



Gating Mechanism
Nov 24th, 2016

Applied Deep Learning

YUN-NUNG (VIVIAN) CHEN WWW.CSIE.NTU.EDU.TW/~YVCHEN/F105-ADL



臺灣大學

National Taiwan University

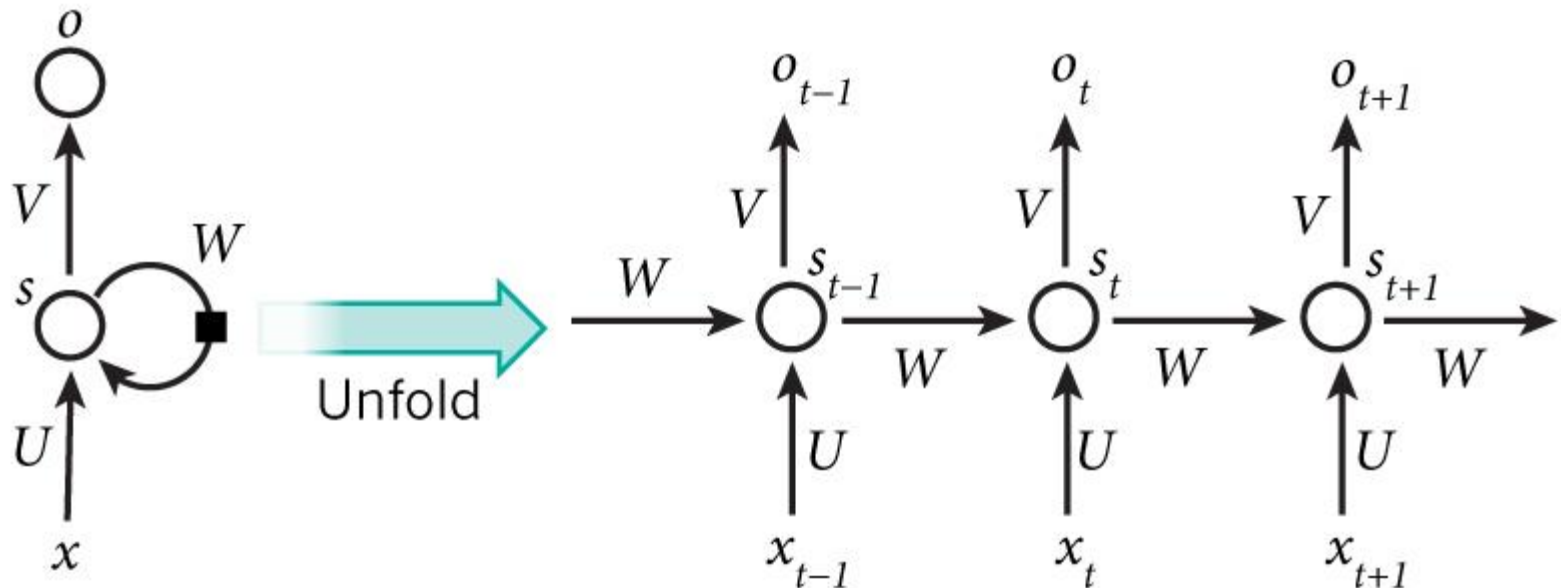
Review

Vanishing Gradient Problem

Recurrent Neural Network Definition

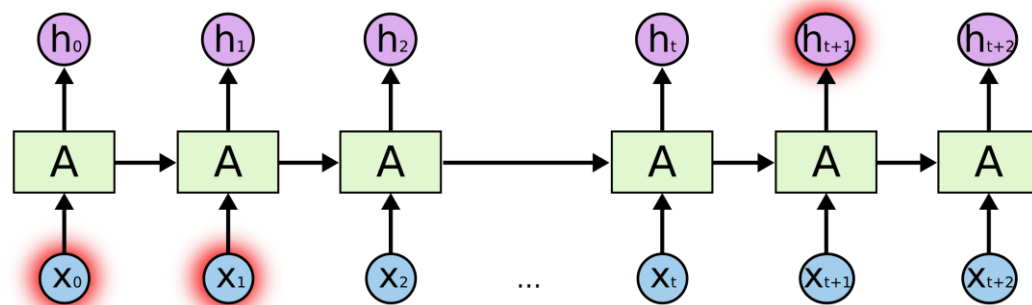
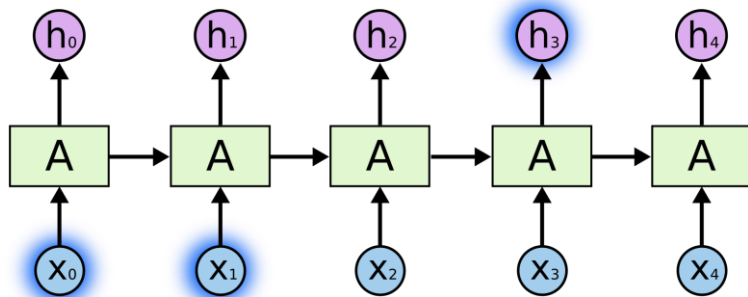
$$s_t = \sigma(W s_{t-1} + U x_t) \quad \sigma(\cdot): \text{tanh, ReLU}$$

