

# Deep NLP 1: Recurrent Neural Networks

Oliver Guhr



# M.Sc. Oliver Guhr

University of Applied Sciences Dresden

Faculty of Computer Science and Mathematics

Department of Artificial Intelligence



[oliver.guhr@htw-dresden.de](mailto:oliver.guhr@htw-dresden.de)

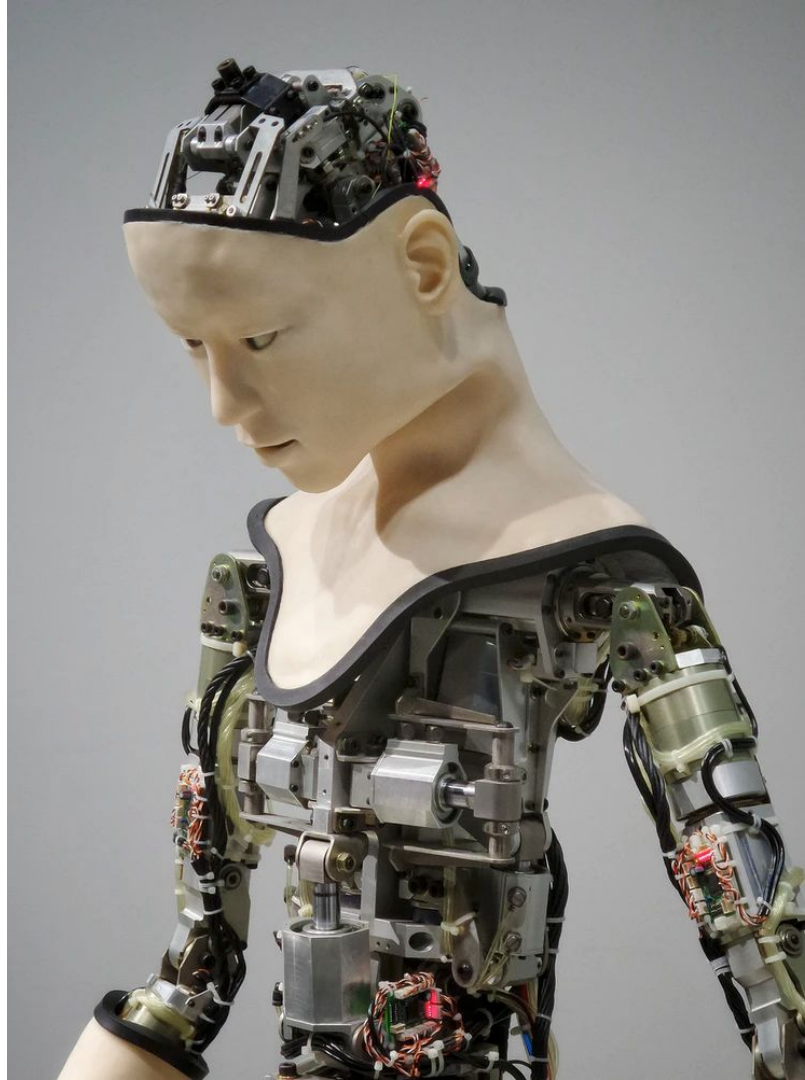
# Our Robots

Tesaro, August  
and Anna



# Goal for today

- Warm Up
- First overview about NLP
- Learn about RNNS
- Train an RNN





Warm Up

# How is the pace of this course?

(to slow, just right, to fast)

**How do you feel about  
this course?**

# **Do you have any questions?**

(Do we talk about model x? etc.)



# AI Buzz



Home > Medien > Fake-News-Generator - Mit Lügen spielen

Medienberufe

15. Mai 2019, 19:00 Uhr Fake-News-Generator

## Mit Lügen spielen



Enorme Datenmengen: Die Forscher beschreiben, dass sich der Bot mit Informationen von mehr als acht Millionen Webseiten nährt. (Foto: Franziska Gabbert/dpa)

**Auf einer neuen Website können Nutzerinnen und Nutzer aus wenigen Stichworten Falschmeldungen generieren lassen. Eigentlich wollten die Entwickler das Werkzeug unter Verschluss halten, um Missbrauch auszuschließen.**

[The Guardian view](#) [Columnists](#) [Cartoons](#) [Opinion videos](#) [Letters](#) [More](#)**Opinion** Artificial intelligence (AI)

● This article is more than 2 months old

## A robot wrote this entire article. Are you scared yet, human?

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

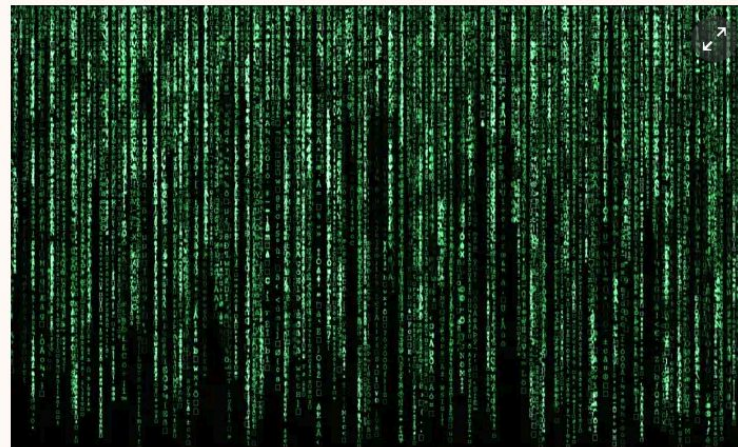
- For more about GPT-3 and how this essay was written and edited, please read our editor's note below

**GPT-3**

Tue 8 Sep 2020 09:45 BST



69.630 1.188



## Artificial intelligence (AI)

Alex Hern

@alexhern

Thu 14 Feb 2019 17:00 GMT

6.473 572

## New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse



## Editorially independent, open to everyone

We chose a different approach – will you support it?

Support The Guardian →

### most viewed



**Live** US-China trade war: Beijing vows to retaliate as tariffs raised - Business live



Anna Sorokin: fake German heiress sentenced to up to 12 years in prison



Freddie Starr: comedian found dead at home in



**Support The**

Available for everyone, fund

[Contribute →](#)

[Subscri](#)

**News**

World UK Scien

**Artificial in  
(AI)**

**Alex Hern**

🐦 @alexhern

Thu 14 Feb

**TE**

Login

INDY/LIFE

AI  
OpenAI let us try :  
generator

**OpenAI built a text generator so good, it's  
considered too dangerous to release**

Zack Whittaker @zackwhittaker / 3 months ago  
art NLP text

**AI TEXT GENERATOR TOO  
DANGEROUS TO RELEASE, SAY  
CREATORS**

Developers cite concerns over fake news proliferation and risk of online impersonation

ially  
endent,  
o everyone

ferent approach –  
ort it?

[Guardian →](#)

US-China trade war:  
ng vows to retaliate as  
fs raised - Business live

a Sorokin: fake German  
ess sentenced to up to  
ars in prison

die Starr: comedian  
d dead at home in

# Natural Language Processing

“processing of texts”

# Natural Language Understanding

“extract information from texts”

# Deep Natural Language Processing

“processing of texts with deep neural networks”



Language is hard.

Let's eat grandpa.

Let's eat grandpa.

Let's eat, grandpa.

## The Pope's Baby Step on Gays

3-4 Minuten



STEFANO SPAZIANI

First, the good news. Pope Francis is already showing himself to be a winsome, endearing and inspiring successor to St. Peter. His trip to Brazil catapulted him to rock-star status, with his care for the poor and the dispossessed, his willingness to engage the throngs with little regard for his security and even with his crowd-pleasing offer of a song on the guitar. This is no formal and aloof bishop but rather a man of and for the people. Justice is on his mind and his lips.

Source NY Time: <http://ti.me/13XReHS>

- Language processing problems are mostly optimisation problems, mostly only approximate solutions exist.
- Different languages differ greatly from each other. Semantic structure, grammar, syntax, word order, text types, stylistics
- Language is ambiguous
- Language is creative
- Language is non-sequential (interactions between widely separated elements in a sentence)
- The context often determines the meaning.
- Further Information: [Dagmar Gromann - Dialog or Dialogue?](#)

# Applications

- Standard Tasks
  - Named Entity Recognition
  - Text Classification
  - Text Summarization
  - Text Generation
  - Question Answering (Text und Multiple Choice)
  - Translation
- newer Tasks
  - Virtual Assi
  - Playing Quiz
  - Dialoge Generation
  - Code Generation
  - HTML Layout Generator
  - Reasoning Questions
  - Fact Checking

# Translations



Übersetze **Englisch** (erkannt) ▼

Übersetze nach **Deutsch** ▼

Anredeform ▼

Glossar

What Donald Trump Liked About Being President

He preferred the parts of the job that combined pomp, splendor and a world amenable to his decisions. In other words, he always seemed to genuinely enjoy pardoning turkeys.

Was Donald Trump am Präsidentenamt schätzte

Er bevorzugte die Teile der Arbeit, die Pomp, Pracht und eine Welt, die seinen Entscheidungen zugänglich ist, miteinander verbinden. Mit anderen Worten, es schien ihm immer wirklich Spaß zu machen, Truthähne zu begnadigen.

>

📄 🔗 ↓

Service: [deepl.com](https://www.deepl.com)



# Question Answering



The 1973 oil crisis began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC, consisting of the Arab members of OPEC plus Egypt and Syria) proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US\$3 per barrel to nearly \$12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "first oil shock", followed by the 1979 oil crisis, termed the "second oil shock."

