

# Lecture 8: RNNs

Alan Ritter

(many slides from Greg Durrett)

# Recall: Training Tips

---

# Recall: Training Tips

---

- ▶ Parameter initialization is critical to get good gradients, some useful heuristics (e.g., Xavier initializer)

# Recall: Training Tips

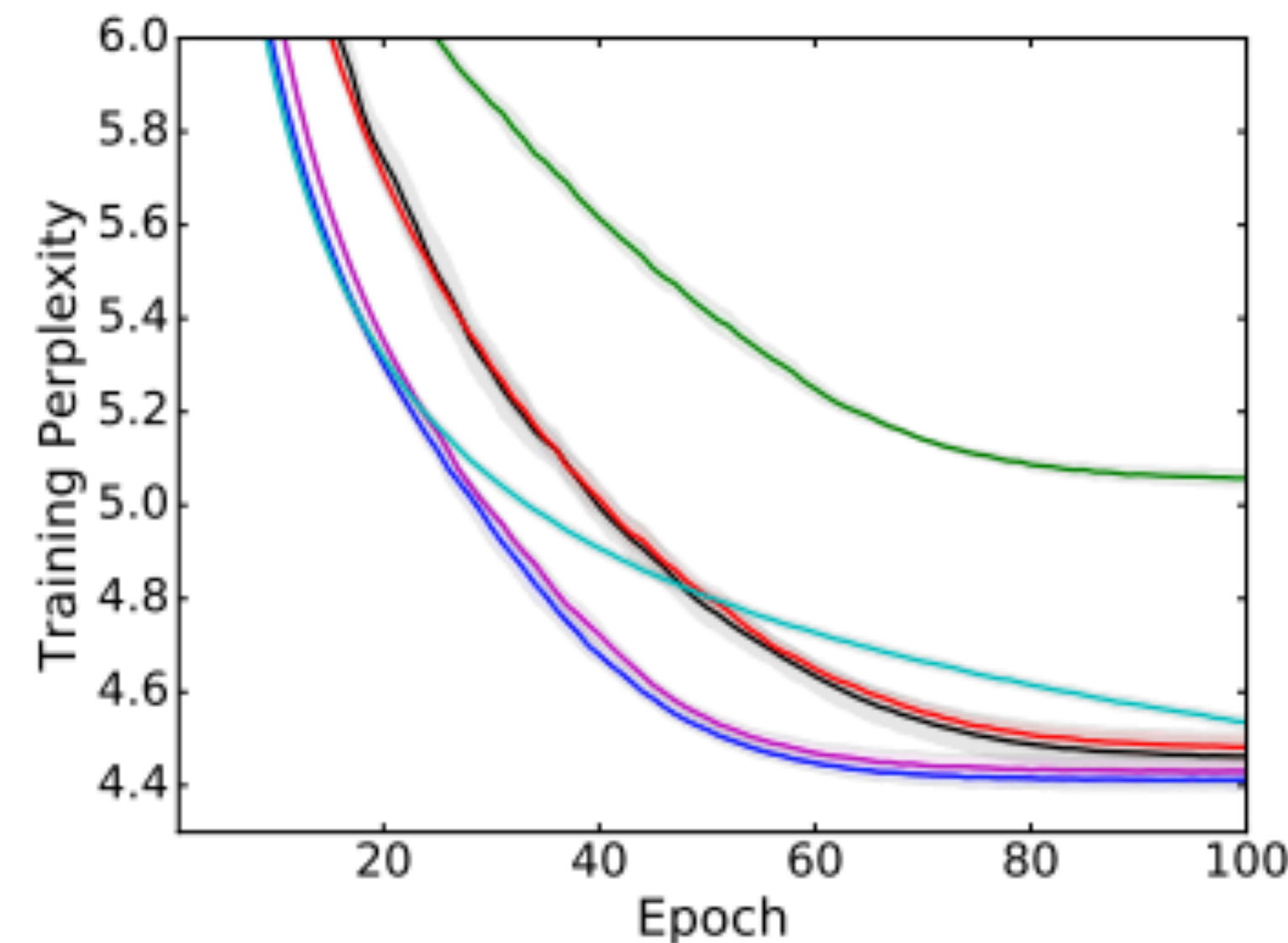
---

- ▶ Parameter initialization is critical to get good gradients, some useful heuristics (e.g., Xavier initializer)
- ▶ Dropout is an effective regularizer

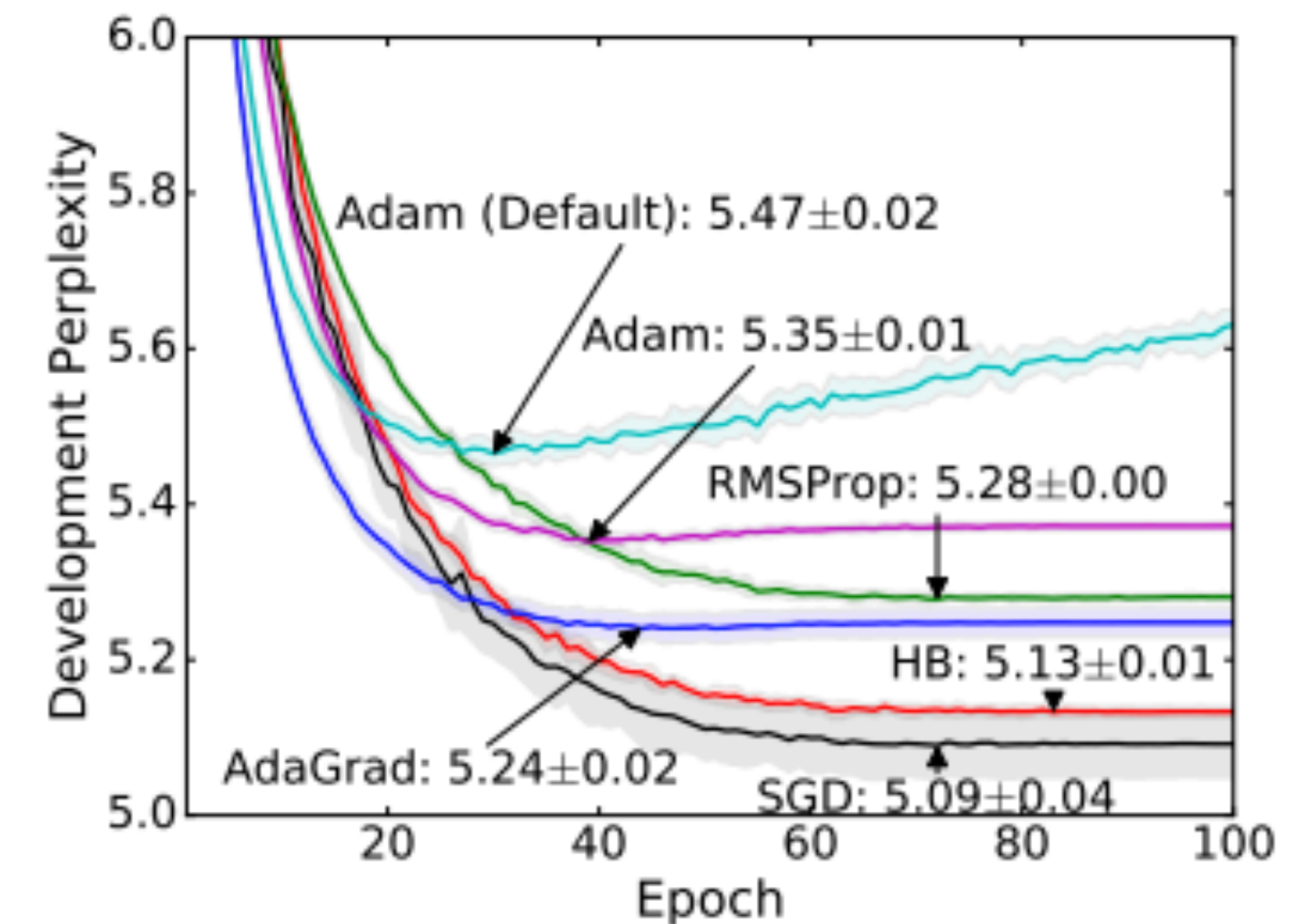
# Recall: Training Tips

- ▶ Parameter initialization is critical to get good gradients, some useful heuristics (e.g., Xavier initializer)
- ▶ Dropout is an effective regularizer

- ▶ Think about your optimizer: Adam or tuned SGD work well



(e) Generative Parsing (Training Set)



(f) Generative Parsing (Development Set)

# Recall: Word Vectors

◆ *the president said that the downturn was over* ◆

<i>president</i>	<i>the __ of</i>
<i>president</i>	<i>the __ said</i> ←
<i>governor</i>	<i>the __ of</i>
<i>governor</i>	<i>the __ appointed</i>
<i>said</i>	<i>sources __</i> ◆
<i>said</i>	<i>president __ that</i>
<i>reported</i>	<i>sources __</i> ◆

*president*  
*governor*

*said*  
*reported*

*the*  
*a*

[Finch and Chater 92, Shuetze 93, many others]



# Recall: Word Vectors

♦ *the president said that the downturn was over* ♦

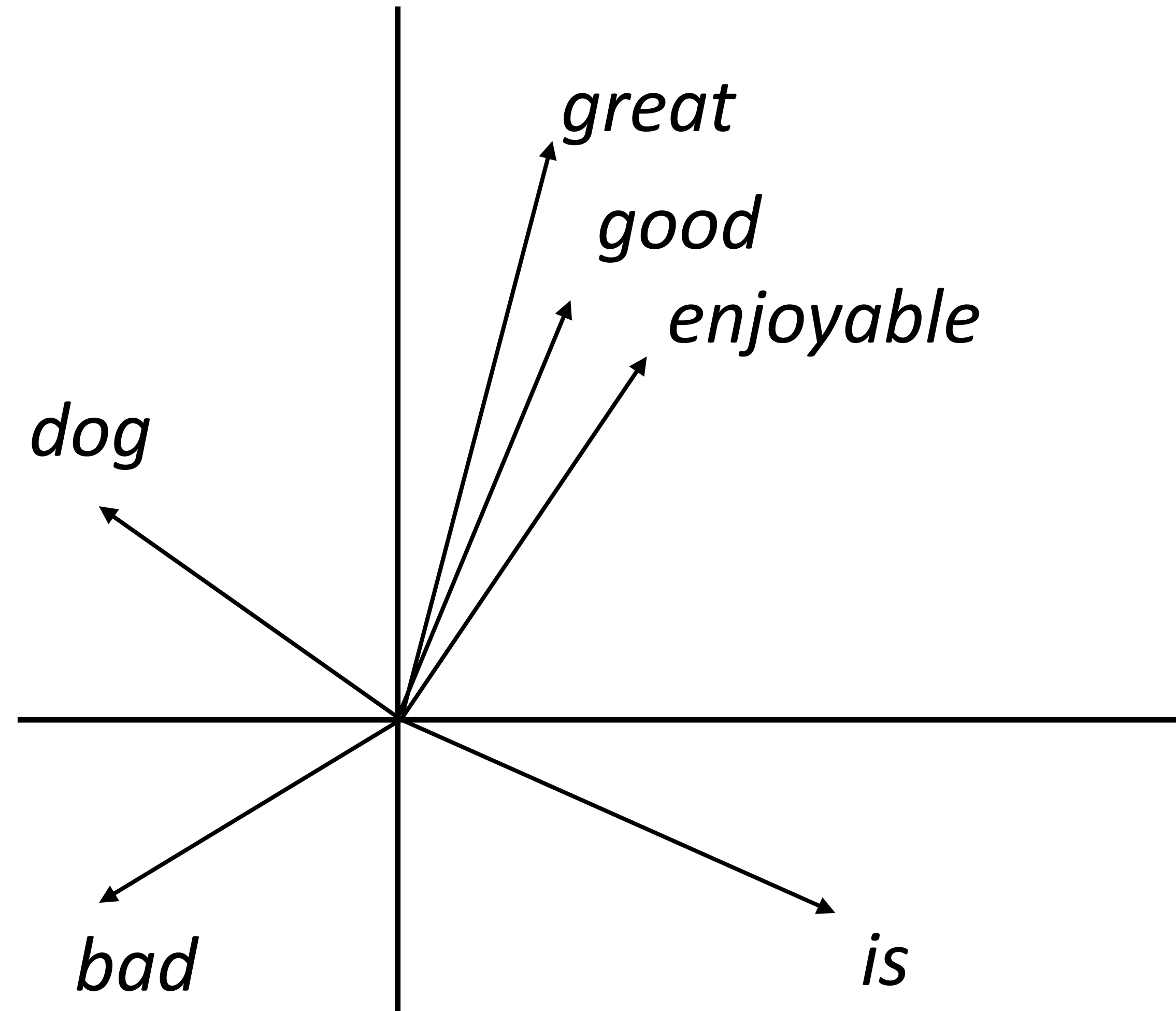
<i>president</i>	<i>the __ of</i>
<i>president</i>	<i>the __ said</i> ←
<i>governor</i>	<i>the __ of</i>
<i>governor</i>	<i>the __ appointed</i>
<i>said</i>	<i>sources __ ♦</i>
<i>said</i>	<i>president __ that</i>
<i>reported</i>	<i>sources __ ♦</i>

*president*  
*governor*

*said*  
*reported*

*the*  
*a*

[Finch and Chater 92, Shuetze 93, many others]



# Recall: Continuous Bag-of-Words

---

- ▶ Predict word from context

*the dog bit the man*



Mikolov et al. (2013)



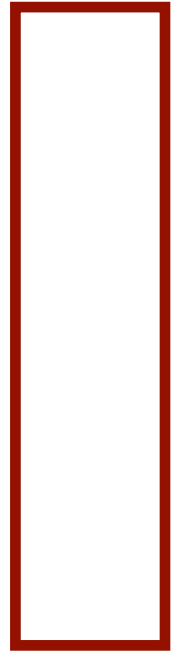
# Recall: Continuous Bag-of-Words

---

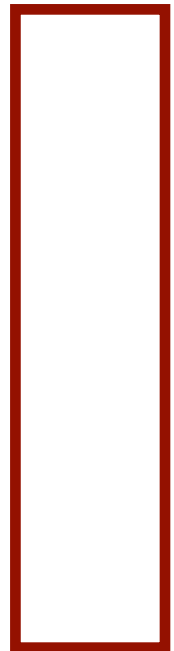
- Predict word from context

*the dog bit the man*

Mikolov et al. (2013)



*dog*



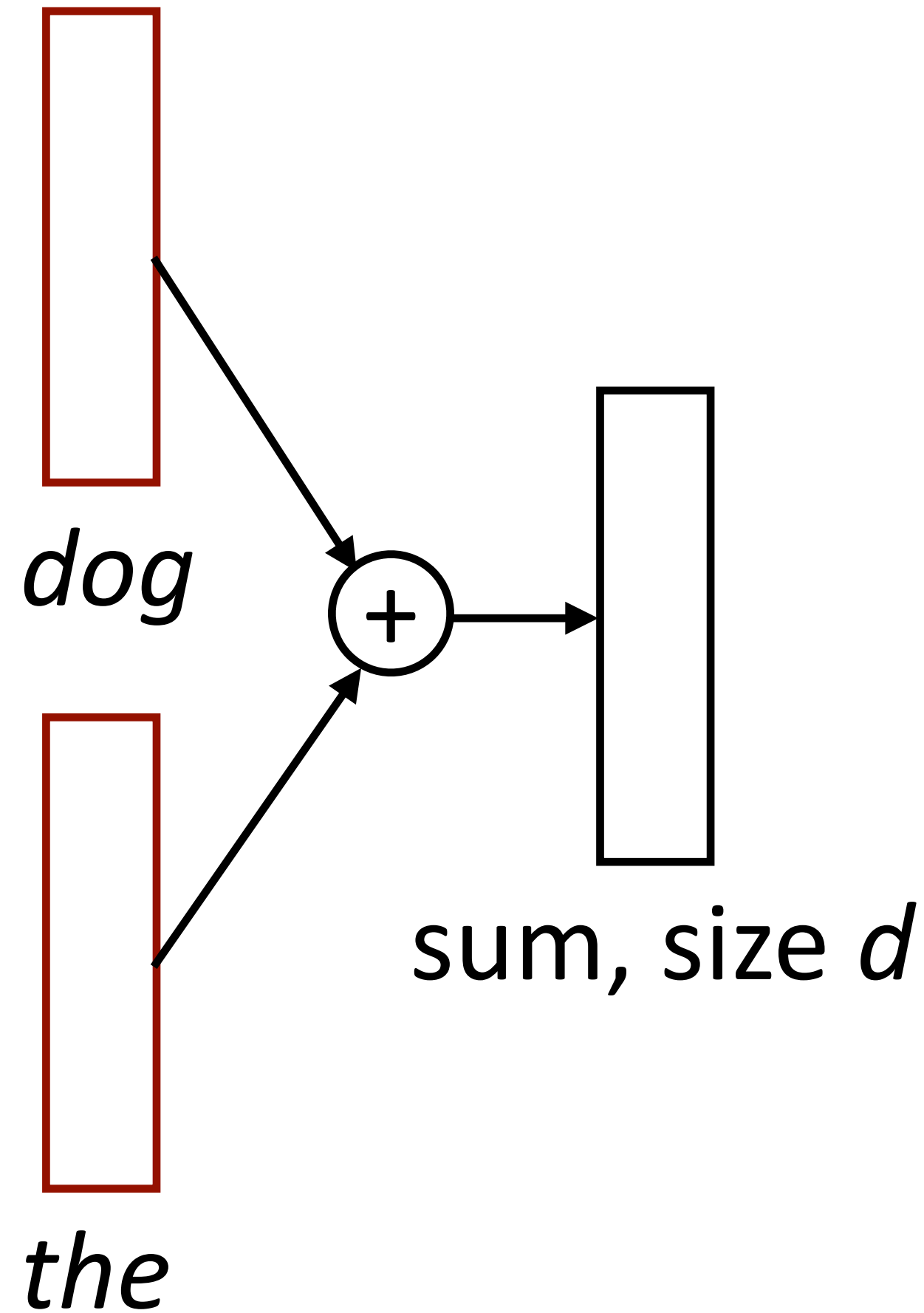
*the*

# Recall: Continuous Bag-of-Words

- Predict word from context

*the dog bit the man*

Mikolov et al. (2013)

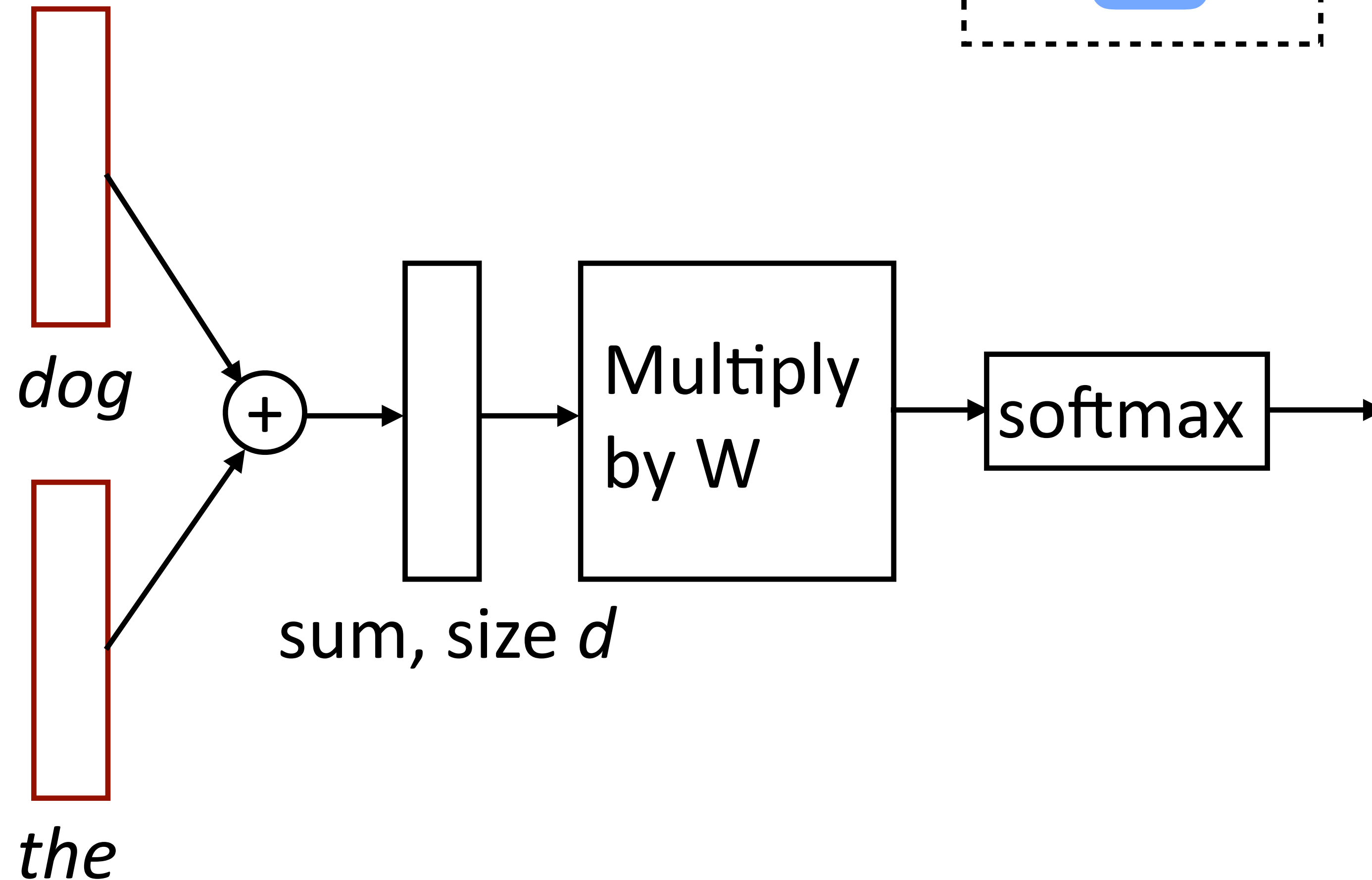


# Recall: Continuous Bag-of-Words

- Predict word from context

*the dog bit the man*

Mikolov et al. (2013)

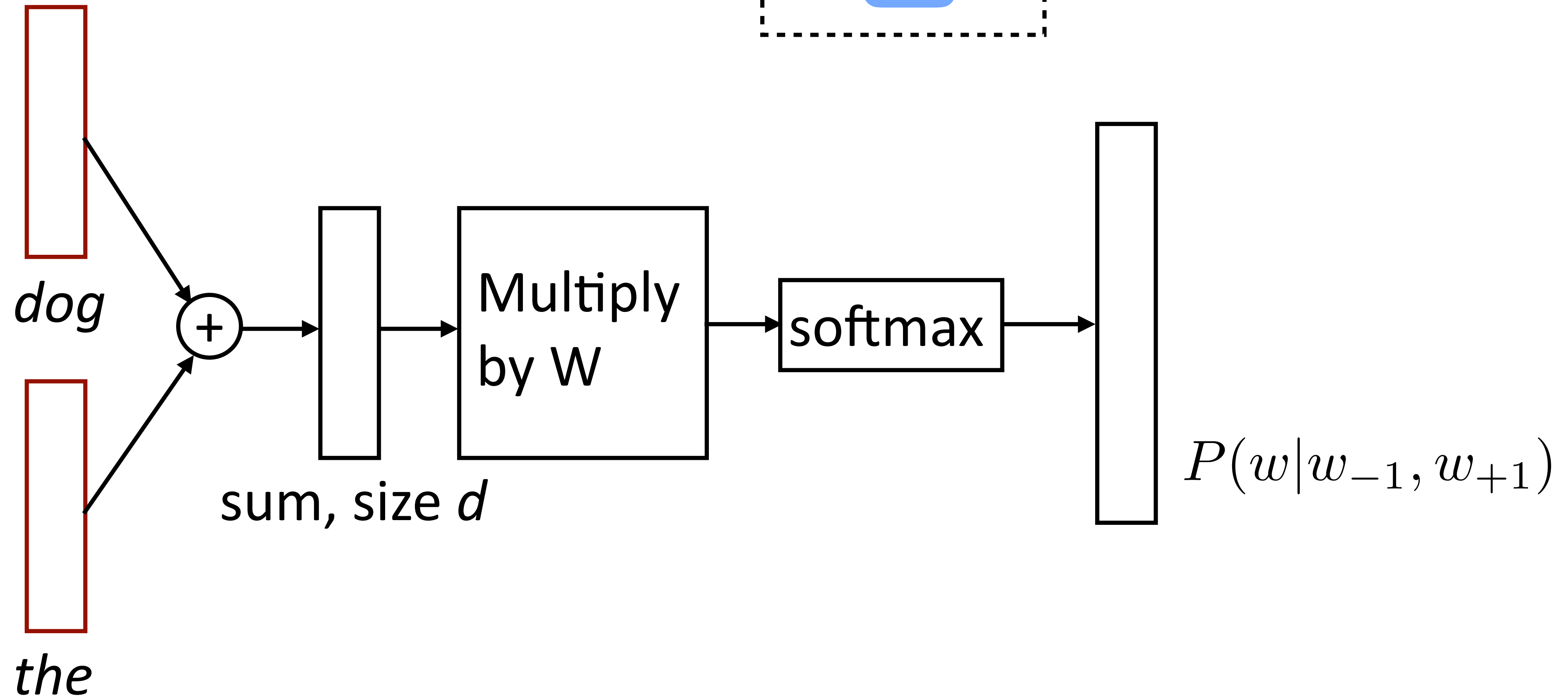


# Recall: Continuous Bag-of-Words

- Predict word from context

*the dog bit the man*

Mikolov et al. (2013)

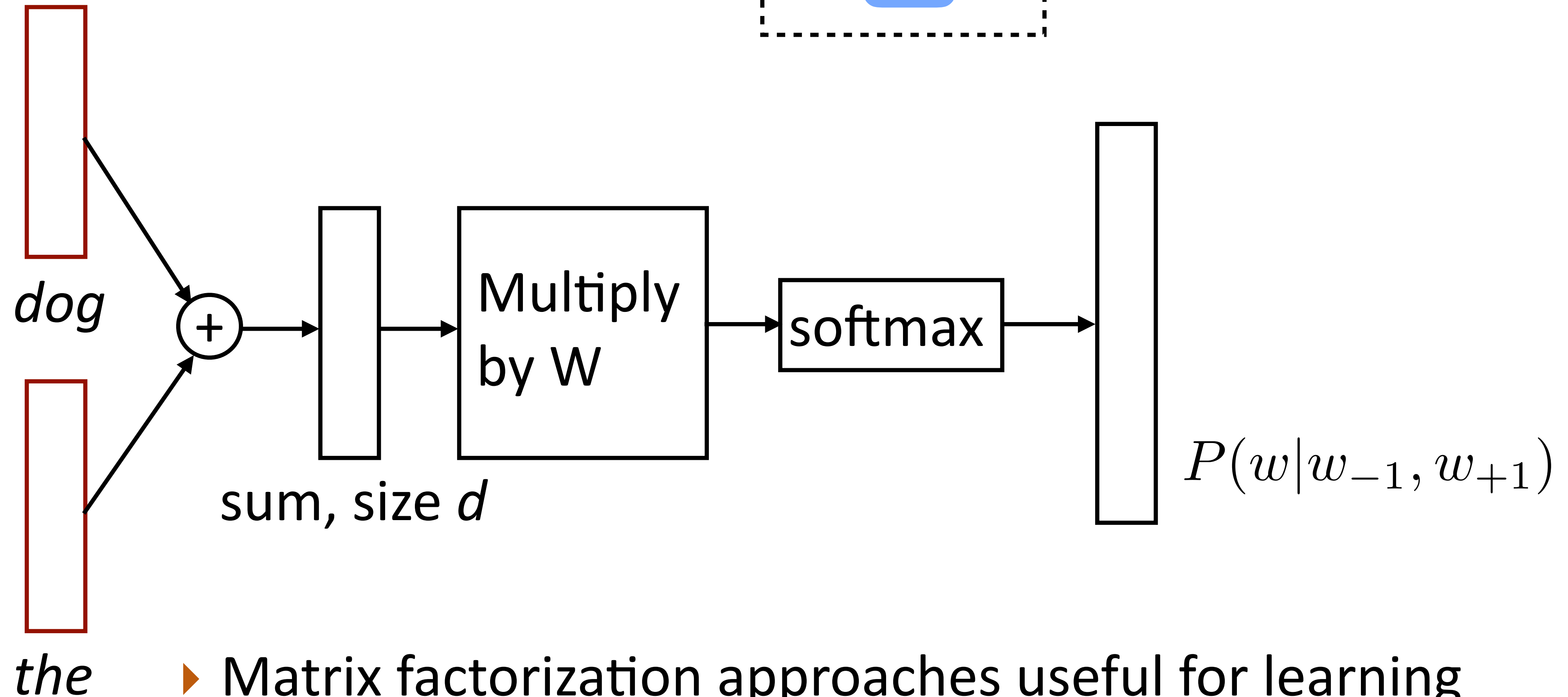


# Recall: Continuous Bag-of-Words

- Predict word from context

*the dog bit the man*

Mikolov et al. (2013)



- Matrix factorization approaches useful for learning vectors from really large data

# Using Word Embeddings

---

# Using Word Embeddings

---

- ▶ Approach 1: learn embeddings directly from data in your neural model, no pretraining
  - ▶ Often works pretty well



# Using Word Embeddings

---

- ▶ Approach 1: learn embeddings directly from data in your neural model, no pretraining
  - ▶ Often works pretty well
- ▶ Approach 2: pretrain using GloVe, keep fixed
  - ▶ Faster because no need to update these parameters
  - ▶ Need to make sure GloVe vocabulary contains all the words you need

# Using Word Embeddings

---

- ▶ Approach 1: learn embeddings directly from data in your neural model, no pretraining
  - ▶ Often works pretty well
- ▶ Approach 2: pretrain using GloVe, keep fixed
  - ▶ Faster because no need to update these parameters
  - ▶ Need to make sure GloVe vocabulary contains all the words you need
- ▶ Approach 3: initialize using GloVe, fine-tune
  - ▶ Not as commonly used anymore

# Compositional Semantics

---

# Compositional Semantics

---

- ▶ What if we want embedding representations for whole sentences?

# Compositional Semantics

---

- ▶ What if we want embedding representations for whole sentences?
- ▶ Skip-*thought* vectors (Kiros et al., 2015), similar to skip-gram generalized to a sentence level (more later)

# Compositional Semantics

---

- ▶ What if we want embedding representations for whole sentences?
- ▶ Skip-*thought* vectors (Kiros et al., 2015), similar to skip-gram generalized to a sentence level (more later)
- ▶ Is there a way we can compose vectors to make sentence representations? Summing? RNNs?

# This Lecture

---

- ▶ Recurrent neural networks
- ▶ Vanishing gradient problem
- ▶ LSTMs / GRUs
- ▶ Applications / visualizations



# RNN Basics

# RNN Motivation

---

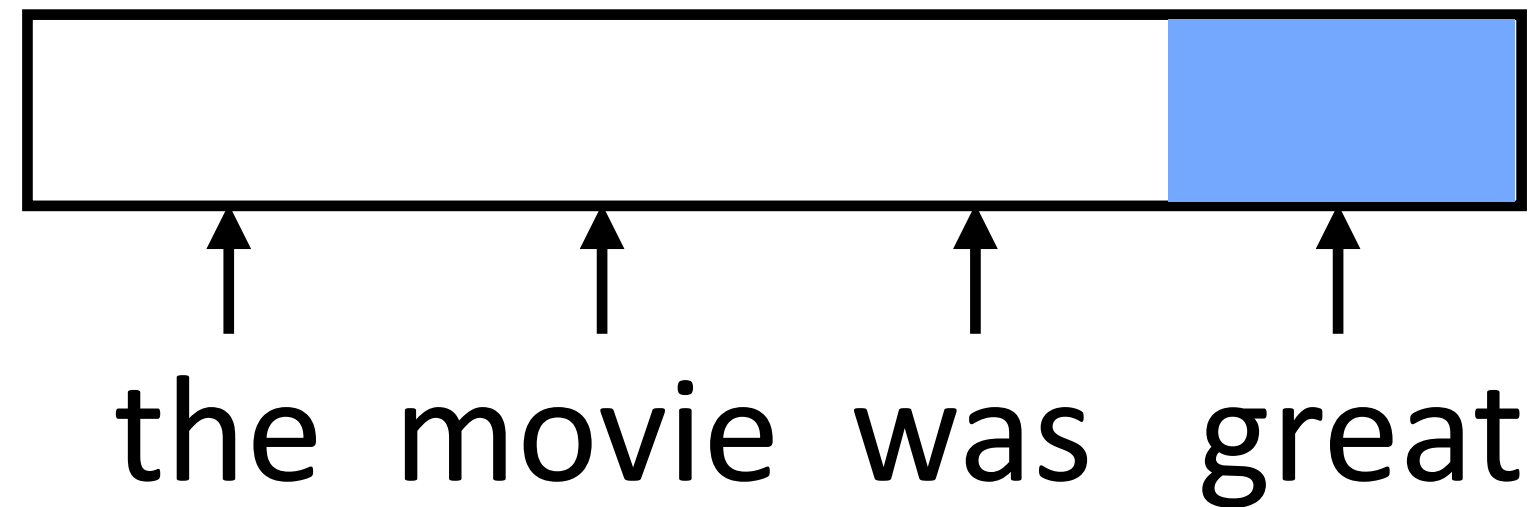
- ▶ Feedforward NNs can't handle variable length input: each position in the feature vector has fixed semantics

the movie was great

# RNN Motivation

---

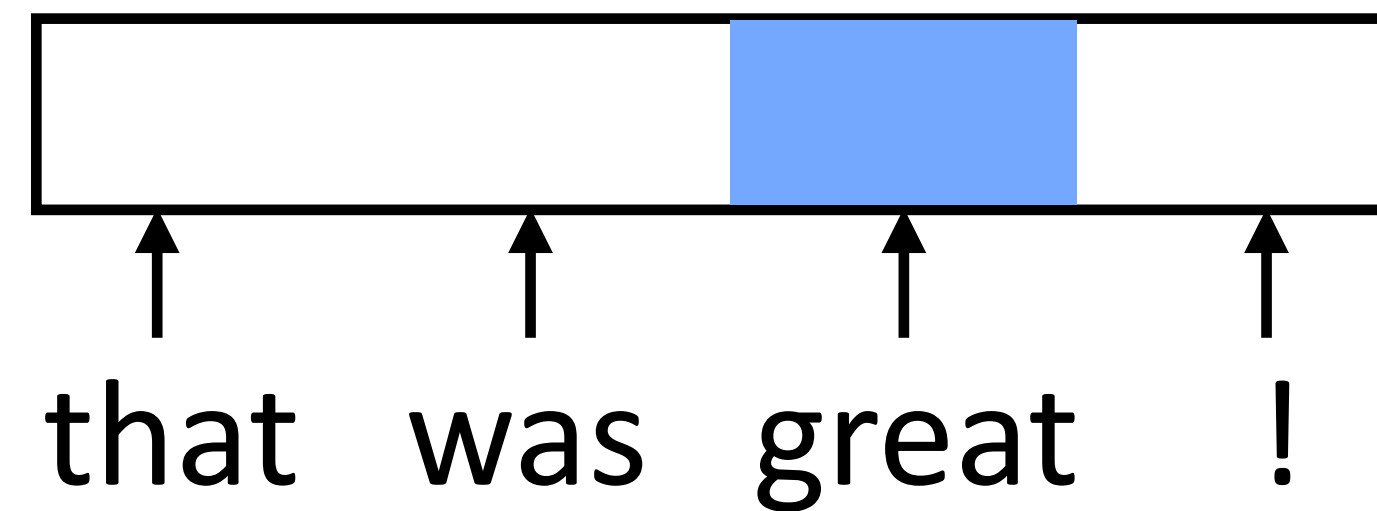
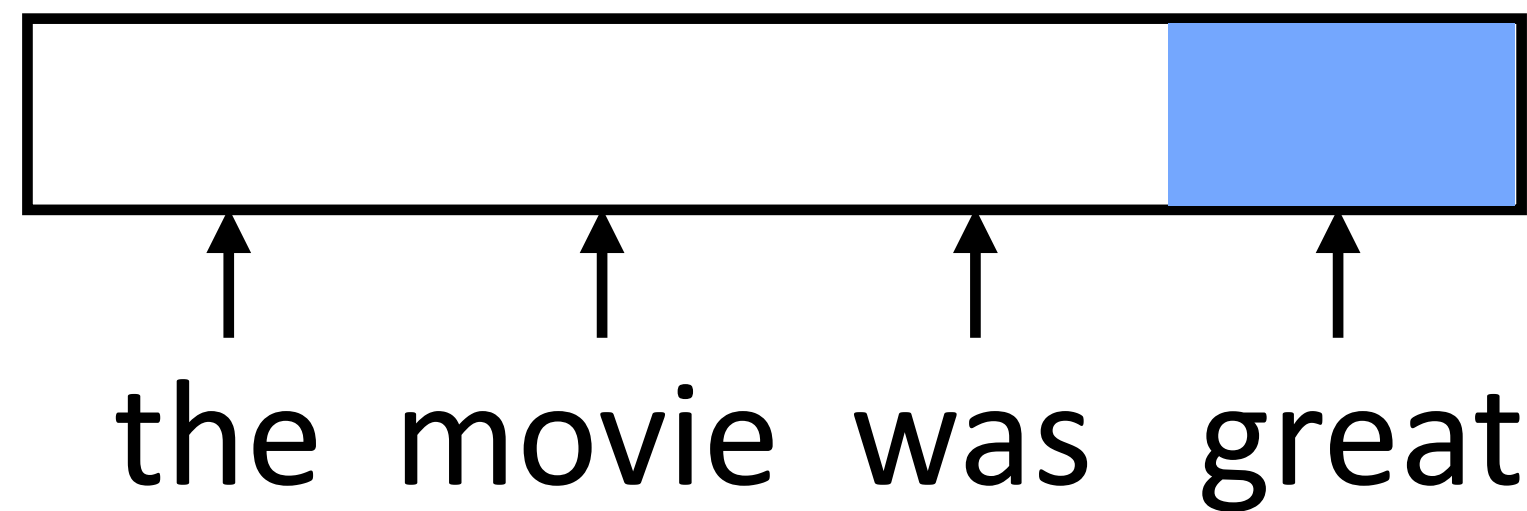
- ▶ Feedforward NNs can't handle variable length input: each position in the feature vector has fixed semantics



# RNN Motivation

---

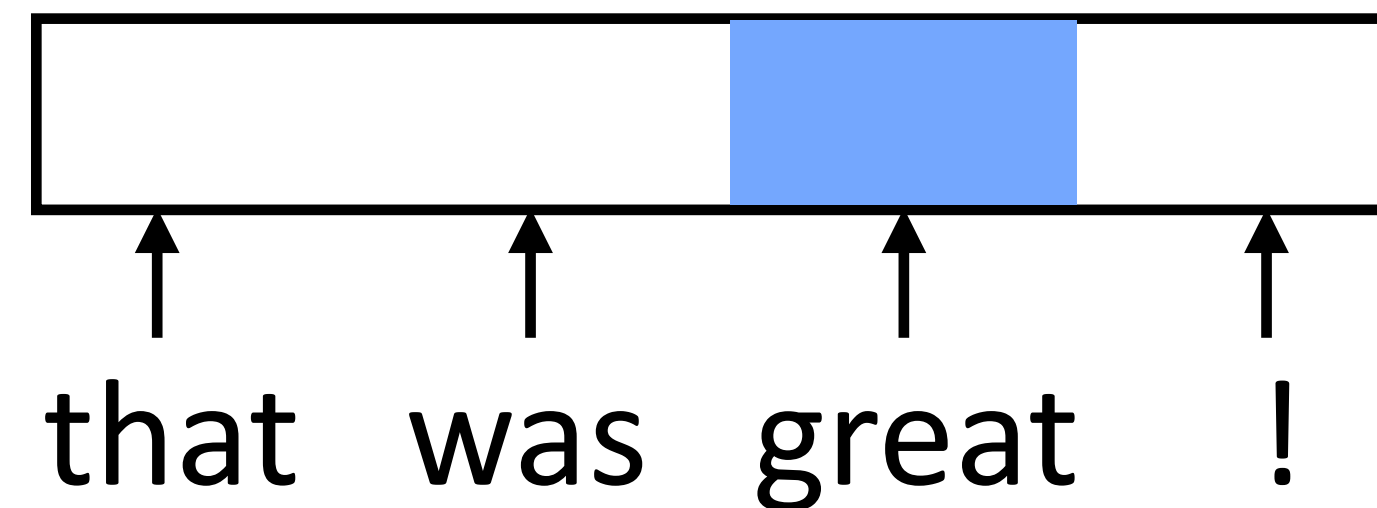
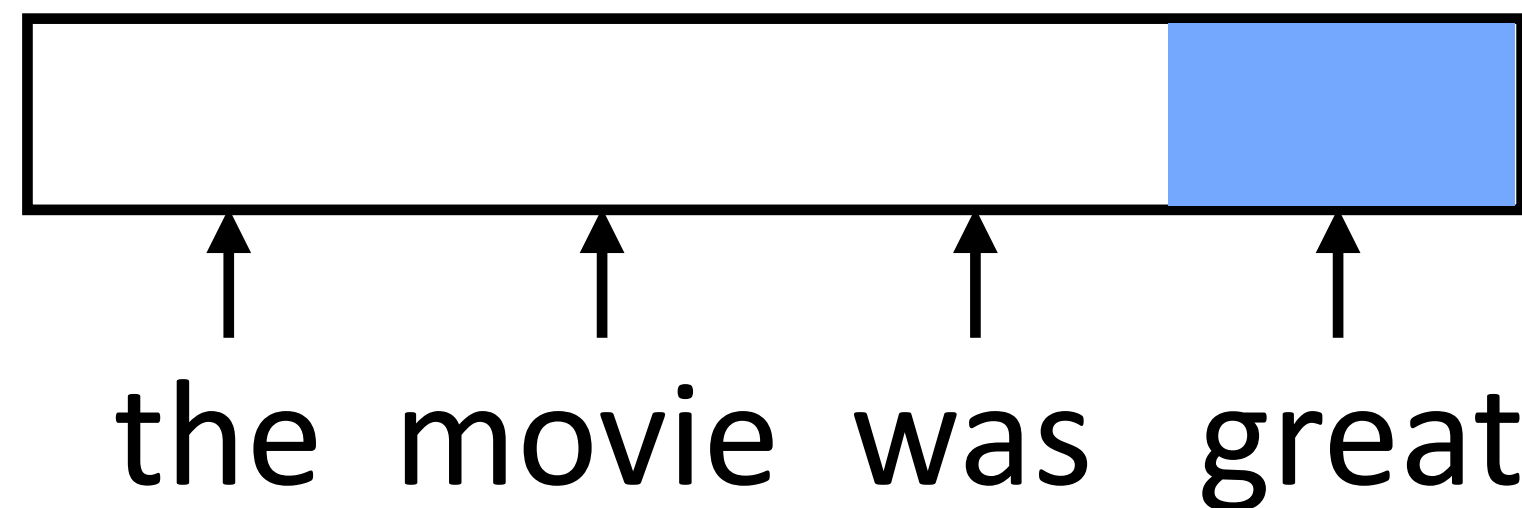
- ▶ Feedforward NNs can't handle variable length input: each position in the feature vector has fixed semantics



