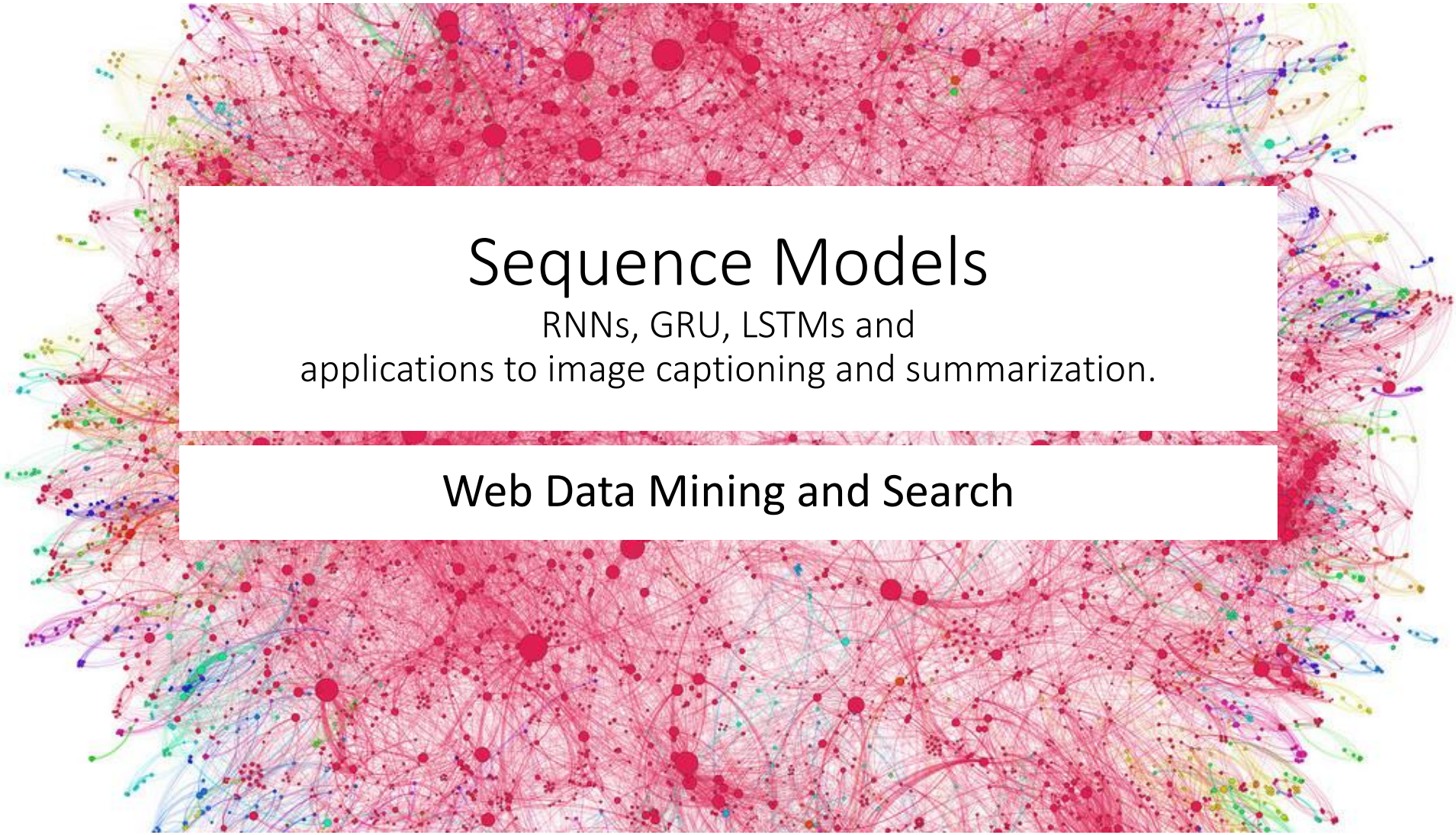


Sequence Models

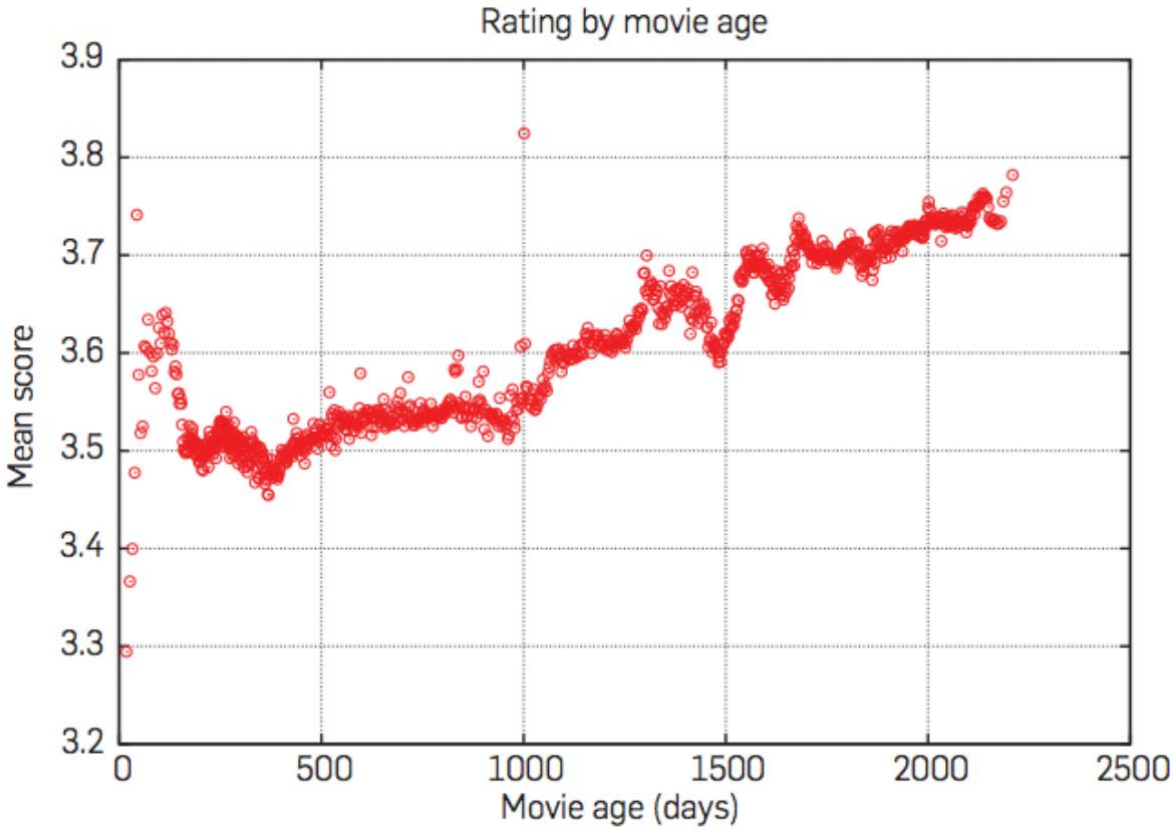
RNNs, GRU, LSTMs and
applications to image captioning and summarization.

Web Data Mining and Search

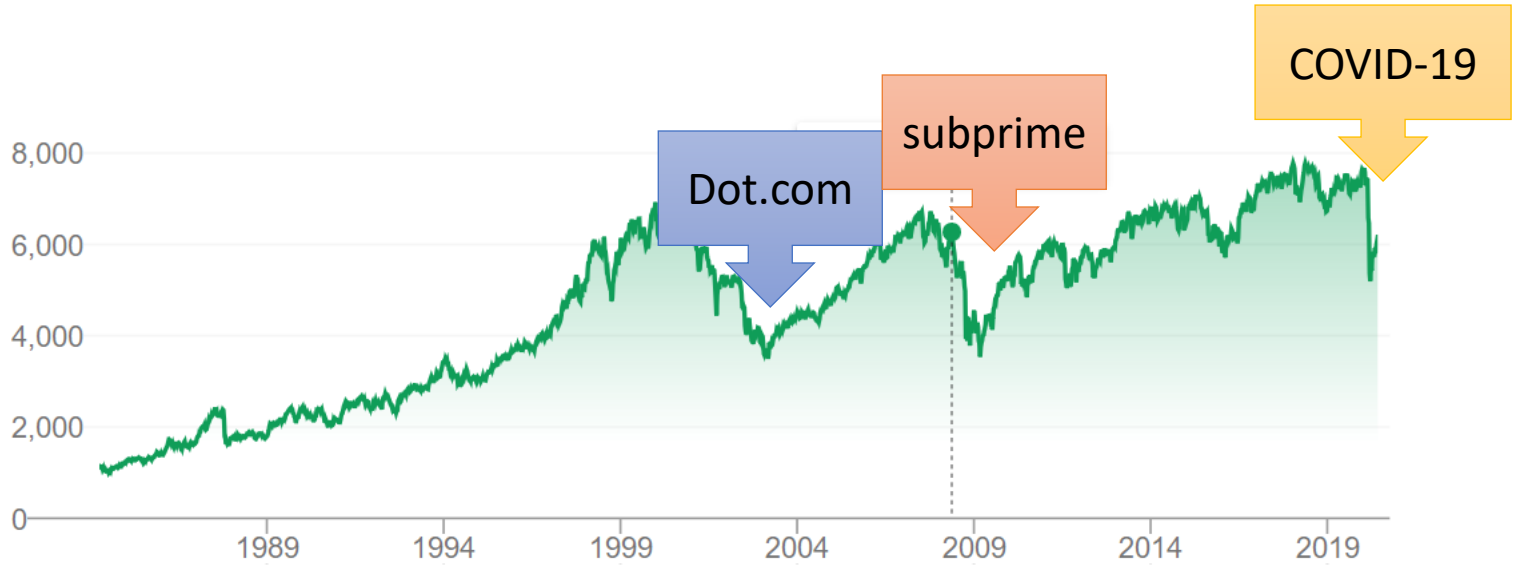
From static data to sequence data

- There are many domains where data samples have a dynamic or unknown size.
- CNNs and traditional multi-layer networks cope well with fixed-size input and output.
- However, there are many domains where
  - Input data is a sequence
  - Output is a sequence

Sequence problems: trend analysis

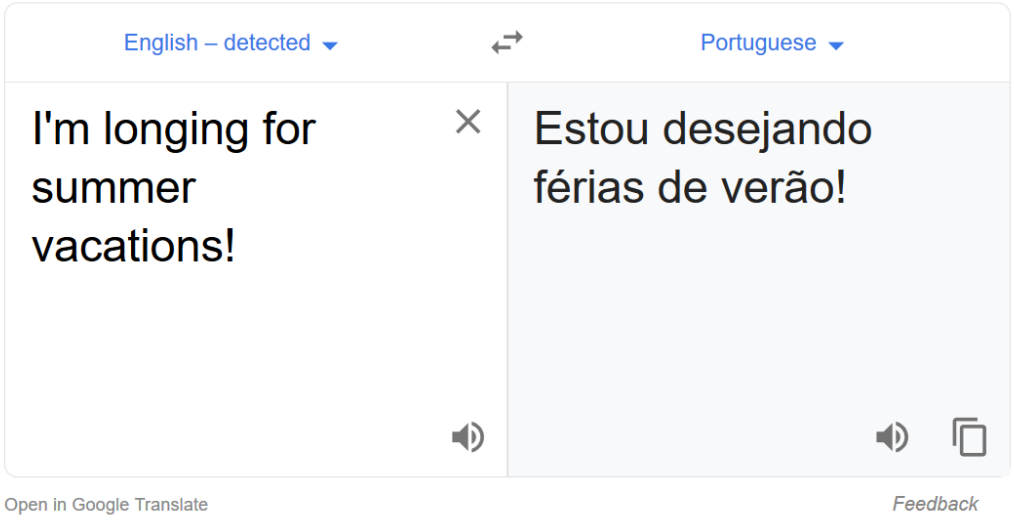


Sequence problems: stock markets



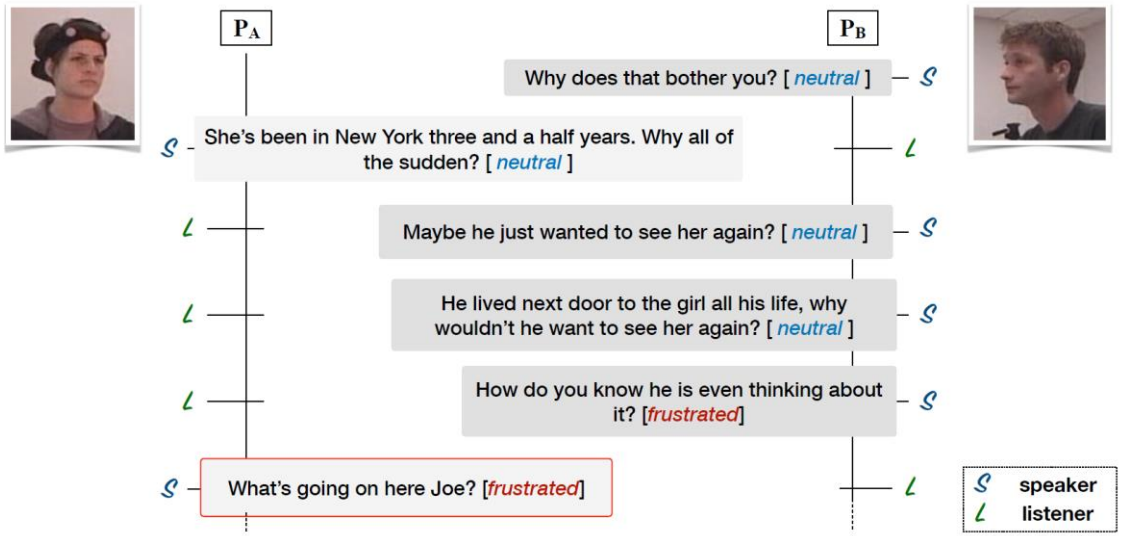
Sequence problems: Machine translation

- Translating a sentence from one language to another language.



Sequence problems: Sentiment analysis

- Detection of excitement, depression, frustration, etc.



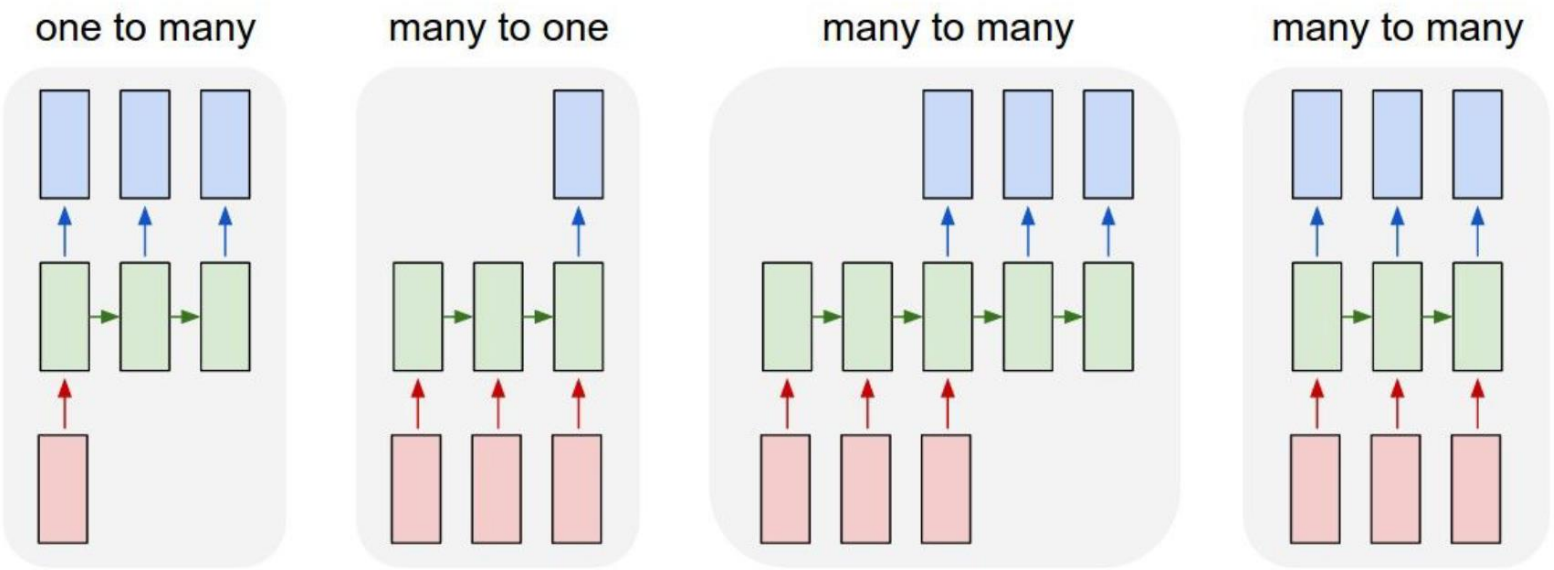
Majumder, Navonil, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. "Dialoguernn: An attentive rnn for emotion detection in conversations." *AAAI* 2019.