$$o_t = \text{softmax}(V s_t)$$



Vanishing Gradient: Gating Mechanism

RNN: keeps temporal sequence information



“I grew up in France...
I speak fluent French.”

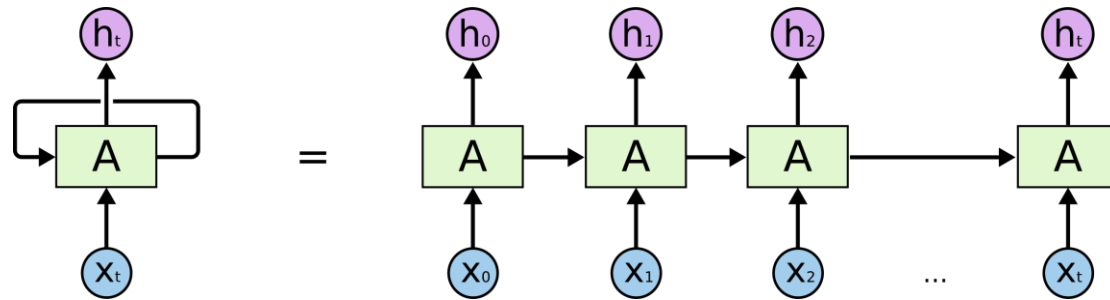
Issue: in theory, RNNs can handle such “long-term dependencies,” but they cannot in practice
→ use gates to directly encode the long-distance information

Long Short-Term Memory

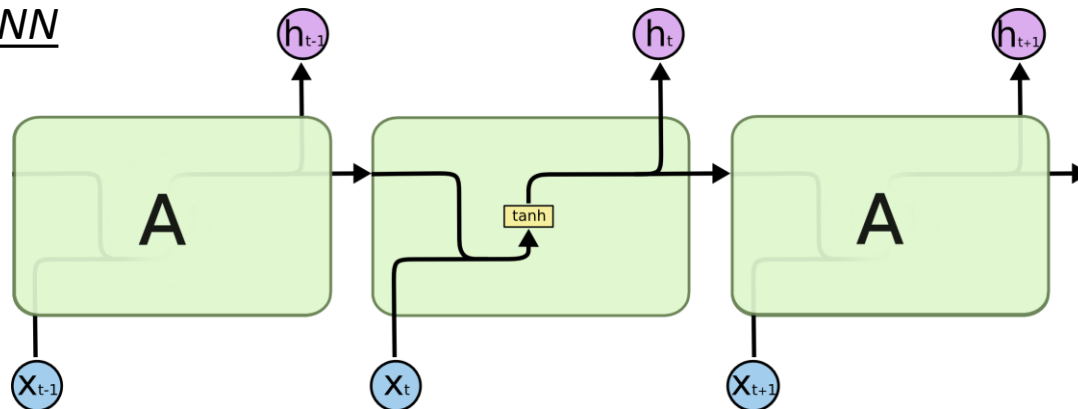
Addressing Vanishing Gradient Problem

Long Short-Term Memory (LSTM)

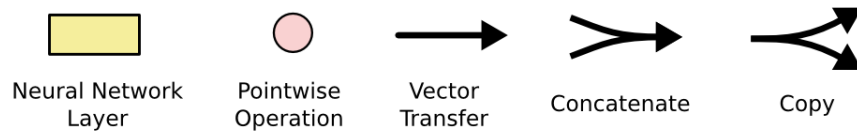
LSTMs are explicitly designed to avoid the long-term dependency problem



