

Lecture 8: Sequence Modeling and Recurrent Neural Networks (RNN)

Olexandr Isayev

Department of Chemistry, CMU

olexandr@cmu.edu

Class Project

- Class project 2-minute topic intro by **March 17**
- Please use a template and fill in ASAP:
- <https://docs.google.com/presentation/d/18MzCutGZ4InECsVgFEKUHarXbazgo3UPaXgpWj3brRQ/edit?usp=sharing>

Project Title

Team:
Student Name
Student Name

A brief idea of the problem you are trying to solve

Dataset (type of data, size, etc.)

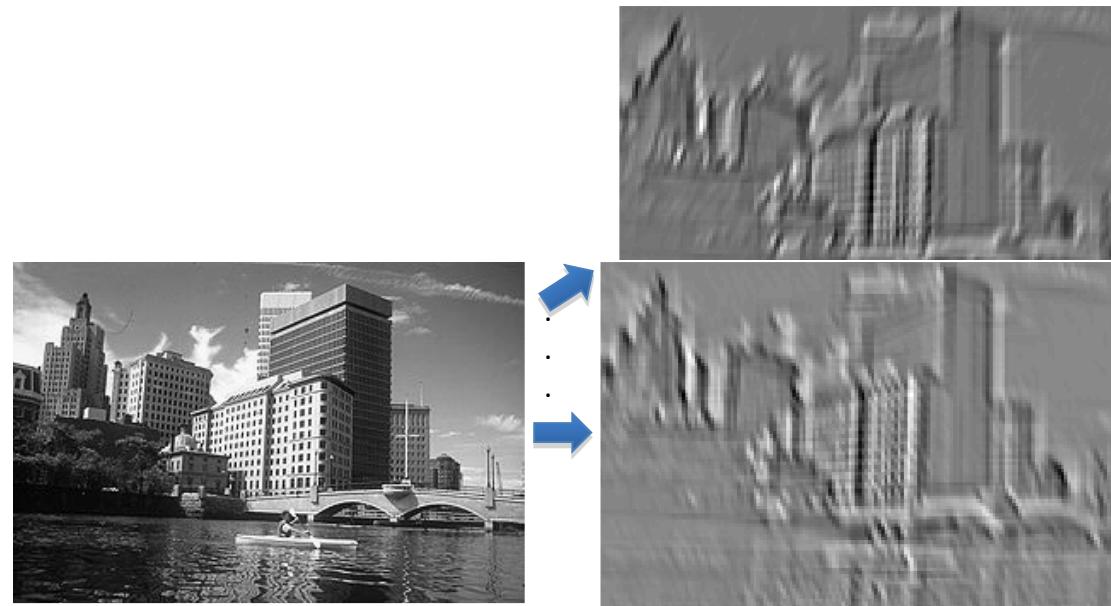
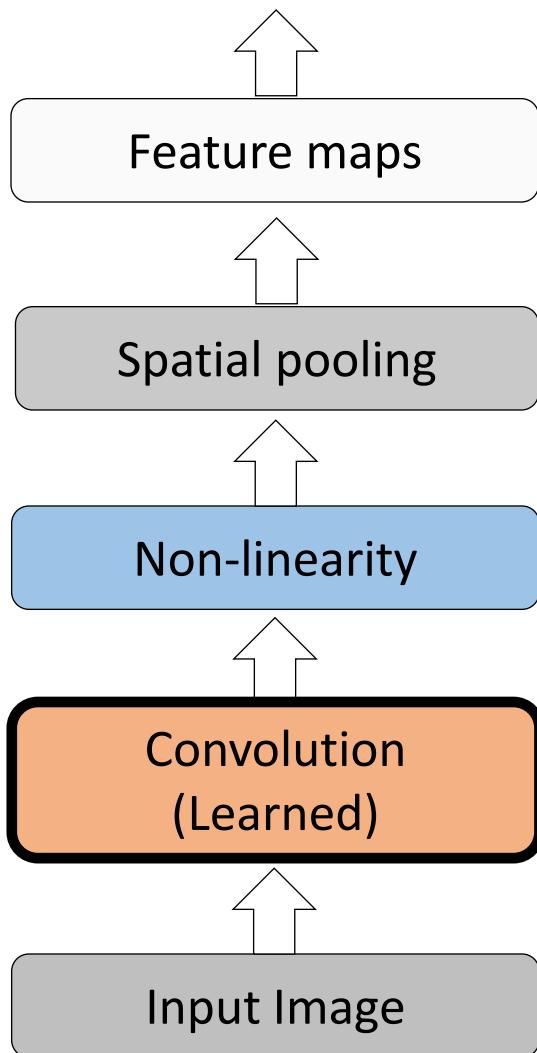
What are you planning to do?

Please add your slides after this template.

HW2 Discussions

Wednesday, March 15

Summary: CNN pipeline



Input

Feature Map

Architectures

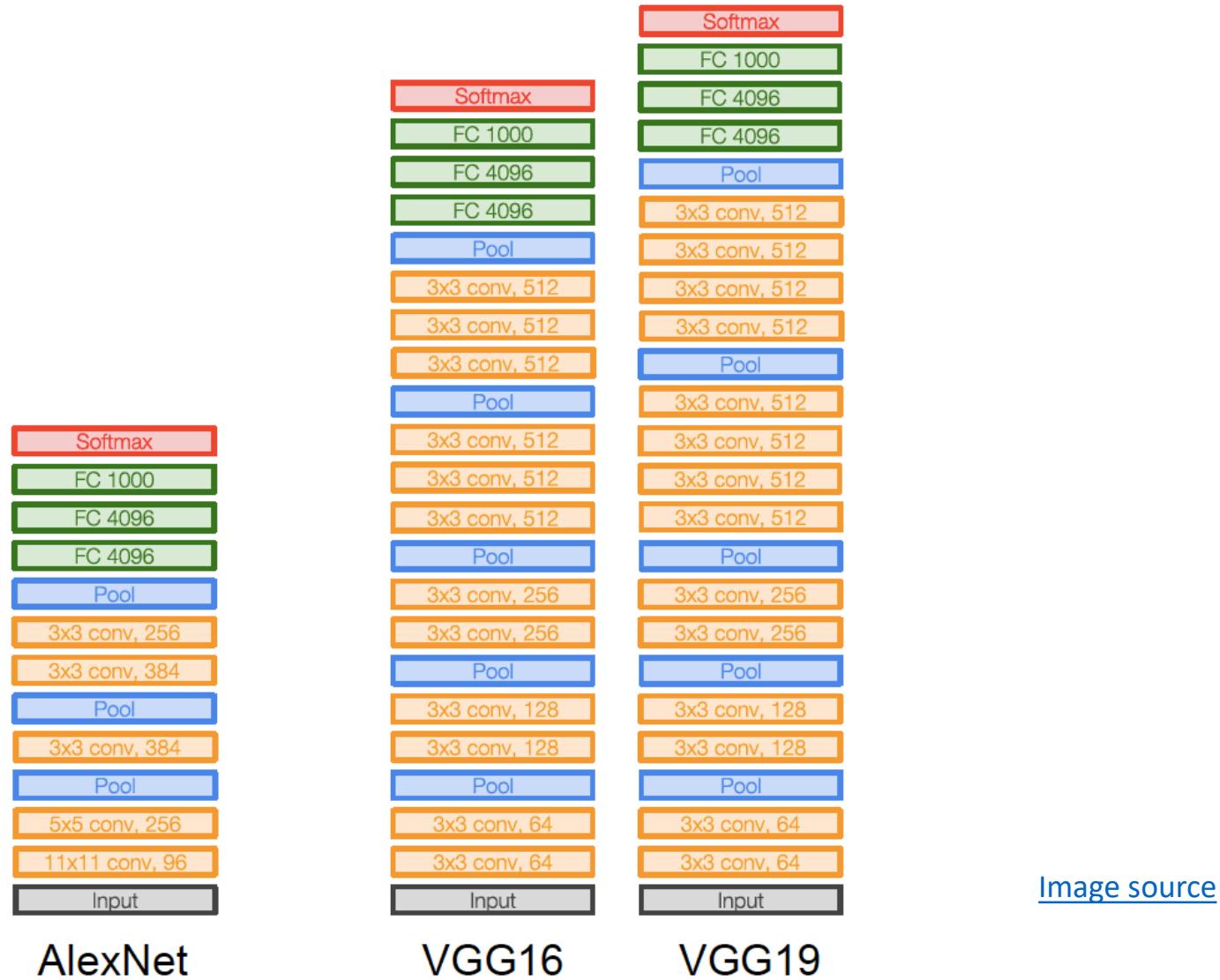
1st generation (2012-2013): AlexNet

2nd generation (2014): VGGNet, GoogLeNet

3rd generation (2015): ResNet

4th generation (2016): Wide ResNet, ResNeXt, DenseNet

VGGNet vs. AlexNet



ResNet: ImageNet 2015 winner

AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)



ResNet, **152 layers**
(ILSVRC 2015)



K. He, X. Zhang, S. Ren, and J. Sun, [Deep Residual Learning for Image Recognition](#),
CVPR 2016 (Best Paper)

A typical phenomenon

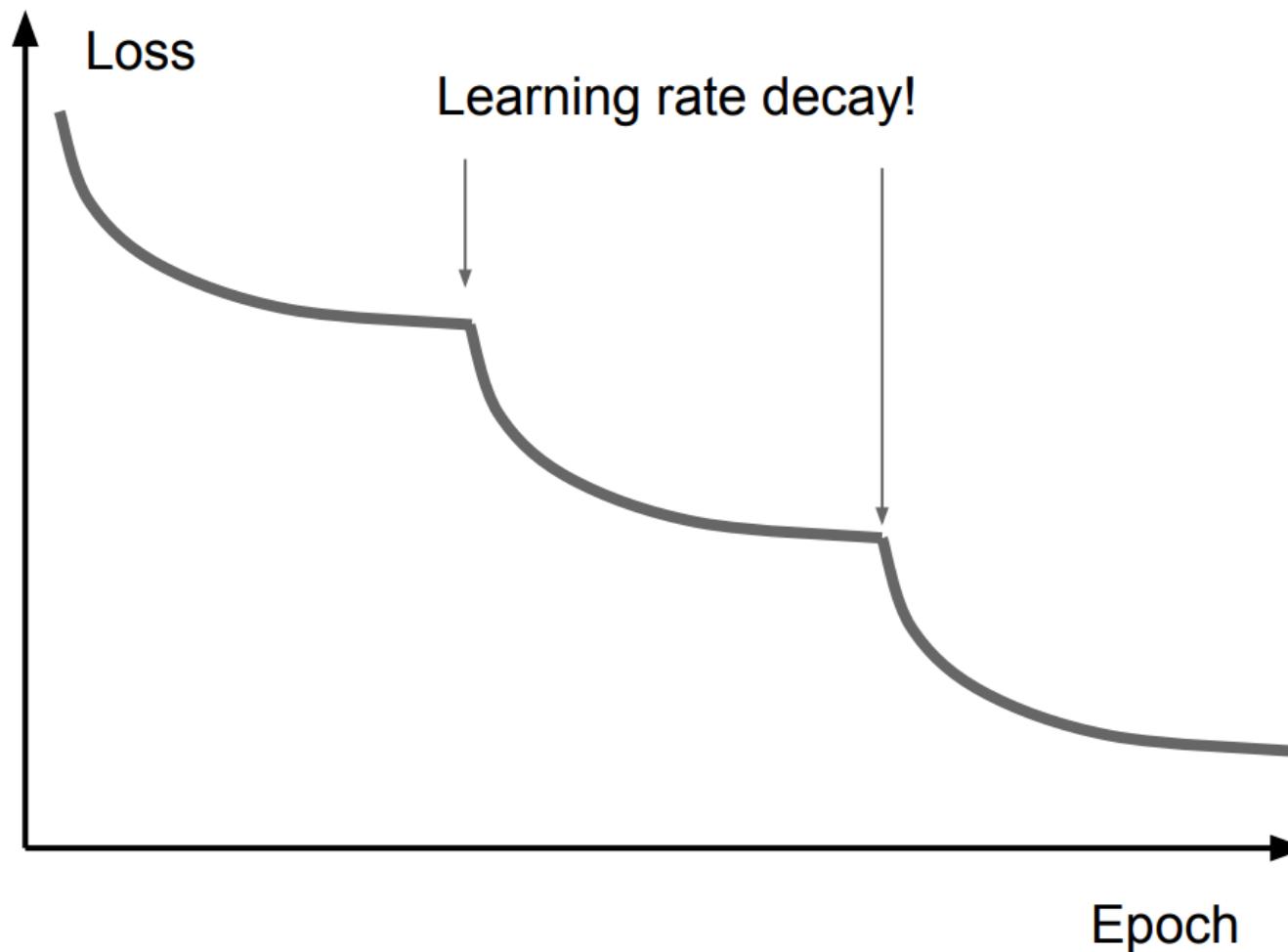


Image source: [Stanford CS231n](#)

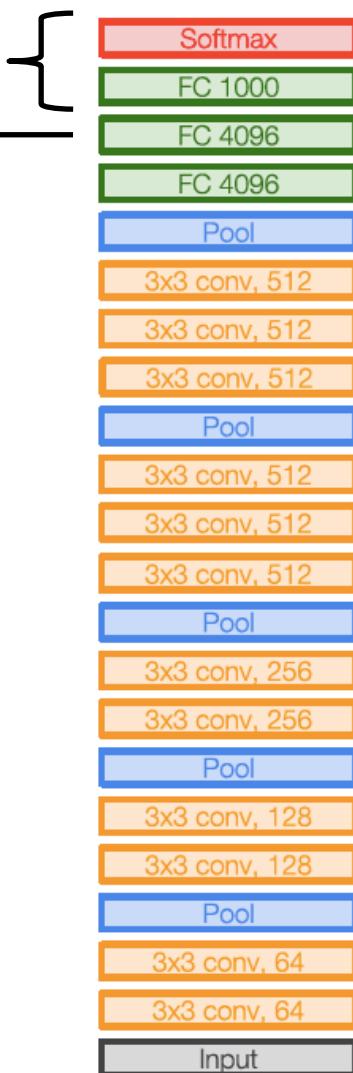
Data preprocessing

- Zero centering
 - Subtract *mean image* – all input images need to have the same resolution
 - Subtract *per-channel means* – images don't need to have the same resolution
- Optional: rescaling – divide each value by (per-pixel or per-channel) standard deviation
- Be sure to apply the same transformation at training and test time!
 - Save training set statistics and apply to test data

How to use a pre-trained network for a new task?

Remove these layers

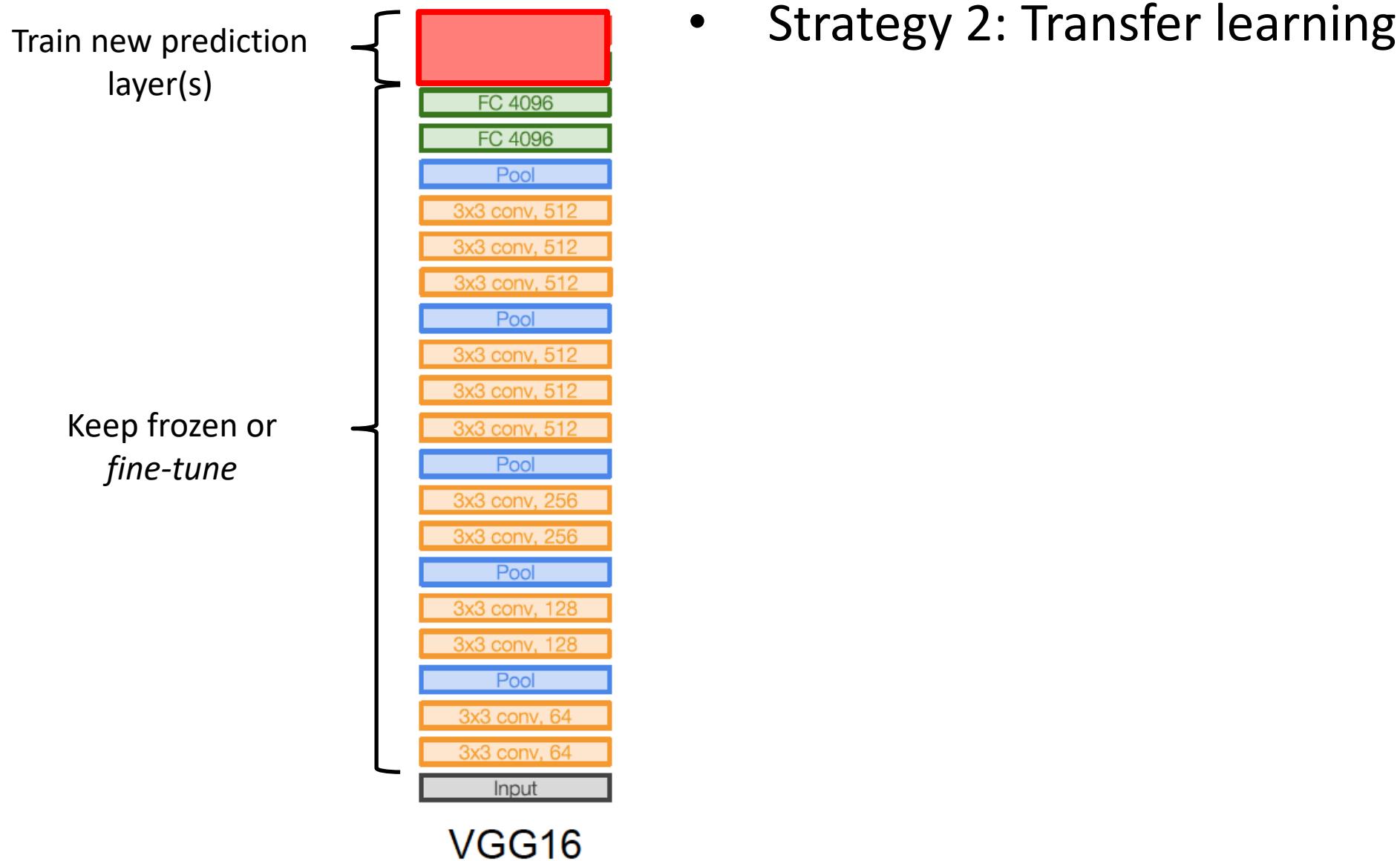
Use as off-the-shelf
feature



- Strategy 1: Use as feature extractor

VGG16

How to use a pre-trained network for a new task?