**When did the 1973 oil crisis begin?**

Ground Truth Answers: October 1973 October 1973 October 1973 October 1973

**What was the price of oil in March of 1974?**

Ground Truth Answers: nearly \$12 \$12 \$12 \$12 \$12

**When was the second oil crisis?**

Ground Truth Answers: 1979 1979 1979 1979 1979

**What was another term used for the oil crisis?**

Ground Truth Answers: first oil shock shock shock first oil shock shock

**Who proclaimed the oil embargo?**

Ground Truth Answers: members of the Organization of Arab Petroleum Exporting Countries members of the Organization of Arab Petroleum Exporting Countries Organization of Arab Petroleum Exporting Countries members of the Organization of Arab Petroleum Exporting Countries OAPEC

# Question Answering



- Stanford Question Answering Dataset (SQuAD)
- Standard Datat set for QA
- consists of 150000 questions
- 44 MB of data
- <https://rajpurkar.github.io/SQuAD-explorer/>

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452

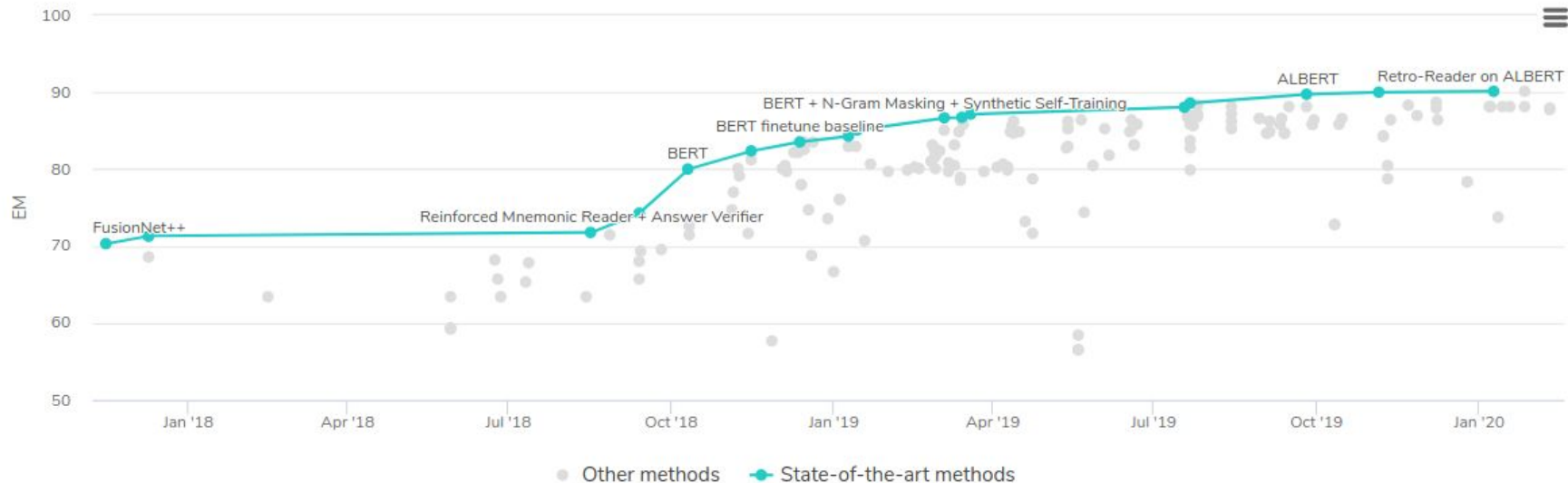
# Question Answering



- Stanford Question Answering Dataset (SQuAD)
- Standard Datat set for QA
- consists of 150000 questions
- 44 MB of data
- <https://rajpurkar.github.io/SQuAD-explorer/>

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) QIANXIN	90.724	93.011
2 May 05, 2020	SA-Net-V2 (ensemble) QIANXIN	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) Shanghai Jiao Tong University <a href="http://arxiv.org/abs/2001.09694">http://arxiv.org/abs/2001.09694</a>	90.578	92.978
3 Jul 31, 2020	ATRLP+PV (ensemble) Hithink RoyalFlush	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) SRCB_DML	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) SRCB_DML	90.420	92.799

# Question Answering on SQuAD 2.0



# Transformer Quiz



question 2

Which 20th century British Prime Minister was born in Portsmouth?

You	winston churchill	NOPE :(	1
T5	James CALLAGHAN	CORRECT!	2

NEXT QUESTION

Correct answer: james callaghan

► See all accepted answers

## Input

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

## Model Output

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

[...]

# Text to Shell



Playground 



Text to command



Please provide a simple natural language description of the give shell command.

Q: `cd ~`

A: go to the home directory

Q: `grep -r "pepsi" *`

A: find occurrences of the string "pepsi" in every file in the current directly recursively

Q: `python -m SimpleHTTPServer`

A: start a simple web server with python

Q: `ls -l -i :8000`

A: get information about the process using port 8000

Q: `ps ax`

A: show all processes

Q: `npm install eslint --save-dev`

A: install the `eslint` package as a development dependency

Q: `git checkout feature-branch-3`

A: switch to the branch feature-branch-3

Q: `rm -rf *`

A: remove all files in the current directory recursively

Q: `head -n 3 testing.txt`

A: print the first three lines of the file testing.txt

SUBMIT 



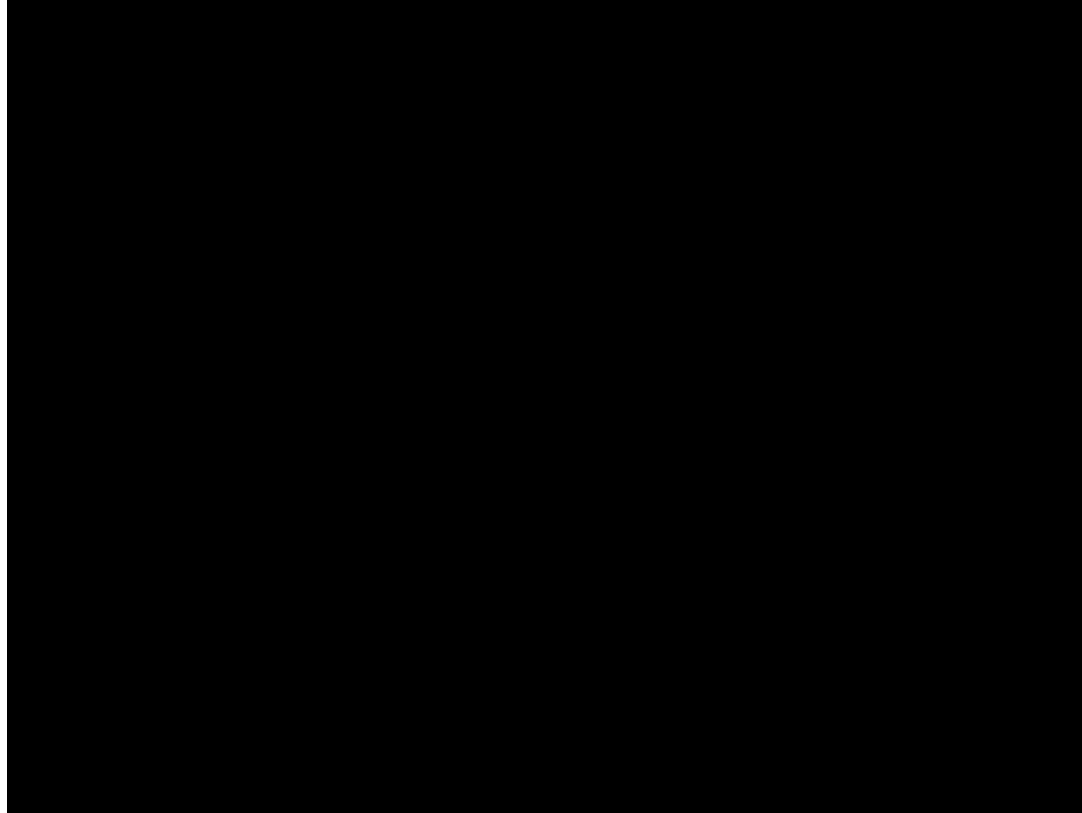
Inject structural text (start sequence, reset sequence)



Quelle Harlan Duman

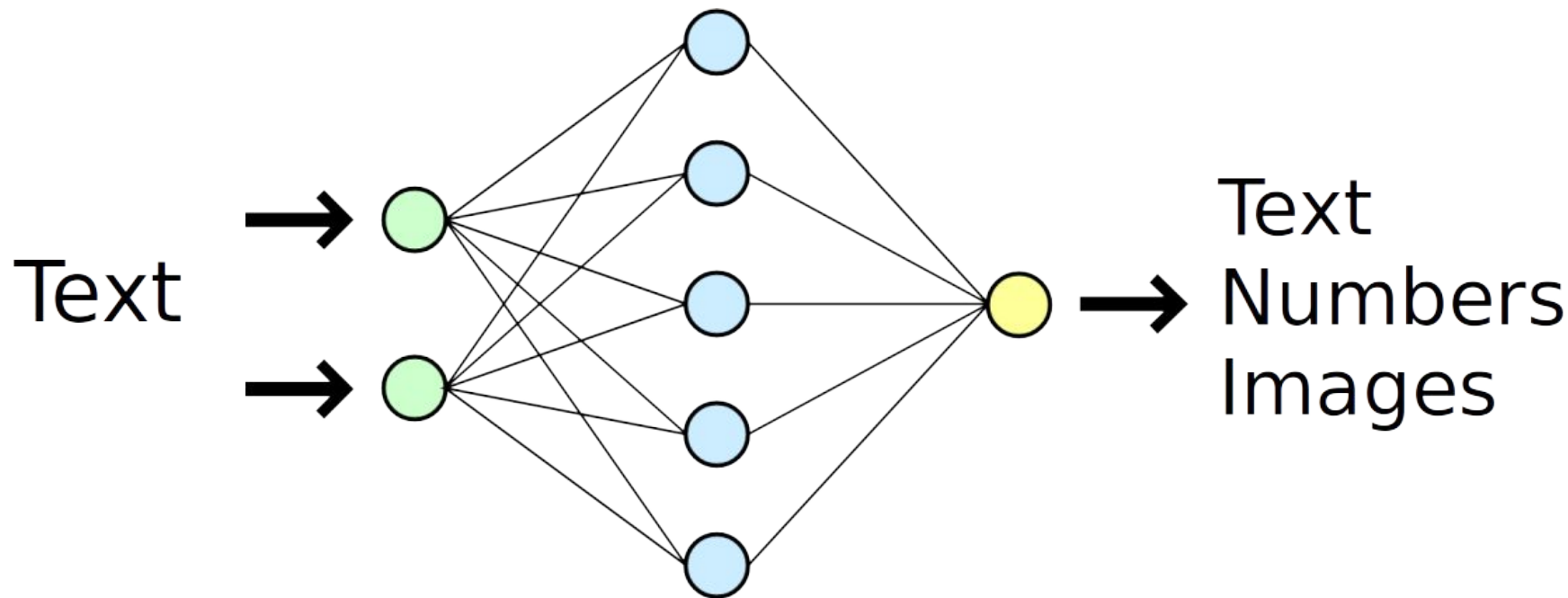


# Text to HTML



[Quelle](#) / Produkt [debuild.co](#)