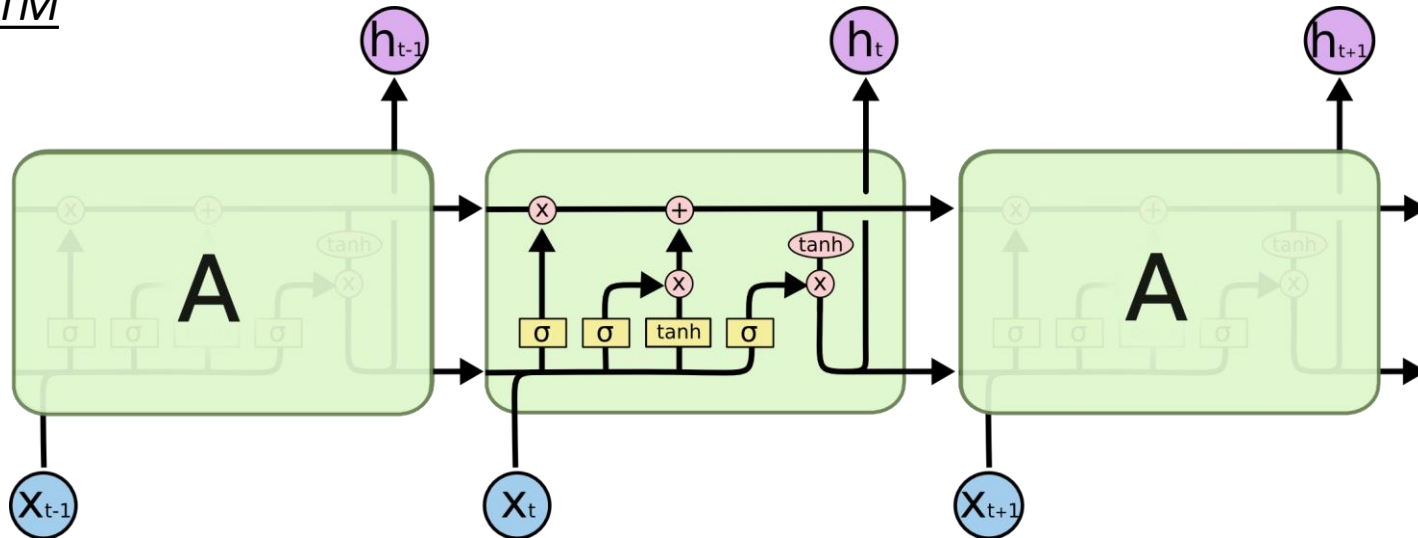
Vanilla RNN



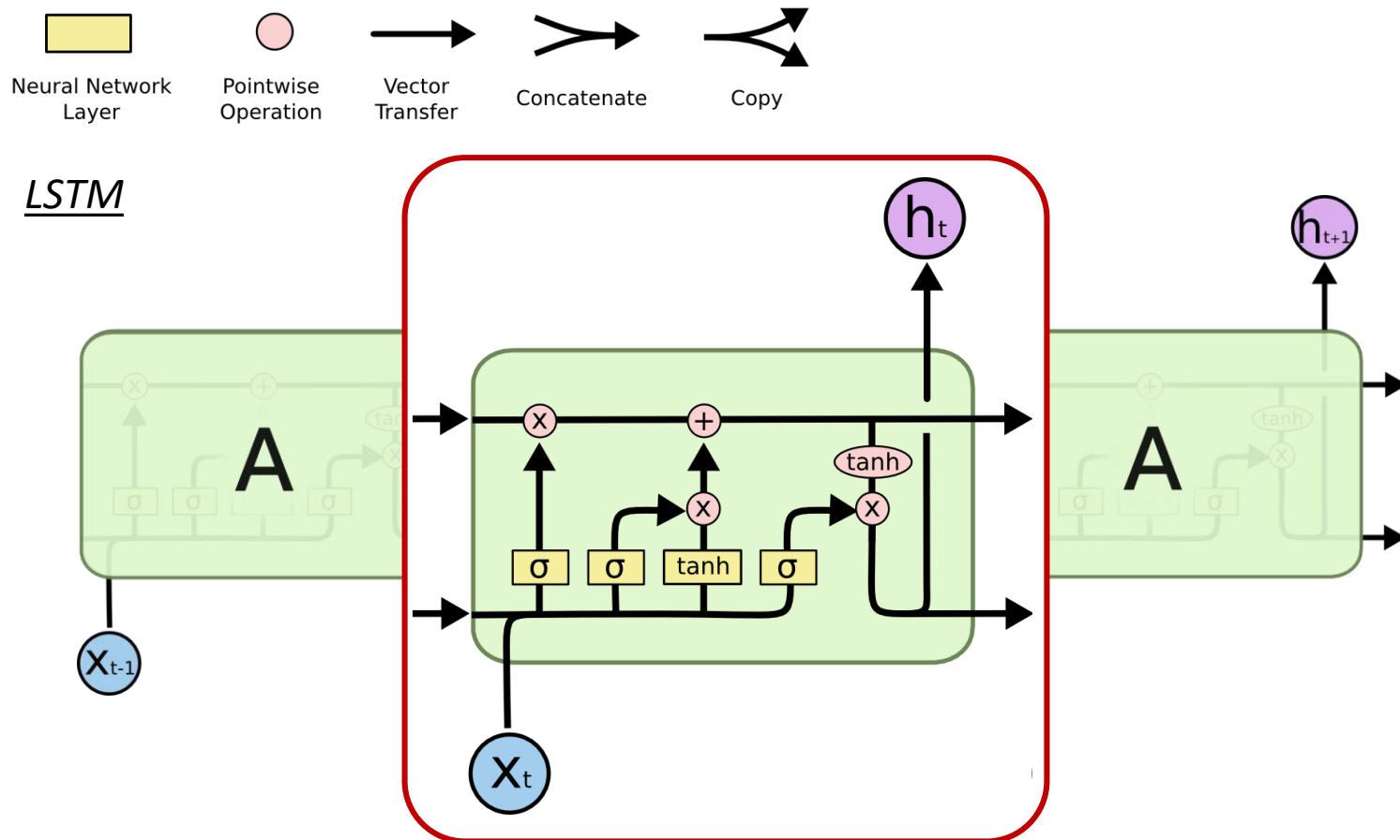
Long Short-Term Memory (LSTM)



