CS2109S Tutorial 8

Recurrent Neural Networks

(AY 25/26 Semester 2)

April 2, 2026

(Prepared by Benson)

Contents

Recurrent Neural Networks

Q1. Model Selection for Sequential Data
Q2. RNN Tracing
Q3. Bidirectional RNNs
Q4. 1D Convolution vs. RNNs

Why CNNs/RNNs?

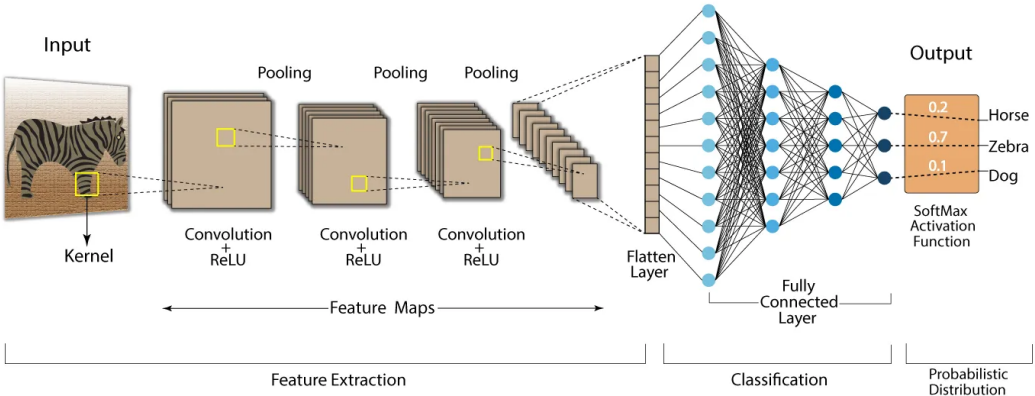
Should we simply whack a *very very very* deep Neural Network?

Problem context is important.

Idea: Exploit characteristics in the problem domain.

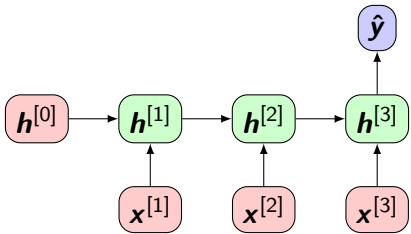
Why CNNs/RNNs?

Convolutional Neural Networks: Exploiting **locality**.



Why CNNs/RNNs?

Recurrent Neural Networks: Exploiting **sequential nature**.



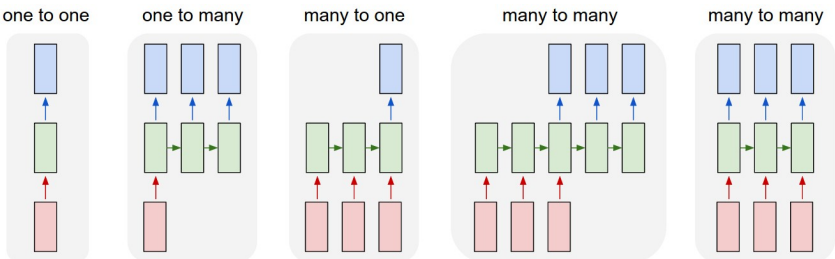
Q1. Model Selection for Sequential Data

What type of model should we use?

(a) Image Captioning — One-to-many model

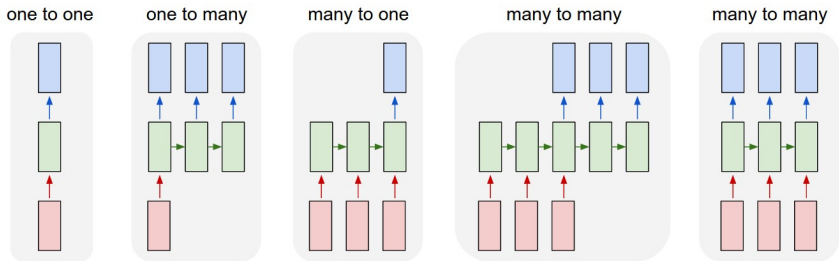
(b) Stock market prediction — Many-to-one model

(c) Language Translation — Many-to-many model



Q1. Model Selection for Sequential Data

Exercise: For music generation with a prompt (e.g. the style), what type of RNN model should we use?