# How do neural networks process texts?



**How do we encode characters  
for a neural network?**

# One Hot Encoding



Lets encode the word „hello“

h	0	0	0	1
e	0	0	1	0
l	0	1	0	0
o	1	0	0	0

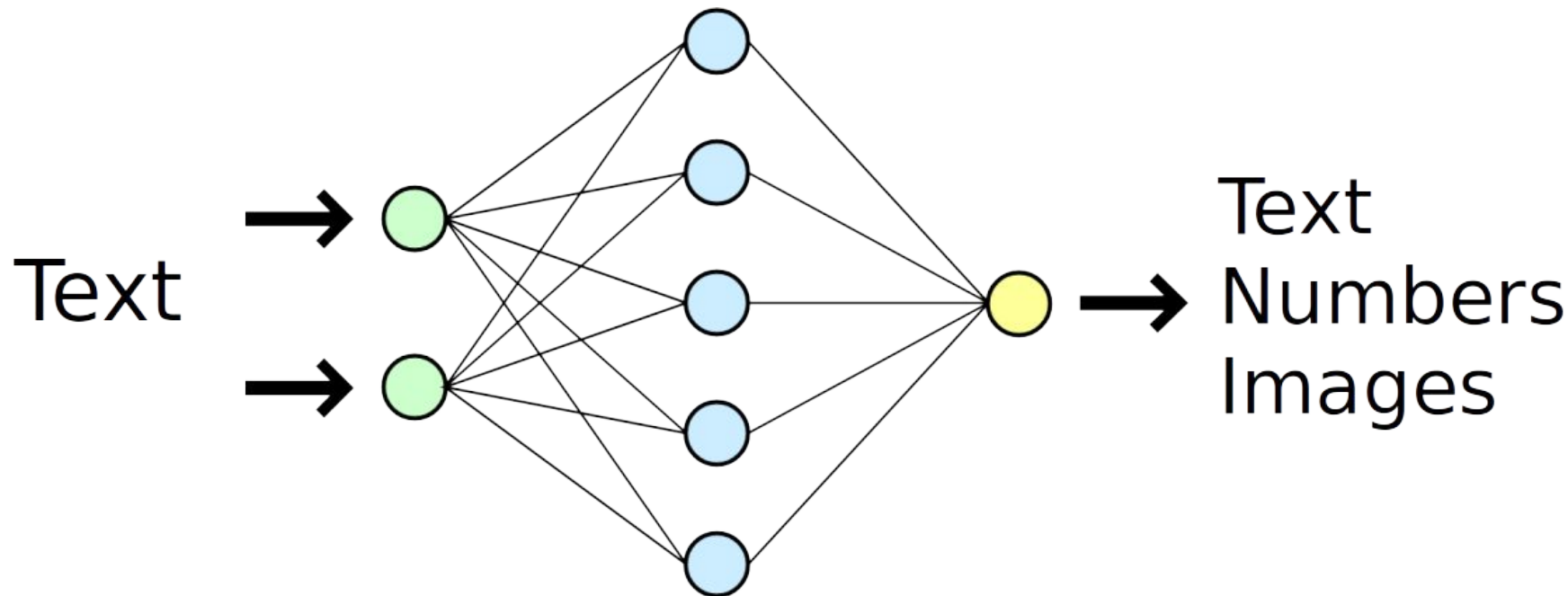
# One Hot Encoding

Let's encode the word „hello“

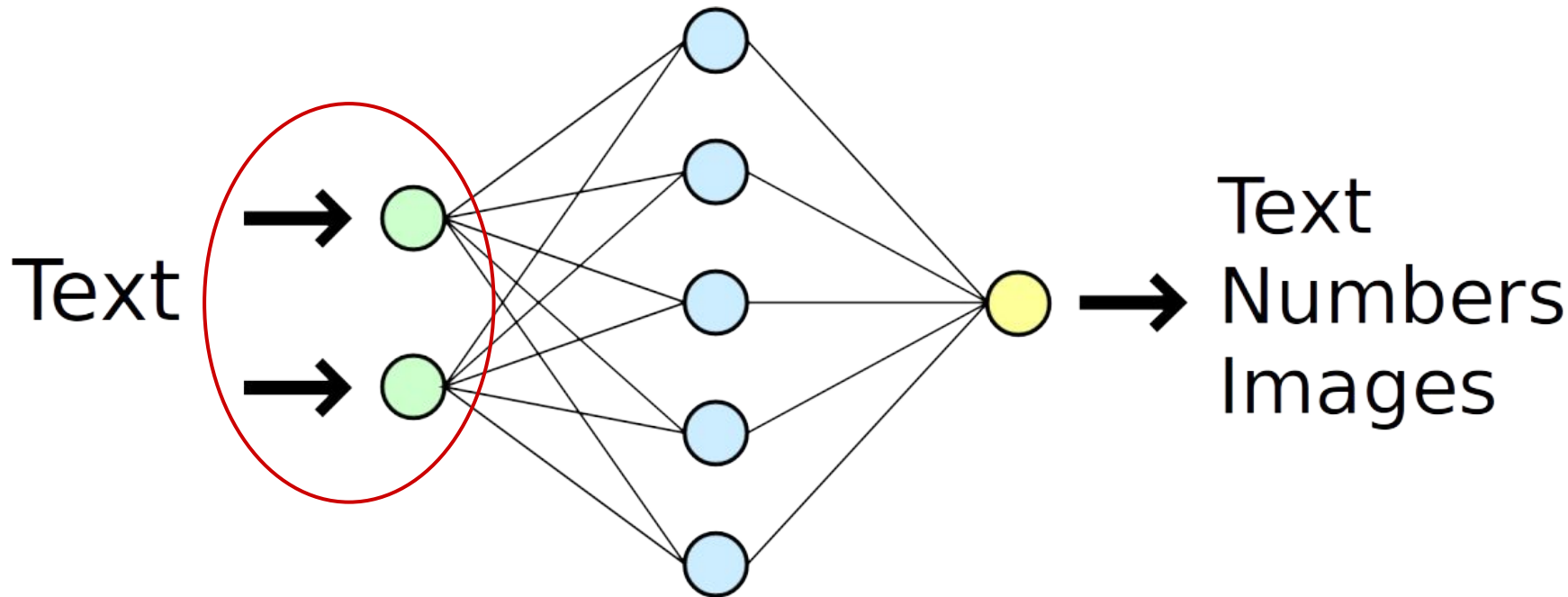
h	0	0	0	1
e	0	0	1	0
l	0	1	0	0
o	1	0	0	0

$$v^h = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad v^e = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \dots \quad \longrightarrow \quad V^{hello} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Can you spot the issue?



Can you spot the issue?





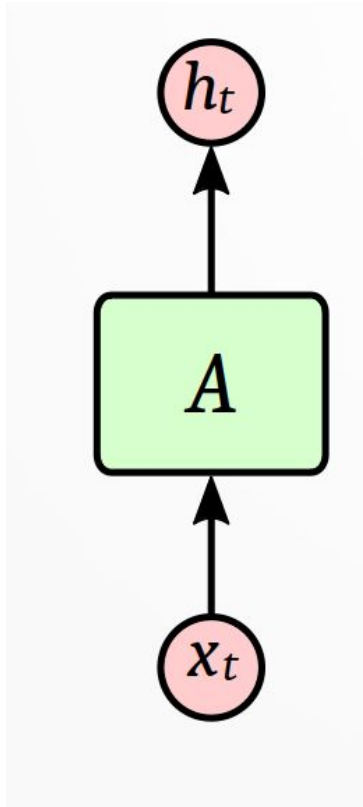
**We need to handle inputs of arbitrary length.**

# Recurrent Neural Networks

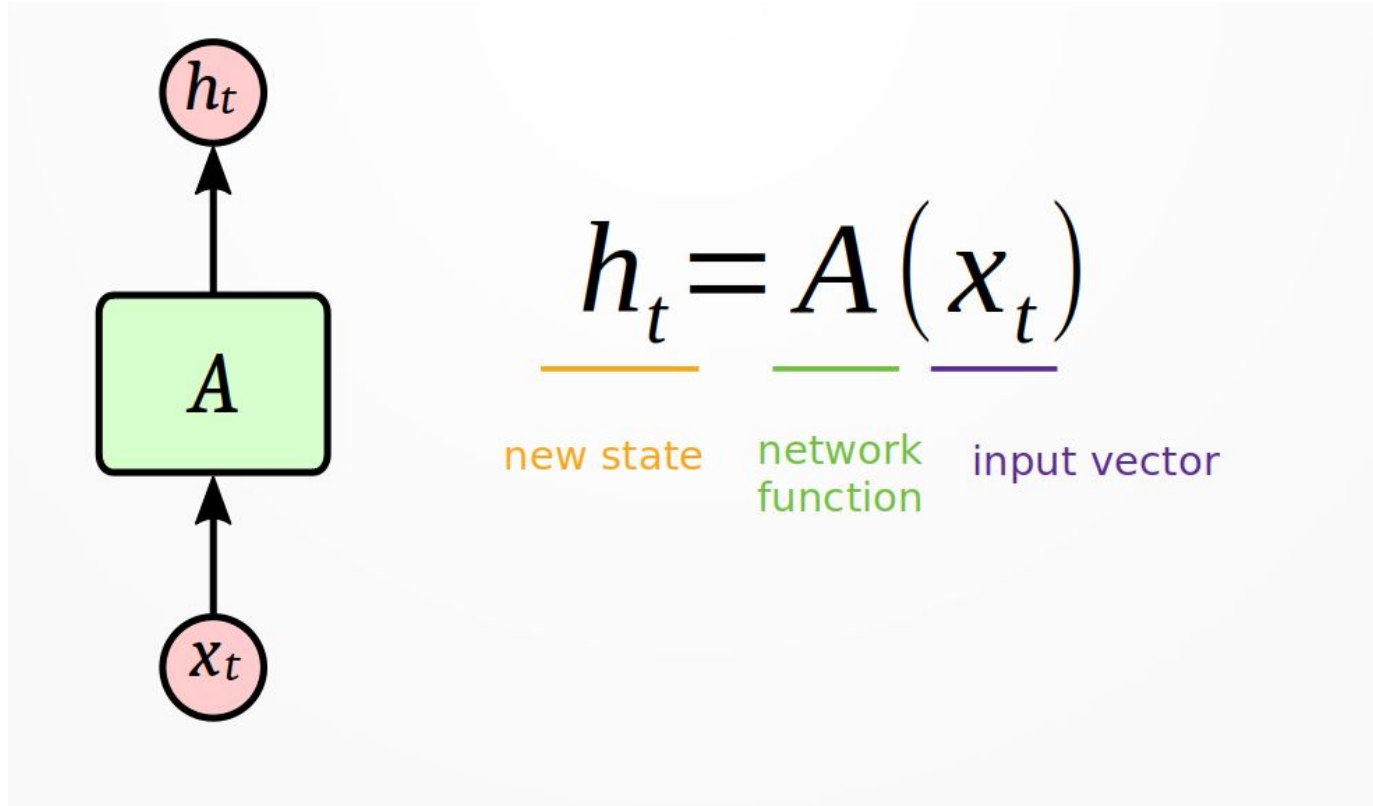
RNNs are networks with loops, allowing information to persist.