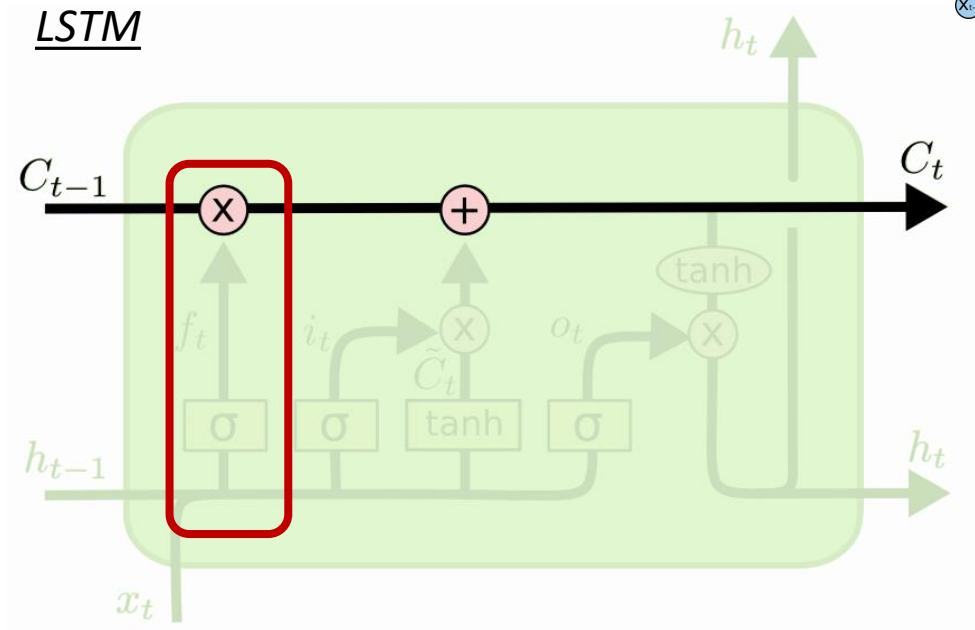
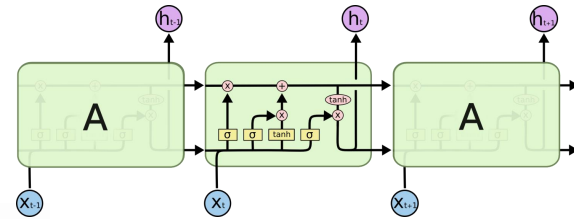
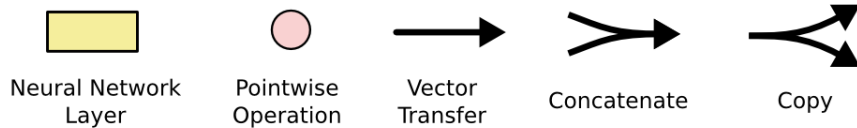
LSTM