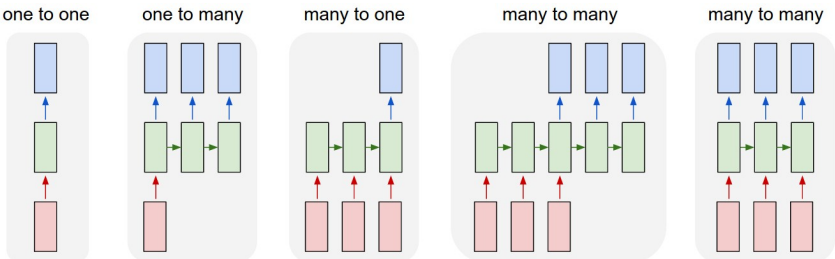
Q2. RNN Tracing

(a) We would like to use simple RNN model to predict the last word in the following incomplete sentence made of 2 words:

"I love..."

What type of RNN model is needed?

- Many-to-one.

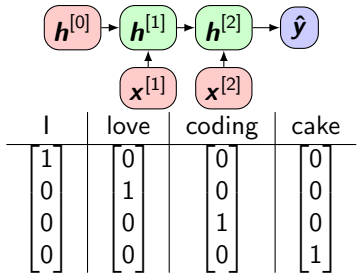


Q2. RNN Tracing

(b) Compute the hidden states $\boldsymbol{h}^{[1]}, \boldsymbol{h}^{[2]}$.

$$\boldsymbol{h}^{[1]} = (\boldsymbol{W}^{[xh]})^\top \boldsymbol{x}^{[1]} + (\boldsymbol{W}^{[hh]})^\top \boldsymbol{h}^{[0]}$$
$$= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\boldsymbol{h}^{[2]} = (\boldsymbol{W}^{[xh]})^\top \boldsymbol{x}^{[2]} + (\boldsymbol{W}^{[hh]})^\top \boldsymbol{h}^{[1]}$$
$$= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$



$$\boldsymbol{W}^{[xh]} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{W}^{[hh]} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\boldsymbol{W}^{[hy]} = \begin{bmatrix} 1 & 0 & 2 & -1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \boldsymbol{h}^{[0]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Q2. RNN Tracing

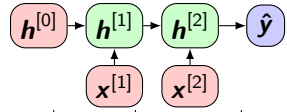
(c) Compute the most probable next word.

$$\hat{\boldsymbol{y}} = \text{softmax}\left((\boldsymbol{W}^{[hy]})^\top \boldsymbol{h}^{[2]}\right)$$

$$= \text{softmax}\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 1 \\ -1 & 0 \end{bmatrix}^\top \begin{bmatrix} 1 \\ 2 \end{bmatrix}\right)$$

$$= \text{softmax}\left(\begin{bmatrix} 1 \\ 2 \\ 4 \\ -1 \end{bmatrix}\right) = \begin{bmatrix} 0.04 \\ 0.11 \\ 0.84 \\ 0.01 \end{bmatrix}$$

$\therefore$ The most probable next word is "coding".

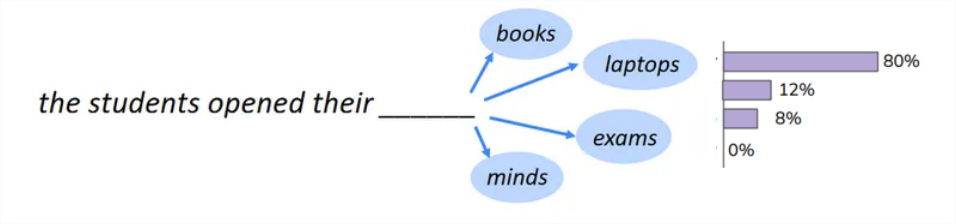


$$\boldsymbol{W}^{[xh]} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{W}^{[hh]} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\boldsymbol{W}^{[hy]} = \begin{bmatrix} 1 & 0 & 2 & -1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \boldsymbol{h}^{[0]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Extra: Greedy Search vs Sampling

Extra Slide



Think: LLMs like ChatGPT allow you to specify a **temperature**. How does temperature affect the sampling?

Q2. RNN Tracing

(d) What happens if we input "love I" instead?

- The order of computing the hidden states will change.
- The predictions will be different.

(e) What is the potential issue of using one-hot encoding to represent the input vector?

- High dimensional.
- Two semantically similar words (e.g., "trend" and "pattern") have zero similarity in one-hot space.