[Rummelhart et al. 1986]

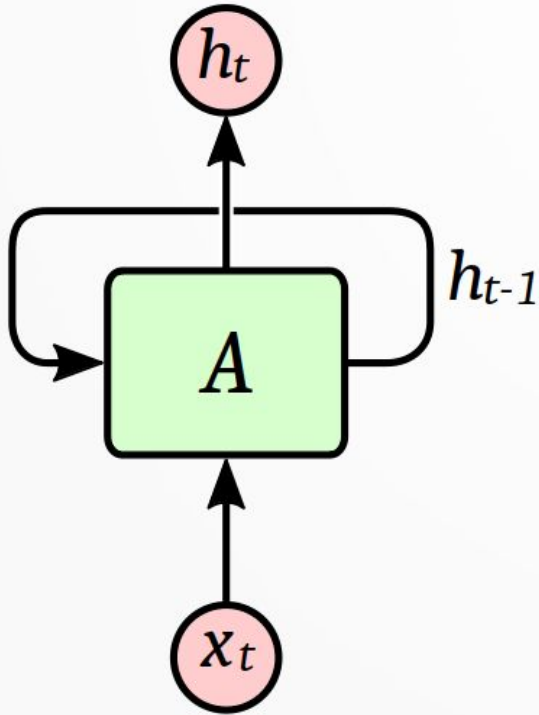
# A simple neural network



# A simple neural network

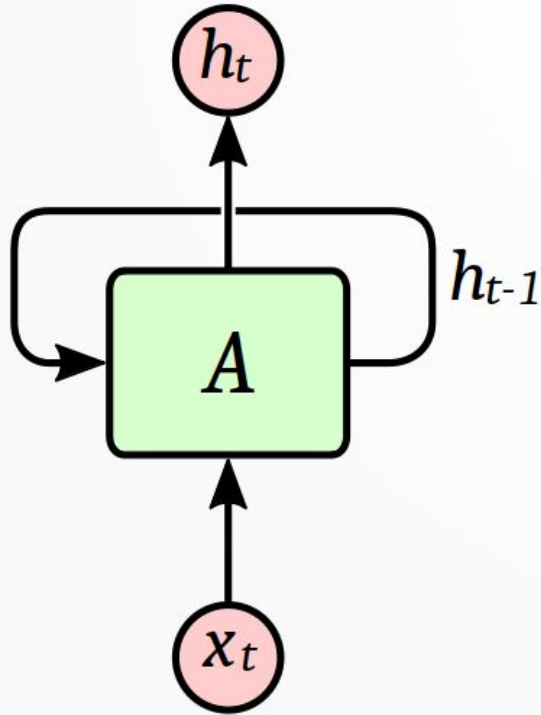


# Recurrent Neural Networks



By passing the previous hidden state, the network can keep an „internal state“ as the input sequence is processed.