Long Short-Term Memory (LSTM)



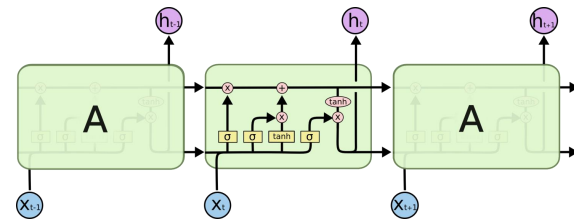
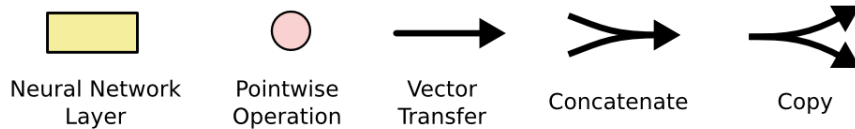
Long Short-Term Memory (LSTM)



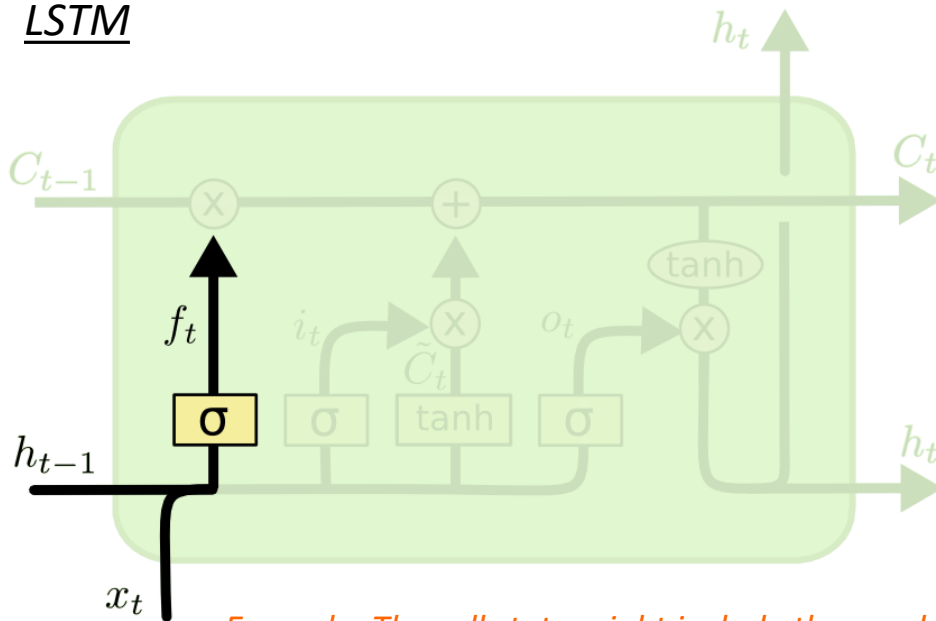
runs straight down the chain with minor linear interactions
 → easy for information to flow along it unchanged

Gates are a way to optionally let information through
 → composed of a sigmoid and a pointwise multiplication operation

Long Short-Term Memory (LSTM)