# RNN Motivation

---

- ▶ Feedforward NNs can't handle variable length input: each position in the feature vector has fixed semantics

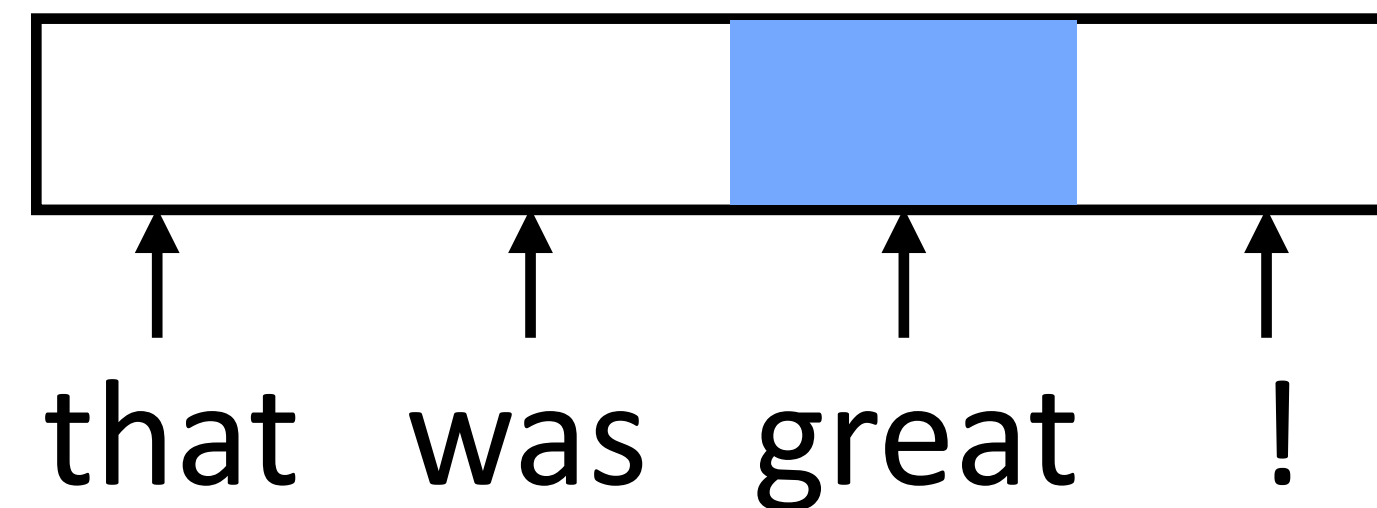
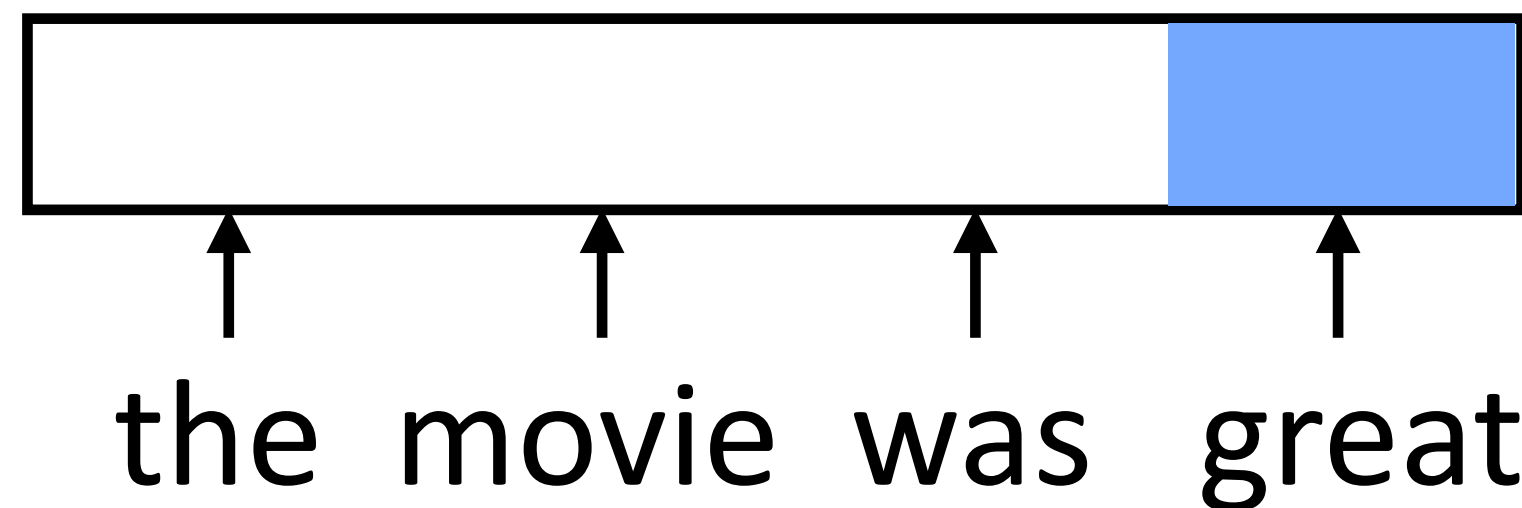


- ▶ These don't look related (*great* is in two different orthogonal subspaces)

# RNN Motivation

---

- ▶ Feedforward NNs can't handle variable length input: each position in the feature vector has fixed semantics

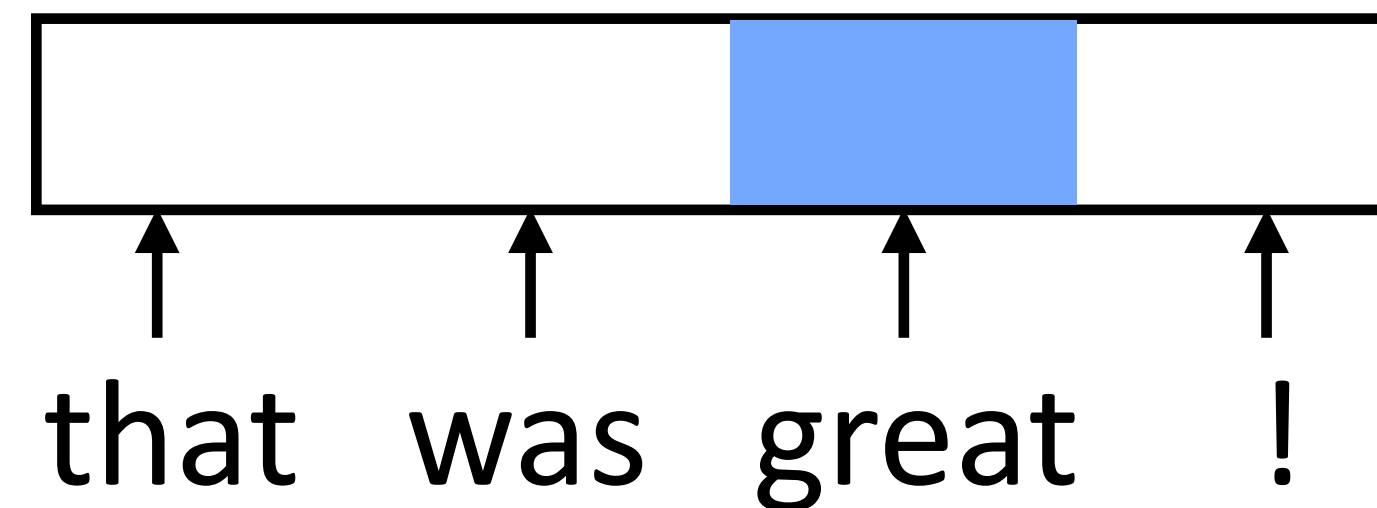
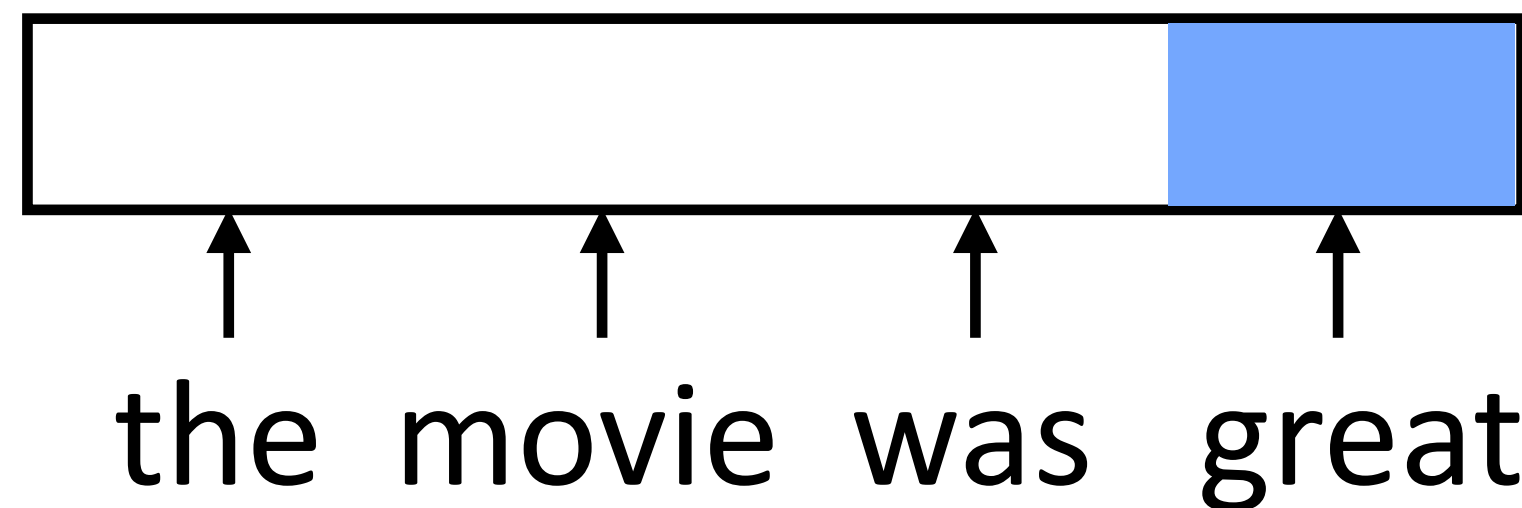


- ▶ These don't look related (*great* is in two different orthogonal subspaces)
- ▶ Instead, we need to:

# RNN Motivation

---

- ▶ Feedforward NNs can't handle variable length input: each position in the feature vector has fixed semantics



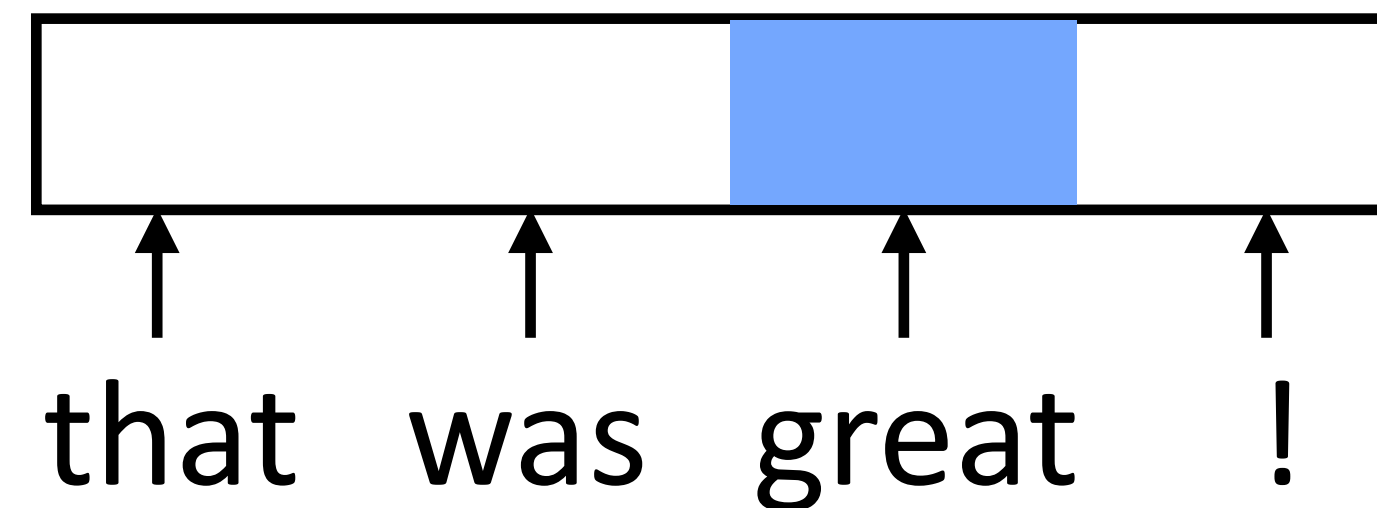
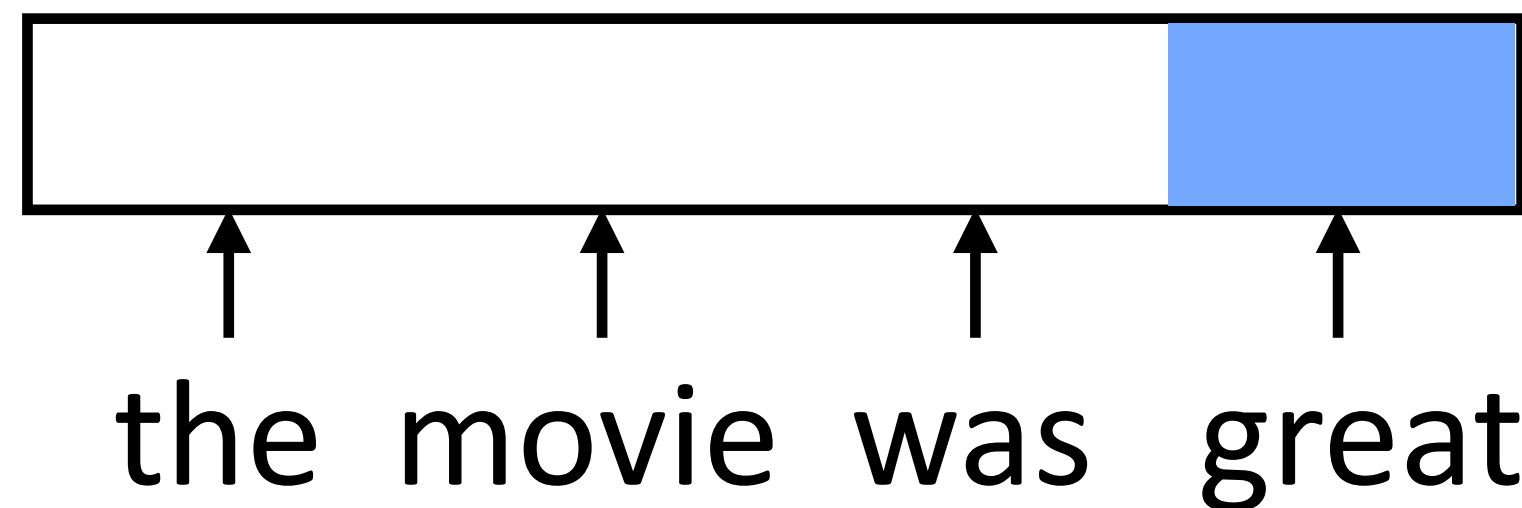
- ▶ These don't look related (*great* is in two different orthogonal subspaces)
- ▶ Instead, we need to:
  - 1) Process each word in a uniform way



# RNN Motivation

---

- ▶ Feedforward NNs can't handle variable length input: each position in the feature vector has fixed semantics

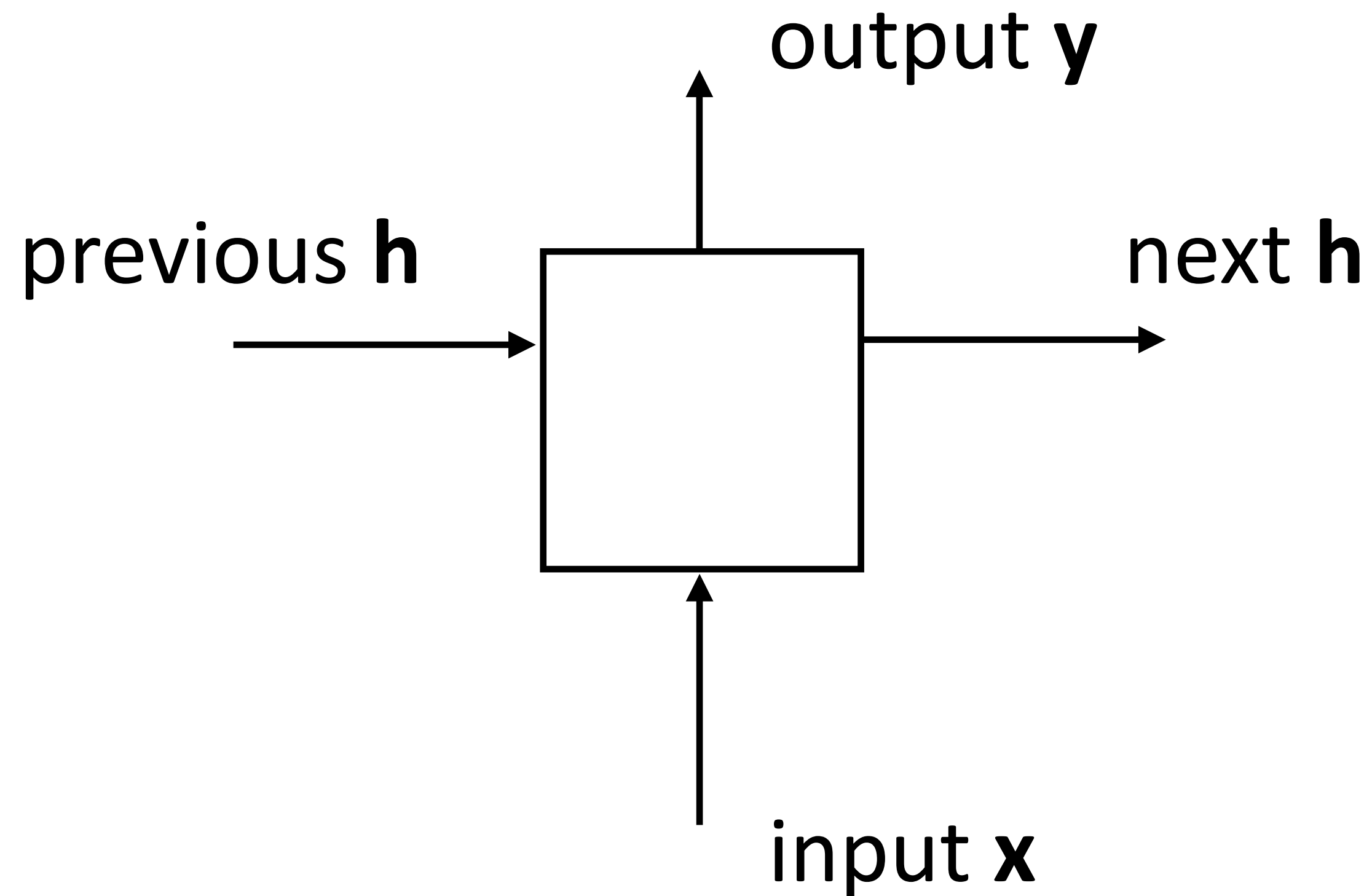


- ▶ These don't look related (*great* is in two different orthogonal subspaces)
- ▶ Instead, we need to:
  - 1) Process each word in a uniform way
  - 2) ...while still exploiting the context that that token occurs in

# RNN Abstraction

---

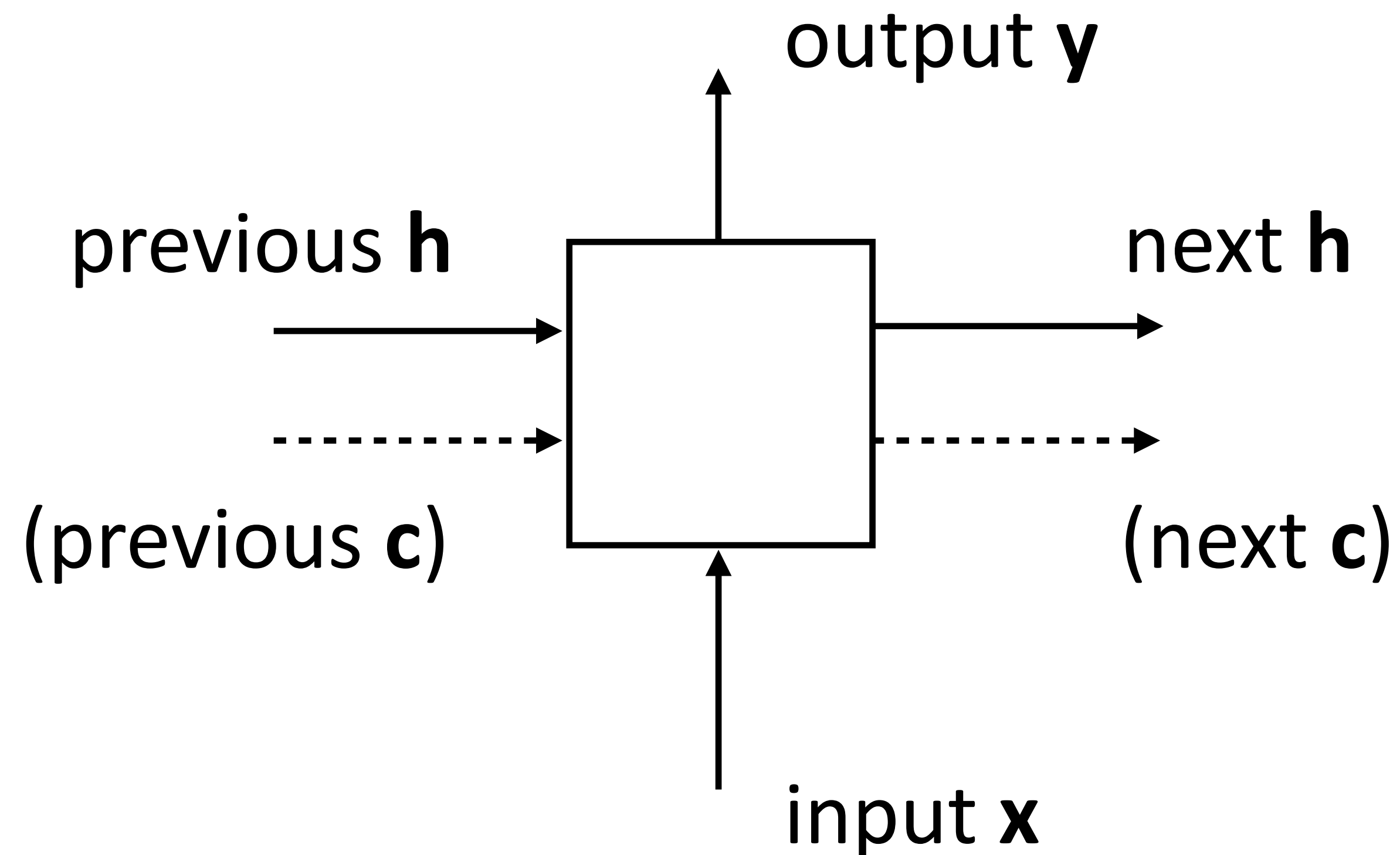
- ▶ Cell that takes some input  $\mathbf{x}$ , has some hidden state  $\mathbf{h}$ , and updates that hidden state and produces output  $\mathbf{y}$  (all vector-valued)



# RNN Abstraction

---

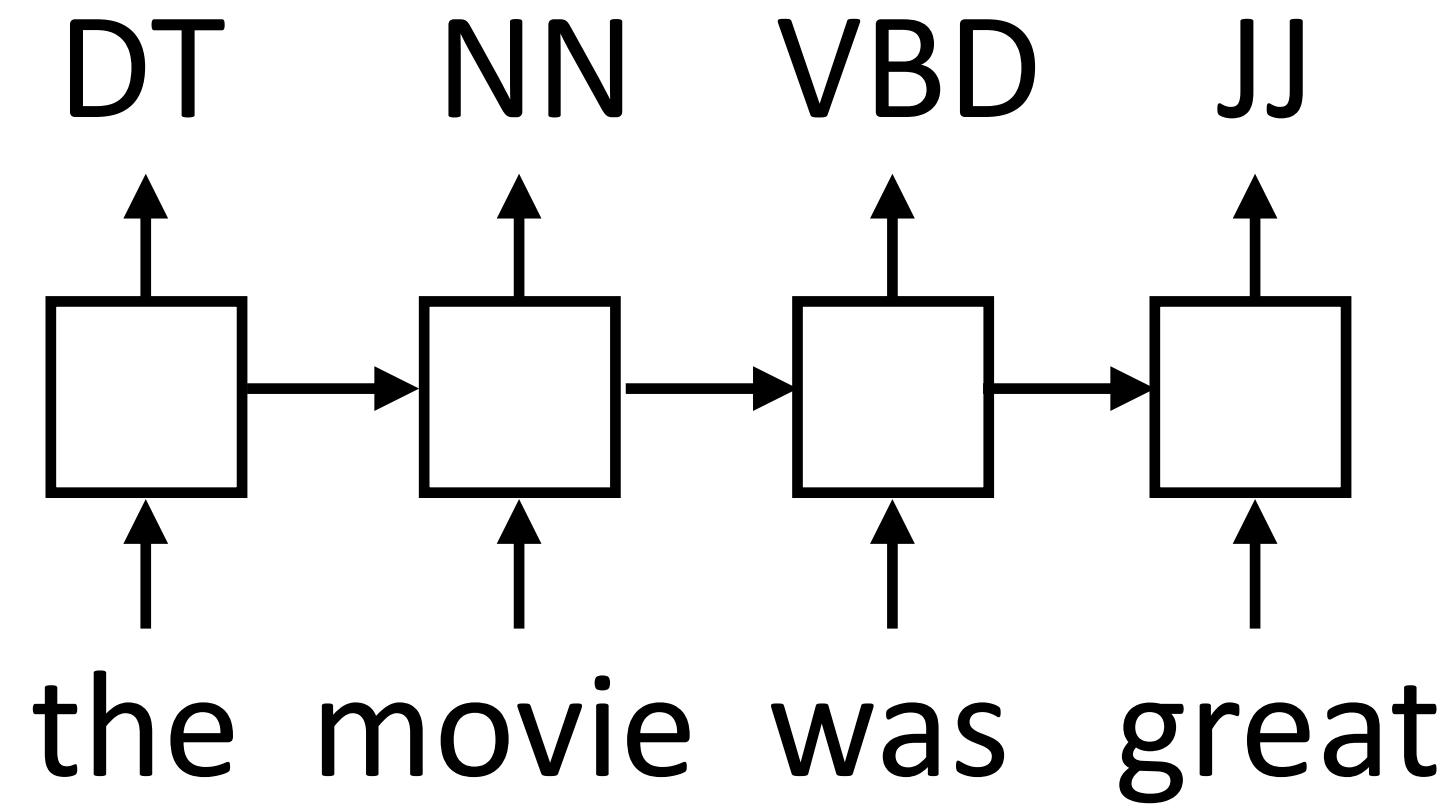
- ▶ Cell that takes some input  $\mathbf{x}$ , has some hidden state  $\mathbf{h}$ , and updates that hidden state and produces output  $\mathbf{y}$  (all vector-valued)



# RNN Uses

---

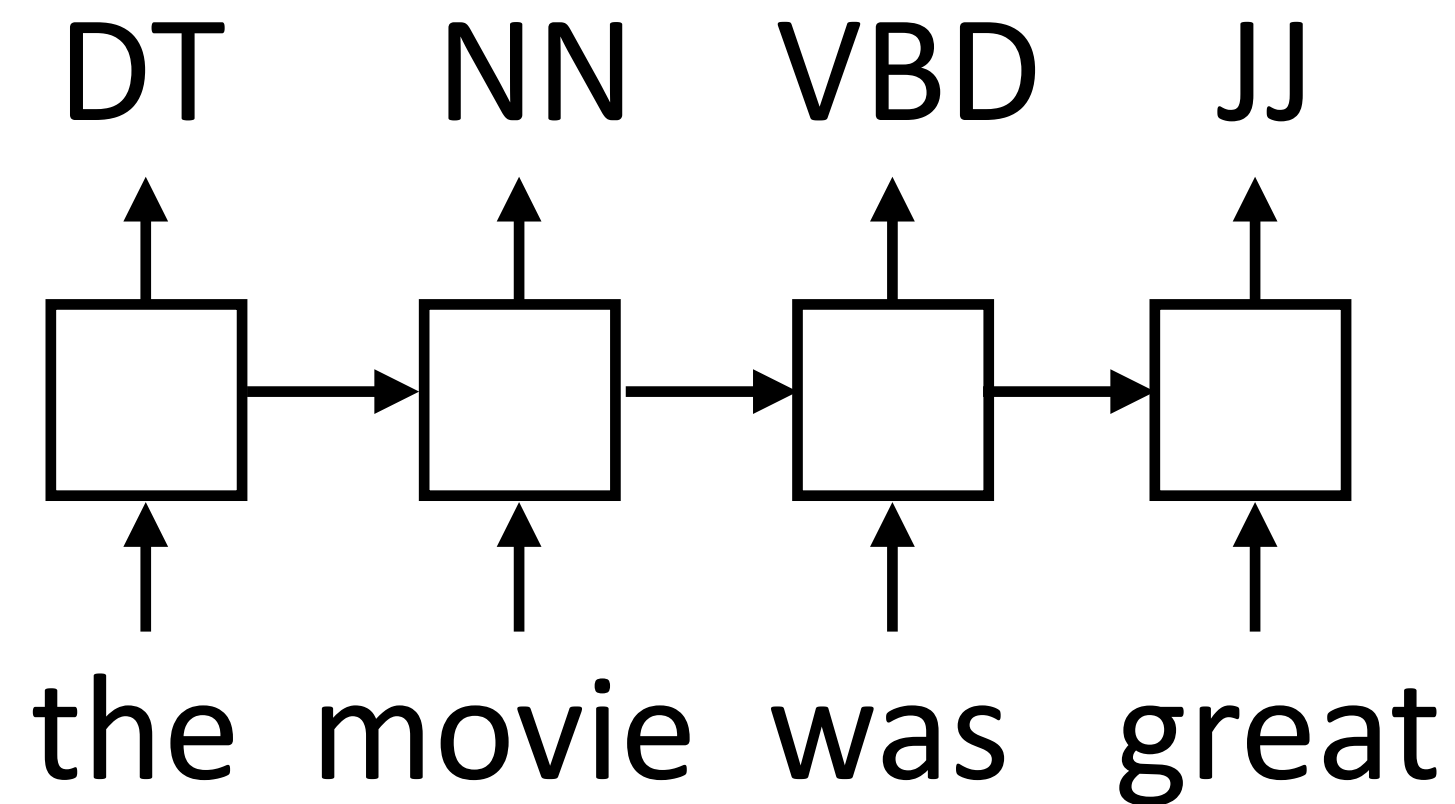
- ▶ Transducer: make some prediction for each element in a sequence



output  $\mathbf{y}$  = score for each tag, then softmax

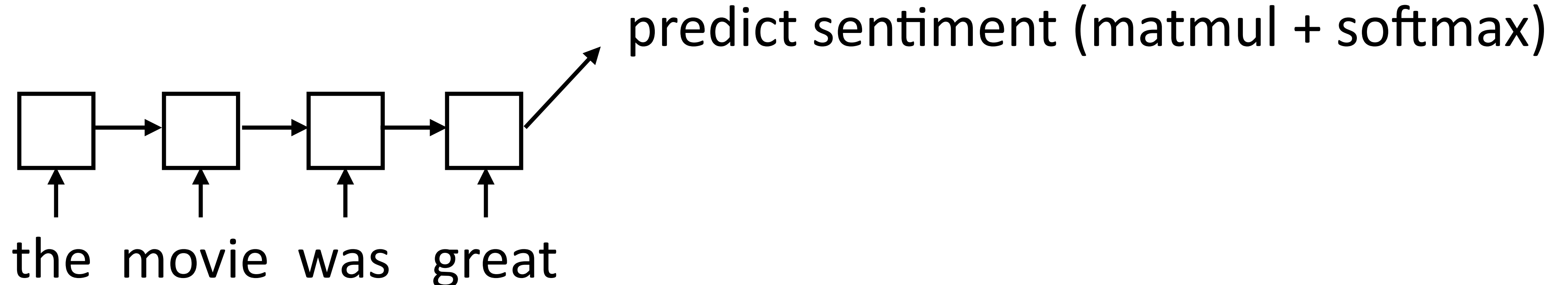
# RNN Uses

- ▶ Transducer: make some prediction for each element in a sequence



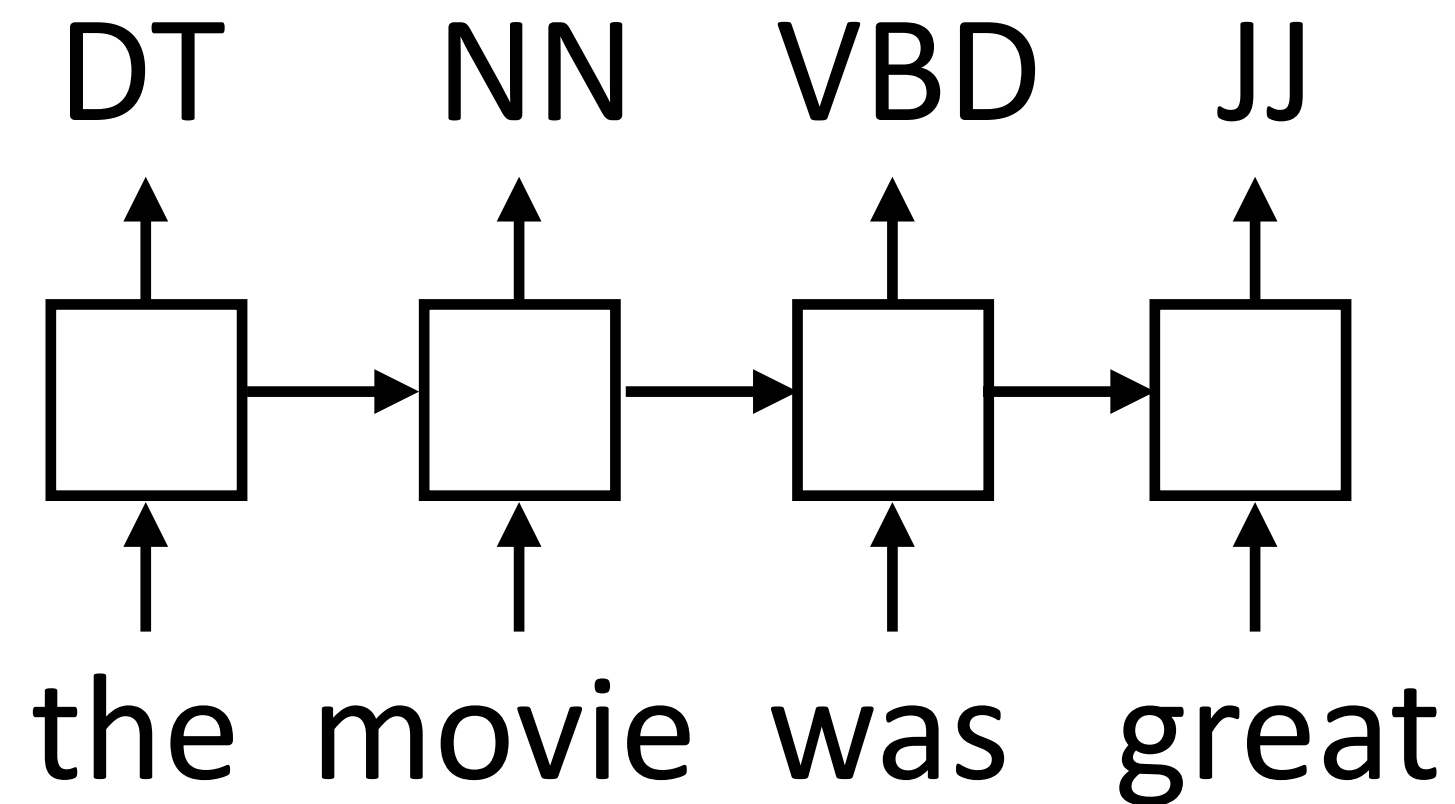
output  $\mathbf{y}$  = score for each tag, then softmax

- ▶ Acceptor/encoder: encode a sequence into a fixed-sized vector and use that for some purpose



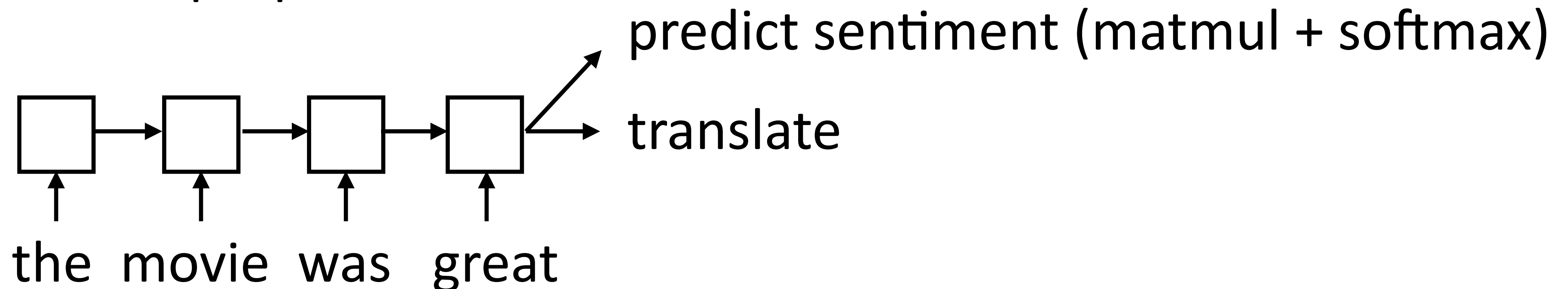
# RNN Uses

- ▶ Transducer: make some prediction for each element in a sequence



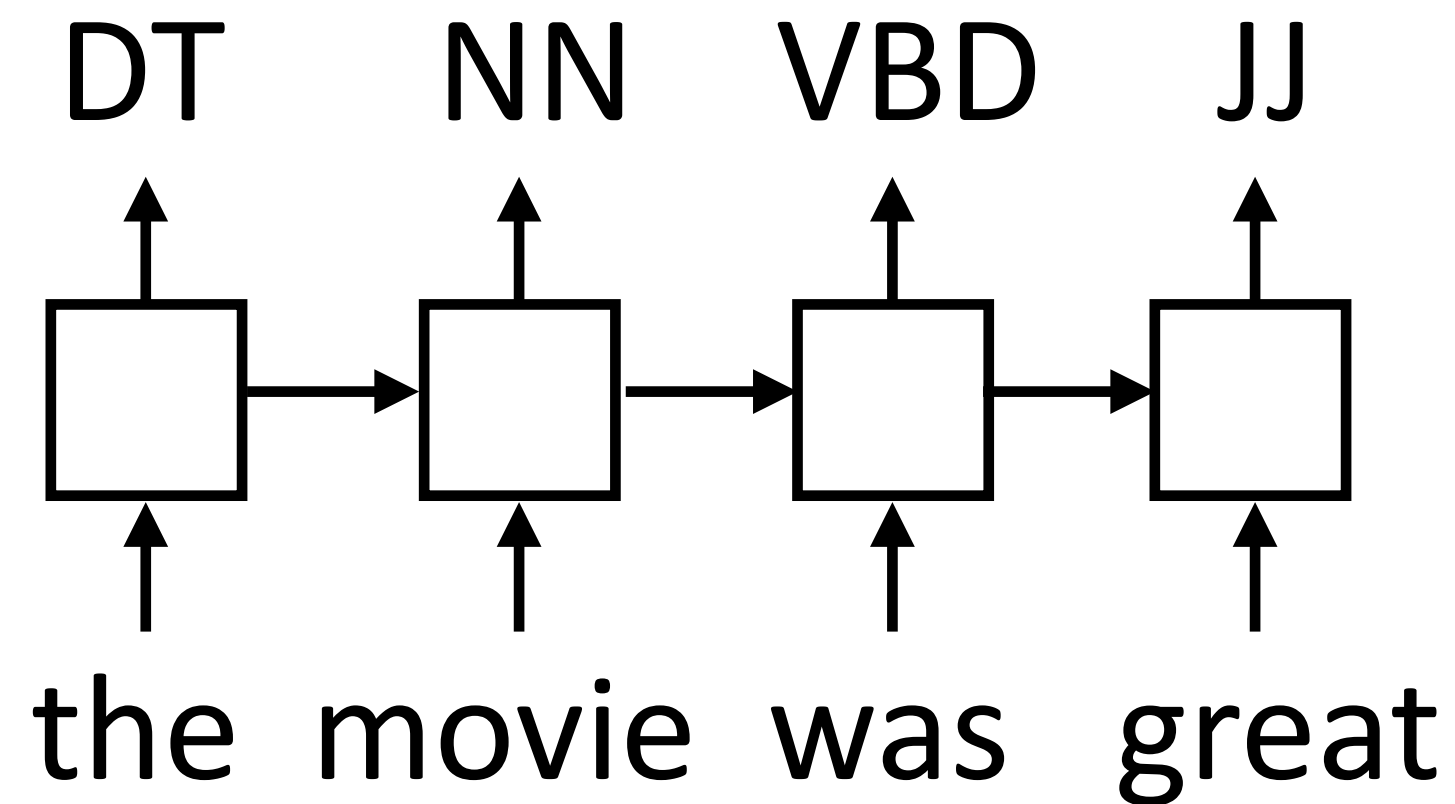
output  $\mathbf{y}$  = score for each tag, then softmax

- ▶ Acceptor/encoder: encode a sequence into a fixed-sized vector and use that for some purpose