Sequence Modeling



Sequence Modeling



How to take a **variable length sequence** as input?

How to predict a **variable length sequence** as output?

Embeddings

An embedding is a mapping of a discrete — categorical — variable to a vector of continuous numbers.

In the context of neural networks, embeddings are *low-dimensional, learned* continuous vector representations of discrete variables.

Neural network embeddings are useful because they can *reduce the dimensionality* of categorical variables and *meaningfully represent* categories in the transformed space.

Embeddings

Neural network embeddings have 3 primary purposes:

- Finding nearest neighbors in the embedding space. These can be used to make recommendations based on user interests or cluster categories.
- As input to a machine learning model for a supervised task.
- For visualization of concepts and relations between categories.

Learning Embeddings

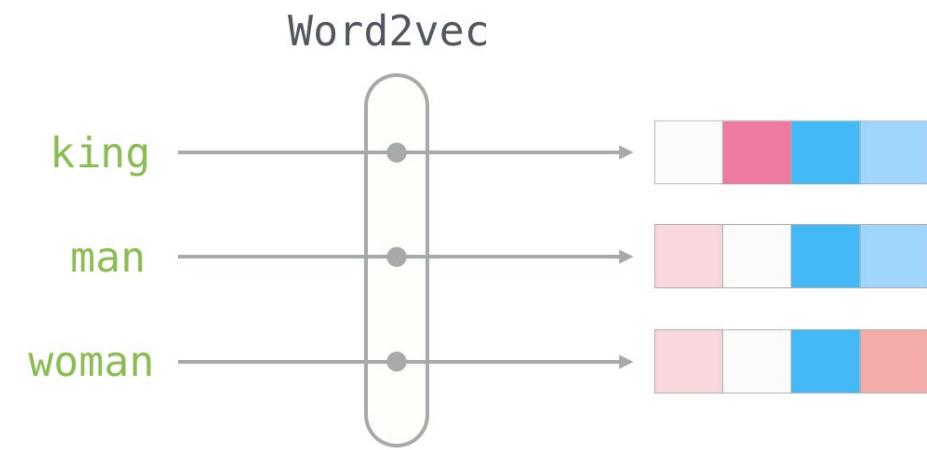
The main issue with one-hot encoding is that the transformation does not rely on any supervision.

We can greatly improve embeddings by *learning* them using a neural network on a supervised task.

The resulting embedded vectors are representations of categories where similar categories — relative to the task — are closer to one another.

Word2Vec

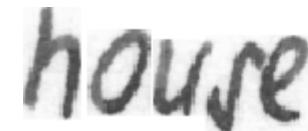
Distributed vector representation for words



Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111-3119. 2013.

Neural Nets for Sequence Modeling

Sequence Modeling: Handwritten Text Translation

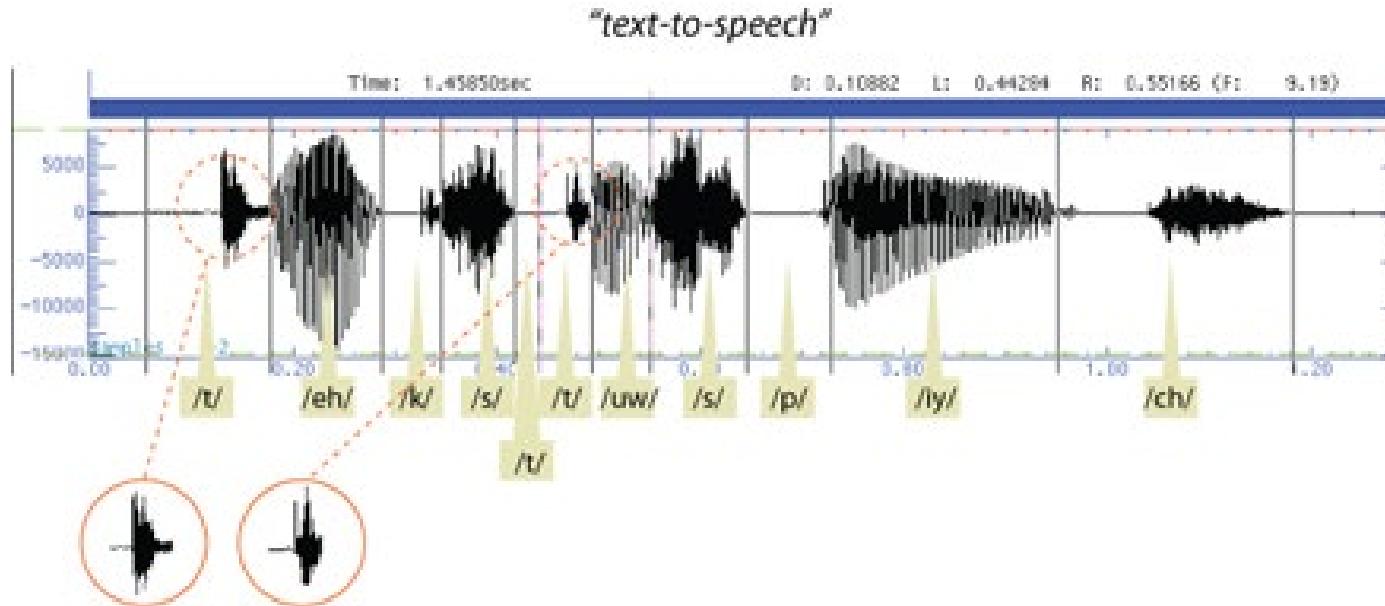
 → "house"

Winter is here. Go to
the store and buy some
snow shovels.

Winter is here. Go to the store and buy
some snow shovels.

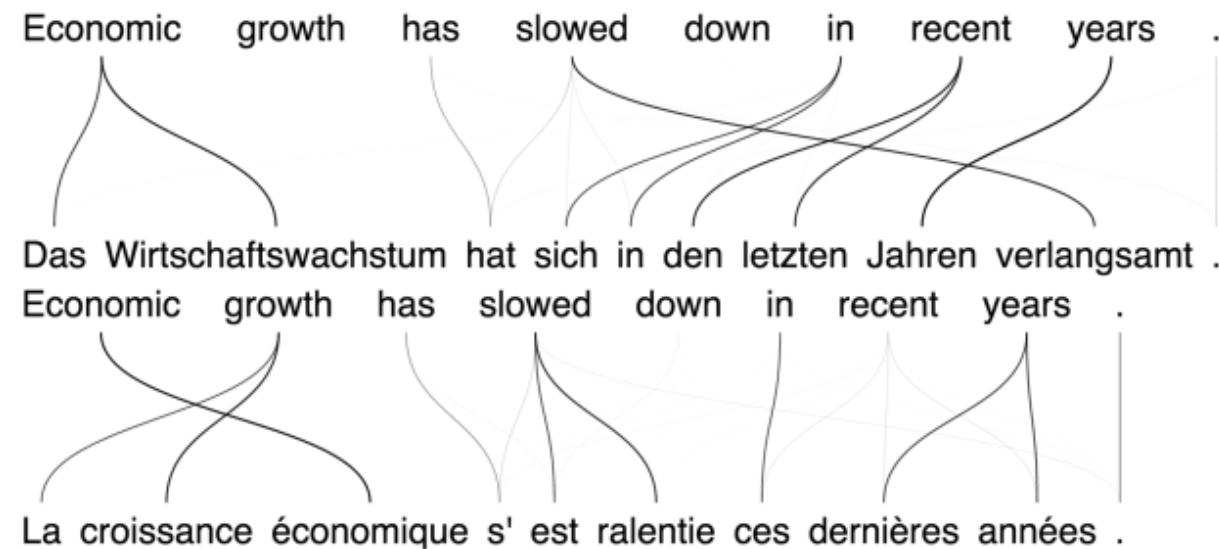
Input : Image
Output: Text

Sequence Modeling: Text-to-Speech



Input : Audio
Output: Text

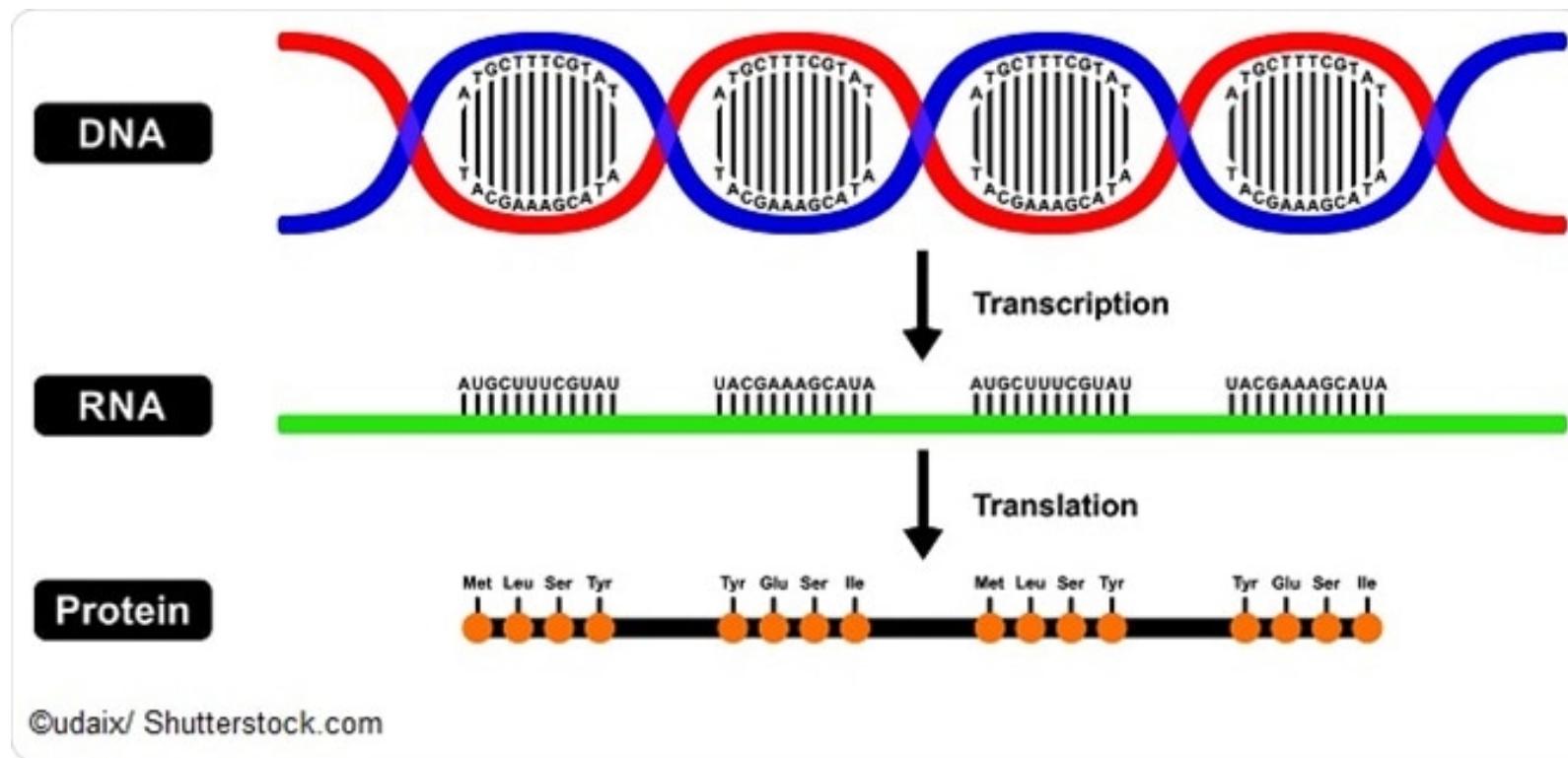
Sequence Modeling: Machine Translation



Input : Text

Output: Translated Text

Biological sequences



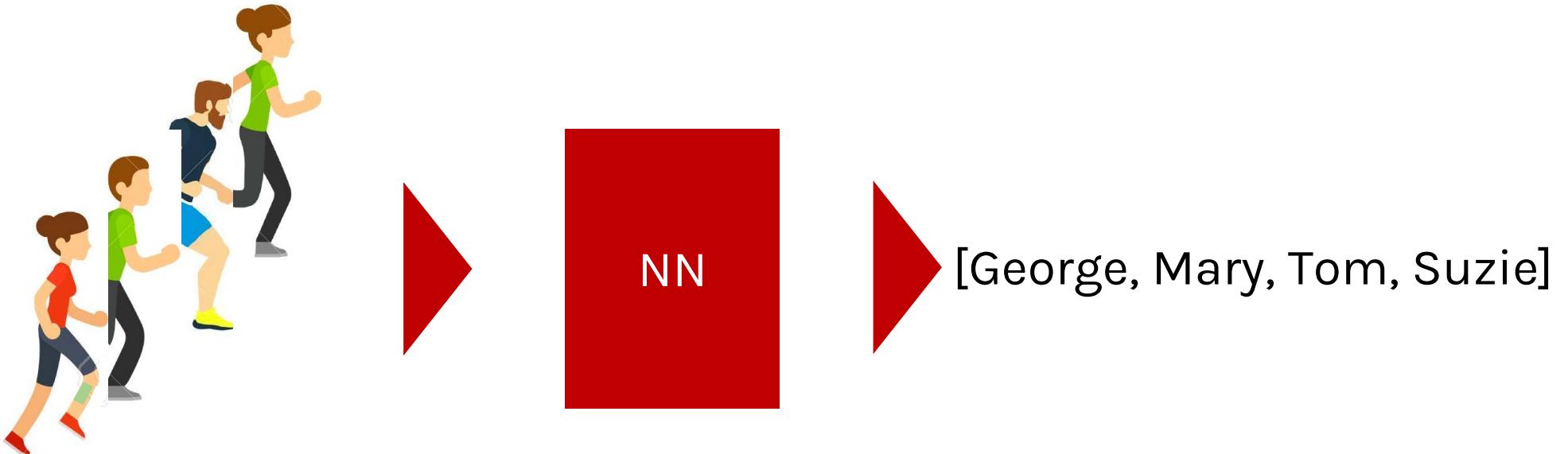
Input : Sequence

Output: biological function, site of methylation, site of translation, 3D structure, ...

Recurrent Neural Networks (RNN)

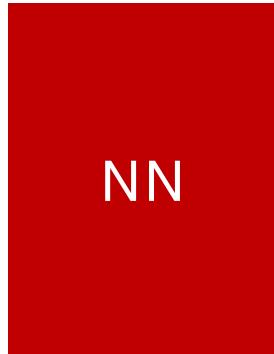
What can my NN do?

Training: Present to the NN examples and learn from them.

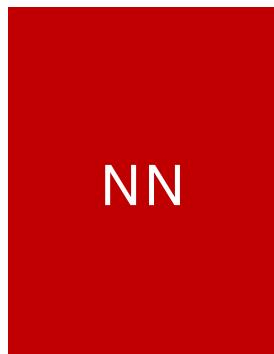


What can my NN do?