LSTM



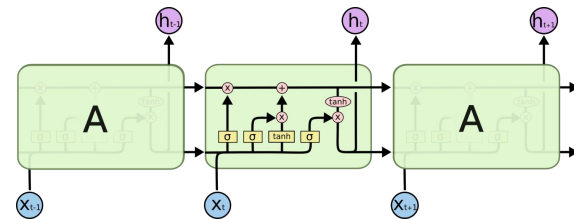
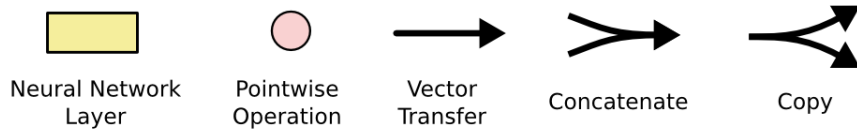
forget gate (a sigmoid layer): decides what information we're going to throw away from the cell state

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

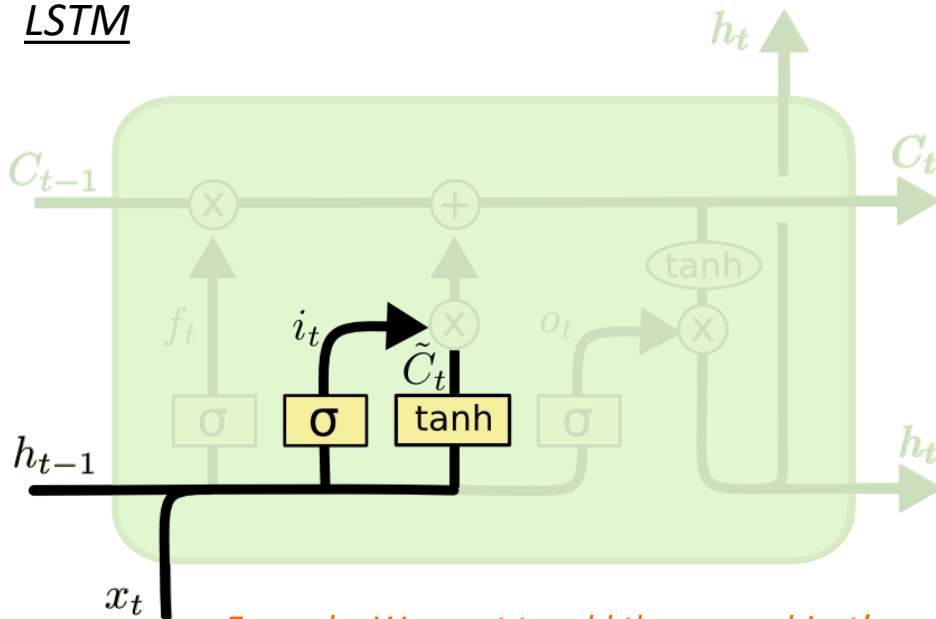
- 1: "completely keep this"
- 0: "completely get rid of this"

Example: The cell state might include the gender of the present subject, so that the correct pronouns can be used. When seeing a new subject, we want to forget the old subject's gender.

Long Short-Term Memory (LSTM)



LSTM



input gate (a sigmoid layer): decides what new information we're going to store in the cell state

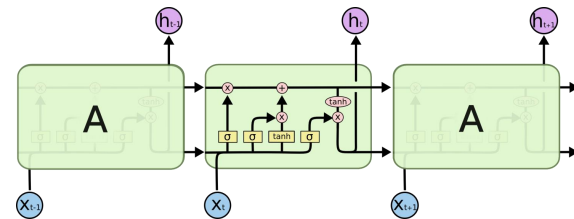
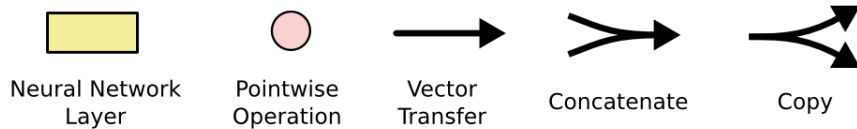
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