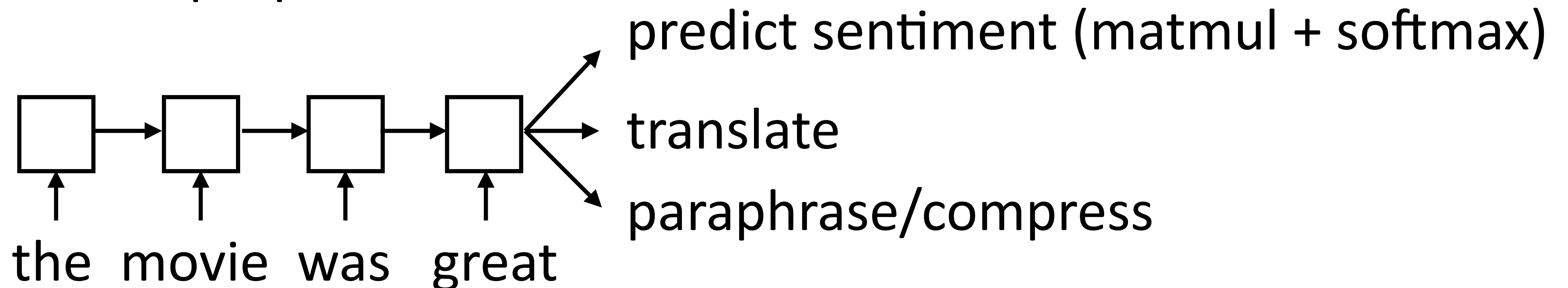
# RNN Uses

- ▶ Transducer: make some prediction for each element in a sequence



output  $\mathbf{y}$  = score for each tag, then softmax

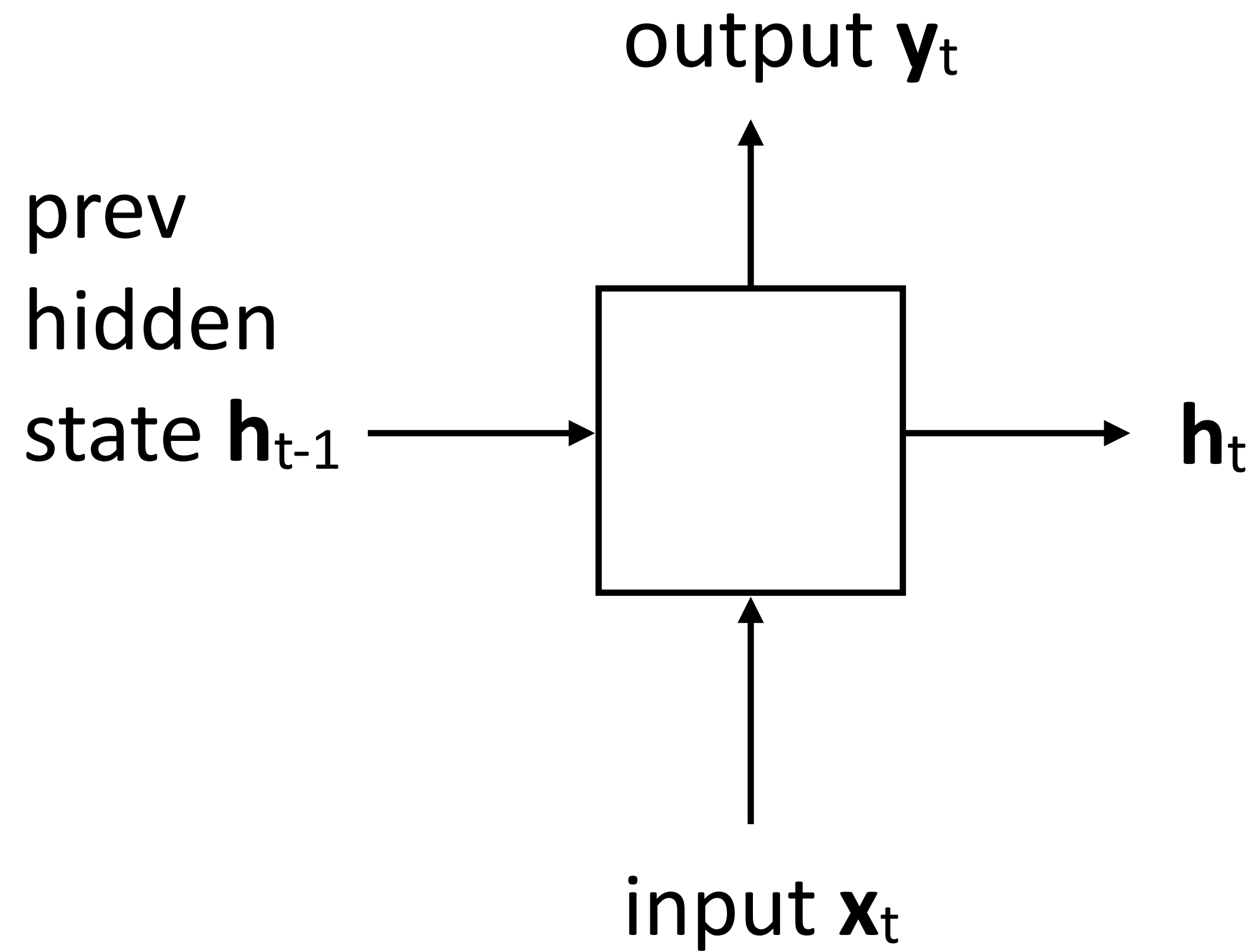
- ▶ Acceptor/encoder: encode a sequence into a fixed-sized vector and use that for some purpose





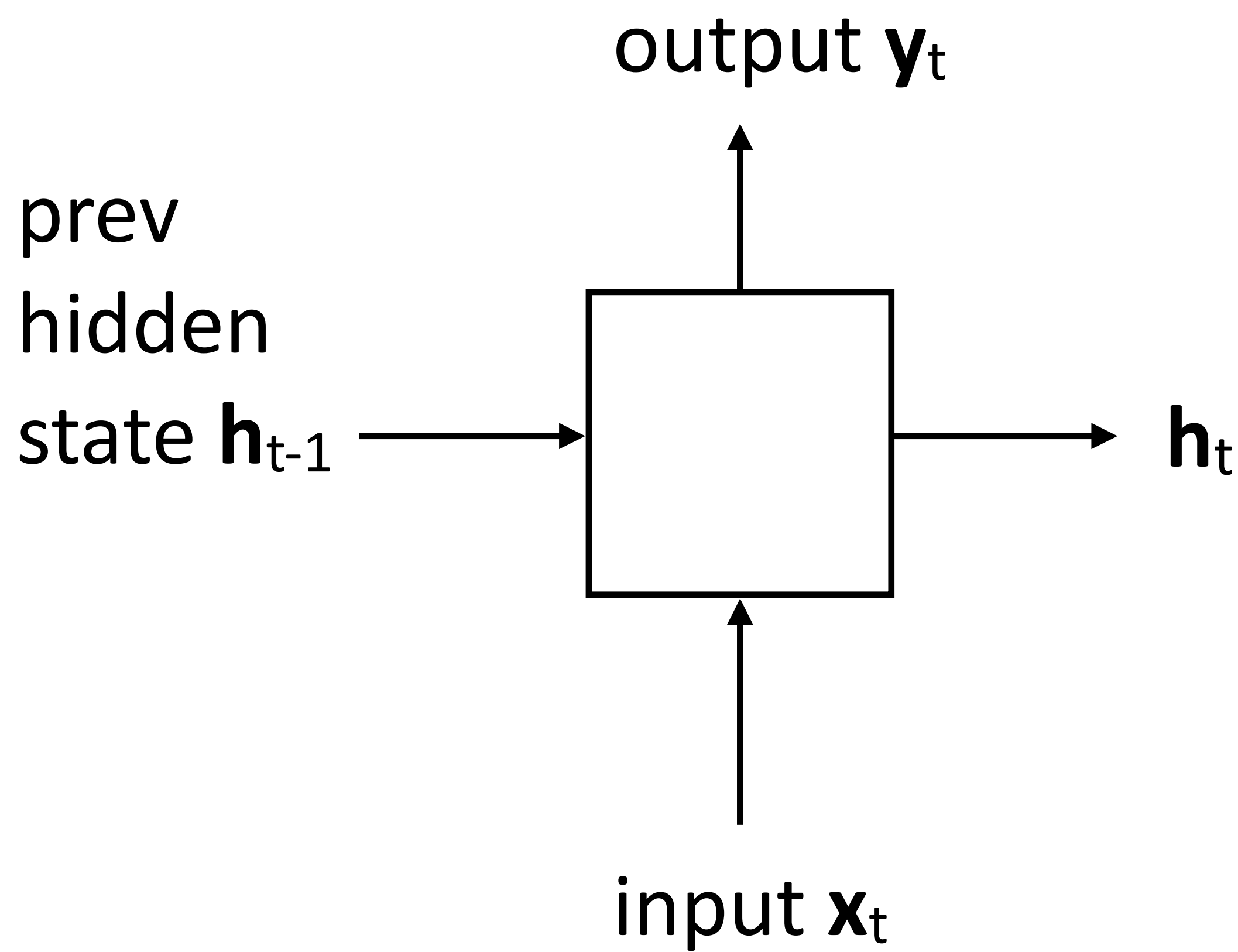
# Elman Networks

---



# Elman Networks

---

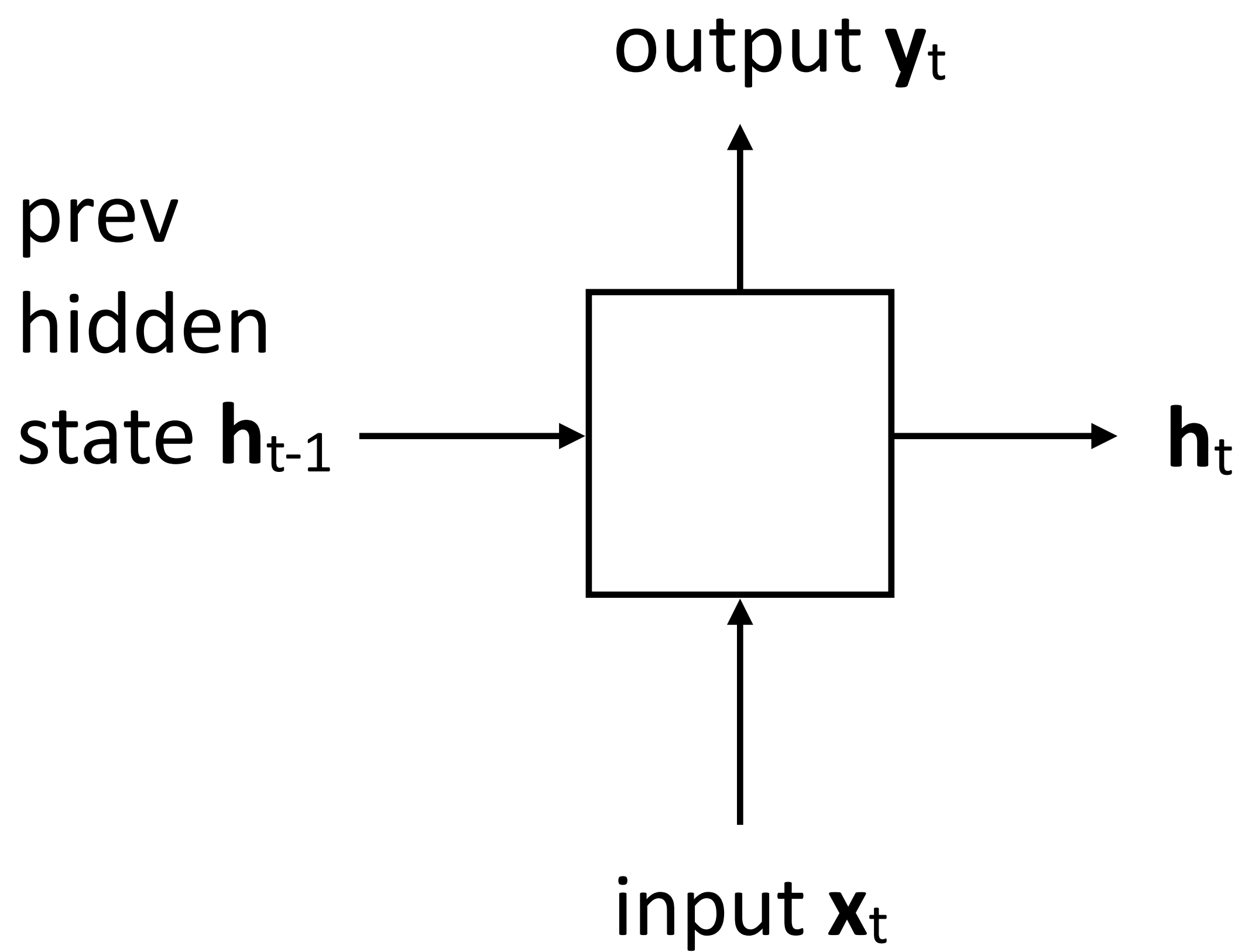


$$\mathbf{h}_t = \tanh(W\mathbf{x}_t + V\mathbf{h}_{t-1} + \mathbf{b}_h)$$

- Updates hidden state based on input and current hidden state

# Elman Networks

---



$$\mathbf{h}_t = \tanh(W\mathbf{x}_t + V\mathbf{h}_{t-1} + \mathbf{b}_h)$$

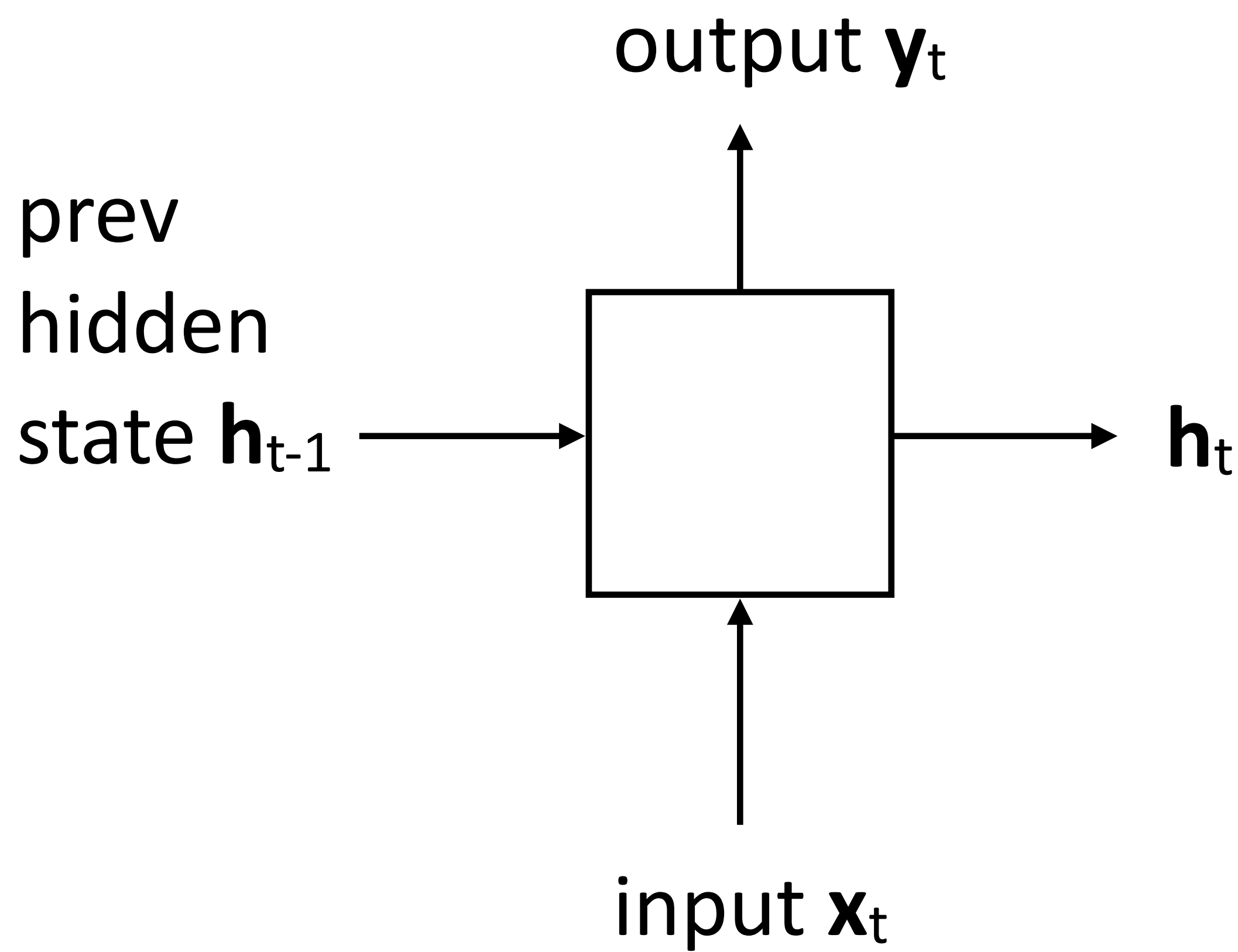
- Updates hidden state based on input and current hidden state

$$\mathbf{y}_t = \tanh(U\mathbf{h}_t + \mathbf{b}_y)$$

- Computes output from hidden state

# Elman Networks

---



$$\mathbf{h}_t = \tanh(W\mathbf{x}_t + V\mathbf{h}_{t-1} + \mathbf{b}_h)$$

- Updates hidden state based on input and current hidden state

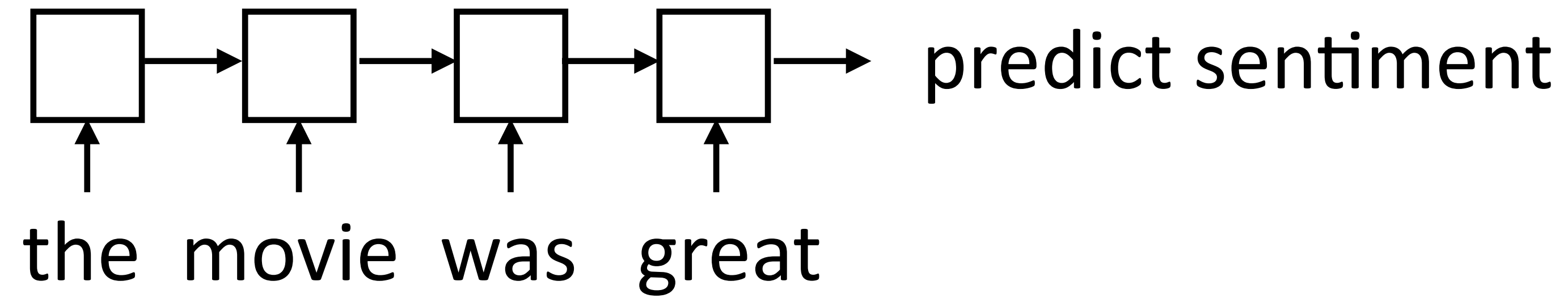
$$\mathbf{y}_t = \tanh(U\mathbf{h}_t + \mathbf{b}_y)$$

- Computes output from hidden state

- Long history! (invented in the late 1980s)

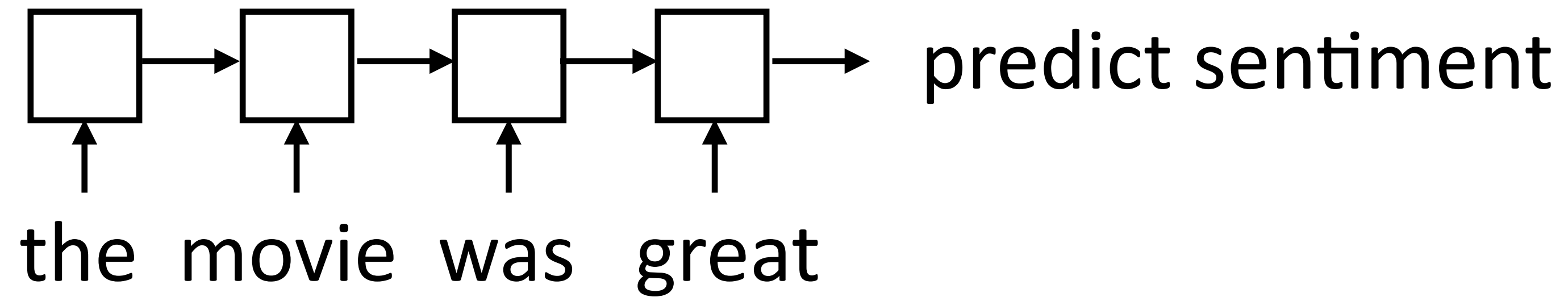
# Training Elman Networks

---



# Training Elman Networks

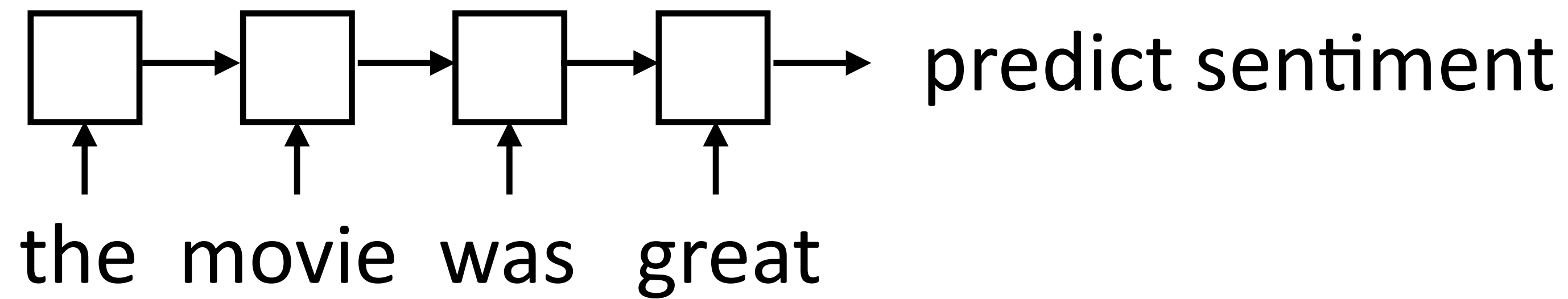
---



- ▶ “Backpropagation through time”: build the network as one big computation graph, some parameters are shared

# Training Elman Networks

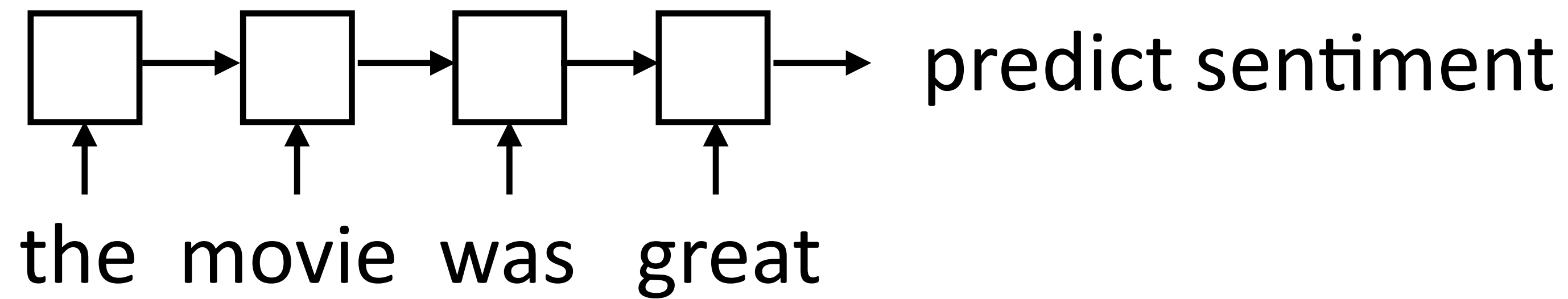
---



- ▶ “Backpropagation through time”: build the network as one big computation graph, some parameters are shared
  - ▶ RNN potentially needs to learn how to “remember” information for a long time!
- it was my **favorite** movie of 2016, though it wasn't without **problems** -> **+**

# Training Elman Networks

---

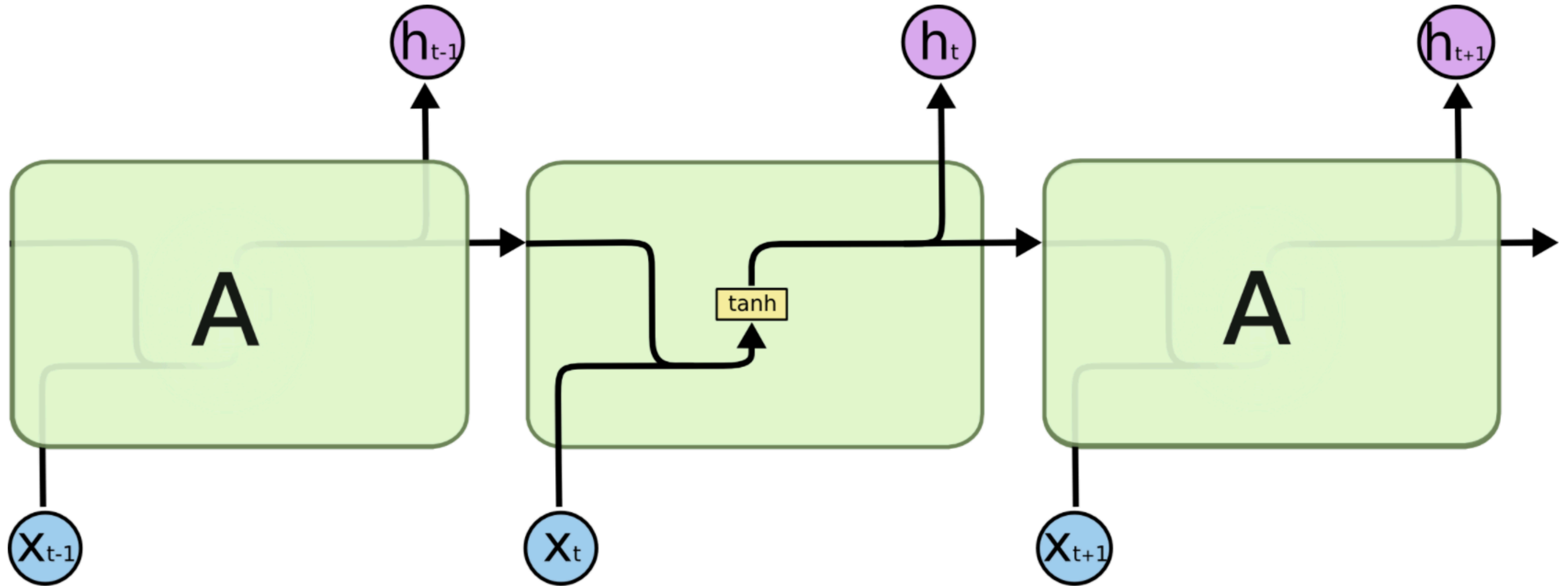


- ▶ “Backpropagation through time”: build the network as one big computation graph, some parameters are shared
  - ▶ RNN potentially needs to learn how to “remember” information for a long time!
- it was my **favorite** movie of 2016, though it wasn't without **problems** -> +
- ▶ “Correct” parameter update is to do a better job of remembering the sentiment of *favorite*

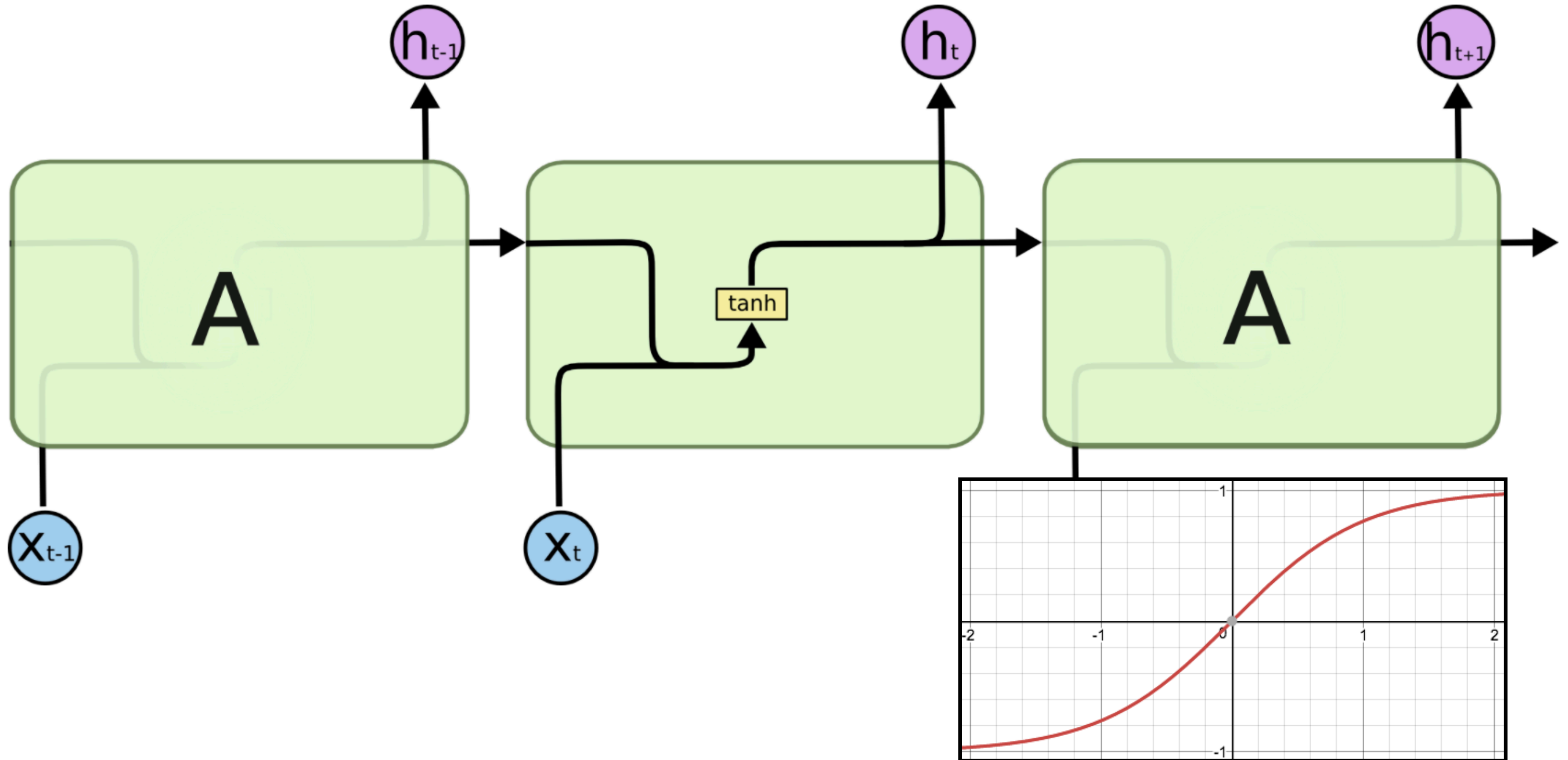


# Vanishing Gradient

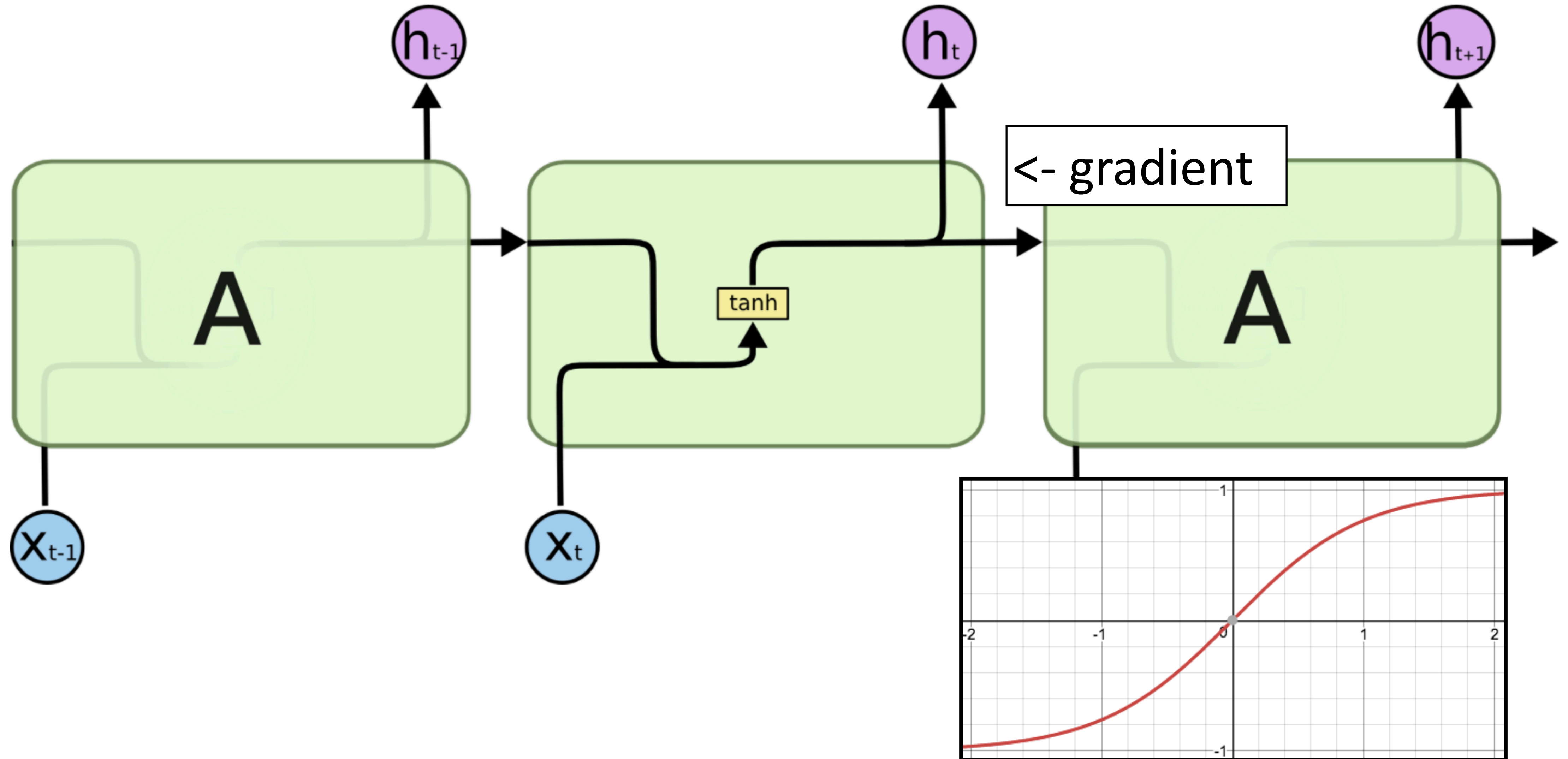
---



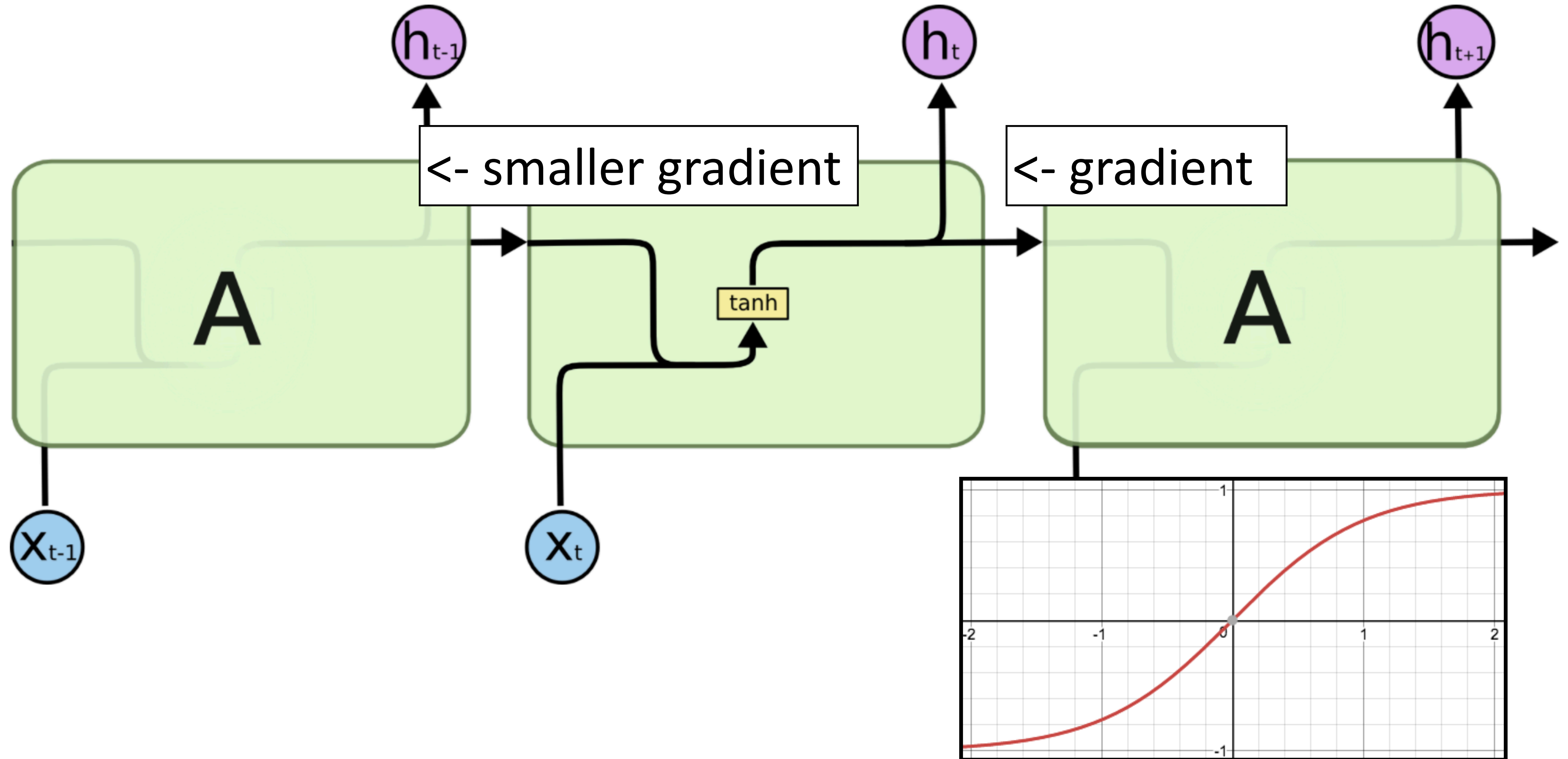
# Vanishing Gradient



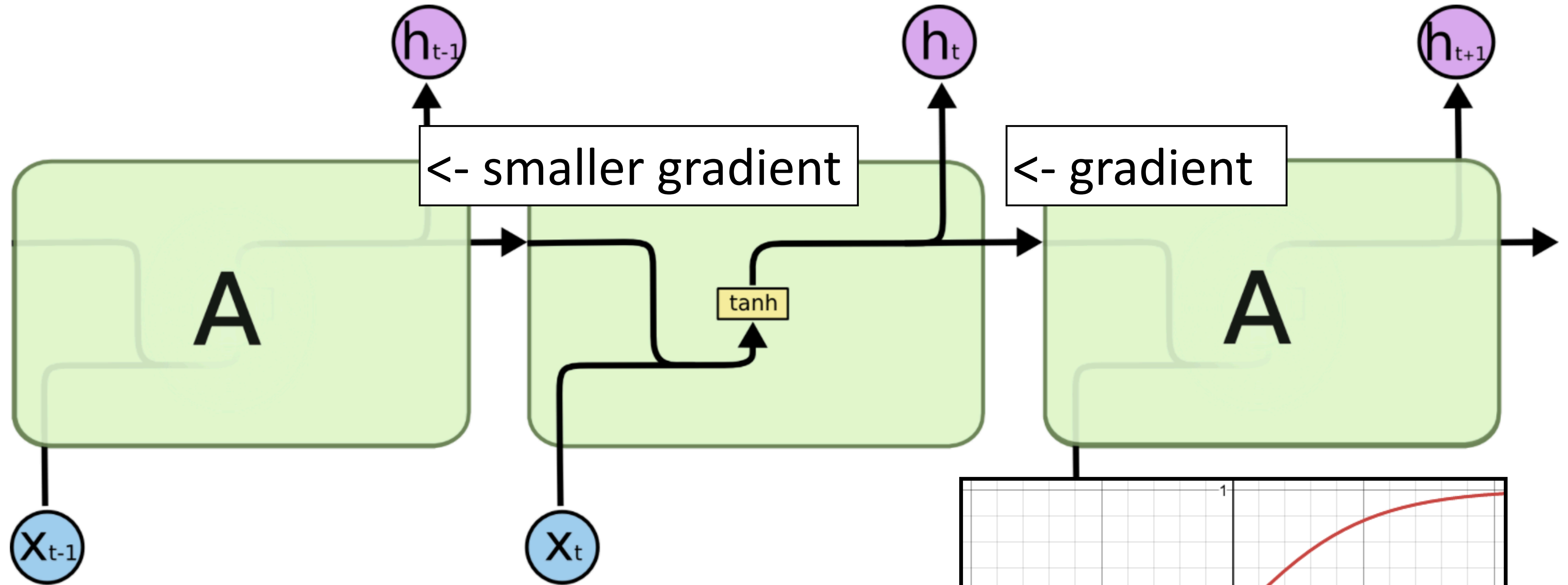
# Vanishing Gradient



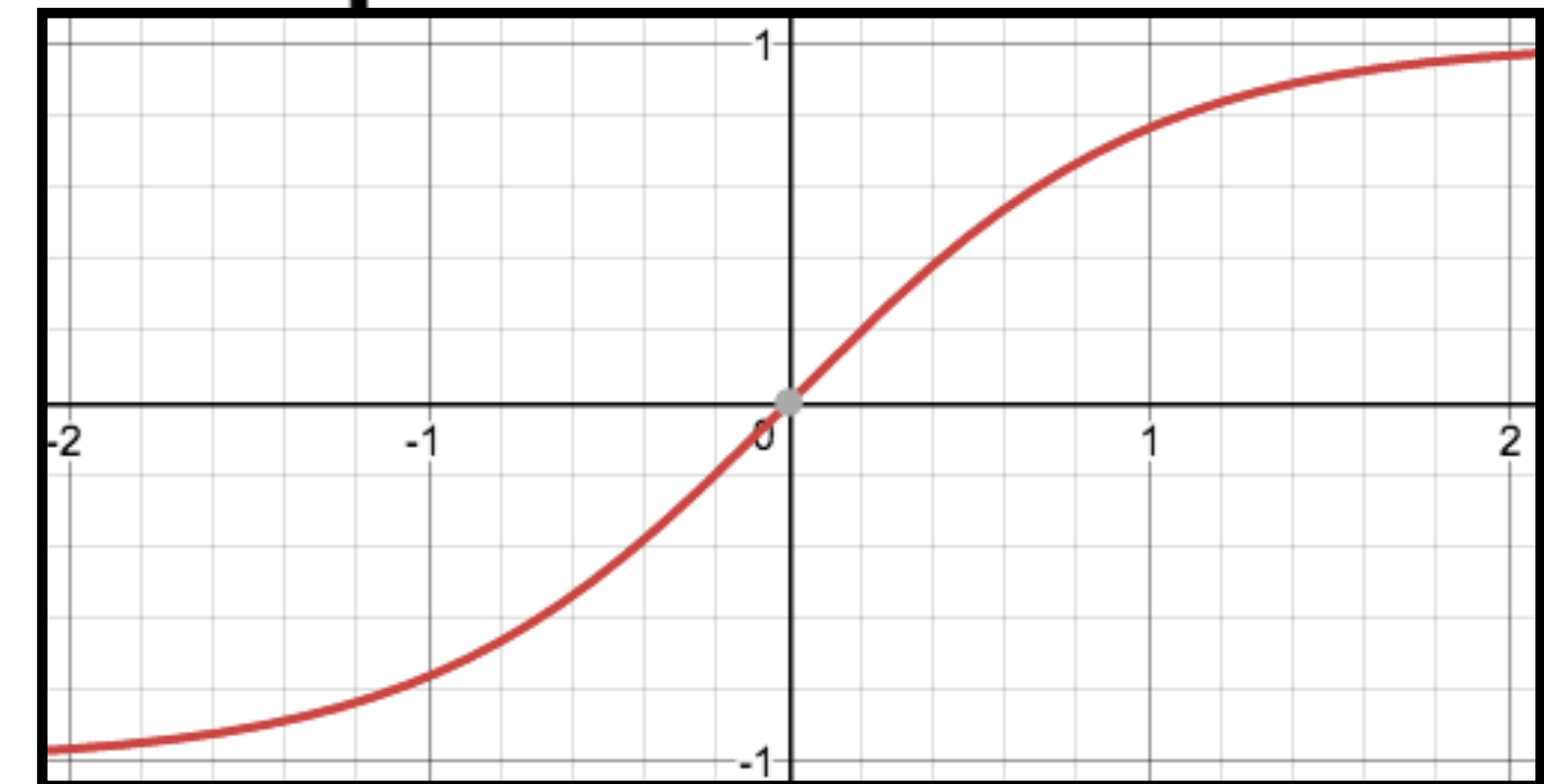
# Vanishing Gradient



# Vanishing Gradient

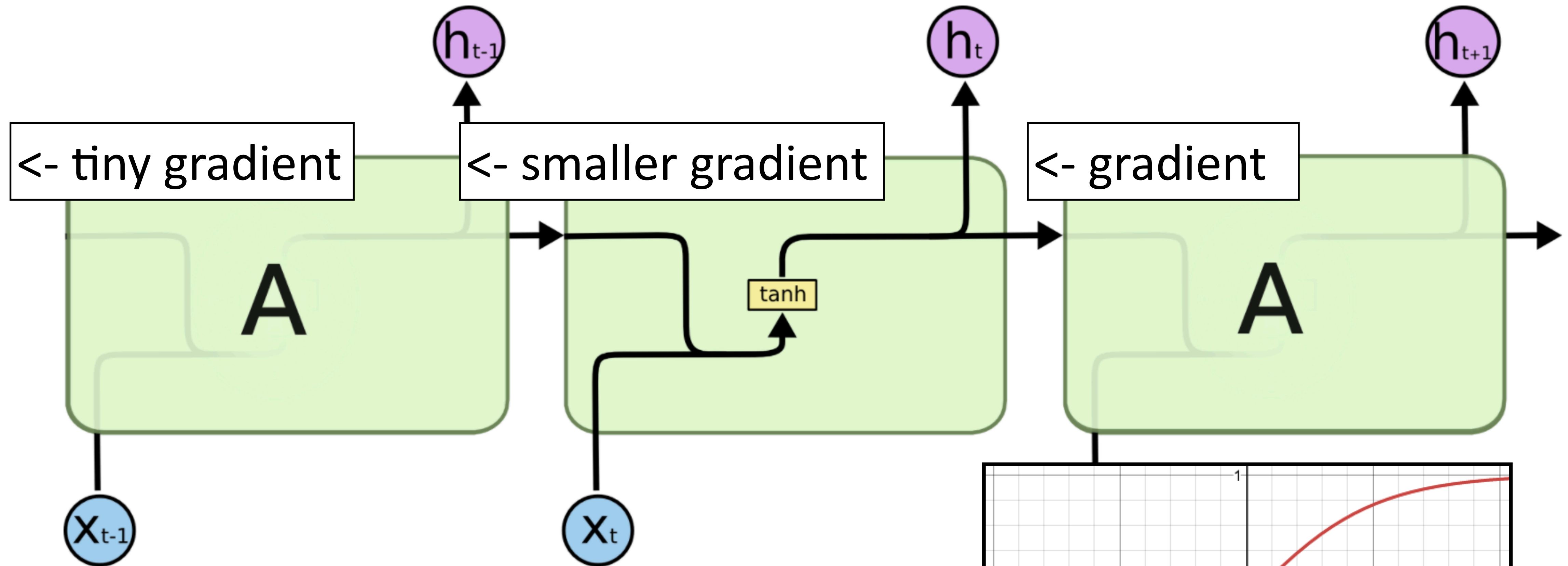


- ▶ Gradient diminishes going through tanh; if not in  $[-2, 2]$ , gradient is almost 0

