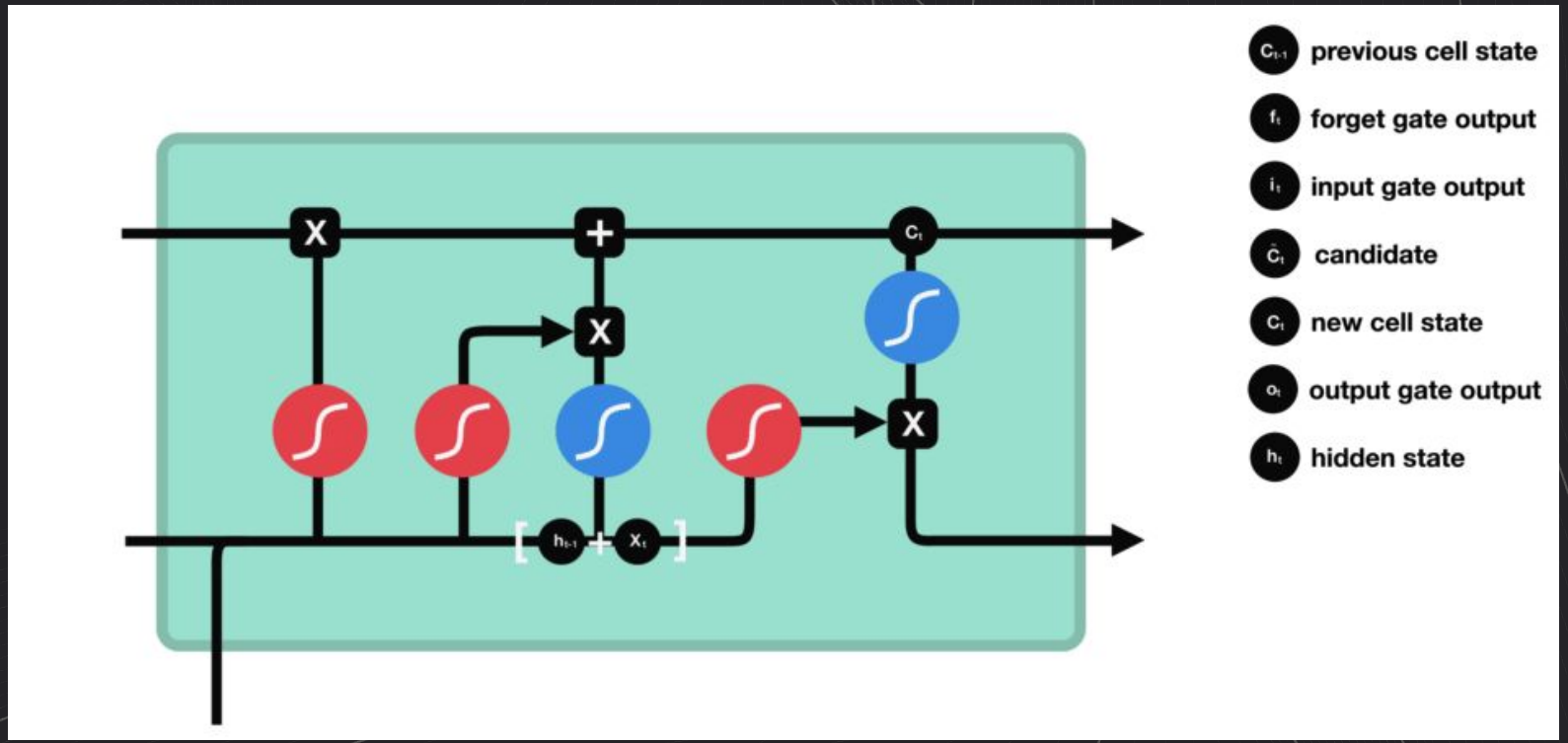
LSTM
Input gate



LSTM
Cell state

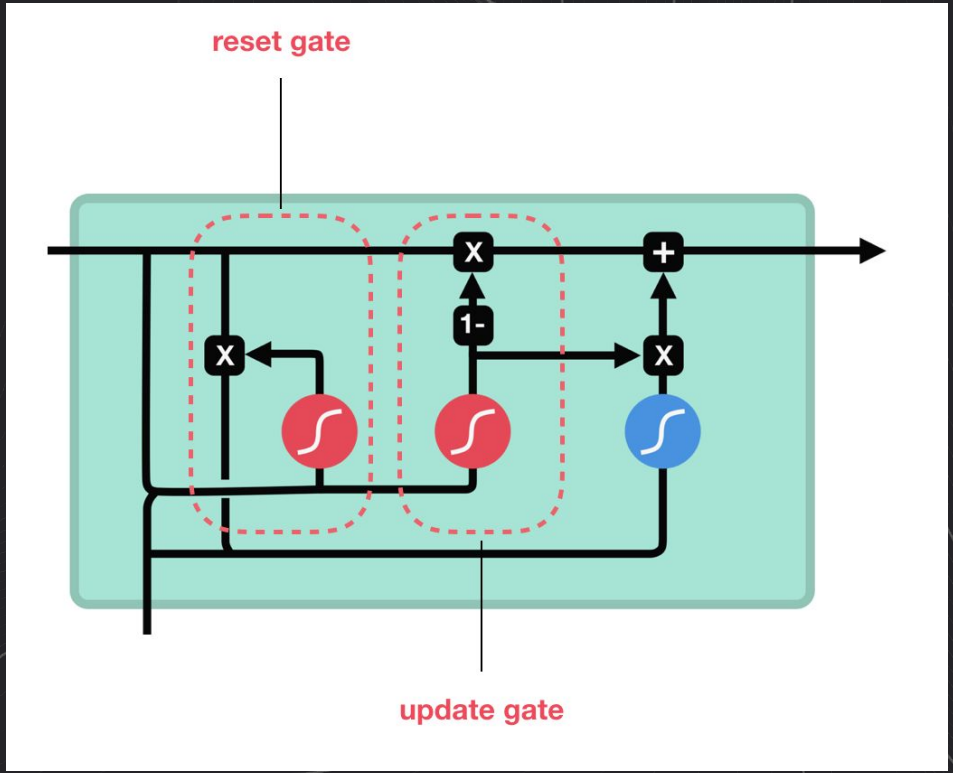


LSTM
Output gate

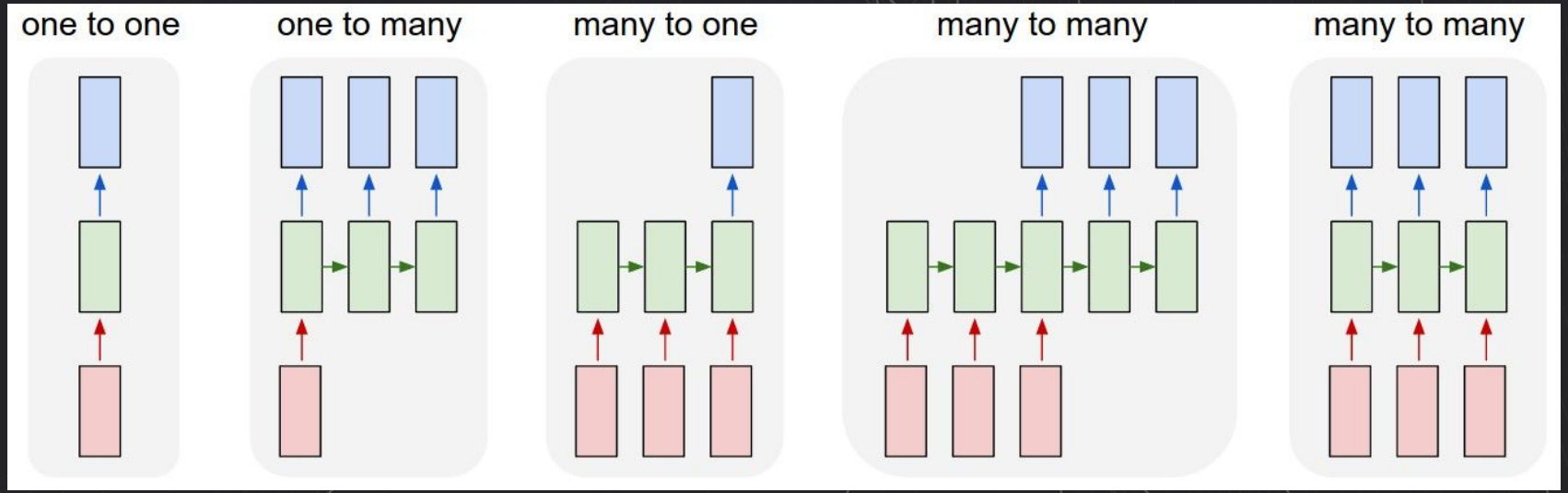


GRU

GRU



RNNs applications



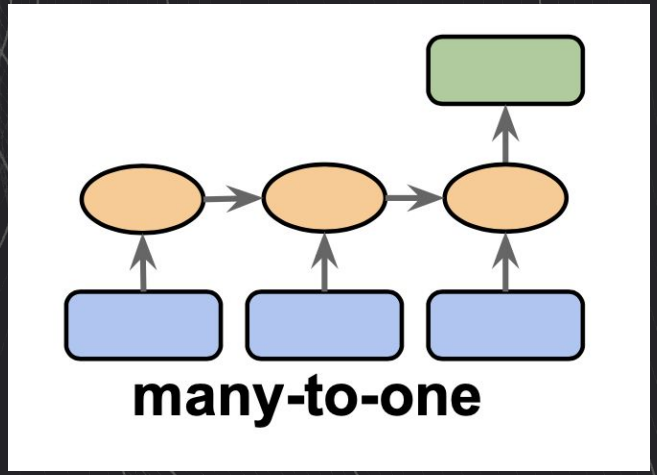