# Recurrent Neural Networks

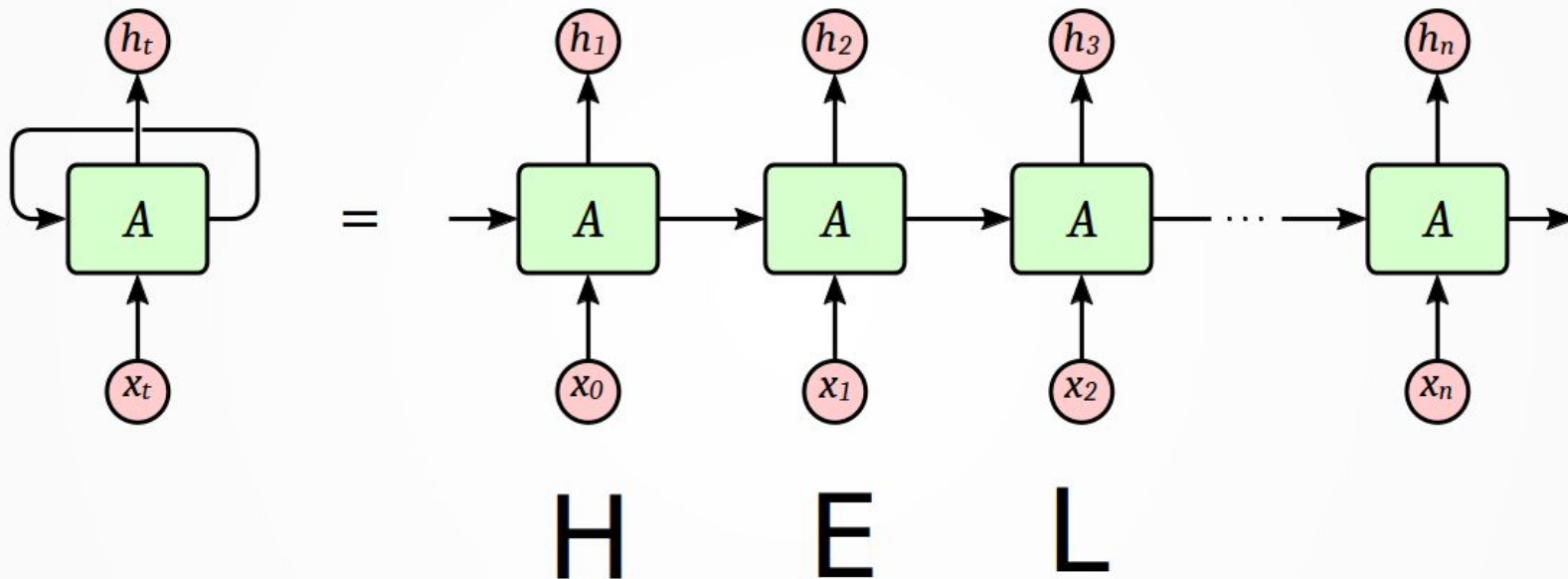


$$\underline{h_t} = \underline{A}(\underline{h_{t-1}}, \underline{x_t})$$

new state      network function      previous state      input vector

**How do we use this network  
to process text?**

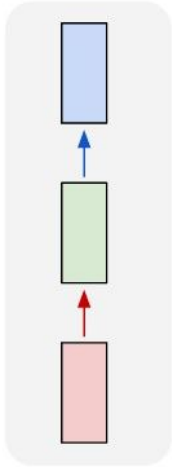
# Unfolding in Time





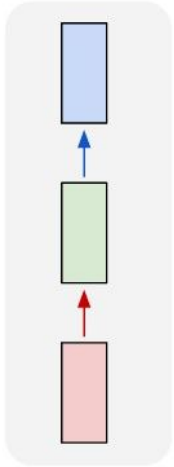
# Network Architectures

one to one

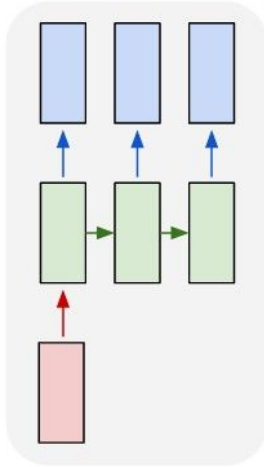


# Simple Neural Network

one to one



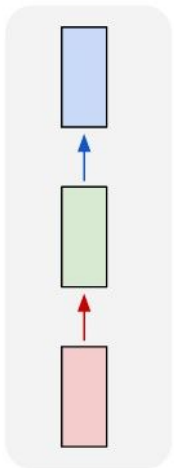
one to many



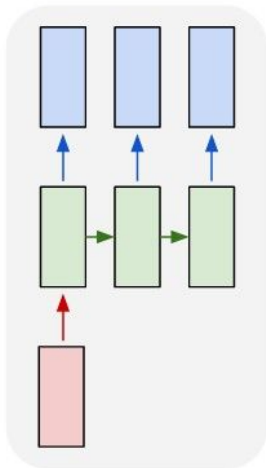
## **image captioning**

Image to a sequence of words

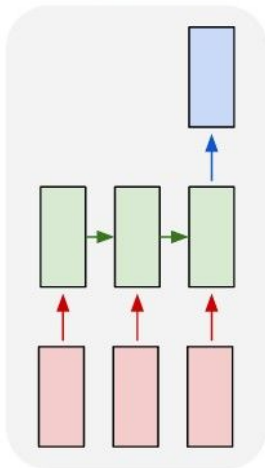
one to one



one to many



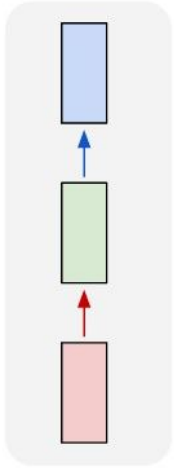
many to one



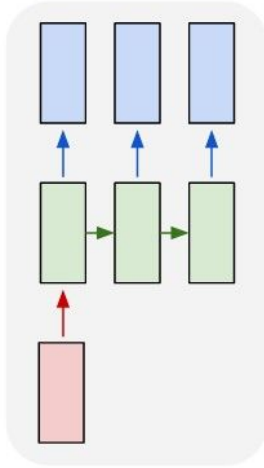
## **classification**

sequence of words to a class

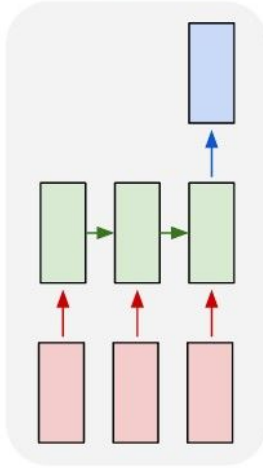
one to one



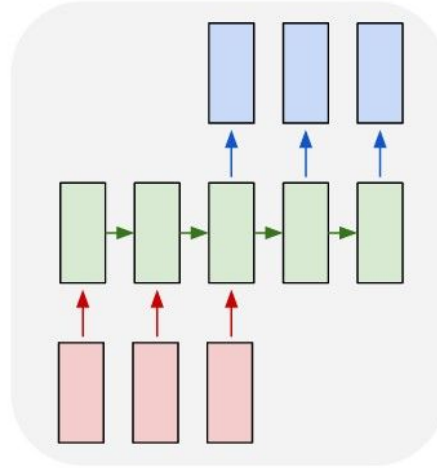
one to many



many to one



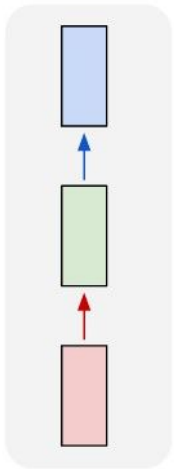
many to many



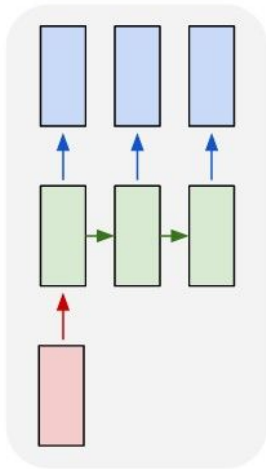
## **machine translation**

sequence of words to sequence of words

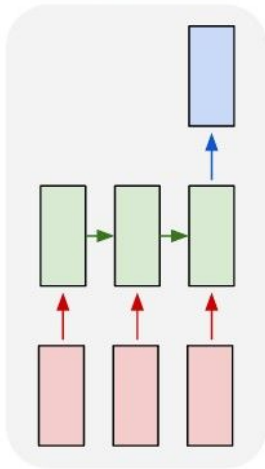
one to one



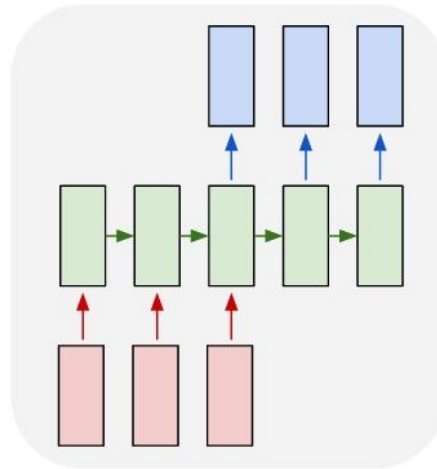
one to many



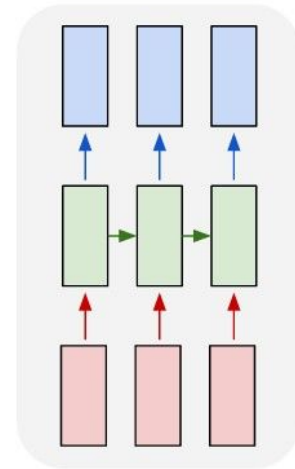
many to one



many to many



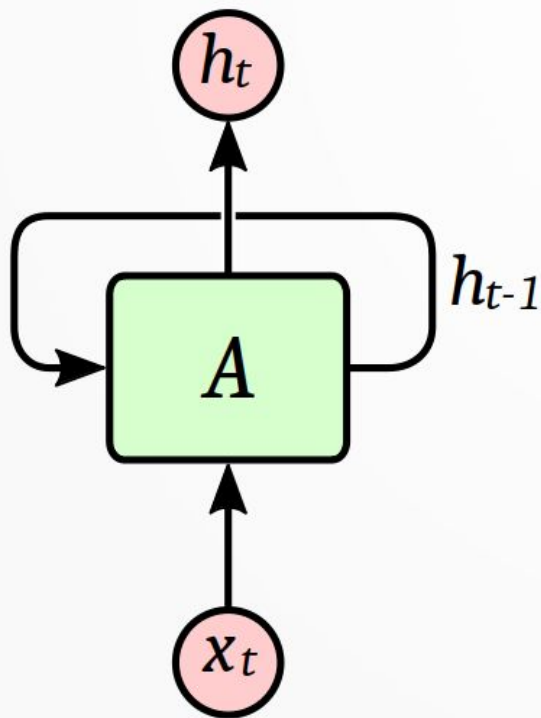
many to many



## Video classification

A list of frames to a list of classes

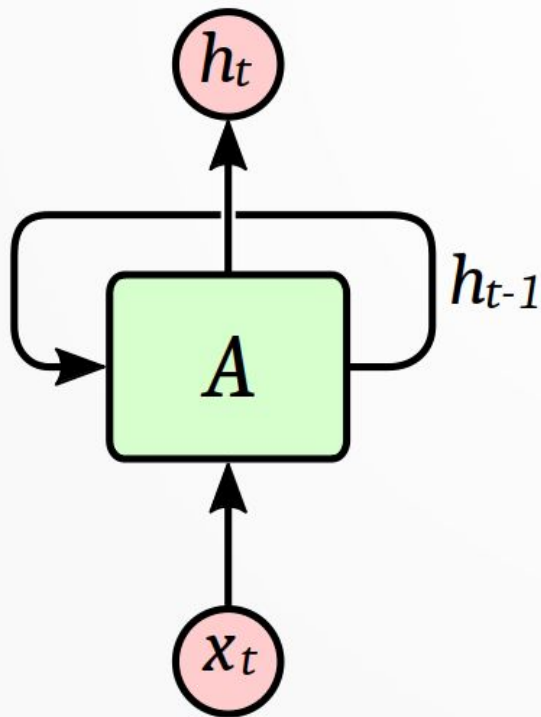
# Training RNNs



How can we apply  
backpropagation  
to nets  
with loops?

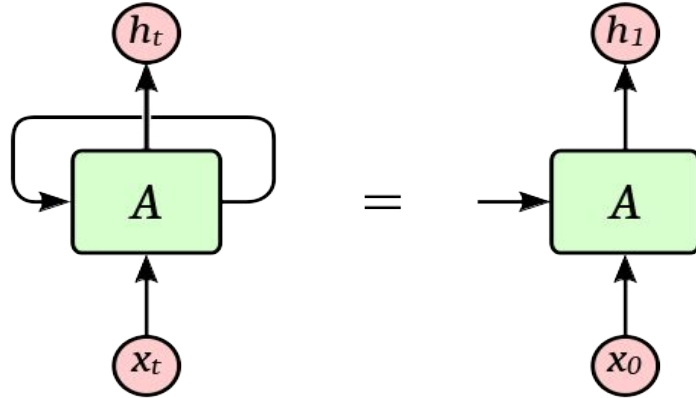


# Training RNNs

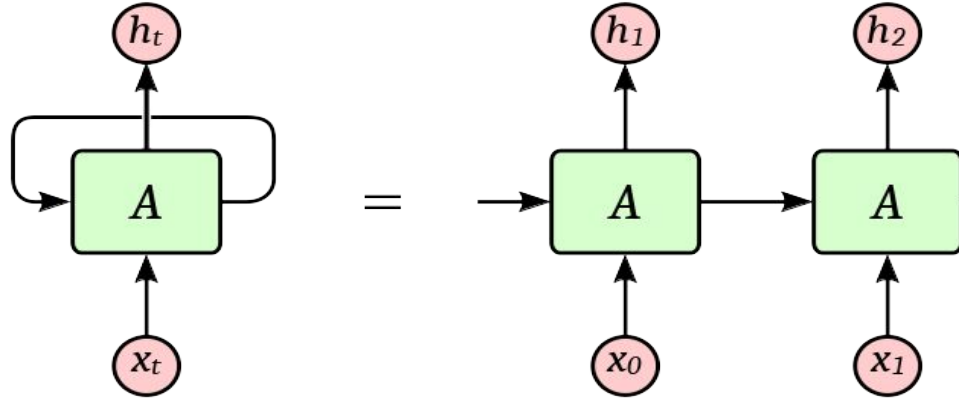


with backpropagation  
through time (BPTT)

## Training RNNs



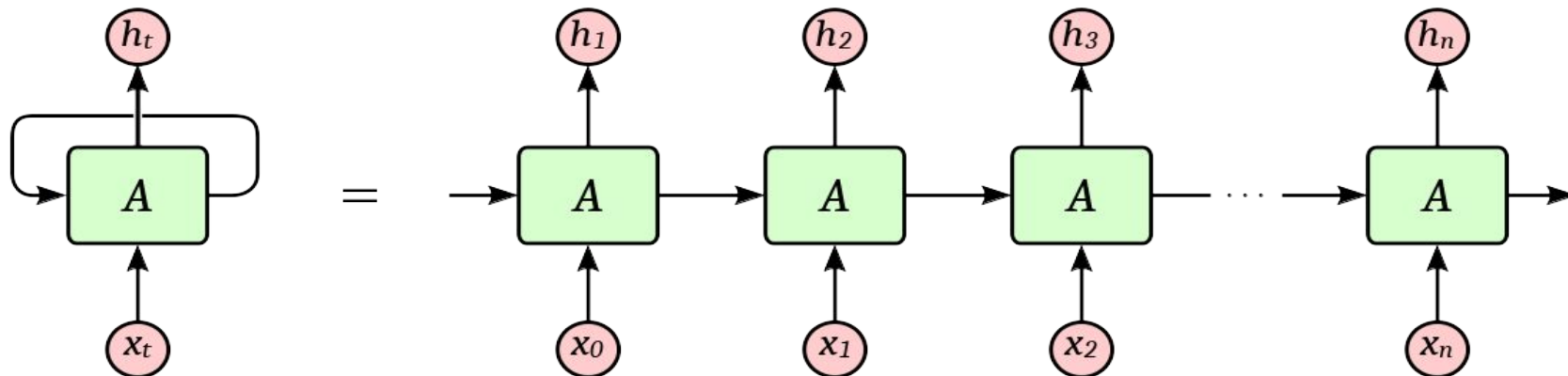
# Training RNNs



## Training RNNs



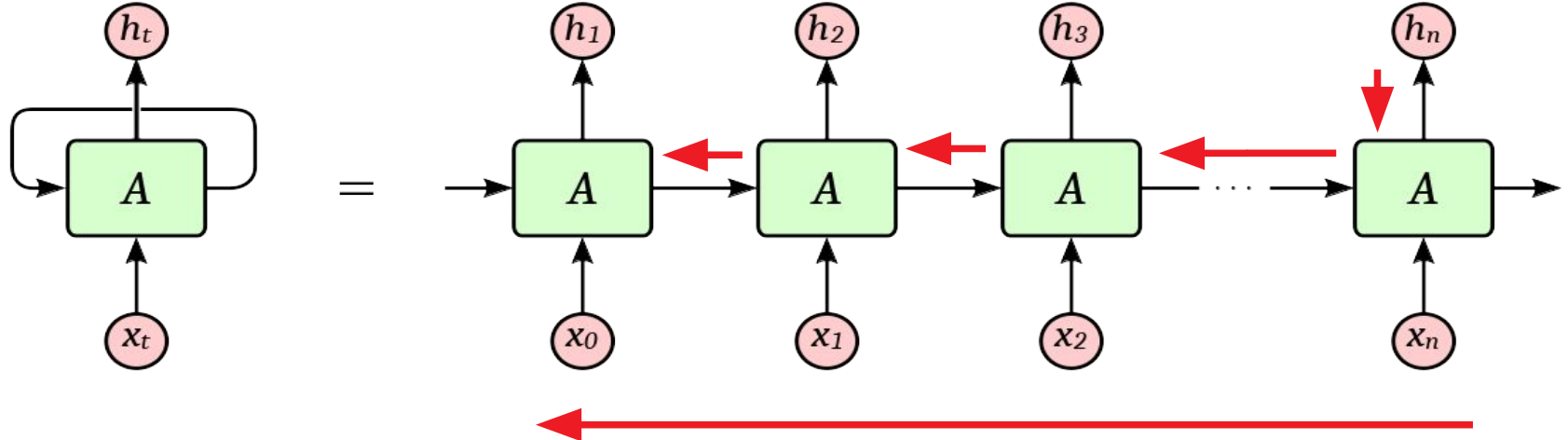
1. forward-propagate the inputs over the unfolded network