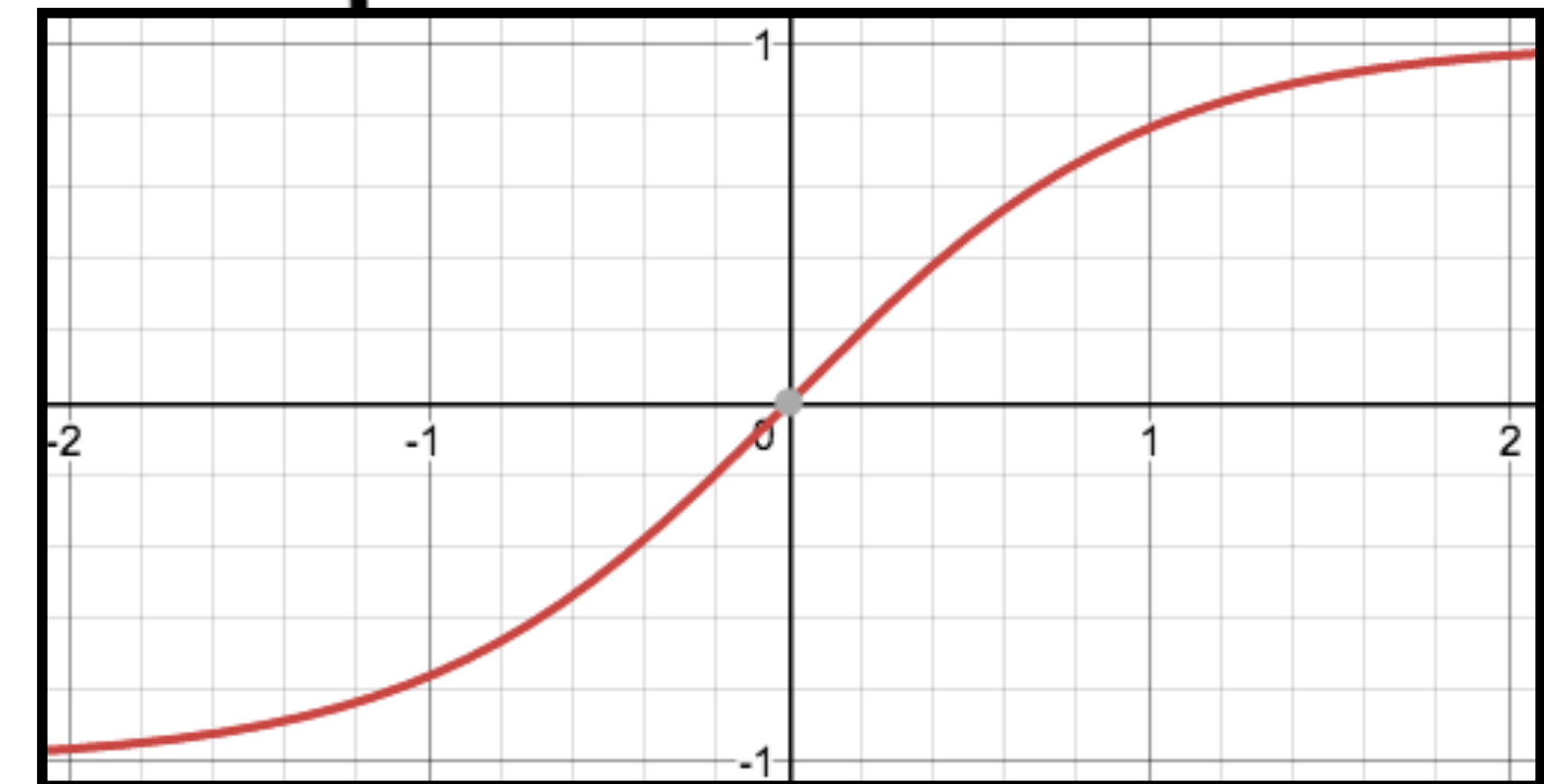




# Vanishing Gradient



- ▶ Gradient diminishes going through tanh; if not in  $[-2, 2]$ , gradient is almost 0



LSTMs/GRUs

# Gated Connections

---

- ▶ Designed to fix “vanishing gradient” problem using *gates*

$$\mathbf{h}_t = \mathbf{h}_{t-1} \odot \mathbf{f} + \text{func}(\mathbf{x}_t)$$

gated

$$\mathbf{h}_t = \tanh(W\mathbf{x}_t + V\mathbf{h}_{t-1} + \mathbf{b}_h)$$

Elman



# Gated Connections

---

- ▶ Designed to fix “vanishing gradient” problem using *gates*

$$\mathbf{h}_t = \mathbf{h}_{t-1} \odot \mathbf{f} + \text{func}(\mathbf{x}_t)$$

gated

$$\mathbf{h}_t = \tanh(W\mathbf{x}_t + V\mathbf{h}_{t-1} + \mathbf{b}_h)$$

Elman

- ▶ Vector-valued “forget gate”  $\mathbf{f}$  computed based on input and previous hidden state

$$\mathbf{f} = \sigma(W^{xf}\mathbf{x}_t + W^{hf}\mathbf{h}_{t-1})$$

- ▶ Sigmoid: elements of  $\mathbf{f}$  are in  $(0, 1)$

# Gated Connections

- ▶ Designed to fix “vanishing gradient” problem using *gates*

$$\mathbf{h}_t = \mathbf{h}_{t-1} \odot \mathbf{f} + \text{func}(\mathbf{x}_t)$$

gated

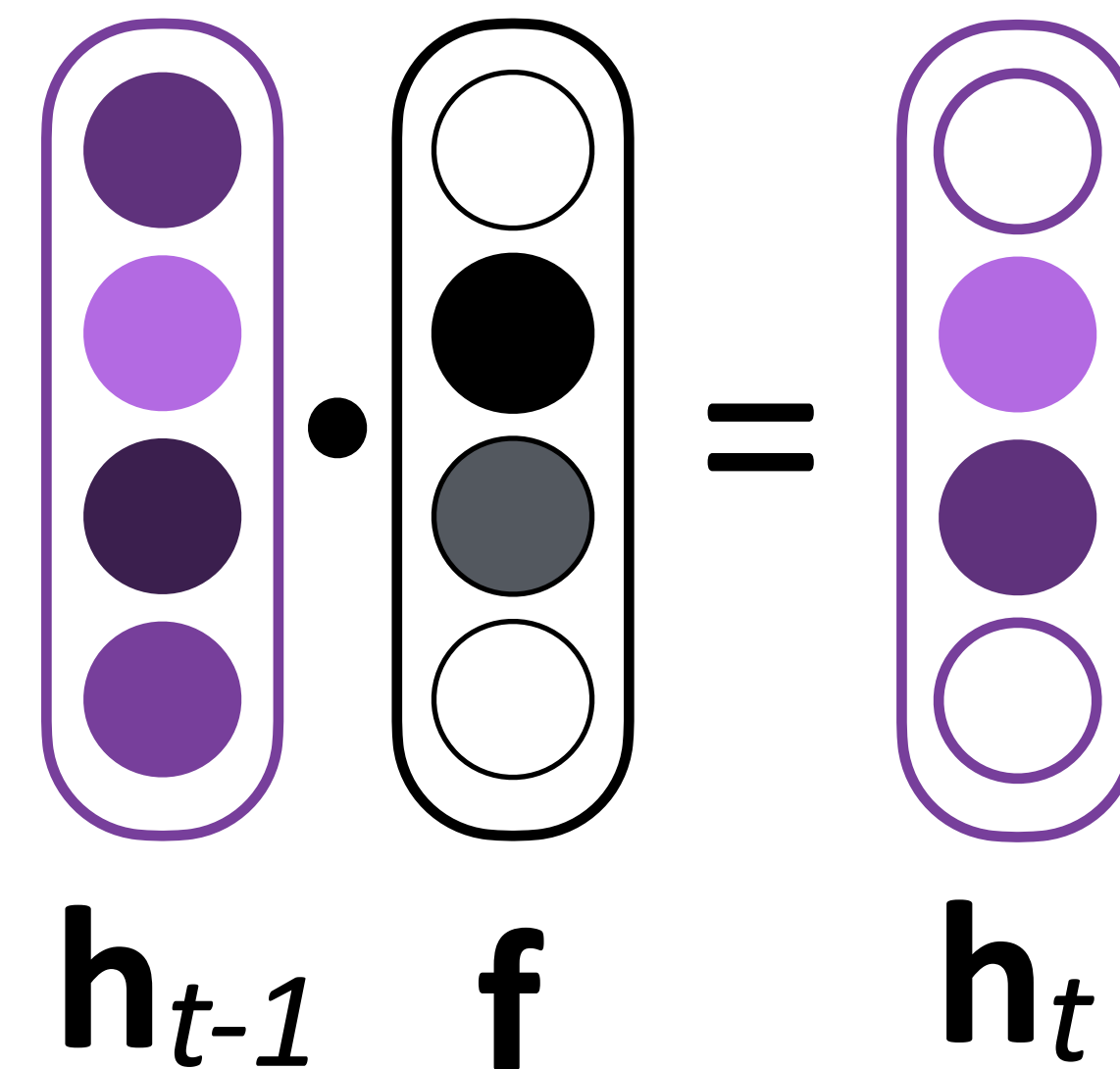
$$\mathbf{h}_t = \tanh(W\mathbf{x}_t + V\mathbf{h}_{t-1} + \mathbf{b}_h)$$

Elman

- ▶ Vector-valued “forget gate”  $\mathbf{f}$  computed based on input and previous hidden state

$$\mathbf{f} = \sigma(W^{xf}\mathbf{x}_t + W^{hf}\mathbf{h}_{t-1})$$

- ▶ Sigmoid: elements of  $\mathbf{f}$  are in  $(0, 1)$



# Gated Connections

- ▶ Designed to fix “vanishing gradient” problem using *gates*

$$\mathbf{h}_t = \mathbf{h}_{t-1} \odot \mathbf{f} + \text{func}(\mathbf{x}_t)$$

gated

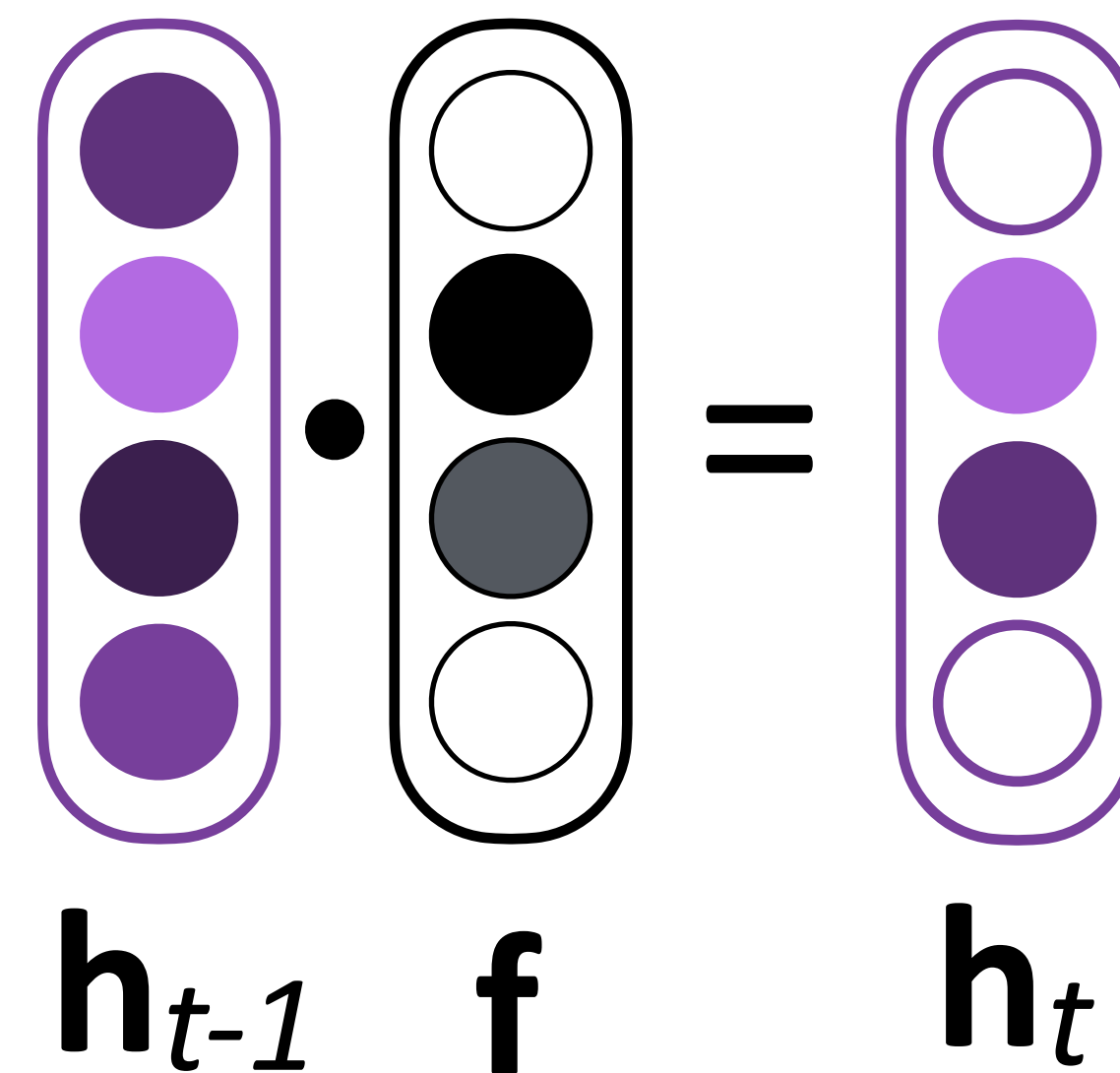
$$\mathbf{h}_t = \tanh(W\mathbf{x}_t + V\mathbf{h}_{t-1} + \mathbf{b}_h)$$

Elman

- ▶ Vector-valued “forget gate”  $\mathbf{f}$  computed based on input and previous hidden state

$$\mathbf{f} = \sigma(W^{xf}\mathbf{x}_t + W^{hf}\mathbf{h}_{t-1})$$

- ▶ Sigmoid: elements of  $\mathbf{f}$  are in  $(0, 1)$
- ▶ If  $\mathbf{f} \approx \mathbf{1}$ , we simply sum up a function of all inputs — gradient doesn’t vanish!



# LSTMs

---

- ▶ “Cell”  $\mathbf{c}$  in addition to hidden state  $\mathbf{h}$

$$\mathbf{c}_t = \mathbf{c}_{t-1} \odot \mathbf{f} + \text{func}(\mathbf{x}_t, \mathbf{h}_{t-1})$$

# LSTMs

---

- ▶ “Cell”  $\mathbf{c}$  in addition to hidden state  $\mathbf{h}$

$$\mathbf{c}_t = \mathbf{c}_{t-1} \odot \mathbf{f} + \text{func}(\mathbf{x}_t, \mathbf{h}_{t-1})$$

- ▶ Vector-valued forget gate  $\mathbf{f}$  depends on the  $\mathbf{h}$  hidden state

$$\mathbf{f} = \sigma(W^{xf}\mathbf{x}_t + W^{hf}\mathbf{h}_{t-1})$$

# LSTMs

---

- ▶ “Cell”  $\mathbf{c}$  in addition to hidden state  $\mathbf{h}$

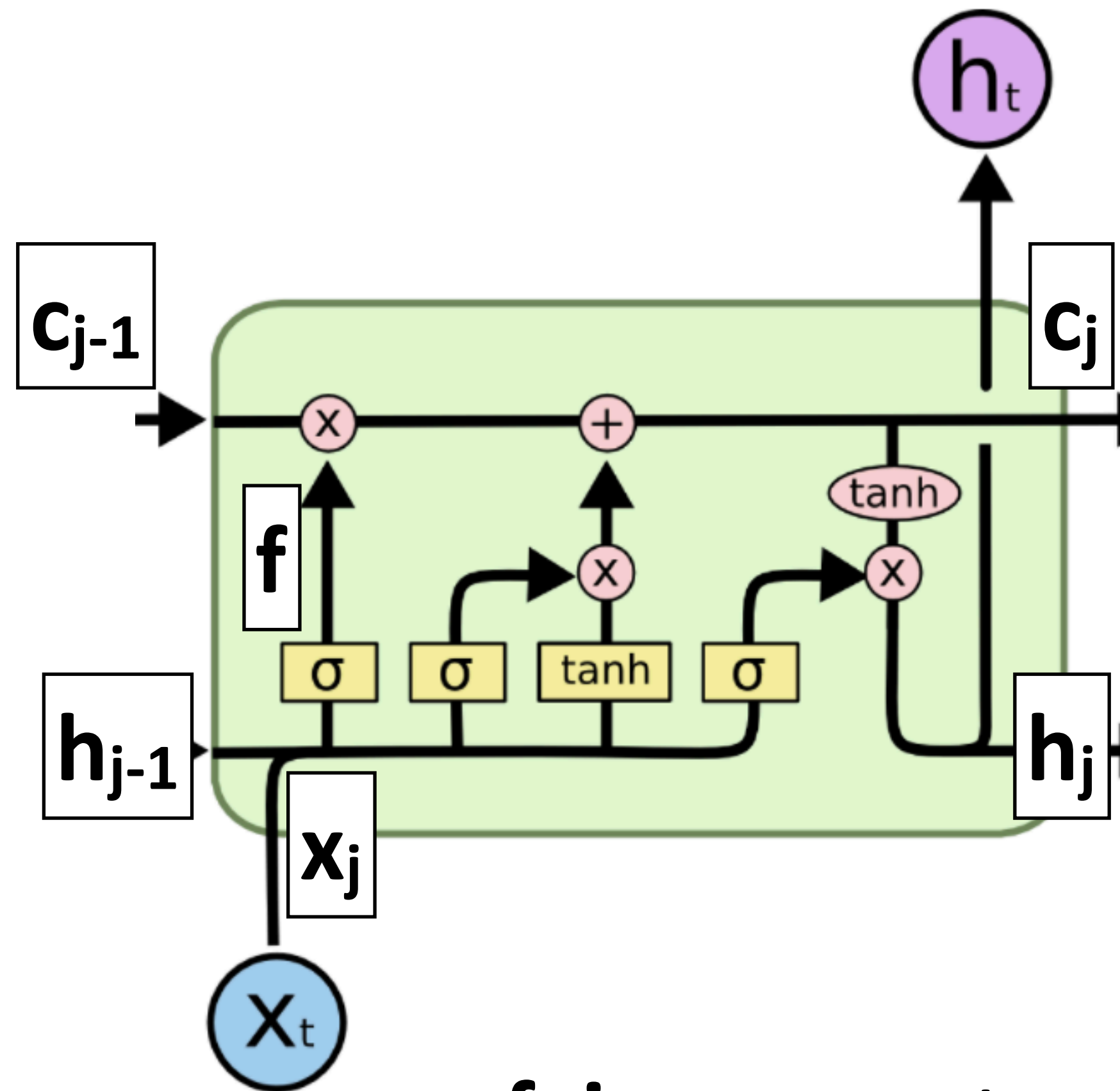
$$\mathbf{c}_t = \mathbf{c}_{t-1} \odot \mathbf{f} + \text{func}(\mathbf{x}_t, \mathbf{h}_{t-1})$$

- ▶ Vector-valued forget gate  $\mathbf{f}$  depends on the  $\mathbf{h}$  hidden state

$$\mathbf{f} = \sigma(W^{xf}\mathbf{x}_t + W^{hf}\mathbf{h}_{t-1})$$

- ▶ Basic communication flow:  $\mathbf{x} \rightarrow \mathbf{c} \rightarrow \mathbf{h} \rightarrow \text{output}$ , each step of this process is gated in addition to gates from previous timesteps

# LSTMs



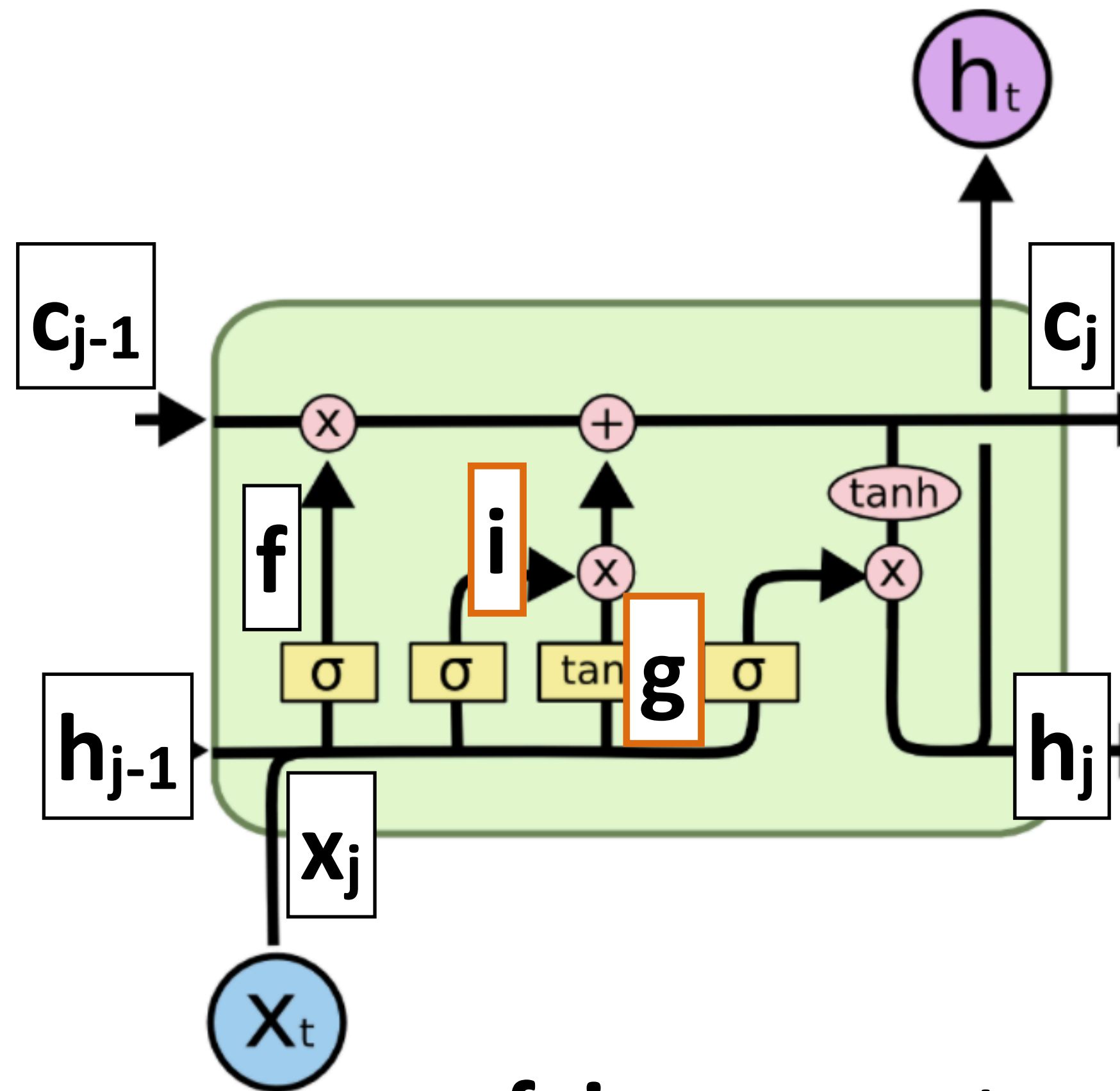
$$c_j = c_{j-1} \odot f + \boxed{g \odot i}$$

$$f = \sigma(x_j W^{xf} + h_{j-1} W^{hf})$$

- ▶  $f, i, o$  are gates that control information flow
- ▶  $g$  reflects the main computation of the cell



# LSTMs



$$c_j = c_{j-1} \odot f + g \odot i$$

$$f = \sigma(x_j W^{xf} + h_{j-1} W^{hf})$$

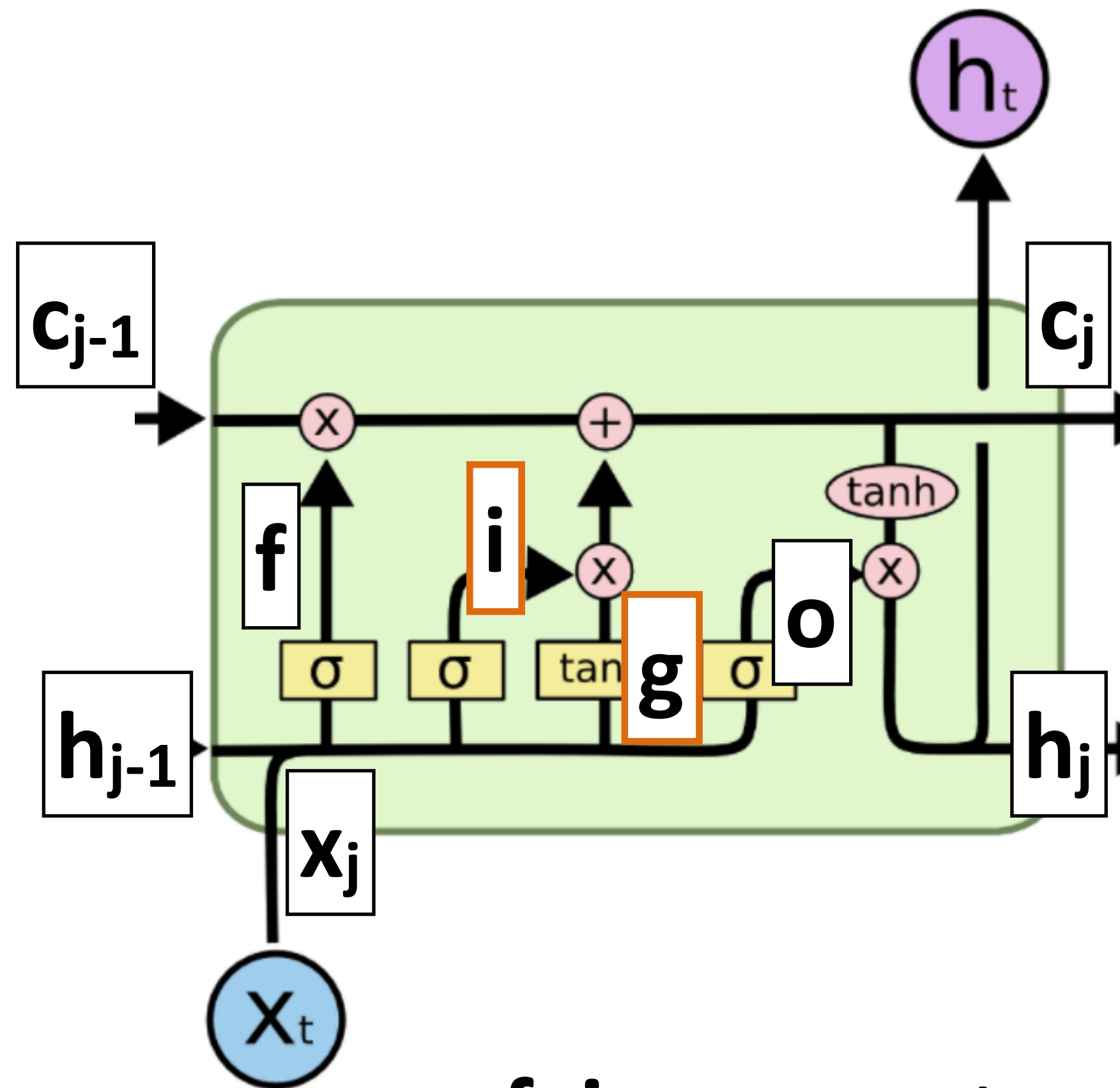
$$g = \tanh(x_j W^{xg} + h_{j-1} W^{hg})$$

$$i = \sigma(x_j W^{xi} + h_{j-1} W^{hi})$$

- ▶  $f, i, o$  are gates that control information flow
- ▶  $g$  reflects the main computation of the cell



# LSTMs



$$c_j = c_{j-1} \odot f + g \odot i$$

$$f = \sigma(x_j W^{xf} + h_{j-1} W^{hf})$$

$$g = \tanh(x_j W^{xg} + h_{j-1} W^{hg})$$

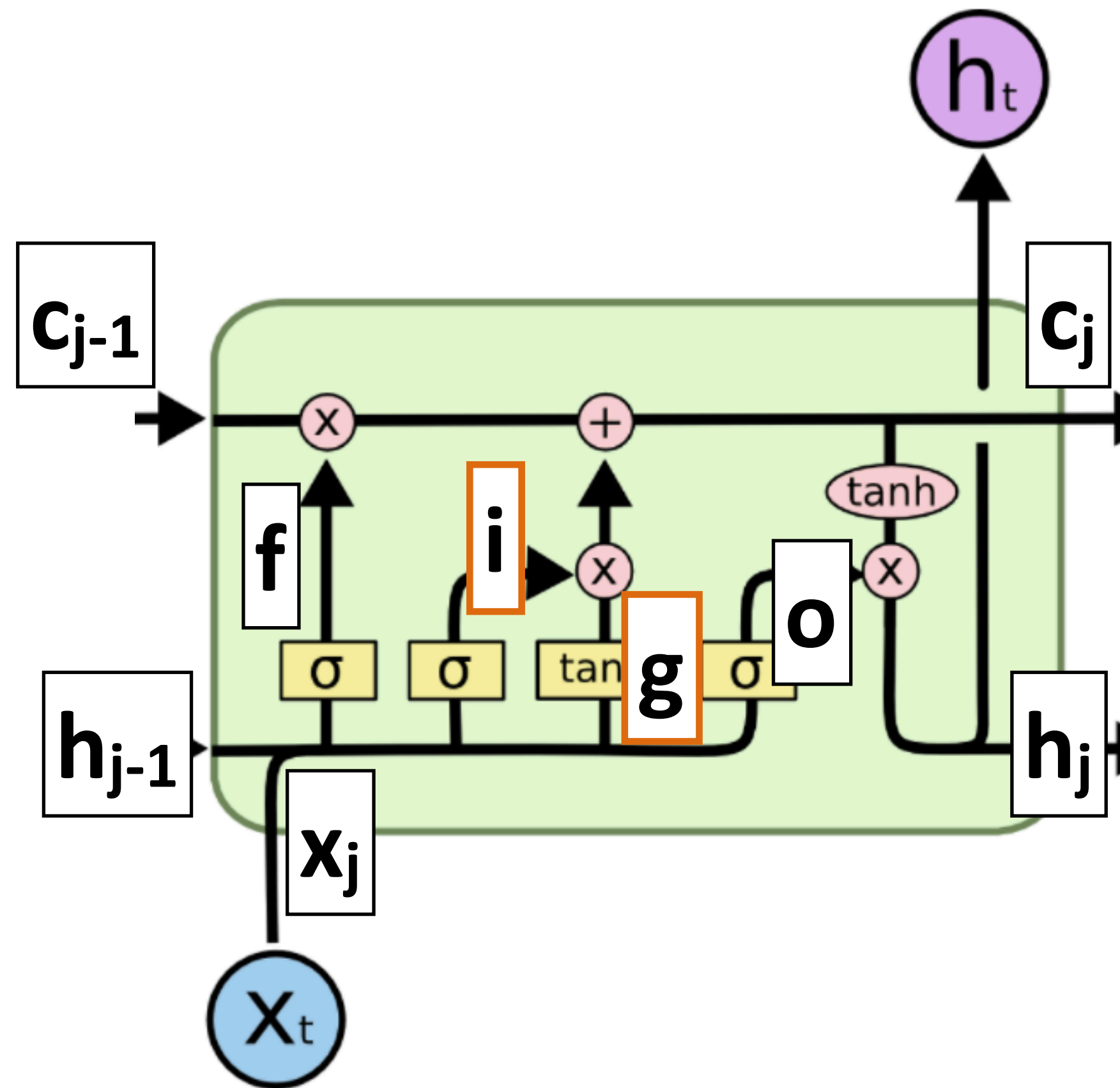
$$i = \sigma(x_j W^{xi} + h_{j-1} W^{hi})$$

$$h_j = \tanh(c_j) \odot o$$

$$o = \sigma(x_j W^{xo} + h_{j-1} W^{ho})$$

- ▶  $f$ ,  $i$ ,  $o$  are gates that control information flow
- ▶  $g$  reflects the main computation of the cell

# LSTMs



$$c_j = c_{j-1} \odot f + g \odot i$$

$$f = \sigma(x_j W^{xf} + h_{j-1} W^{hf})$$

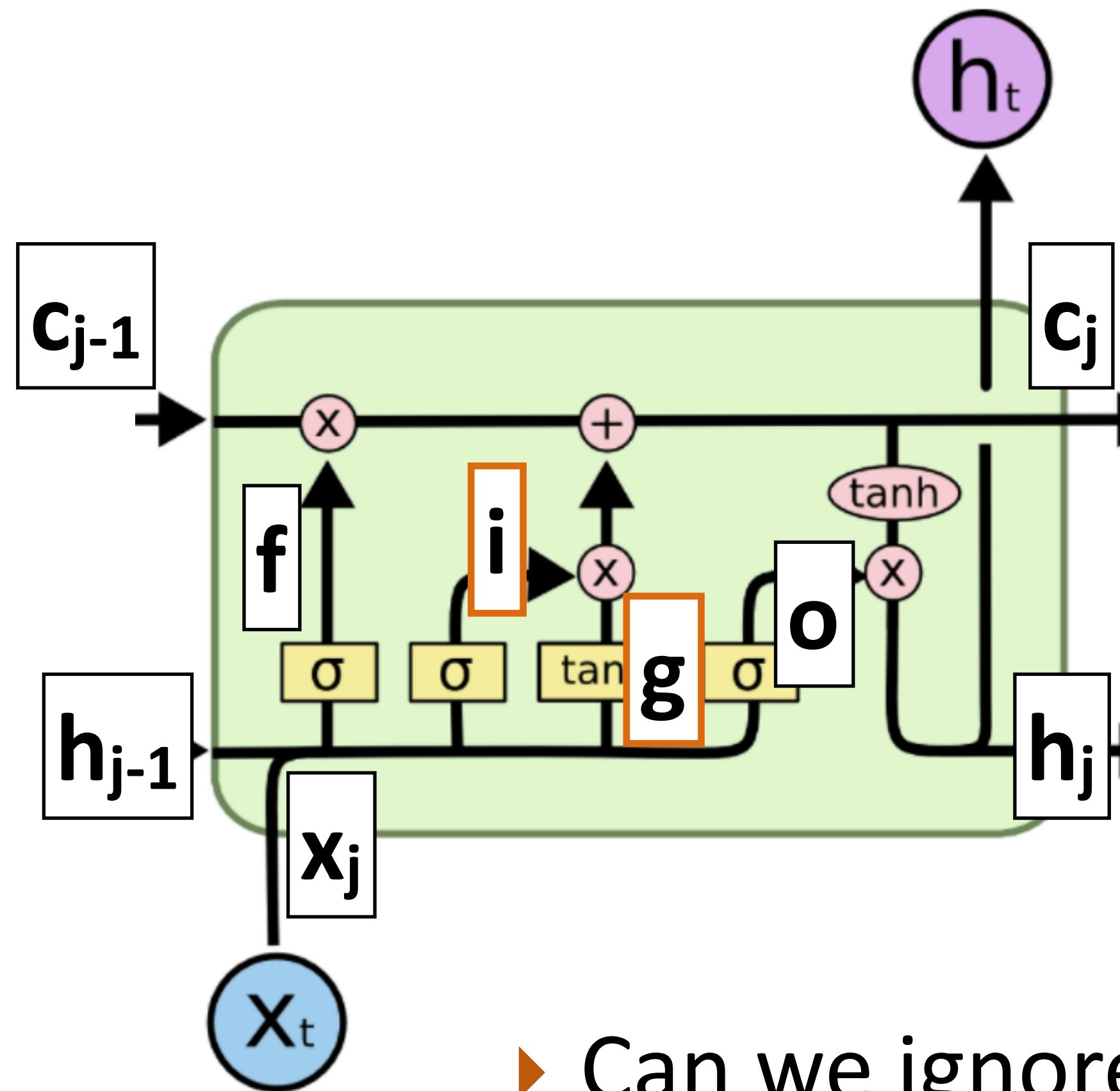
$$g = \tanh(x_j W^{xg} + h_{j-1} W^{hg})$$

$$i = \sigma(x_j W^{xi} + h_{j-1} W^{hi})$$

$$h_j = \tanh(c_j) \odot o$$

$$o = \sigma(x_j W^{xo} + h_{j-1} W^{ho})$$

# LSTMs



$$c_j = c_{j-1} \odot f + g \odot i$$

$$f = \sigma(x_j W^{xf} + h_{j-1} W^{hf})$$

$$g = \tanh(x_j W^{xg} + h_{j-1} W^{hg})$$

$$i = \sigma(x_j W^{xi} + h_{j-1} W^{hi})$$

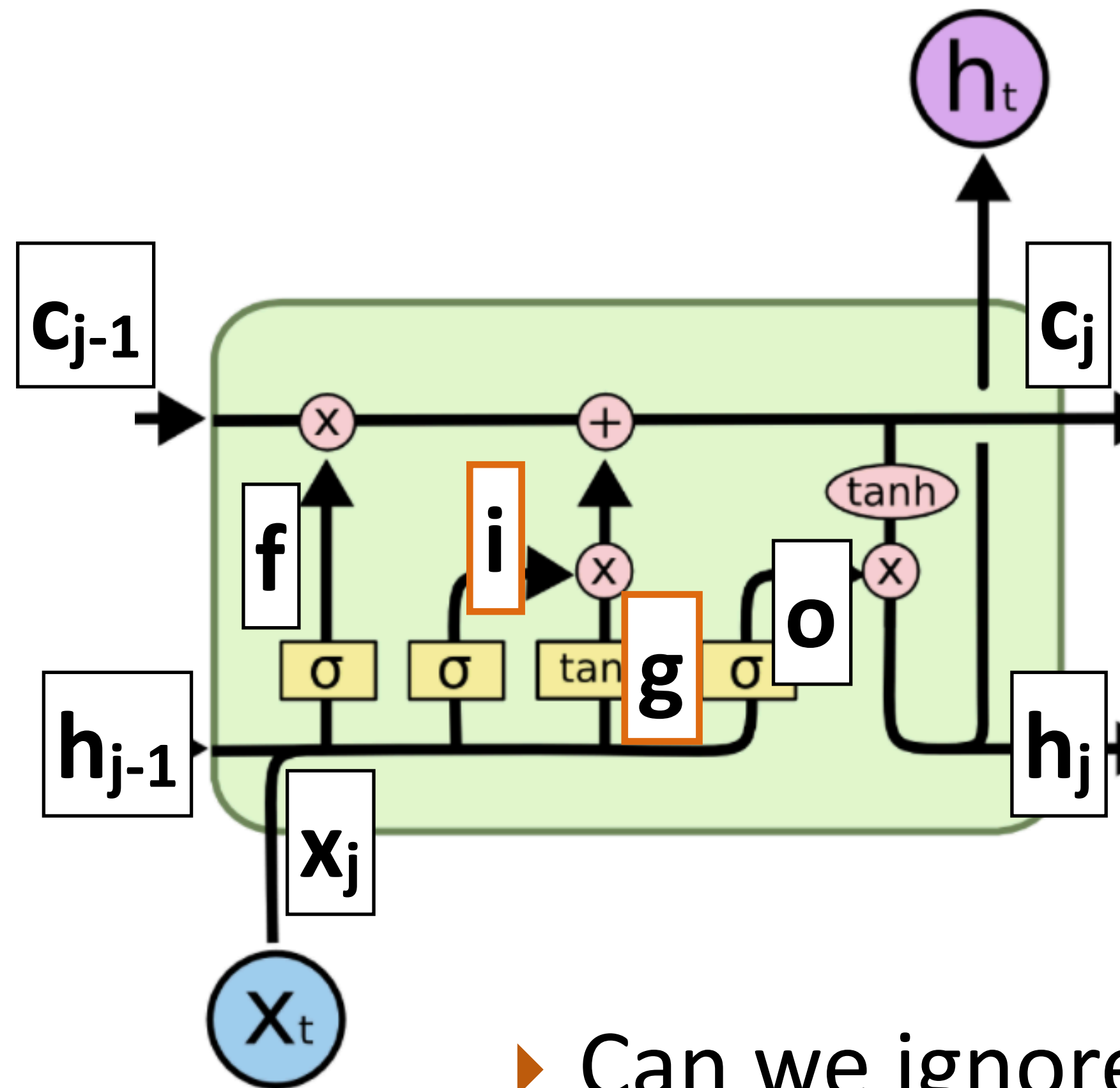
$$h_j = \tanh(c_j) \odot o$$

$$o = \sigma(x_j W^{xo} + h_{j-1} W^{ho})$$

- Can we ignore the old value of  $c$  for this timestep?