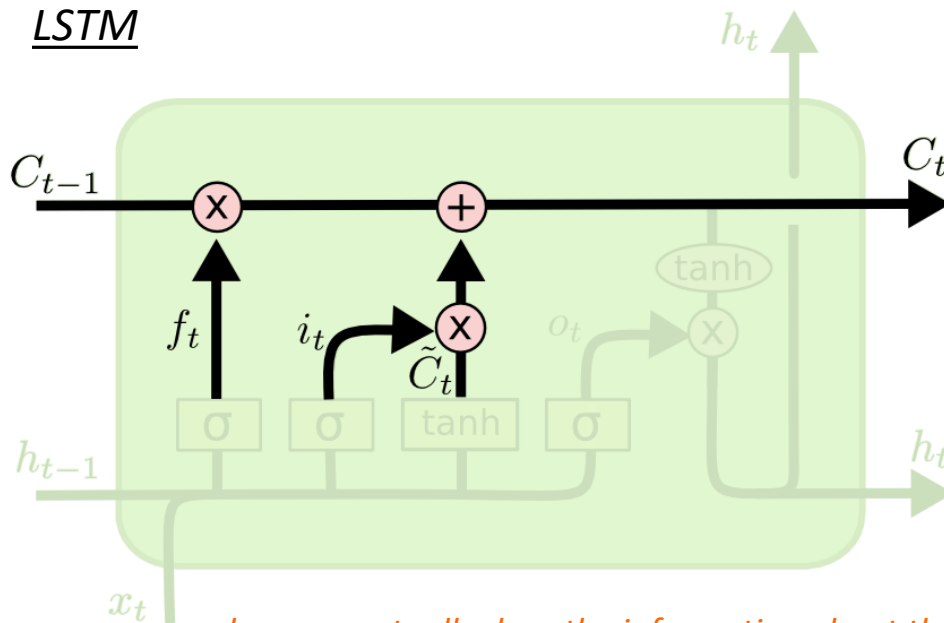
Vanilla RNN

Example: We want to add the new subject's gender to the cell state for replacing the old one.

Long Short-Term Memory (LSTM)



LSTM



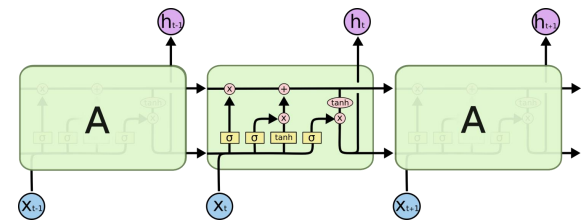
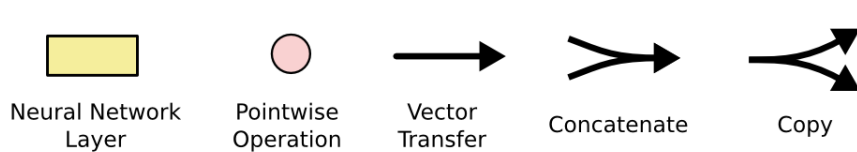
where we actually drop the information about the old subject's gender and add the new information

cell state update: forgets the things we decided to forget earlier and add the new candidate values, scaled by how much we decided to update each state value

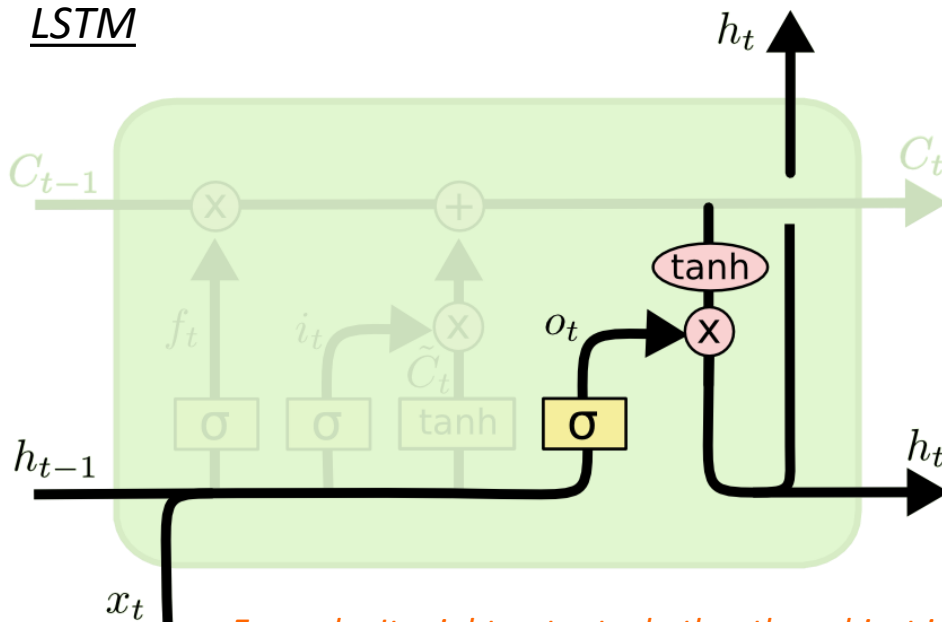
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