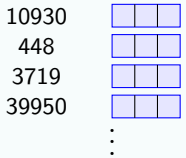
Extra: Token Embeddings

Demo: GPT4 Tokenizer

Write an email apolog izing to Sarah for the tragic gardening mish ap . Explain how it happened .

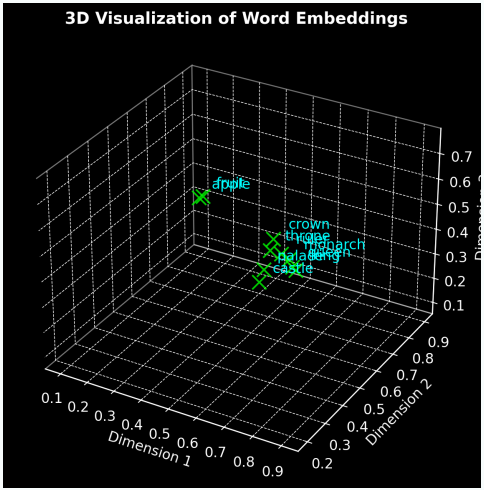
```
[10930, 448, 3719,
39950, 6396, 316, 32145,
395, 290, 62374, 66241,
80785, 403, 13, 115474,
1495, 480, 12570, 13]
```

Token Embeddings

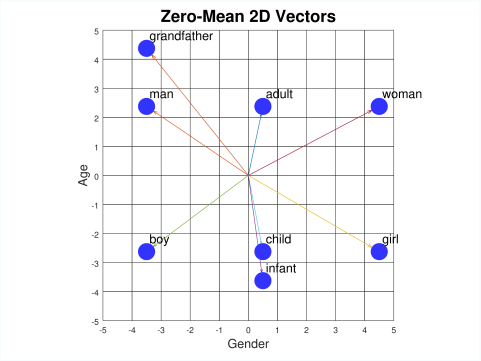


Extra: Token Embeddings

Think: What are some ideal properties of token embeddings?



Capture semantic similarities



Meaningful linear substructures

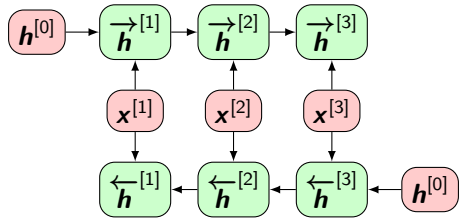
Q3. Bidirectional RNNs

(a) Compute the hidden states $\overrightarrow{\boldsymbol{h}}^{[1]}$, $\overrightarrow{\boldsymbol{h}}^{[2]}$, $\overrightarrow{\boldsymbol{h}}^{[3]}$ and $\overleftarrow{\boldsymbol{h}}^{[1]}$, $\overleftarrow{\boldsymbol{h}}^{[2]}$, $\overleftarrow{\boldsymbol{h}}^{[3]}$.

$$\begin{aligned}\overrightarrow{\boldsymbol{h}}^{[1]} &= (\overrightarrow{\boldsymbol{W}}^{[xh]})^\top \boldsymbol{x}^{[1]} + (\overrightarrow{\boldsymbol{W}}^{[hh]})^\top \boldsymbol{h}^{[0]} \\ &= \begin{bmatrix}1 & 2 & 3\end{bmatrix}\begin{bmatrix}1 & 0 & 0\end{bmatrix}^\top + \begin{bmatrix}1\end{bmatrix}\begin{bmatrix}0\end{bmatrix} = \begin{bmatrix}1\end{bmatrix}\end{aligned}$$

$$\begin{aligned}\overrightarrow{\boldsymbol{h}}^{[2]} &= (\overrightarrow{\boldsymbol{W}}^{[xh]})^\top \boldsymbol{x}^{[2]} + (\overrightarrow{\boldsymbol{W}}^{[hh]})^\top \overrightarrow{\boldsymbol{h}}^{[1]} \\ &= \begin{bmatrix}1 & 2 & 3\end{bmatrix}\begin{bmatrix}0 & 1 & 0\end{bmatrix}^\top + \begin{bmatrix}1\end{bmatrix}\begin{bmatrix}1\end{bmatrix} = \begin{bmatrix}3\end{bmatrix}\end{aligned}$$