Prediction: Given an example



George

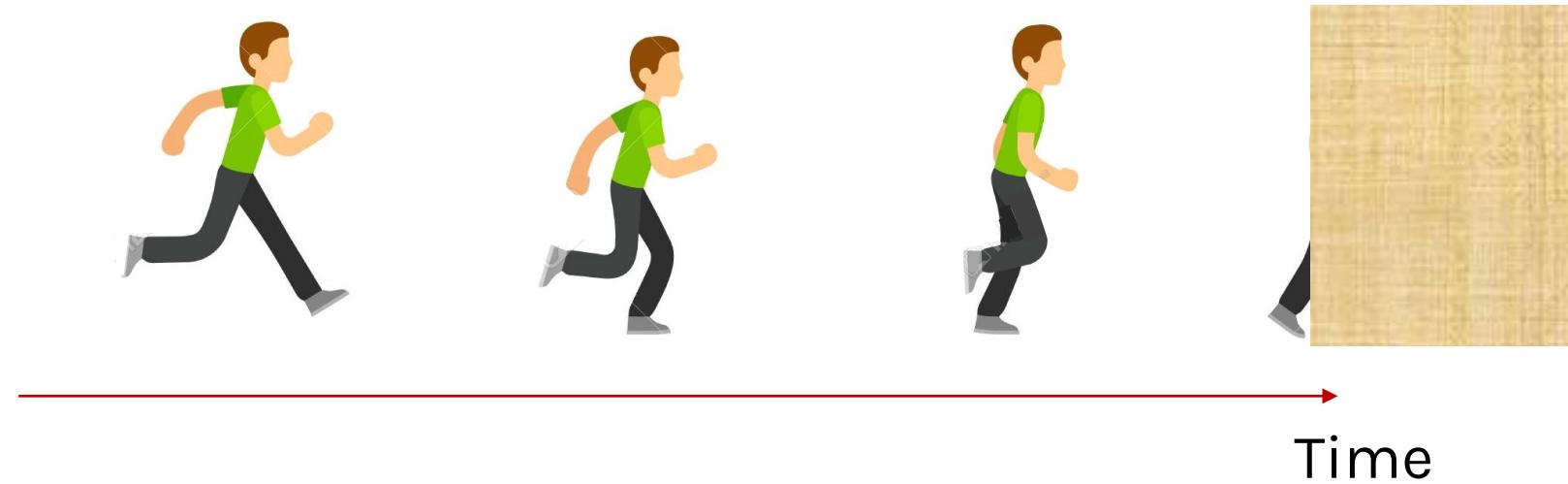


Mary

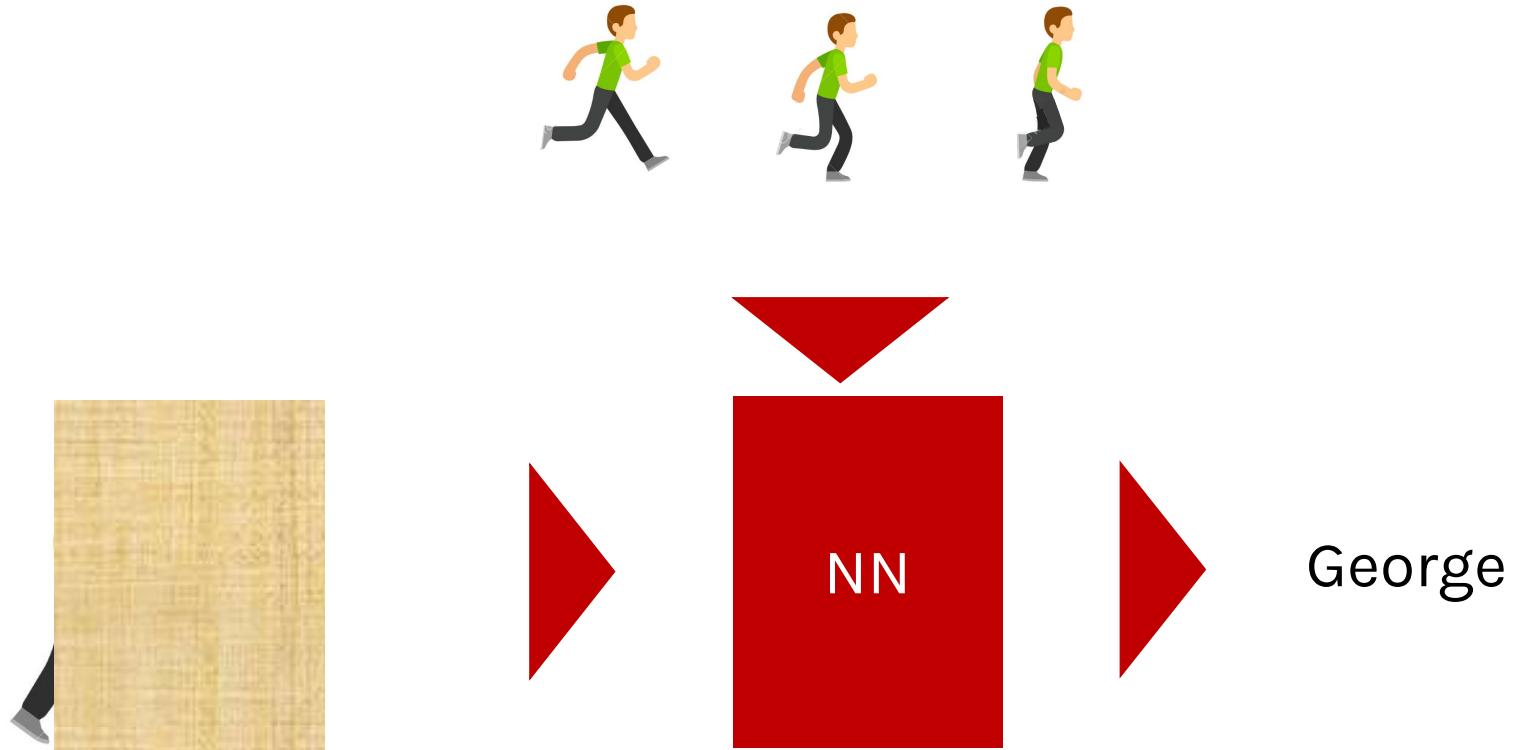
What my NN can NOT do?



Learn from previous examples

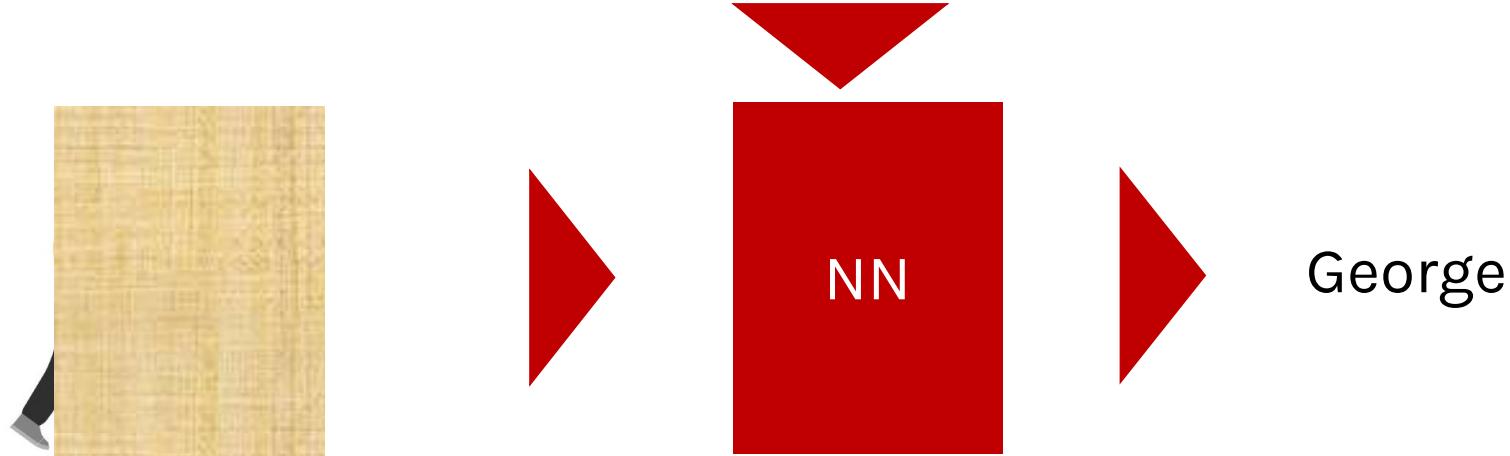


Recurrent Neural Network (RNN)



Recurrent Neural Network (RNN)

I have seen George
moving in this way
before.



- RNNs recognize the data's sequential characteristics and use patterns to predict the next likely scenario.

Recurrent Neural Network (RNN)

He told me I could have it



I do not know. I need to know who said that and what he said before. Can you tell me more?

- Our model requires context - or contextual information - to understand the subject (he) and the direct object (it) in the sentence.

RNN – Another Example with Text

- Hellen: Nice sweater Joe.
- Joe: Thanks, Hellen. It used to belong to my brother and **he told me I could have it.**



WHO IS
HE?



I see what you mean now!
The noun “he” stands for
Joe’s brother while “it” for
the sweater.

- After providing sequential information, the model understood the subject (Joe’s brother) and the direct object (sweater) in the sentence .

Sequences

- We want a machine learning model to understand sequences, not isolated samples.
- Can MLP do this?
- Assume we have a sequence of temperature measurements and we want to take 3 sequential measurements and predict the next one

	features
samples	
1	35
2	32
3	45
4	48
5	41
6	39
7	36
...	...

Sequences

- We want a machine learning model to understand sequences, not isolated samples.
- Can MLP do this?
- Assume we have a sequence of temperature measurements and we want to take 3 sequential measurements and predict the next one

	features
samples	
1	35
2	32
3	45
4	48
5	41
6	39
7	36
...	...

1	35
2	32
3	45
4	48

Sequences

- We want a machine learning model to understand sequences, not isolated samples.
- Can MLP do this?
- Assume we have a sequence of temperature measurements and we want to take 3 sequential measurements and predict the next one

	features
samples	
1	35
2	32
3	45
4	48
5	41
6	39
7	36
...	...

1	35
2	32
3	45
4	48

2	32
3	45
4	48

3	45
4	48
5	41

Sequences

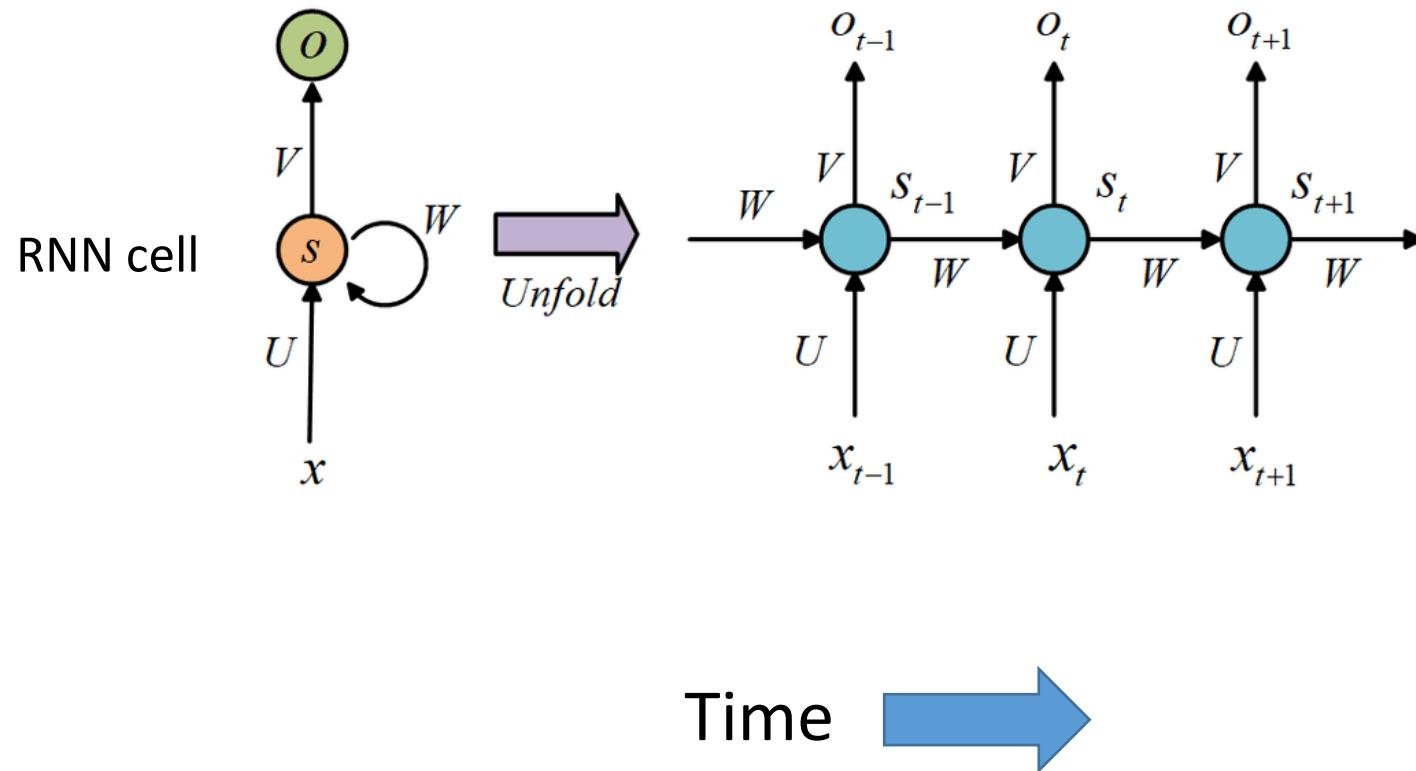
- We want a machine learning model to understand sequences, not isolated samples.
- Can MLP do this?
- Assume we have a sequence of temperature measurements and we want to take 3 sequential measurements and predict the next one

	samples	features
1		35
2		32
3		45
4		48
5		41
6		39
7		36
...		...
1		35
2		32
3		45
4	4	48
2		32
3		45
4	5	41
3		45
4		48
5		41
6	6	39

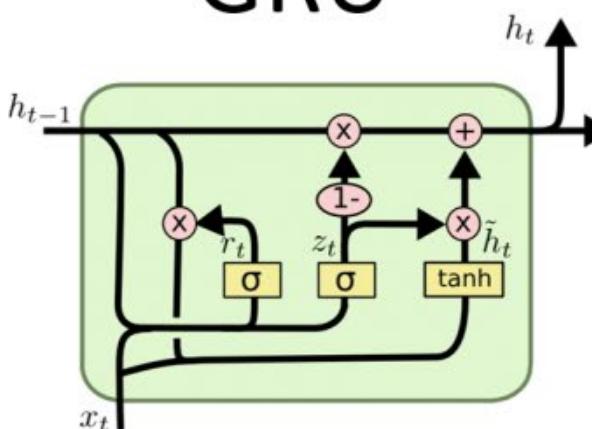
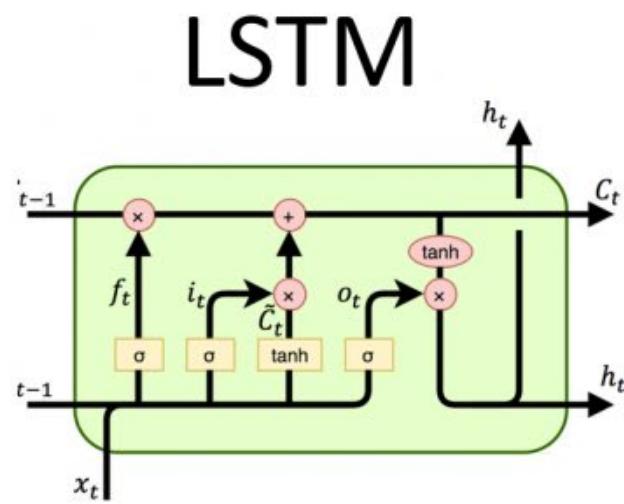
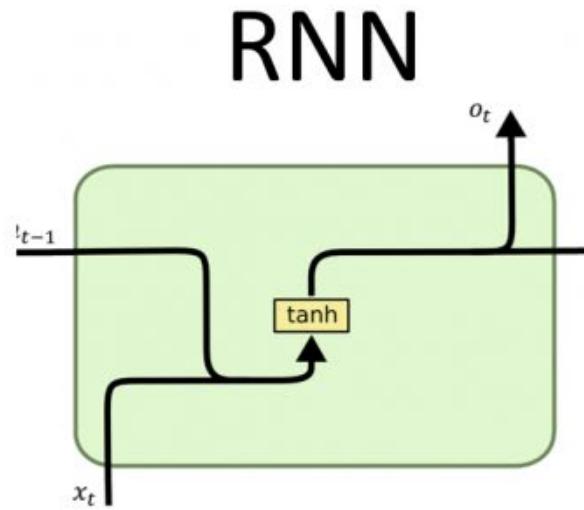
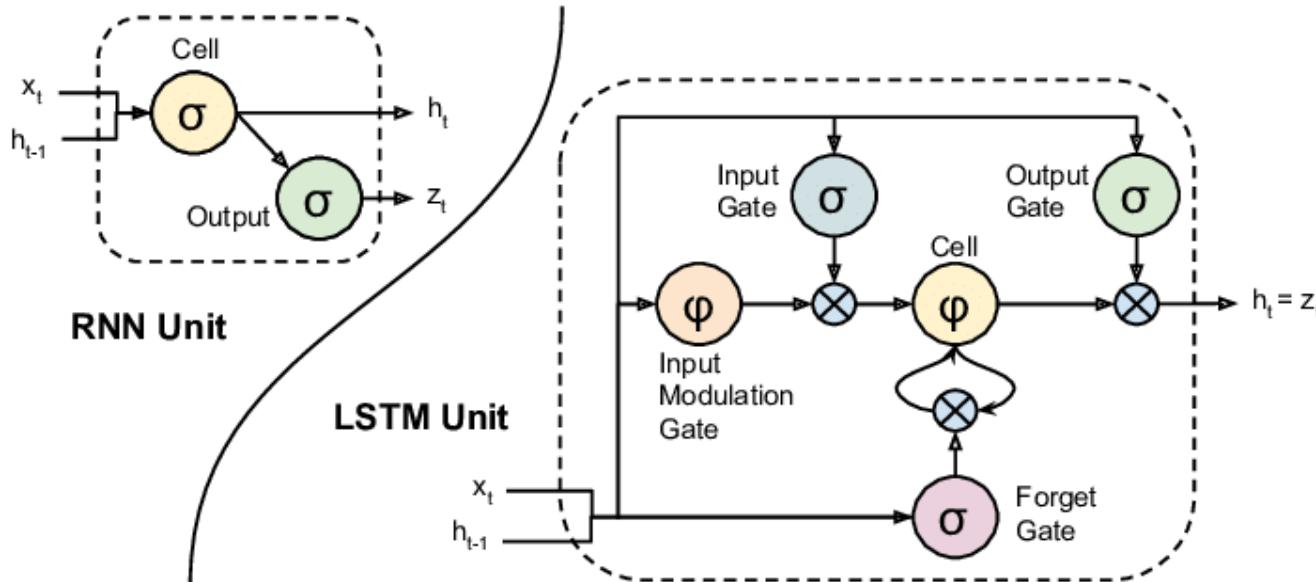
Why not CNNs or MLPs?

