Machine Learning for Large-Scale Data Analysis and Decision Making (MATH80629A)
Fall 2021

Week #8 - **Summary**

Announcement

- **Next Class:** Project meeting (10-12 min each team)
- **Study plan**: due **October 18, 2021**
- **Homework 1:** due **November 5, 2021**

Today

- **Fourth Quiz** on Gradescope!
- Summary of Recurrent Neural networks and Convolutional neural networks
- Q&A
- Hands-on session

Quiz 3

Login to your Gradescope account

RNNs

Temporal dependencies

Analyzing temporal dependencies ⇒ Improved decisions

Frame 0	Frame 1	Frame 2	Frame 3	Frame 4
Stem: seen Petals: hidden	Stem: seen Petals: hidden	Stem: seen Petals: partial	Stem: partial Petals: partial	Stem: hidden Petals: seen

$P(Iris)$: 0.1 $P(\neg Iris)$: 0.9	$P(Iris)$: 0.11 $P(\neg Iris)$: 0.89	$P(Iris)$: 0.2 $P(\neg Iris)$: 0.8	$P(Iris)$: 0.45 $P(\neg Iris)$: 0.55	$\mathbf{P(Iris)}$: **0.9** $P(\neg Iris)$: 0.1

Decision on sequence of observations

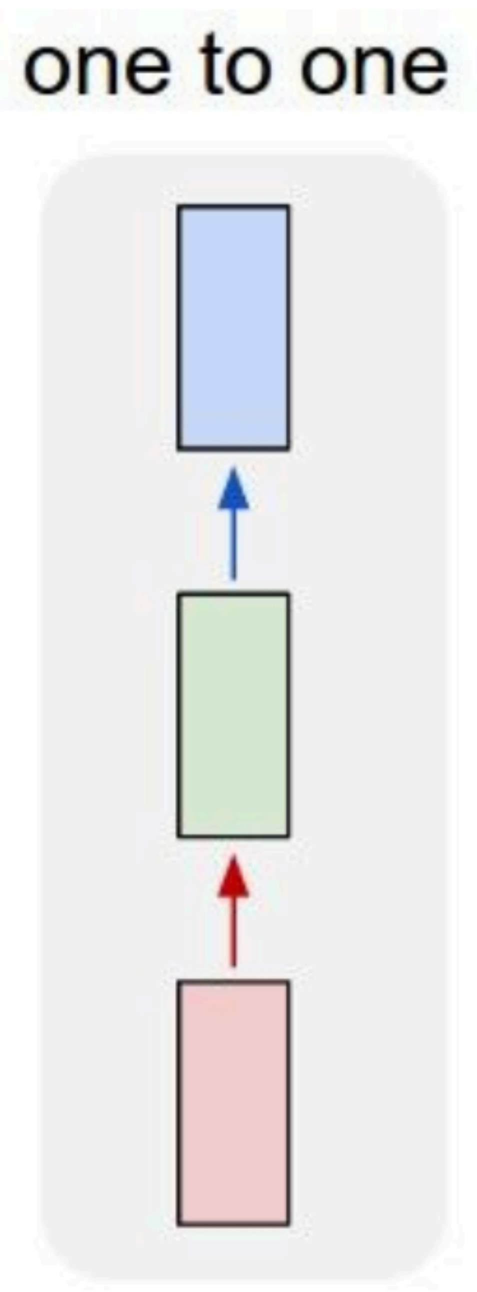
Sequential Data

- Sometimes the sequence of data matters.
  - Text generation
  - Stock price prediction
- **For example: The clouds are in the .... ?**
  - **sky**

Sequential Data

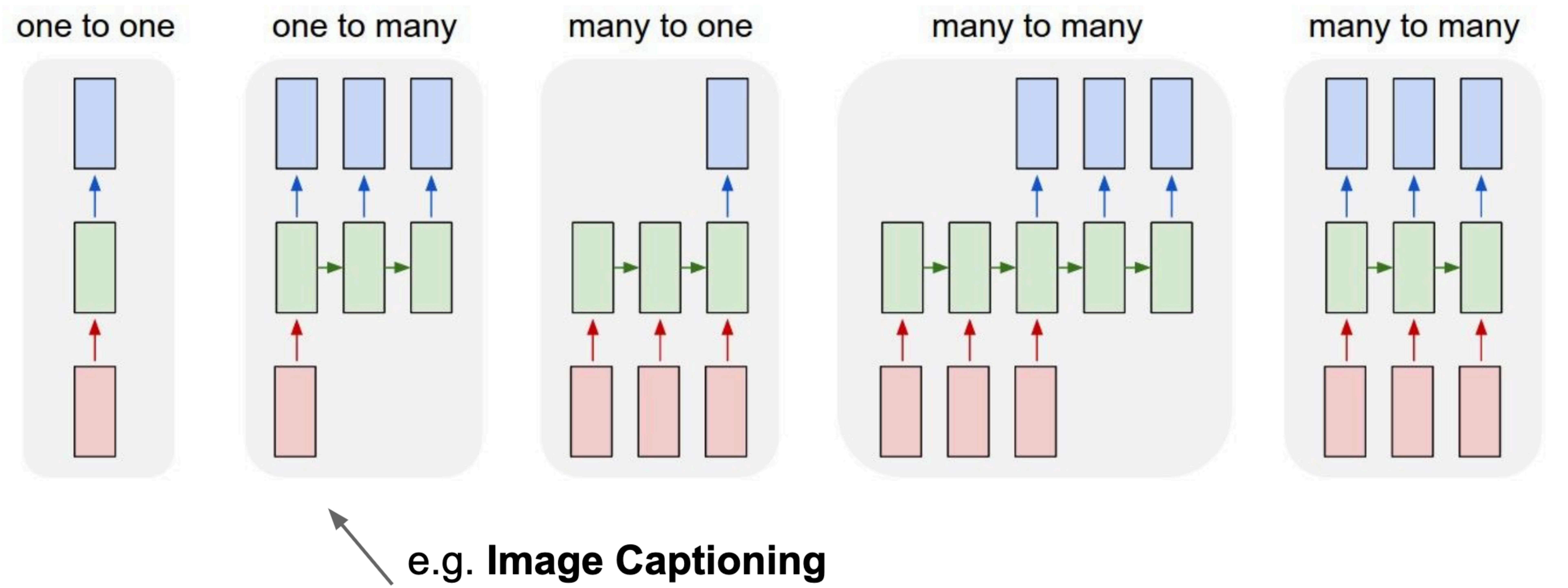
- Sometimes the sequence of data matters.
  - Text generation
  - Stock price prediction
- **For example: The clouds are in the .... ?**
  - **sky**
- Simple solution: Neural networks?
  - Fixed input/output size
  - Fixed number of steps