Tang, Duyu, Bing Qin, and Ting Liu. "Aspect level sentiment classification with deep memory network." *arXiv* 2016.

Usually, data is not

Independent and Identically Distributed (IID).

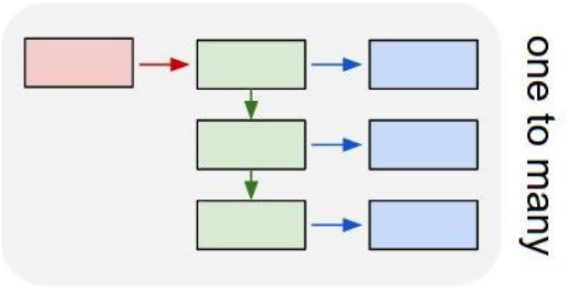
There are several possible architectures



Web data sequence modeling tasks

Input **Output**

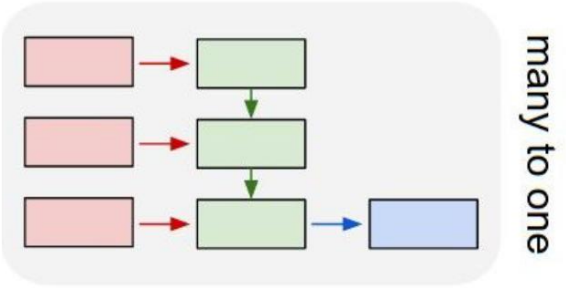
Image Captioning



A person riding a motorbike on dirt road

Sentiment Analysis

RNNs are awesome.



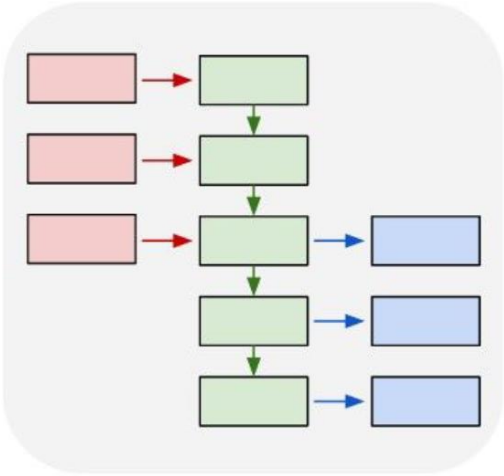
Positive

Web data sequence modeling tasks

Input **Output**

Machine Translation

Happy birthday!



ਜਨਮਦਿਨ ਮੁਬਾਰਕ

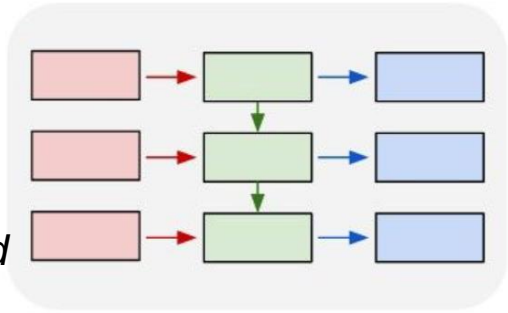
Janamadina mubāraka

Conversational search

Find me nearby restaurant

Chinese

Tell me more about the 2nd



Of what type of food?

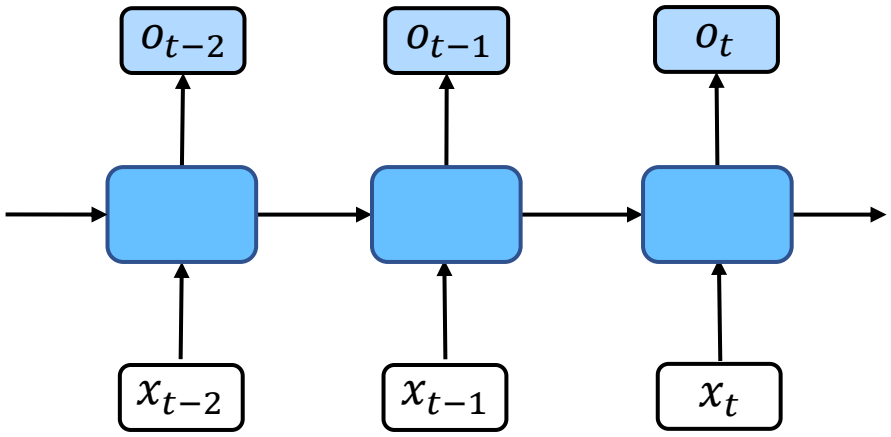
I found these ones: ….

Sure! Here you are ….

Auto regressive models

- In auto-regressive models the current output depends on the current input and a limited span of past inputs.

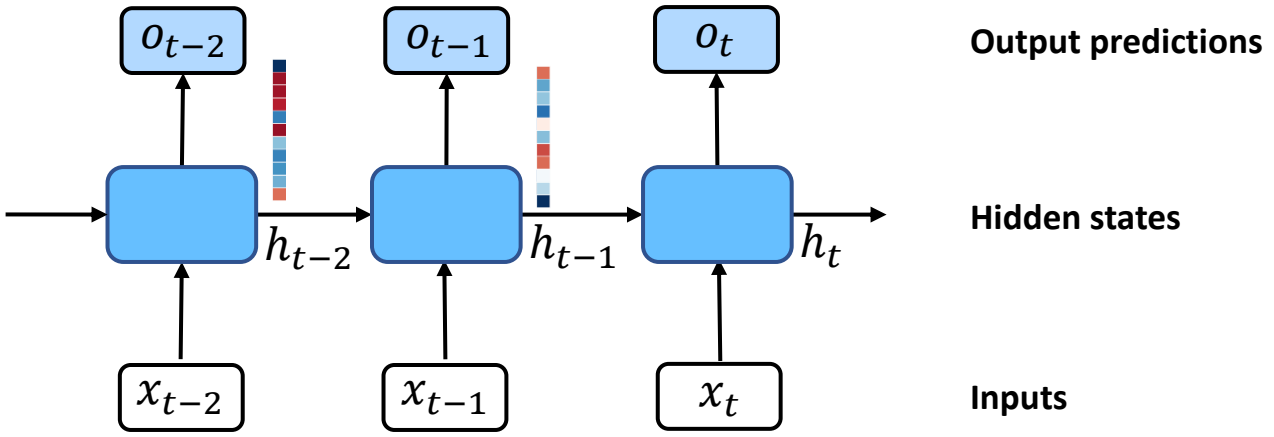
$$o_t = p(x_t \,|x_{t-1}, x_{t-2}, \ldots, x_{t-\tau})$$



Latent auto regressive models

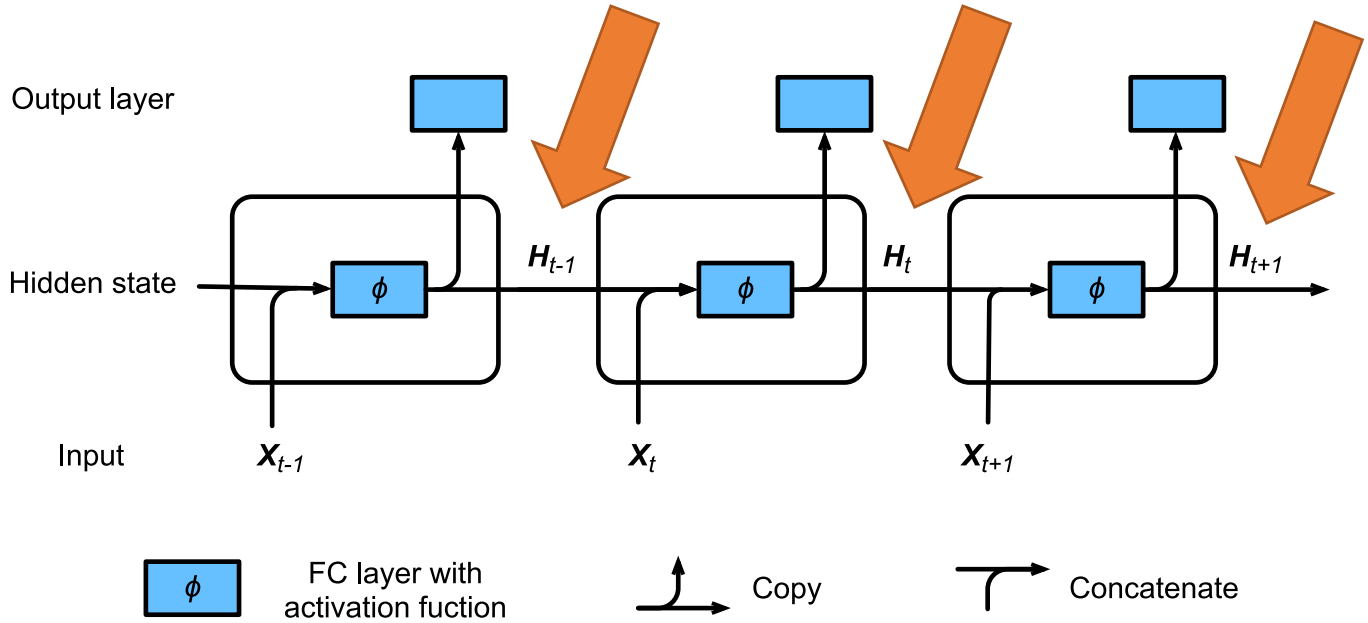
- In latent auto-regressive models, the model depends on the current input and a hidden state, capturing the past inputs:

$$o_t =\sim p(x_t|x_{t-1}, h_t)$$

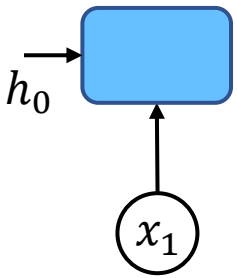


Hidden state

- The hidden state:
  - is propagated from state to state, and
  - it works as a memory to help decisions in later parts of the sequence.