1. MLPs/CNNs require fixed input and output size
2. MLPs/CNNs can't classify inputs in multiple places

Main Concept of RNNs

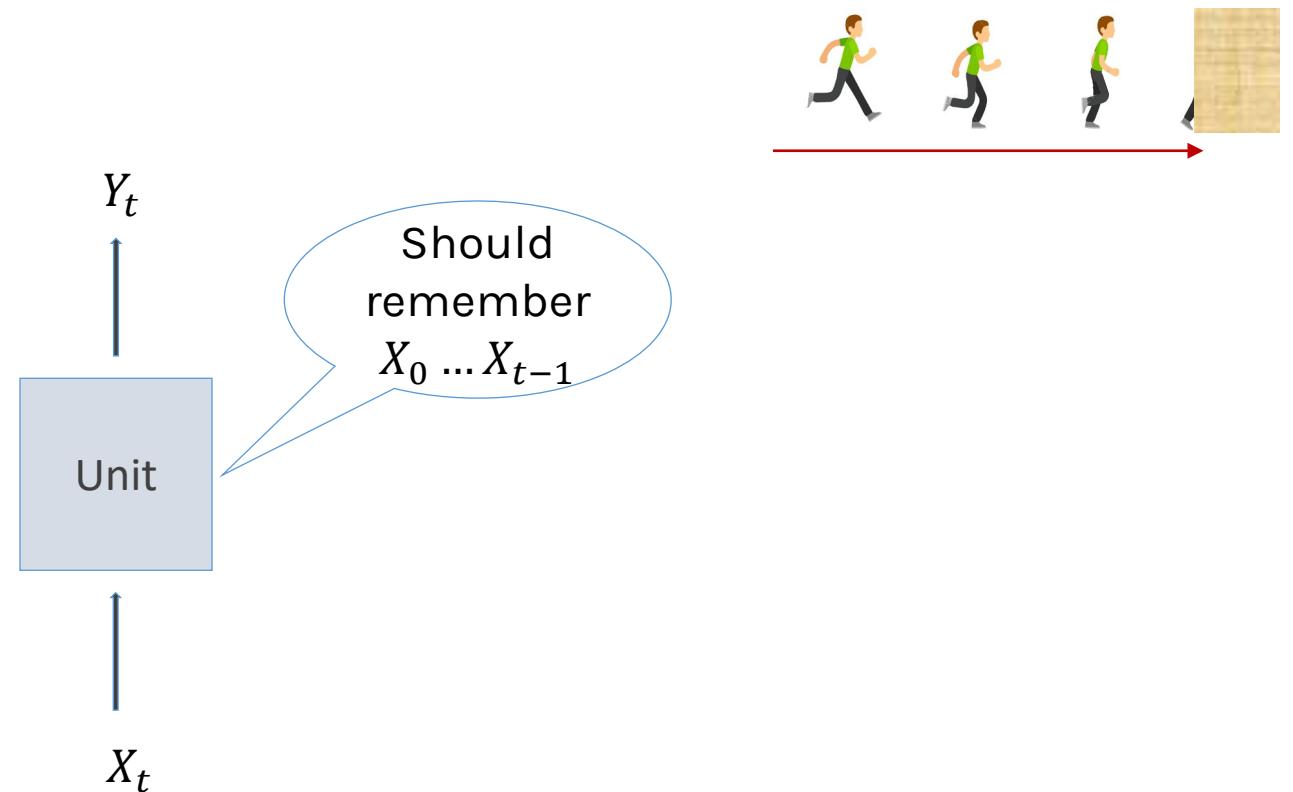


RNN Cells



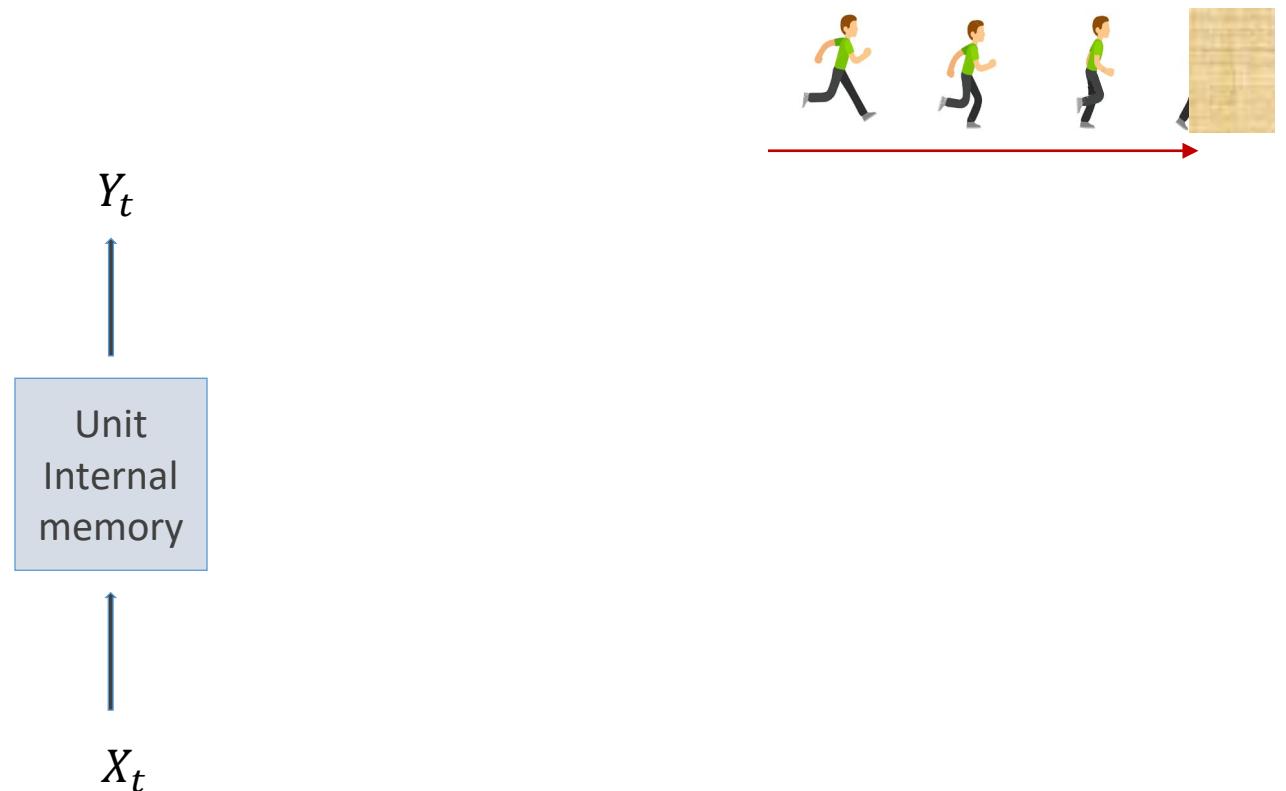
Memory

- Somehow the computational unit should remember what it has seen before.



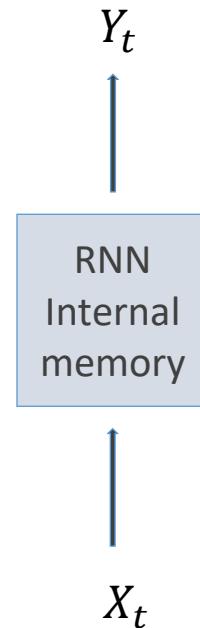
Memory

- Somehow the computational unit should remember what it has seen before.



Memory

- Somehow the computational unit should remember what it has seen before.
- We'll call the information the unit's **state**.

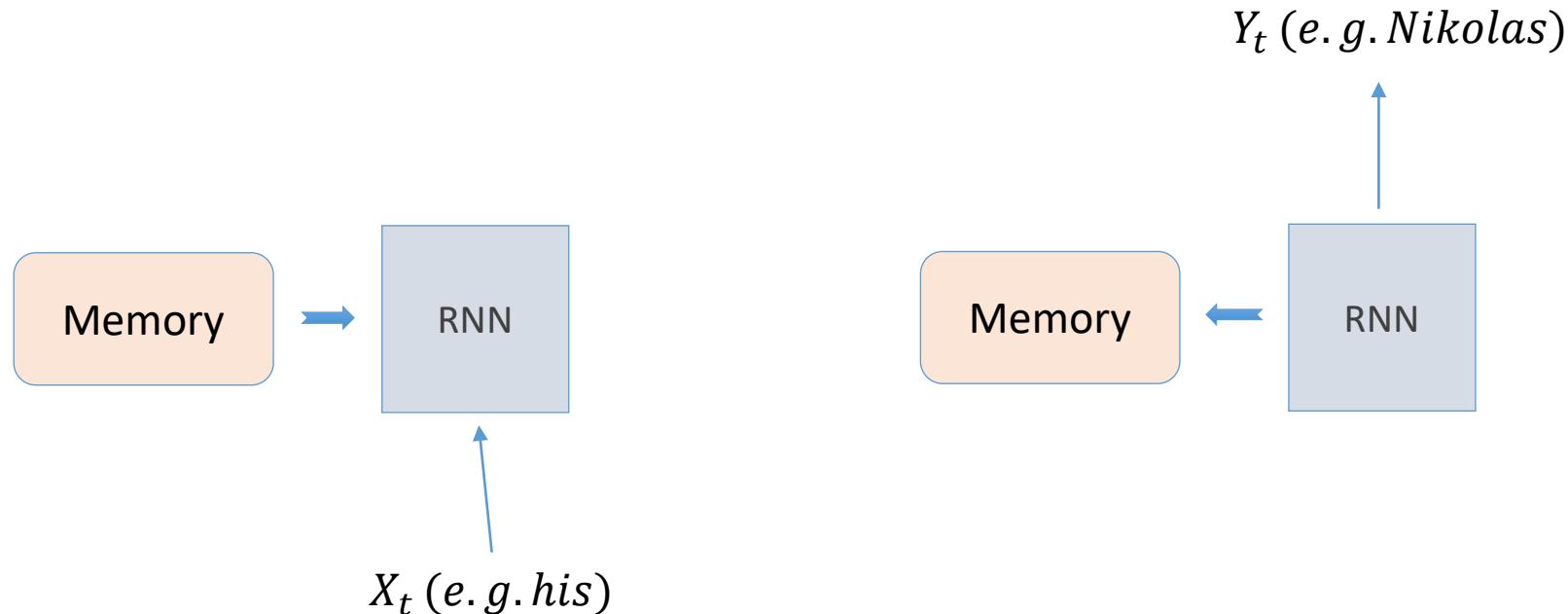


Memory

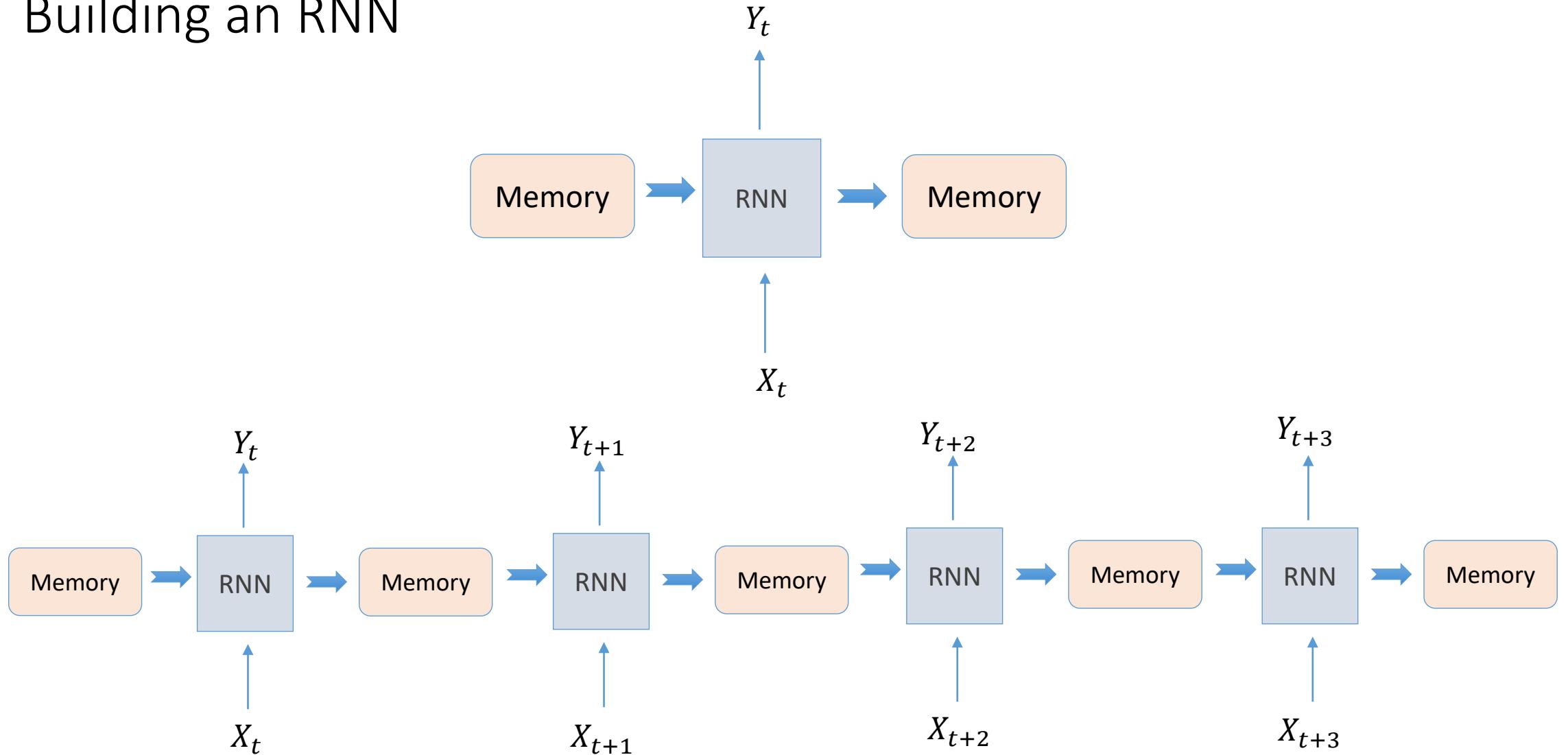
- In neural networks, once training is over, the weights do not change. This means that the network is done learning and done changing.
- Then, we feed in values, and it simply applies the operations that make up the network, using the values it has learned.
- But the RNN units are able to remember new information after training has completed.
- **That is, they're able to keep changing after training is over.**

Memory

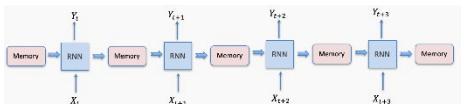
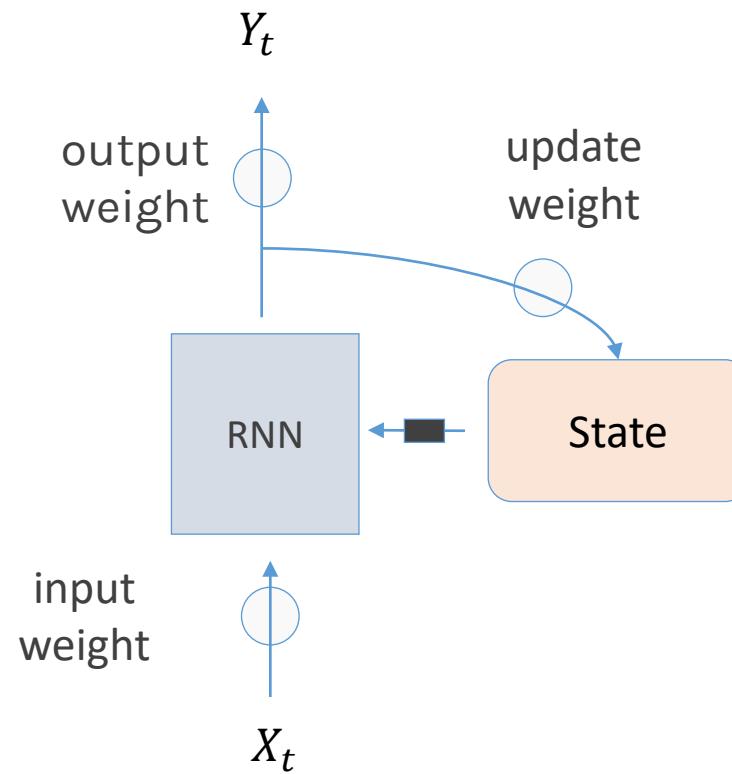
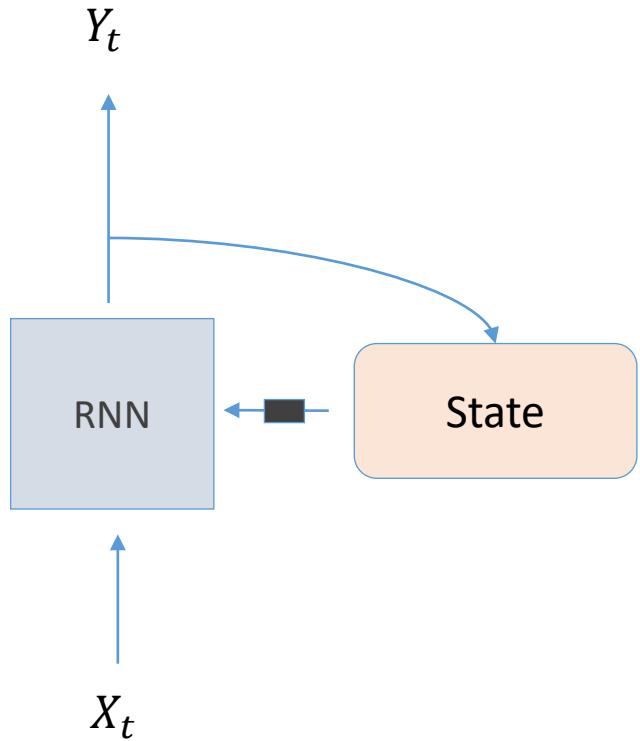
- Question: How can we do this? How can build a unit that remembers the past?
- The memory or **state** can be written to a file but in RNNs, we keep it inside the recurrent unit.
- In an array or in a vector!



Building an RNN



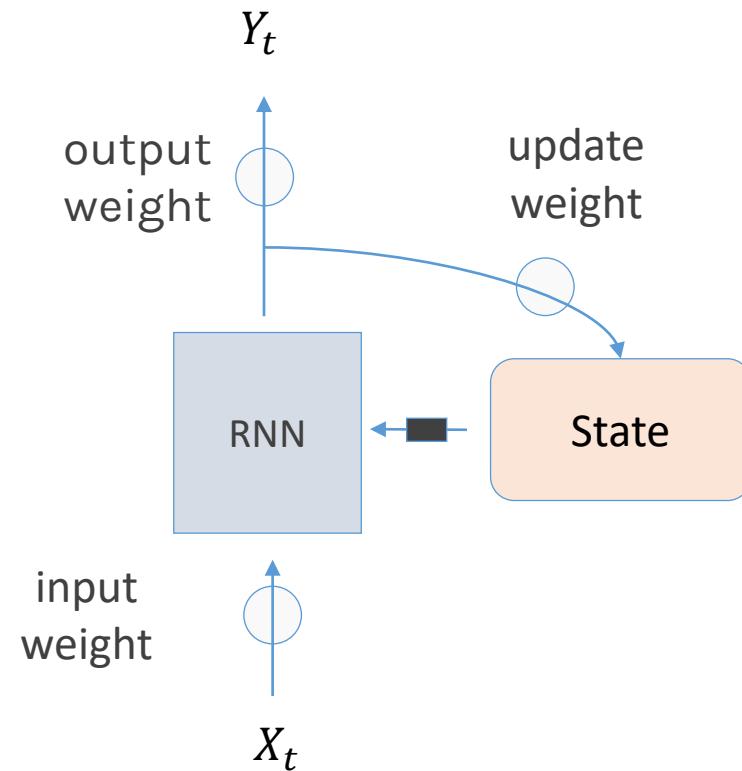
Structure of an RNN cell



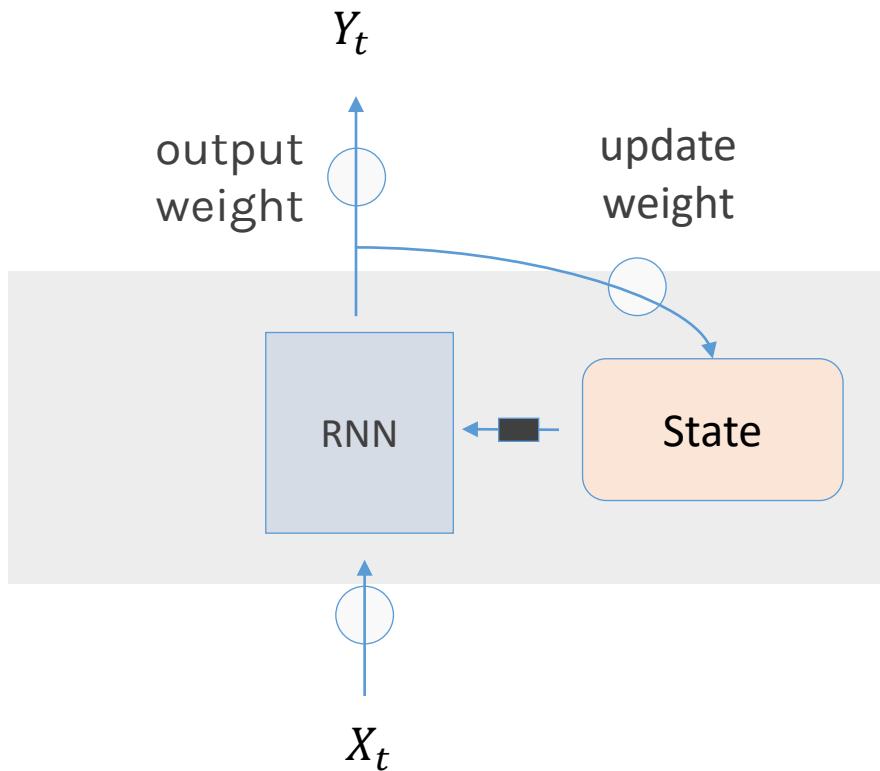
Backprop Through Time

- For each input, unfold network for the sequence length T
- Back-propagation: apply forward and backward pass on unfolded network
- Memory cost: $O(T)$

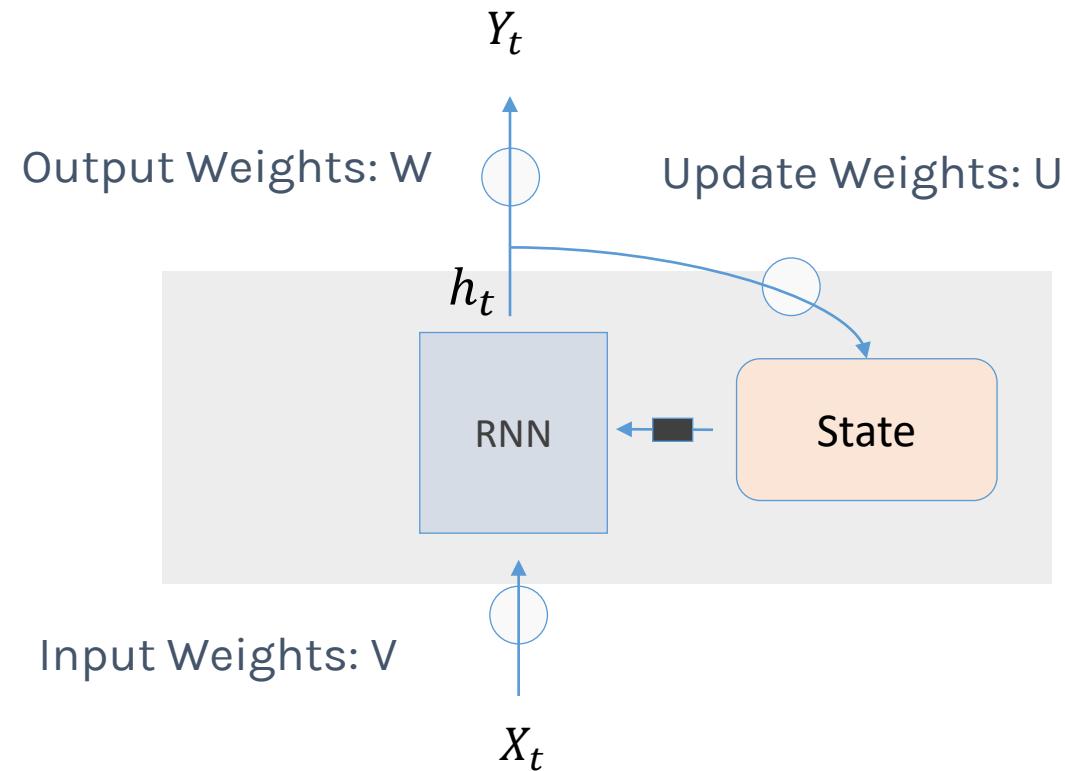
Backprop Through Time



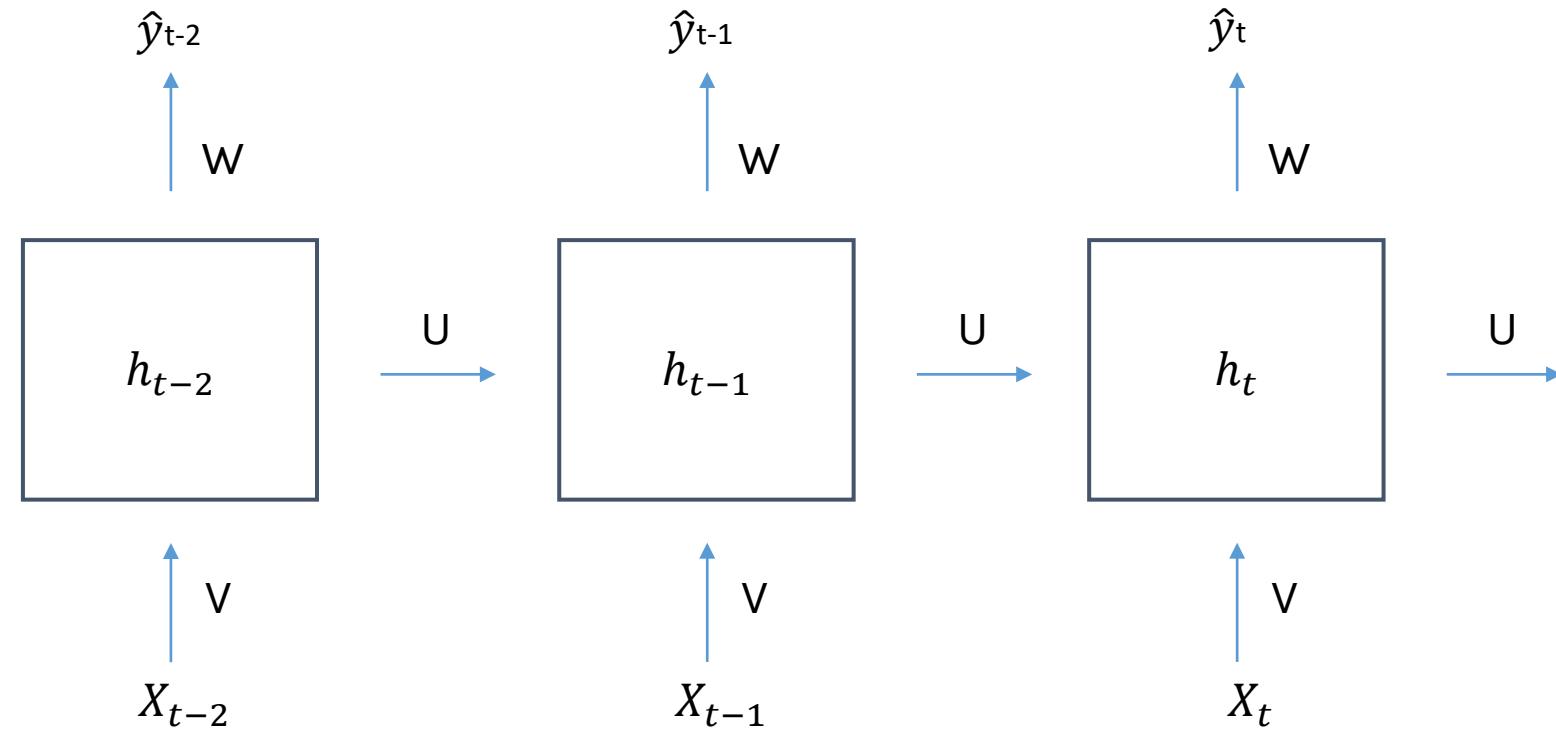
Backprop Through Time



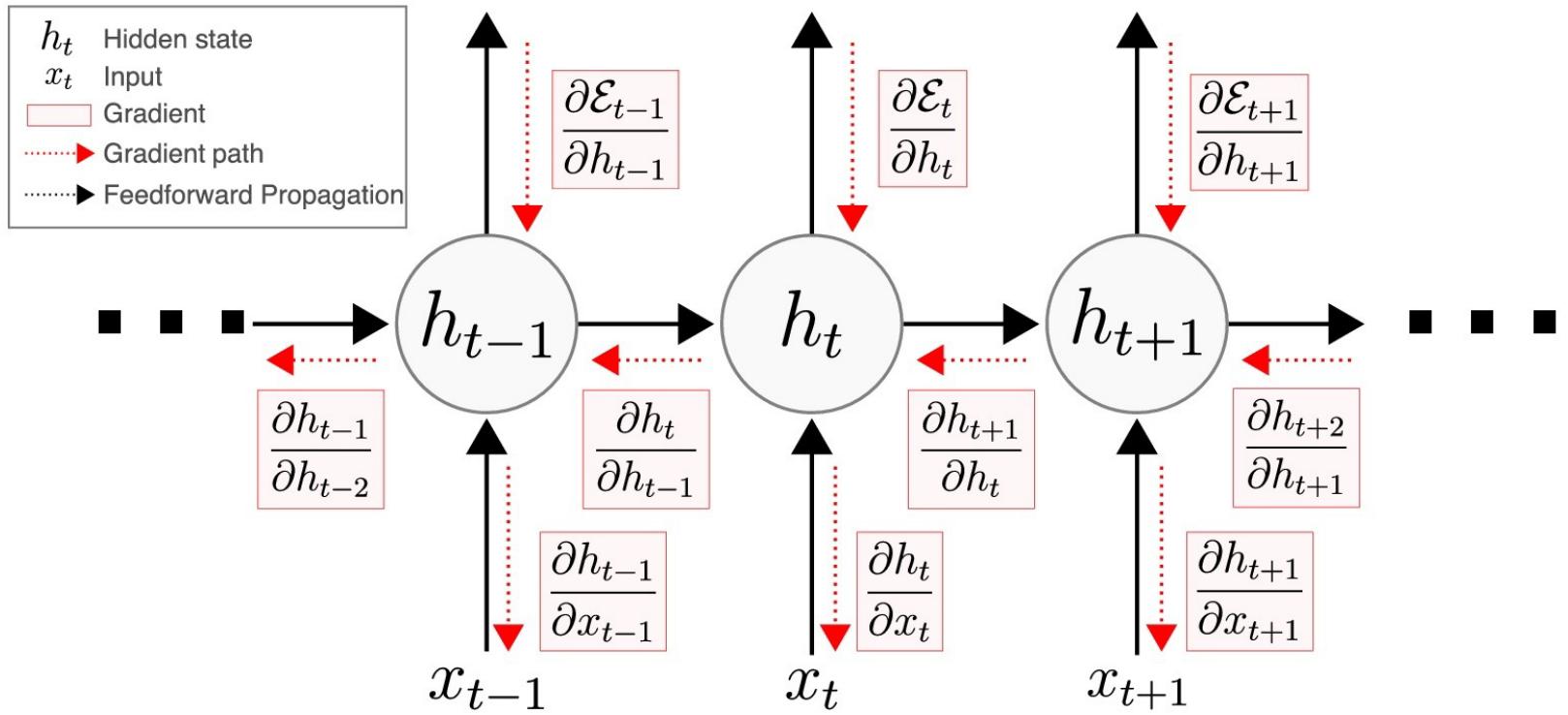
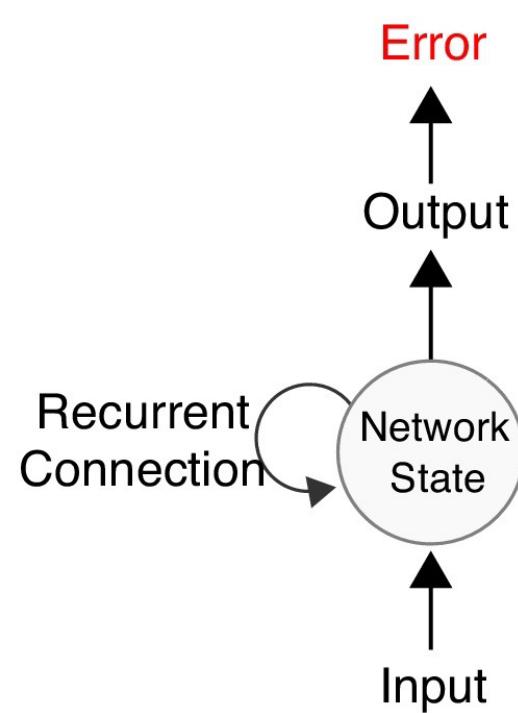
Backprop Through Time



Backprop Through Time



You have two activation functions g_h which serves as the activation for the hidden state and g_y which is the activation of the output. In the example shown before g_y was the identity.



Current Opinion in Neurobiology

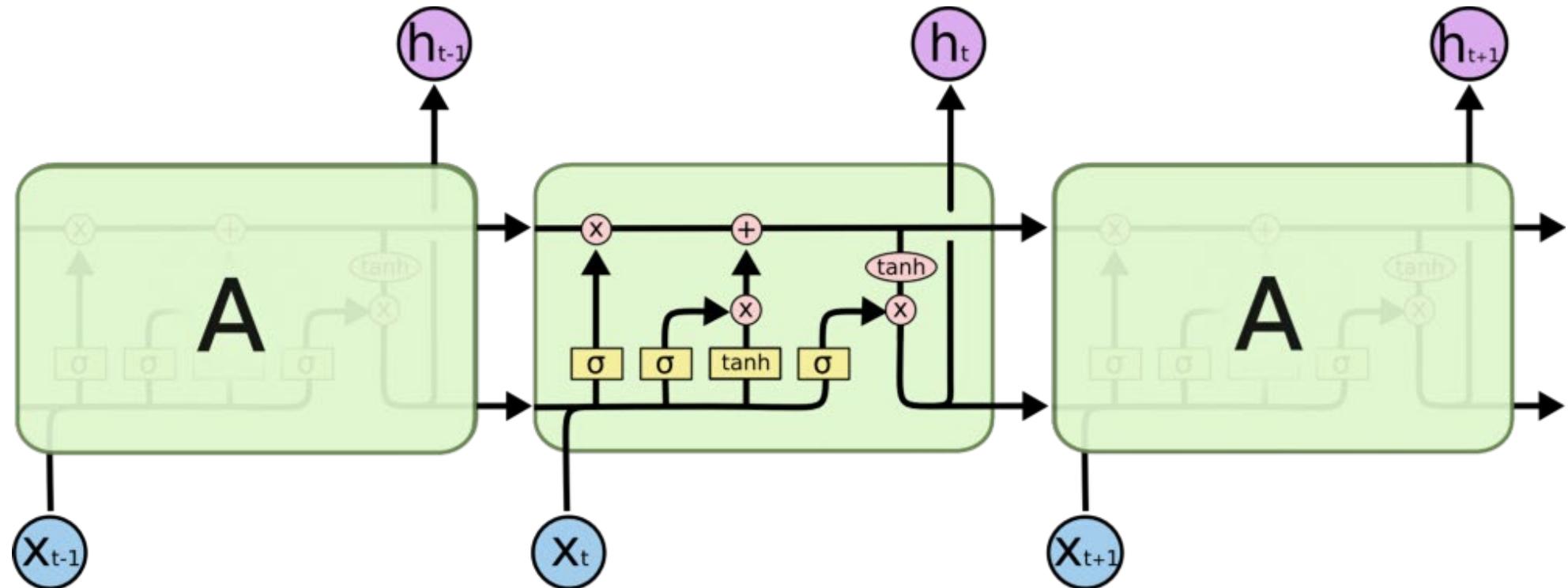
Long-term Dependencies

- Unfolded networks can be very deep
- Long-term interactions are given exponentially smaller weights than small-term interactions
- Gradients tend to either *vanish* or *explode*

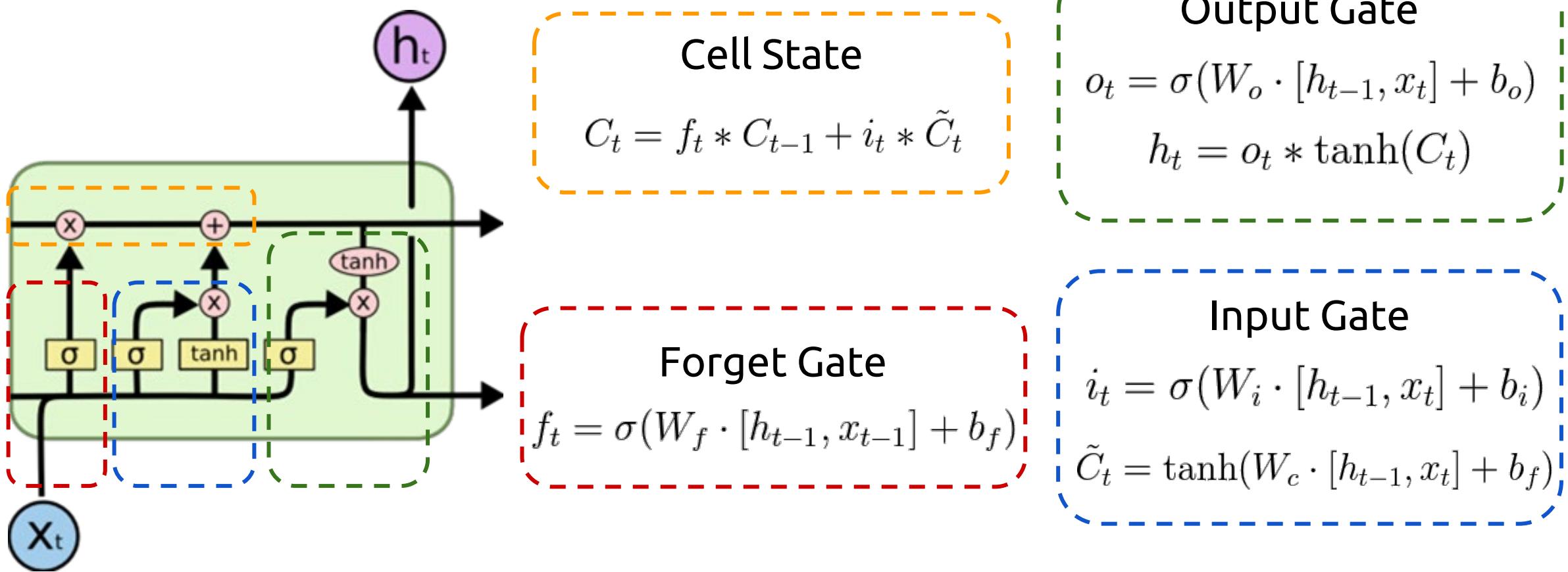
Long Short-Term Memory

- Handles long-term dependencies
- Leaky units where weight on self-loop α is context-dependent
- Allow network to decide whether to accumulate or forget past info

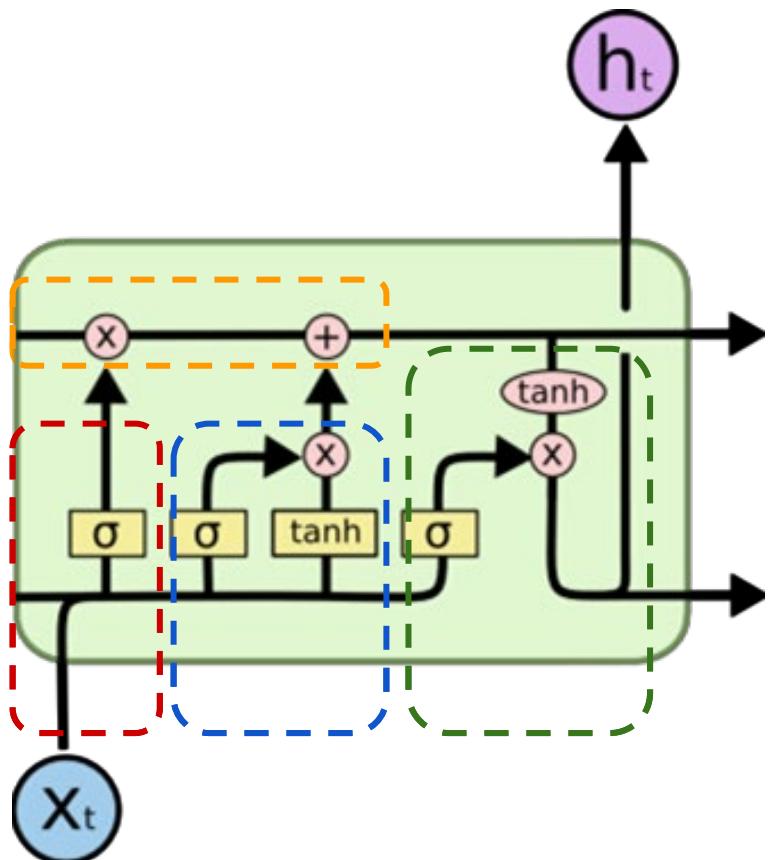
LSTM: Long short term memory



LSTM big picture ...

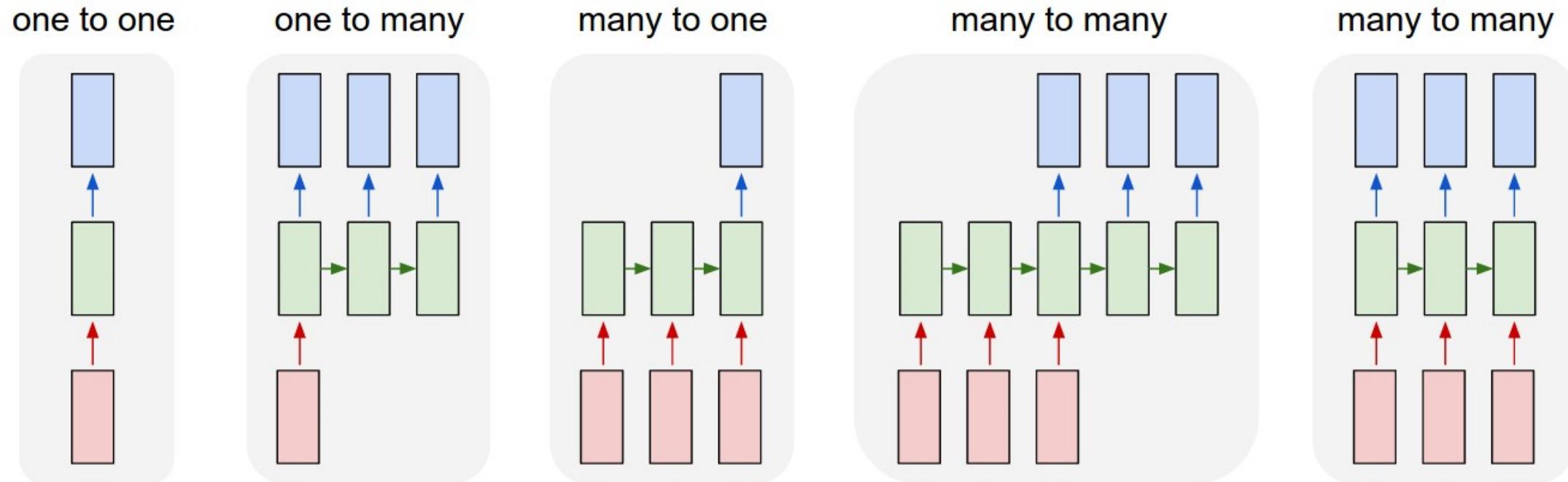


LSTM big picture ...

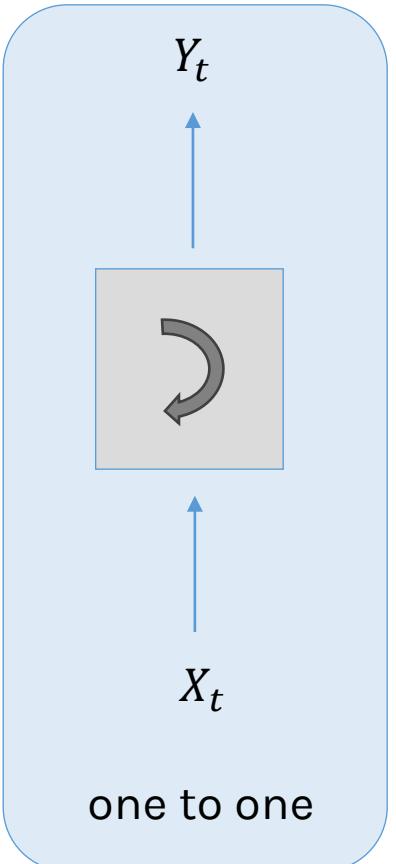


1. LSTM are recurrent neural network with a cell and a hidden state, boths of these are updated in each step and can be thought as memories.
2. Cell states work as a long term memory and the updates depends on the relation between the hidden state in $t - 1$ and the input.
3. The hidden state of the next step is a transformation of the cell state and the output (which is the section that is in general used to calculate our loss, ie information that we want in a short memory).

What makes Recurrent Networks so special?

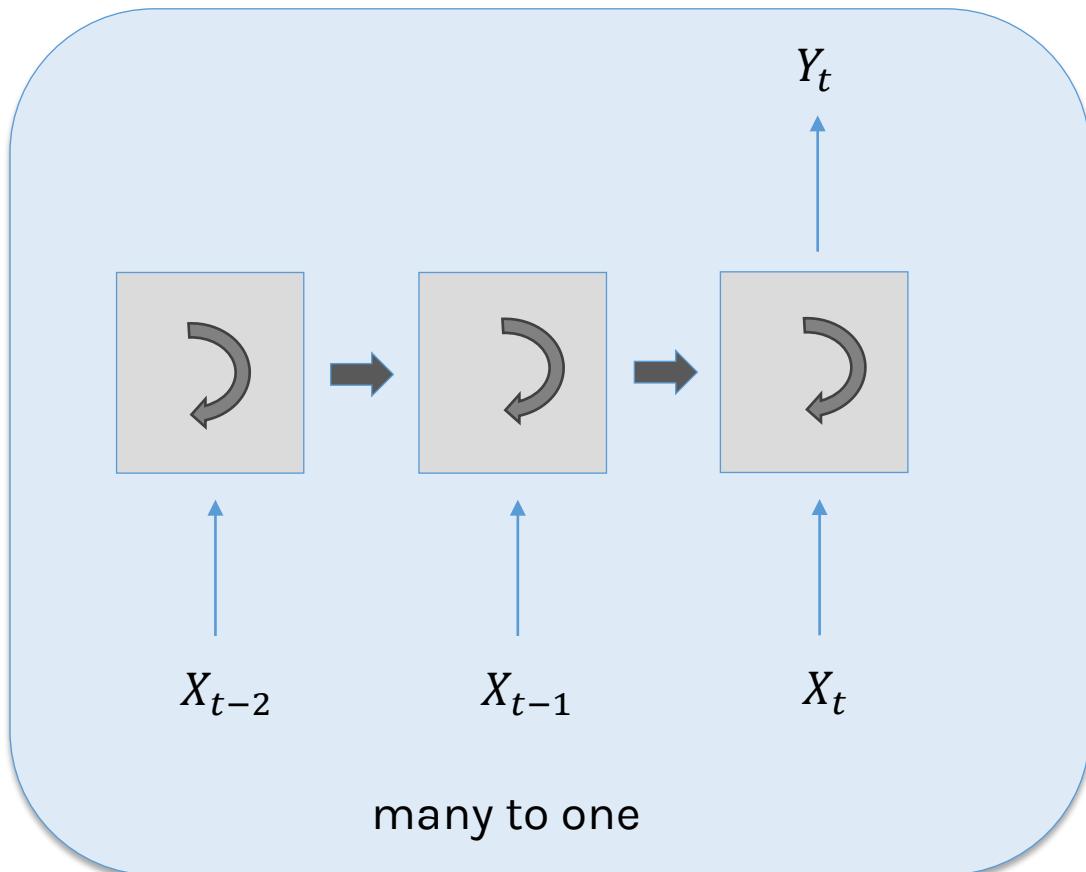


RNN Structures



- The **one to one** structure is useless.
- It takes a single input and it produces a single output.
- Not useful because the RNN cell is making little use of its unique ability to remember things about its input sequence

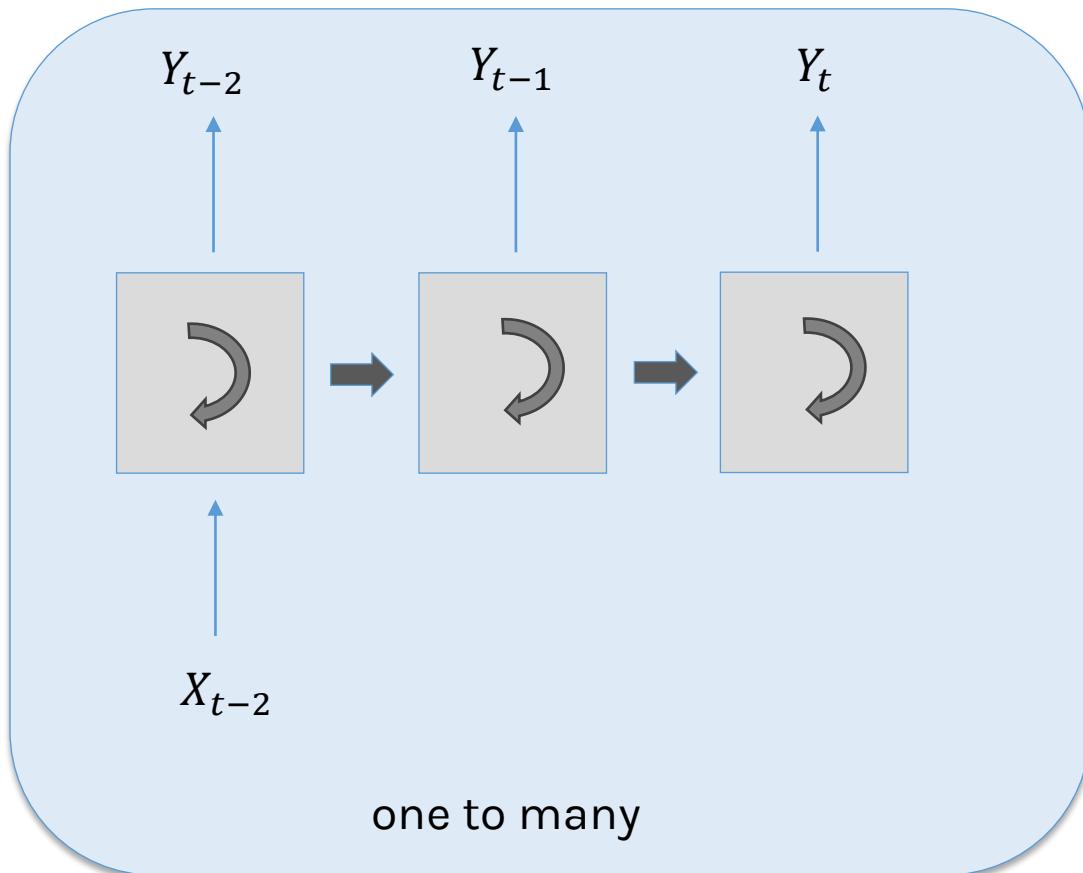
RNN Structures (cont)



The **many to one** structure reads in a sequence and gives us back a single value.

Example: Sentiment analysis, where the network is given a piece of text and then reports on some quality inherent in the writing. A common example is to look at a movie review and determine if it was positive or negative.

RNN Structures (cont)

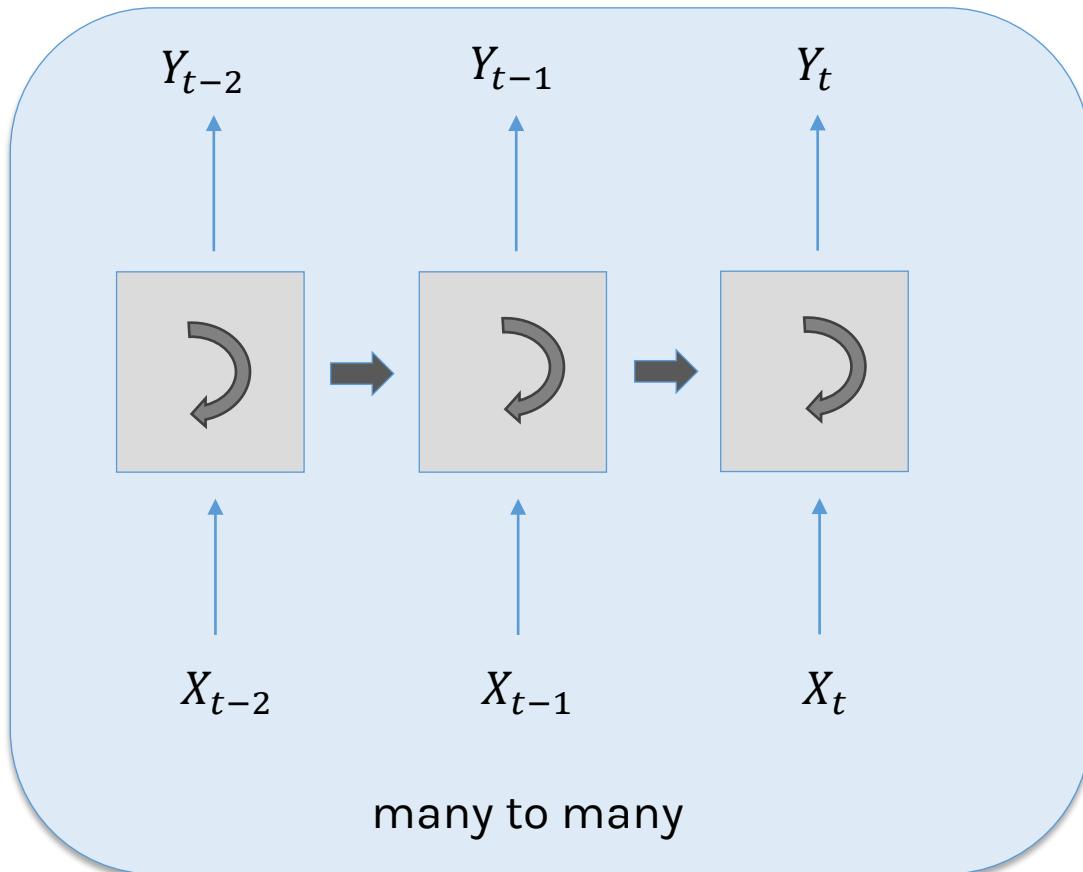


The **one to many** takes in a single piece of data and produces a sequence.

For example we give it the starting note for a song, and the network produces the rest of the melody for us.

Sequence generative models

RNN Structures (cont)



The **many to many** structures are in some ways the most interesting. used for machine translation.

Seq2seq models

Machine Translation

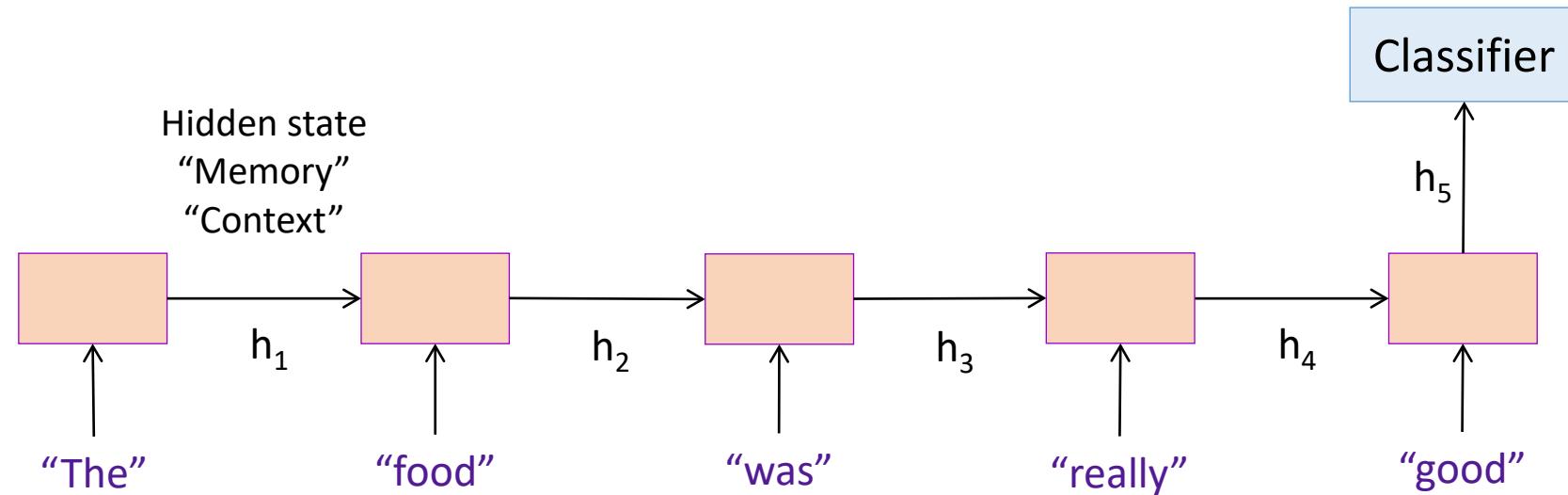
...

Ref

- [Chen17b] Qiming Chen, Ren Wu, “CNN Is All You Need”, arXiv 1712.09662, 2017. <https://arxiv.org/abs/1712.09662>
- [Chu17] Hang Chu, Raquel Urtasun, Sanja Fidler, “Song From PI: A Musically Plausible Network for Pop Music Generation”, arXiv preprint, 2017. <https://arxiv.org/abs/1611.03477>
- [Johnson17] Daniel Johnson, “Composing Music with Recurrent Neural Networks”, Heahedria, 2017. <http://www.hexahedria.com/2015/08/03/composing-music-with-recurrent-neural-networks/>
- [Deutsch16b] Max Deutsch, “Silicon Valley: A New Episode Written by AI”, Deep Writing blog post, 2017. <https://medium.com/deep-writing/silicon-valley-a-new-episode-written-by-ai-a8f832645bc2>
- [Fan16] Bo Fan, Lijuan Wang, Frank K. Soong, Lei Xie “Photo-Real Talking Head with Deep Bidirectional LSTM”, Multimedia Tools and Applications, 75(9), 2016. https://www.microsoft.com/en-us/research/wp-content/uploads/2015/04/icassp2015_fanbo_1009.pdf

Example 1: Sentiment classification

- Recurrent model:



Example 2: Text generation

RNN Bible
@RNN_Bible
Random bible verses generated using Recurrent Neural Networks (char-rnn).
Joined May 2015

Tweets Following Followers
2,197 1 485

Tweets **Tweets & replies**

RNN Bible @RNN_Bible · 20 Jun 2016
24:11 Thus saith the LORD of hosts; Ask now this stones are for the righteous and the children of Israel.
1 2 3

RNN Bible @RNN_Bible · 19 Jun 2016
24:16 And they took up twelve stones out of the city of David, and discomfit Jordan.
1 2 1

RNN Bible @RNN_Bible · 19 Jun 2016
3:20 And the LORD shall send a proverb against the LORD thy God, and shalt not each laugh.
1 5 3

RNN Bible @RNN_Bible · 19 Jun 2016
23:2 And the vision of the breaking thereof shall be in rubbick, and they shall take away the stones out of the land.
1 2 1

DeepDrumpf
@DeepDrumpf
I'm a Neural Network trained on Trump's transcripts. Priming text in []s. Donate (gofundme.com/deepdrumpf) to interact! Created by @hayesbh.
Joined March 2016
7 Following 24.6K Followers

Tweets **Tweets & replies** **Media** **Likes**

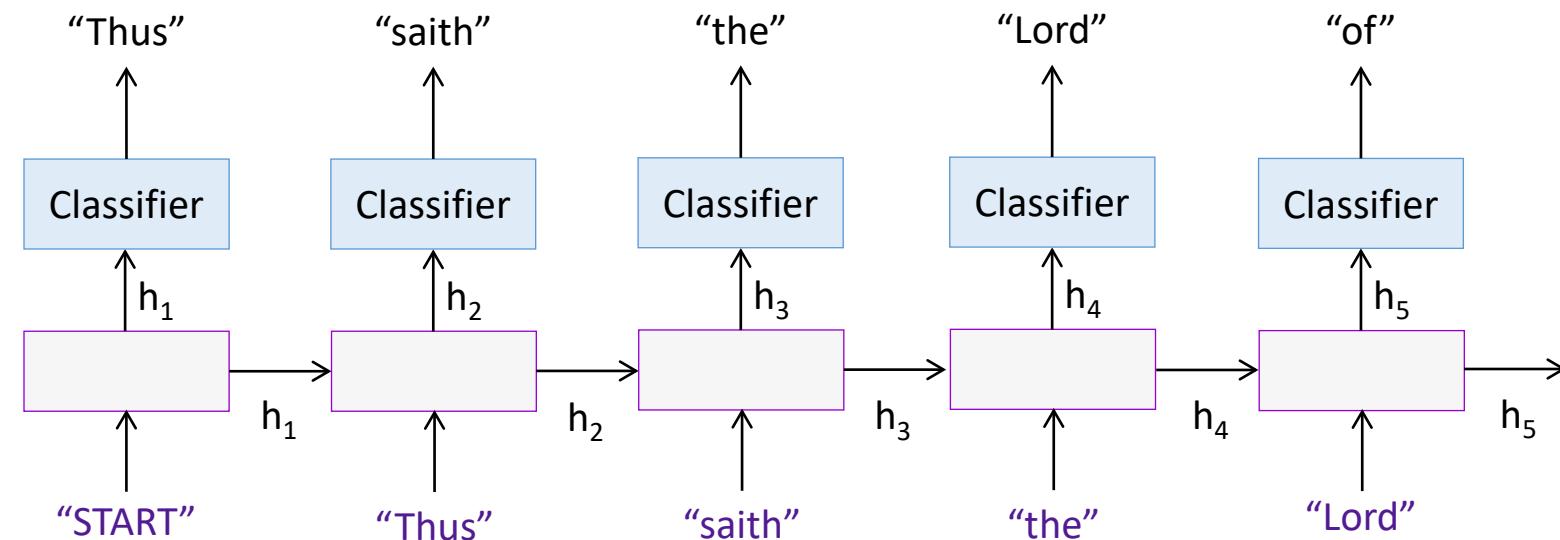
DeepDrumpf @DeepDrumpf · May 31, 2017
[Despite the negative press #covfefe] look at what's going on. They shoot media. Usually that's a bad sign of things to come.
6 38 124

DeepDrumpf @DeepDrumpf · Apr 7, 2017
When I have to build a hotel, we're bombing the hell out of them. Lots of money. To those suffering, I say vote for Donald. #SyriaStrikes
1 71 173

DeepDrumpf @DeepDrumpf · Mar 20, 2017
Replying to @Thomas1774Paine
There will be no amnesty. It is going to pass because the people are going to be gone. I'm giving a mandate. #ComeyHearing @Thomas1774Paine

Example 2: Text generation

- Sample from the distribution of a given text corpus (also known as language modeling)
- Can be done one character or one word at a time:



Example 3: Image caption generation



A cat sitting on a suitcase on the floor



A cat is sitting on a tree branch



A dog is running in the grass with a frisbee



A white teddy bear sitting in the grass



Two people walking on the beach with surfboards



A tennis player in action on the court



Two giraffes standing in a grassy field

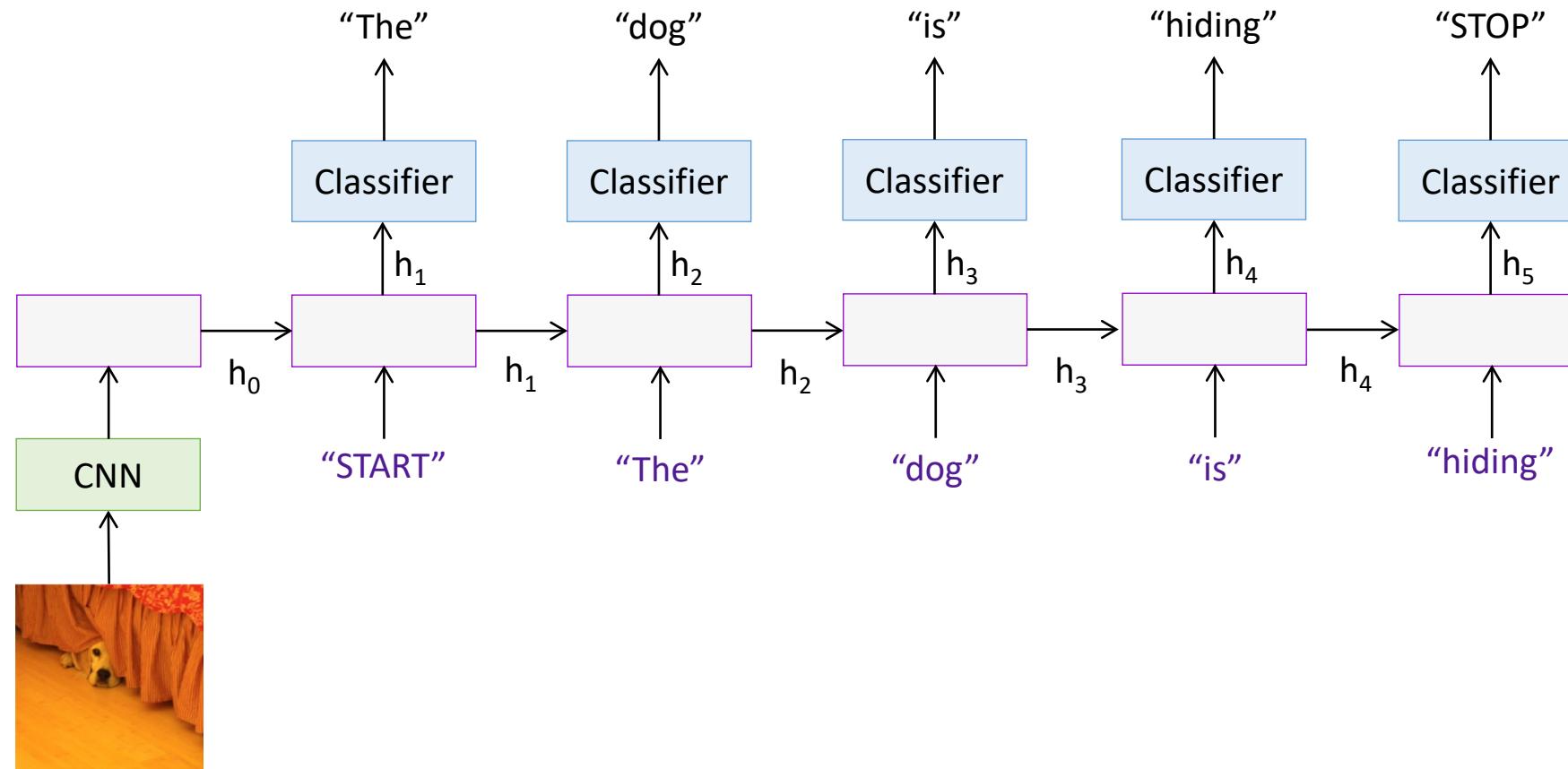


A man riding a dirt bike on a dirt track

Source: [J. Johnson](#)

Captions generated using [neuraltalk2](#)

Example 3: Image caption generation



Example 4: Machine translation

The screenshot shows the Google Translate interface. The source text is in French, and the target text is in English. The French text is a poem by Charles Baudelaire:

Correspondances
La Nature est un temple où de vivants piliers
Laisquent parfois sortir de confuses paroles;
L'homme y passe à travers des forêts de symboles
Qui l'observent avec des regards familiers.
Comme de longs échos qui de loin se confondent
Dans une ténèbreuse et profonde unité,
Vaste comme la nuit et comme la clarté,
Les parfums, les couleurs et les sons se répondent.
Il est des parfums frais comme des chairs d'enfants,
Doux comme les hautbois, verts comme les prairies,
— Et d'autres, corrompus, riches et triomphants,
Ayant l'expansion des choses infinies,
Comme l'ambre, le musc, le benzoin et l'encens,
Qui chantent les transports de l'esprit et des sens.
— Charles Baudelaire

The English translation is:

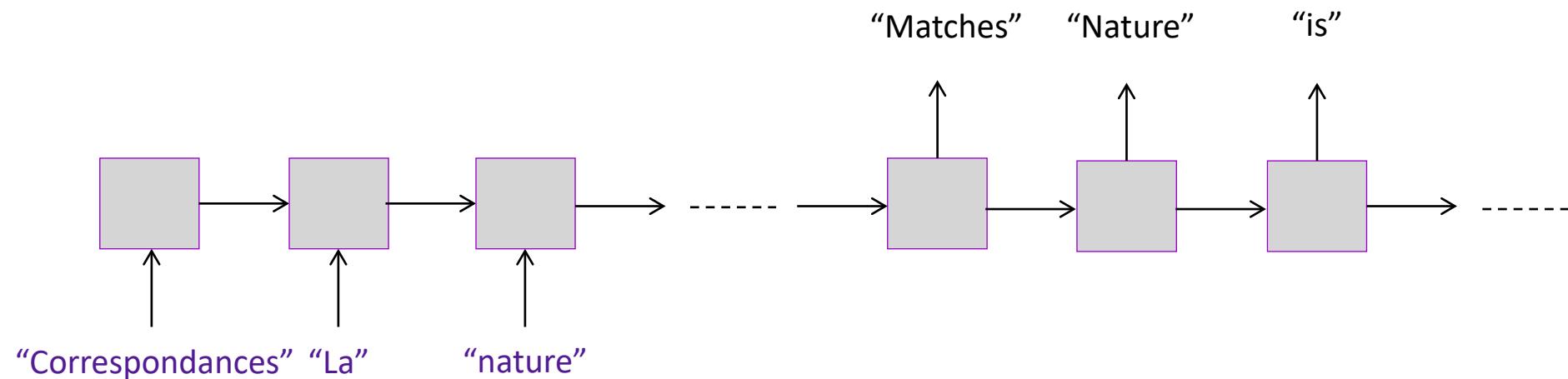
Matches
Nature is a temple where living pillars
Sometimes let out confused words;
Man goes through symbol forests
Which observe him with familiar eyes.
Like long echoes that by far merge
In a dark and deep unity,
As vast as the night and as clarity,
The perfumes, the colors and the sounds answer each other.
There are fresh perfumes like children's flesh,
Sweet like oboes, green like meadows,
- And others, corrupt, rich and triumphant,
Having the expansion of infinite things,
Like amber, musk, benzoin and incense,
Who sing the transports of the mind and the senses.
- Charles Baudelaire

At the bottom left, there are icons for microphone, keyboard, and a dropdown menu. At the bottom center, it says "693/5000". On the right side, there are icons for star, square, and a pencil.

<https://translate.google.com/>

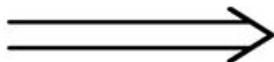
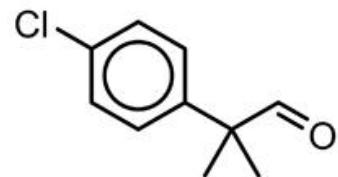
Example 4: Machine translation

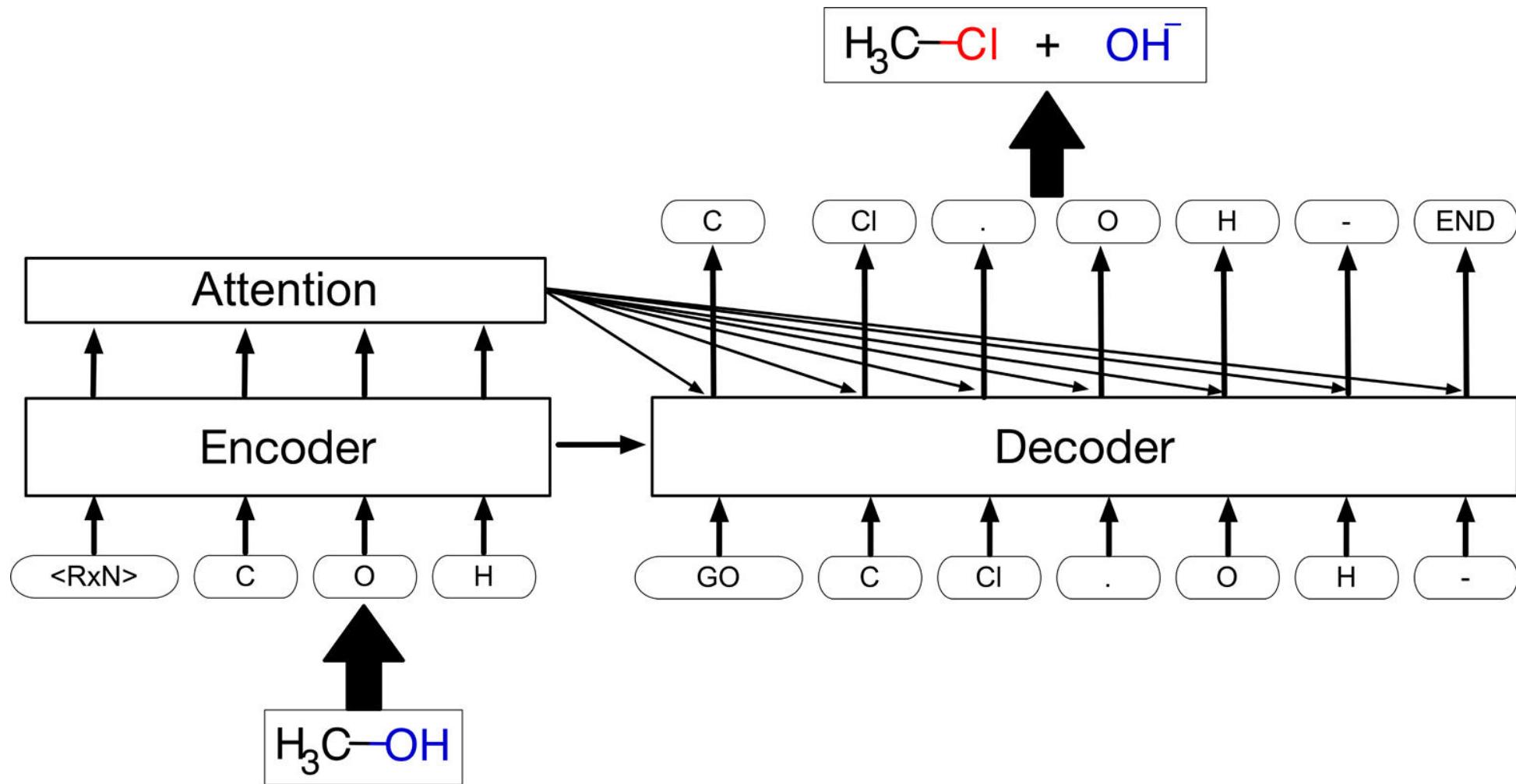
- Multiple input – multiple output (or sequence to sequence) scenario:



Synthesis prediction

(target)



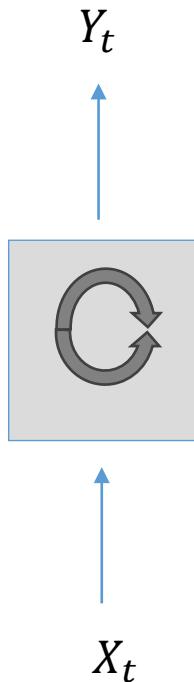


Bidirectional

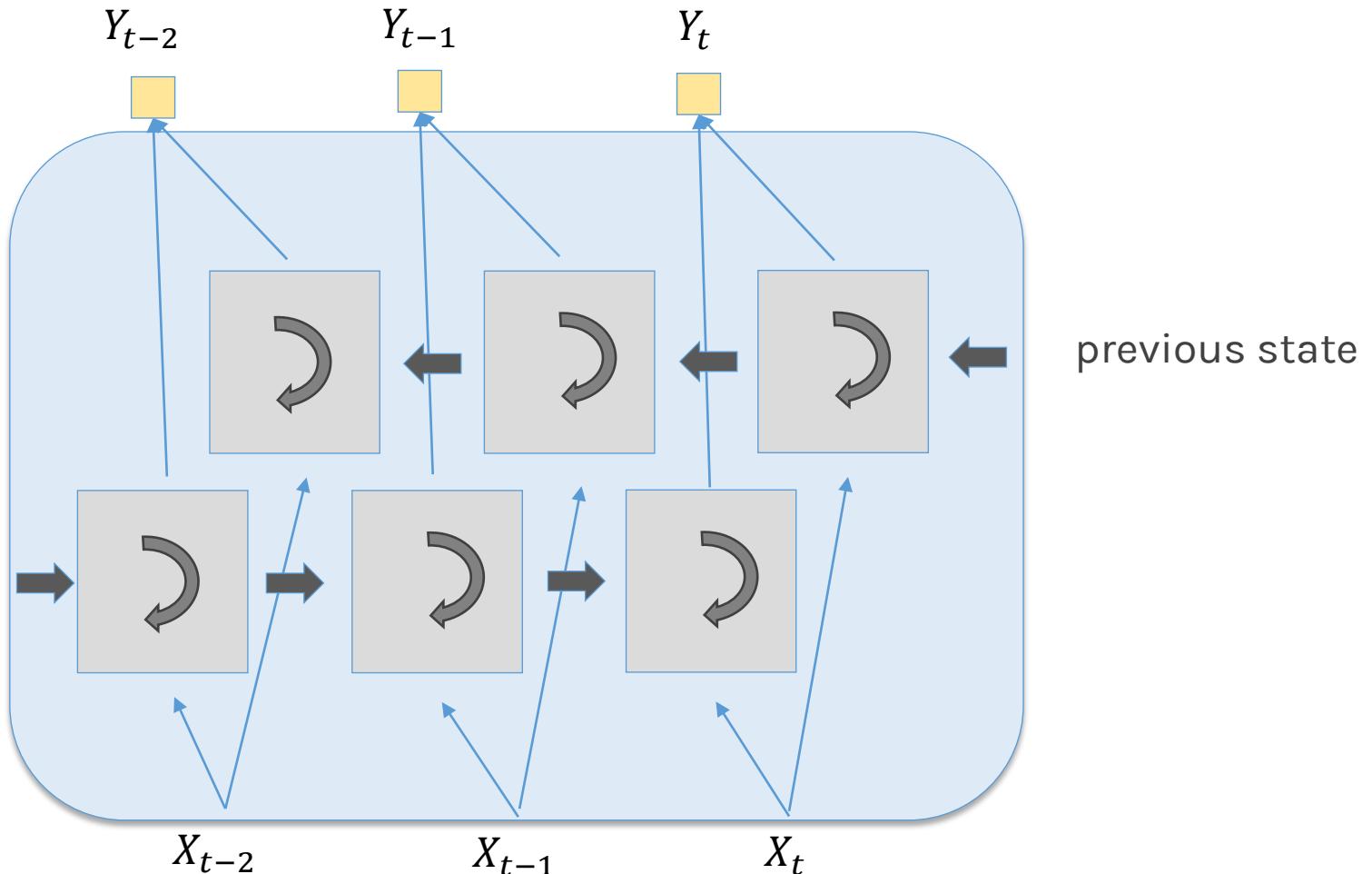
- LSTM and RNN are designed to analyze sequence of values.
- For example: *Patrick said he needs a vacation.*
- *he* here means *Patrick* and we know this because *Patrick* was before the word *he*.
- However consider the following sentence:
- *He needs to work more, Peter said about Patrick.*
- Bidirectional RNN or BRNN or bidirectional LSTM or BLSTM when using LSTM units.

Bidirectional (cond)

symbol for a BRNN



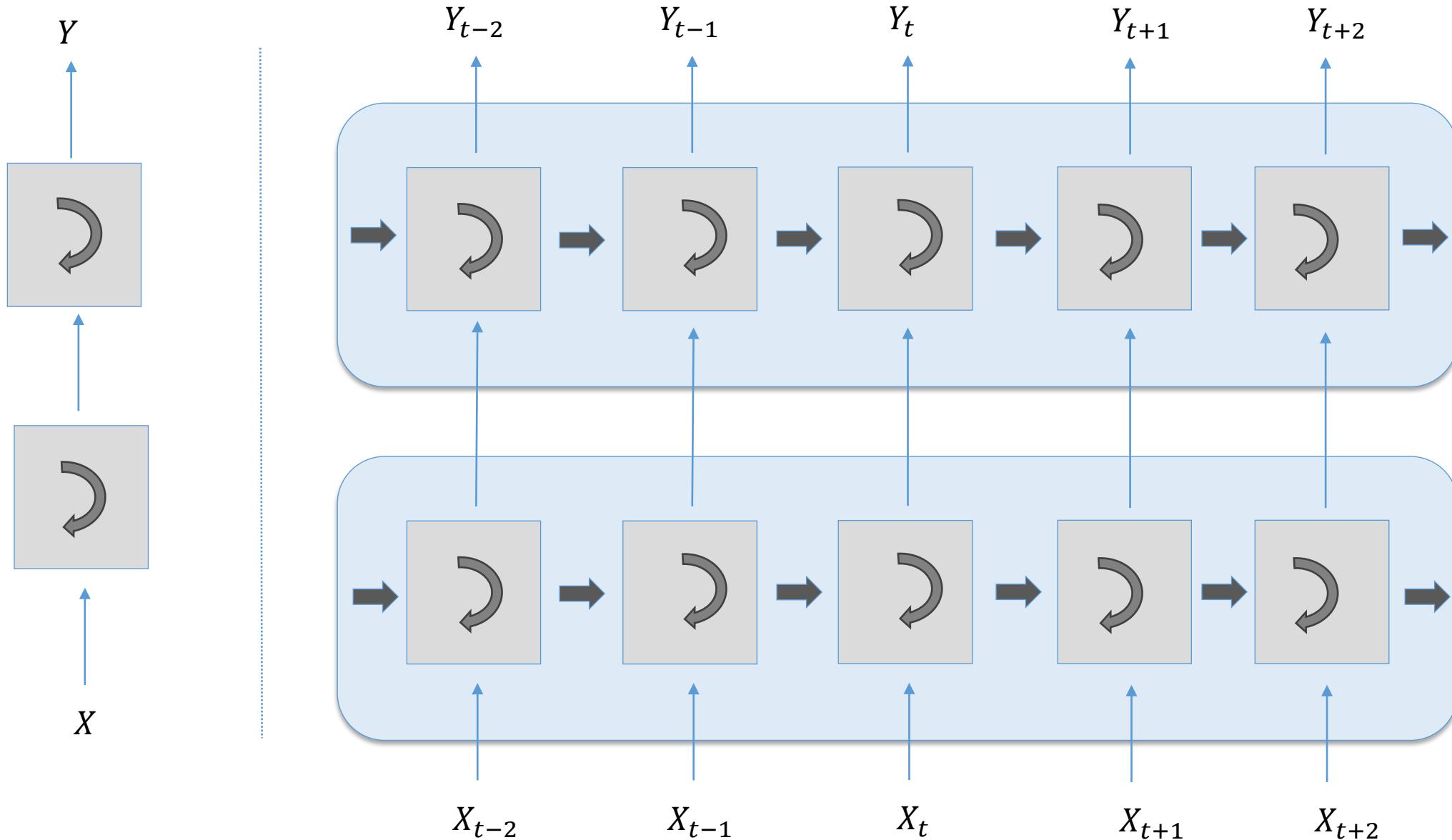
previous state



Deep RNN

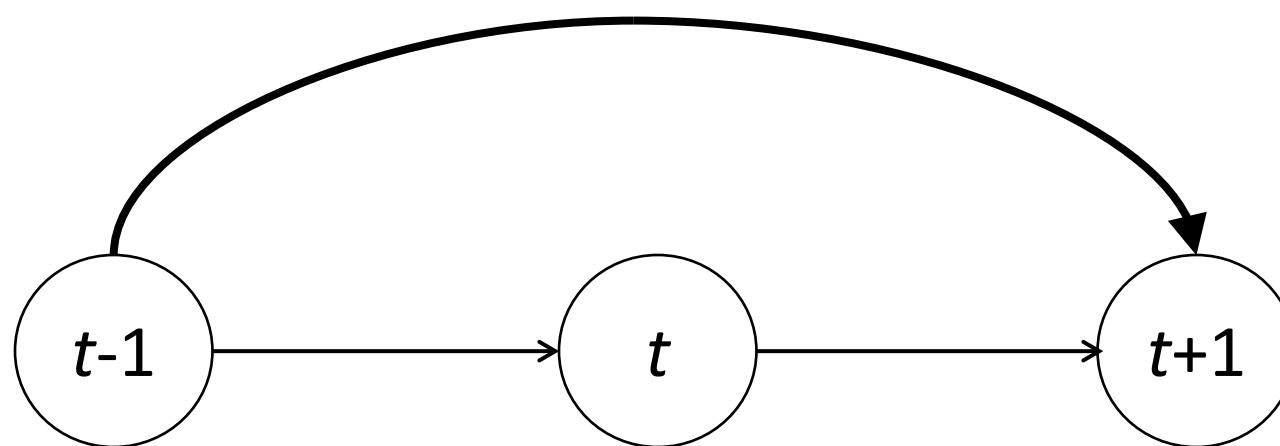
- LSTM units can be arranged in layers, so that each the output of each unit is the input to the other units. This is called a **deep RNN**, where the adjective “deep” refers to these multiple layers.
- Each layer feeds the LSTM on the next layer
- First time step of a feature is fed to the first LSTM, which processes that data and produces an output (and a new state for itself).
- That output is fed to the next LSTM, which does the same thing, and the next, and so on.
- Then the second time step arrives at the first LSTM, and the process repeats.

Deep RNN



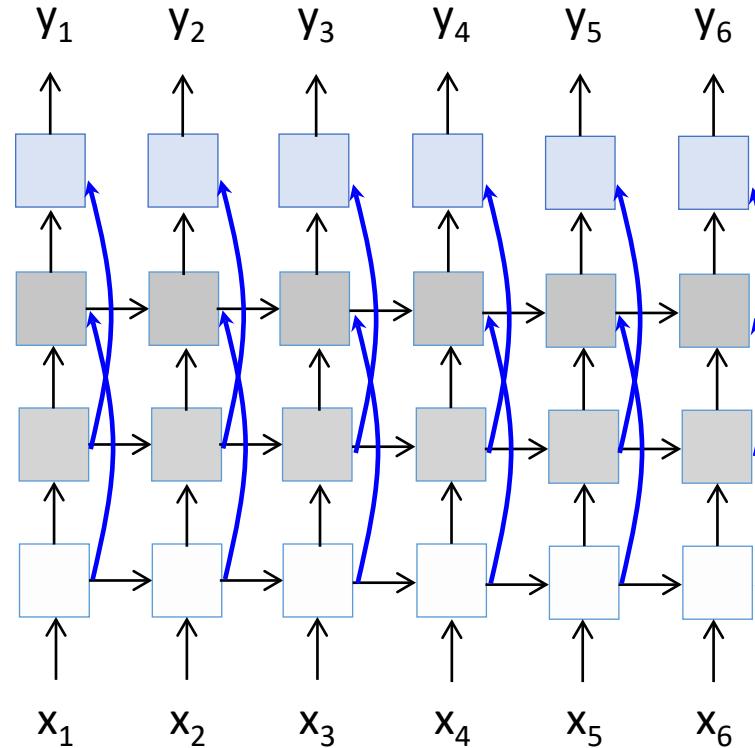
Skip Connections

- Add additional **connections between units d time steps apart**
- Creating paths through time where gradients neither vanish or explode



Multi-layer RNNs

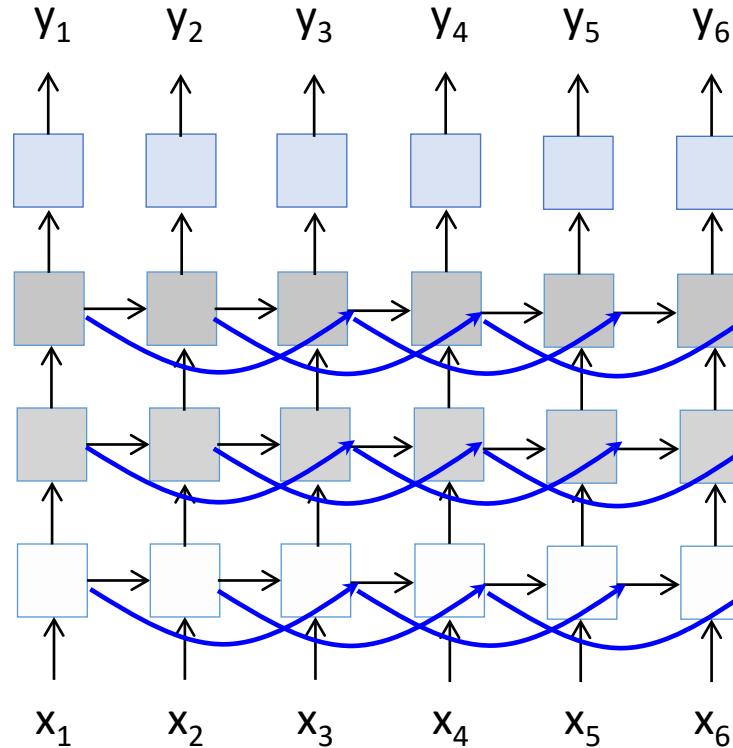
- We can of course design RNNs with multiple hidden layers



- Anything goes: skip connections across layers, across time, ...

Multi-layer RNNs

- We can of course design RNNs with multiple hidden layers



- Anything goes: skip connections across layers, across time, ...