## Training RNNs



1. forward-propagate the inputs over the unfolded network

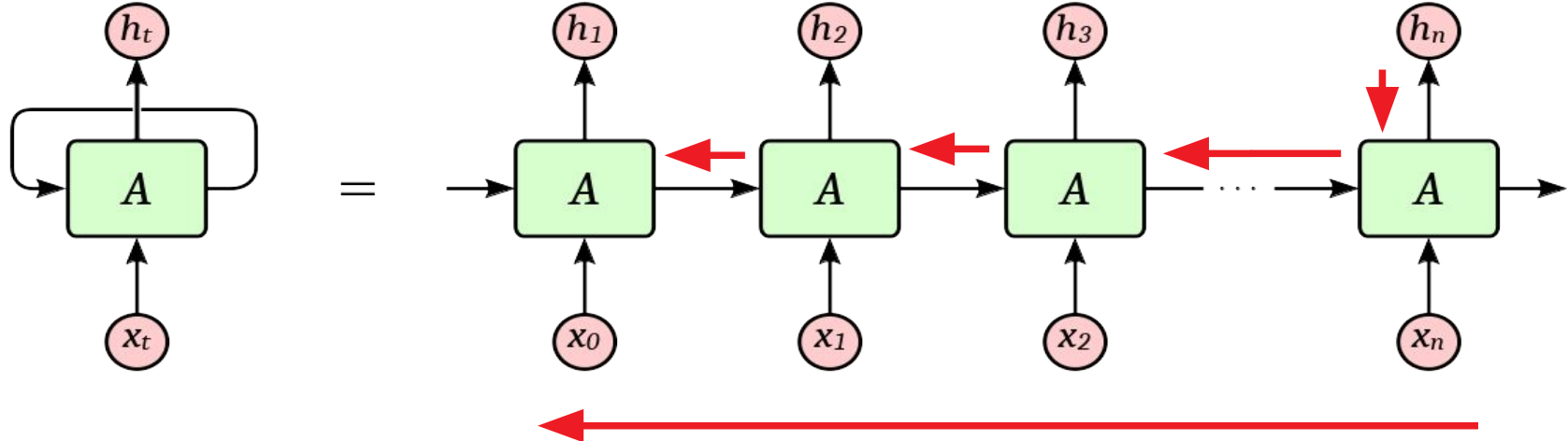


2. back-propagate the error, back across the unfolded network

## Training RNNs



1. forward-propagate the inputs over the unfolded network



2. back-propagate the error, back across the unfolded network

3. sum the weight changes and update all weights

# Further Information



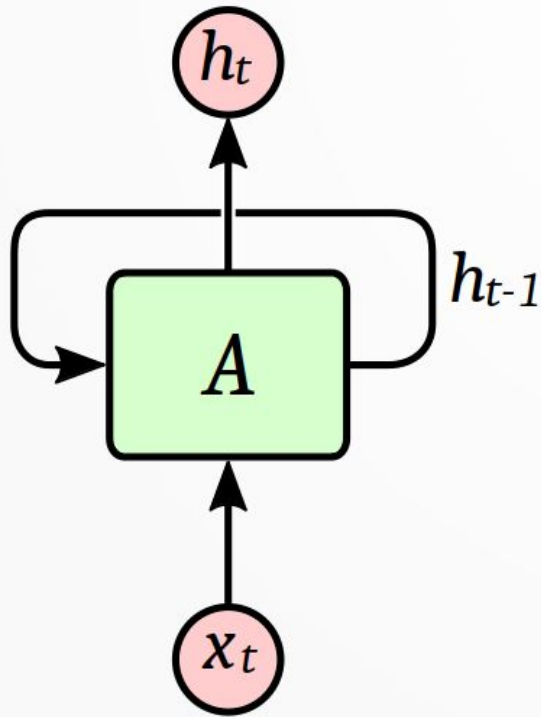
- A well explained implementation of BPTT can be found [here](#)
- [Andrew Ng explaining BPTT](#)

# Learning long-term dependencies with gradient descent is difficult

Y. Bengio, P. Simard and P. Frasconi in IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157-166, March 1994.



# Recurrent Neural Networks



$$\underline{h_t} = \underline{A}(\underline{h_{t-1}}, \underline{x_t})$$

new state      network function      previous state      input vector

# Recurrent Neural Networks



$$\underline{h_t} = \underline{A} \left( \underline{h_{t-1}}, \underline{x_t} \right)$$

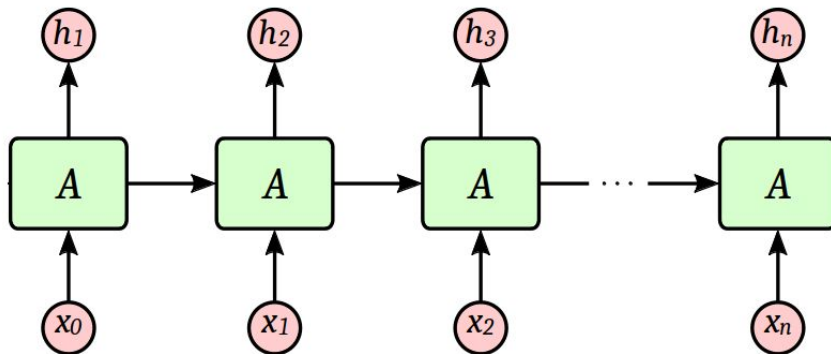
new state      network function      previous state      input vector

A blue curved arrow points from the general equation  $h_t = A(h_{t-1}, x_t)$  down to the specific Elman network equation  $h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$ .

A simple example of a Elman-network

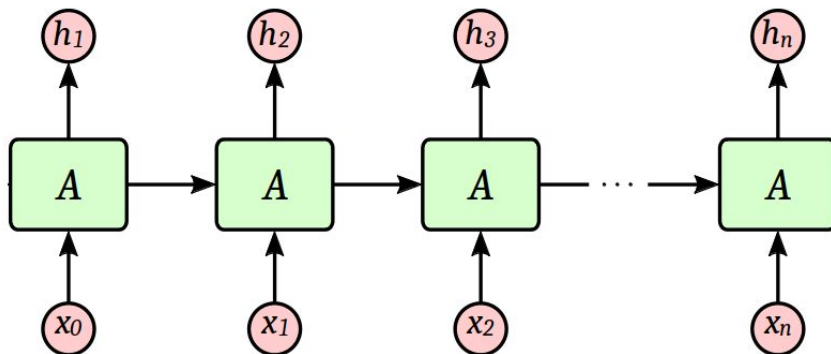
$$\underline{h_t} = \underline{\tanh} \left( \underline{W_{hh} h_{t-1} + W_{xh} x_t} \right)$$
$$y_t = W_{yh} h_t$$

# The vanishing gradient problem



$$h_1 = \tanh(W_{hh}h_0 + W_{xh}x_1)$$

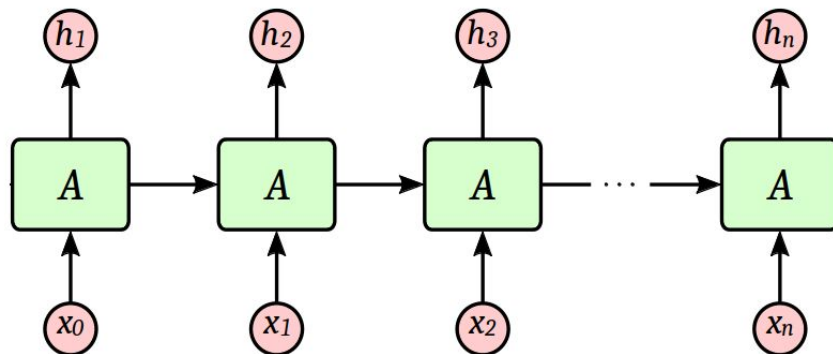
# The vanishing gradient problem



$$h_1 = \tanh(W_{hh} h_0 + W_{xh} x_1)$$

$$h_2 = \tanh(W_{hh} (\tanh(W_{hh} h_0 + W_{xh} x_1)) + W_{xh} x_2)$$

# The vanishing gradient problem

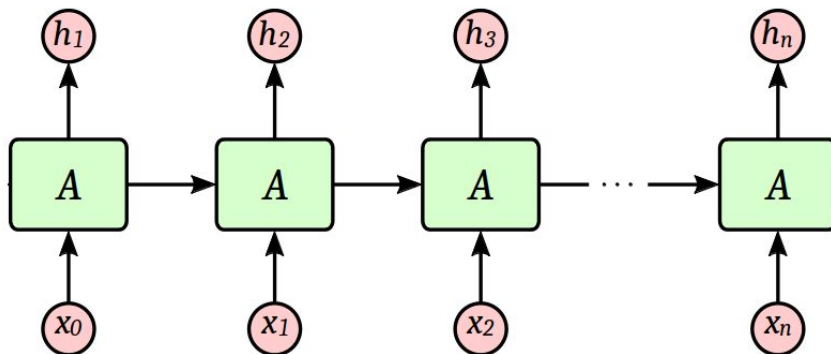


$$h_1 = \tanh(W_{hh} h_0 + W_{xh} x_1)$$

$$h_2 = \tanh(W_{hh} (\tanh(W_{hh} h_0 + W_{xh} x_1)) + W_{xh} x_2)$$

$$h_3 = \tanh(W_{hh} (\tanh(W_{hh} (\tanh(W_{hh} h_0 + W_{xh} x_1)) + W_{xh} x_2)) + W_{xh} x_3)$$

# The vanishing gradient problem



$$h_1 = \tanh(W_{hh} h_0 + W_{xh} x_1)$$

$$h_2 = \tanh(W_{hh}(\tanh(W_{hh} h_0 + W_{xh} x_1)) + W_{xh} x_2)$$

$$h_3 = \tanh(W_{hh}(\tanh(W_{hh}(\tanh(W_{hh} h_0 + W_{xh} x_1)) + W_{xh} x_2)) + W_{xh} x_3)$$

$$h_4 = \tanh(W_{hh}(\tanh(W_{hh}(\tanh(W_{hh}(\tanh(W_{hh} h_0 + W_{xh} x_1)) + W_{xh} x_2)) + W_{xh} x_3)))$$



Backpropagating this recursive function leads to exploding or vanishing gradients.