Feedforward

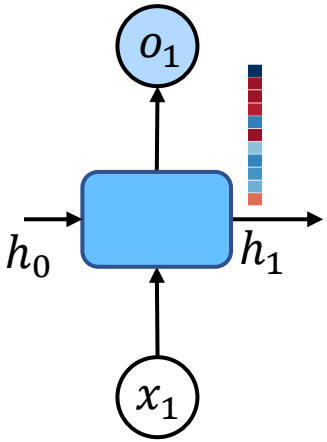


Output predictions

States

Inputs

Feedforward

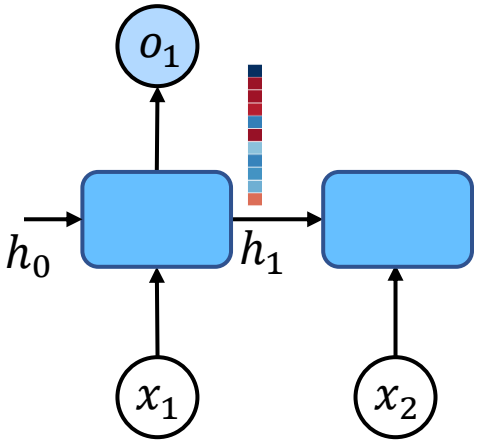


Output predictions

States

Inputs

Feedforward

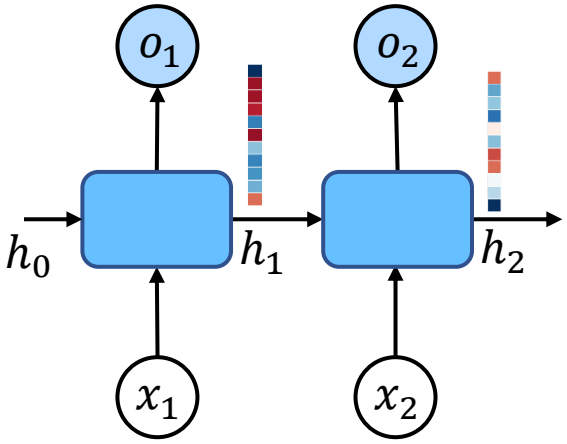


Output predictions

States

Inputs

Feedforward

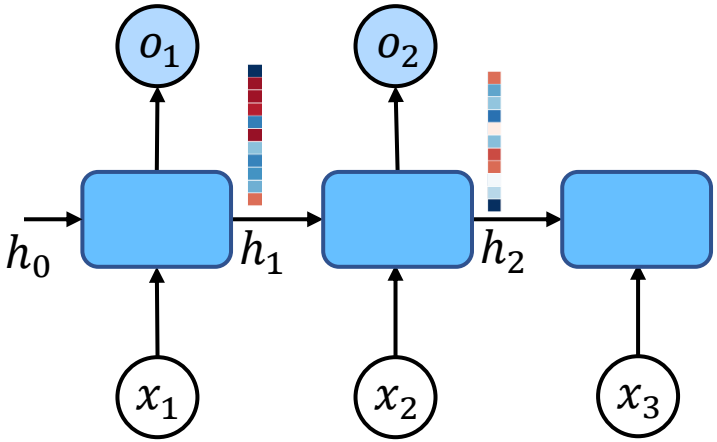


Output predictions

States

Inputs

Feedforward

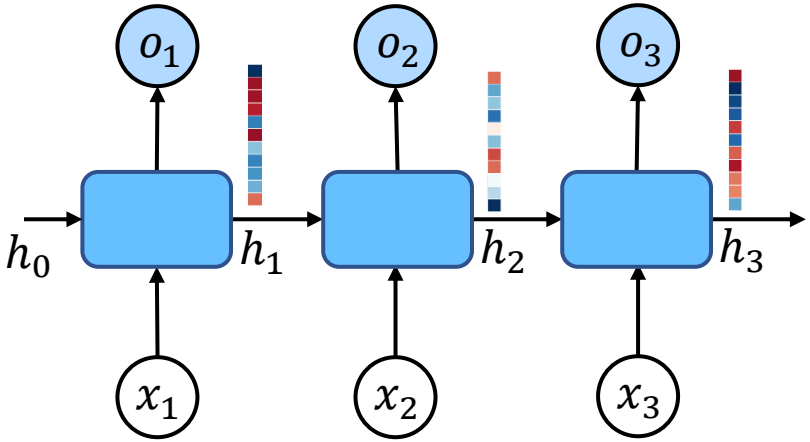


Output predictions

States

Inputs

Feedforward

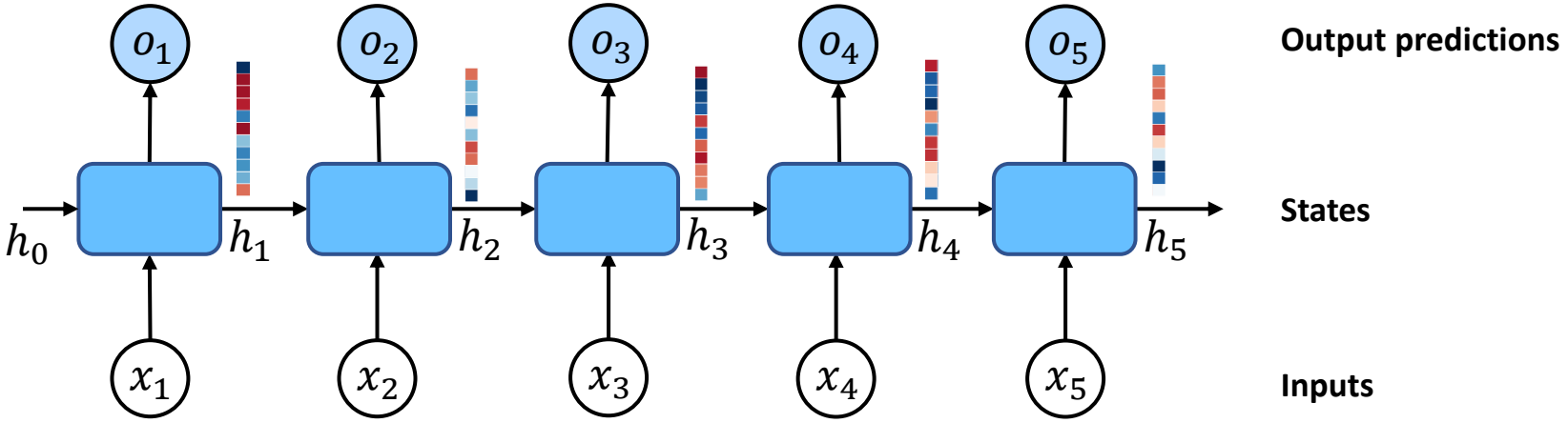


Output predictions

States

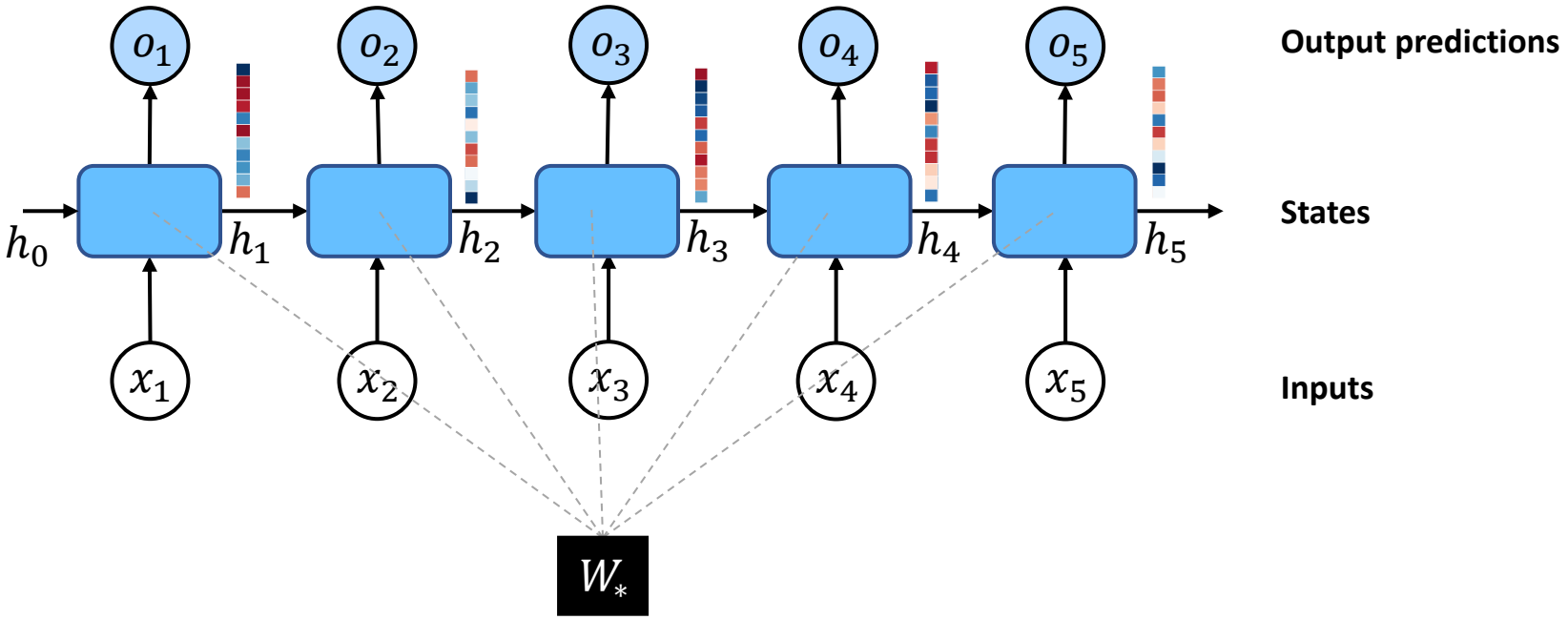
Inputs

Feedforward

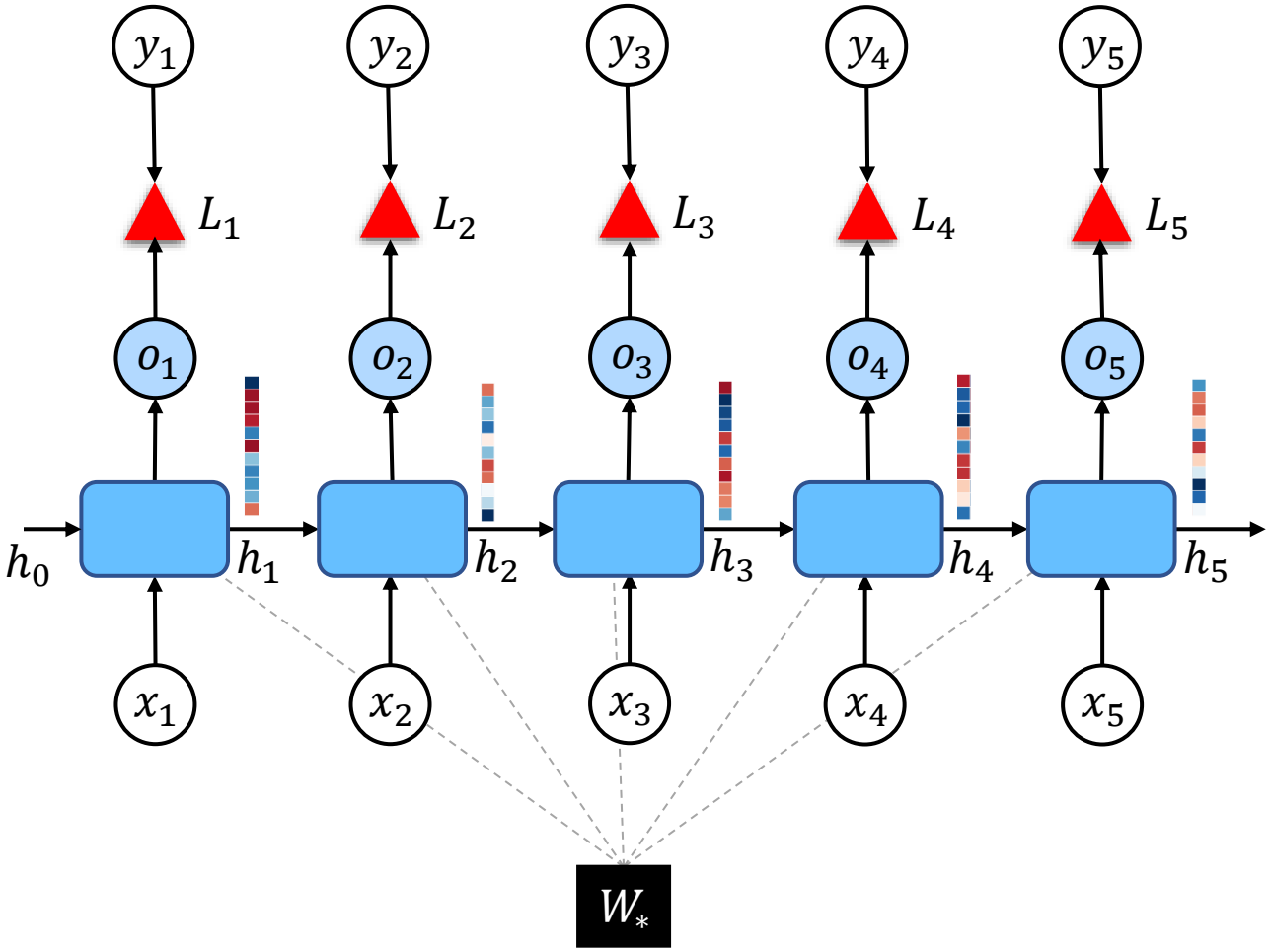


Parameter sharing

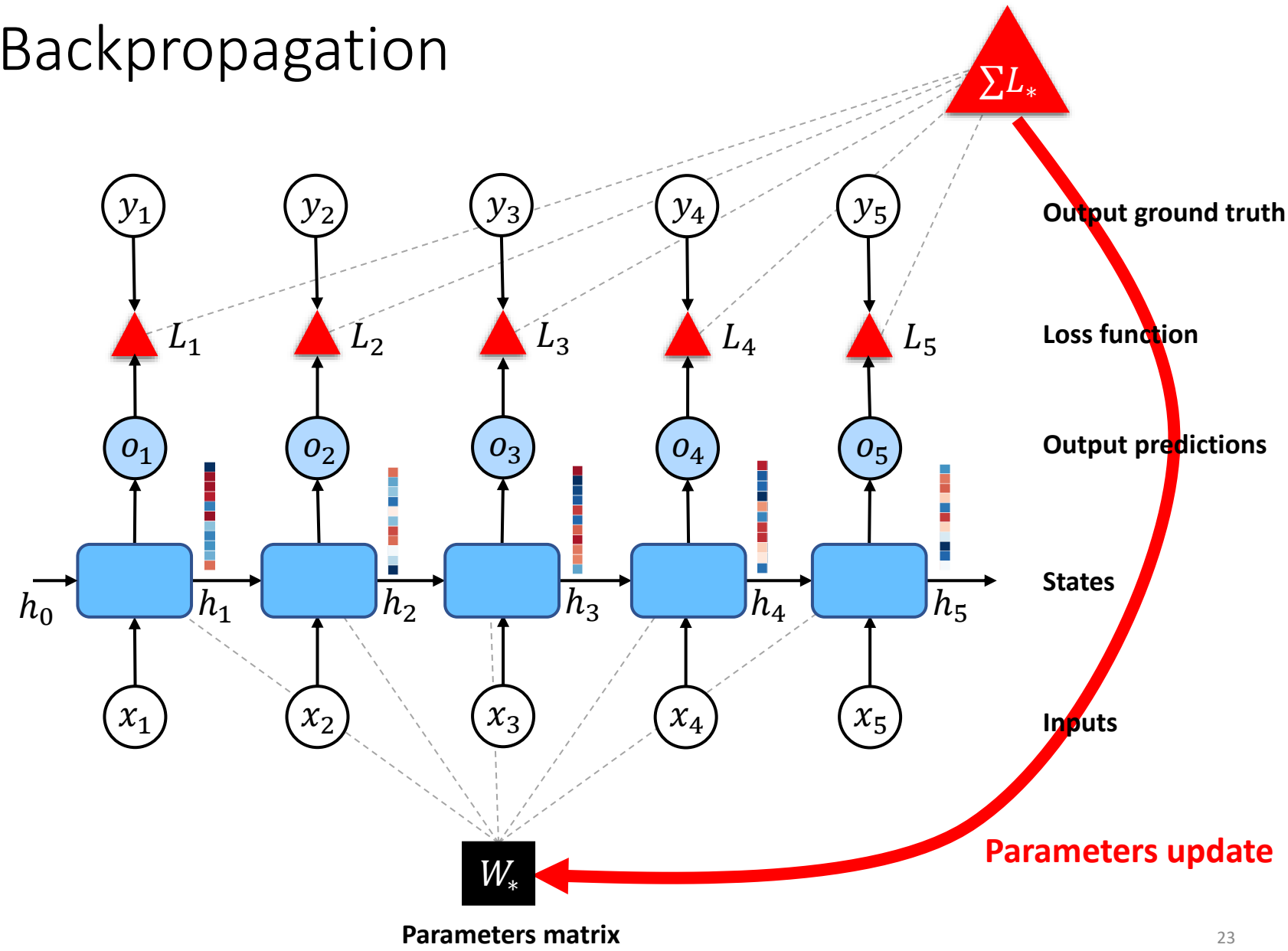
- There is only one RNN that runs through the sequence
- It has only one set of parameters



Training loss



Backpropagation



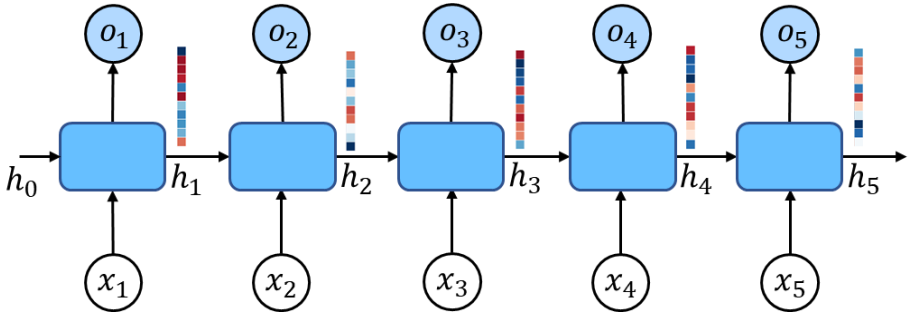
Types of RNNs

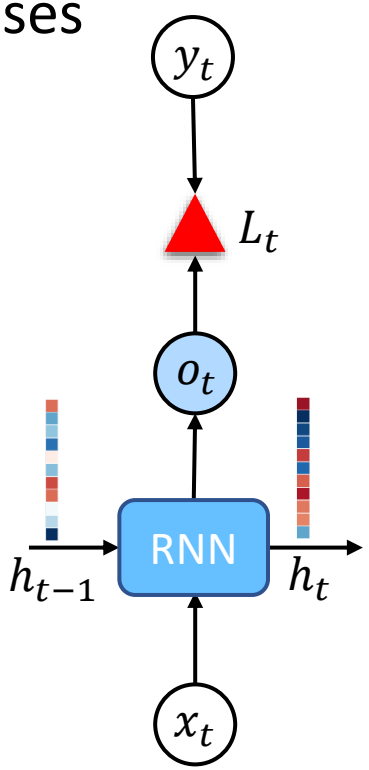
- Recurrent Neural Networks
- Gated Recurrent Neural Network
- Long-term Short-Term Memory

  - All three types of RNNs, try to tackle the memory problem of RNNs, also known as vanishing or exploding gradient problem.
  - There are many variations of each one of these three types of RNNs.

Elman's Recurrent Neural Network

- Elman's Recurrent Neural Network (RNN) processes sequences of data by processing each step t.
- The RNN unit is a recurrent function.
- The goal is to predict the output y at each step t.





Elman's Recurrent Neural Network

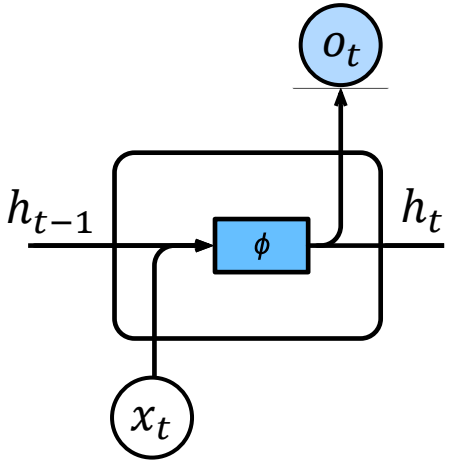
- The RNN unit preserves information from:
  - previous state h_{t-1};
  - current input data input x_t.

- The RNN unit is a recurrent function:

$$h_t = \phi\left([W_{hh}\ W_{xh}] \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} + b_h\right)$$

  - W_{xh} and W_{hh} are the RNN parameters matrix
  - The output at each step t is:

$$o_t = h_t \cdot W_{hq} + b_q$$

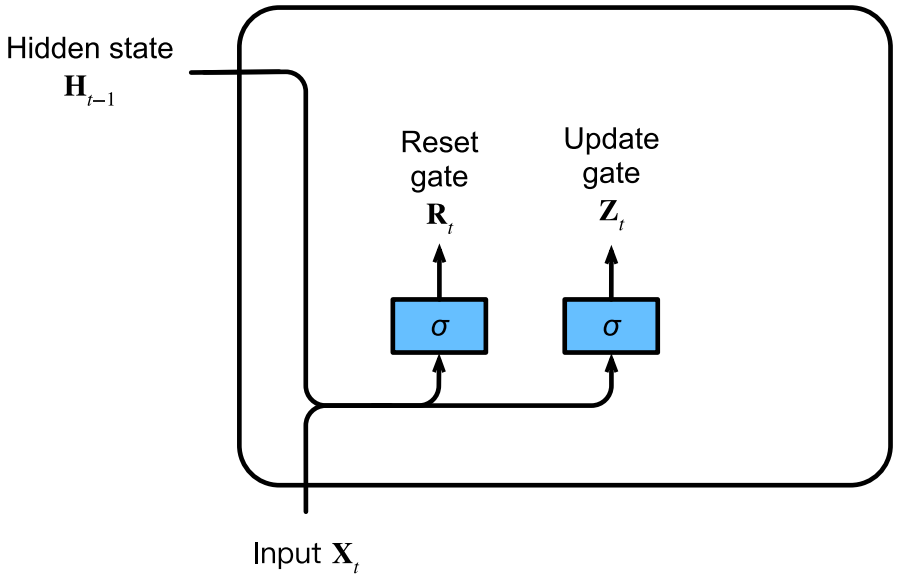


State and output are too tightly connected

- In Elman's RNN the output and the state are derived from the same variable.
- A unit's state should give more emphasis to the current data or the previous state.
- State passing mechanism should be able to control the amount of information from data and/or previous state that is encoded in a state
  - A better approach is to make a stronger separation between state and output.
- GRUs and LSTMs are the best examples of such idea.