# LSTMs



$$c_j = c_{j-1} \odot f + g \odot i$$

$$f = \sigma(x_j W^{xf} + h_{j-1} W^{hf})$$

$$g = \tanh(x_j W^{xg} + h_{j-1} W^{hg})$$

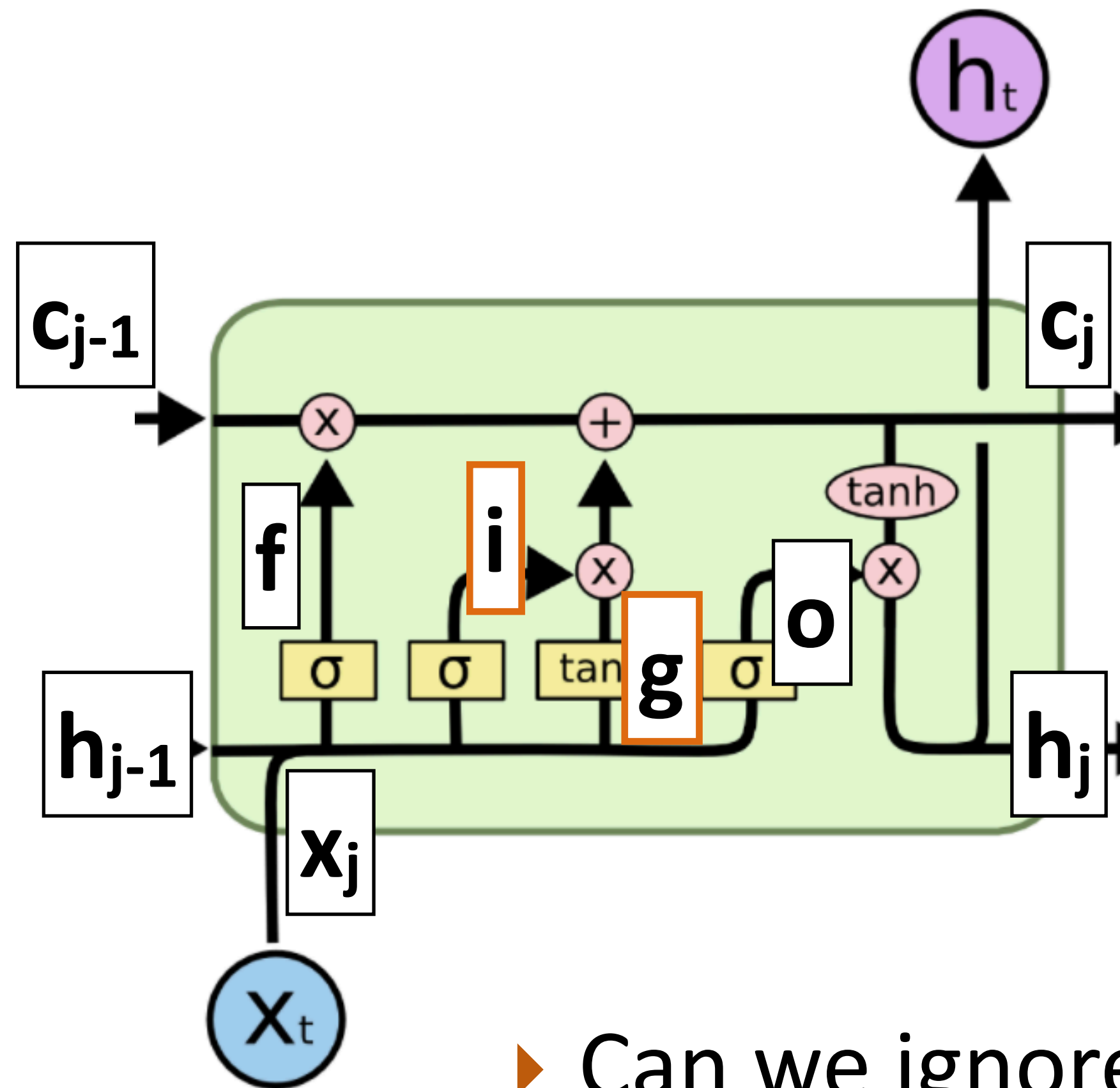
$$i = \sigma(x_j W^{xi} + h_{j-1} W^{hi})$$

$$h_j = \tanh(c_j) \odot o$$

$$o = \sigma(x_j W^{xo} + h_{j-1} W^{ho})$$

- ▶ Can we ignore the old value of  $c$  for this timestep?
- ▶ Can an LSTM sum up its inputs  $x$ ?

# LSTMs



$$c_j = c_{j-1} \odot f + g \odot i$$

$$f = \sigma(x_j W^{xf} + h_{j-1} W^{hf})$$

$$g = \tanh(x_j W^{xg} + h_{j-1} W^{hg})$$

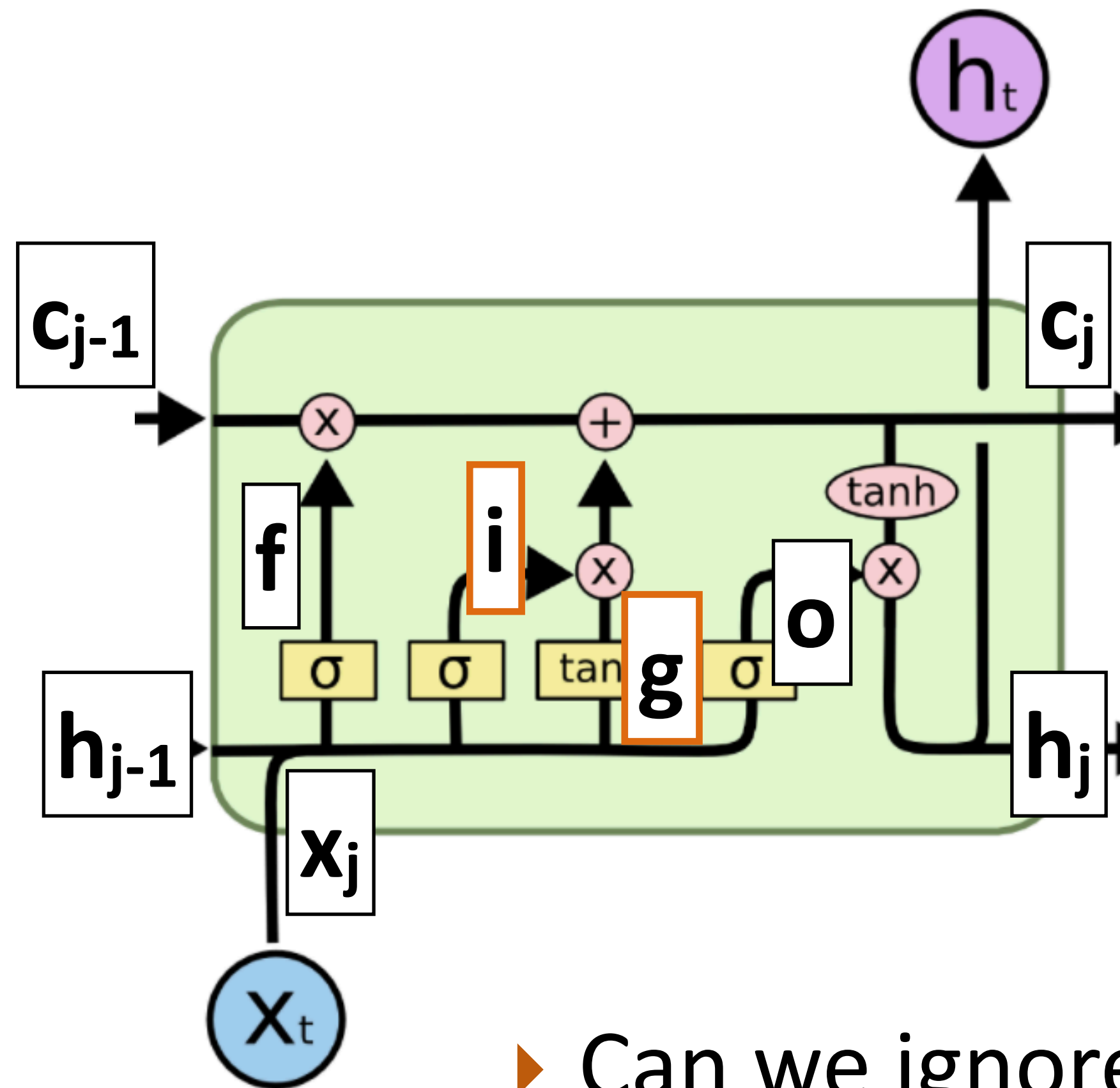
$$i = \sigma(x_j W^{xi} + h_{j-1} W^{hi})$$

$$h_j = \tanh(c_j) \odot o$$

$$o = \sigma(x_j W^{xo} + h_{j-1} W^{ho})$$

- ▶ Can we ignore the old value of  $c$  for this timestep?
- ▶ Can an LSTM sum up its inputs  $x$ ?
- ▶ Can we ignore a particular input  $x$ ?

# LSTMs



$$c_j = c_{j-1} \odot f + g \odot i$$

$$f = \sigma(x_j W^{xf} + h_{j-1} W^{hf})$$

$$g = \tanh(x_j W^{xg} + h_{j-1} W^{hg})$$

$$i = \sigma(x_j W^{xi} + h_{j-1} W^{hi})$$

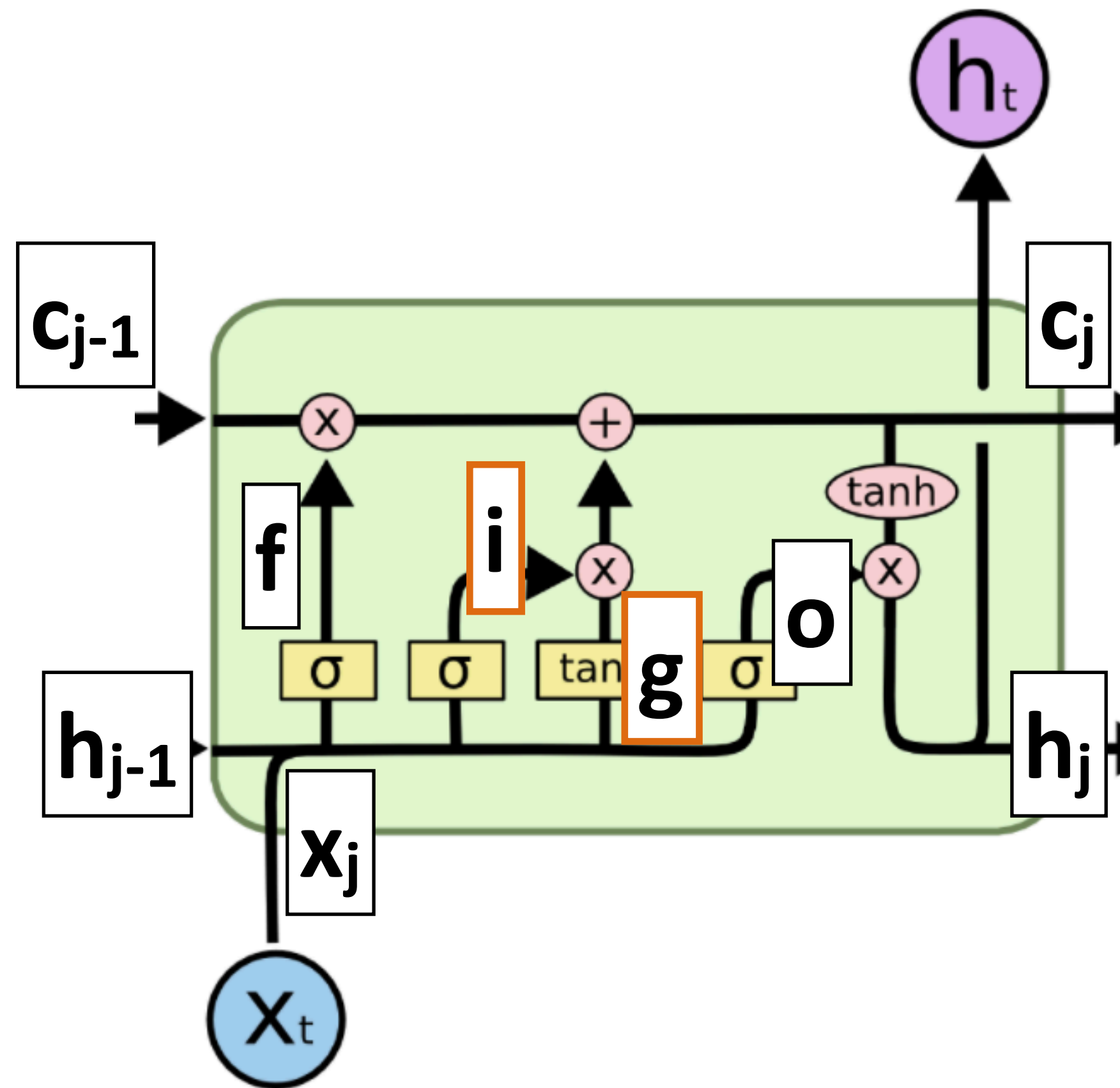
$$h_j = \tanh(c_j) \odot o$$

$$o = \sigma(x_j W^{xo} + h_{j-1} W^{ho})$$

- ▶ Can we ignore the old value of  $c$  for this timestep?
- ▶ Can an LSTM sum up its inputs  $x$ ?
- ▶ Can we ignore a particular input  $x$ ?
- ▶ Can we output something without changing  $c$ ?

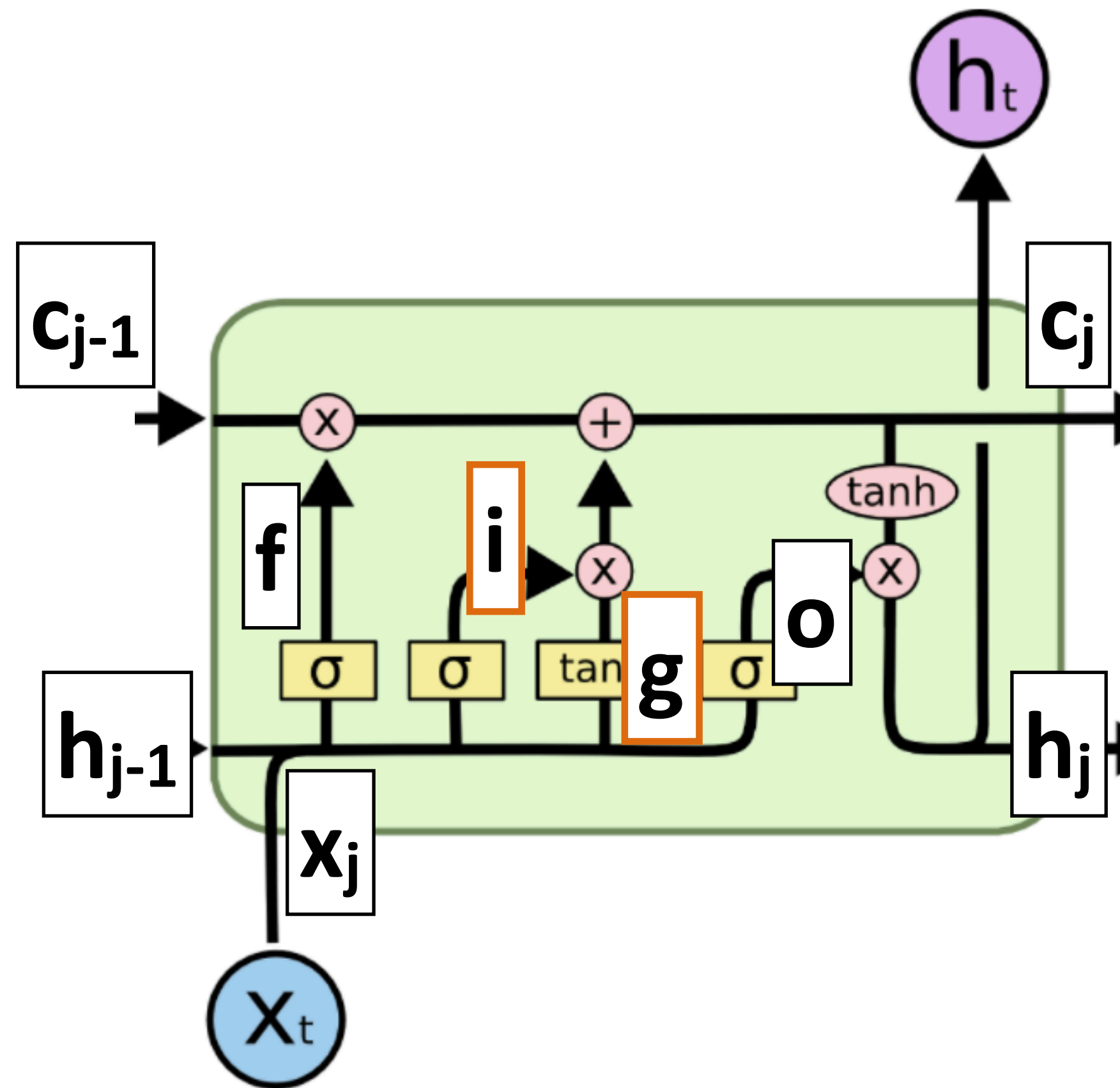


# LSTMs



- ▶ Ignoring recurrent state entirely:
  - ▶ Lets us get feedforward layer over token

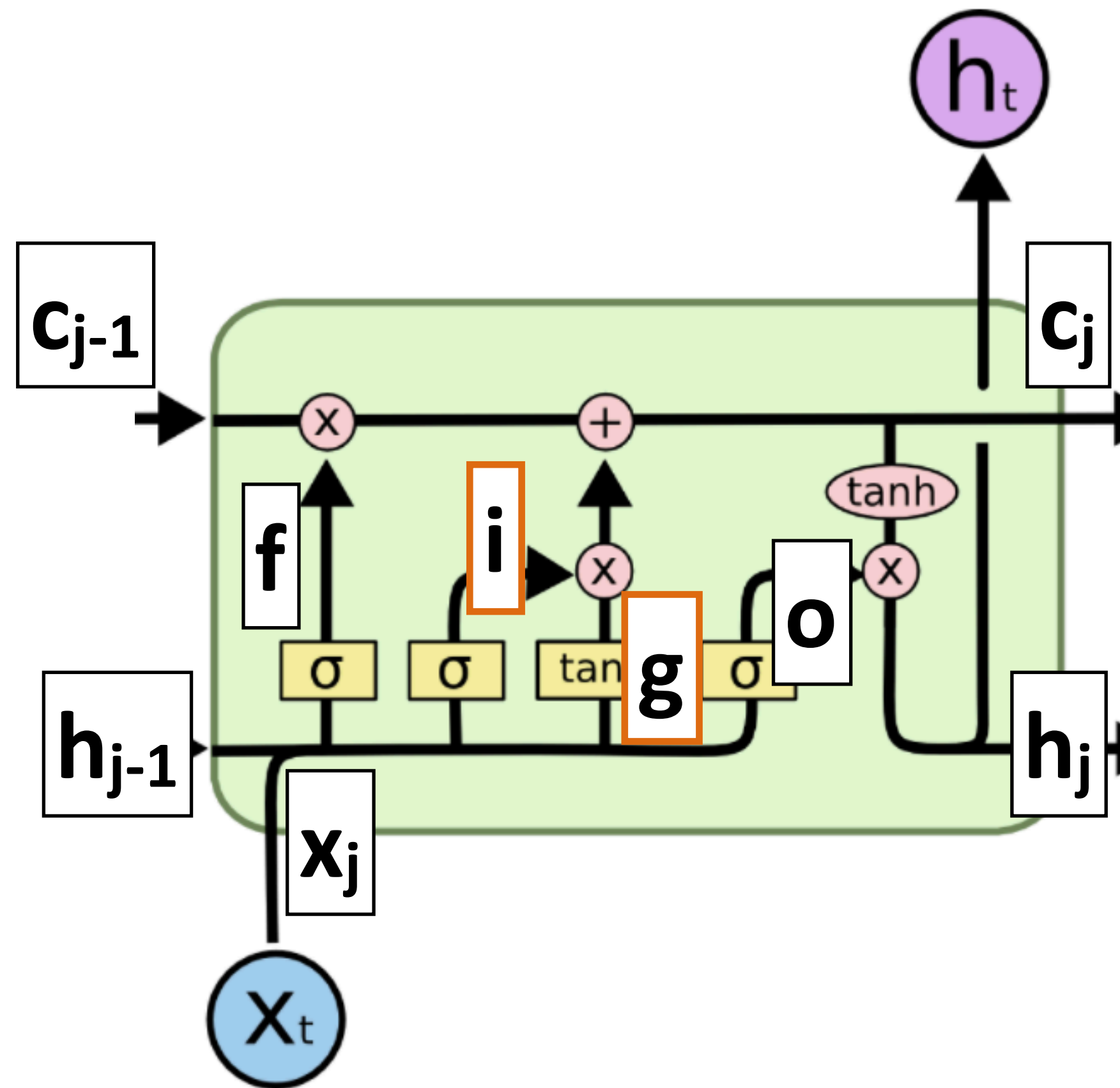
# LSTMs



- ▶ Ignoring recurrent state entirely:
  - ▶ Lets us get feedforward layer over token
- ▶ Ignoring input:
  - ▶ Lets us discard stopwords

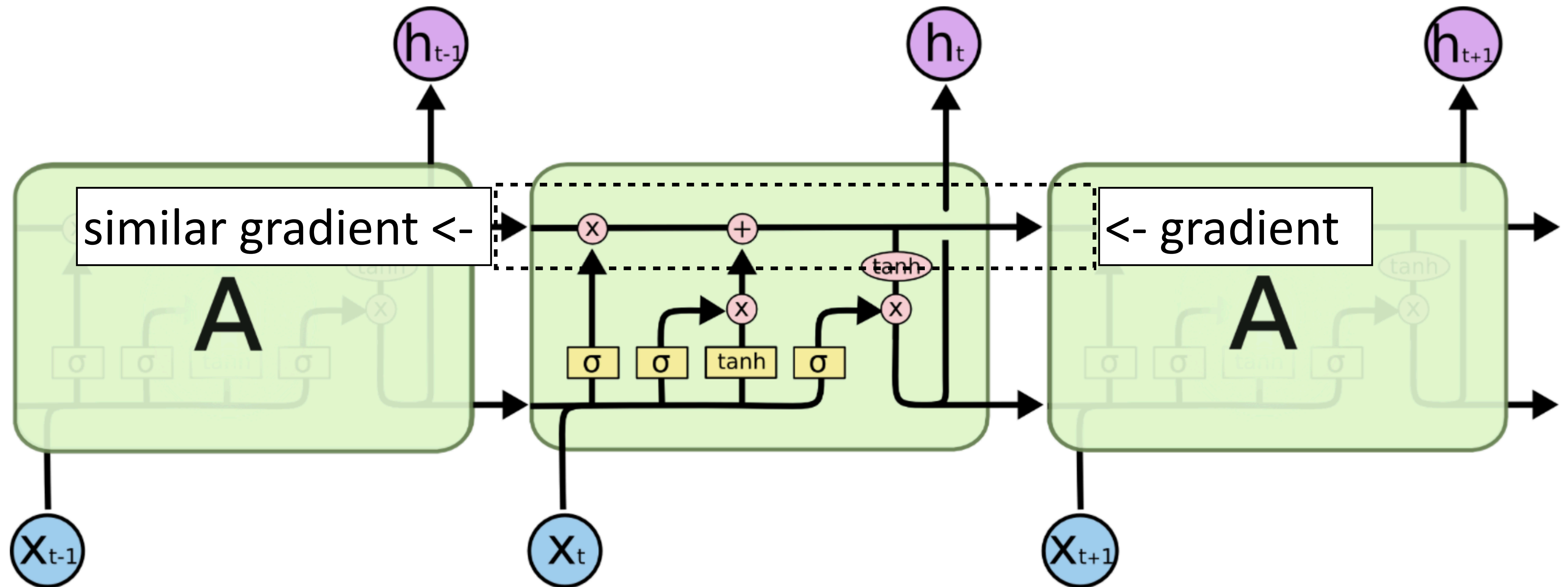


# LSTMs

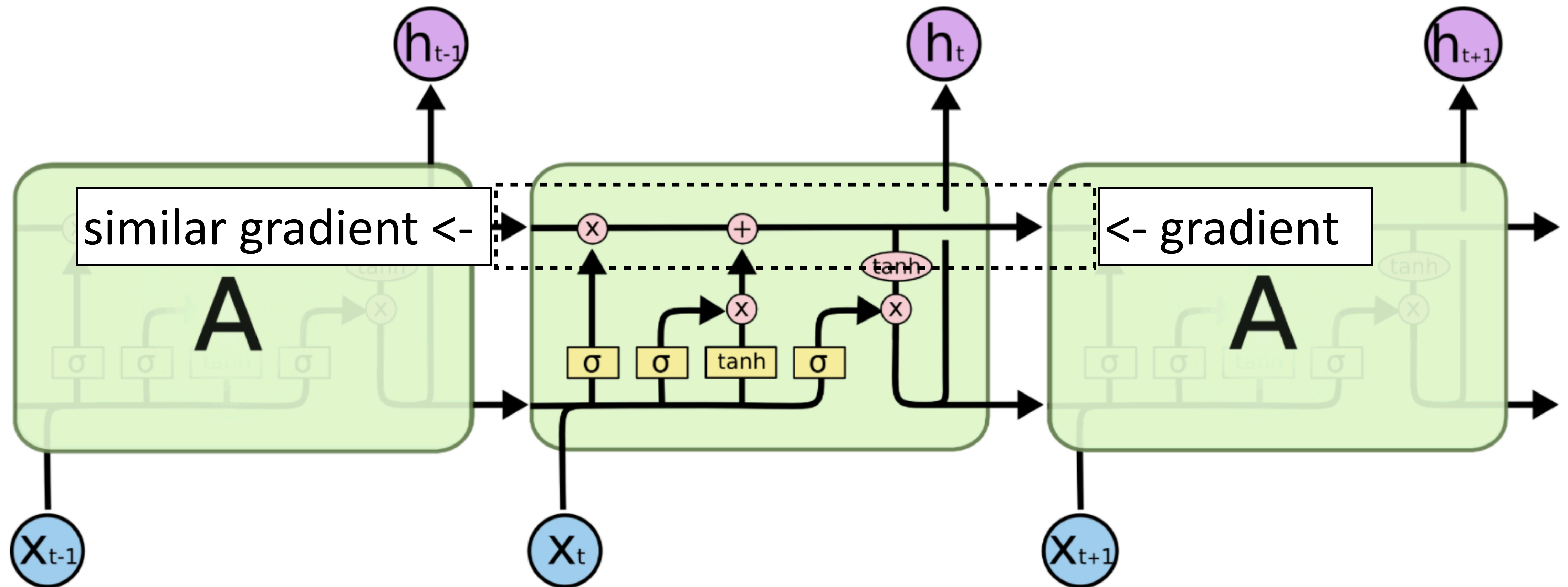


- ▶ Ignoring recurrent state entirely:
  - ▶ Lets us get feedforward layer over token
- ▶ Ignoring input:
  - ▶ Lets us discard stopwords
- ▶ Summing inputs:
  - ▶ Lets us compute a bag-of-words representation

# LSTMs

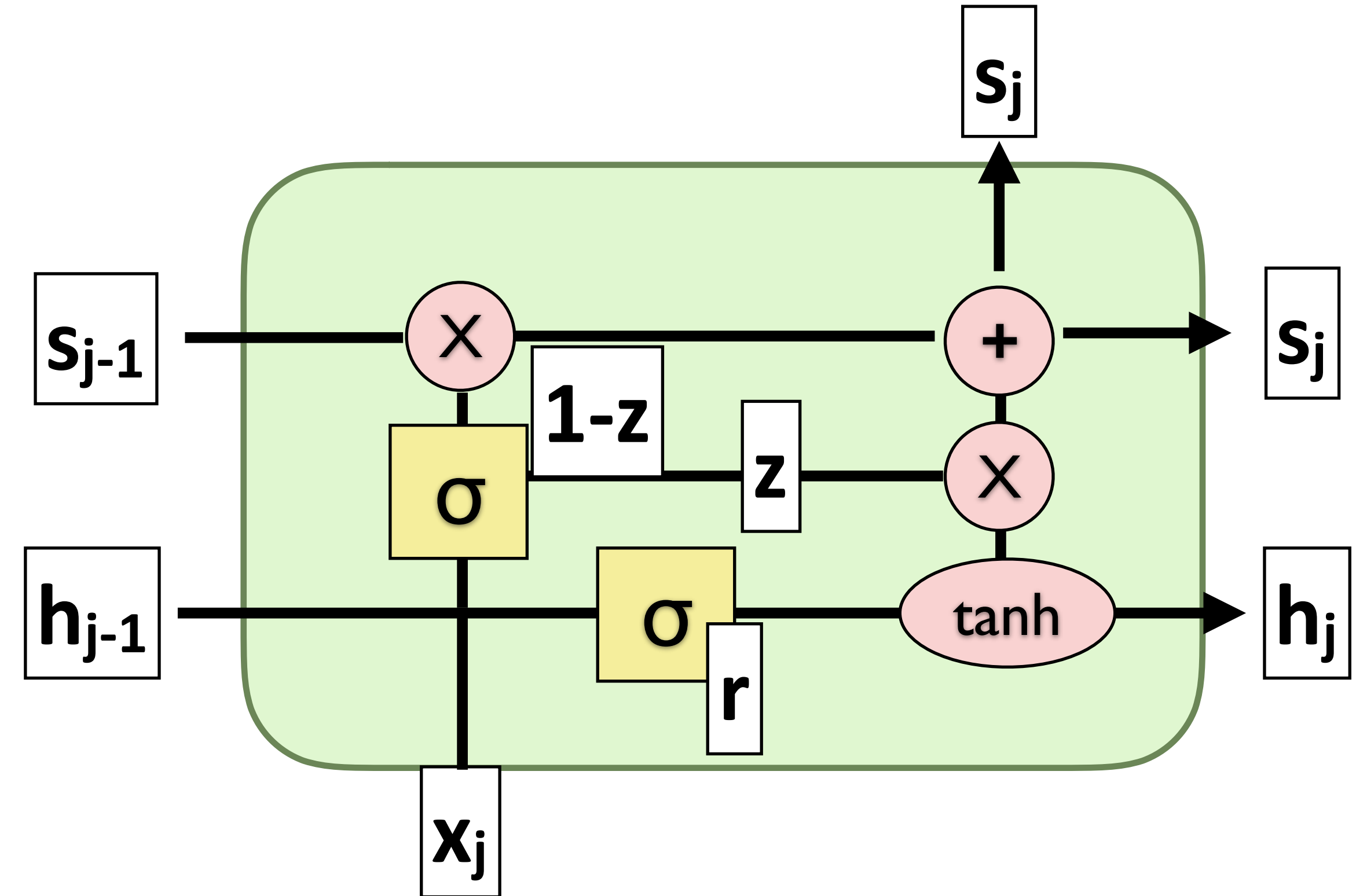
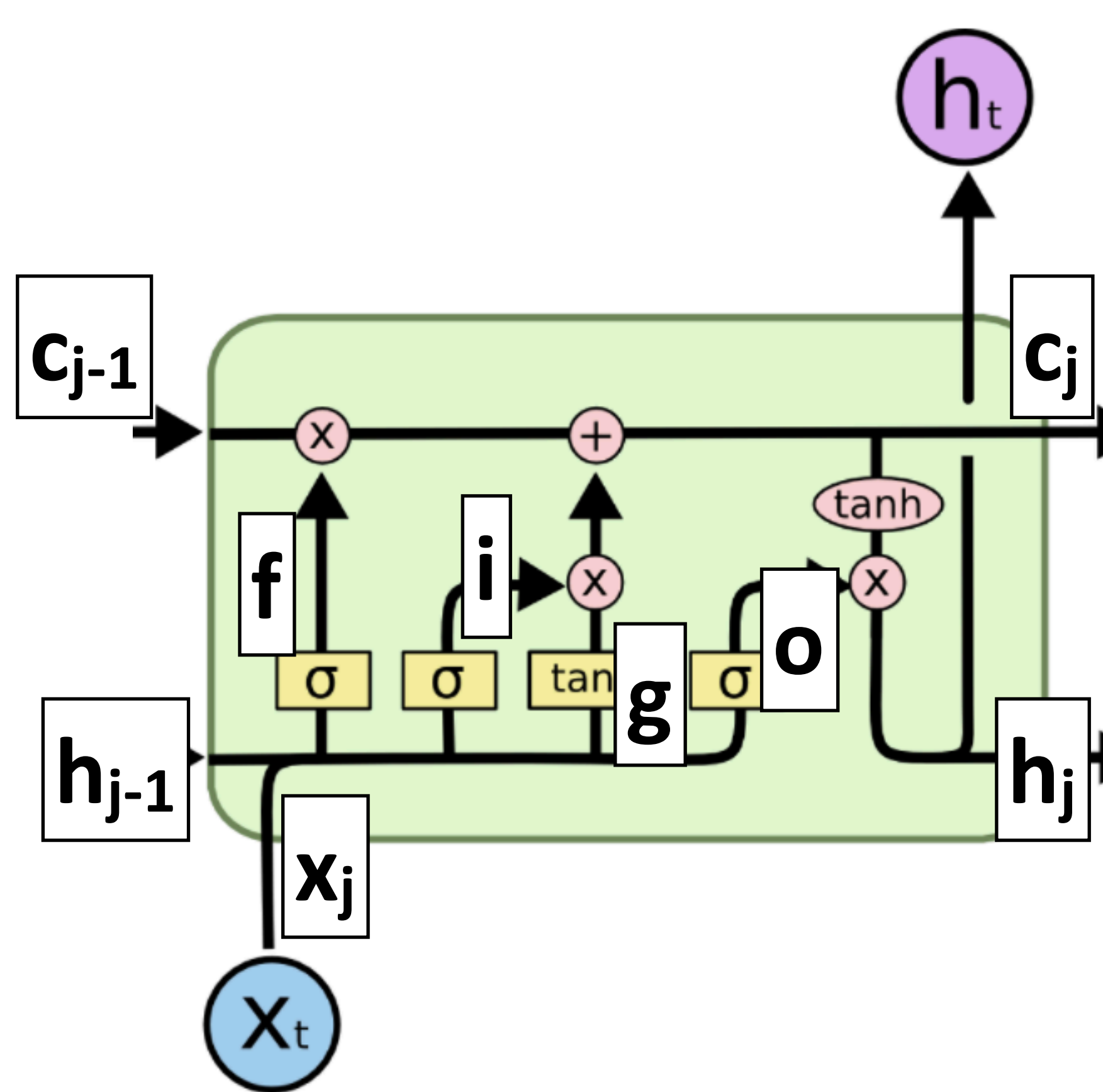


# LSTMs



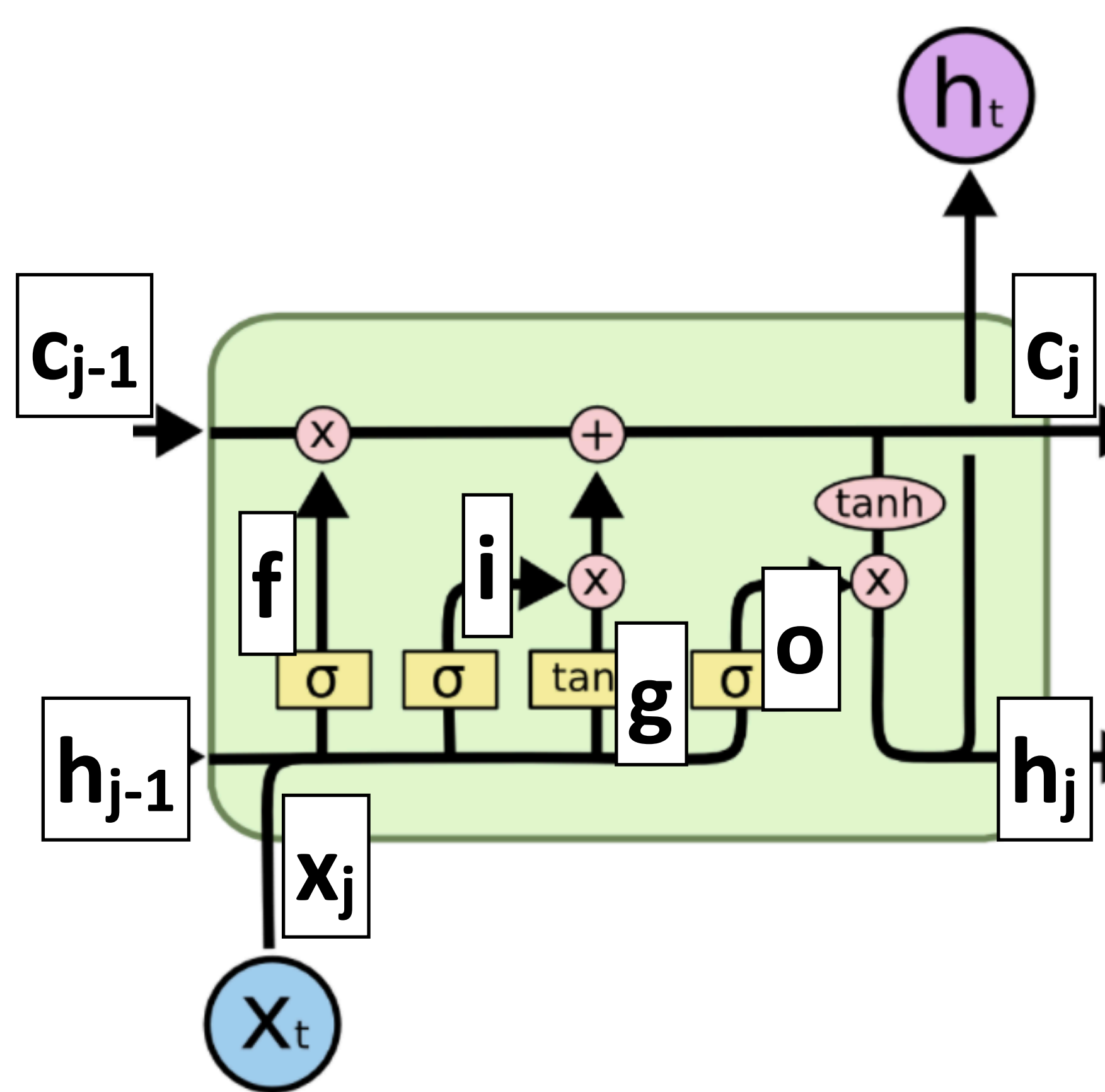
- ▶ Gradient still diminishes, but in a controlled way and generally by less — usually initialize forget gate = 1 to remember everything to start