# Papers



On the difficulty of training recurrent neural networks

Pascanu, Mikolov and Bengio, 2013

<http://proceedings.mlr.press/v28/pascanu13.pdf>

Learning long-term dependencies with gradient descent is difficult

Bengio, Simard and Frasconi, 1994

<https://ieeexplore.ieee.org/document/279181>

Untersuchungen zu dynamischen neuronalen Netzen

Hochreiter, 1991

<http://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf>

# Solutions for this problem:

- Limiting the number of past time steps (Hochreiter, 1991)
- Exploding gradient can be fixed with gradient clipping
- Vanishing gradients can be controlled different architectures (LSTM)
- New: Not using recursions :)

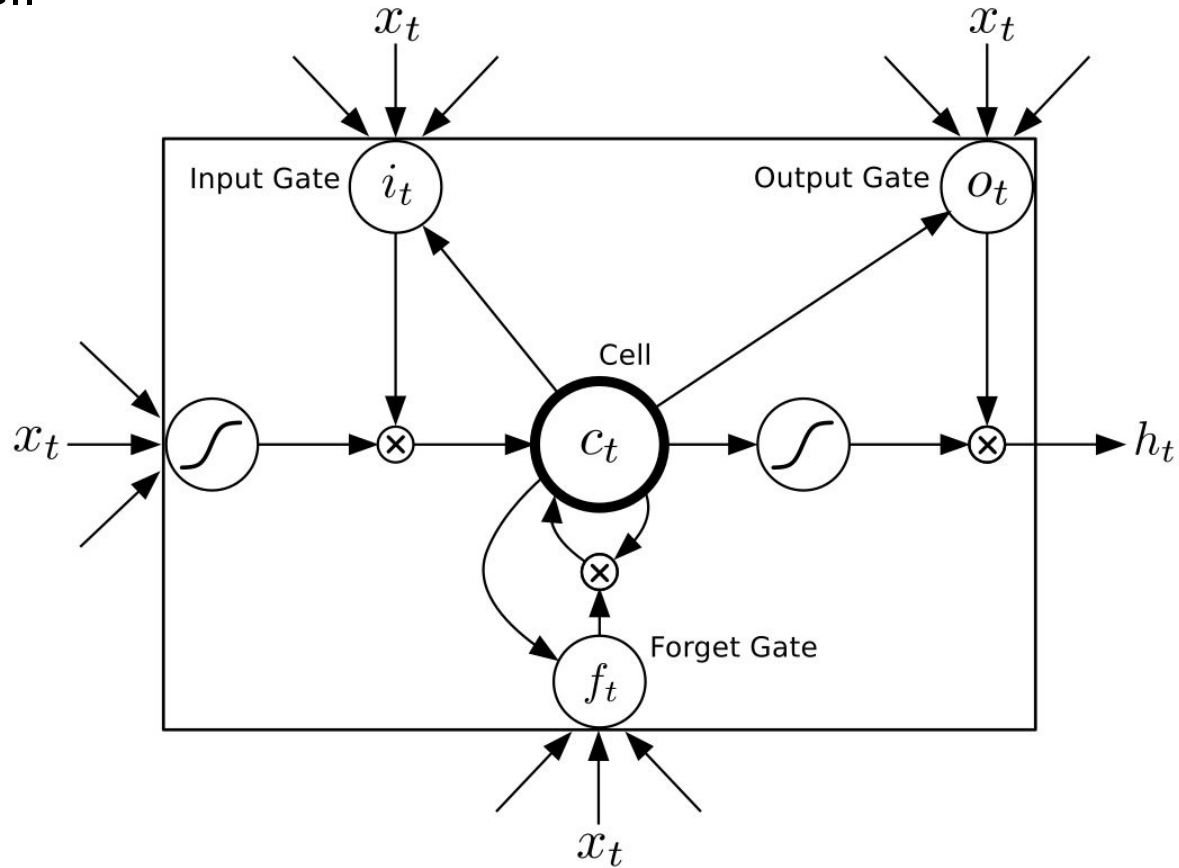


# **LSTM**

## **Long short-term memory**

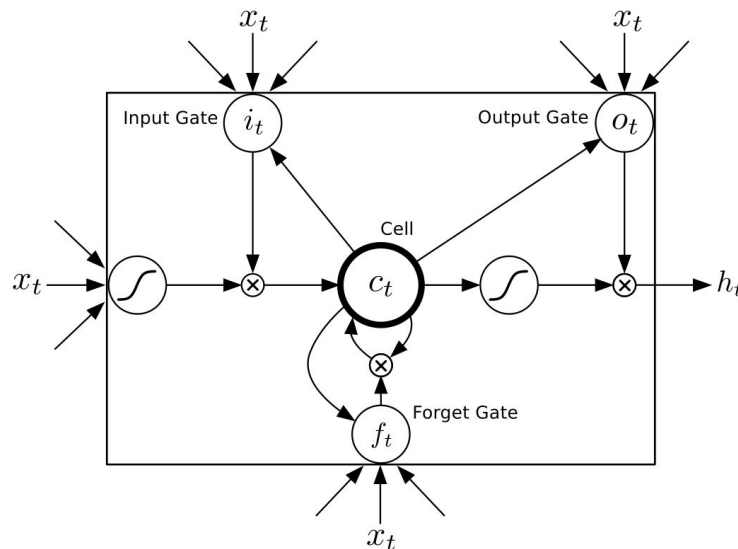
**[Hochreiter et al., 1997]**

# A LSTM Cell



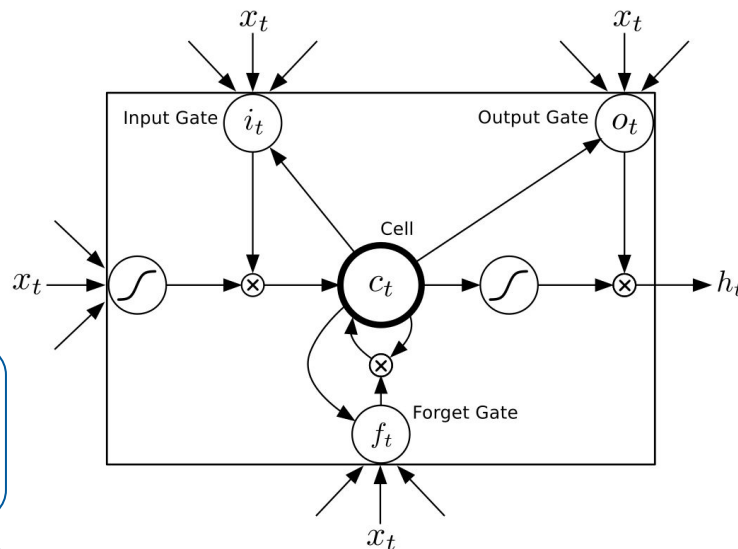
Graves et al. 2013, Speech Recognition with Deep Recurrent Neural Networks

# A LSTM Cell



$$\begin{aligned}i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\h_t &= o_t \tanh(c_t)\end{aligned}$$

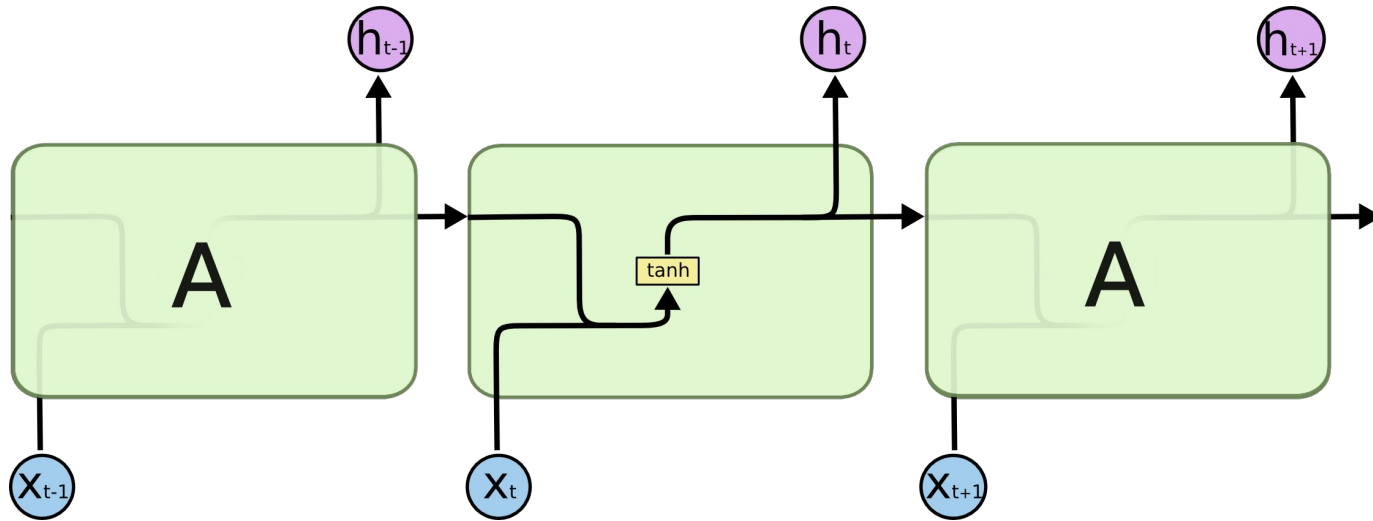
# A LSTM Cell



The sigmoid function outputs a number between 0 and 1

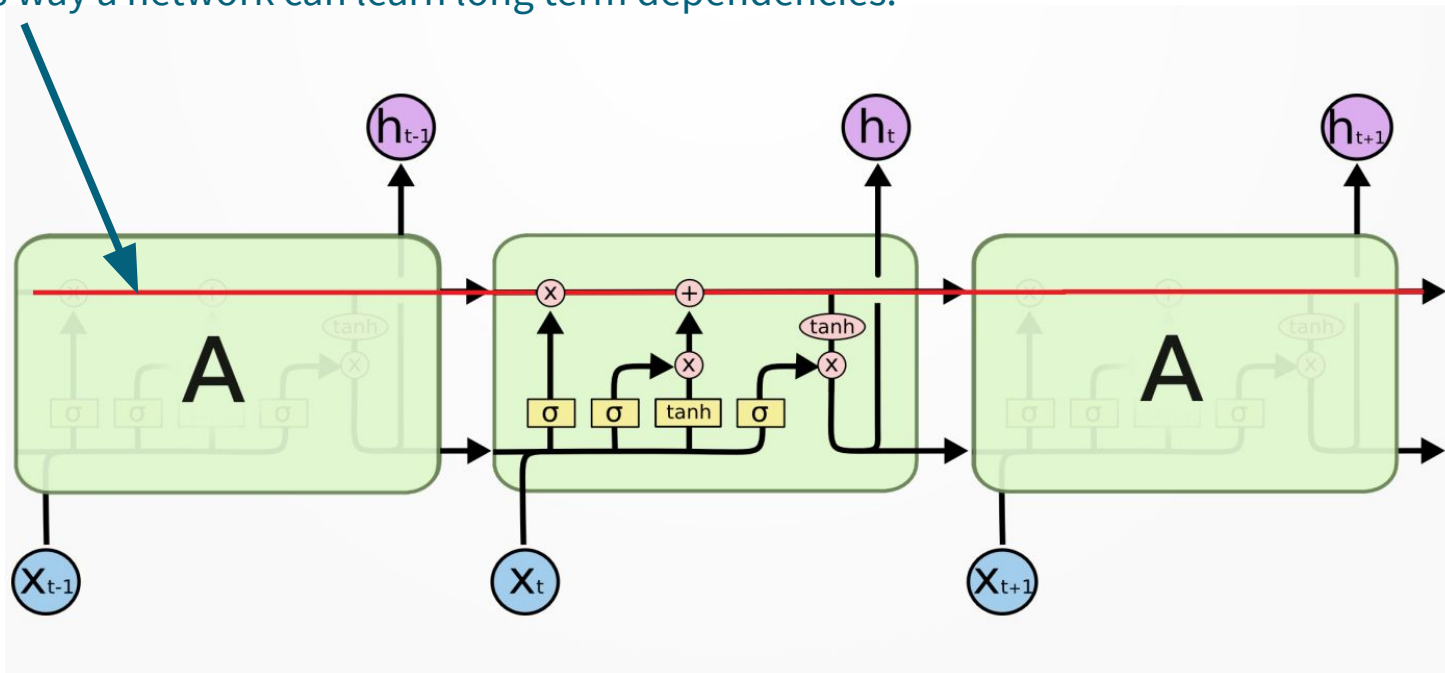
$$\begin{aligned}i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\h_t &= o_t \tanh(c_t)\end{aligned}$$

# Standard RNN



# LSTM

A LSTM cell can keep its internal state unchanged over many timesteps.  
This way a network can learn long term dependencies.