Neural Networks



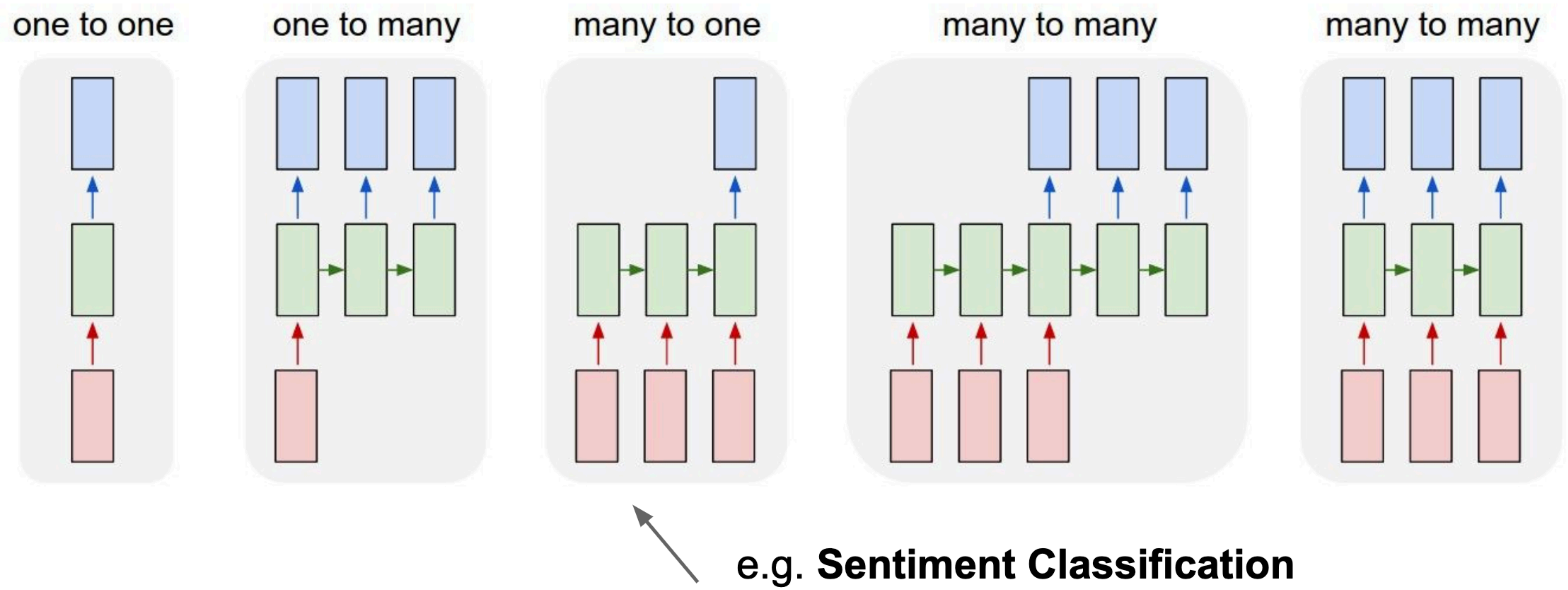
Vanilla Neural Networks

Recurrent Neural Networks



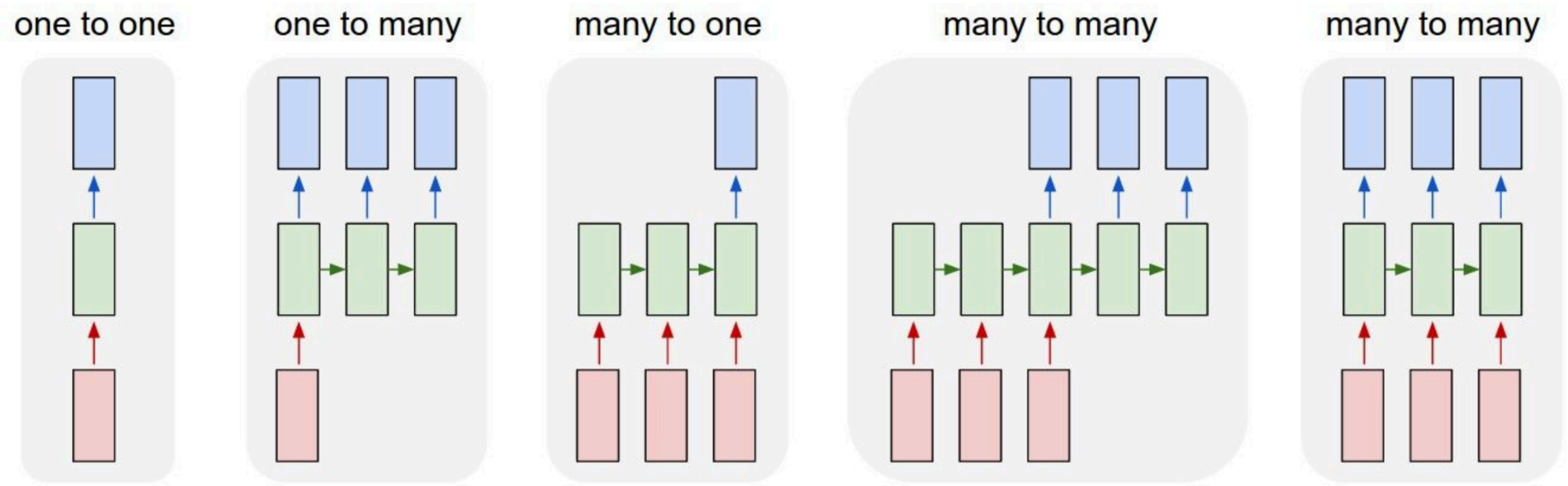
e.g. **Image Captioning**
image -> sequence of words

Recurrent Neural Networks



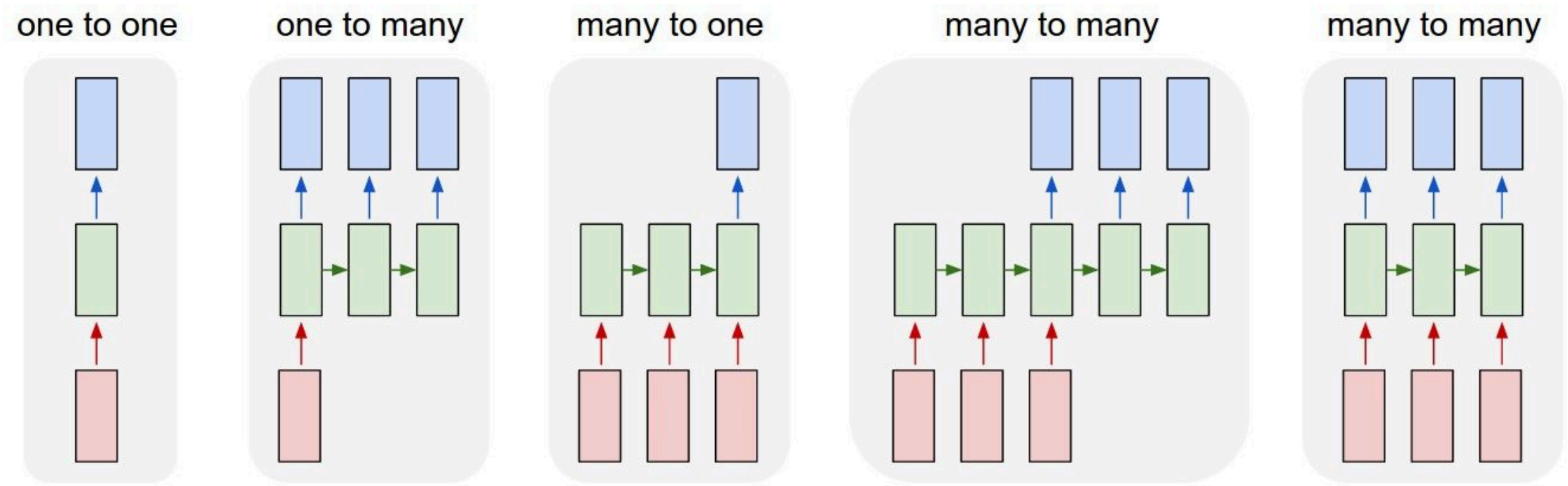
e.g. **Sentiment Classification**
sequence of words -> sentiment

Recurrent Neural Networks



e.g. **Machine Translation**
seq of words -> seq of words

Recurrent Neural Networks

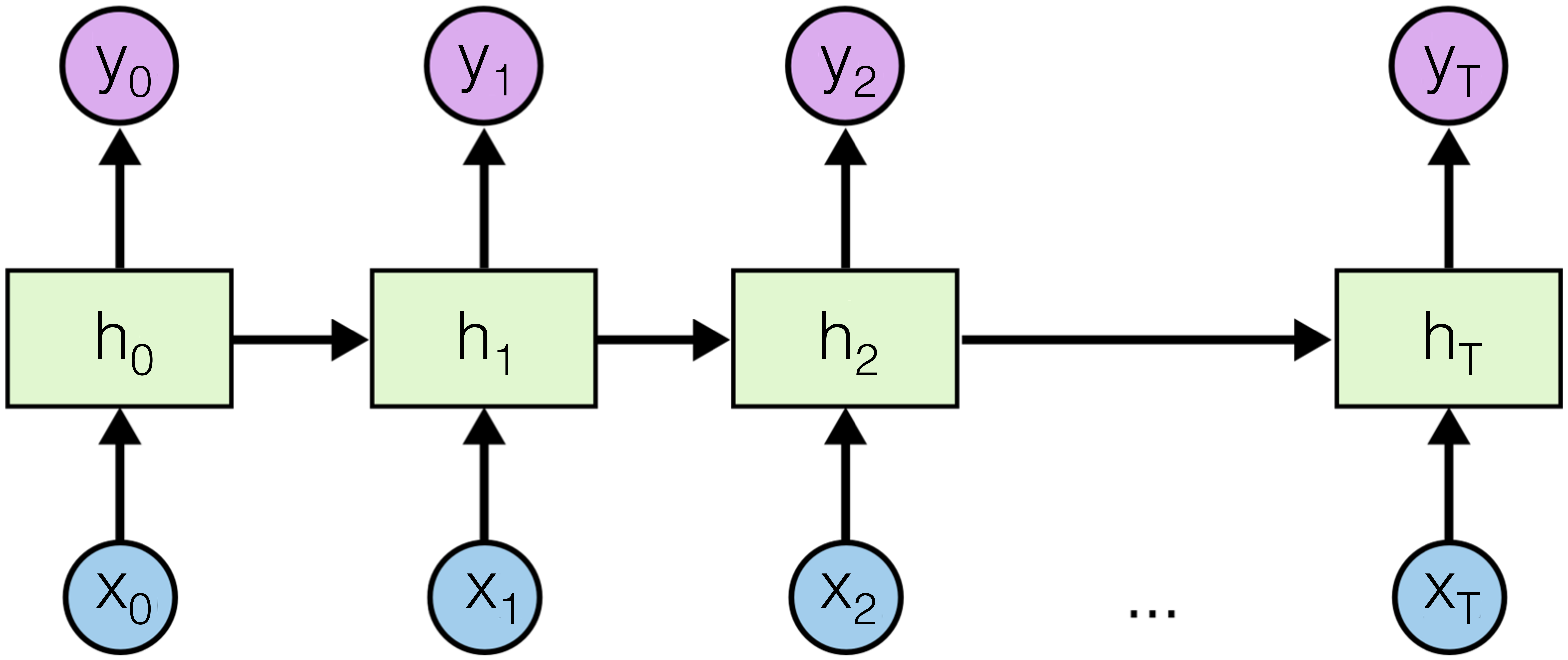


e.g. **Video classification on frame level**

Basic idea

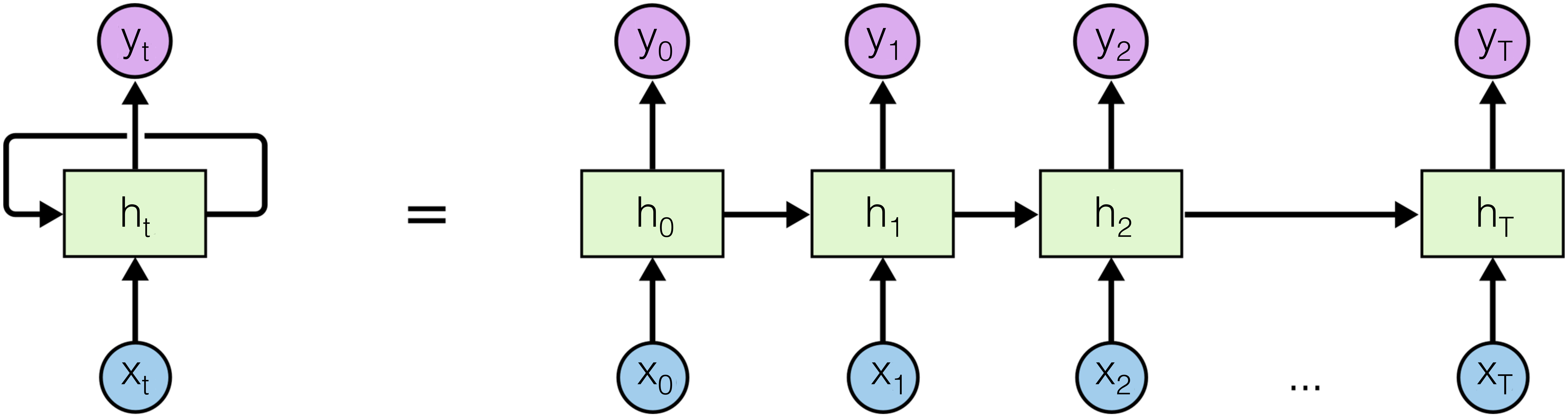
A many-to-many task

- **x**: inputs
- **y**: outputs
- **h**: hidden layers
- subscripts index time



[http://colah.github.io/]