# GRUs

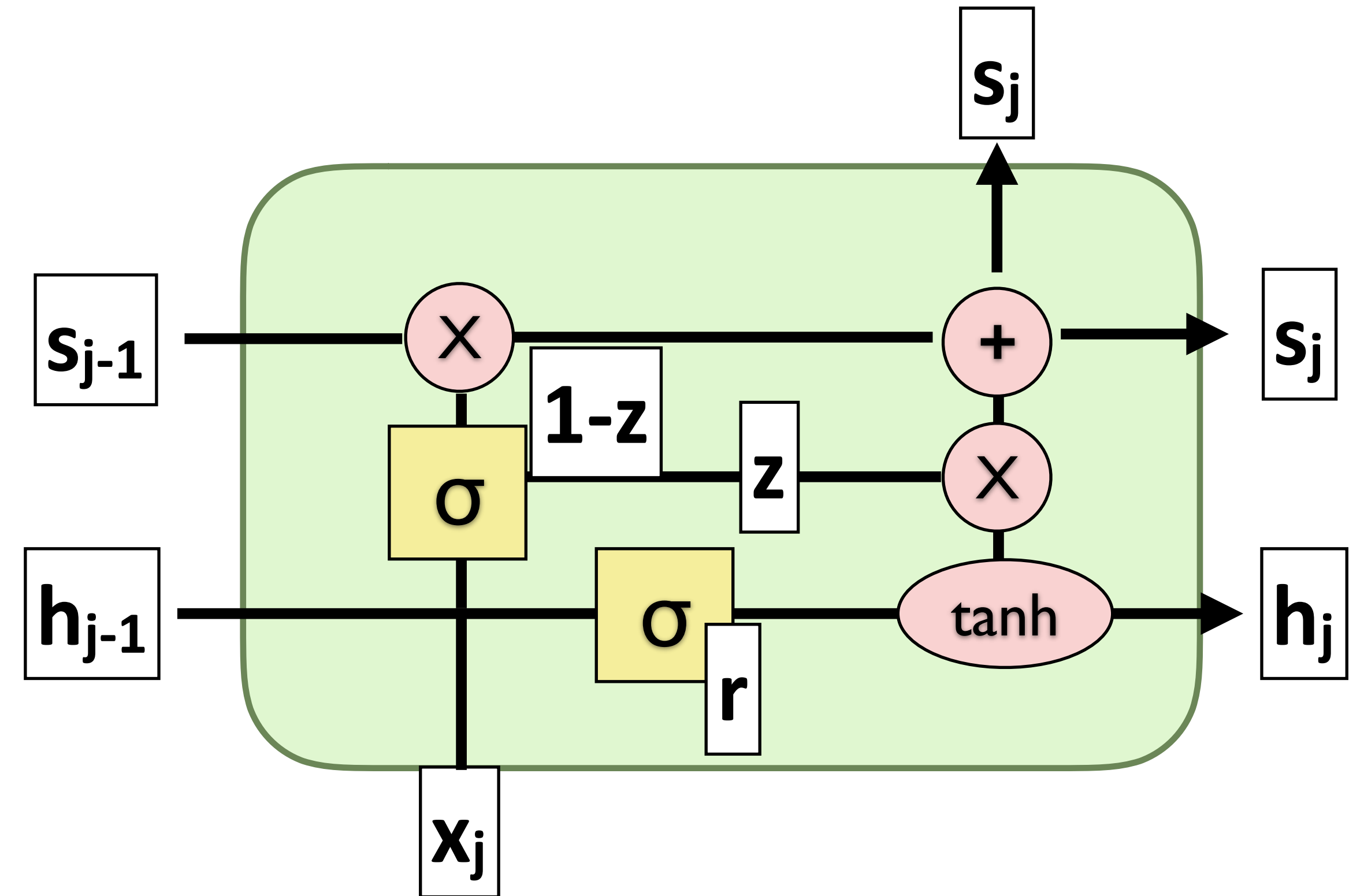


- ▶ LSTM: more complex and slower, may work a bit better

# GRUs



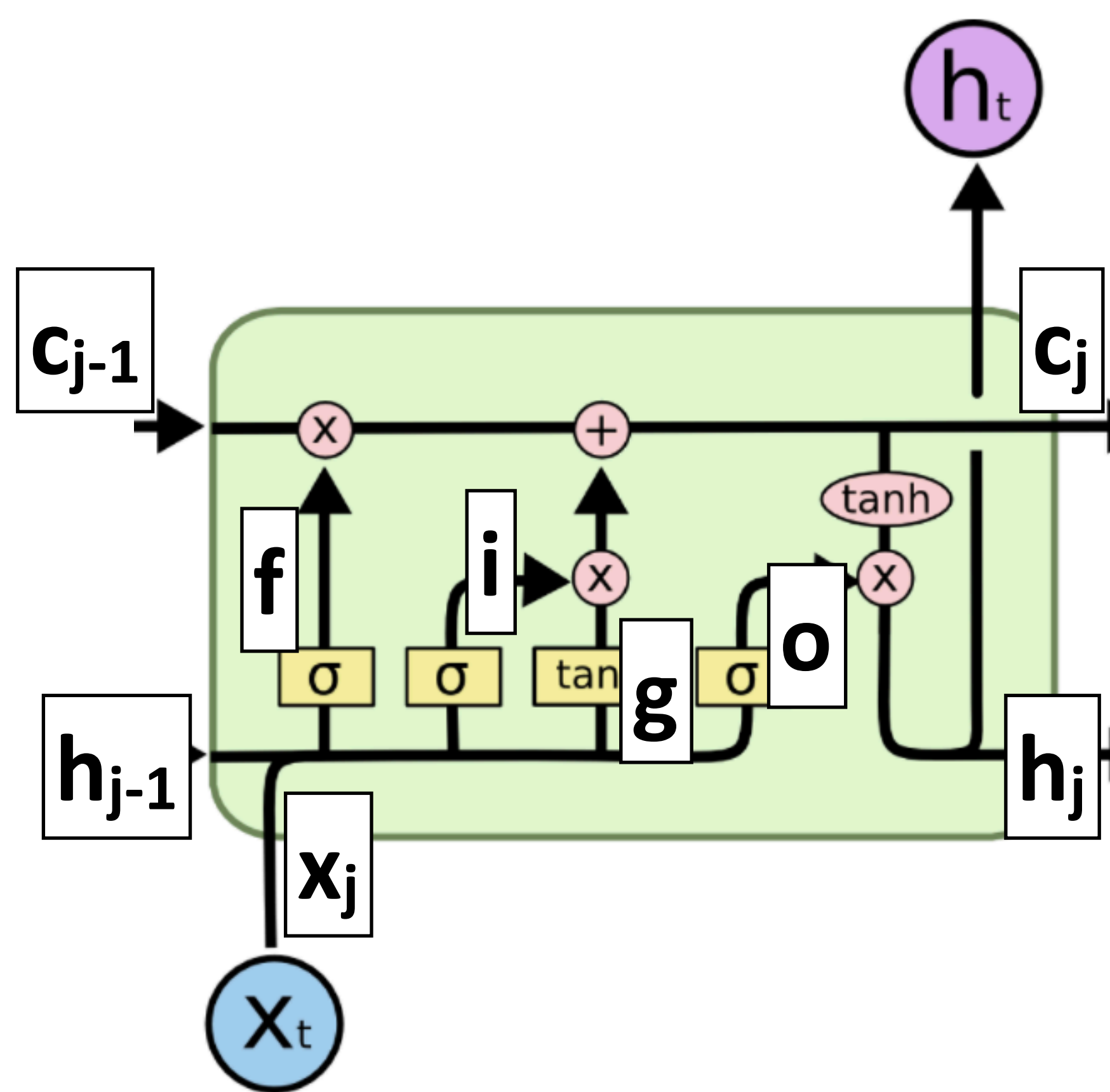
- ▶ LSTM: more complex and slower, may work a bit better



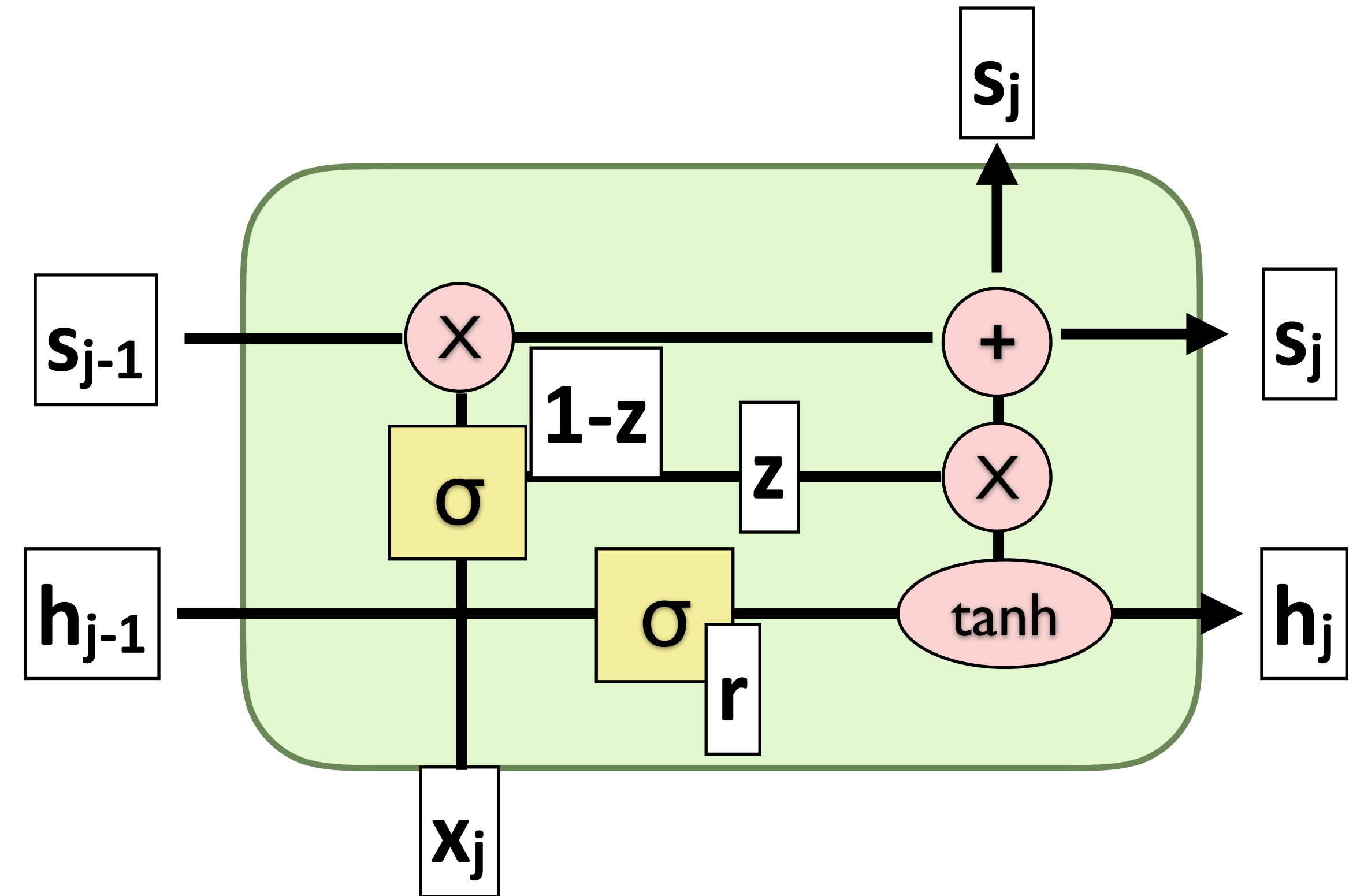
- ▶ GRU: faster, a bit simpler



# GRUs



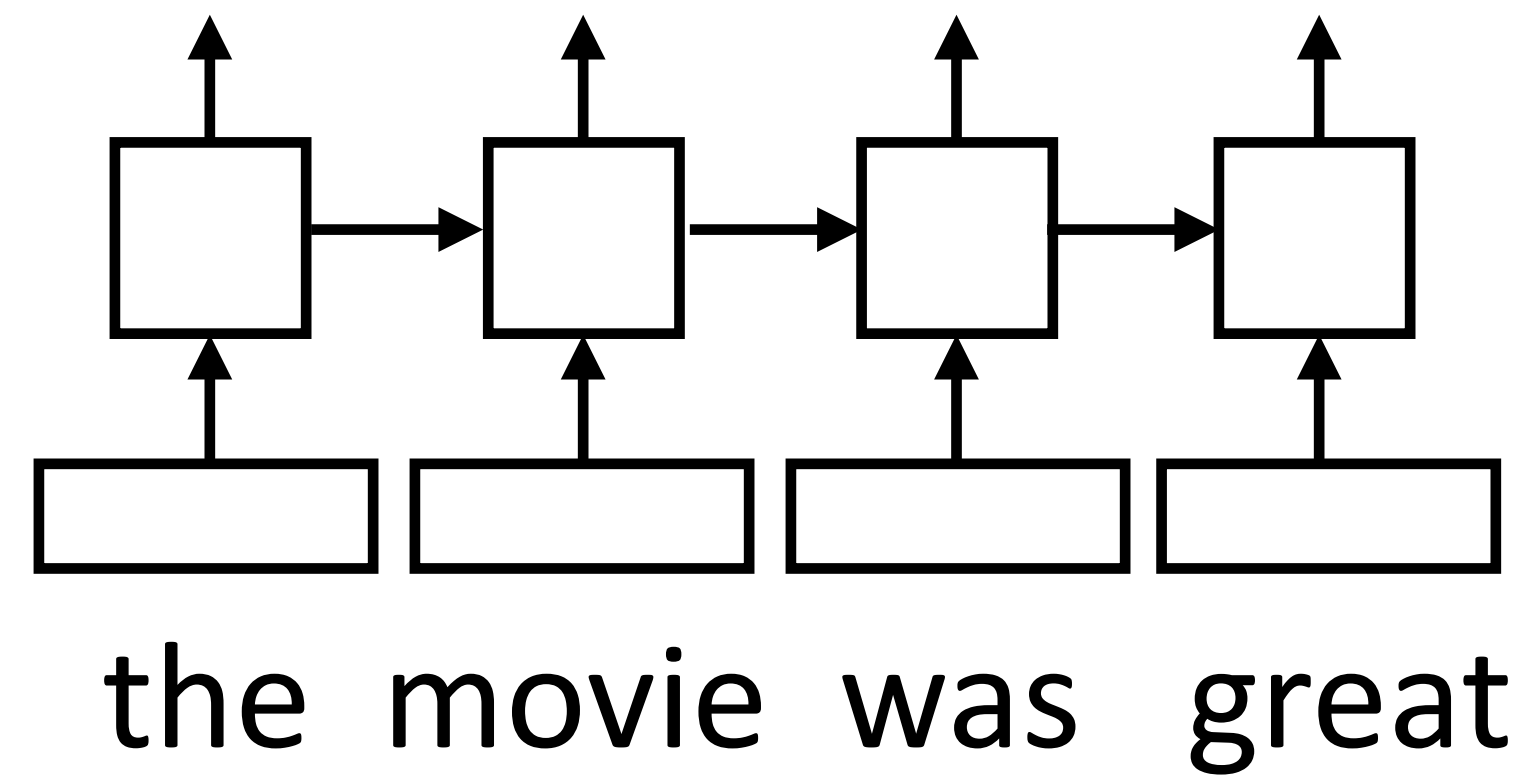
- ▶ LSTM: more complex and slower, may work a bit better



- ▶ GRU: faster, a bit simpler
- ▶ Two gates:  $z$  (forget, mixes  $s$  and  $h$ ) and  $r$  (mixes  $h$  and  $x$ )

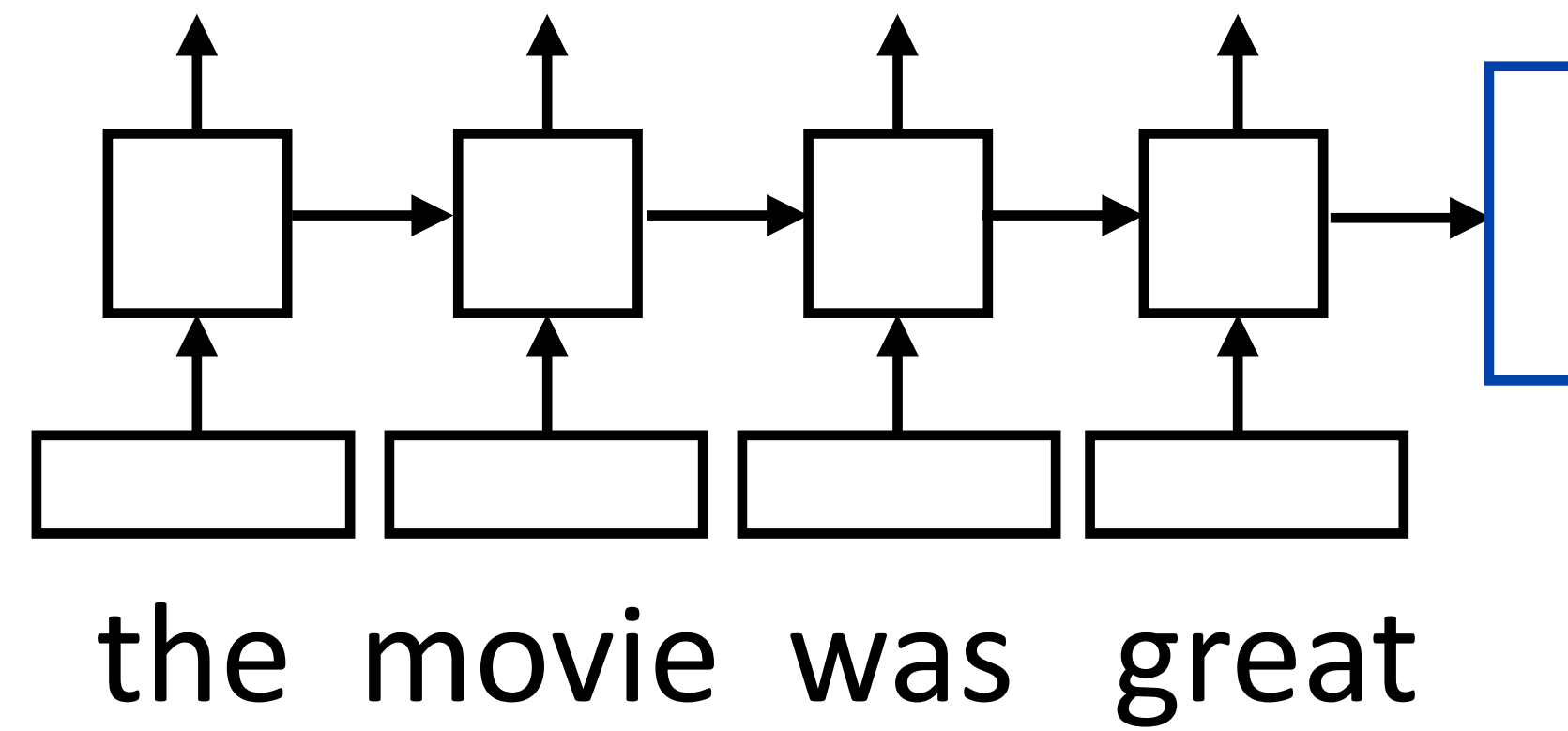
# What do RNNs produce?

---



# What do RNNs produce?

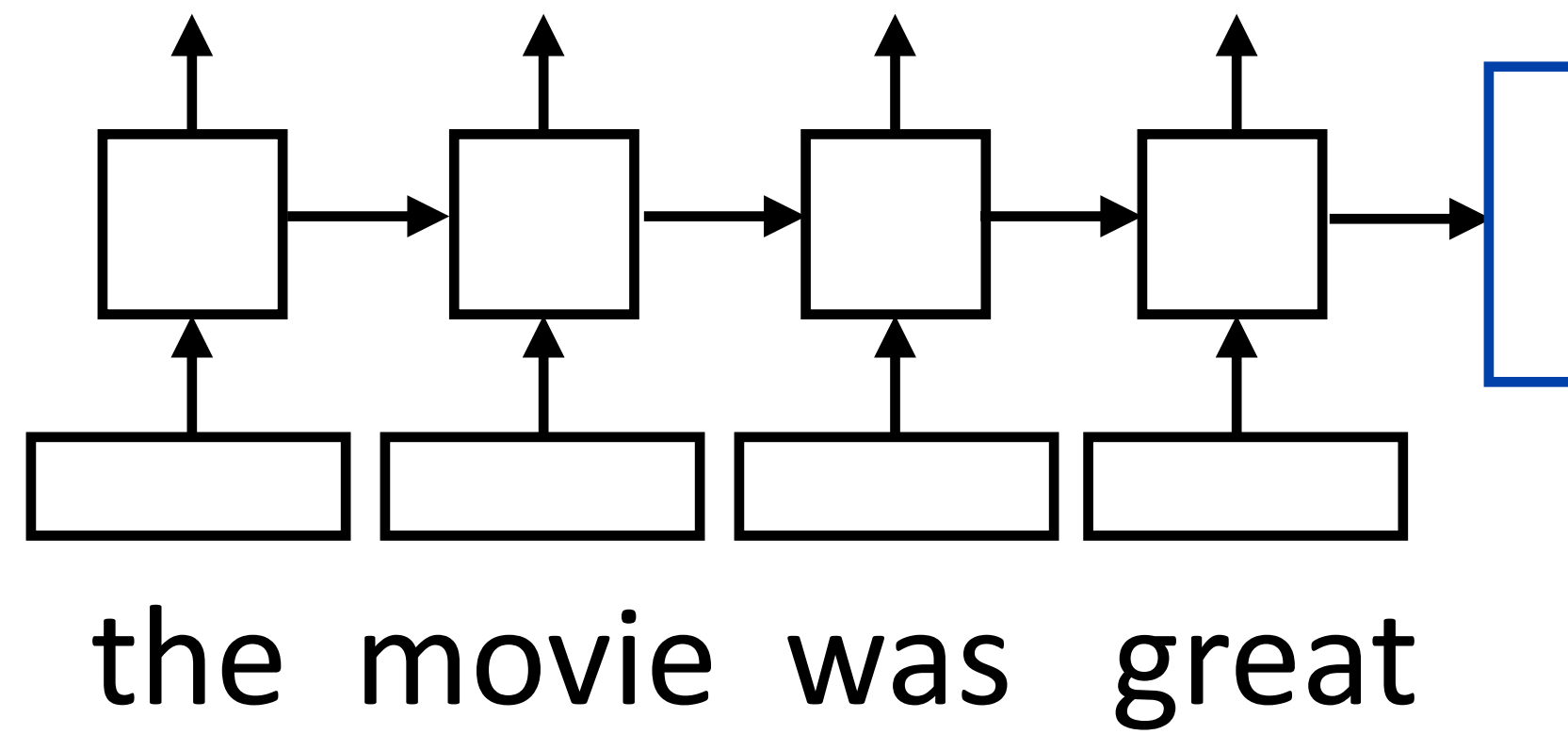
---





# What do RNNs produce?

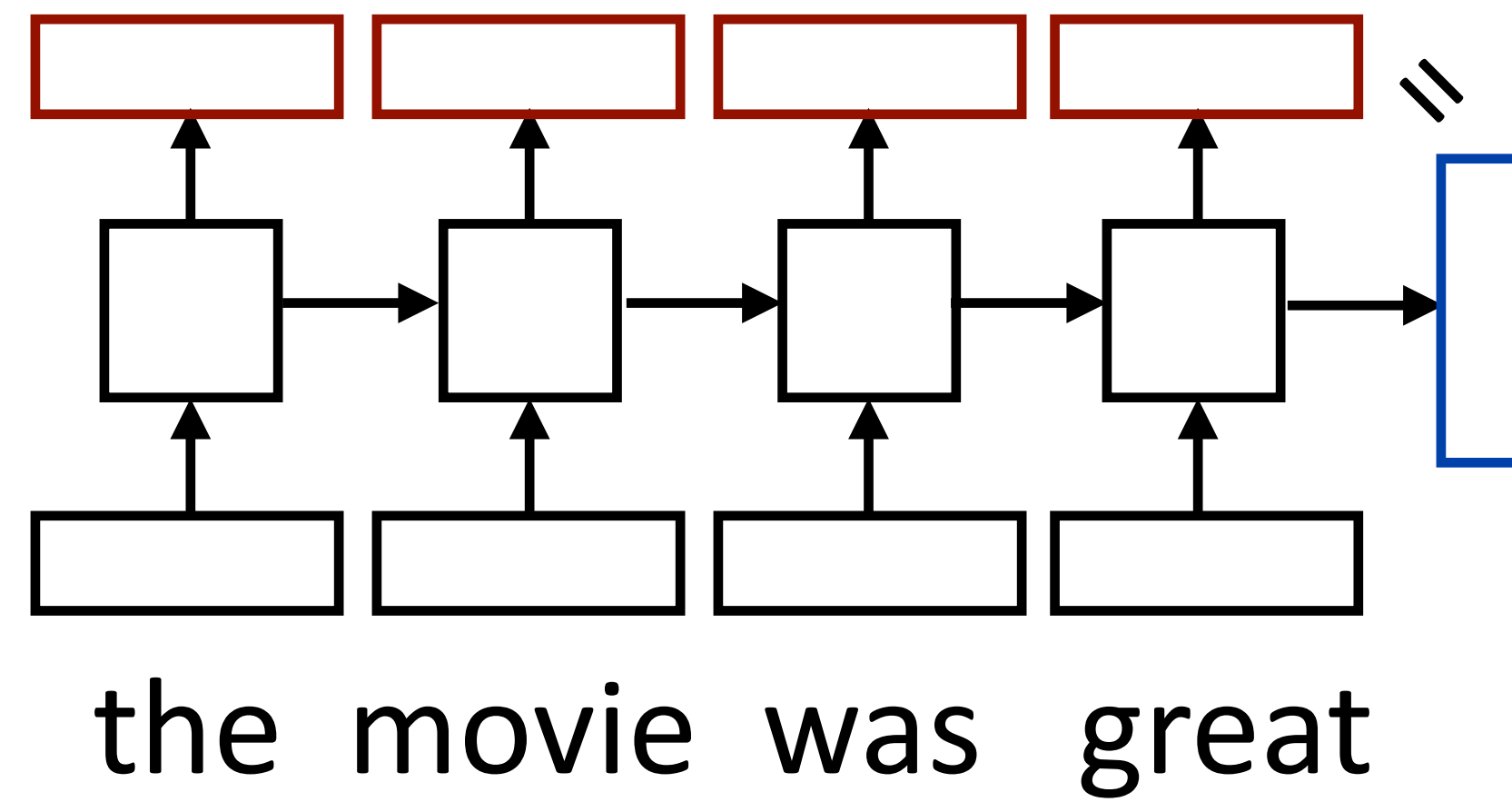
---



- **Encoding of the sentence** — can pass this a decoder or make a classification decision about the sentence

# What do RNNs produce?

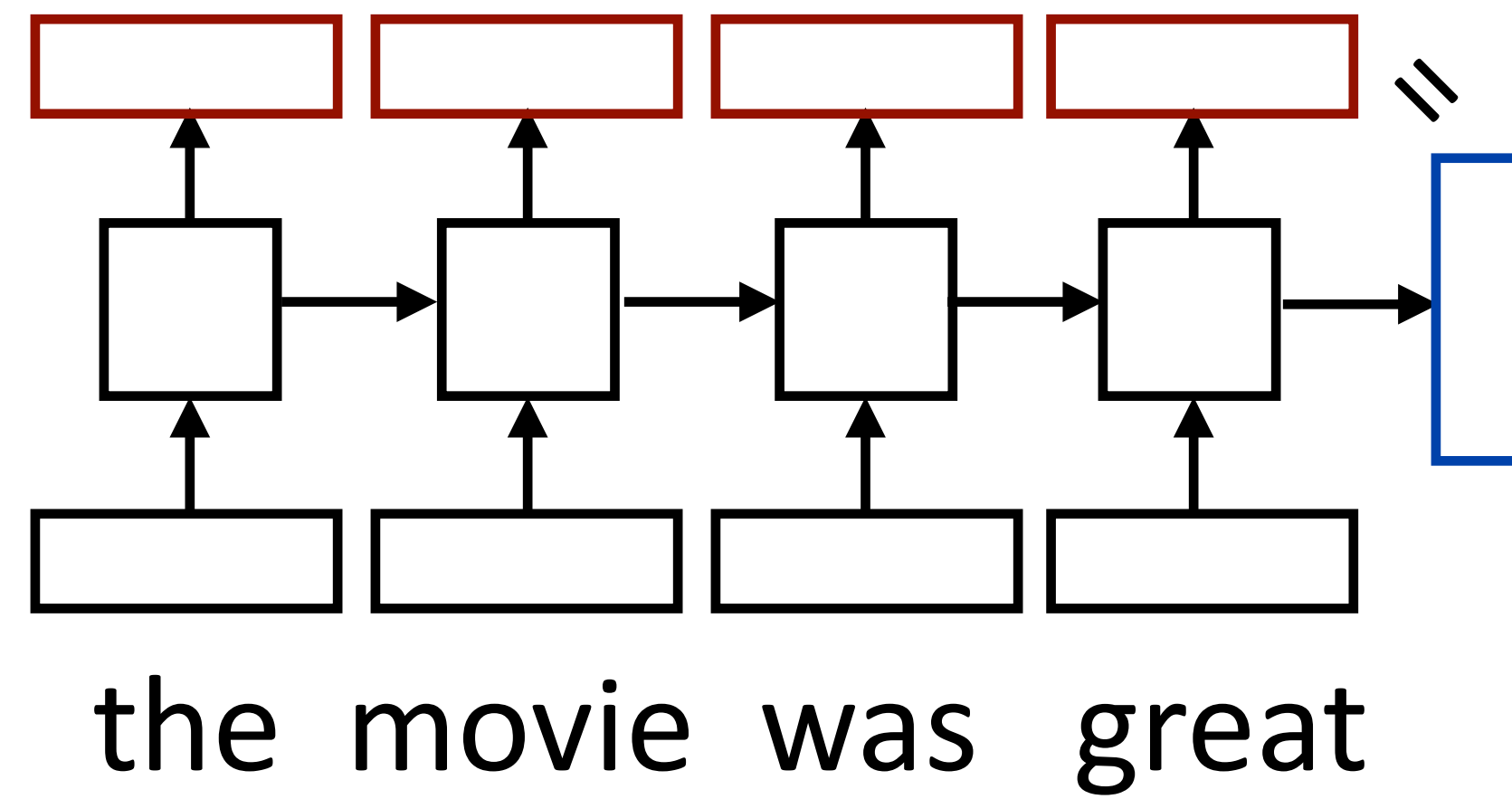
---



- **Encoding of the sentence** — can pass this a decoder or make a classification decision about the sentence

# What do RNNs produce?

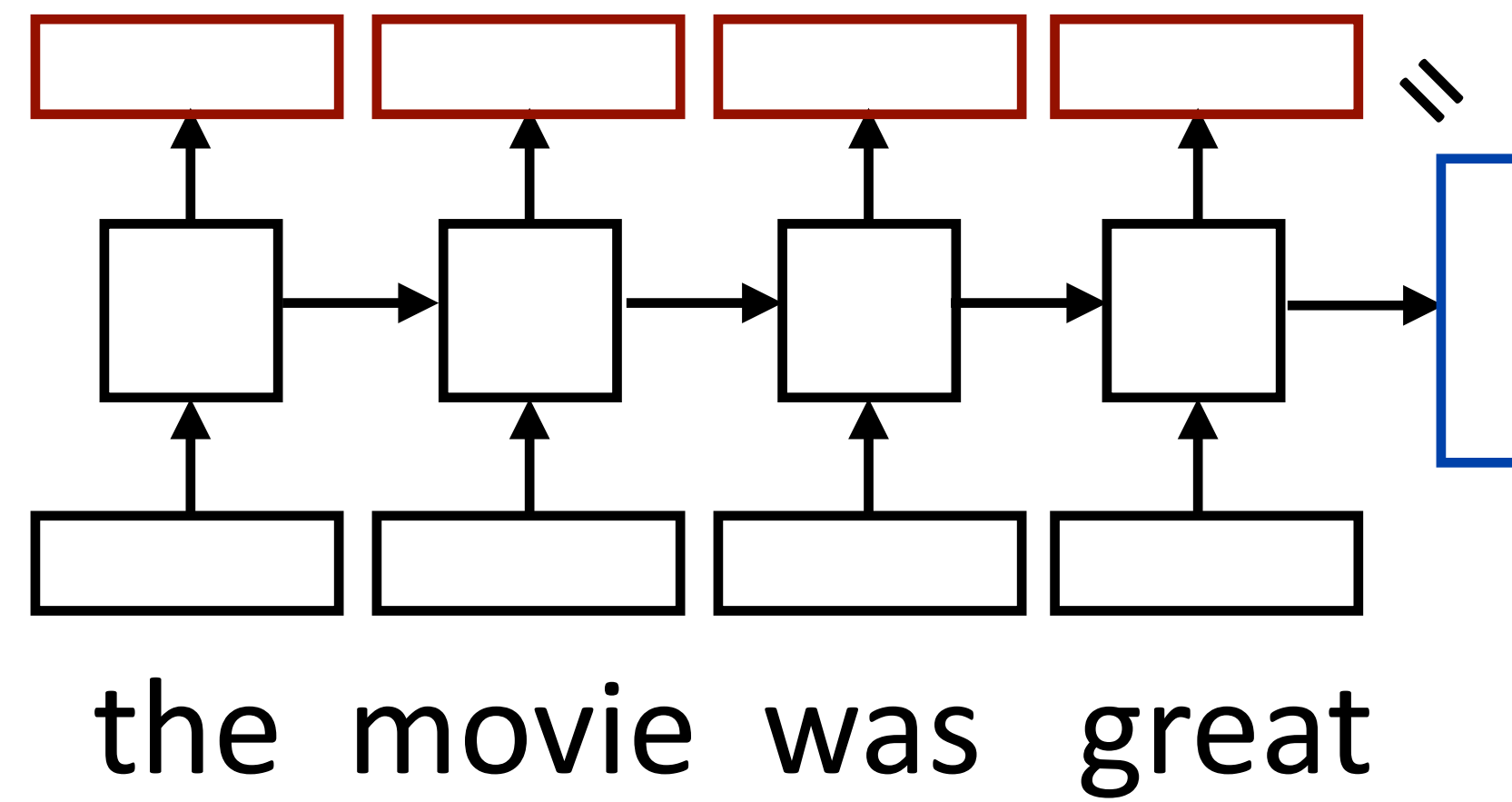
---



- ▶ **Encoding of the sentence** — can pass this a decoder or make a classification decision about the sentence
- ▶ **Encoding of each word** — can pass this to another layer to make a prediction (can also pool these to get a different sentence encoding)

# What do RNNs produce?

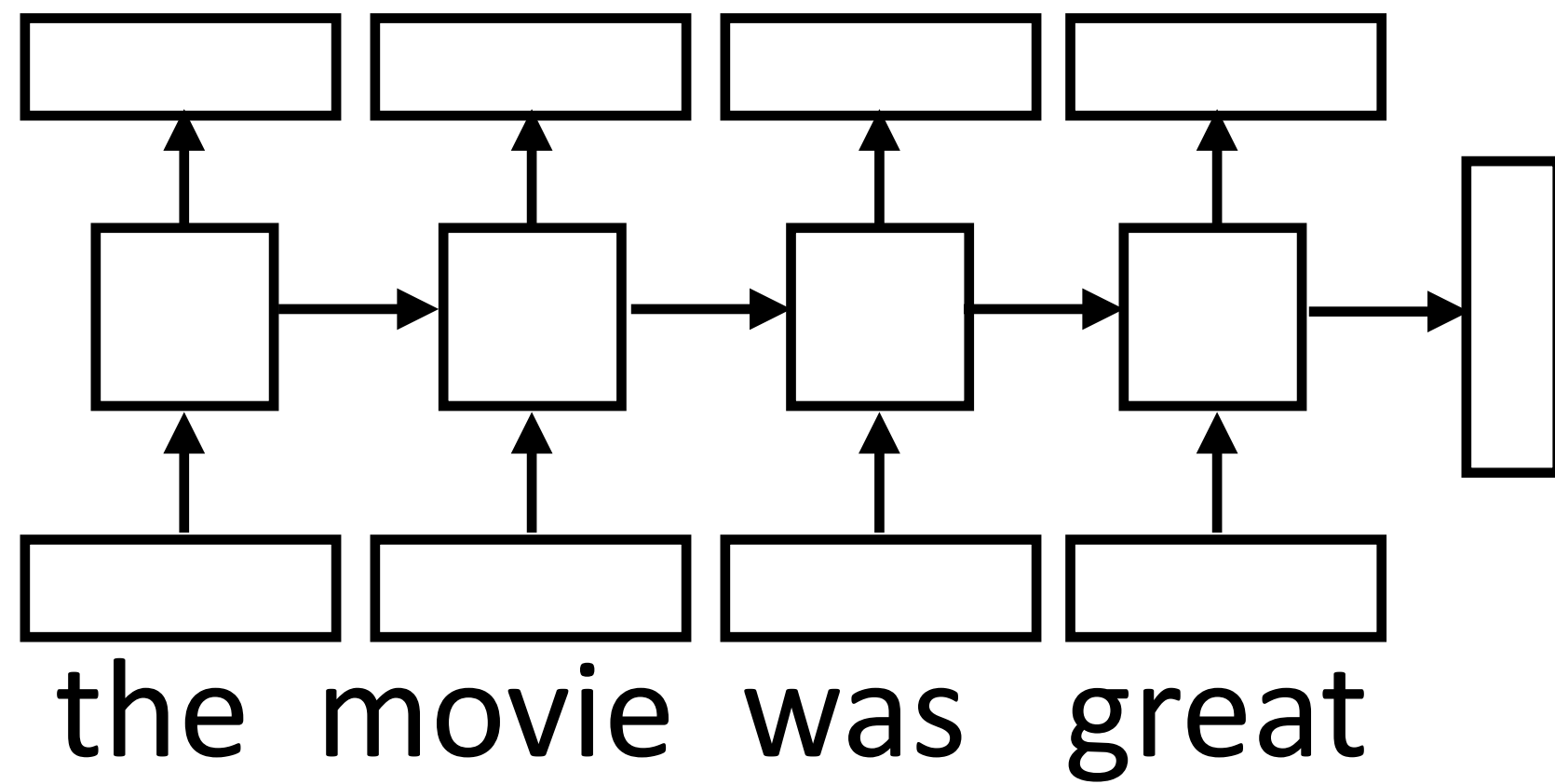
---



- ▶ **Encoding of the sentence** — can pass this a decoder or make a classification decision about the sentence
- ▶ **Encoding of each word** — can pass this to another layer to make a prediction (can also pool these to get a different sentence encoding)
- ▶ RNN can be viewed as a transformation of a sequence of vectors into a sequence of context-dependent vectors

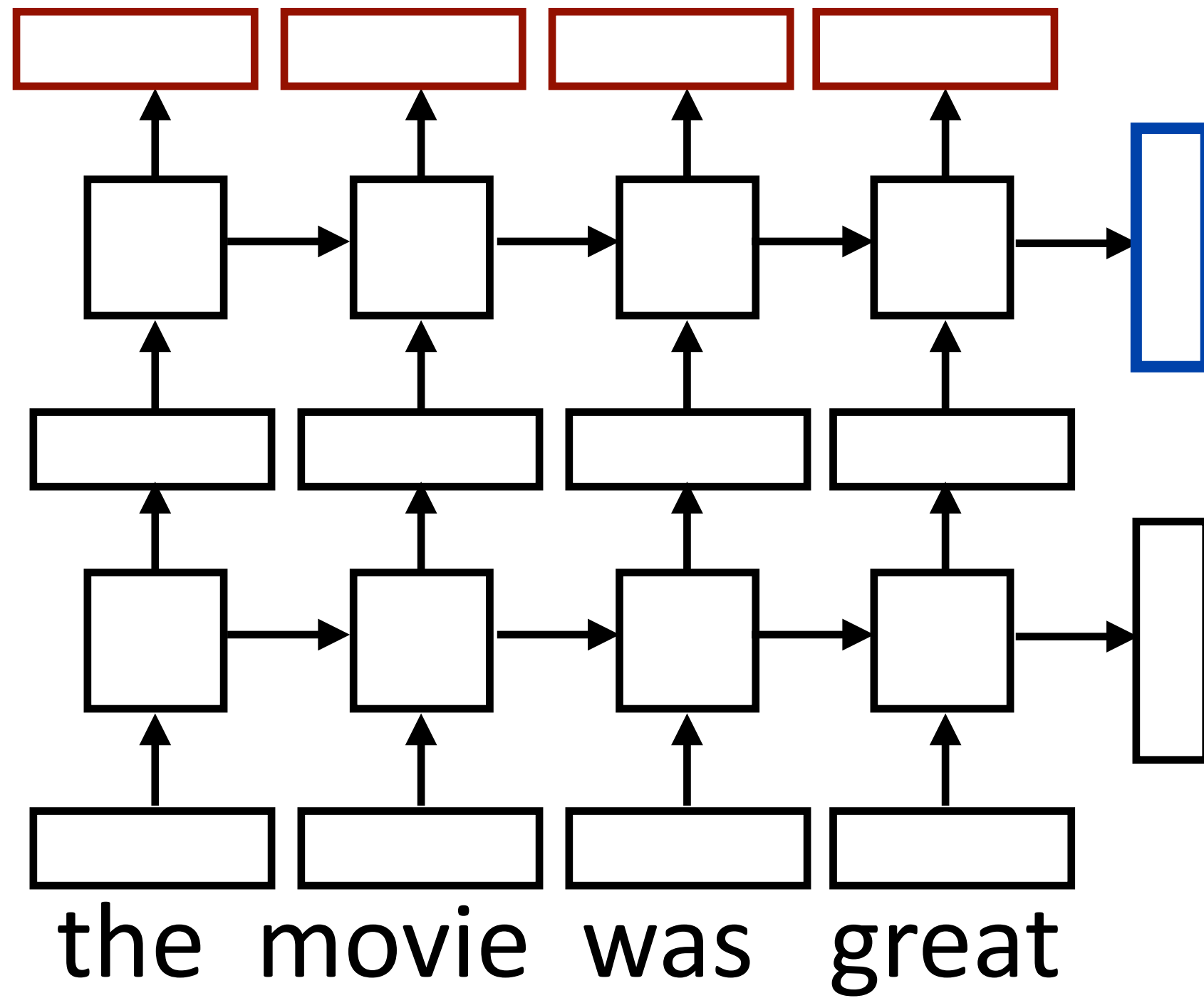
# Multilayer Bidirectional RNN

---

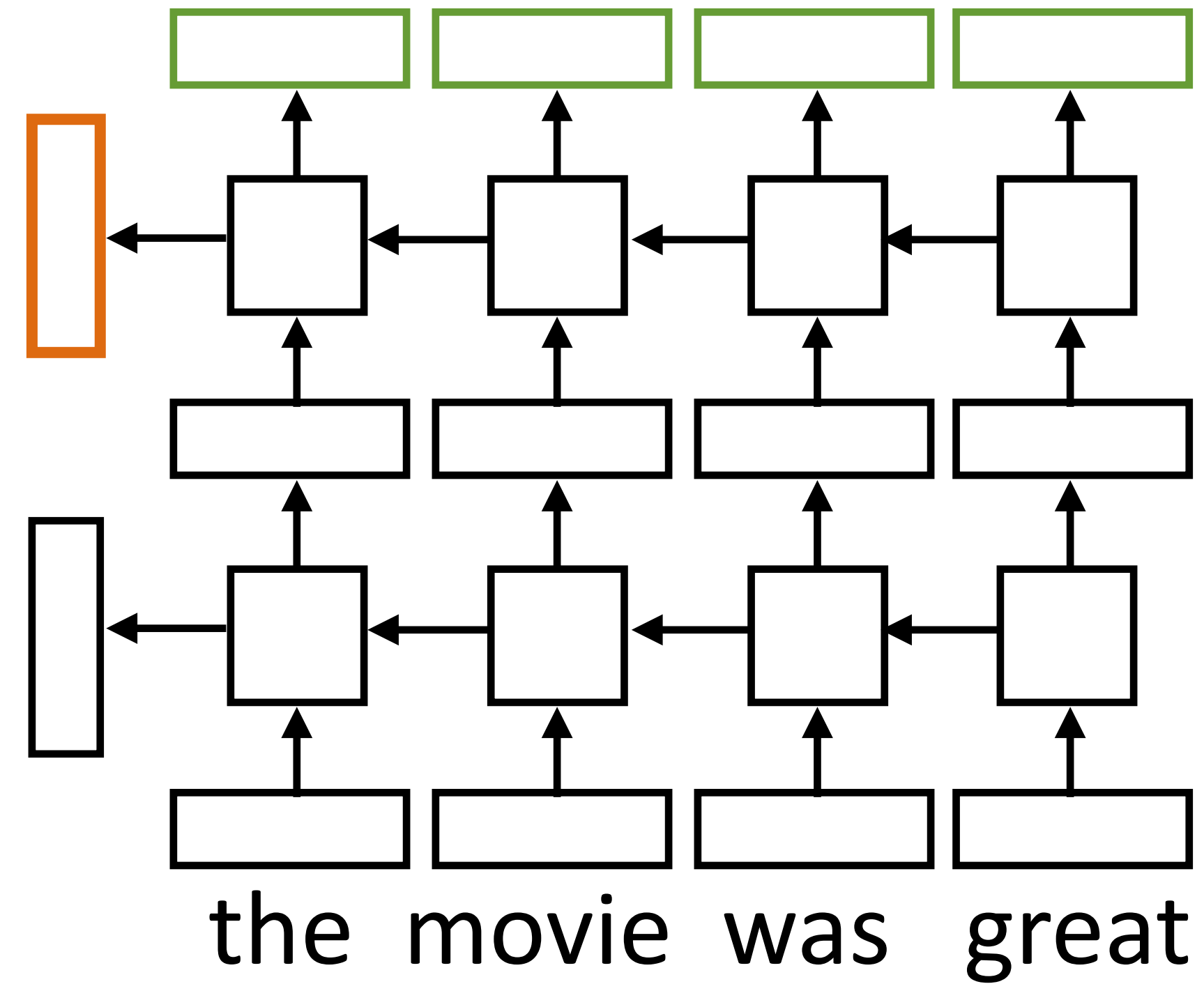
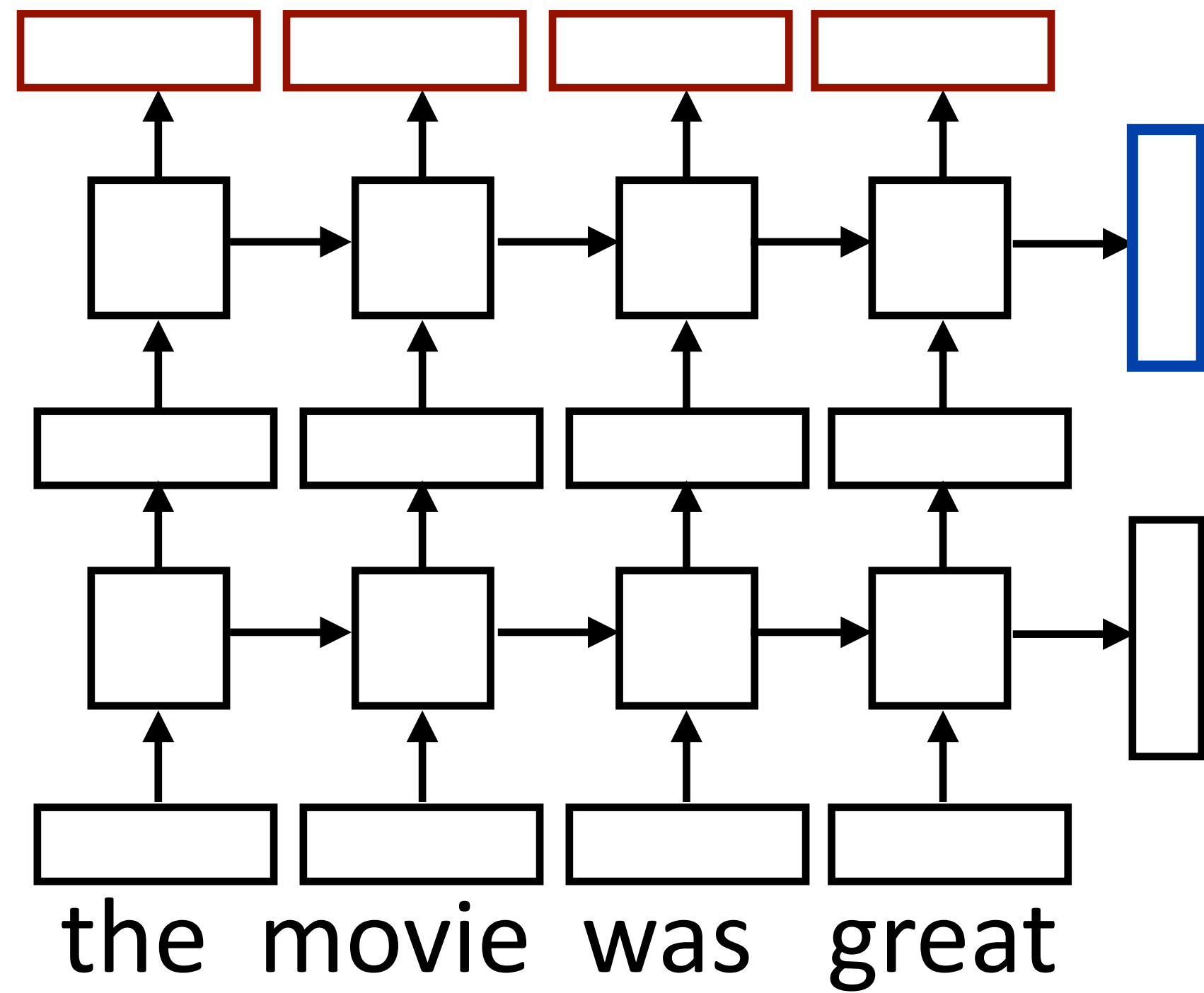


# Multilayer Bidirectional RNN

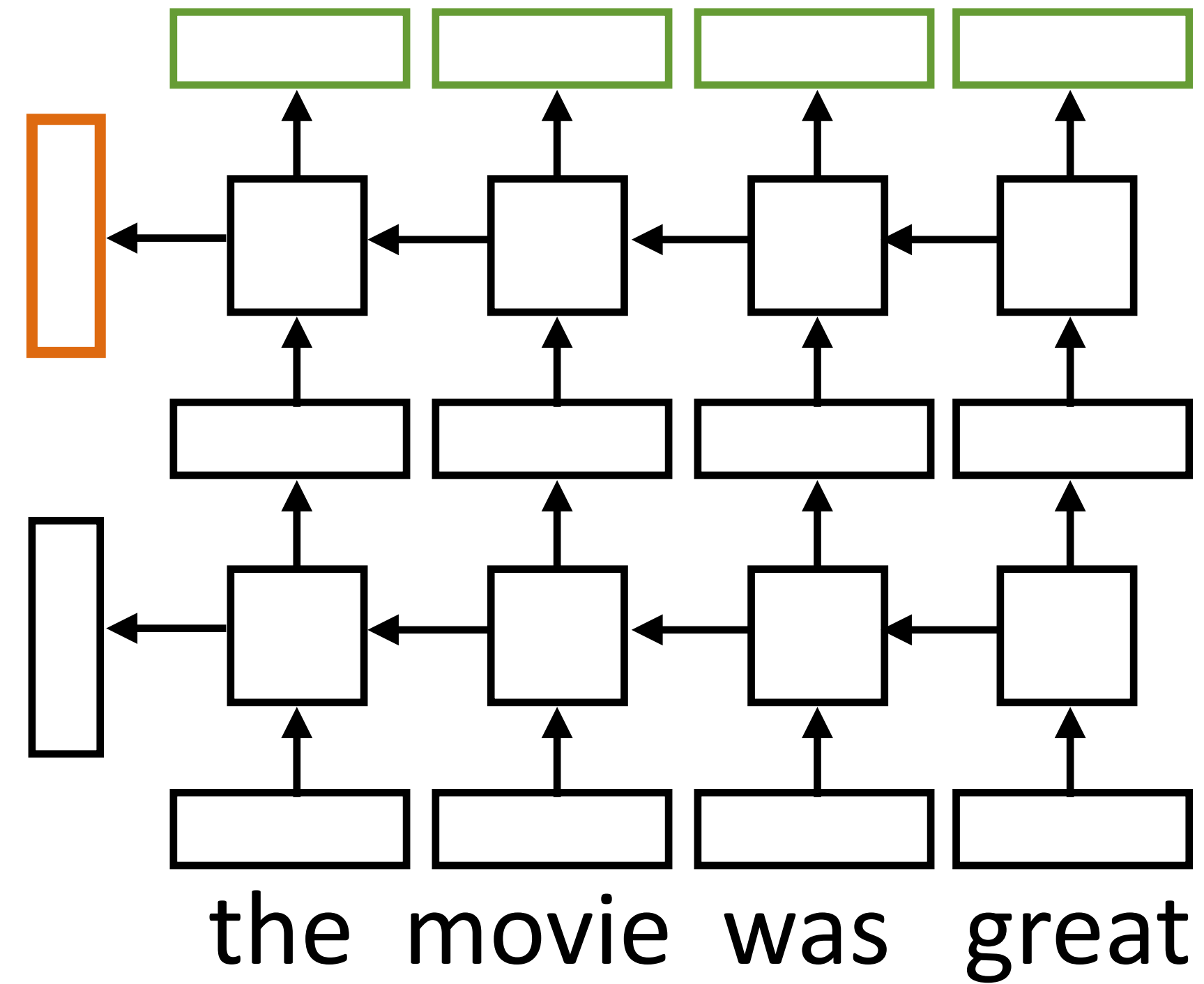
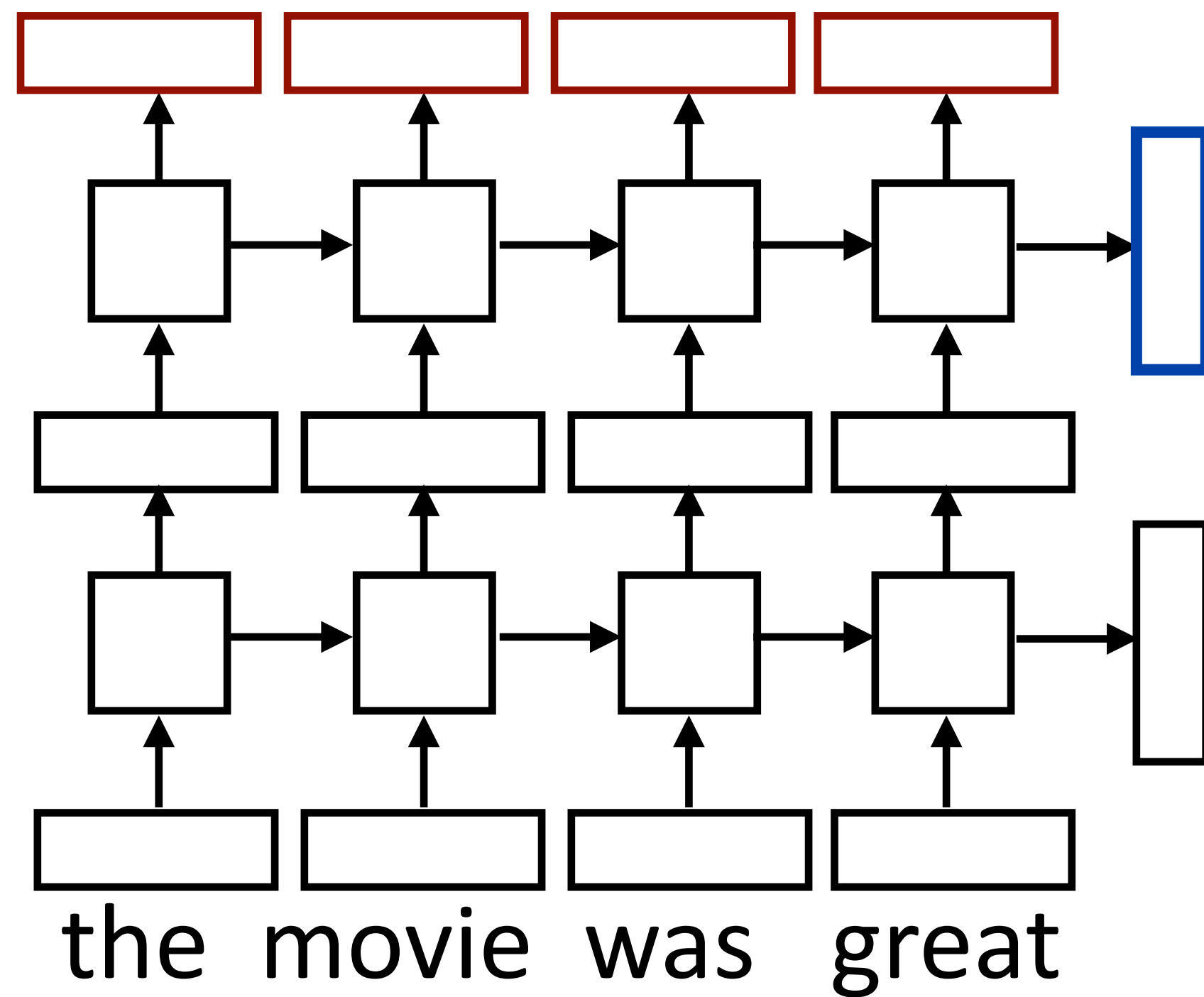
---



# Multilayer Bidirectional RNN



# Multilayer Bidirectional RNN

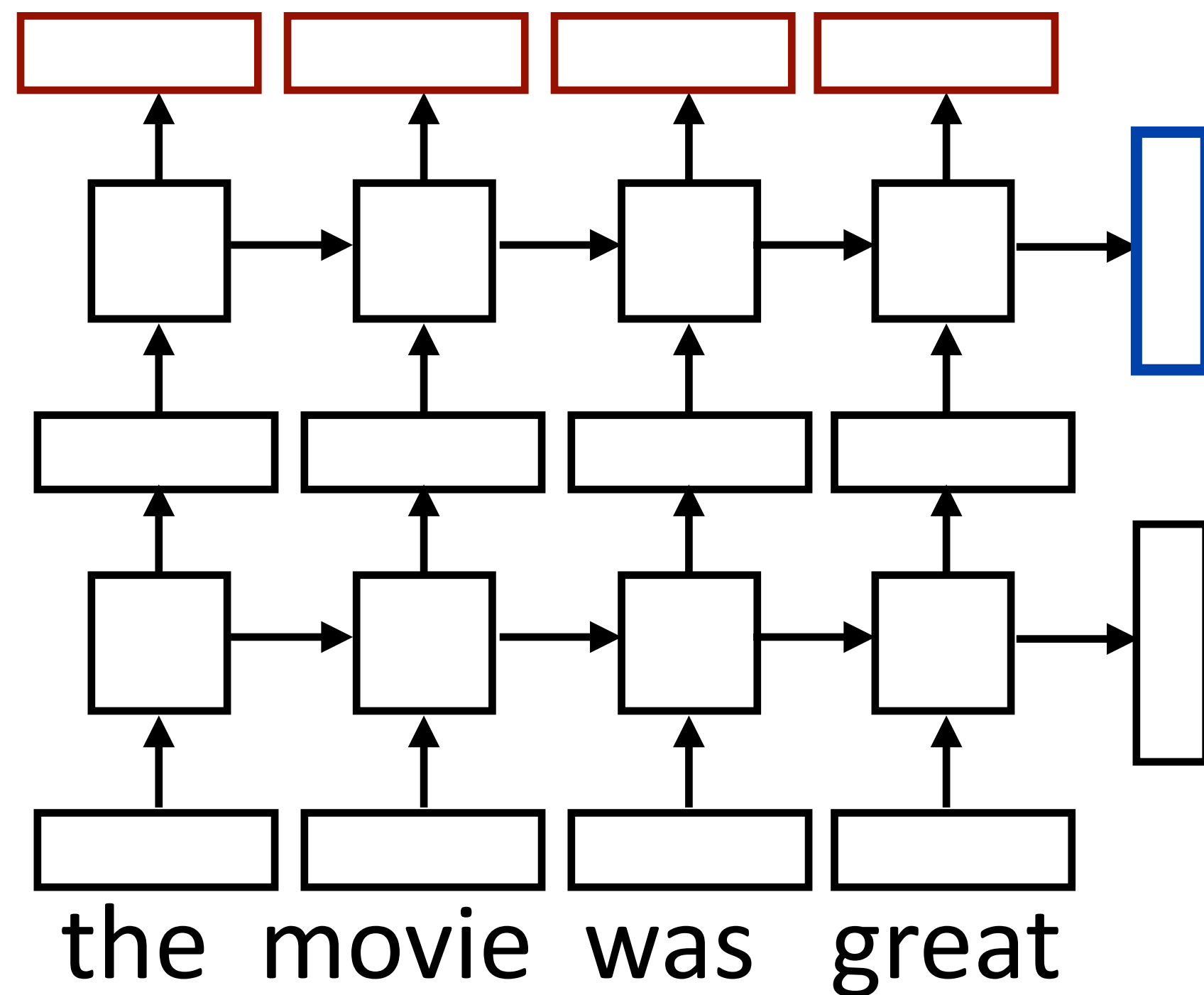


- Sentence classification based on concatenation of both final outputs

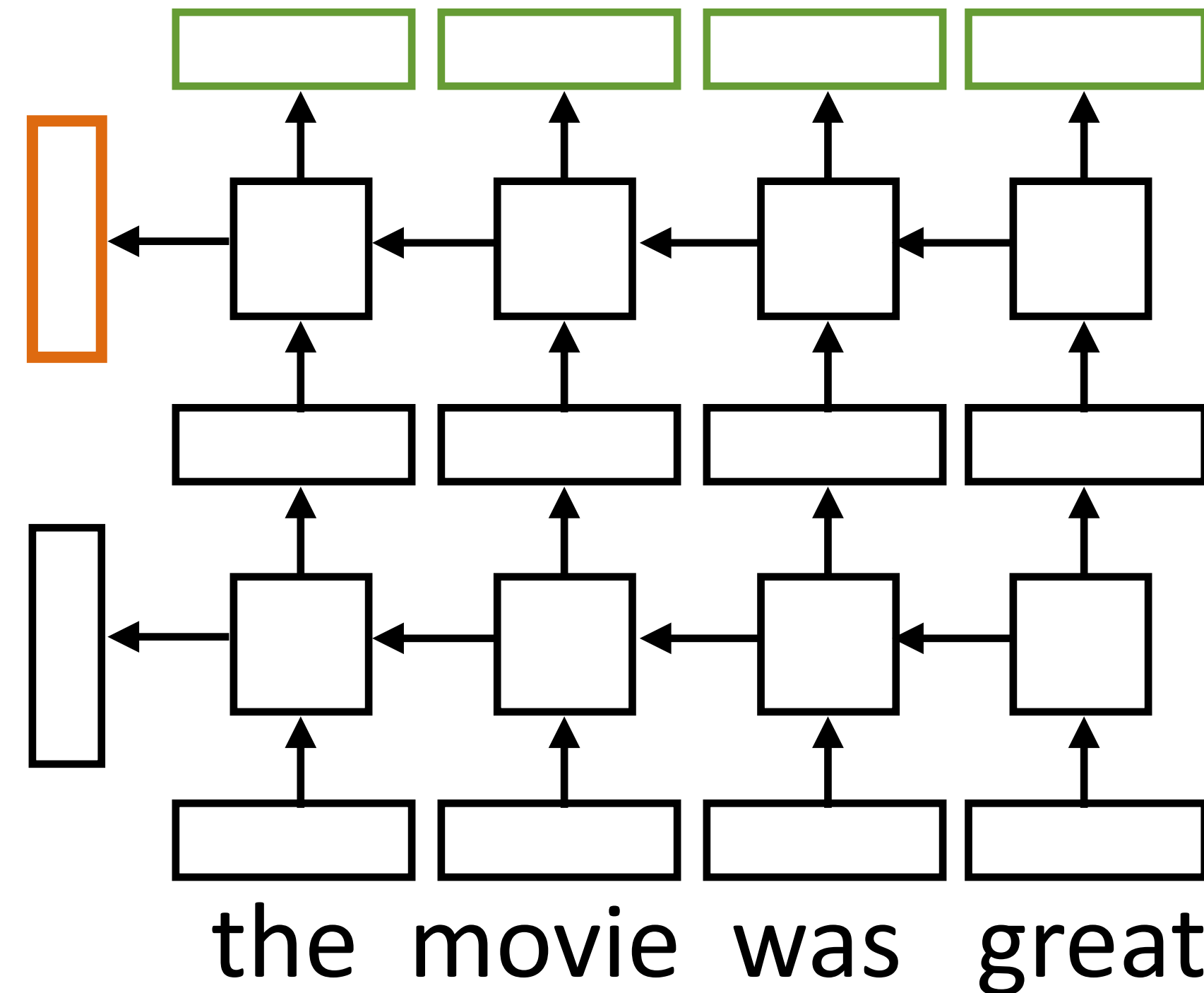




# Multilayer Bidirectional RNN



- ▶ Sentence classification based on concatenation of both final outputs

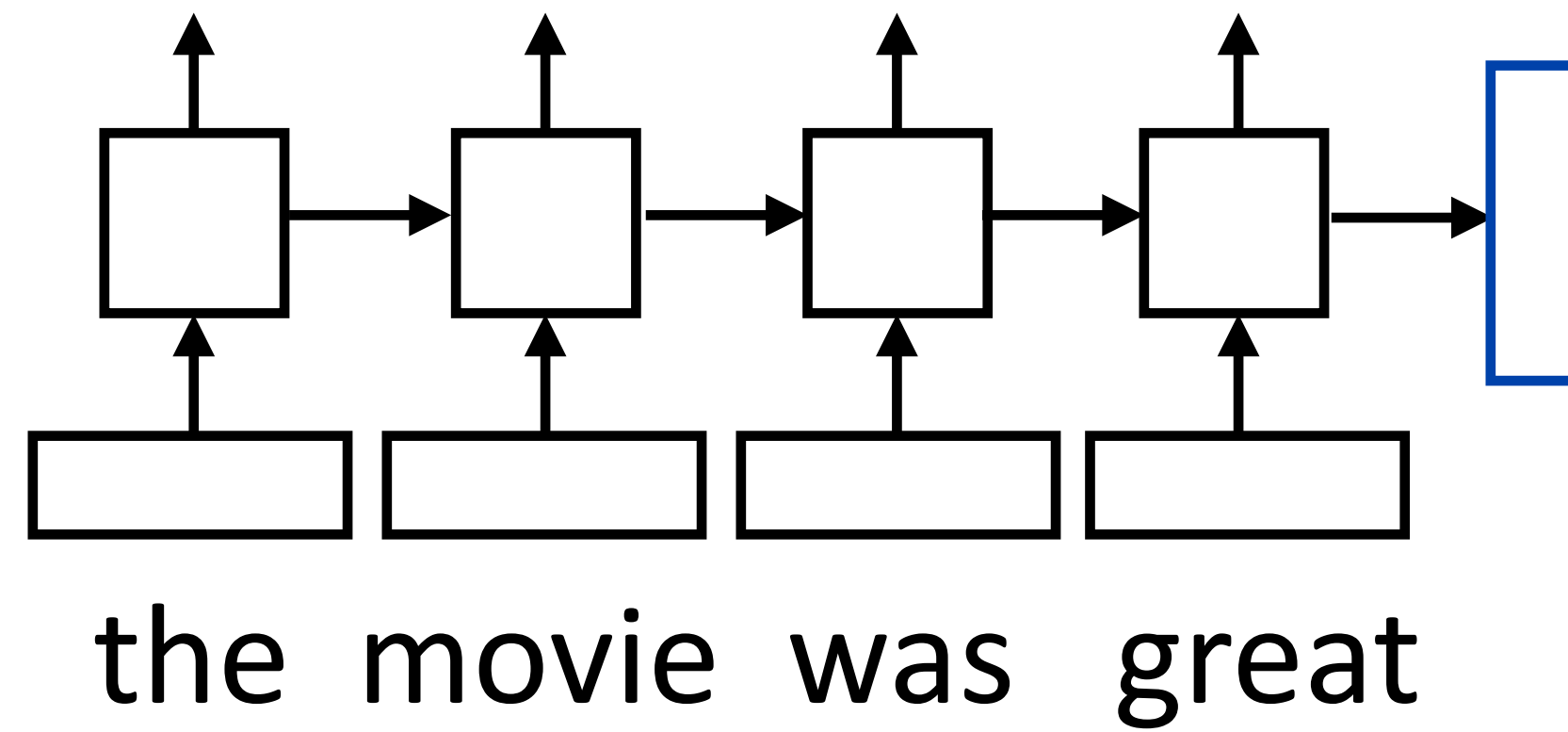


- ▶ Token classification based on concatenation of both directions' token representations



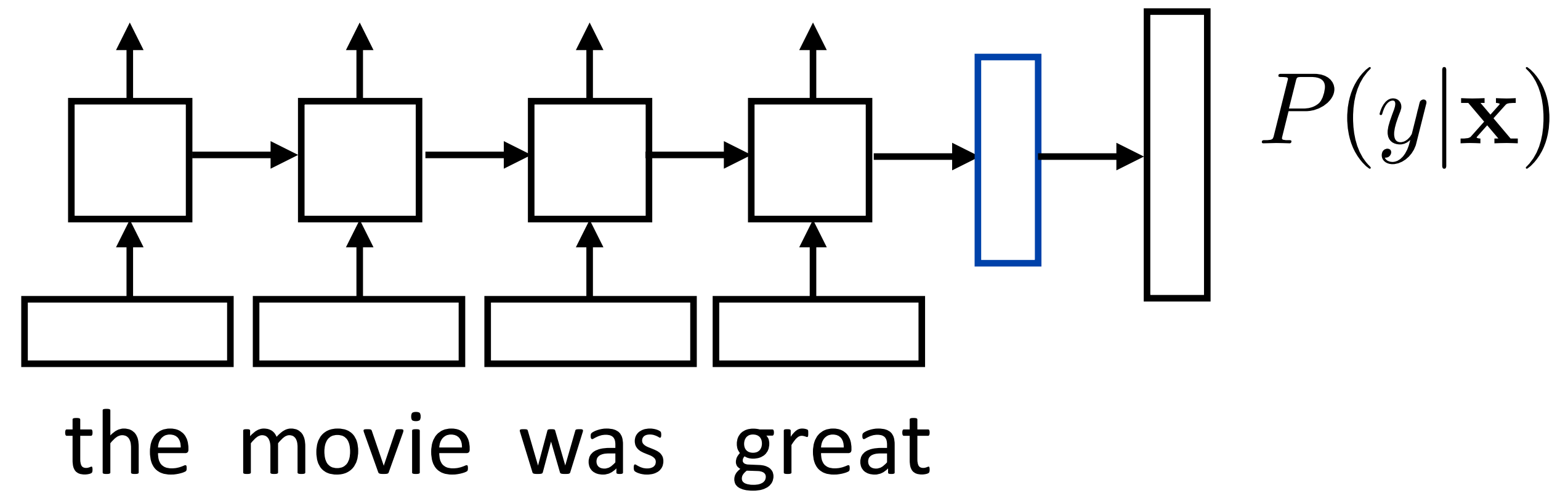
# Training RNNs

---



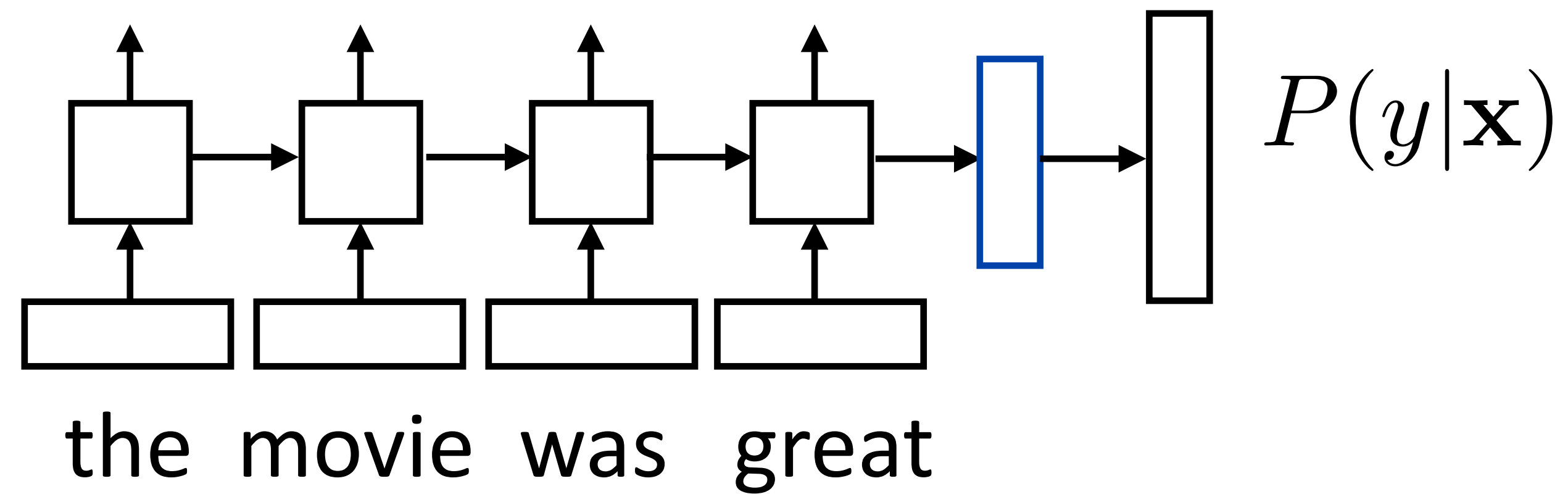
# Training RNNs

---



# Training RNNs

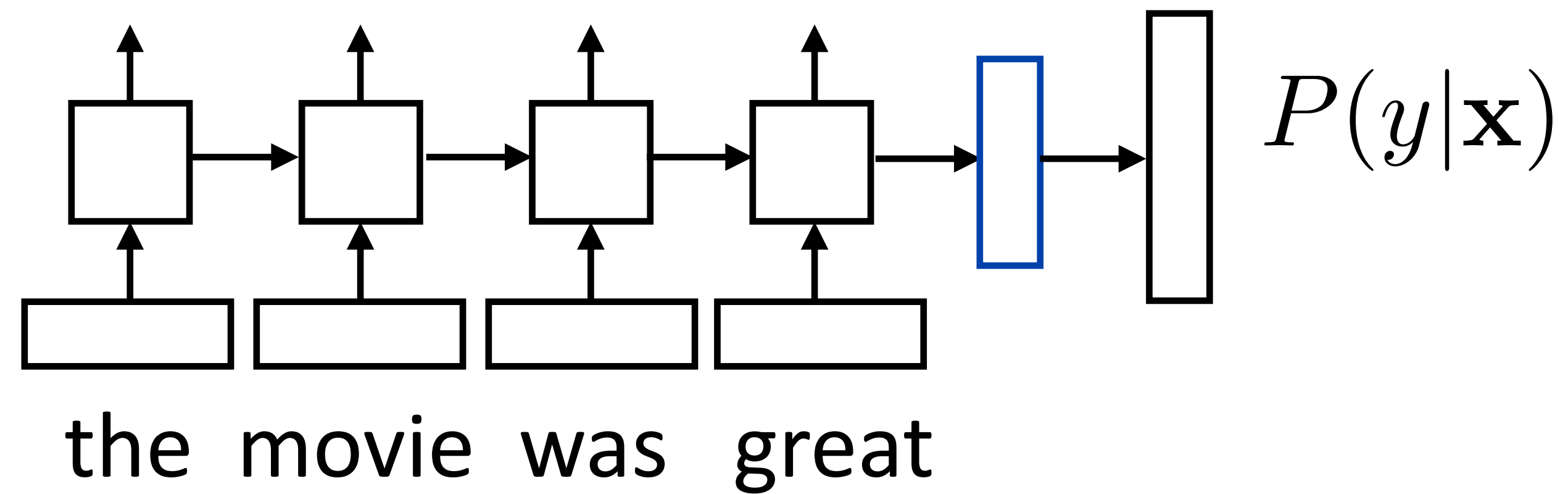
---



- Loss = negative log likelihood of probability of gold label (or use SVM or other loss)

# Training RNNs

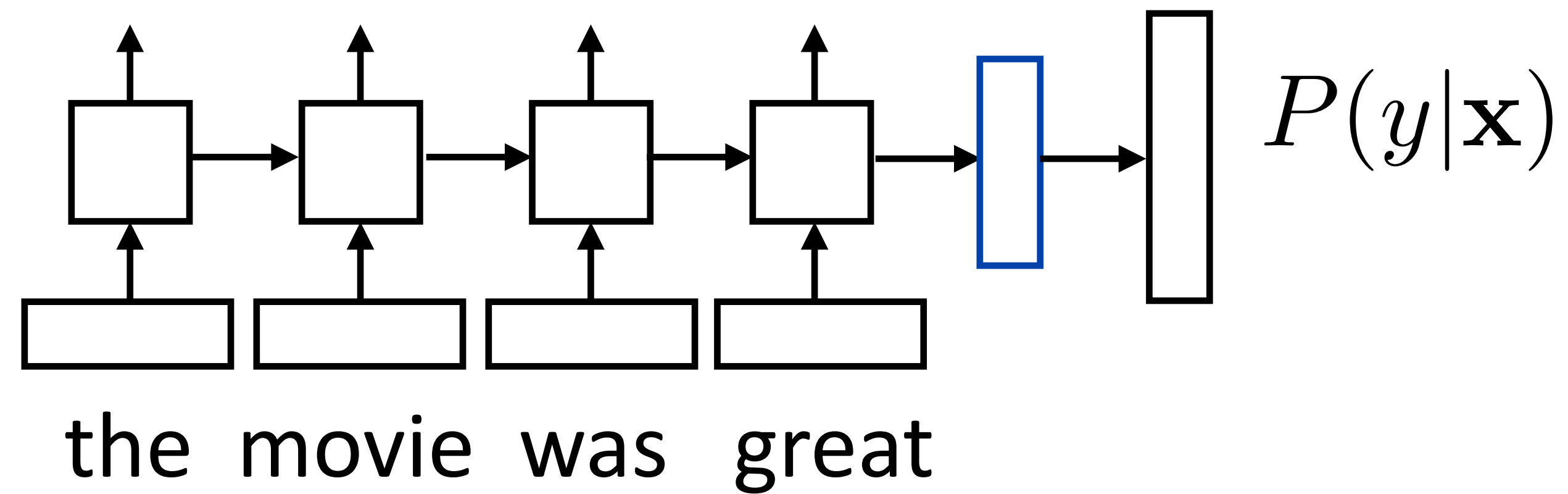
---



- ▶ Loss = negative log likelihood of probability of gold label (or use SVM or other loss)
- ▶ Backpropagate through entire network

# Training RNNs

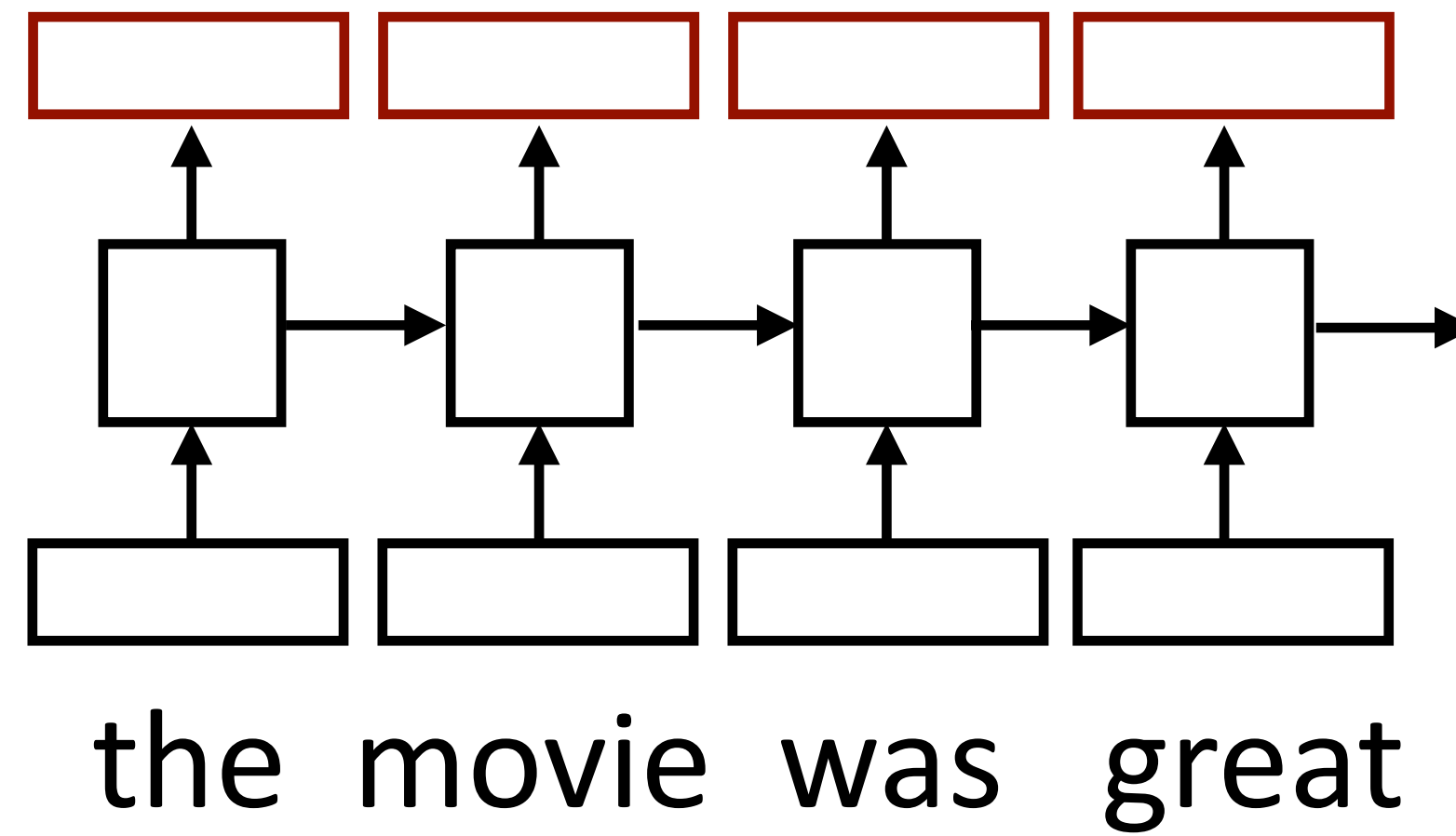
---



- ▶ Loss = negative log likelihood of probability of gold label (or use SVM or other loss)
- ▶ Backpropagate through entire network
- ▶ Example: sentiment analysis