- Chris Olah: [Understanding LSTM Networks](#)
- Jürgen Schmidhuber: [Tutorial on LSTM Recurrent Networks](#)
- **LSTM: A Search Space Odyssey**  
Klaus Greff, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, Jürgen Schmidhuber 2015
- **Speech Recognition with Deep Recurrent Neural Networks**  
Graves et al. 2013
- **Long Short-term Memory**  
Sepp Hochreiter, Jürgen Schmidhuber

# **What does the network learn?**



# Visualizing the cell activations



Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

# Visualizing the cell activations



Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

# Visualizing the cell activations



Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
                           siginfo_t *info)
{
    int sig = next_signal(pending, mask);
    if (sig) {
        if (current->notifier) {
            if (sigismember(current->notifier_mask, sig)) {
                if (!current->notifier(current->notifier_data)) {
                    clear_thread_flag(TIF_SIGPENDING);
                    return 0;
                }
            }
        }
        collect_signal(sig, pending, info);
    }
    return sig;
}
```

# Visualizing the cell activations



A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
    char *str;
    if (!*bufp || (len == 0) || (len > *remain))
        return ERR_PTR(-EINVAL);
    /* Of the currently implemented string fields, PATH_MAX
     * defines the longest valid length.
     */
}
```

# What does the network learn?



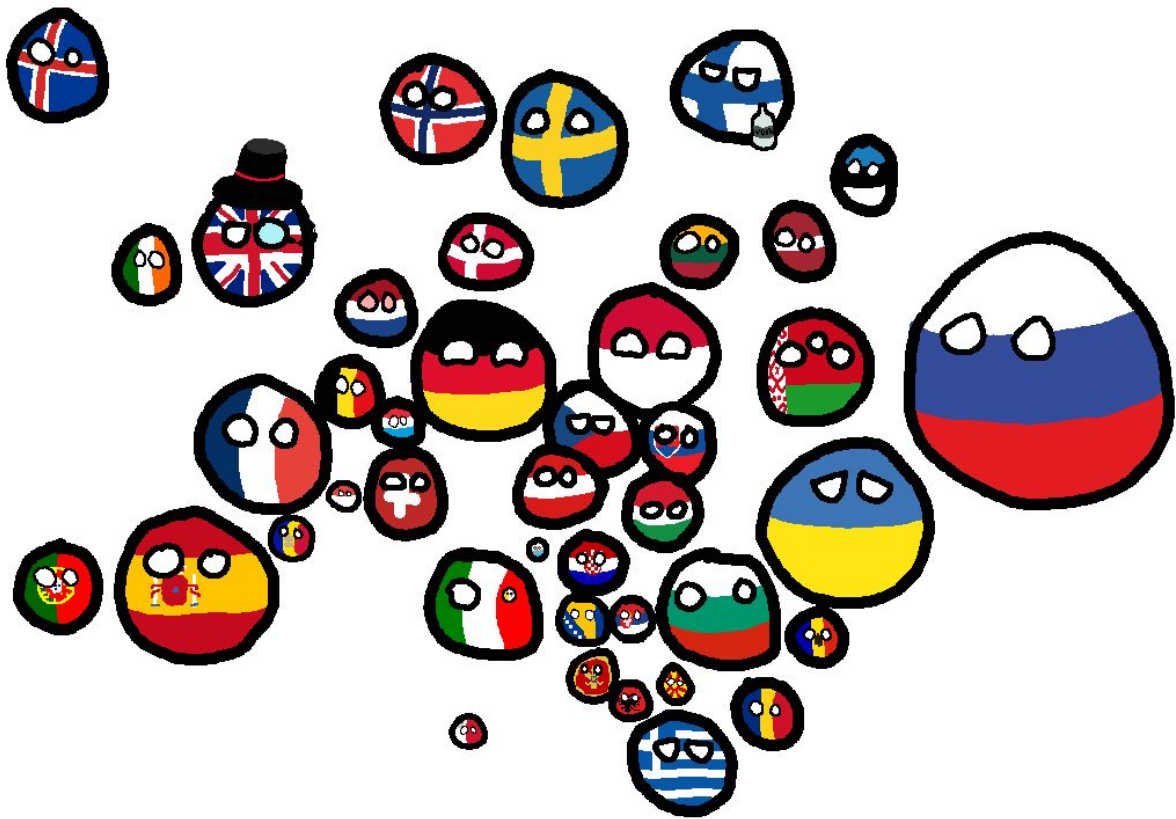
- Visualizing the predictions and the “neuron” firings in the RNN
- Set the background color based on the neurons activation
- Literature:
  - <http://karpathy.github.io/2015/05/21/rnn-effectiveness>
- Paper:
  - Karpathy, Johnson, and Fei-Fei: Visualizing and Understanding Recurrent Networks, ICLR Workshop 2016

- Time Series Prediction
- Speech Recognition
  - Speech Recognition with Deep Recurrent Neural Networks Graves et al. 2013
- Drawing (Pictures, handwriting)
  - Generating Sequences With Recurrent Neural Networks Graves 2013
- Music Generation
  - Song From PI: A Musically Plausible Network for Pop Music Generation

- RNNs are used to process sequential data
- They have applications in many other domains:
  - Speech Recognition
  - Time Series Prediction
  - Drawing (Pictures, Handwriting)
- During Backpropagation gradients can explode or vanish
- Processing sequential information is a hot topic of research (Bert, GTP-3)

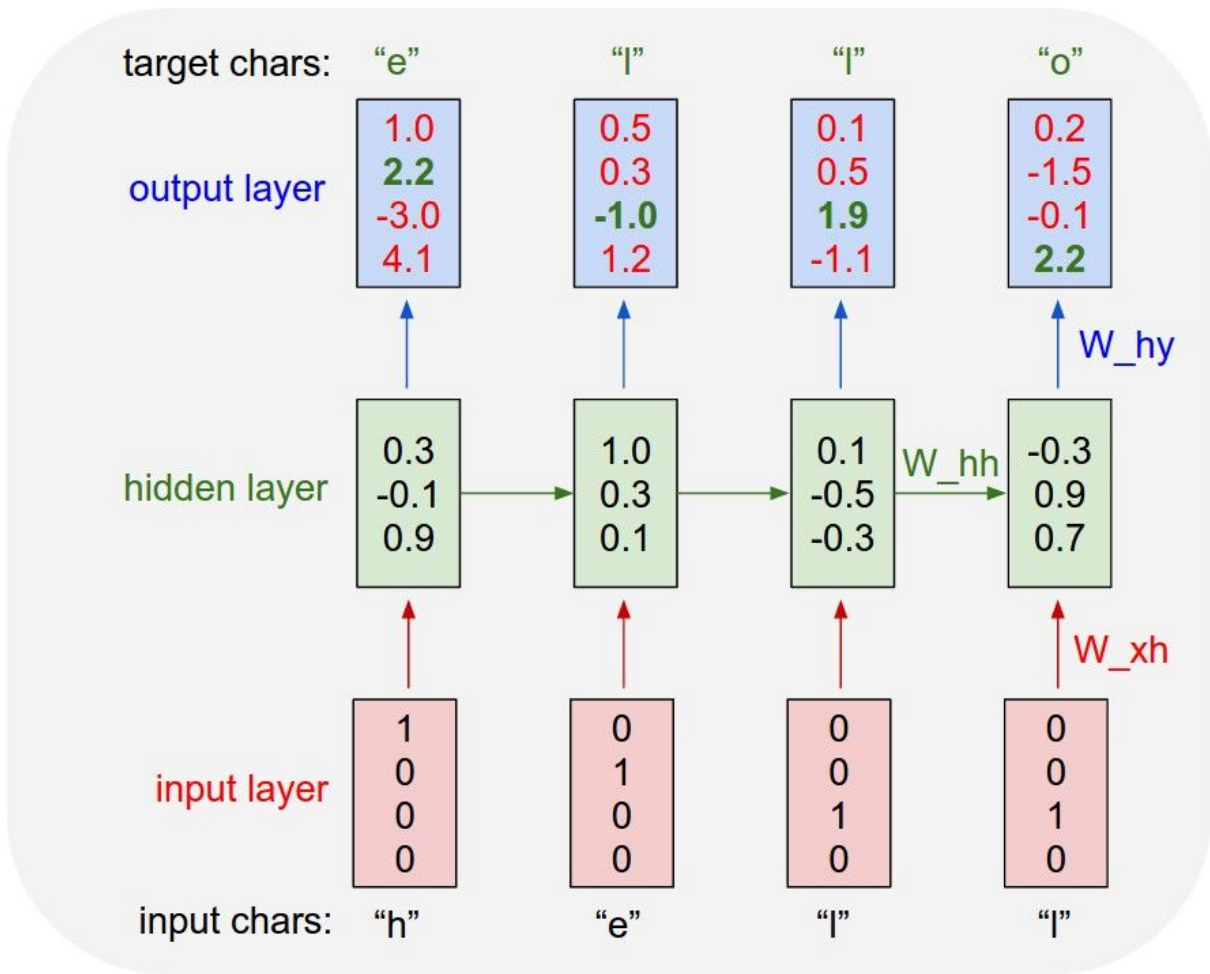
# Tutorial

Guess where is  
my name from?





- You will classify names using a character-level RNN
- Take a look at the [notebook](#) for your tasks
- What you will learn:
  - One Hot Encoding
  - Character level RNNs
  - How to use LSTM and GRUs



# Recurrent Neural Networks



$$\underbrace{h_t}_{\text{new state}} = \underbrace{A}_{\text{network function}} \left( \underbrace{h_{t-1}}_{\text{previous state}}, \underbrace{x_t}_{\text{input vector}} \right)$$

A simple example of a Elman-network

$$\underbrace{h_t}_{\text{new state}} = \tanh \left( \underbrace{W_{hh} h_{t-1} + W_{xh} x_t}_{\text{network function}} \right)$$
$$y_t = W_{yh} h_t$$