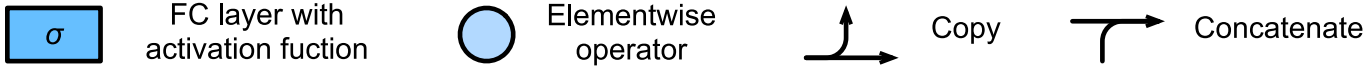
Gated Recurrent Unit: Old state information

- Introduces a reset memory and update state functions



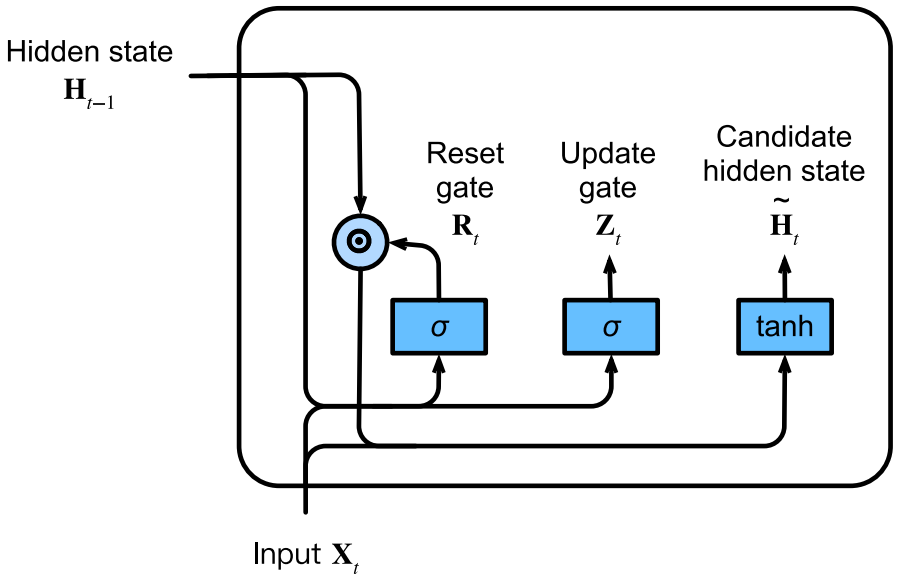
$$R_t = \sigma_r\left(W_r \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$Z_t = \sigma_t\left(W_u \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$



Gated Recurrent Unit: New candidate state

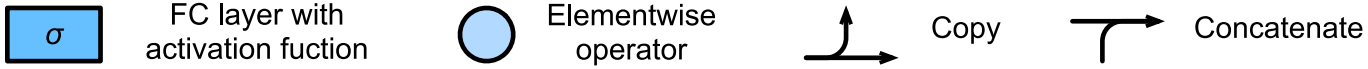
- Introduces a reset memory and update state functions
- Computes a candidate hidden state



$$R_t = \sigma_r\left(W_r \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

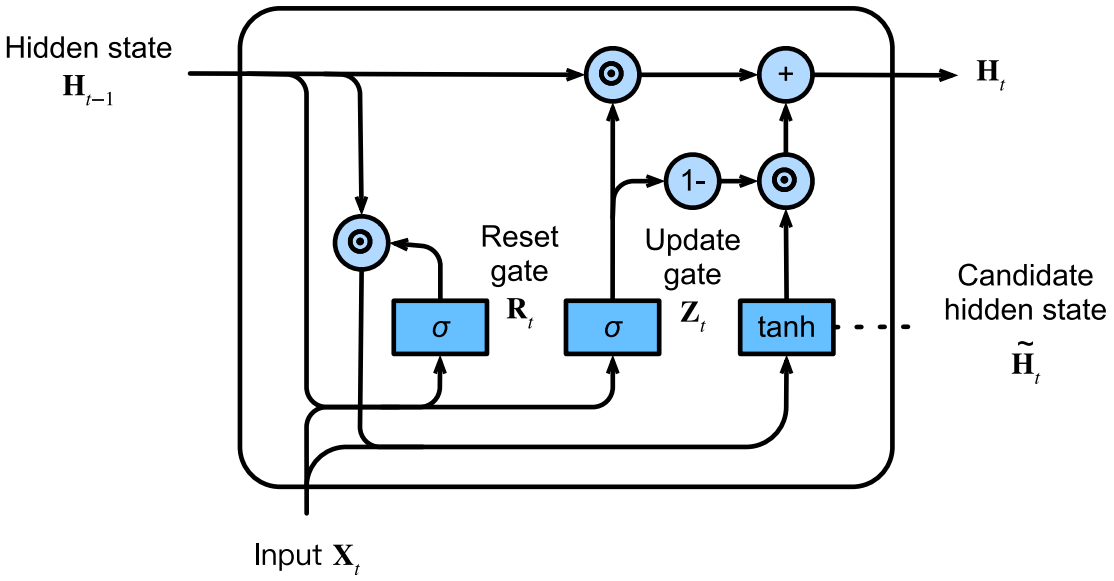
$$Z_t = \sigma_t\left(W_u \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$\tilde{h}_t = tanh\left(W_c \cdot \begin{bmatrix} R_t \odot h_{t-1} \\ x_t \end{bmatrix}\right)$$



Gated Recurrent Unit

- The new hidden state is a mixture of the candidate hidden state and the previous hidden state.

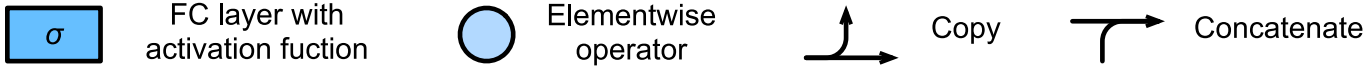


$$R_t = \sigma_r\left(W_r \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$Z_t = \sigma_t\left(W_u \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$\tilde{h}_t = tanh\left(W_c \cdot \begin{bmatrix} R_t \odot h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$h_t = Z_t \odot \tilde{h}_t + (1 - Z_t) \odot h_{t-1}$$



Gated Recurrent Unit

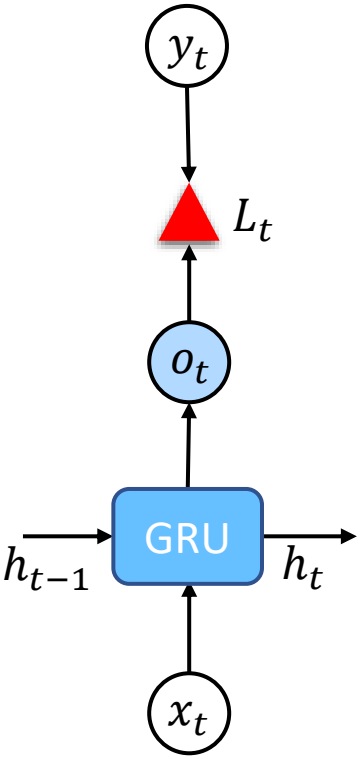
- The reset gate controls the **information that is deleted** from the previous state:

$$R_t = \sigma_r\left(W_r \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right) \qquad \tilde{h}_t = tanh\left(W_c \cdot \begin{bmatrix} R_t \odot h_{t-1} \\ x_t \end{bmatrix}\right)$$

- The update gate controls the **information that is preserved** from the previous state:

$$h_t = Z_t \odot \tilde{h}_t + (1 - Z_t) \odot h_{t-1} \qquad Z_t = \sigma_t\left(W_u \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

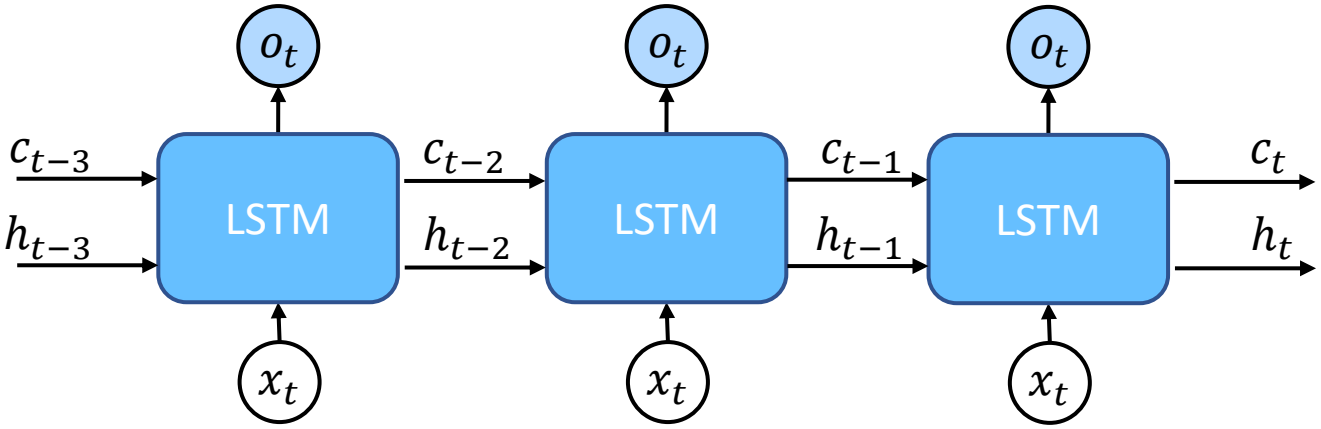
- The output is computed from the cell memory: $y_t = g_y(c_t)$



Long-Short Term Memory

- **Key idea 1:** Separate **state** from **memory**.
  - This allows to better preserve memory from past states and still generate the correct output from the hidden state.

- **Key idea 2:** Put the memory along an **uninterrupted path**.
  - It avoids the vanishing gradient problem and lets information propagate backwards.



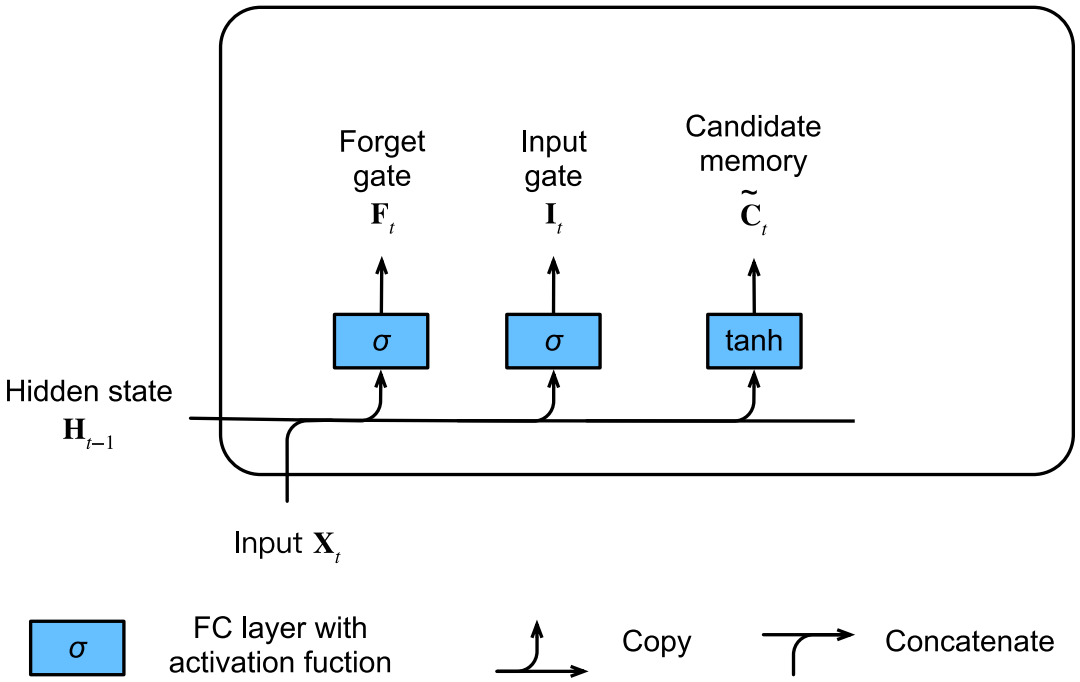
Long-Short Term Memory

- Similarly to the GRU, there are several gates to control the information flow.

$$F_t = \sigma\left(W_f \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$I_t = \sigma\left(W_i \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$\widehat{c_t} = tanh\left(W_c \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$



LSTM – Memory output

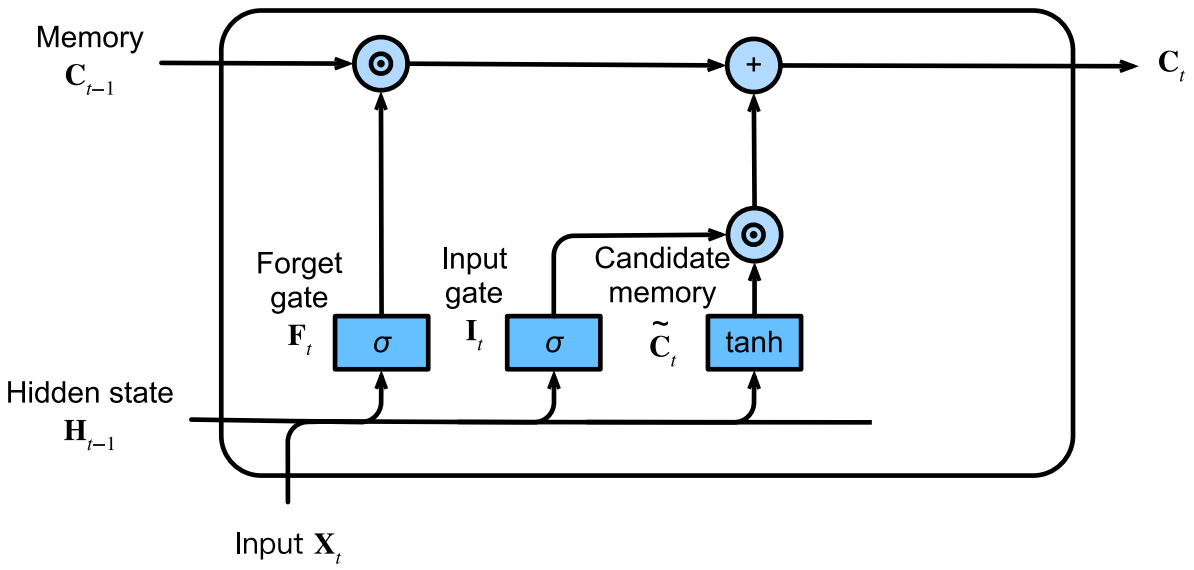
- The memory flows along an uninterrupted path.