$$\begin{aligned}\overrightarrow{\boldsymbol{h}}^{[3]} &= (\overrightarrow{\boldsymbol{W}}^{[xh]})^\top \boldsymbol{x}^{[3]} + (\overrightarrow{\boldsymbol{W}}^{[hh]})^\top \overrightarrow{\boldsymbol{h}}^{[2]} \\ &= \begin{bmatrix}1 & 2 & 3\end{bmatrix}\begin{bmatrix}0 & 0 & 1\end{bmatrix}^\top + \begin{bmatrix}1\end{bmatrix}\begin{bmatrix}3\end{bmatrix} = \begin{bmatrix}6\end{bmatrix}\end{aligned}$$



not	very	good
$\begin{bmatrix}1\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\1\end{bmatrix}$

$$\overrightarrow{\boldsymbol{W}}^{[xh]} = \begin{bmatrix}1\\2\\3\end{bmatrix}, \overleftarrow{\boldsymbol{W}}^{[xh]} = \begin{bmatrix}0\\2\\3\end{bmatrix}$$

$$\overrightarrow{\boldsymbol{W}}^{[hh]} = \begin{bmatrix}1\end{bmatrix}, \overleftarrow{\boldsymbol{W}}^{[hh]} = \begin{bmatrix}1\end{bmatrix}, \boldsymbol{h}^{[0]} = \begin{bmatrix}0\end{bmatrix}$$

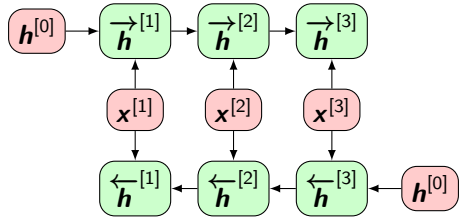
Q3. Bidirectional RNNs

(a) Compute the hidden states $\overrightarrow{\boldsymbol{h}}^{[1]}, \overrightarrow{\boldsymbol{h}}^{[2]}, \overrightarrow{\boldsymbol{h}}^{[3]}$ and $\overleftarrow{\boldsymbol{h}}^{[1]}, \overleftarrow{\boldsymbol{h}}^{[2]}, \overleftarrow{\boldsymbol{h}}^{[3]}$.

$$\begin{aligned}\overleftarrow{\boldsymbol{h}}^{[3]} &= (\overleftarrow{\boldsymbol{W}}^{[xh]})^\top \boldsymbol{x}^{[3]} + (\overleftarrow{\boldsymbol{W}}^{[hh]})^\top \boldsymbol{h}^{[0]} \\ &= \begin{bmatrix}0 & 2 & 3\end{bmatrix}\begin{bmatrix}0 & 0 & 1\end{bmatrix}^\top + \begin{bmatrix}1\end{bmatrix}\begin{bmatrix}0\end{bmatrix} = \begin{bmatrix}3\end{bmatrix}\end{aligned}$$

$$\begin{aligned}\overleftarrow{\boldsymbol{h}}^{[2]} &= (\overleftarrow{\boldsymbol{W}}^{[xh]})^\top \boldsymbol{x}^{[2]} + (\overleftarrow{\boldsymbol{W}}^{[hh]})^\top \overleftarrow{\boldsymbol{h}}^{[3]} \\ &= \begin{bmatrix}0 & 2 & 3\end{bmatrix}\begin{bmatrix}0 & 1 & 0\end{bmatrix}^\top + \begin{bmatrix}1\end{bmatrix}\begin{bmatrix}3\end{bmatrix} = \begin{bmatrix}5\end{bmatrix}\end{aligned}$$

$$\begin{aligned}\overleftarrow{\boldsymbol{h}}^{[1]} &= (\overleftarrow{\boldsymbol{W}}^{[xh]})^\top \boldsymbol{x}^{[1]} + (\overleftarrow{\boldsymbol{W}}^{[hh]})^\top \overleftarrow{\boldsymbol{h}}^{[2]} \\ &= \begin{bmatrix}0 & 2 & 3\end{bmatrix}\begin{bmatrix}1 & 0 & 0\end{bmatrix}^\top + \begin{bmatrix}1\end{bmatrix}\begin{bmatrix}5\end{bmatrix} = \begin{bmatrix}5\end{bmatrix}\end{aligned}$$



not	very	good
$\begin{bmatrix}1\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\1\end{bmatrix}$

$$\overrightarrow{\boldsymbol{W}}^{[xh]} = \begin{bmatrix}1\\2\\3\end{bmatrix}, \overleftarrow{\boldsymbol{W}}^{[xh]} = \begin{bmatrix}0\\2\\3\end{bmatrix}$$

$$\overrightarrow{\boldsymbol{W}}^{[hh]} = \begin{bmatrix}1\end{bmatrix}, \overleftarrow{\boldsymbol{W}}^{[hh]} = \begin{bmatrix}1\end{bmatrix}, \boldsymbol{h}^{[0]} = \begin{bmatrix}0\end{bmatrix}$$