RNNs applications

The input data is a sequence

The output data is a fixed-size vector

Examples:

- sentiment analysis
- text classification
- etc



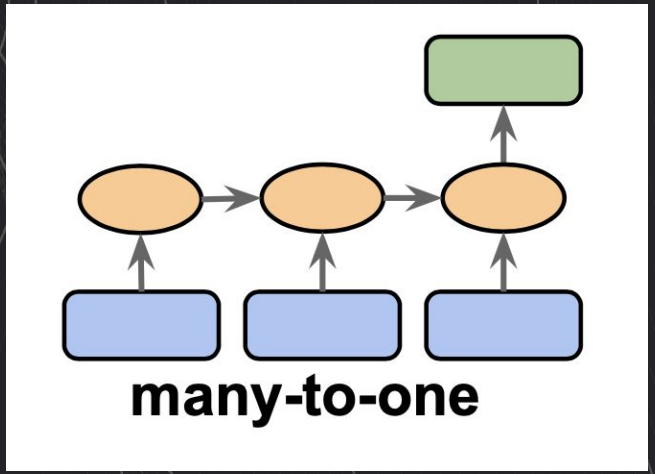
RNNs applications

The input data is a sequence.

The output data is a fixed-size vector.

Examples:

- sentiment analysis
- text classification



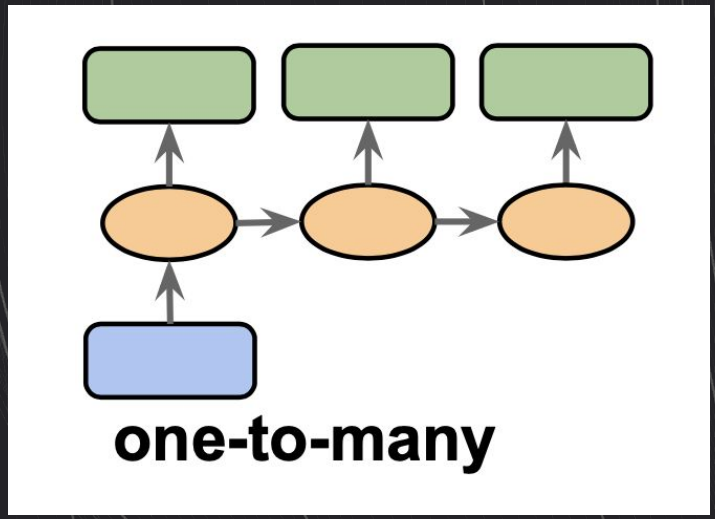
RNNs applications

The input data is in a standard format(not a sequence).

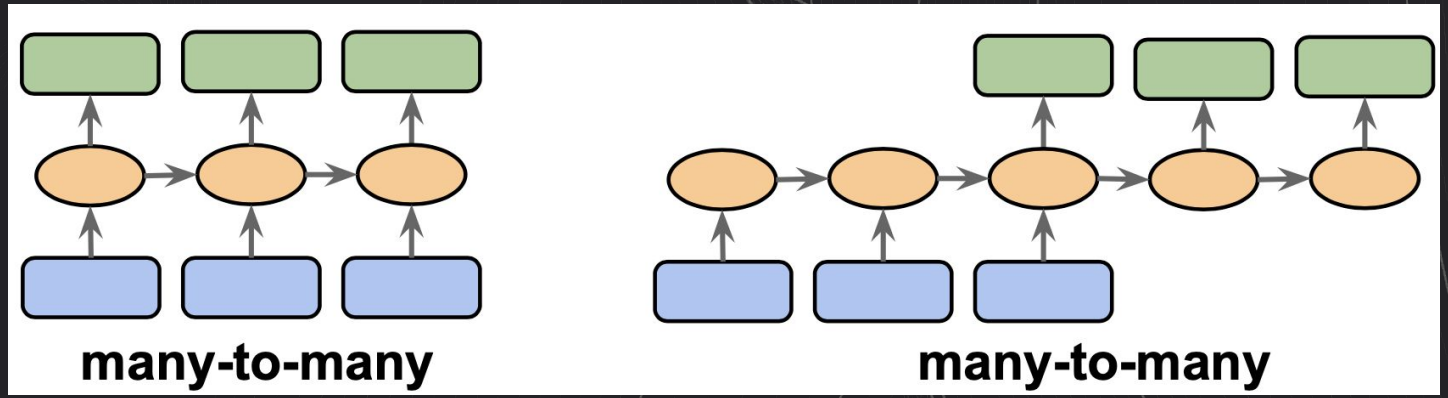
The output data is a sequence.