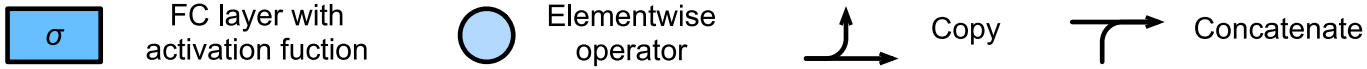
$$F_t = \sigma\left(W_f \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$I_t = \sigma\left(W_i \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$\widehat{c_t} = tanh\left(W_c \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$c_t = I_t \odot \widehat{c_t} + F_t \odot c_{t-1}$$





LSTM – Output gate

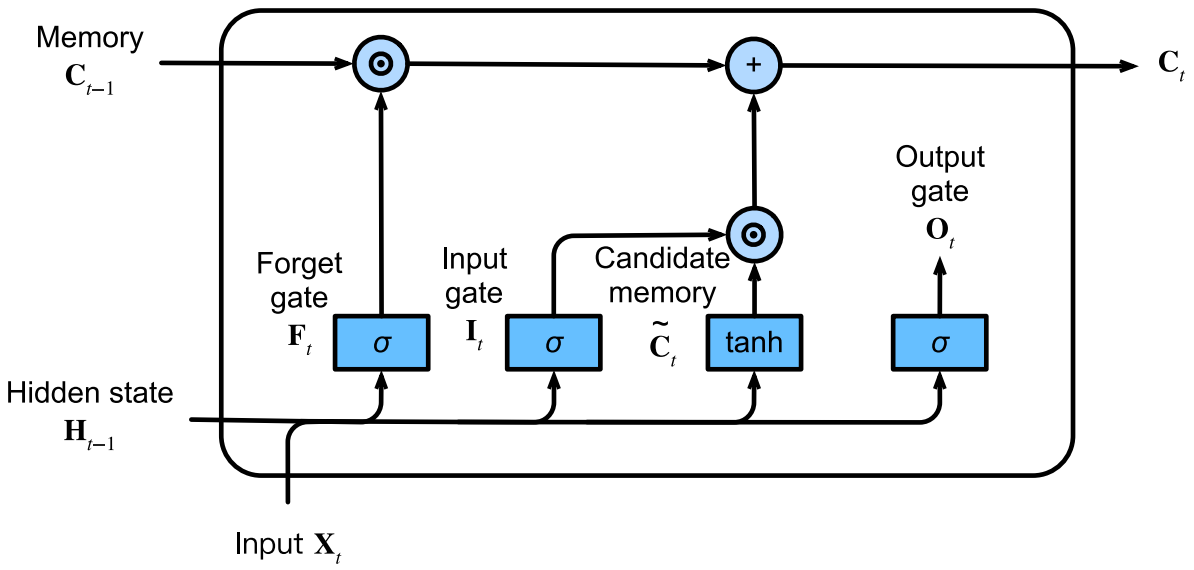
$$F_t = \sigma\left(W_f \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$I_t = \sigma\left(W_i \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$\widehat{c_t} = tanh\left(W_c \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$c_t = I_t \odot \widehat{c_t} + F_t \odot c_{t-1}$$

$$O_t = \sigma\left(W_o \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$



σ FC layer with activation fuction

Elementwise operator

Copy

Concatenate

LSTM – State output

The **new state** will be a combination of the memory t and the relevant part of data at t and state t-1

$$F_t = \sigma\left(W_f \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

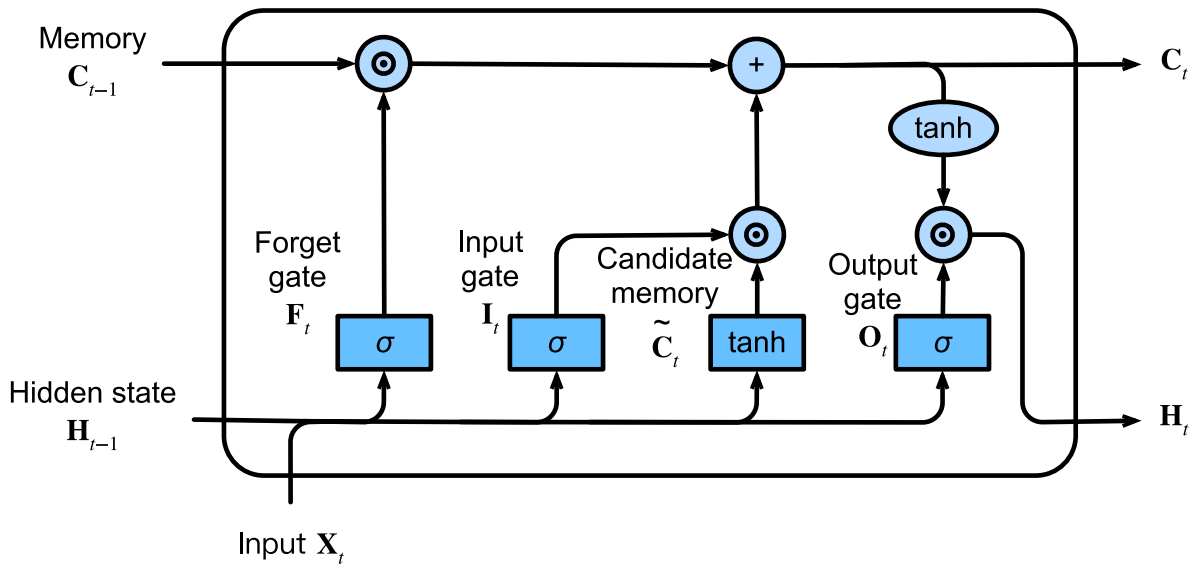
$$I_t = \sigma\left(W_i \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$\widehat{c_t} = tanh\left(W_c \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$c_t = I_t \odot \widehat{c_t} + F_t \odot c_{t-1}$$

$$O_t = \sigma\left(W_o \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$\mathrm{h}_t = \tanh(c_t) \odot O_t$$



σ FC layer with activation fuction

Elementwise operator

Copy

Concatenate

Long Short-Term Memory

Memory output

$$F_t = \sigma\left(W_u \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$I_t = \sigma\left(W_c \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

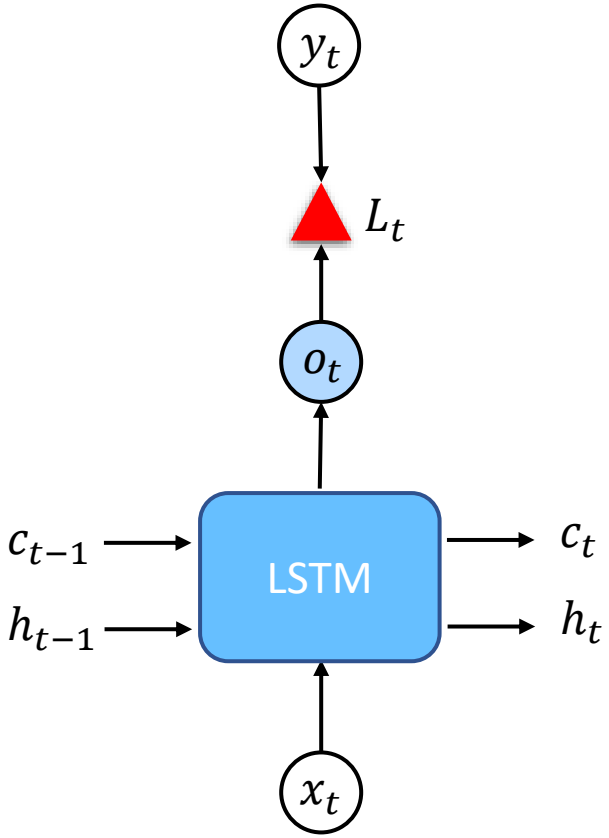
$$\widehat{c_t} = tanh\left(W_f \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$c_t = I_t * \widehat{c_t} + F_t * c_{t-1}$$

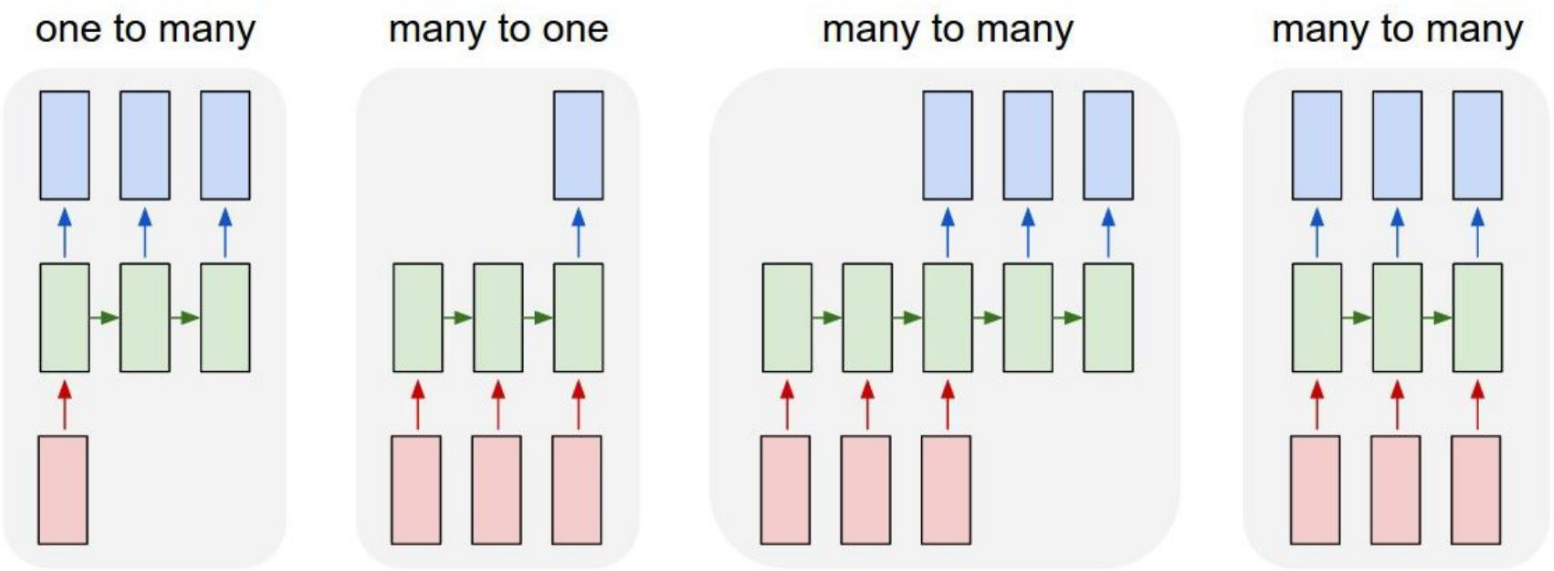
Hidden state output

$$O_t = \sigma\left(W_f \cdot \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix}\right)$$

$$\mathrm{h}_t = \tanh(c_t) * O_t$$

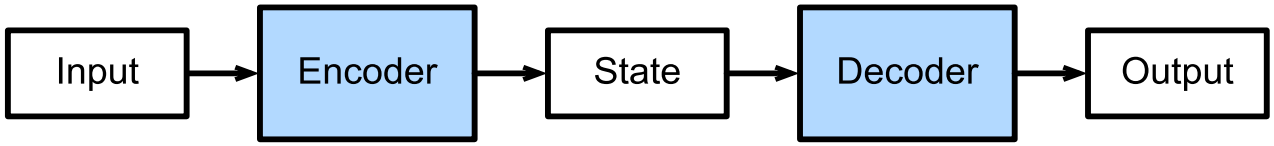


Any type of RNN can be used in any sequence task



Encoder-decoder

- The encoder-decoder is a design pattern.
- The encoder's role is to encode the inputs into state, which often contains several tensors.
- Then the state is passed into the decoder to generate the outputs.



Encoder-decoder

- The encoder and the decoder can have different architectures.
- In image captioning the encoder is a CNN and the decoder is an RNN.

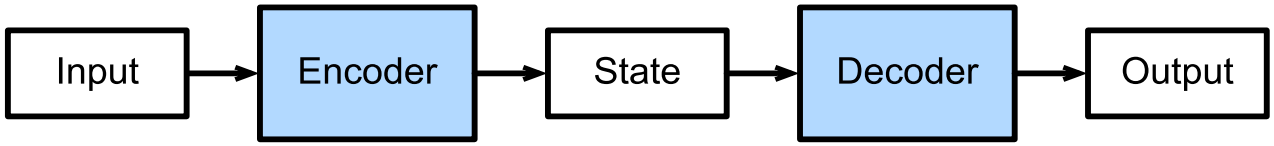


Image captioning example

This layer captures a good high level embedding of the image semantic content.