Unroll through time

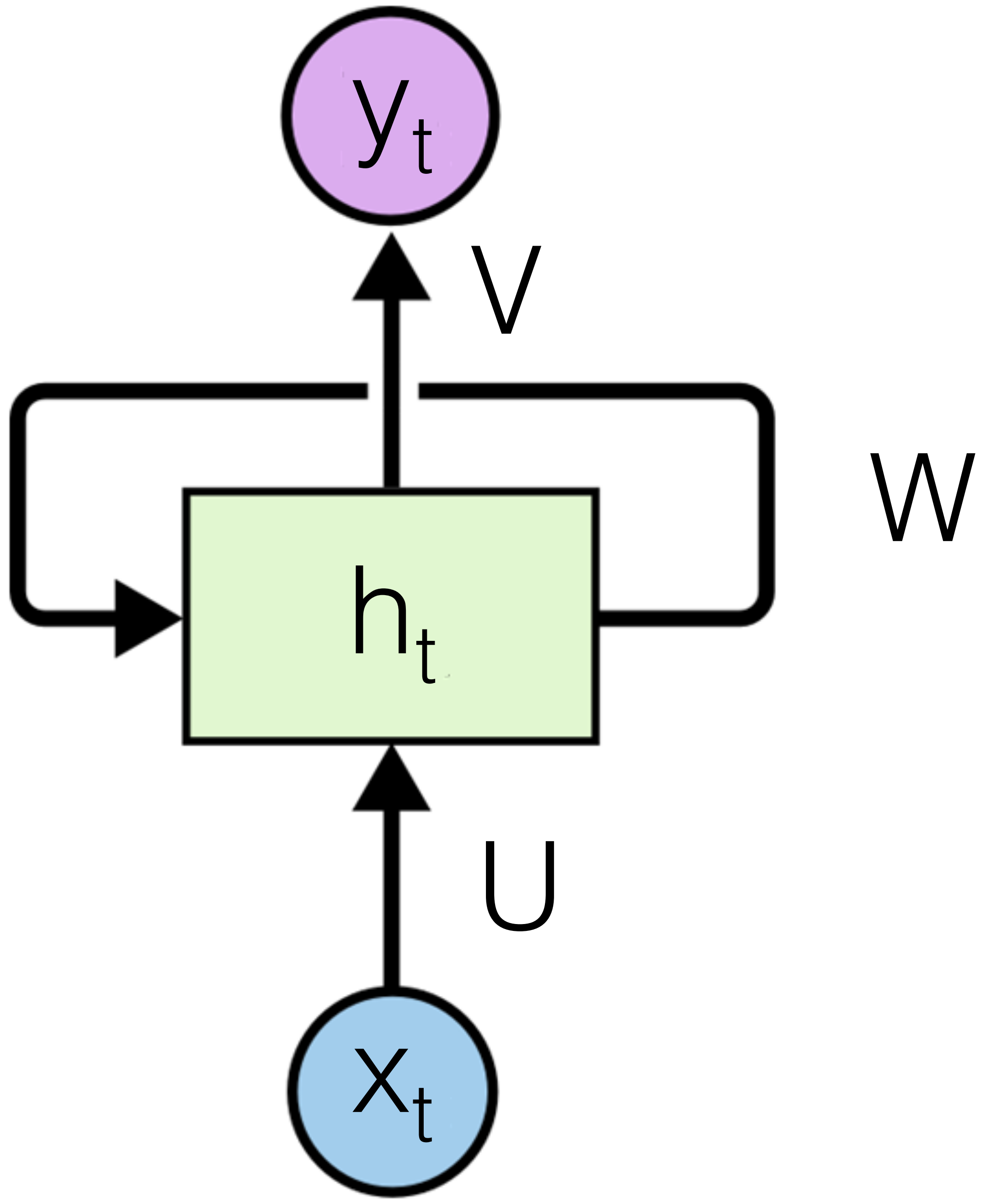


http://colah.github.io/

Basic idea (with parameters)

Process through time (t)

- U, V, W: parameters
  - Shared through time
- Simplest parametrization



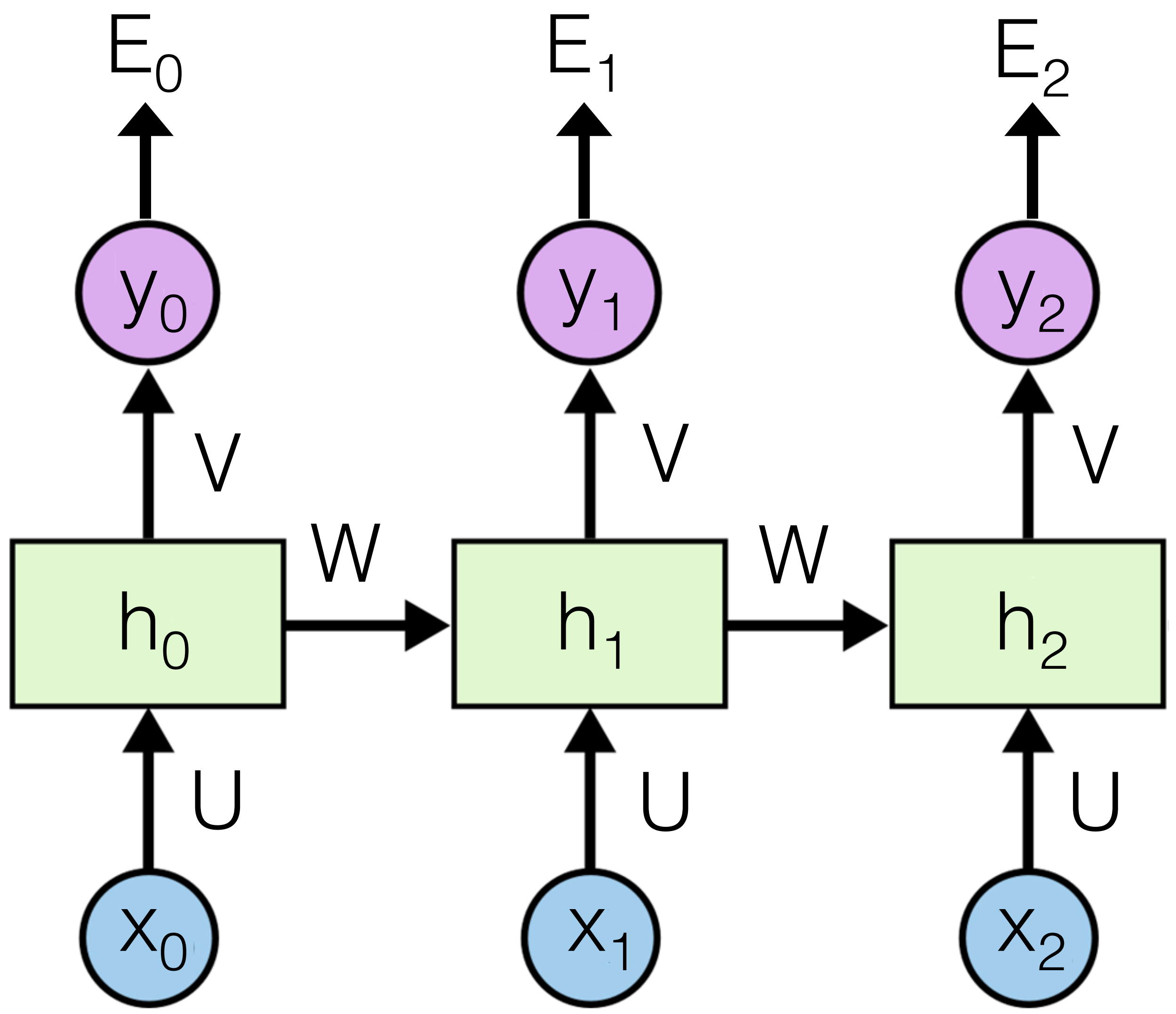
$$\boldsymbol{h_t} = \tanh(\boldsymbol{Ux_t} + \boldsymbol{Wh_{t-1}})$$
$$\boldsymbol{y_t} = \boldsymbol{f}(\boldsymbol{Vh_t})$$

[http://colah.github.io/]

How does gradient flow in RNN?

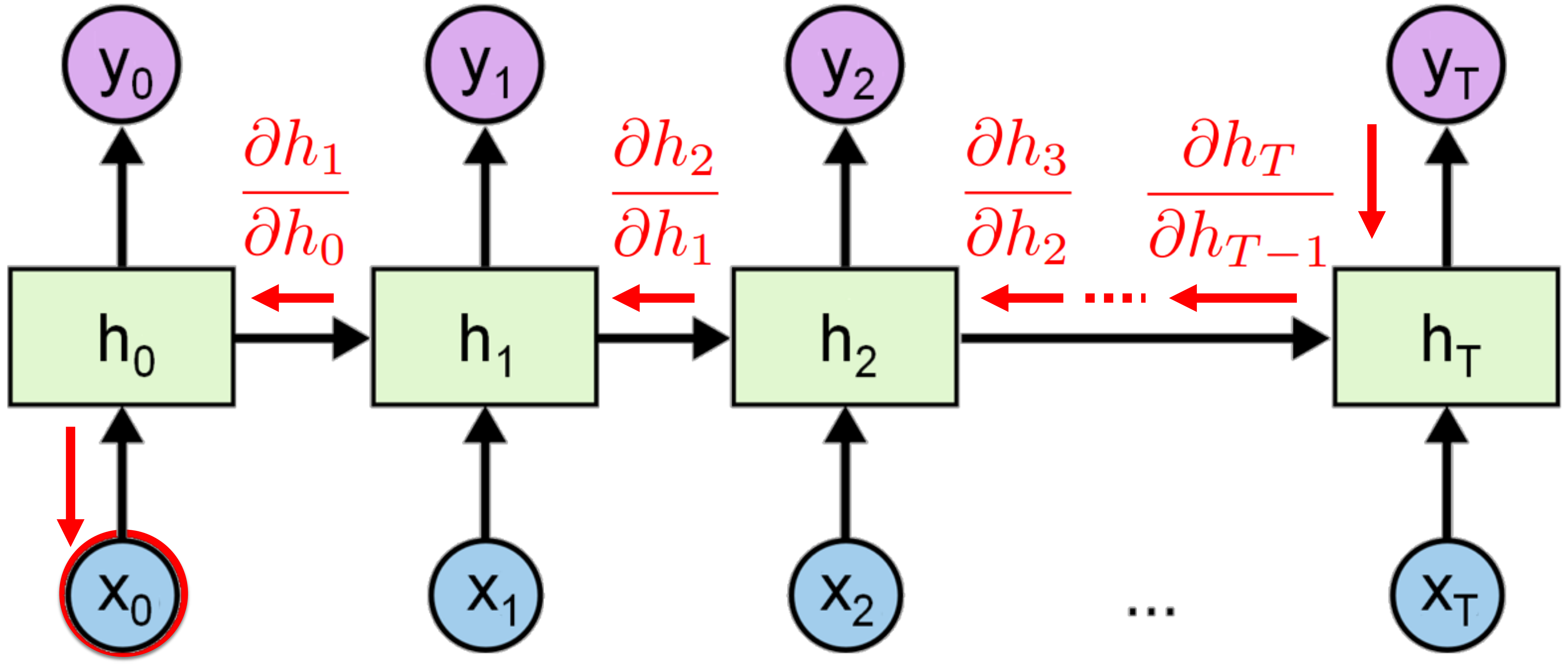
Training RNNs

- **Gradient descent from the loss** $$E = \sum_t (y_t - \hat{y}_t)^2$$
- **Following the structure the gradient is back propagated through time**



RNN Gradient Flow

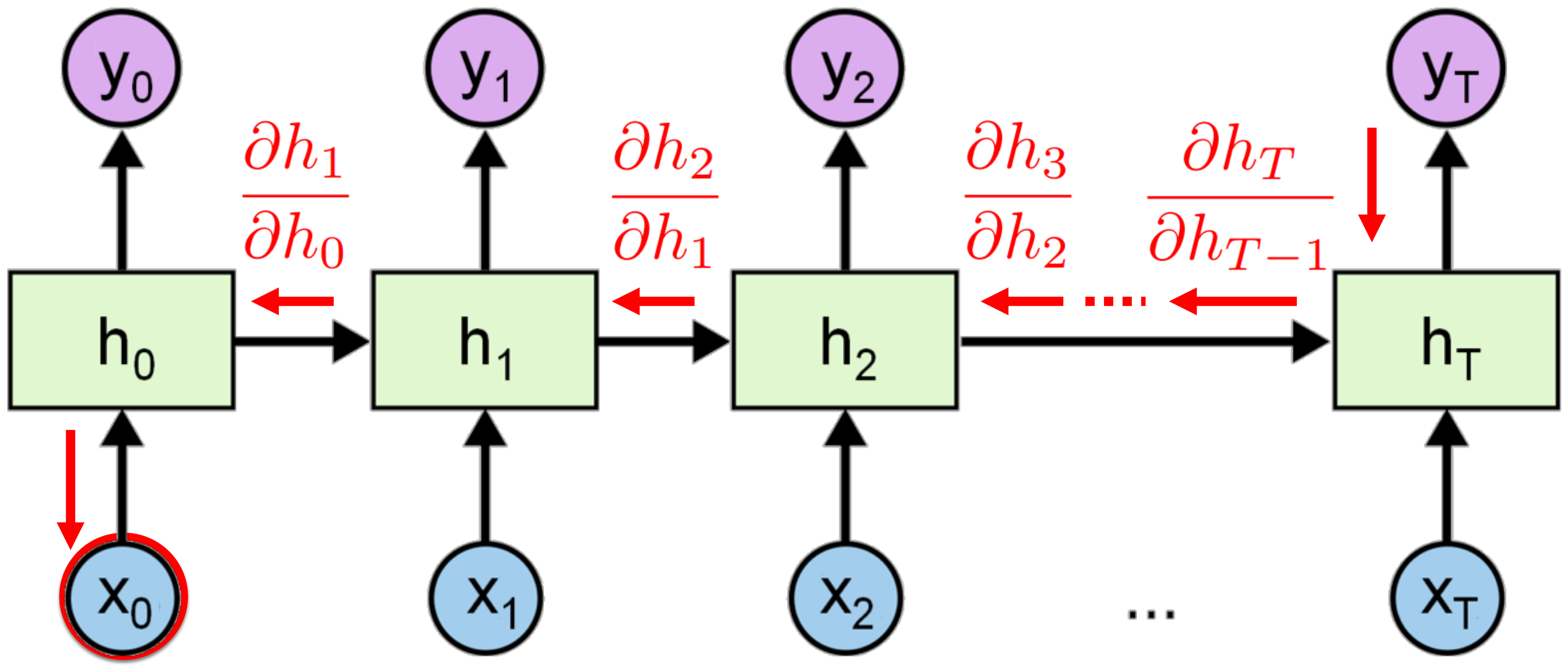
Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994
Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013



Largest singular value > 1:
Exploding gradients

Largest singular value < 1:
Vanishing gradients

RNN Gradient Flow



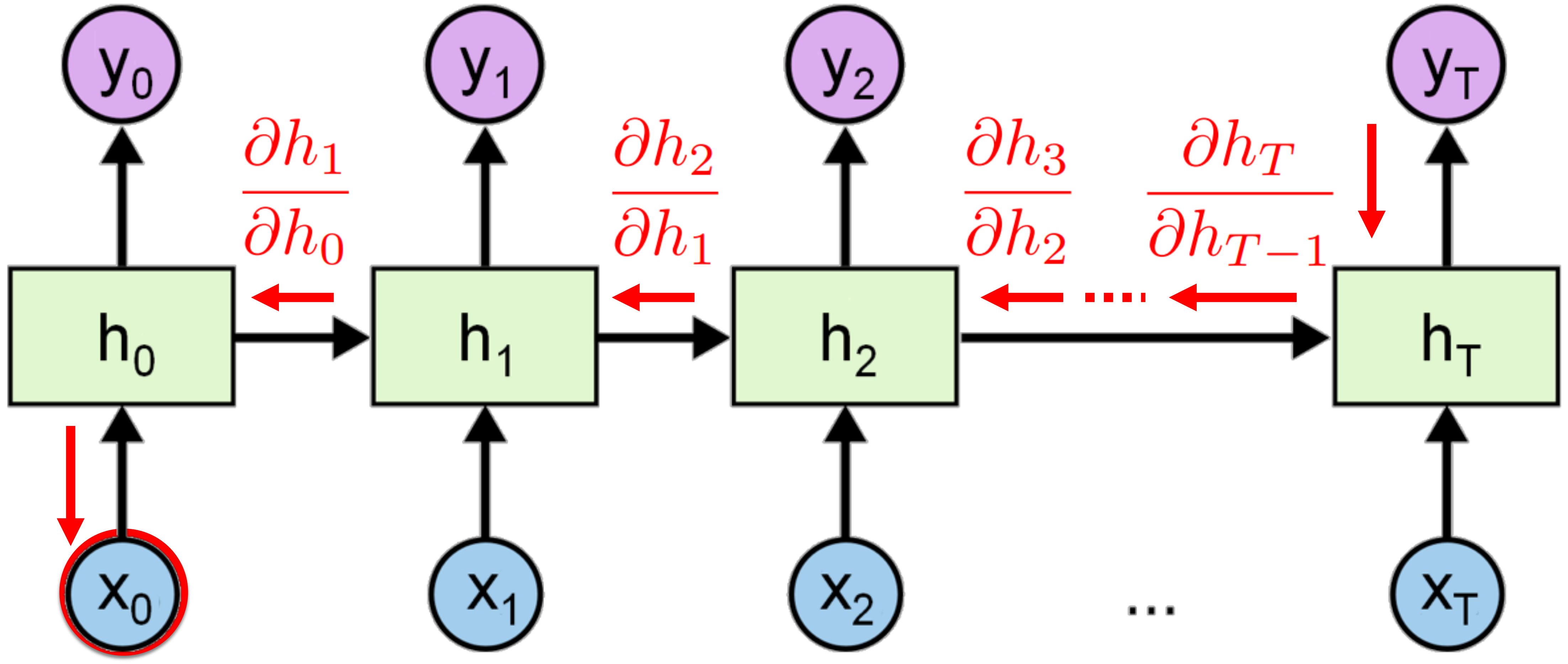
Largest singular value > 1:
Exploding gradients

Largest singular value < 1:
Vanishing gradients

Gradient clipping: Scale gradient if its norm is too big

```
grad_norm = np.sum(grad * grad)
if grad_norm > threshold:
  grad *= (threshold / grad_norm)
```

RNN Gradient Flow



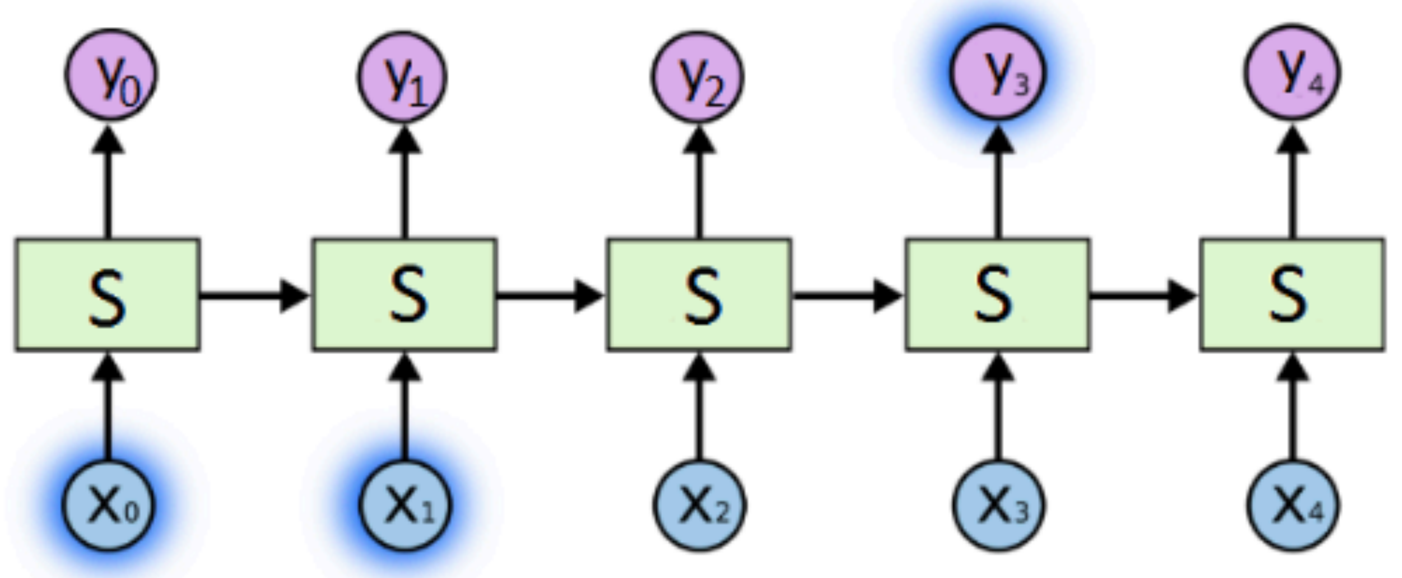
Largest singular value > 1:
Exploding gradients