Examples:

- image captioning



RNNs applications



The input and output data is sequences. Can be direct or delayed.

Examples:

- video-captioning (direct)
- text-translation (delayed)

NER

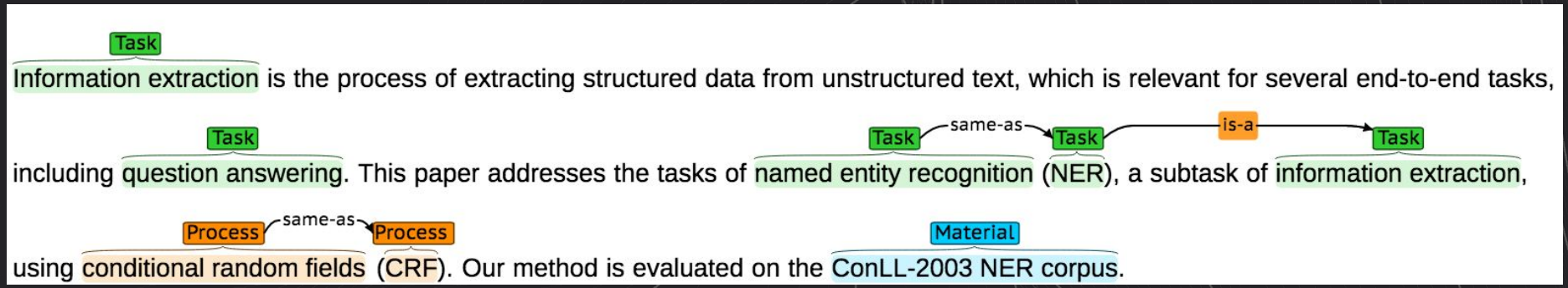
Named Entity Recognition (NER) - NLP task, which classify each word to be some sort of entity.

When Sebastian Thrun PERSON started at Google ORG in 2007 DATE , few people outside of the company took him seriously. “I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun PERSON , now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE .

A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

NER

Named Entity Recognition (NER) - NLP task, which classify each word to be some sort of entity.



NER applications

- processing of accounting documents
- detection of personal information
- speech recognition
- sentiment analysis
- text masking
- complex text analysis
- part of speech tagging
- etc

NER labeling (IO encoding)

On April 4, 1975 childhood friends Bill Gates and Paul Allen found Microsoft.

On	April	4,	1975	childhood	friends	Bill	Gates	and	Paul	Allen	found	Microsoft.
O	D	D	D	O	O	PER	PER	O	PER	PER	O	ORG

Inside - detect entity

Outside - detect other token

NER labeling (BIO encoding)

On April 4, 1975 childhood friends Bill Gates and Paul Allen found Microsoft.

On	April	4,	1975	childhood	friends	Bill	Gates	and	Paul	Allen	found	Microsoft.
O	B-D	I-D	I-D	O	O	B-PER	I-PER	O	B-PER	I-PER	O	B-ORG

Begin - detect beginning of entity

Inside - detect continuation of entity

Outside - detect other token

NER labeling (BILUO encoding)

On April 4, 1975 childhood friends Bill Gates and Paul Allen found Microsoft.

On	April	4,	1975	childhood	friends	Bill	Gates	and	Paul	Allen	found	Microsoft.
O	B-D	I-D	L-D	O	O	B-PER	L-PER	O	B-PER	L-PER	O	U-ORG

Begin - detect beginning of entity

Inside - detect continuation of entity

Last - detect end of entity

Unit - detect single entity

Outside - detect other token

Terminology

- RNN - Recurrent Neural Network
- FCNN - Fully Connected Neural Network
- BOW - Bag Of Words
- LSTM - Long Short Time Memory
- Tanh - Hyperbolic tangent
- sigmoid - sigmoid function
- softmax - softmax function
- vanishing gradient
- exploding gradient
- gate (information flow element)
- GRU - Gated Recurrent Unit
- NER - Named Entity Recognition