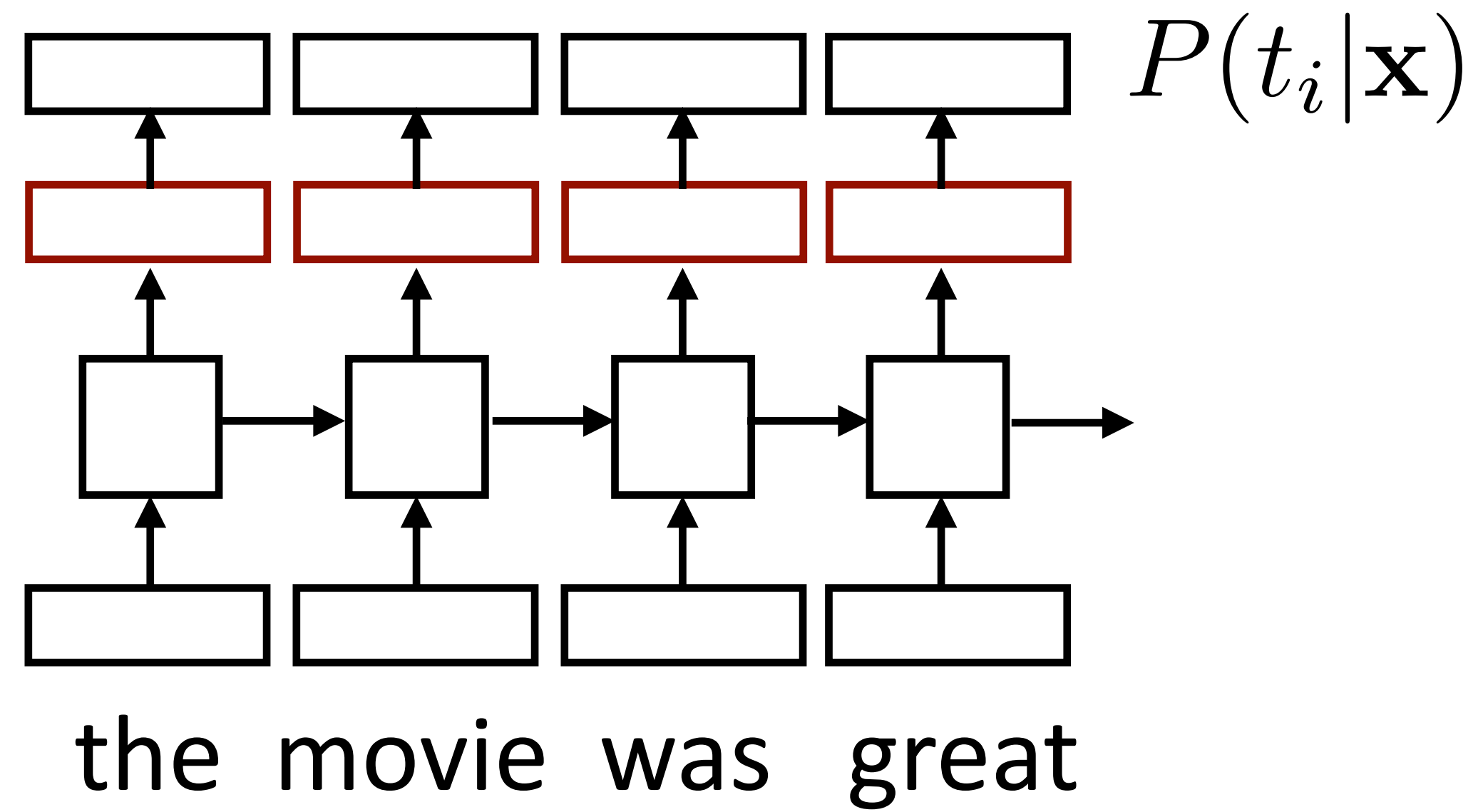
# Training RNNs

---



# Training RNNs

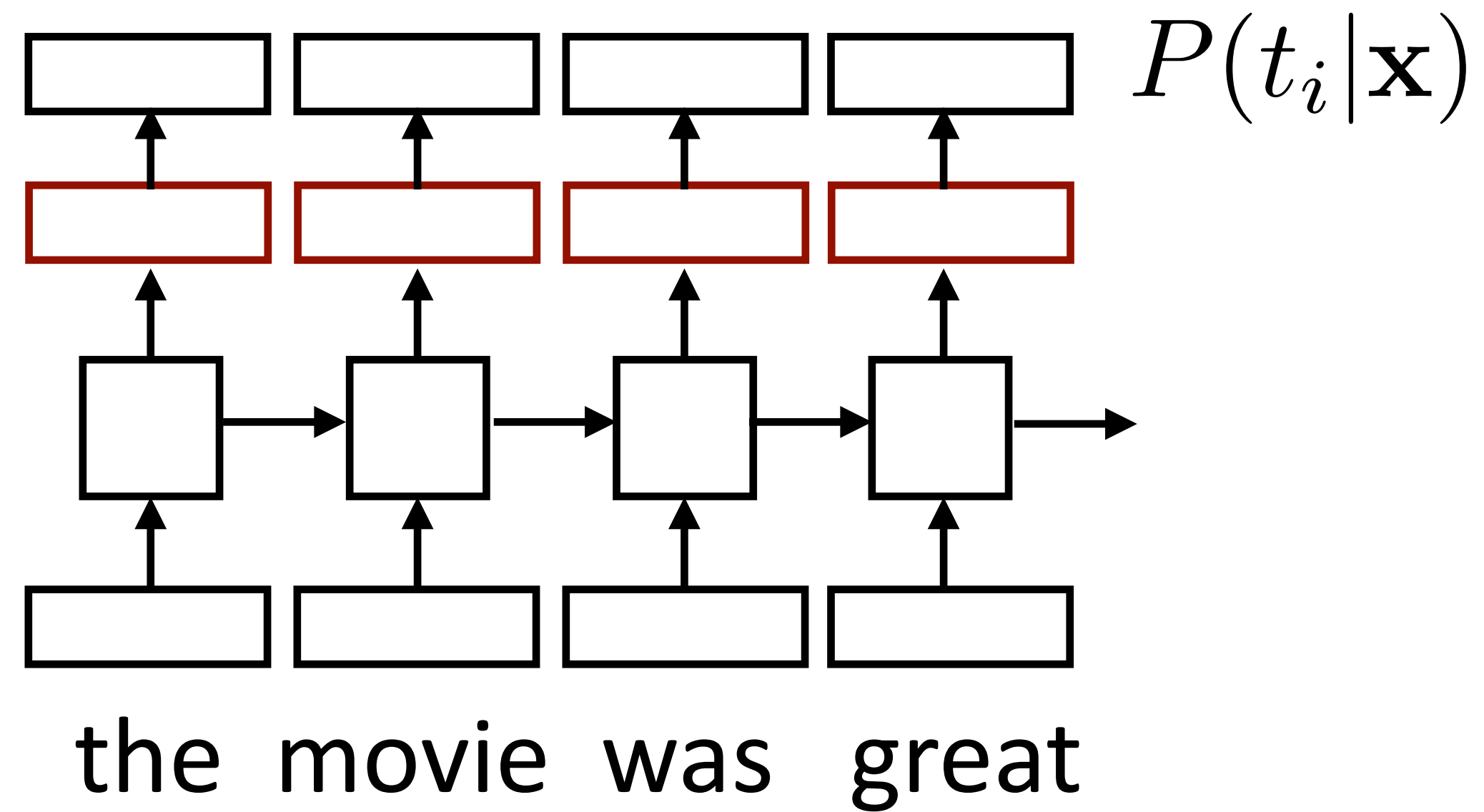
---





# Training RNNs

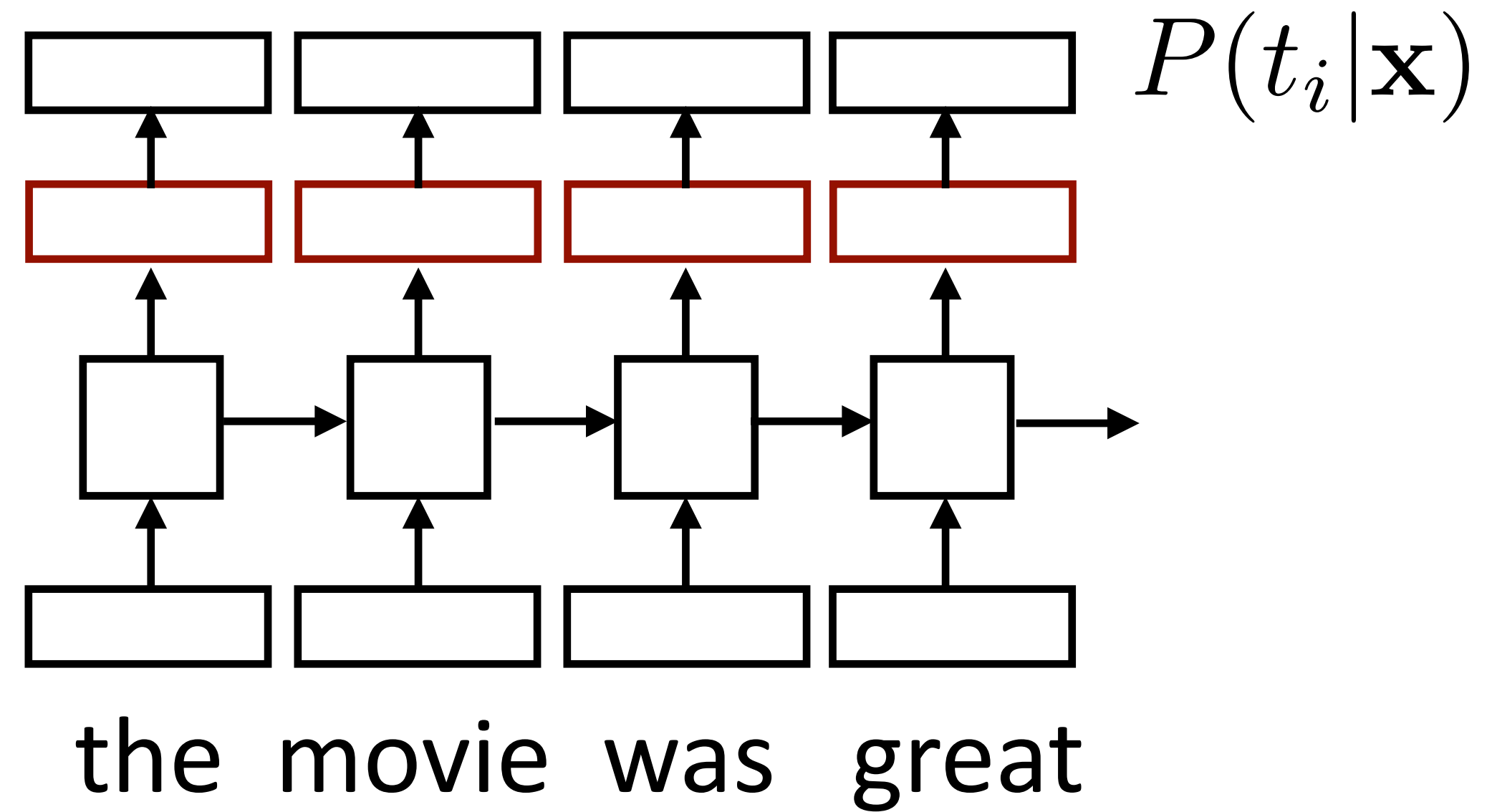
---



- Loss = negative log likelihood of probability of gold predictions, summed over the tags

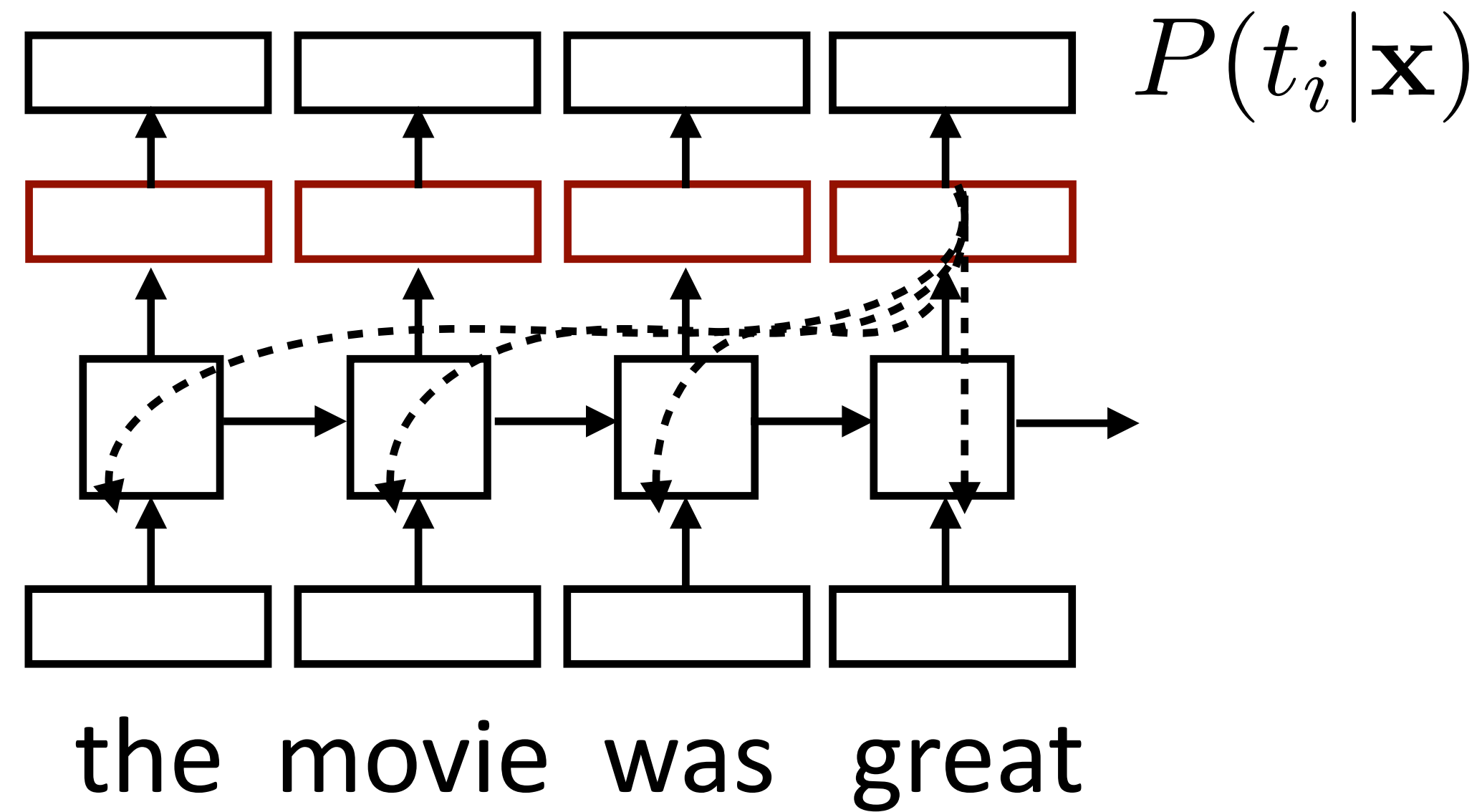
# Training RNNs

---



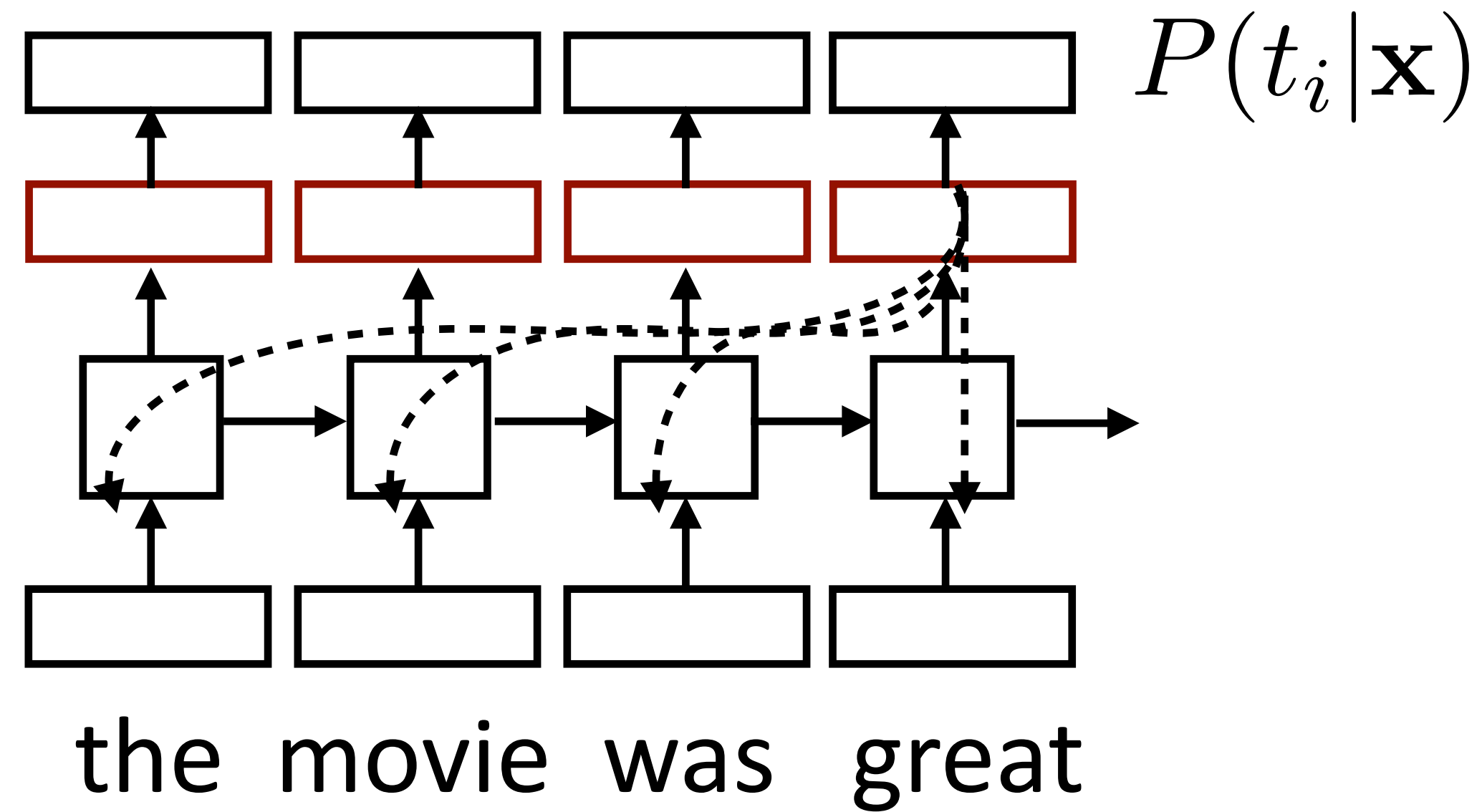
- ▶ Loss = negative log likelihood of probability of gold predictions, summed over the tags
- ▶ Loss terms filter back through network

# Training RNNs



- ▶ Loss = negative log likelihood of probability of gold predictions, summed over the tags
- ▶ Loss terms filter back through network

# Training RNNs



- ▶ Loss = negative log likelihood of probability of gold predictions, summed over the tags
- ▶ Loss terms filter back through network
- ▶ Example: language modeling (predict next word given context)

# Applications

# What can LSTMs model?

---

# What can LSTMs model?

---

- ▶ Sentiment
  - ▶ Encode one sentence, predict
- ▶ Language models
  - ▶ Move left-to-right, per-token prediction

# What can LSTMs model?

---

- ▶ Sentiment
  - ▶ Encode one sentence, predict
- ▶ Language models
  - ▶ Move left-to-right, per-token prediction
- ▶ Translation



# What can LSTMs model?

---

- ▶ Sentiment
  - ▶ Encode one sentence, predict
- ▶ Language models
  - ▶ Move left-to-right, per-token prediction
- ▶ Translation
  - ▶ Encode sentence + then decode, use token predictions for attention weights (later in the course)

# Visualizing LSTMs

---

# Visualizing LSTMs

---

- ▶ Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code

# Visualizing LSTMs

---

- ▶ Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- ▶ Visualize activations of specific cells (components of **c**) to understand them

# Visualizing LSTMs

- ▶ Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- ▶ Visualize activations of specific cells (components of **c**) to understand them

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.



# Visualizing LSTMs

- ▶ Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- ▶ Visualize activations of specific cells (components of **c**) to understand them
- ▶ Counter: know when to generate \n

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

# Visualizing LSTMs

---

- ▶ Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- ▶ Visualize activations of specific cells to see what they track

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."



# Visualizing LSTMs

---

- ▶ Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- ▶ Visualize activations of specific cells to see what they track
- ▶ Binary switch: tells us if we're in a quote or not

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."



# Visualizing LSTMs

---

- ▶ Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- ▶ Visualize activations of specific cells to see what they track

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
    int i;
    if (classes[class]) {
        for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
            if (mask[i] & classes[class][i])
                return 0;
    }
    return 1;
}
```

# Visualizing LSTMs

---

- ▶ Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- ▶ Visualize activations of specific cells to see what they track
- ▶ Stack: activation based on indentation

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
    int i;
    if (classes[class]) {
        for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
            if (mask[i] & classes[class][i])
                return 0;
    }
    return 1;
}
```

# Visualizing LSTMs

- ▶ Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- ▶ Visualize activations of specific cells to see what they track

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
    char *str;
    if (!*bufp || (len == 0) || (len > *remain))
        return ERR_PTR(-EINVAL);
    /* Of the currently implemented string fields, PATH_MAX
     * defines the longest valid length.
     */
}
```



# Visualizing LSTMs

- ▶ Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- ▶ Visualize activations of specific cells to see what they track
- ▶ Uninterpretable: probably doing double-duty, or only makes sense in the context of another activation

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
    char *str;
    if (!*bufp || (len == 0) || (len > *remain))
        return ERR_PTR(-EINVAL);
    /* Of the currently implemented string fields, PATH_MAX
     * defines the longest valid length.
     */
}
```

# What can LSTMs model?

---

- ▶ Sentiment
  - ▶ Encode one sentence, predict
- ▶ Language models
  - ▶ Move left-to-right, per-token prediction
- ▶ Translation
  - ▶ Encode sentence + then decode, use token predictions for attention weights (next lecture)

# What can LSTMs model?

---

- ▶ Sentiment
  - ▶ Encode one sentence, predict
- ▶ Language models
  - ▶ Move left-to-right, per-token prediction
- ▶ Translation
  - ▶ Encode sentence + then decode, use token predictions for attention weights (next lecture)
- ▶ Textual entailment

# What can LSTMs model?

---

- ▶ Sentiment
  - ▶ Encode one sentence, predict
- ▶ Language models
  - ▶ Move left-to-right, per-token prediction
- ▶ Translation
  - ▶ Encode sentence + then decode, use token predictions for attention weights (next lecture)
- ▶ Textual entailment
  - ▶ Encode two sentences, predict

# Natural Language Inference

---

Premise

A boy plays in the snow

Hypothesis

A boy is outside



# Natural Language Inference

---

Premise

Hypothesis

A boy plays in the snow

*entails*

A boy is outside

# Natural Language Inference

---

Premise

Hypothesis

A boy plays in the snow

*entails*

A boy is outside

A man inspects the uniform of a figure

The man is sleeping

# Natural Language Inference

---

Premise

Hypothesis

A boy plays in the snow

*entails*

A boy is outside

A man inspects the uniform of a figure

*contradicts*

The man is sleeping

# Natural Language Inference

---

Premise

Hypothesis

A boy plays in the snow

*entails*

A boy is outside

A man inspects the uniform of a figure

*contradicts*

The man is sleeping

An older and younger man smiling

Two men are smiling and  
laughing at cats playing

# Natural Language Inference

---

Premise

Hypothesis

A boy plays in the snow

*entails*

A boy is outside

A man inspects the uniform of a figure

*contradicts*

The man is sleeping

An older and younger man smiling

*neutral*

Two men are smiling and  
laughing at cats playing

# Natural Language Inference

---

Premise

Hypothesis

A boy plays in the snow

*entails*

A boy is outside

A man inspects the uniform of a figure

*contradicts*

The man is sleeping

An older and younger man smiling

*neutral*

Two men are smiling and  
laughing at cats playing

- ▶ Long history of this task: “Recognizing Textual Entailment” challenge in 2006 (Dagan, Glickman, Magnini)

# Natural Language Inference

---

Premise

Hypothesis

A boy plays in the snow

*entails*

A boy is outside

A man inspects the uniform of a figure

*contradicts*

The man is sleeping

An older and younger man smiling

*neutral*

Two men are smiling and  
laughing at cats playing

- ▶ Long history of this task: “Recognizing Textual Entailment” challenge in 2006 (Dagan, Glickman, Magnini)
- ▶ Early datasets: small (hundreds of pairs), very ambitious (lots of world knowledge, temporal reasoning, etc.)

# SNLI Dataset

---

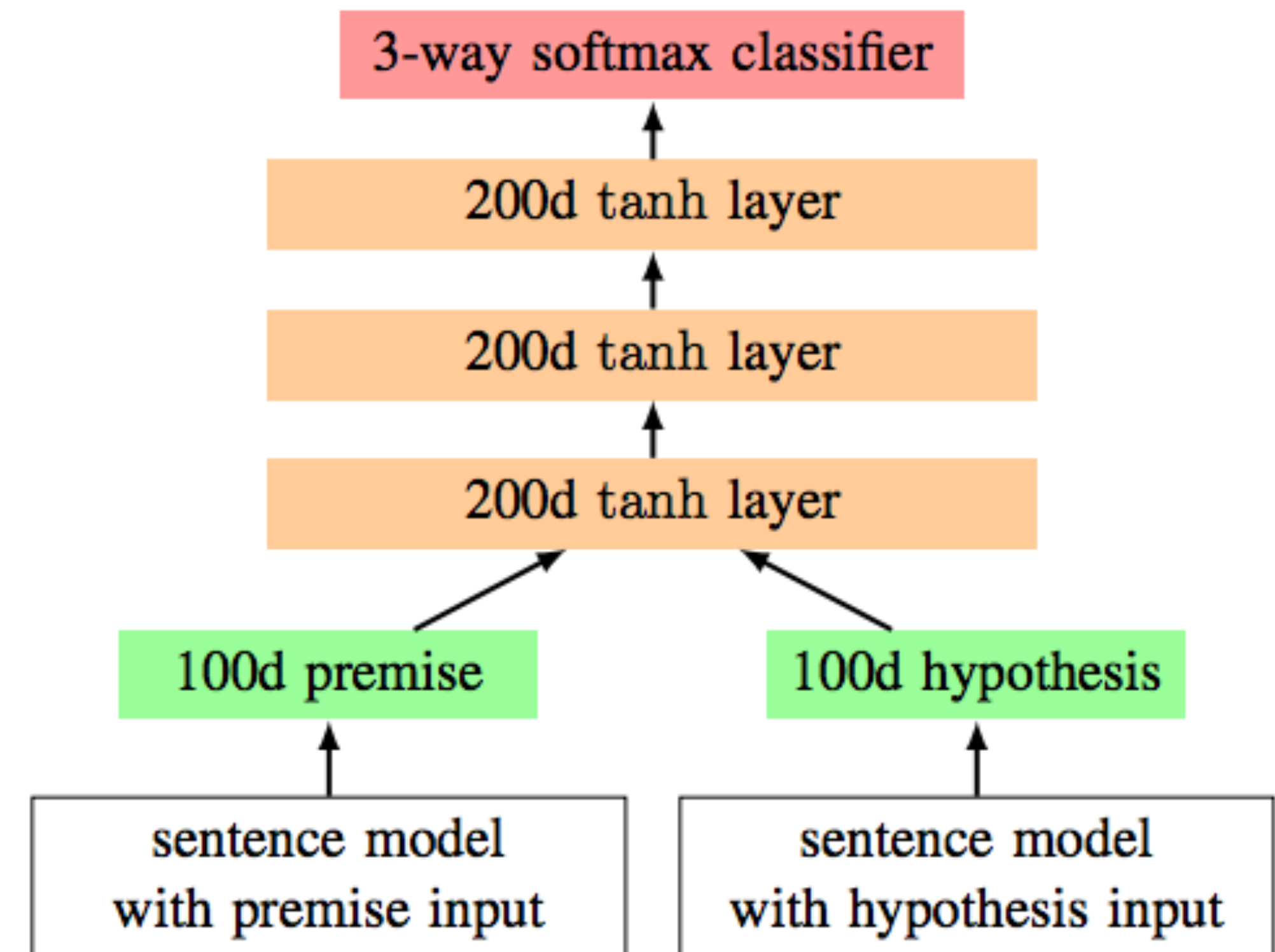
- ▶ Show people captions for (unseen) images and solicit entailed / neural / contradictory statements
- ▶ >500,000 sentence pairs



# SNLI Dataset

---

- ▶ Show people captions for (unseen) images and solicit entailed / neural / contradictory statements
- ▶ >500,000 sentence pairs
- ▶ Encode each sentence and process

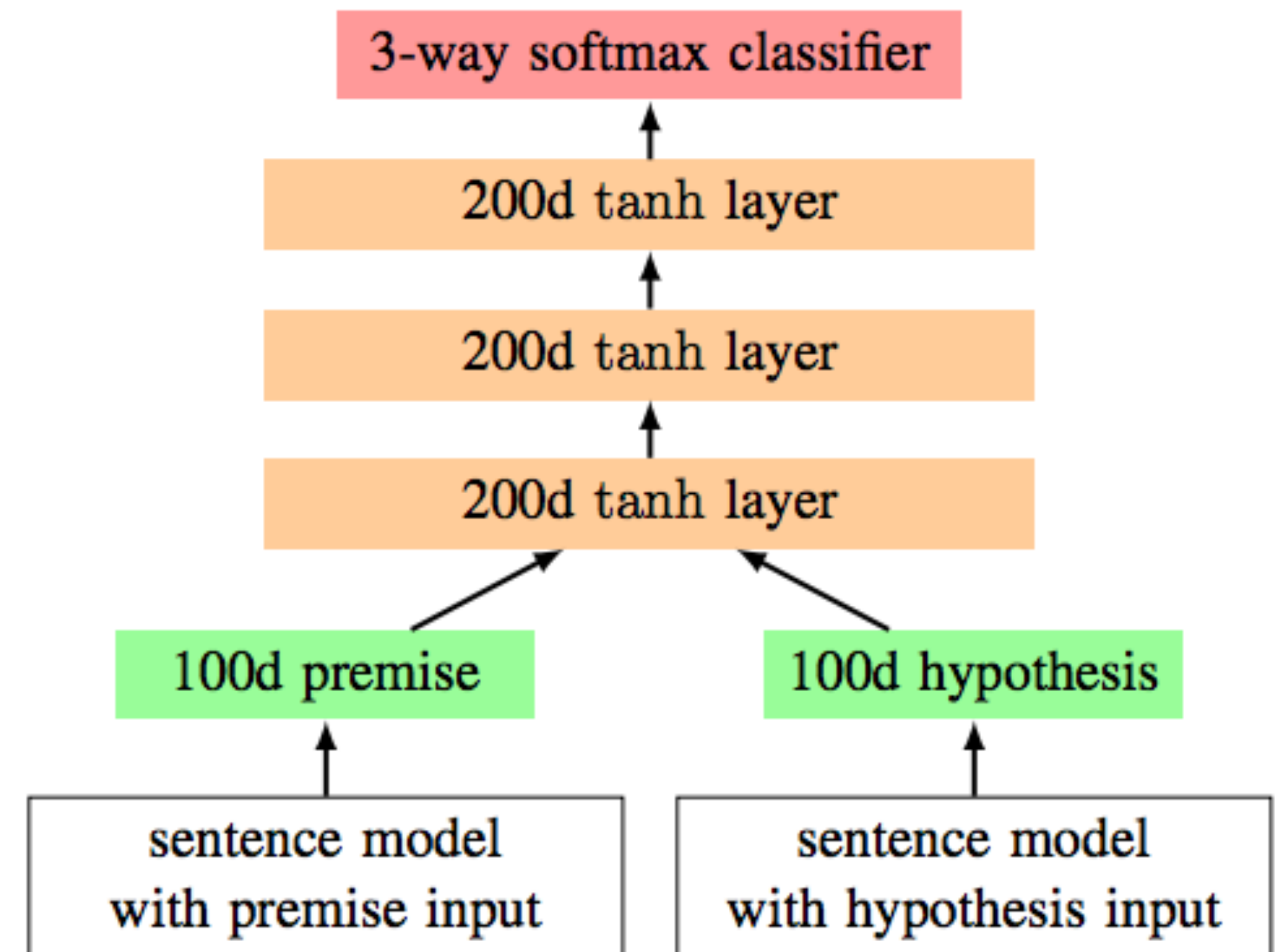


Bowman et al. (2015)

# SNLI Dataset

---

- ▶ Show people captions for (unseen) images and solicit entailed / neural / contradictory statements
  - ▶ >500,000 sentence pairs
  - ▶ Encode each sentence and process
- 100D LSTM: 78% accuracy



Bowman et al. (2015)

# SNLI Dataset

---

- ▶ Show people captions for (unseen) images and solicit entailed / neural / contradictory statements

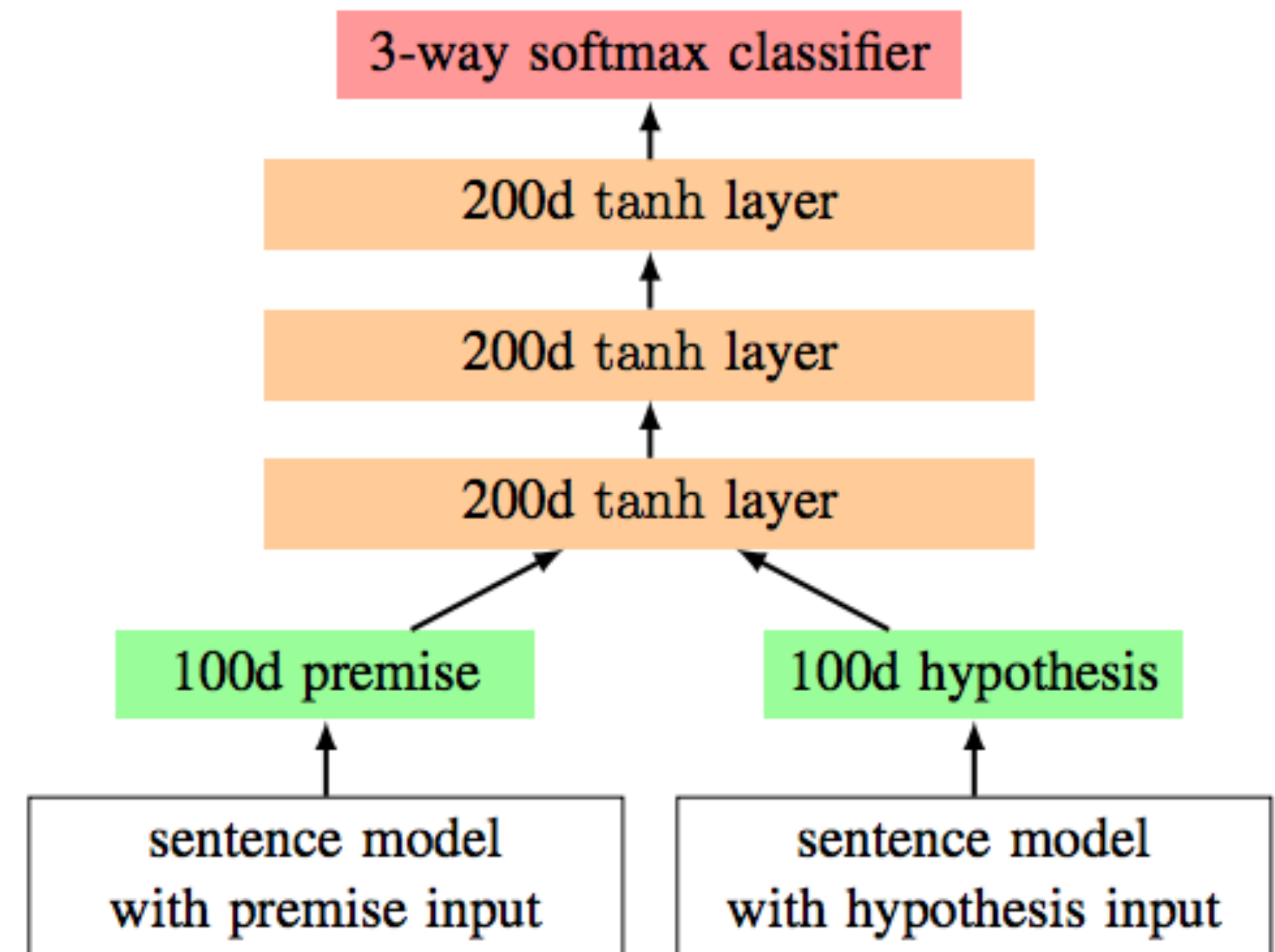
- ▶ >500,000 sentence pairs

- ▶ Encode each sentence and process

100D LSTM: 78% accuracy

300D LSTM: 80% accuracy

(Bowman et al., 2016)



Bowman et al. (2015)

# SNLI Dataset

---

- ▶ Show people captions for (unseen) images and solicit entailed / neural / contradictory statements

- ▶ >500,000 sentence pairs

- ▶ Encode each sentence and process

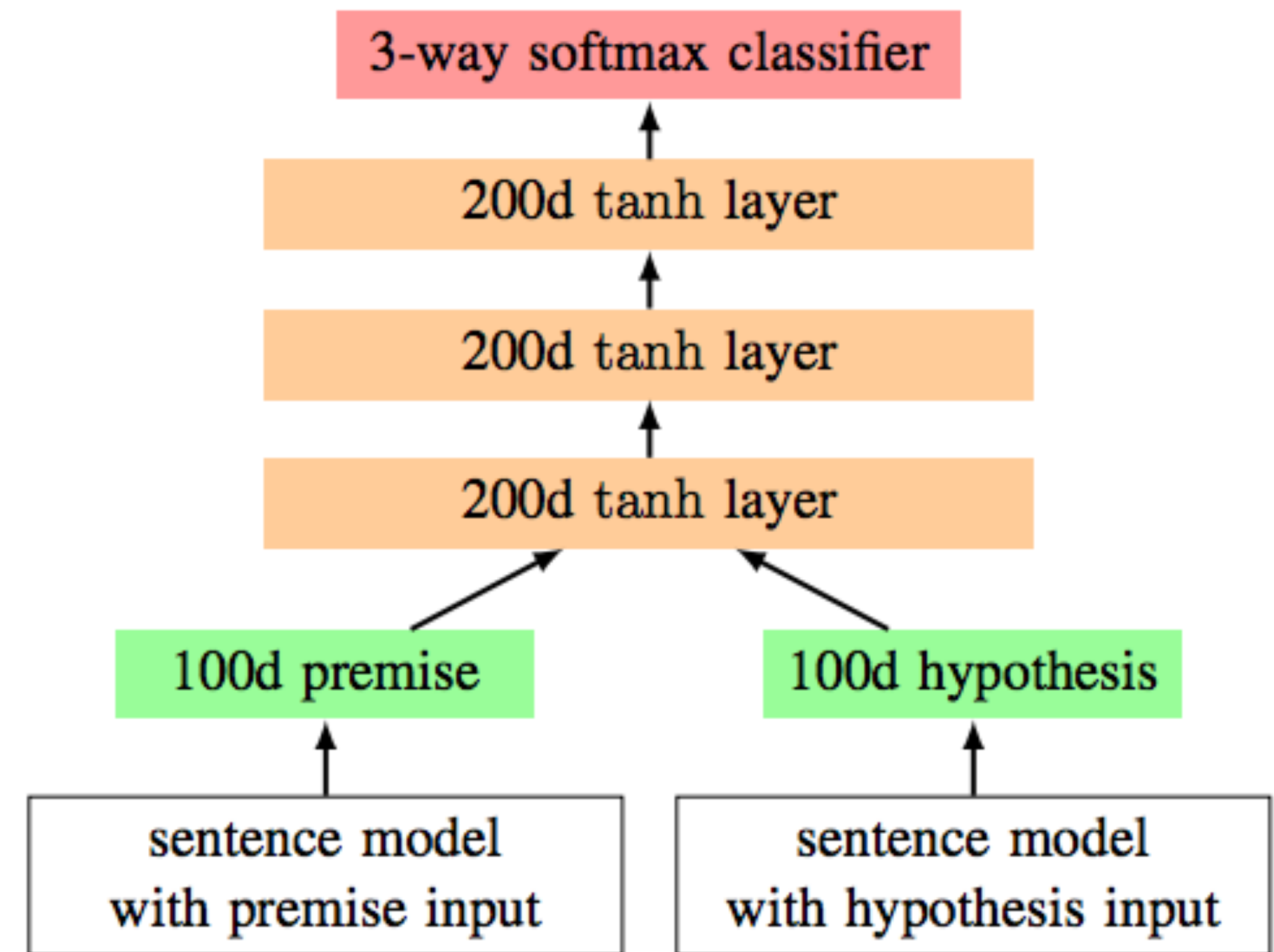
100D LSTM: 78% accuracy

300D LSTM: 80% accuracy

(Bowman et al., 2016)

300D BiLSTM: 83% accuracy

(Liu et al., 2016)



Bowman et al. (2015)



# SNLI Dataset

---

- ▶ Show people captions for (unseen) images and solicit entailed / neural / contradictory statements

- ▶ >500,000 sentence pairs

- ▶ Encode each sentence and process

100D LSTM: 78% accuracy

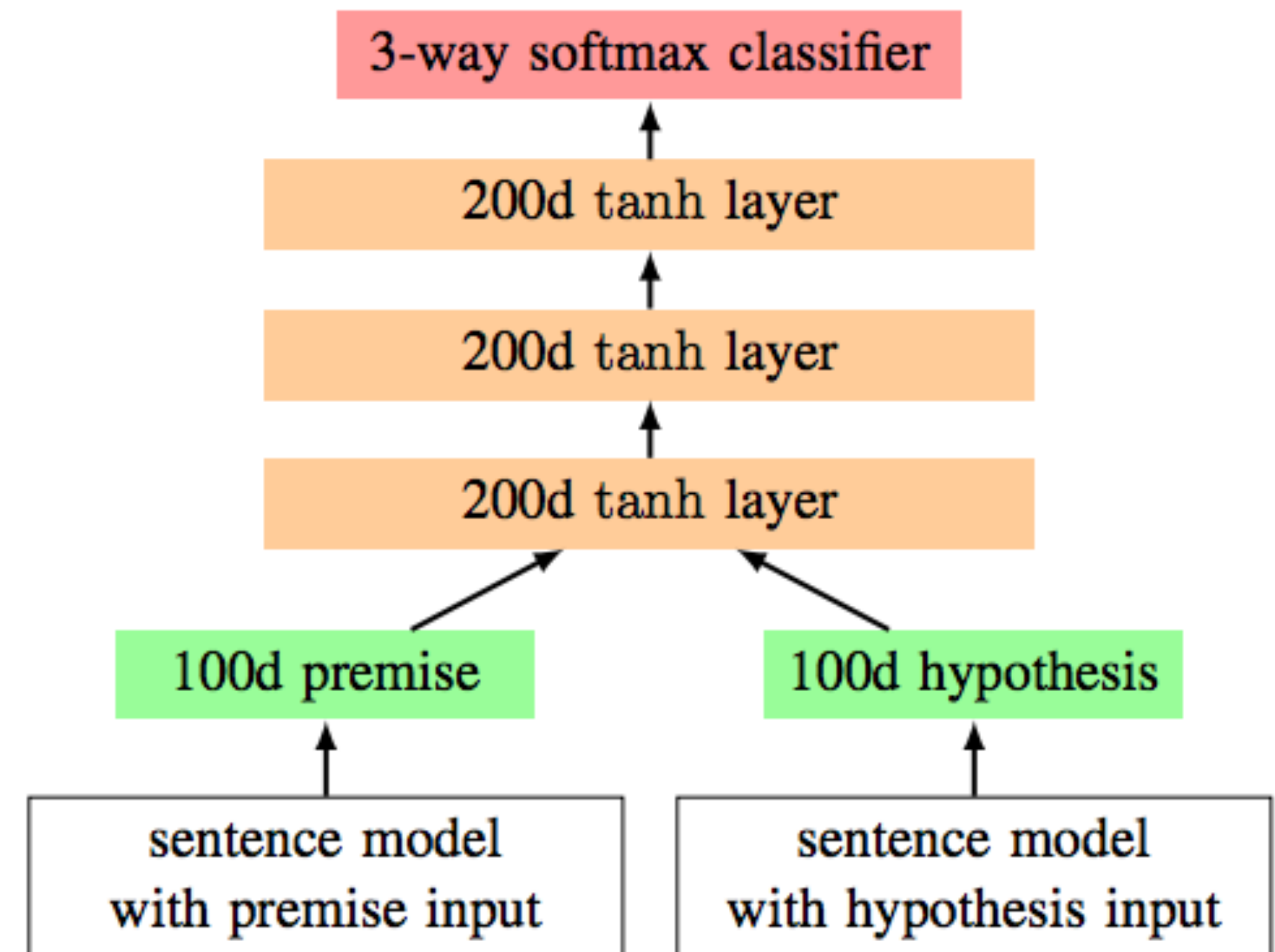
300D LSTM: 80% accuracy

(Bowman et al., 2016)

300D BiLSTM: 83% accuracy

(Liu et al., 2016)

- ▶ Later: better models for this



Bowman et al. (2015)

# Takeaways

---

- ▶ RNNs can transduce inputs (produce one output for each input) or compress the whole input into a vector
- ▶ Useful for a range of tasks with sequential input: sentiment analysis, language modeling, natural language inference, machine translation
- ▶ Next time: CNNs and neural CRFs