Largest singular value < 1:
Vanishing gradients

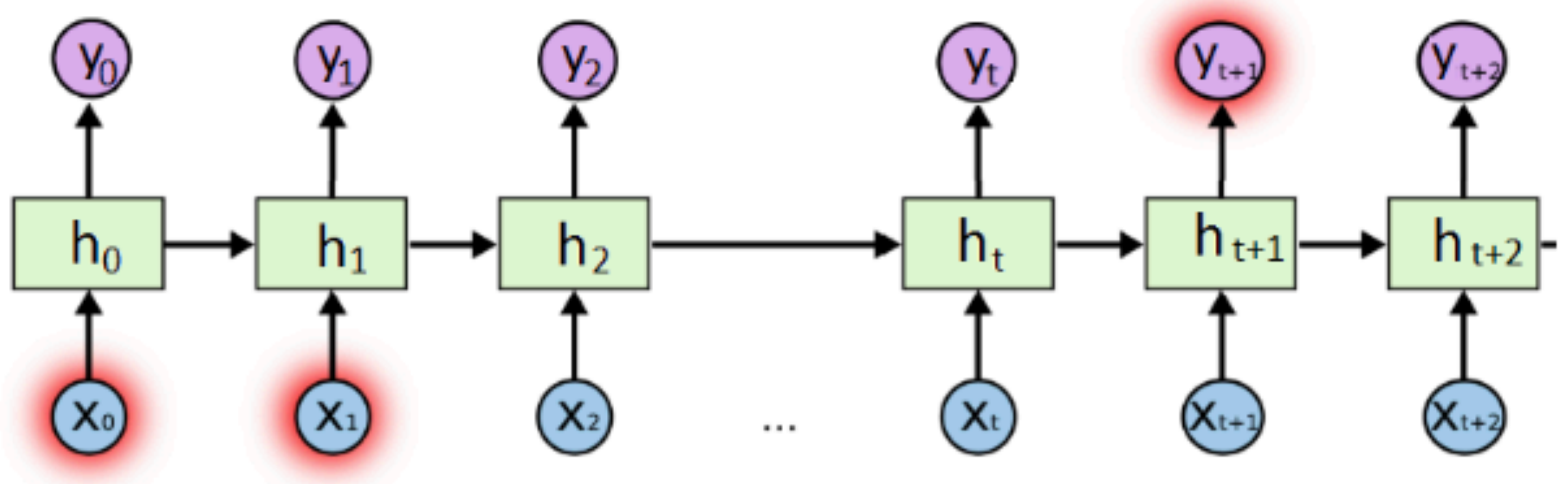
new weight = weight - learning rate*gradient **0**

The Problem of Long-term Dependencies

- RNNs connect previous information to present task:
  - may be enough for predicting the next word for ”the clouds are in the sky”



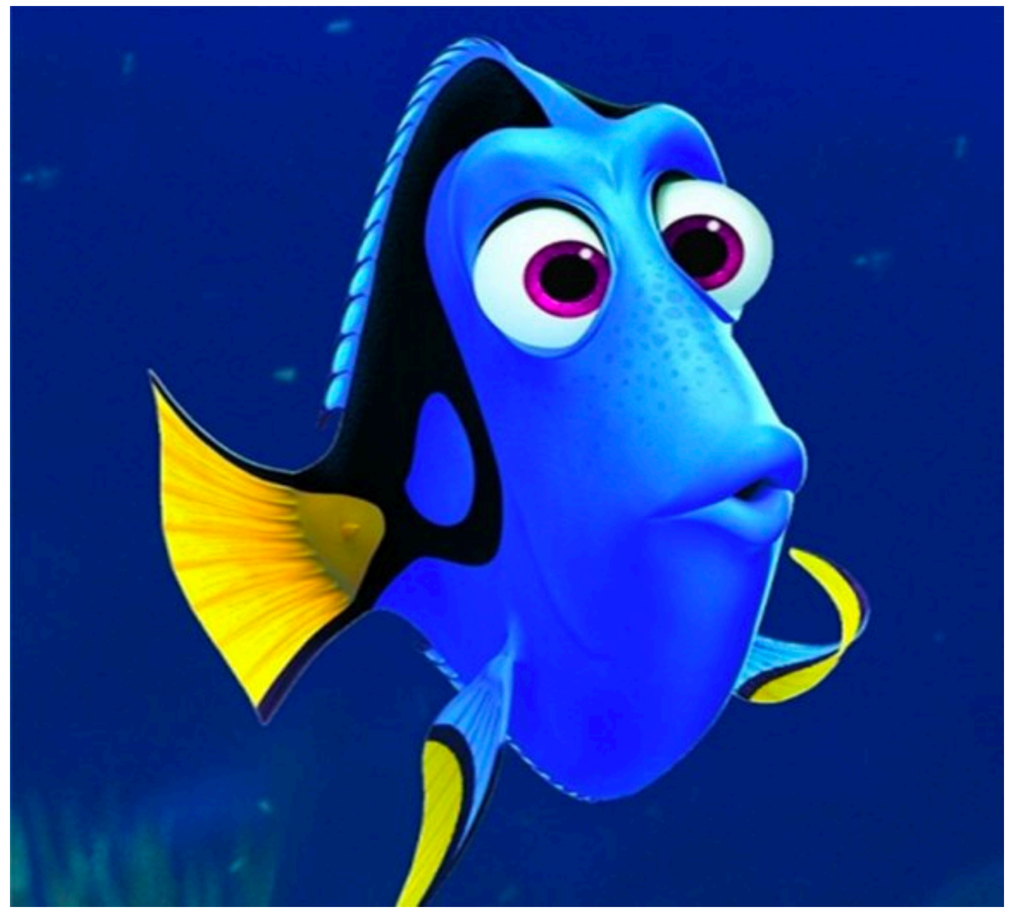
  - may not be enough when more context is needed



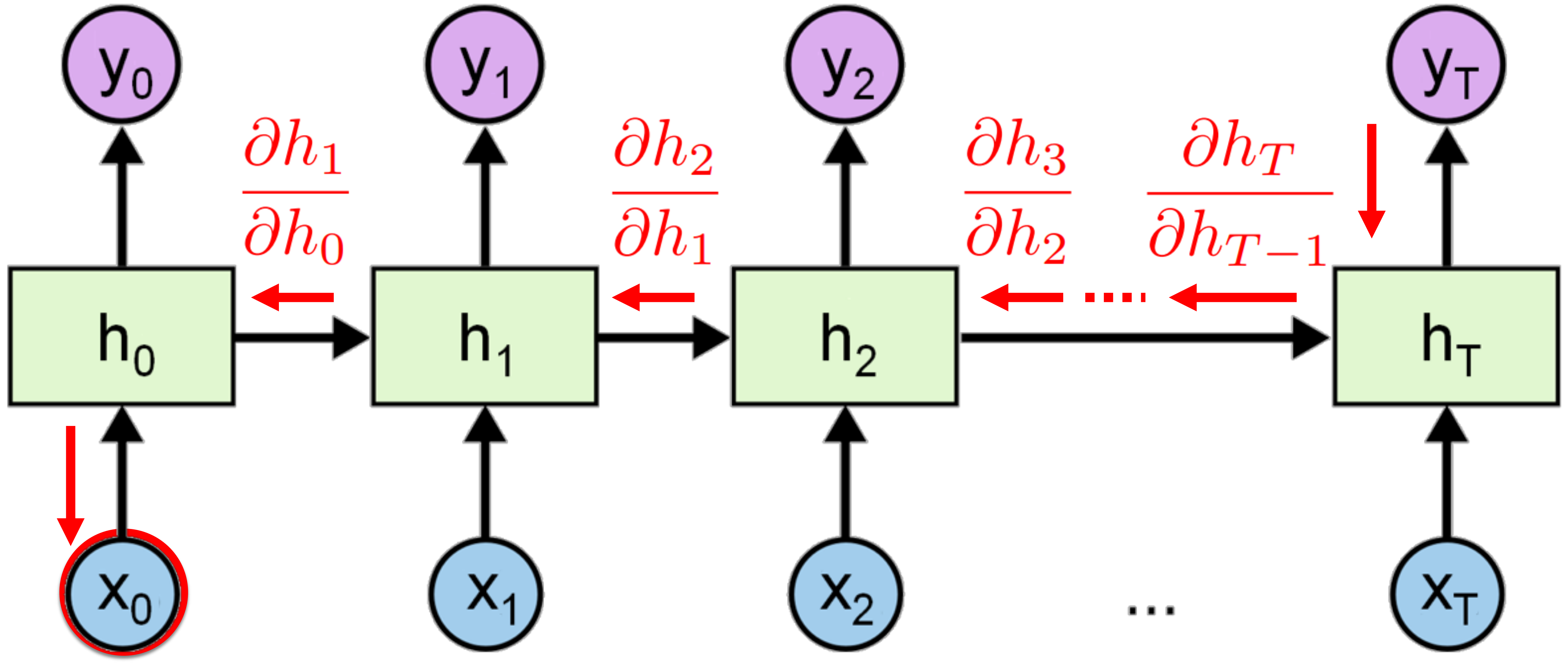
Short-Term memory

- RNNs suffer from what is known as short-term memory!

```
I was born in France, but I have been working in South Africa
working for ... (another 200 words) ... Therefore my mother tongue
is:
```



RNN Gradient Flow



Largest singular value > 1:
Exploding gradients

Largest singular value < 1:
Vanishing gradients

→ Change RNN architecture

0

new weight = weight - learning rate*~~gradient~~

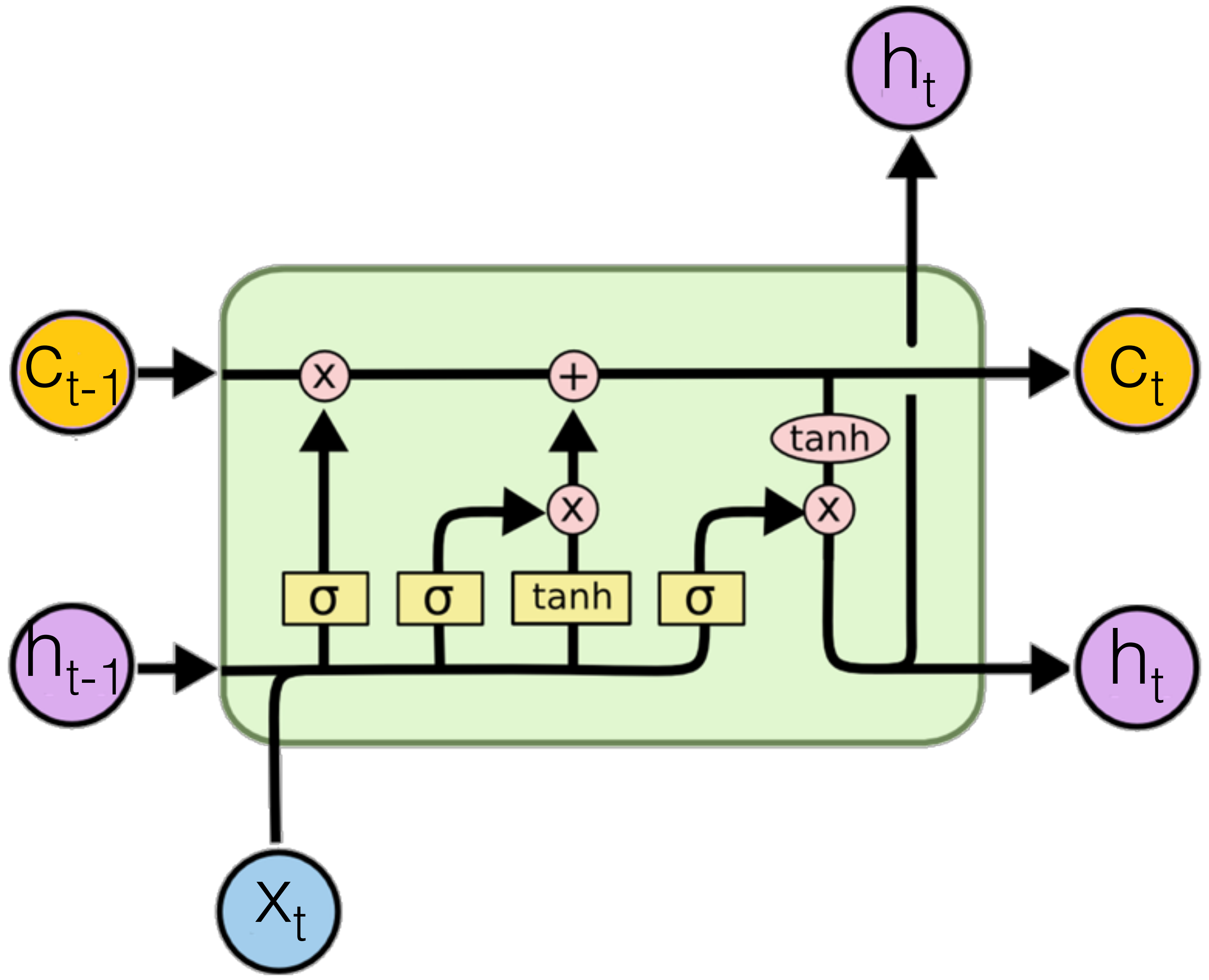
Gradient Vanishing

- **No simple solution**
- **Change the "memory cells" of the RNN**

$$\boldsymbol{h_t} = \tanh(\boldsymbol{U}\boldsymbol{x_t} + \boldsymbol{W}\boldsymbol{h_{t-1}})$$
$$\boldsymbol{y_t} = \boldsymbol{f}(\boldsymbol{V}\boldsymbol{h_t})$$

LSTM

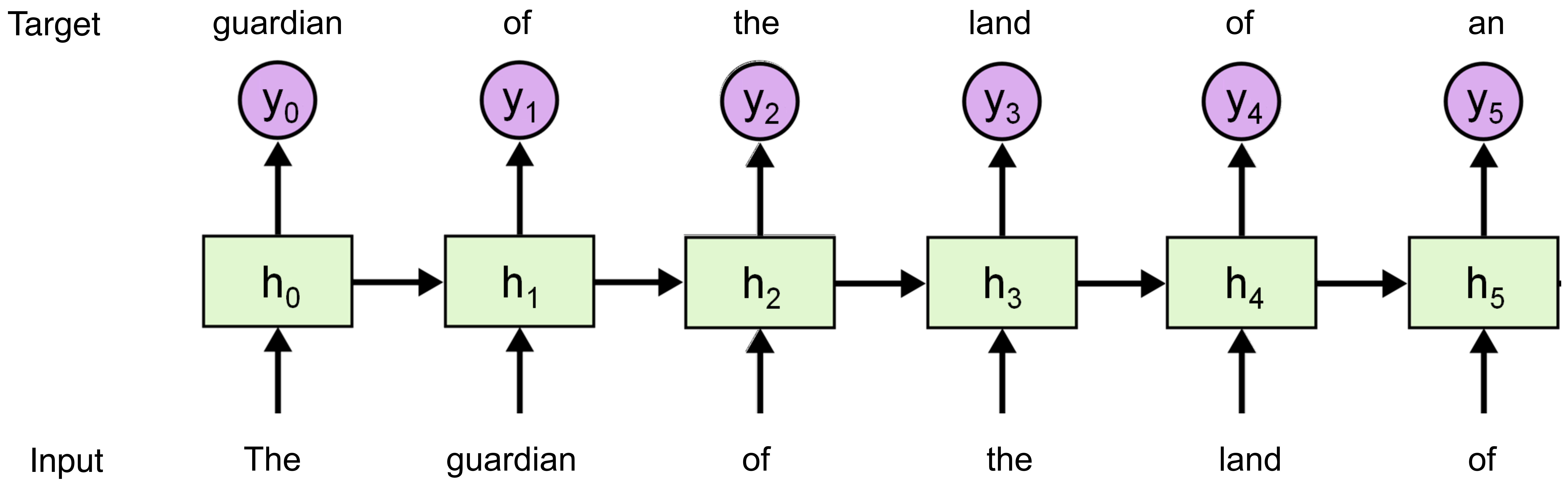
Long short-term memory



http://colah.github.io/

Language Modelling using RNNs

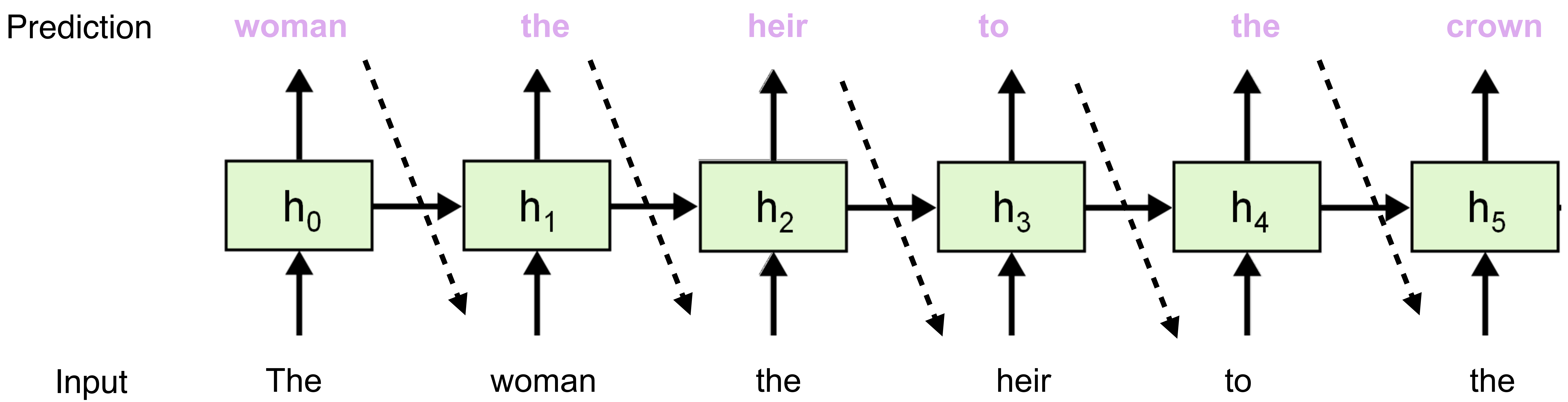
Training.
Given previous words,
predict the next word



http://colah.github.io/

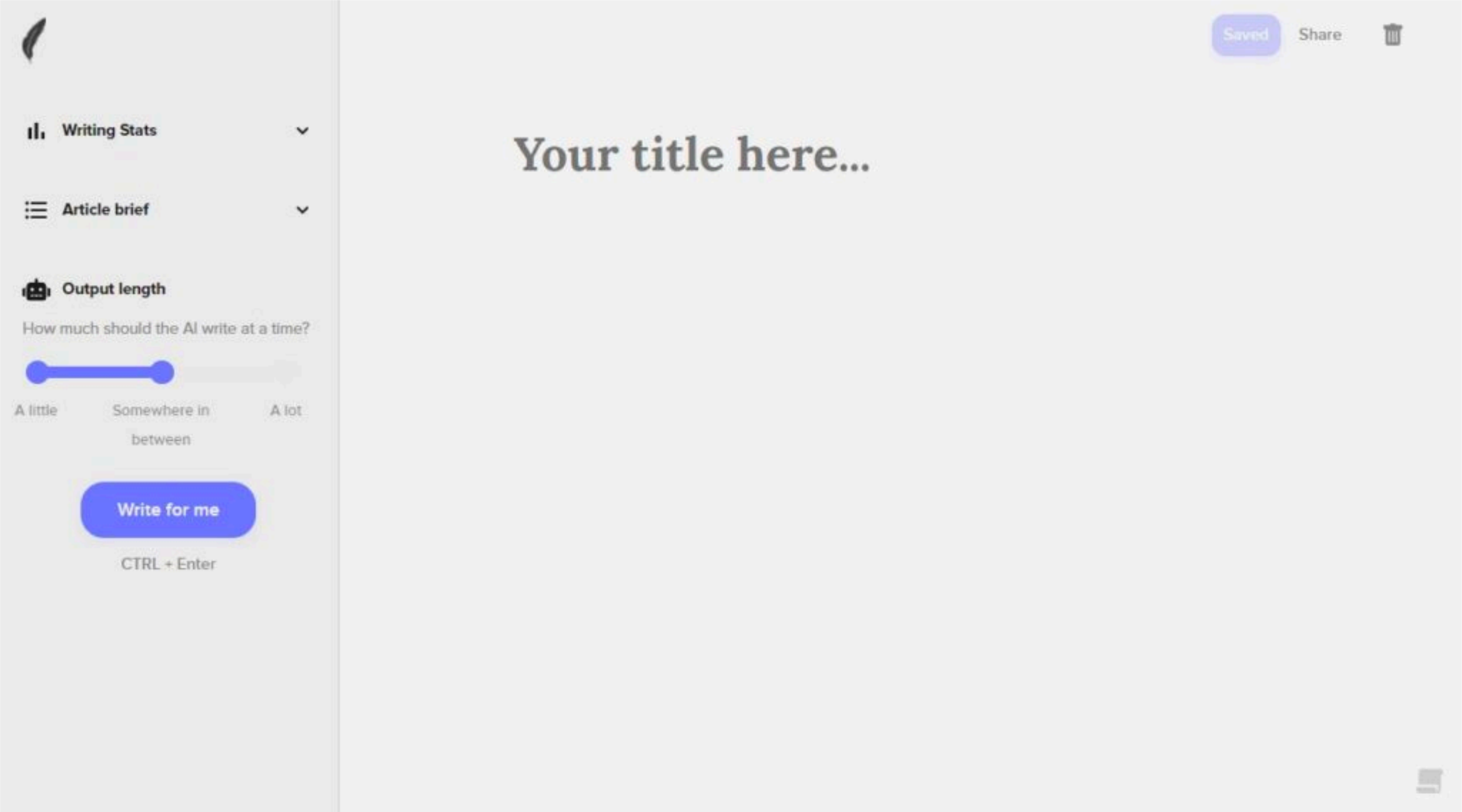
Test.

1. Predict One word at a time.
2. Feed the prediction back to the model



http://colah.github.io/

Fun example to try!



Shortly

https://shortlyai.com/

CNNs

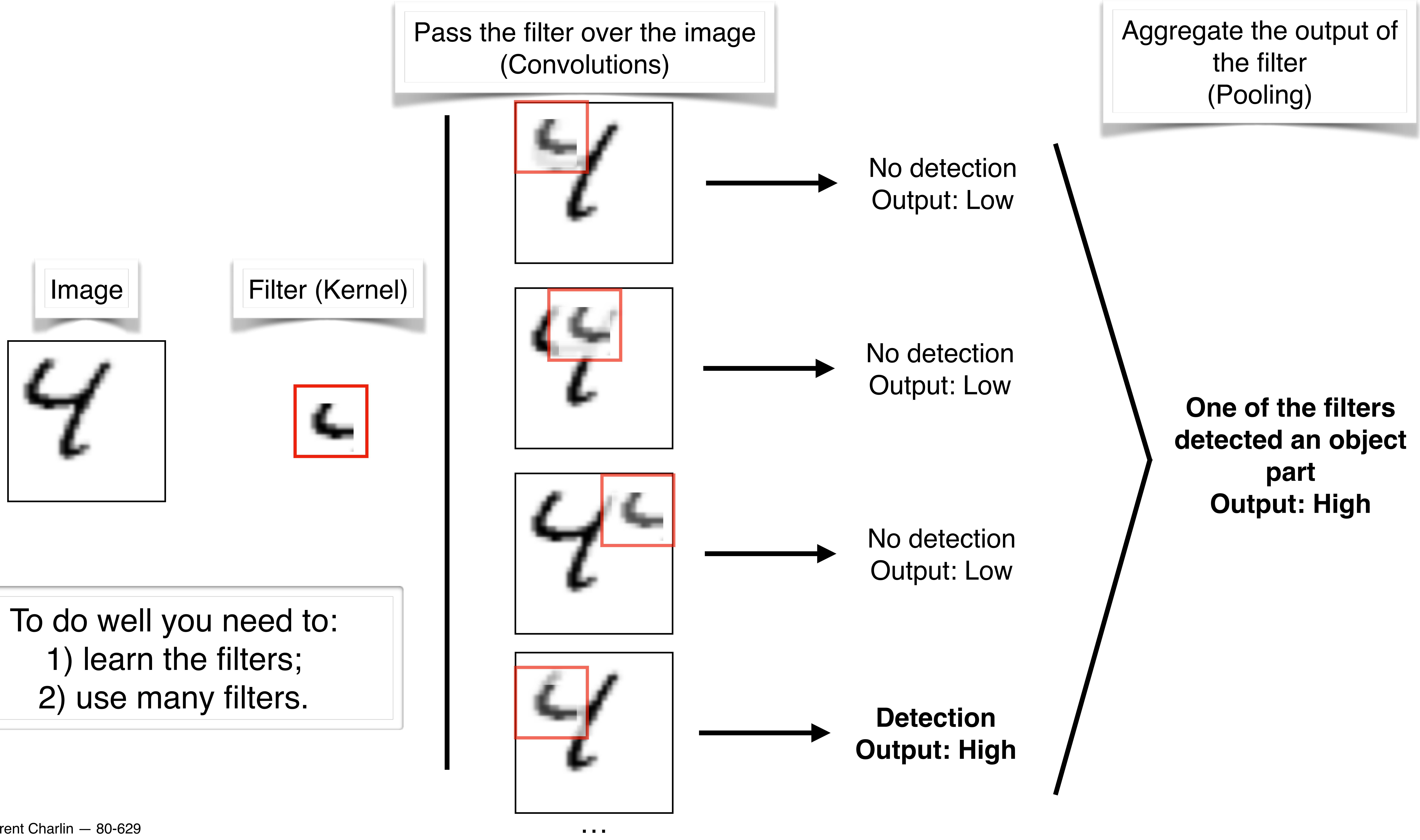
This is how I see

88	126	145	85	123	142	85	123	142	86	124
86	125	142	84	123	140	83	122	139	85	124
85	124	141	82	121	138	82	121	138	84	123
82	119	135	80	117	133	80	117	133	85	122
78	114	128	77	113	127	79	115	129	84	120
79	115	129	78	114	128	80	116	130	83	119
82	118	130	81	117	129	81	117	129	82	118
83	117	129	82	116	128	82	116	128	82	116
79	113	123	79	113	123	80	114	124	81	115
76	108	119	76	108	119	77	109	120	80	112
76	109	118	76	109	118	77	110	119	79	112

This is how my computer sees

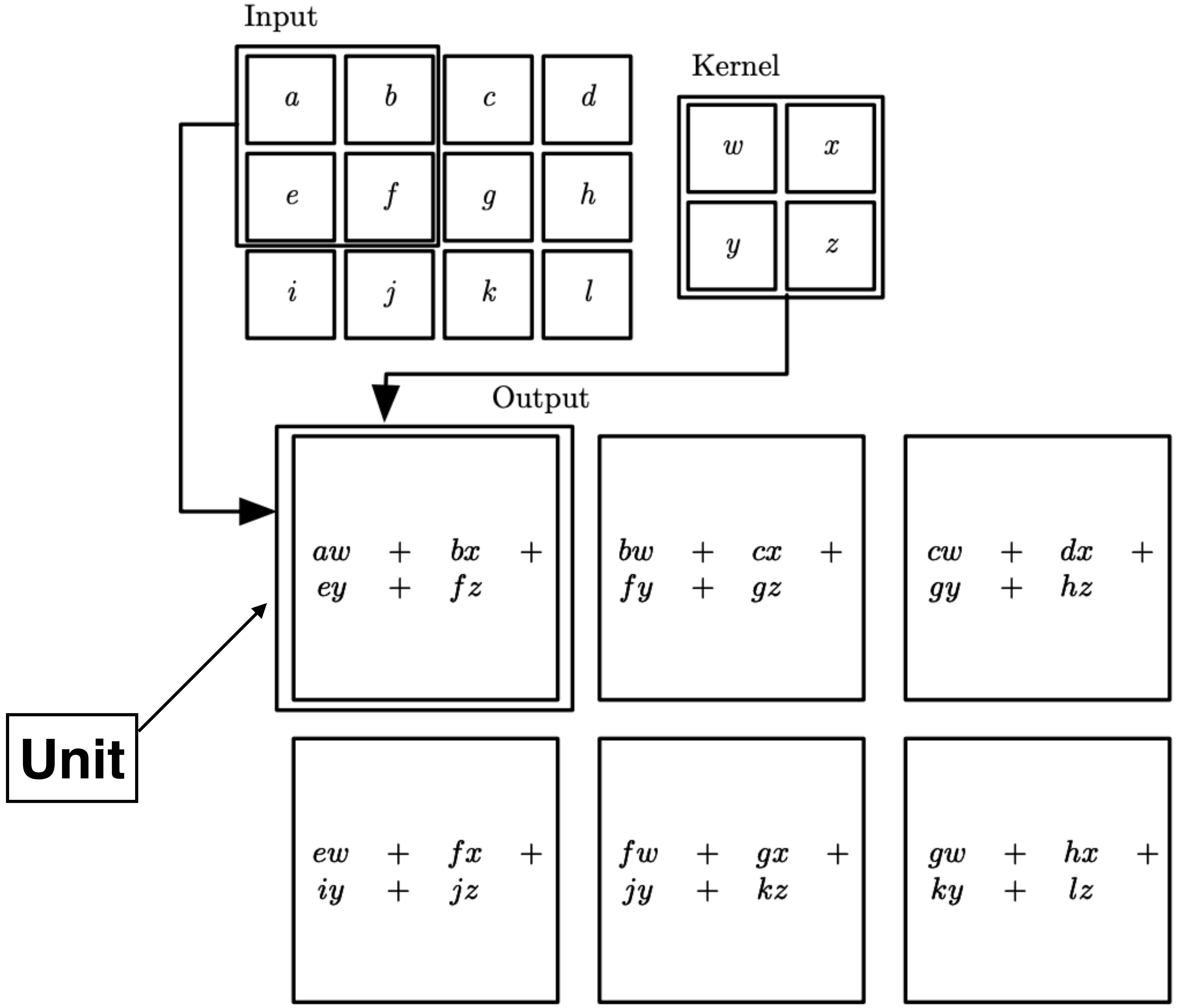
- **A 100 x 100 pixel image has 10,000 dimensions**
    - **Often have a color dimension (data is a tensor)**
- **Modern iPhone (12 MP): 4032 x 3024 pixels**

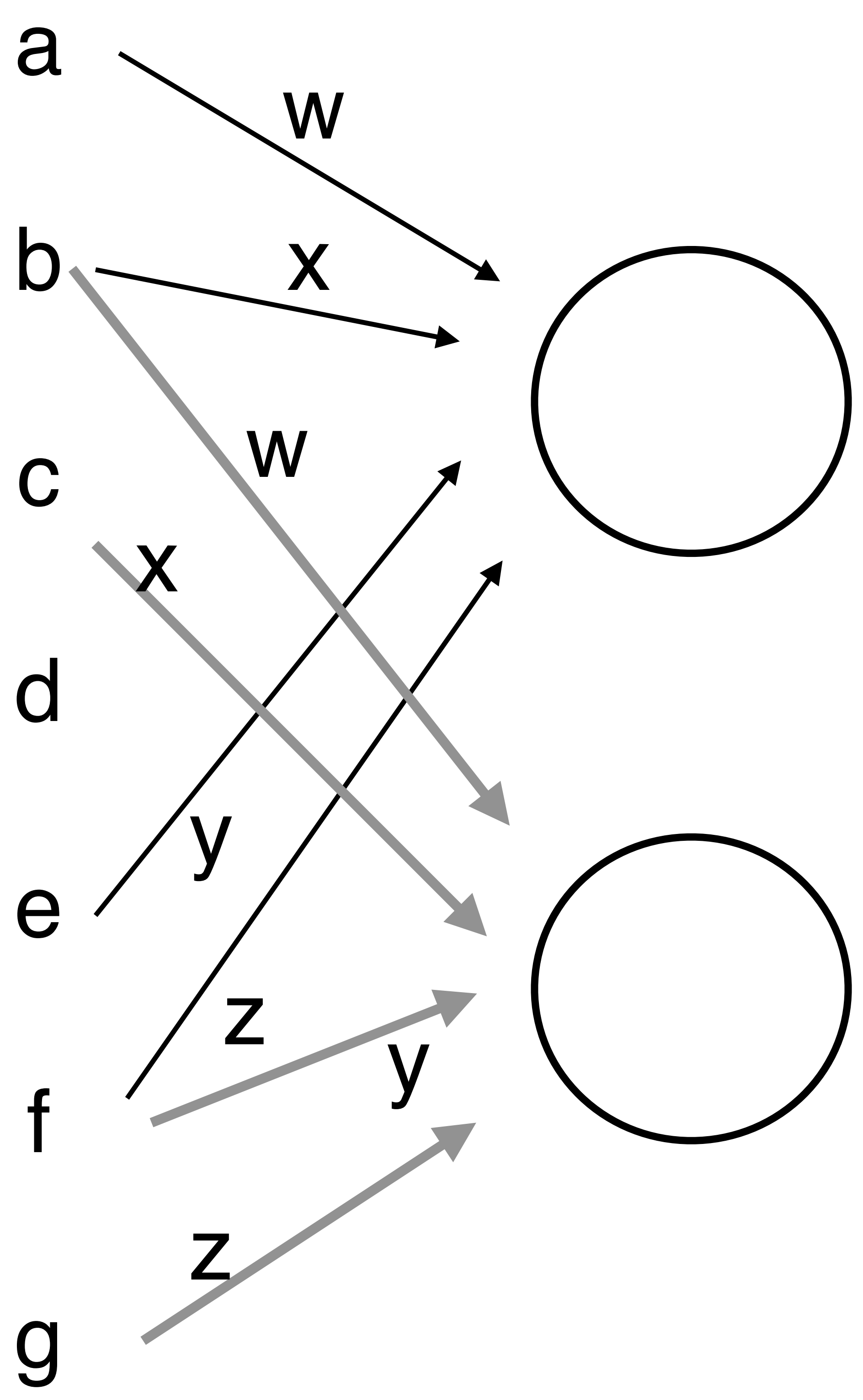
Source: https://medium.com/deep-math-machine-learning-ai/chapter-8-0-convolutional-neural-networks-for-deep-learning-364971e34ab2



To do well you need to:
1) learn the filters;
2) use many filters.

Sparse connections and shared weights





- **Kernels induce sparse and shared connections**
- **Kernels must be small compared to the data**

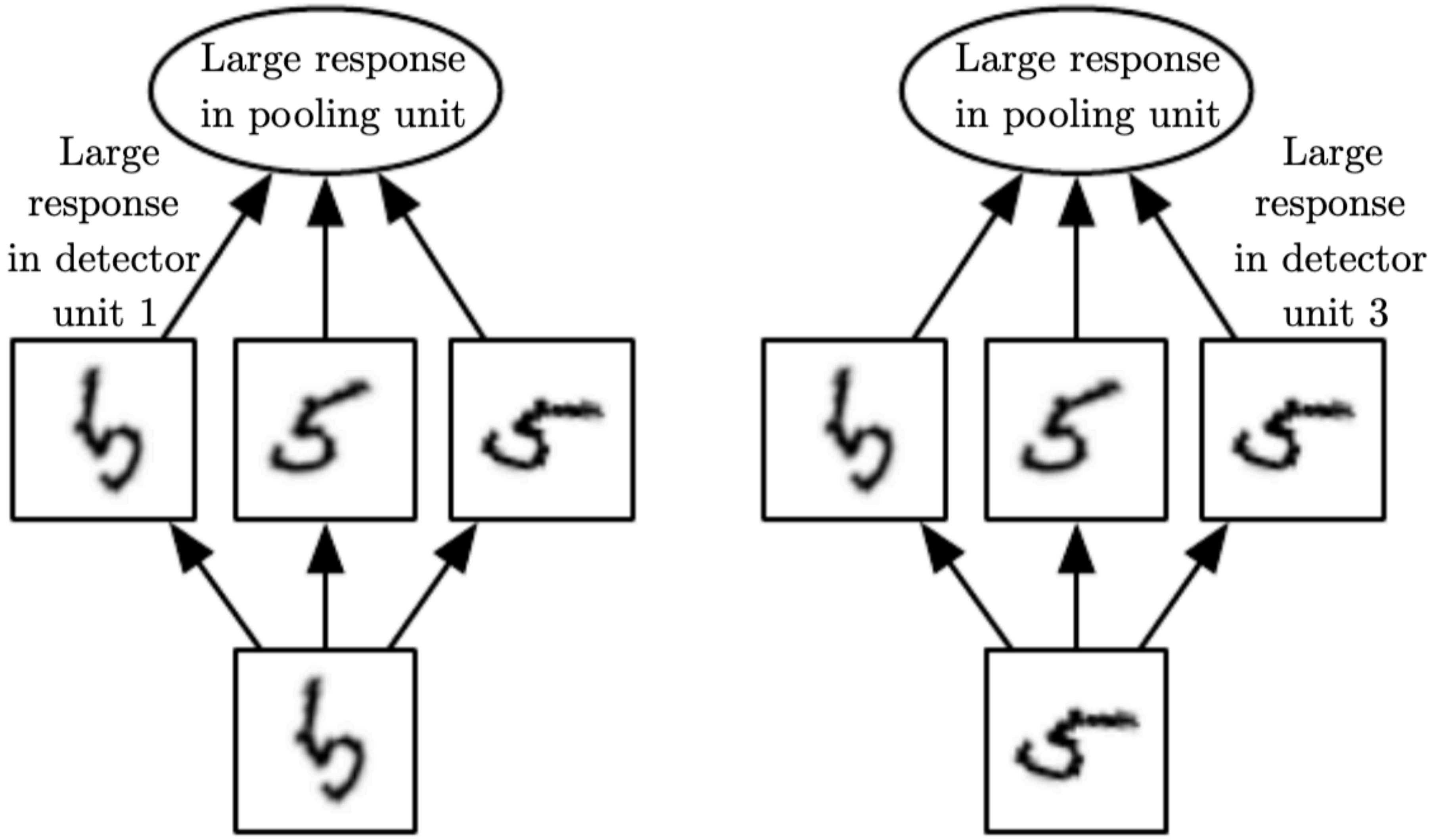
[Figure 9.1, Deep Learning, Book]

Pooling (II)

- **Pooling over different features can learn other types of invariances**
  - **Here, the model learns to be invariant to certain rotations:**

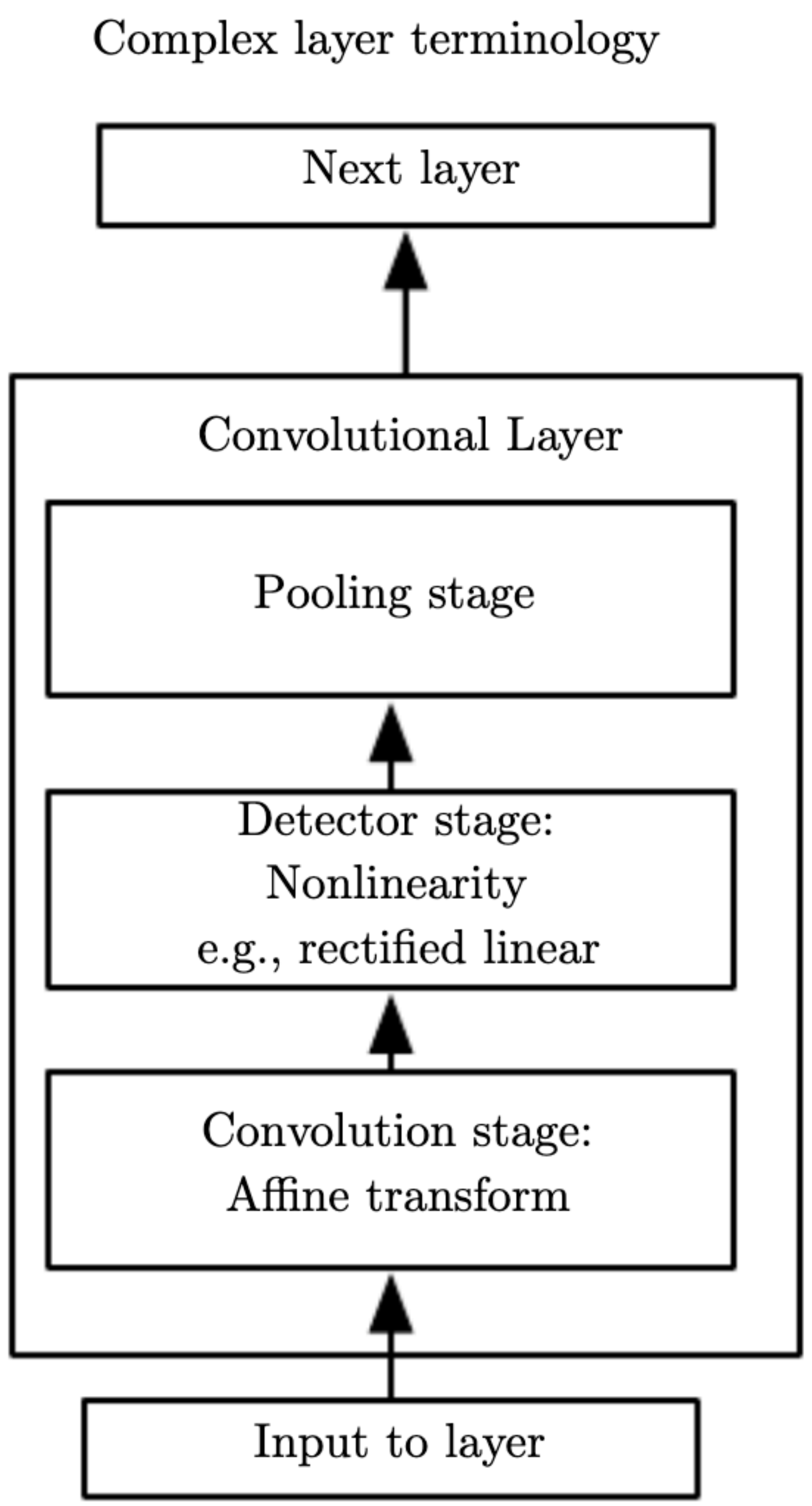
Features / Kernels

Inputs



[Figure 9.9, Deep Learning, Book]

A layer in a CNN



[Figure 9.7, Deep Learning, Book]