Image captioning example

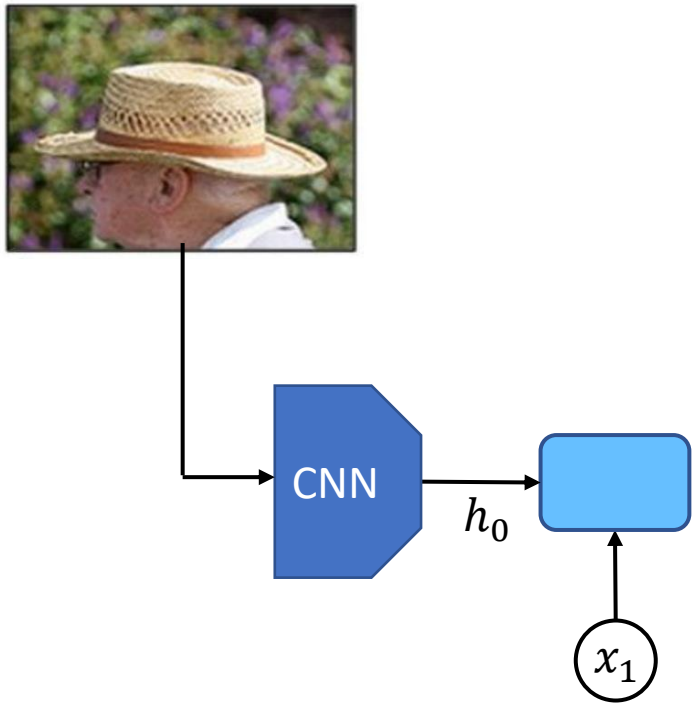


Image captioning example

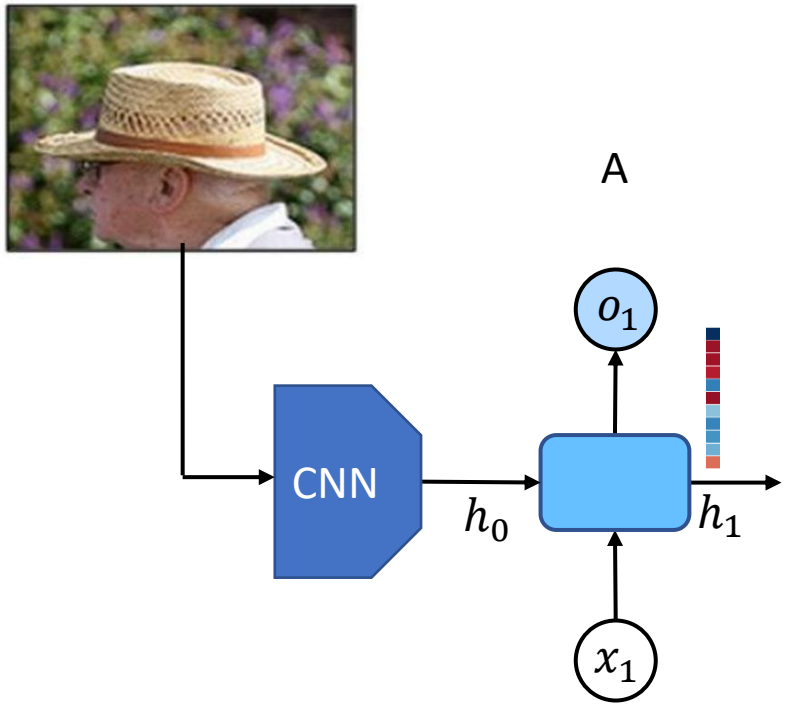


Image captioning example

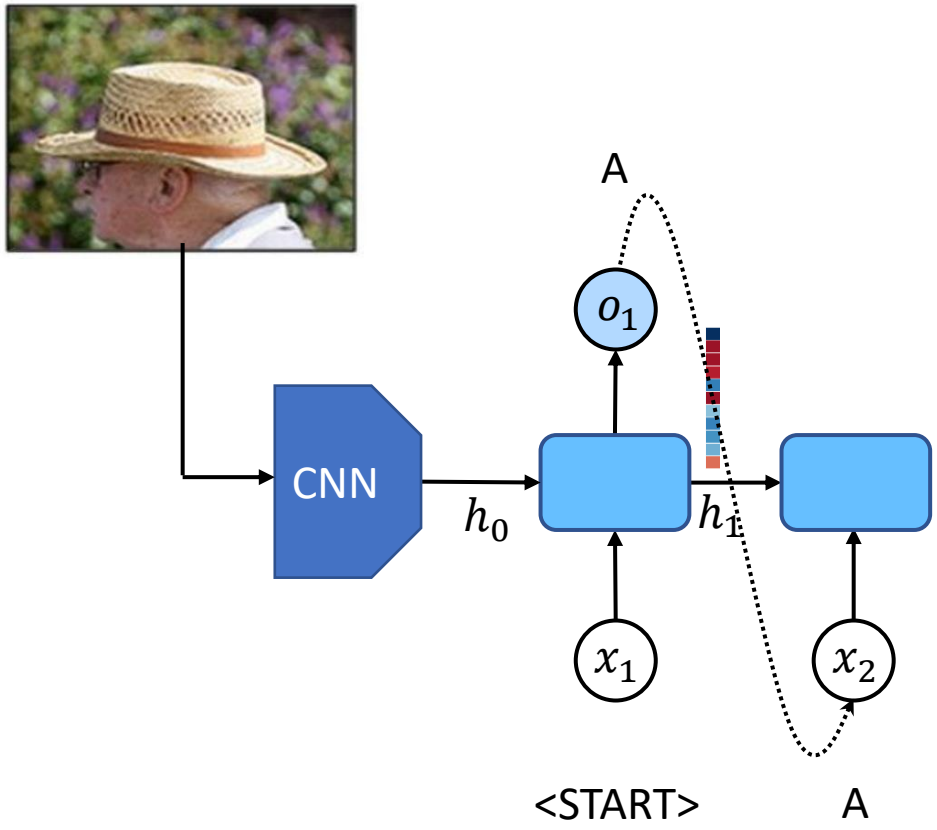


Image captioning example

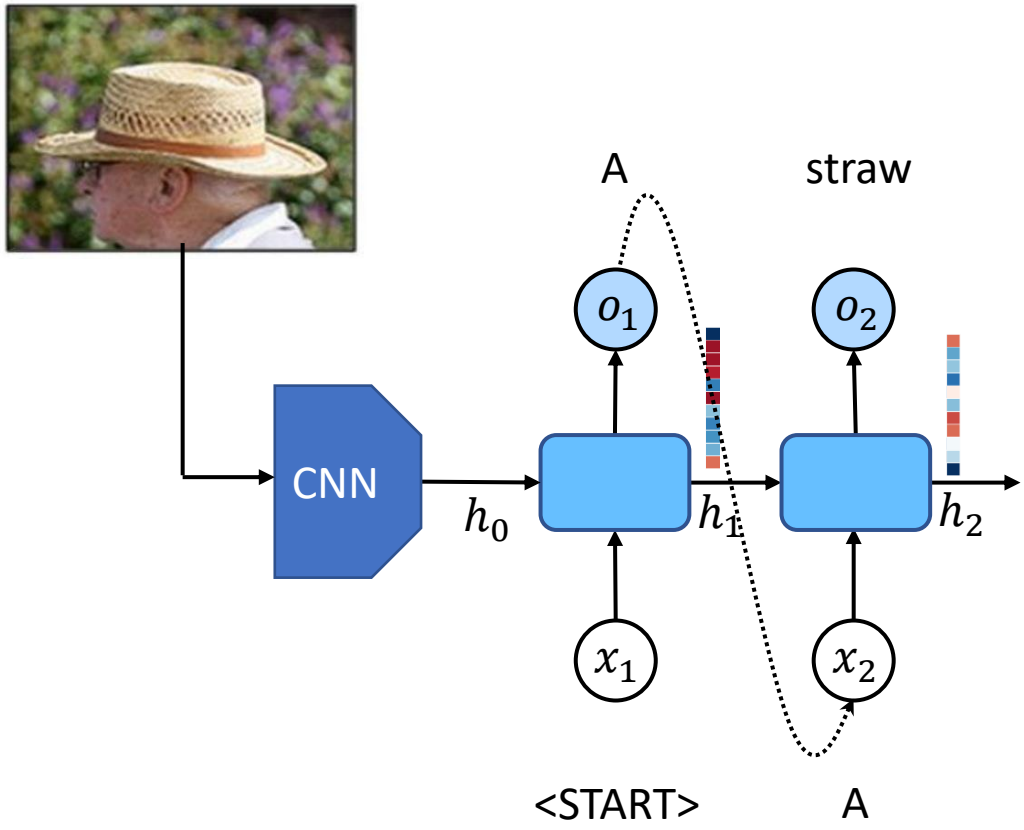


Image captioning example

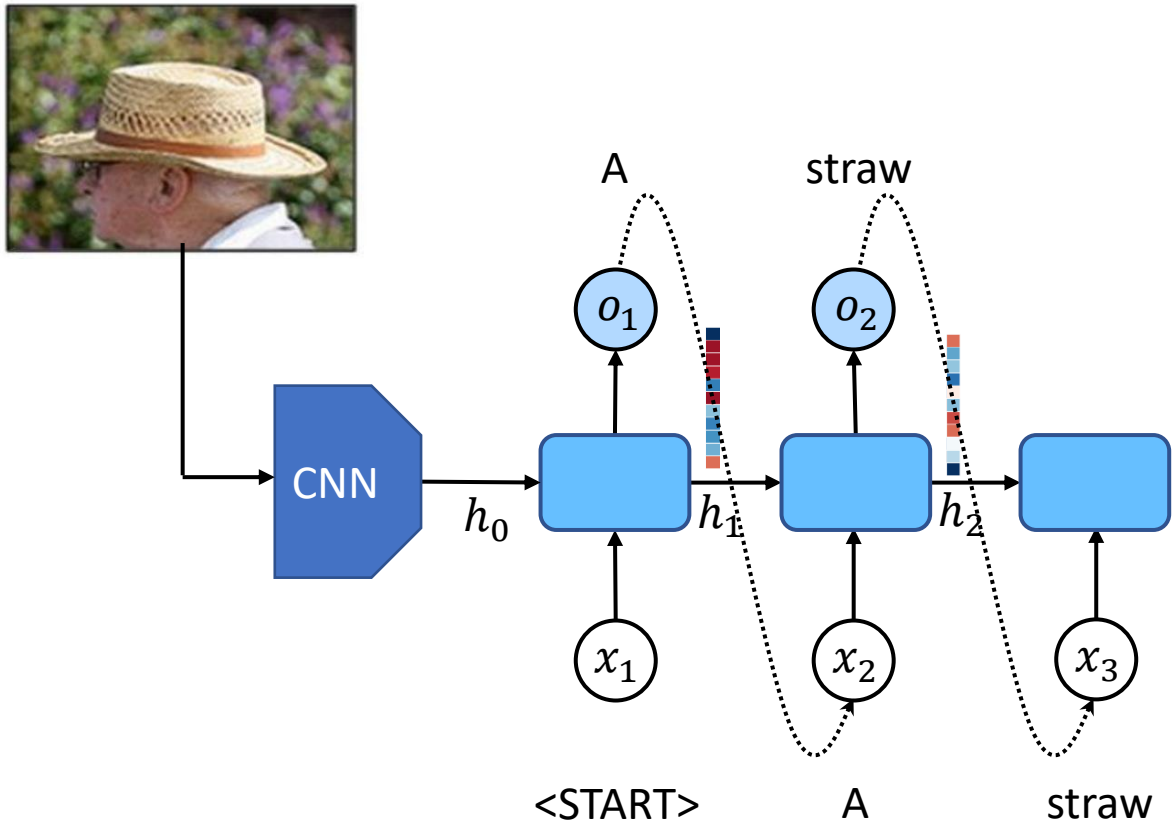


Image captioning example

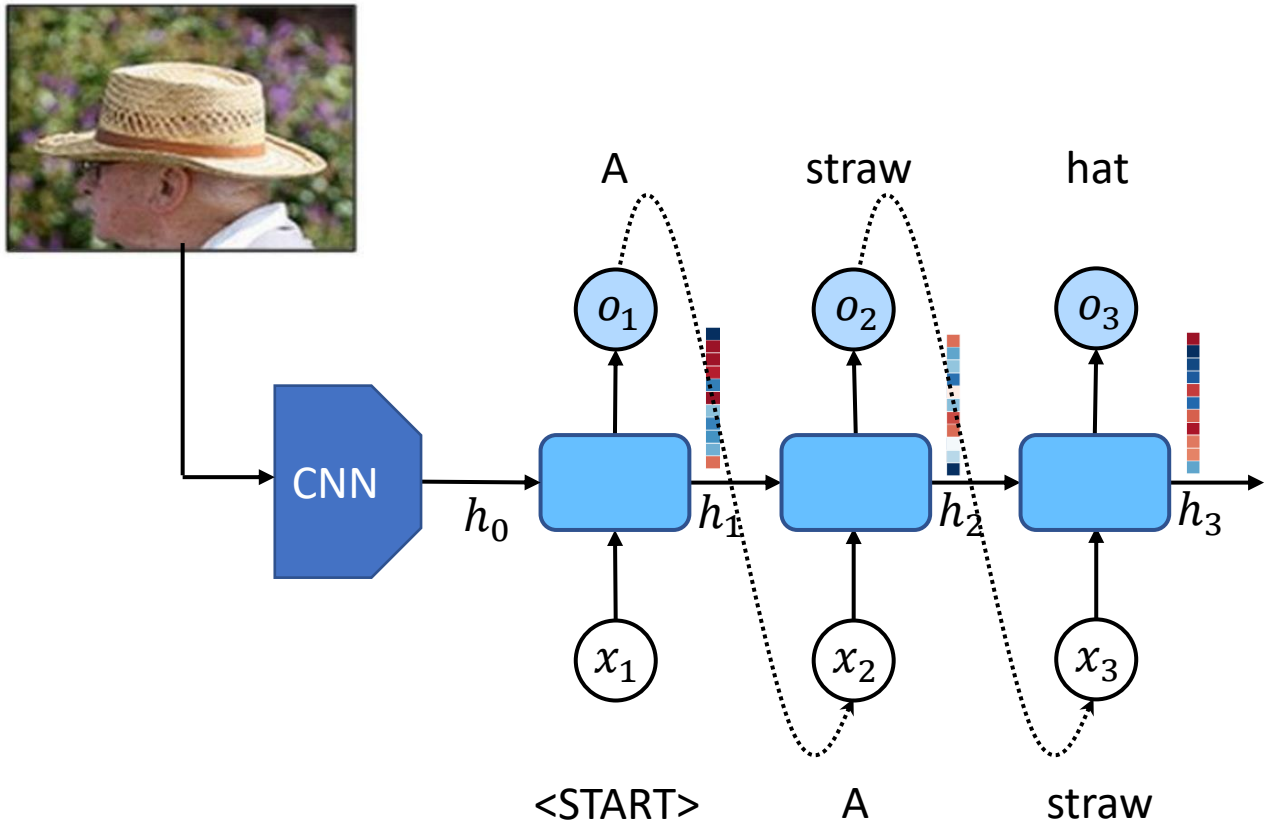


Image captioning example

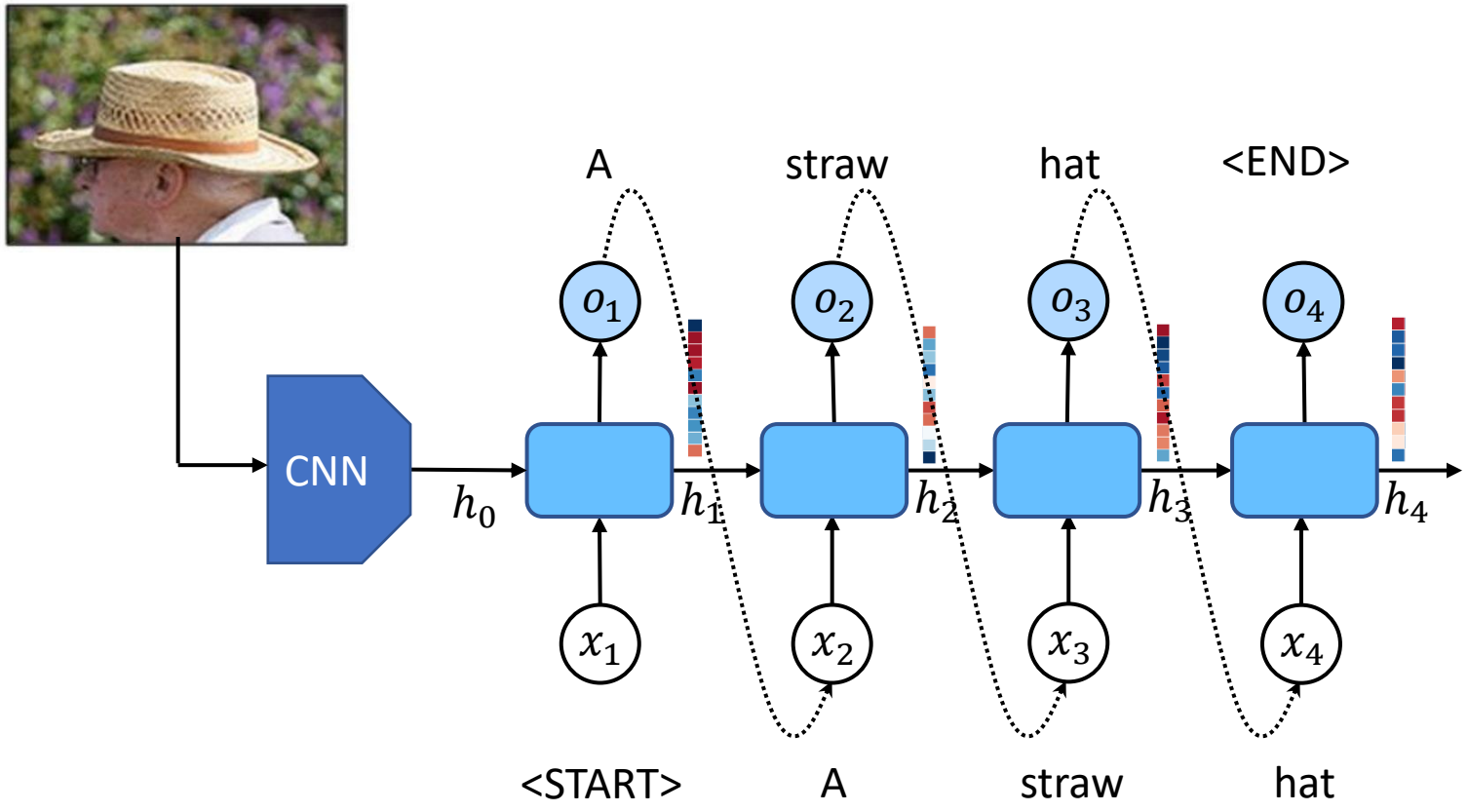


Image captioning examples

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

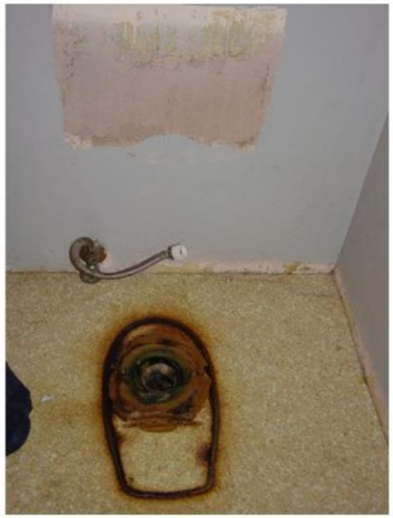
Two people walking on the beach with surfboards

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

Image captioning (bad) examples



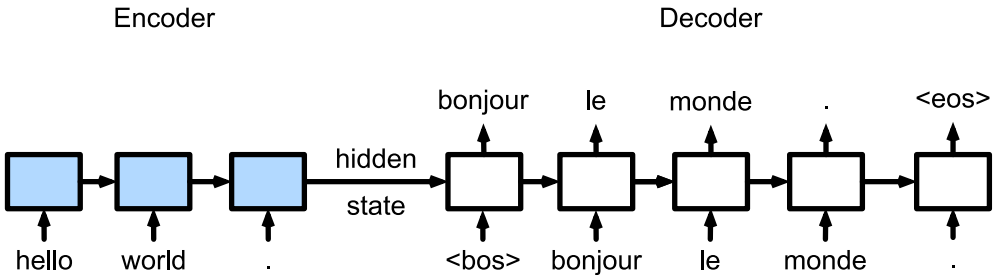
a toilet with a seat up in a bathroom
logprob: -13.44

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

a horse is standing in the middle of a road
logprob: -10.34

Sequence to sequence

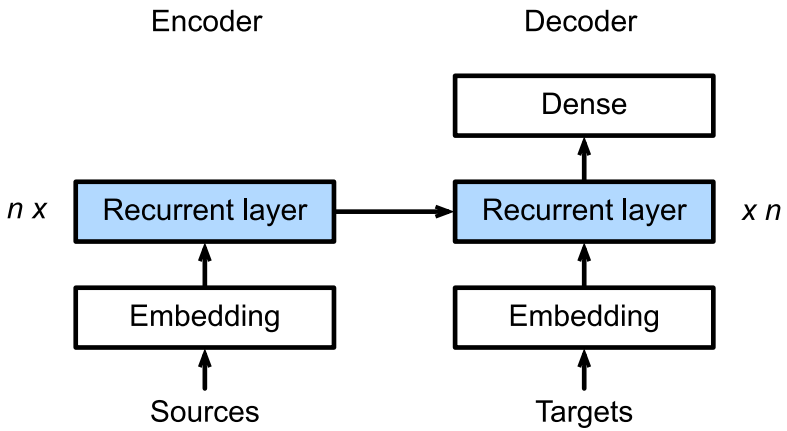
- The sequence to sequence (seq2seq) model is based on the encoder-decoder architecture to generate a sequence output for a sequence input



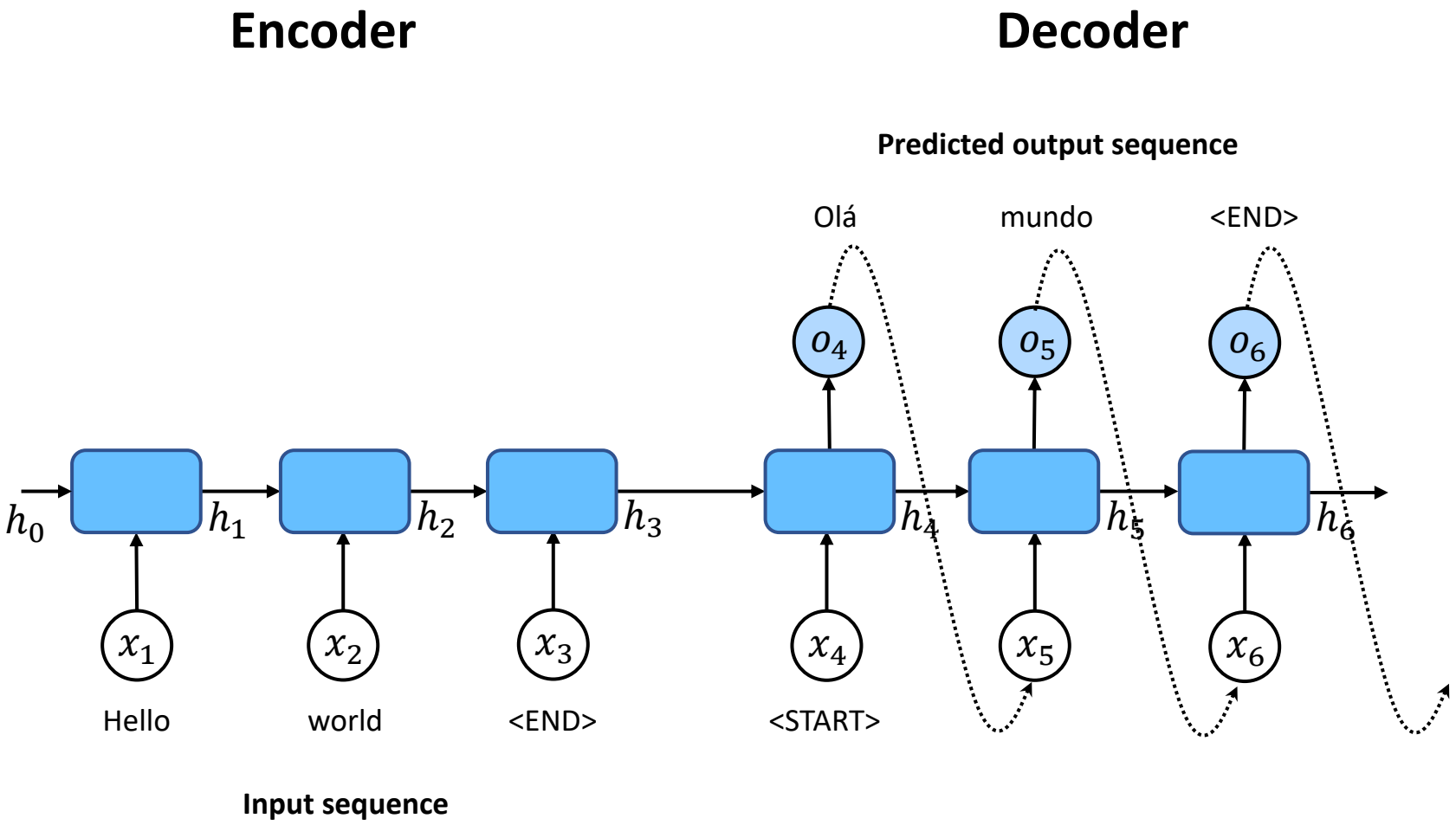
- Both the encoder and the decoder use RNNs to handle sequence inputs of variable length.
- The hidden state of the encoder is used directly to initialize the decoder hidden state.

Sequence to sequence

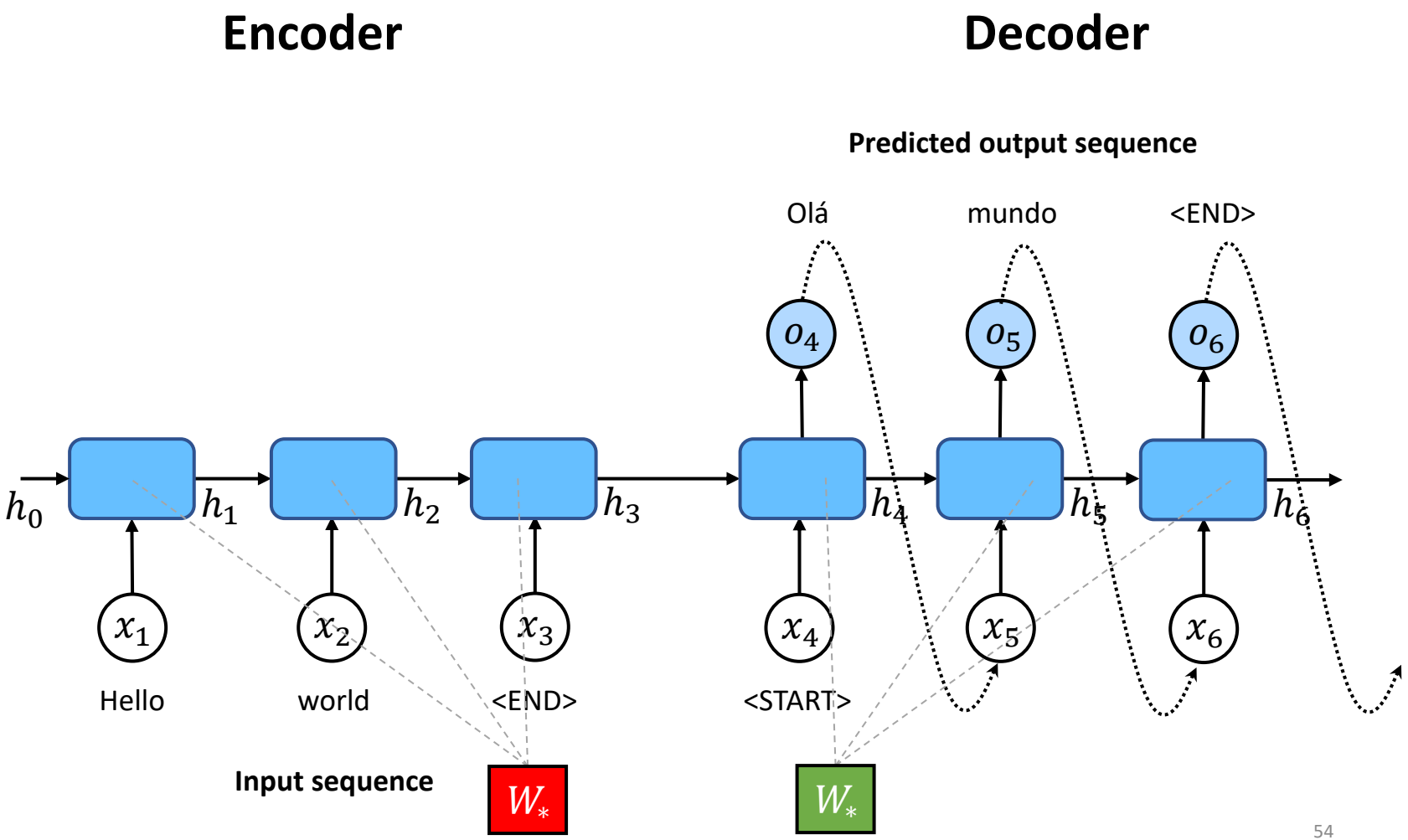
- Sequence to sequence models are a special case of encoder-decoder architectures.
- The hidden state of the encoder is used directly to initialize the decoder hidden state to pass information from the encoder to the decoder.



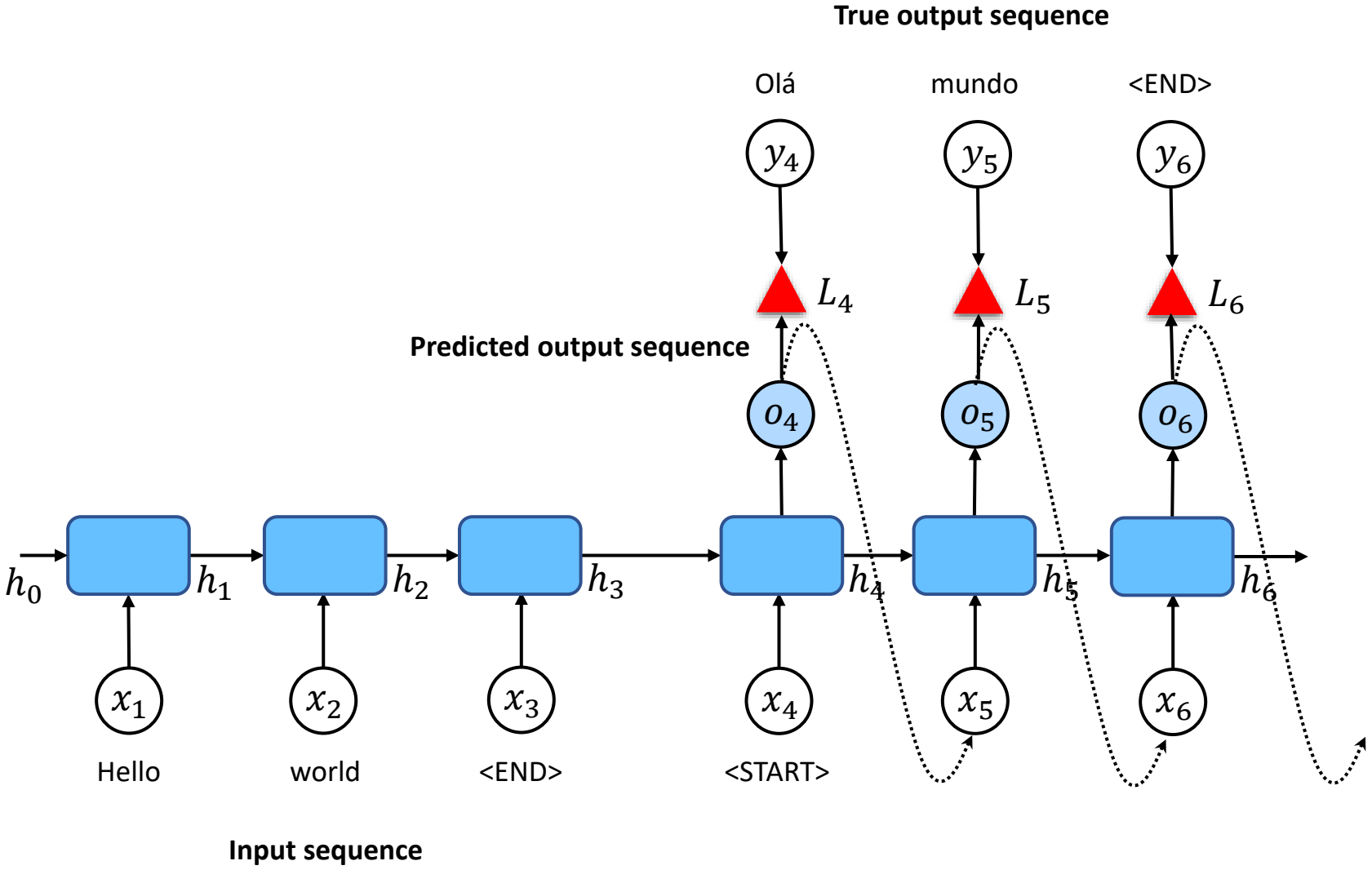
Feedforward in sequence-to-sequence models



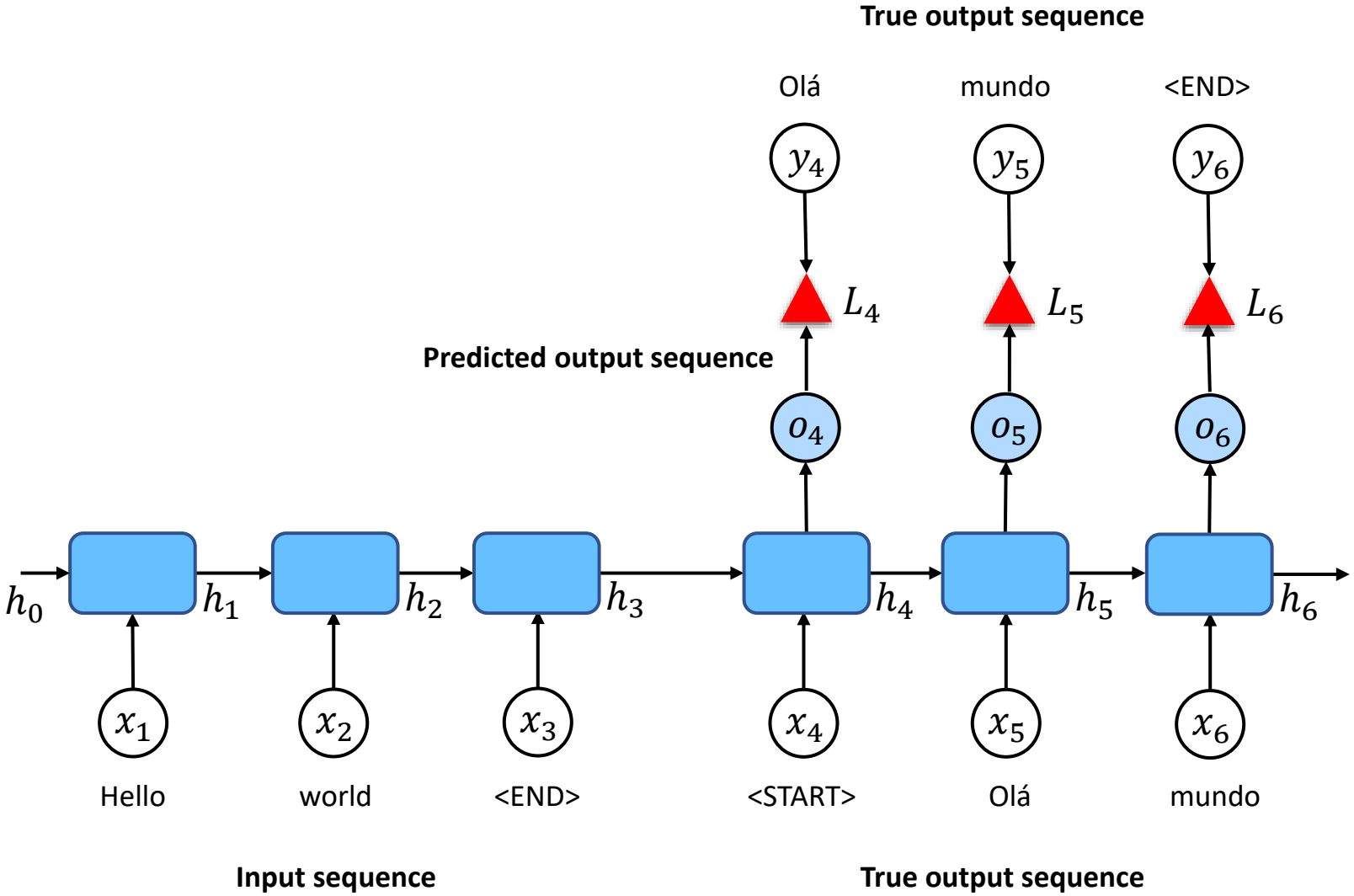
Encoder and decoder have different parameters



Training with reinjection



Training with teacher forcing



Summarization example

Daily Mail

The papal bill Pope Francis insisted on paying himself... before catching the bus home after winning the election

- **Pope Francis insisted on returning to his hotel to settle the bill himself**
- **The pontiff also chose to use a bus instead of a chauffeur driven car**
- **The 76-year-old has eschewed ceremonial traditions for a more humble approach**

With the spiritual wellbeing of the world's 1.2 billion Catholics is on his shoulders he must have quite a to-do list.

But despite his new responsibilities, Francis did not forget to stop off - between engagements - to pay his hotel bill.

Staff at the central Rome priests' residence where Bergoglio was staying before the conclave, were astonished when the newly elected Pope strolled in to collect his luggage and settle the bill.

© Getty Images

Pope Francis insisted on returning to the hotel to collect his luggage and greet the staff before settling the hotel bill himself

'I need to set a good example' he joked.

He was driven to the hotel in a simple car and The Rev. Pawel Rytel-Andrianek, who teaches at the nearby Pontifical Holy Cross University and is staying at the residence, said that workers at the hotel were touched by the Pope's decision to return and bid them farewell.

'He wanted to come here because he wanted to thank the personnel, people who work in this house,' he said. 'He greeted them one by one, no rush, the whole staff, one by one.' Mr Rytel-Andrianek added that Francis apparently knew everyone by name.

A Vatican spokesman said: 'He wanted to get his luggage and the bags. He had left everything there.

'He then stopped in the office, greeted everyone and decided to pay the bill for the room... because he was concerned about giving a good example of what priests and bishops should do.'

Francis is already winning plaudits for his down-to-earth manner.

He has so far refused a motorcade and the official papal Jag for official business. And even on the night of the election he insisted on accompanying the other cardinals back to their lodgings by mini bus, saying: 'I came on the bus, so I'll go home on the bus.'

Meeting cardinals yesterday on his second day of Papal business he eschewed protocol in favour of kissing on two cheeks, shaking hands and hugging.

He told his deputies that old people like himself are 'like good wine, getting better with age', before urging them to impart their wisdom to the young.

Francis began his reign in unorthodox fashion as he shunned public events in order to pray to the Virgin Mary.

During his first Mass since being elected as supreme pontiff, Pope Francis and his cardinals were dressed in simple yellow robes over their cassocks, rather than the formal ceremonial outfits they would normally wear on such a major occasion.

Speaking in Italian without notes, he said: 'We can walk all we want, we can build many things, but if we don't proclaim Jesus Christ, something is wrong. We would become a compassionate NGO and not a Church which is the bride of Christ.

'He who does not pray to the Lord prays to the devil. When we don't proclaim Jesus Christ, we proclaim the worldliness of the devil, the worldliness of the demon.'

'We must always walk in the presence of the Lord, in the light of the Lord, always trying to live in an irreprehensible way,' he said in a heartfelt homily of a parish priest, loaded with biblical references and simple imagery.

'When we walk without the cross, when we build without the cross and when we proclaim Christ without the cross, we are not disciples of the Lord. We are worldly,' he said

'We may be bishops, priests, cardinals, popes, all of this, but we are not disciples of the Lord,' he said.

It was a far simpler message than the dense, three-page discourse Benedict delivered in Latin during his first Mass as pope in 2005.

The difference in style was a sign of Francis' belief that the Catholic Church needs to be at one with the people it serves and not impose its message on a society that often doesn't want to hear it, Francis' authorised biographer, Sergio Rubin, said.

Francis took the helm of the 1.2 billion-member Church at a time of strife and intrigue, with the Vatican rocked by a string of sex abuse scandals, accusations of infighting within its central government and by allegations of financial wrongdoing.

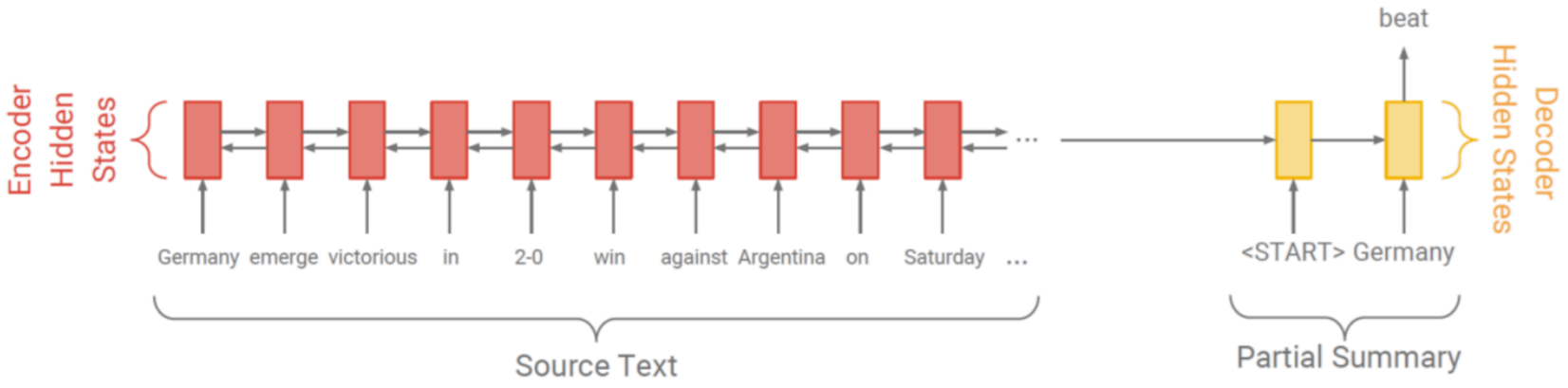
But many within the church believe he could change it for the better.

'It seems to me for now what is certain is it's a great change of style, which for us isn't a small thing,' Mr Rubin said, recalling how the former Cardinal Jorge Bergoglio would celebrate Masses with homeless people and prostitutes in Buenos Aires.

'He believes the church has to go to the streets,' he said, 'to express this closeness of the church and this accompaniment with those who are suffering.'

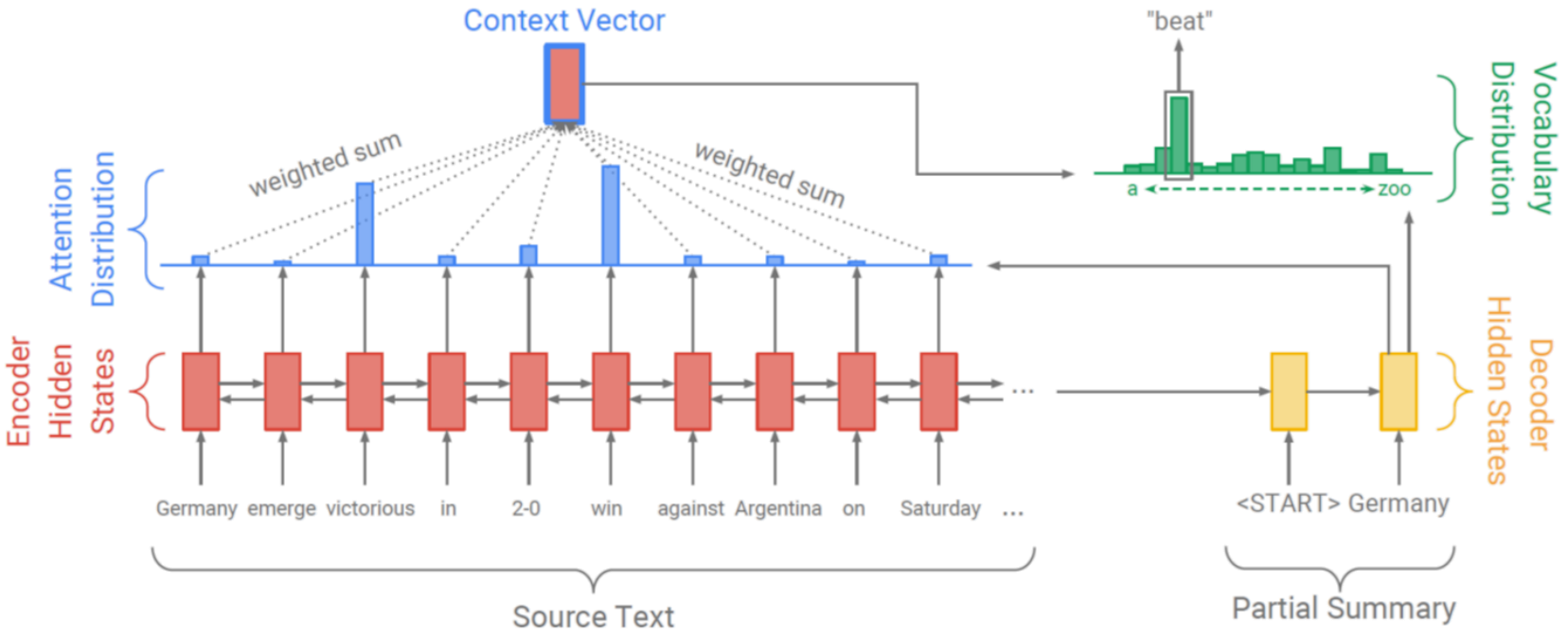
See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In ACL. https://github.com/ymfa/seq2seq-summarizer

Encoder-decoder based summarization



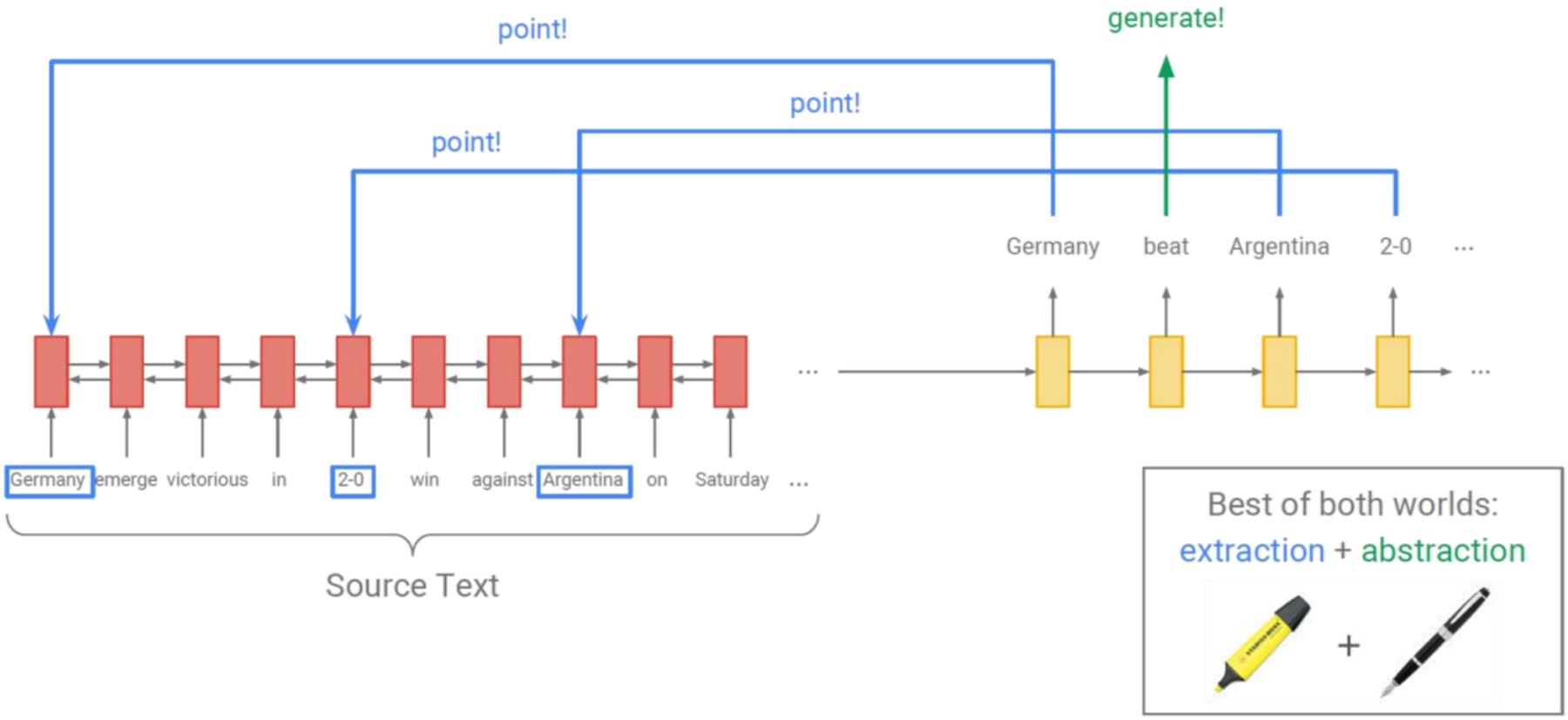
See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In ACL. https://github.com/ymfa/seq2seq-summarizer

Attention based summarization



See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In ACL.
https://github.com/ymfa/seq2seq-summarizer

Get to the point!



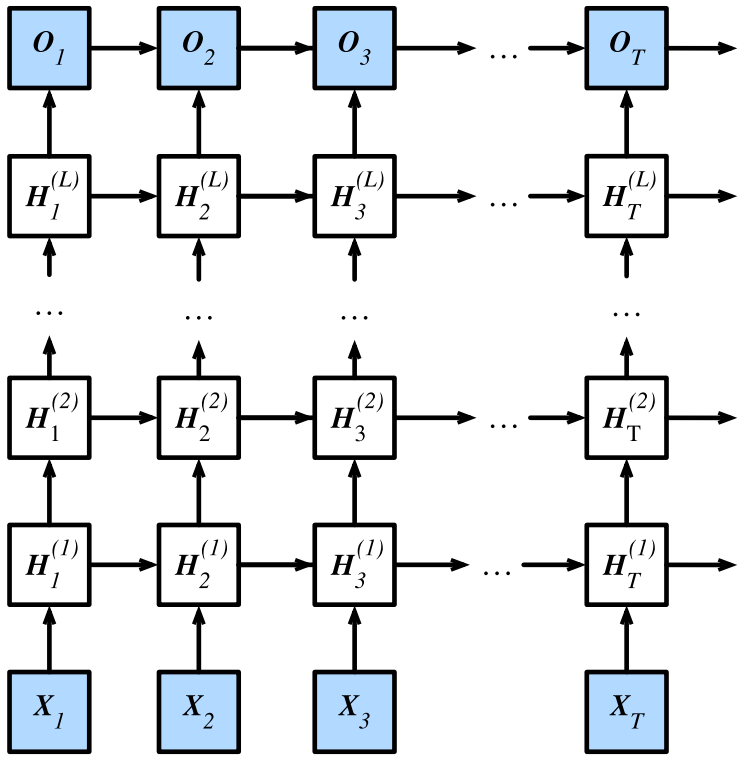
See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In ACL. https://github.com/ymfa/seq2seq-summarizer

Summary

- Sequence models:
  - RNNs can nicely model sequence data.
  - GRUs and LSTMs overcome some of the memory limitations.
- Architectures:
  - Deep archictures to capture complex interactions
  - Bi-direction architectures to capture long-term dependencies
  - Encoder-decoder
  - Sequence to sequence
- **Tasks**: machine translation, image captioning, summarization.
- Dive into Deep Learning chapters 8 and 9:
  - http://d2l.ai/chapter_recurrent-neural-networks/index.html
  - http://d2l.ai/chapter_recurrent-modern/index.html

Deep RNNs

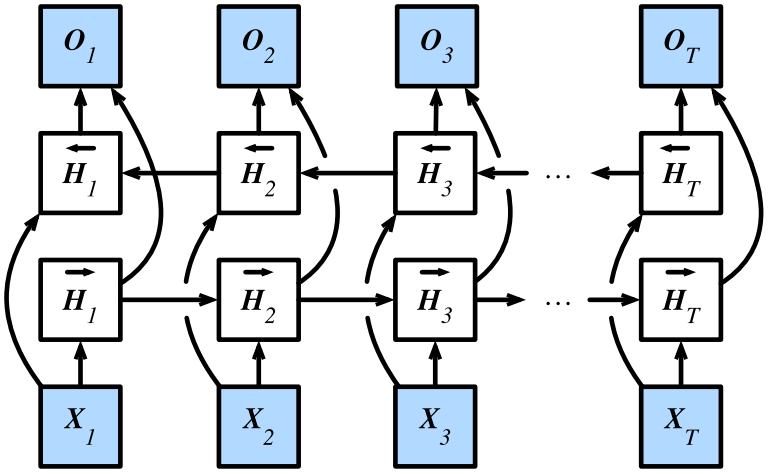
- Stack multiple layers of LSTMs on top of each other.
- This results in a mechanism that is more flexible, due to the combination of several simple layers.
- In particular, data might be relevant at different levels of the stack.
  - For instance, we might want to keep high-level data about financial market conditions (bear or bull market) available, whereas at a lower level we only record shorter-term temporal dynamics.



Bidirectional RNNs

- There are many tasks where the prediction is in the middle of the sequence:
  - I am _____
  - I am _____ very hungry.
  - I am _____ very hungry, I could eat half a pig

- In a bidirectional RNN information from both ends of the sequence is used to estimate the output.



Web Images Captioning / Description

- Karpathy, A., & Fei-Fei, L. “Deep visual-semantic alignments for generating image descriptions”. IEEE CVPR 2015.

- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. “Show, attend and tell: Neural image caption generation with visual attention”. ICML 2015.
    - https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

- Yu, J., Li, J., Yu, Z., & Huang, Q. (2019). Multimodal transformer with multi-view visual representation for image captioning. IEEE Transactions on Circuits and Systems for Video Technology.

- Herdade, S., Kappeler, A., Boakye, K., & Soares, J., “Image Captioning: Transforming Objects into Words”. NIPS 2019.

Readings: Machine translation

- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." In Advances in neural information processing systems, pp. 3104-3112. 2014.

- Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

- https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html

Readings: Web Text Summarization

- Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. In Advances in NIPS.
- Paulus, R., Xiong, C., & Socher, R. (2018). A deep reinforced model for abstractive summarization. In ICLR.
- See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In ACL.
- https://github.com/ymfa/seq2seq-summarizer