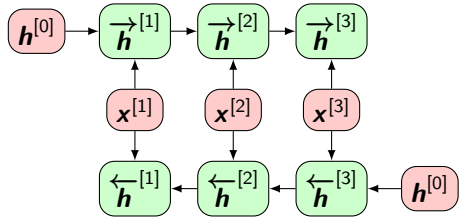
Q3. Bidirectional RNNs

(b) Compare the representation of the word "very" in the standard RNN and the bidirectional RNN.

- Standard RNN: $\boldsymbol{h}^{[2]} = \begin{bmatrix}3\end{bmatrix}$.
- Bidirectional RNN:
  $$\boldsymbol{h}_{bi}^{[2]} = \begin{bmatrix}\overrightarrow{\boldsymbol{h}}^{[2]} \\ \overleftarrow{\boldsymbol{h}}^{[2]}\end{bmatrix} = \begin{bmatrix}3 \\ 5\end{bmatrix}.$$
  Uses both **left and right** contexts.

(c) Advantages and limitations of bidirectional RNN?

- Advantage: Right context would be helpful for some tasks, e.g. sentiment analysis, part-of-speech (POS) tagging.
- Disadvantage: Increases computational cost.



$$\overrightarrow{\boldsymbol{W}}^{[xh]} = \begin{bmatrix}1\\2\\3\end{bmatrix}, \overleftarrow{\boldsymbol{W}}^{[xh]} = \begin{bmatrix}0\\2\\3\end{bmatrix}$$

$$\overrightarrow{\boldsymbol{W}}^{[hh]} = \begin{bmatrix}1\end{bmatrix}, \overleftarrow{\boldsymbol{W}}^{[hh]} = \begin{bmatrix}1\end{bmatrix}, \boldsymbol{h}^{[0]} = \begin{bmatrix}0\end{bmatrix}$$

Q4. 1D Convolution vs. RNNs

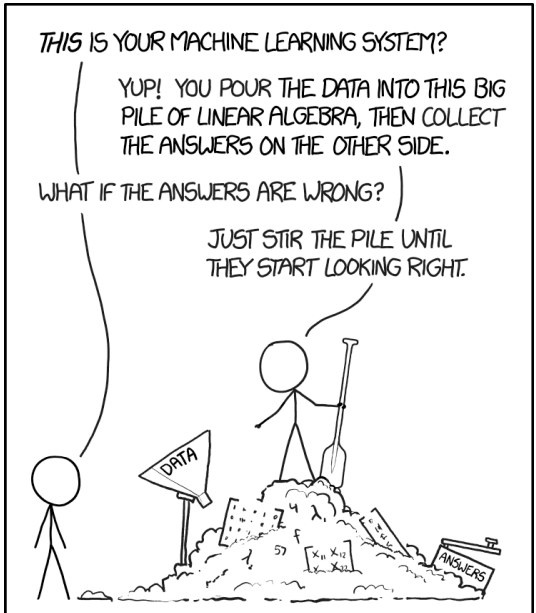
Consider a sequence of sensor readings from a patient's heart beat data:
$\boldsymbol{x} = \begin{bmatrix} 2 & 4 & 1 & -1 & 0 \end{bmatrix}$.

(a) Calculate the output after going through a single kernel $\boldsymbol{w} = \begin{bmatrix} 1 & -1 \end{bmatrix}$.
- $\begin{bmatrix} -2 & 3 & 2 & -1 \end{bmatrix}$.

(b) What is the purpose of the filter?
- Detect sudden drops in the sensor reading.

(c) Compare this with a RNN, in terms of parallelization and long-term context.
- RNNs have sequential dependencies while CNNs can be more effectively parallelized.
- CNNs with $k = 2$ structurally fails to retrieve long-term contexts, while RNNs can theoretically succeed.

That's it!


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS