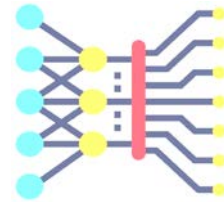


CS60010: Deep Learning

Spring 2023

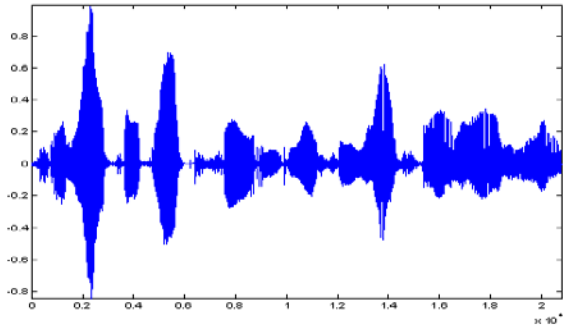
Sudeshna Sarkar

RNN

Sudeshna Sarkar

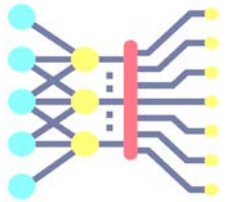
1 Mar 2023

RNN models sequences



- Sequences are everywhere
  - Sentences: Sequence of words
  - Speech: Acoustic features at successive time frames
  - Successive frames in video
  - Rainfall measurements on successive days in Kgp
  - Stock Market: Daily values of current exchange rate

Variable-size inputs



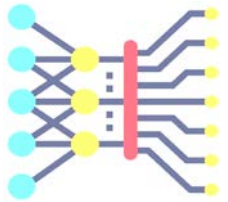
$$x_1 = (x_{1,1}, x_{1,2}, x_{1,3}, x_{1,4})$$

$$x_2 = (x_{2,1}, x_{2,2}, x_{2,3})$$

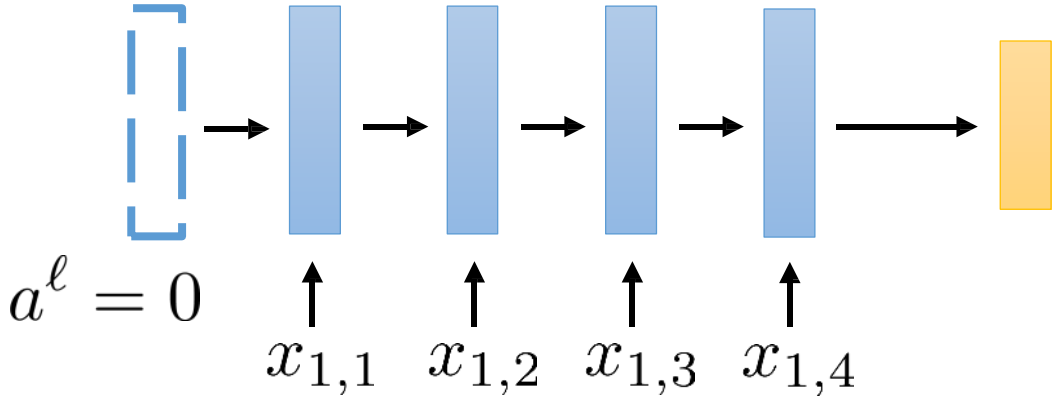
$$x_3 = (x_{3,1}, x_{3,2}, x_{3,3}, x_{3,4}, x_{3,5})$$

1. classifying sentiment for a phrase (sequence of words)
2. recognizing phoneme from sound (sequence of sounds)
3. classifying the activity in a video (sequence of images)

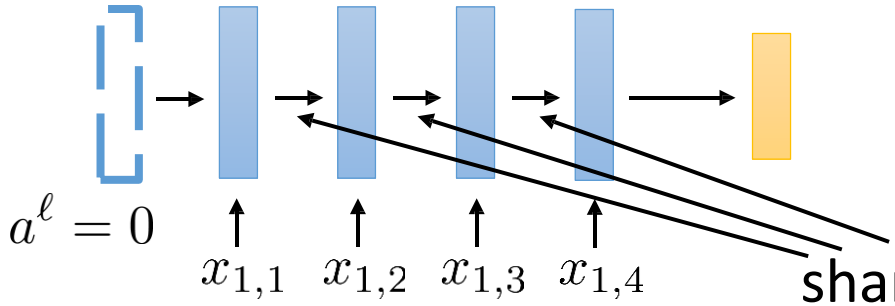
RNN



- RNNs share weights across multiple layers, take an input at each layer, and have a variable number of layers

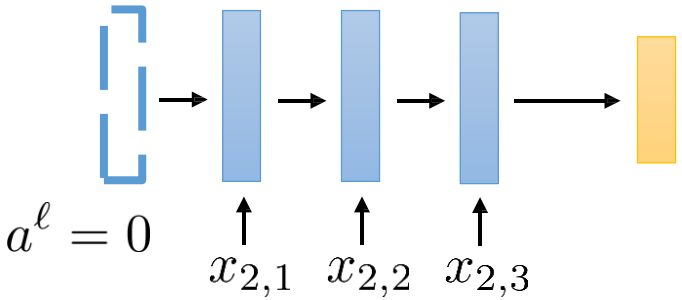


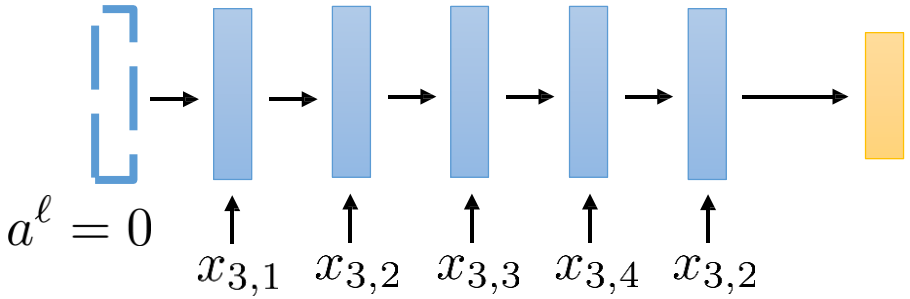
Share weight matrices



shared W_l

i.e., $W^{\ell_i} = W^{\ell_j}$ for all i, j





each layer:

$$\bar{a}^{\ell-1} = \begin{bmatrix} a^{\ell-1} \\ x_{i,t} \end{bmatrix}$$

$$z^\ell = W^\ell \bar{a}^{\ell-1} + b^\ell$$

$$a^\ell = \sigma(z^\ell)$$

This is called a **recurrent** neural network (RNN)

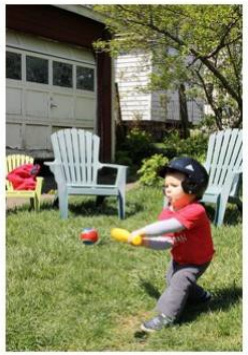
This is a “variable-depth” network

variable-size outputs

Examples:

- generating a text caption for an image
- predicting a sequence of future video frames
- generating an audio sequence

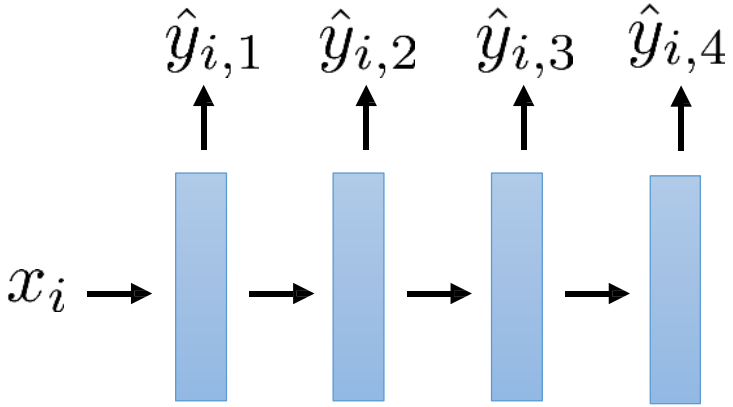
a group of people standing around a room with remotes
logprob: -9.17



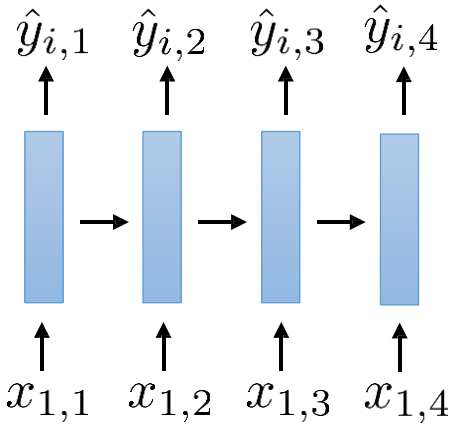
a young boy is holding a baseball bat
logprob: -7.61

a toilet with a seat up in a bathroom
logprob: -13.44

a woman holding a teddy bear in front of a mirror
logprob: -9.65



Inputs and outputs at each step



at each step:

$$\bar{a}^{\ell-1} = \begin{bmatrix} a^{\ell-1} \\ x_{i,t} \end{bmatrix}$$

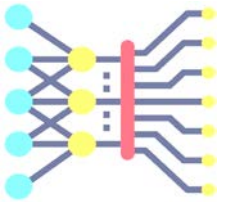
$$z^{\ell} = W^{\ell}\bar{a}^{\ell-1} + b^{\ell}$$

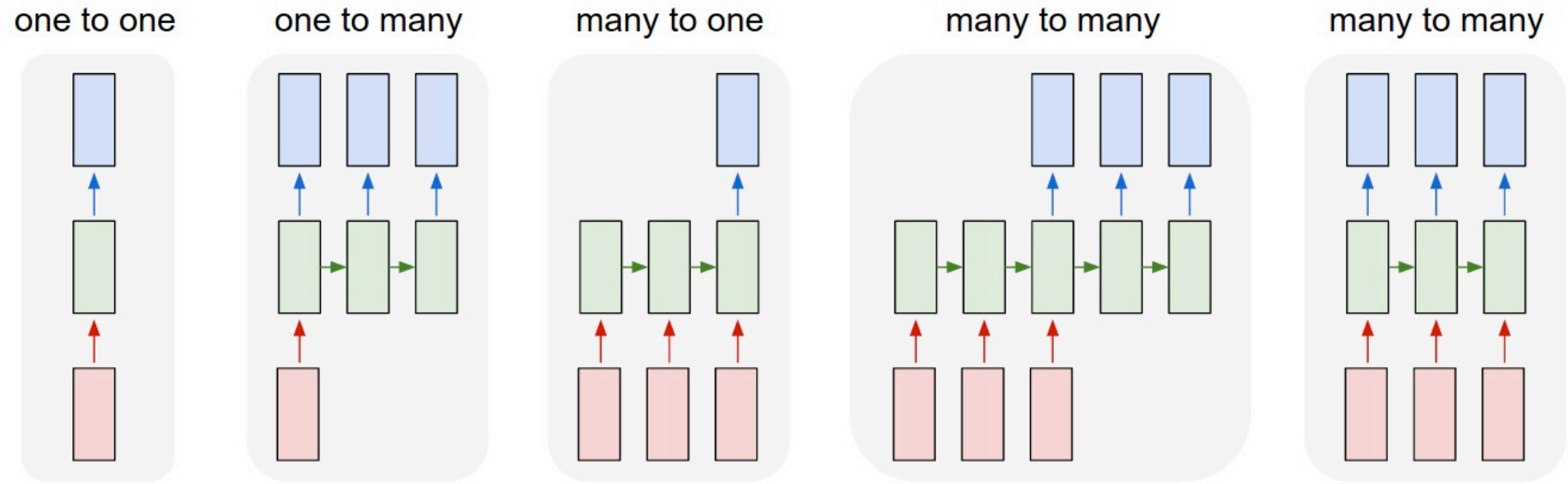
$$a^{\ell} = \sigma(z^{\ell})$$

$$\hat{y}_{\ell} = f(a^{\ell})$$

Machine Translation

Different ways to use RNNs





Input: No sequence
Output: No sequence
Example: “standard” classification / regression problems

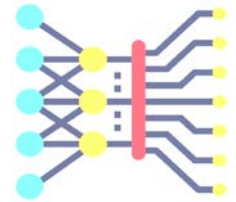
Input: No sequence
Output: Sequence
Example: Im2Caption

Example: activity recognition, sentence classification

Example: machine translation, video captioning, open-ended question answering

Example: frame-level video annotation

Sentiment Classification



- Classify a
  - restaurant review from Yelp! OR
  - movie review from IMDB OR
  - …

  as positive or negative

  - Inputs: Multiple words, one or more sentences
  - Outputs: Positive / Negative classification

1. “The food was really good”
2. “The chicken crossed the road because it was uncooked”

Sentiment Classification

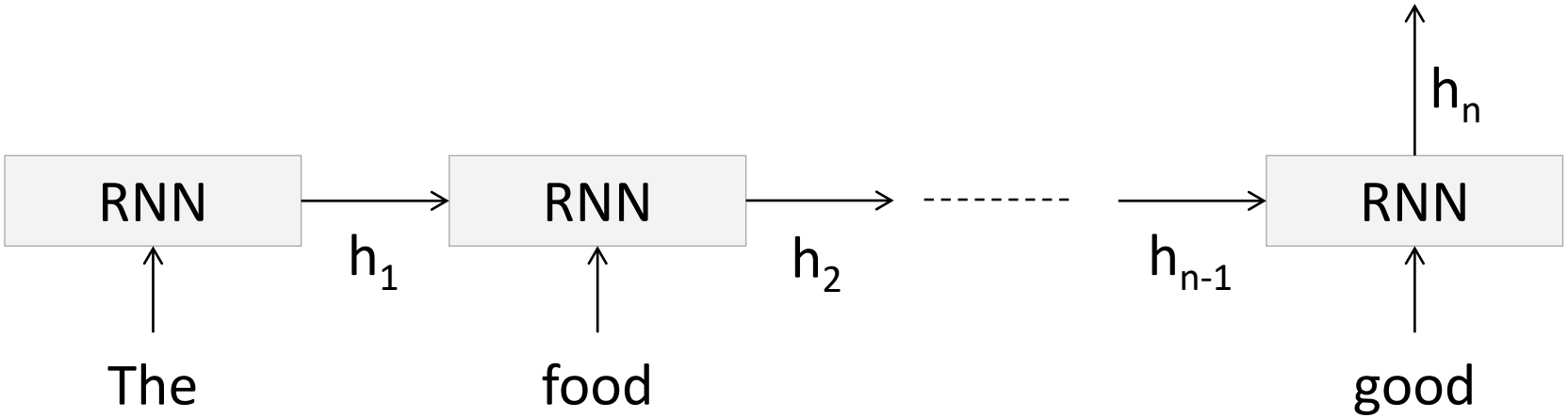
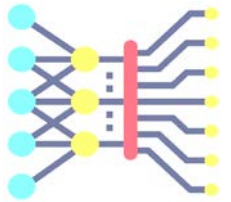


Image Captioning

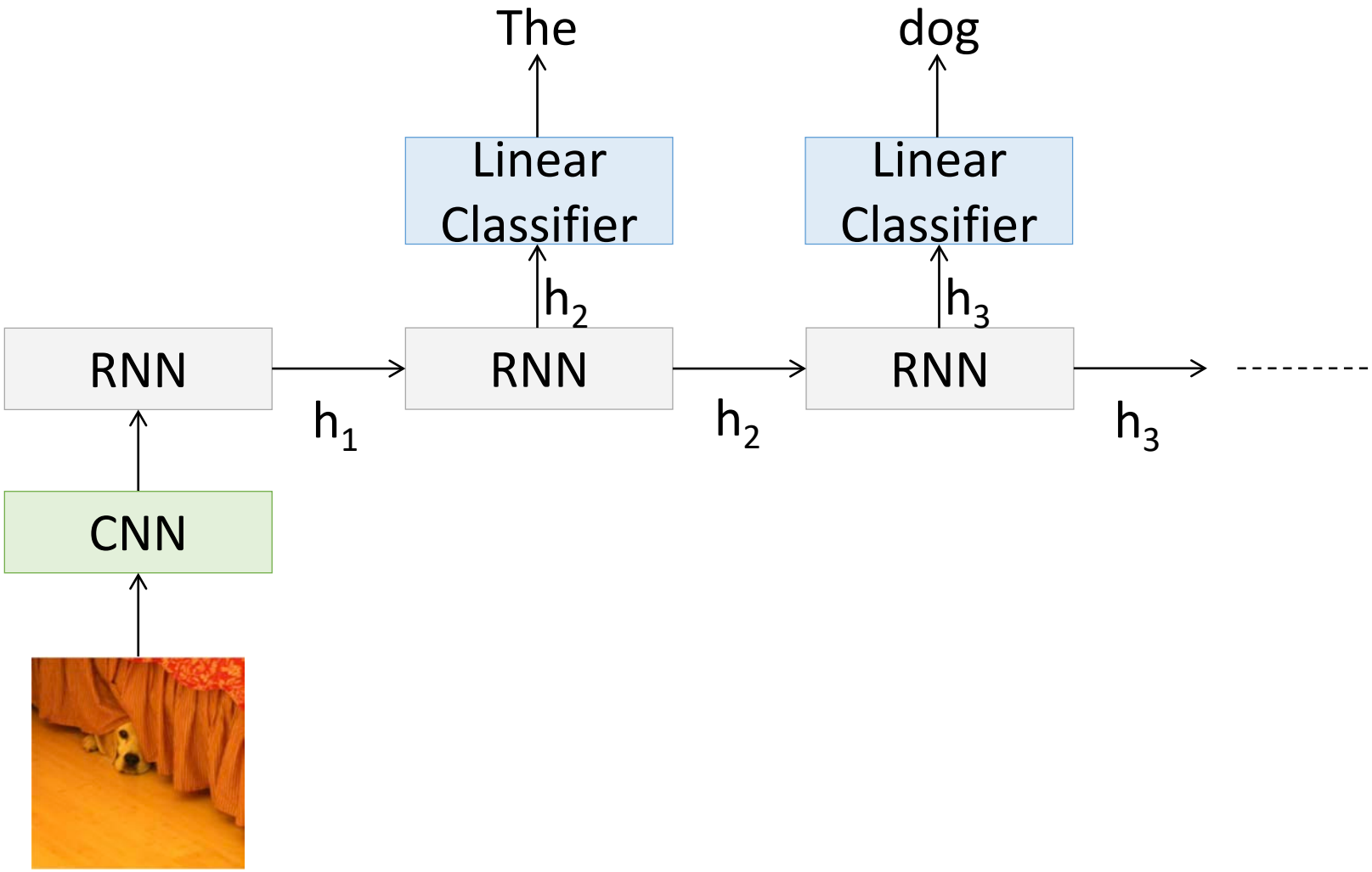


- Given an image, produce a sentence describing its contents

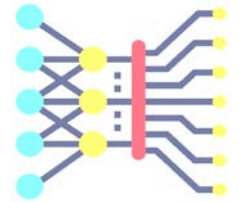
- Inputs: Image feature (from a CNN)
- Outputs: Multiple words (let's consider one sentence)

→ The dog is hiding

Image Captioning



RNN Outputs: Language Modeling



http://karpathy.github.io/2015/05/21/rnn-effectiveness/

VIOLA:

Why, Salisbury must find his flesh and thought

That which I am not aps, not a man and in fire,

To show the reining of the raven and the wars

To grace my hand reproach within, and not a fair are hand,

That Caesar and my goodly father's world;

When I was heaven of presence and our fleets,

We spare with hours, but cut thy council I am great,

Murdered and by thy master's ready there

My power to give thee but so much as hell:

Some service in the noble bondman here,

Would show him to her wine.

KING LEAR:

O, if you were a feeble sight, the courtesy of your law,

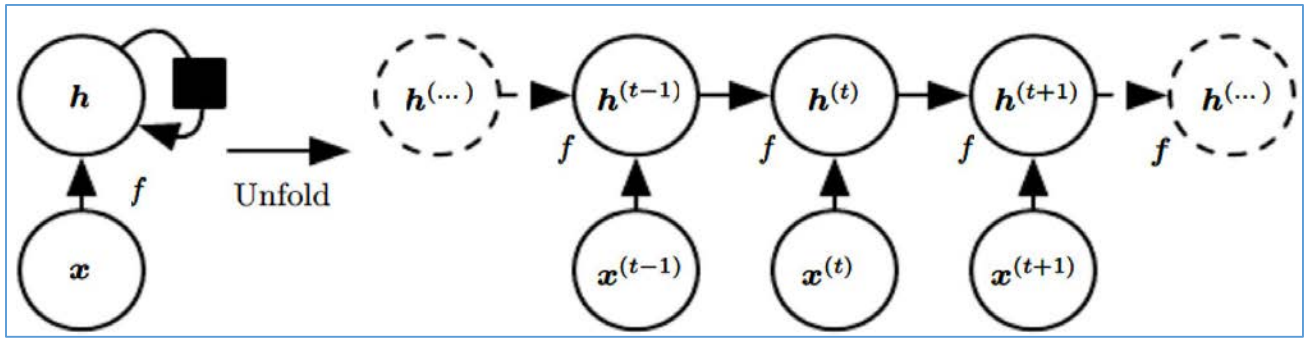
Your sight and several breath, will wear the gods

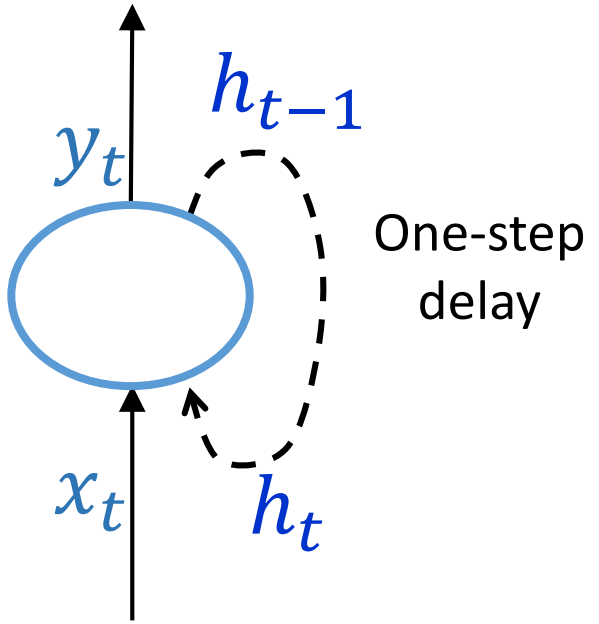
With his heads, and my hands are wonder'd at the deeds,

So drop upon your lordship's head, and your opinion

Shall be against your honour.

Output prediction by RNN





Task : To predict the future from the past

- The network learns to use $\boldsymbol{h}^{(t)}$ as a summary of the task-relevant aspects of the past sequence of inputs upto t
- The summary is in general lossy since it maps a sequence of arbitrary length to a fixed length vector $\boldsymbol{h}^{(t)}$

RNN structure

Often layers are stacked vertically (deep RNNs):