- f_t : decides which to forget
- i_t : decide which to update

Long Short-Term Memory (LSTM)



LSTM



output gate (a sigmoid layer):
decides what new information we're going to output

$$o_t = \sigma (W_o [h_{t-1}, x_t] + b_o)$$

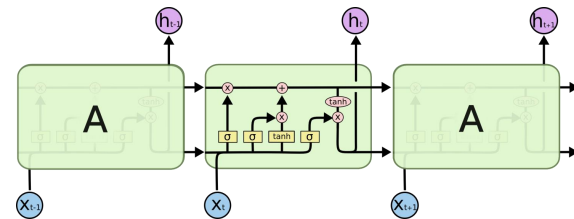
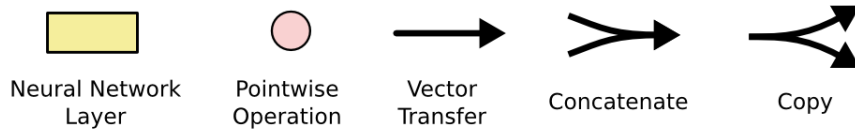
$$h_t = o_t * \tanh (C_t)$$

Example: It might output whether the subject is singular or plural, so that we know what form a verb should be conjugated into if that's what follows next.

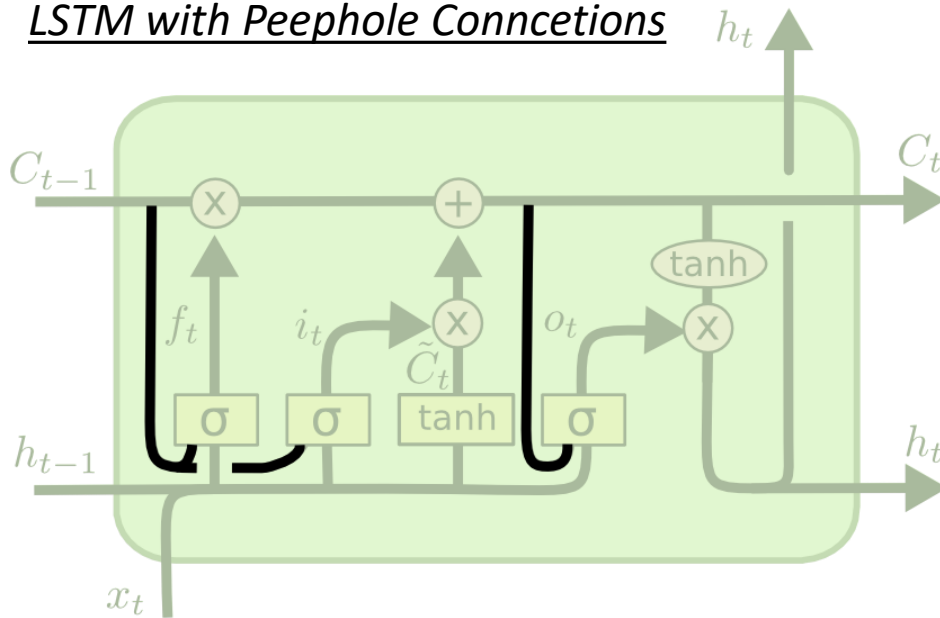
Variants on LSTM

Addressing Vanishing Gradient Problem

LSTM with Peephole Connections



LSTM with Peephole Connections



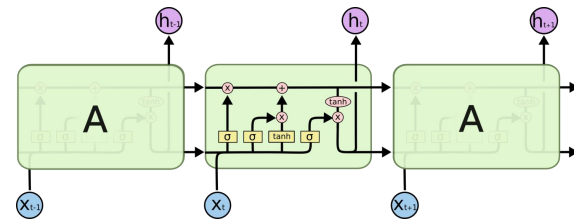
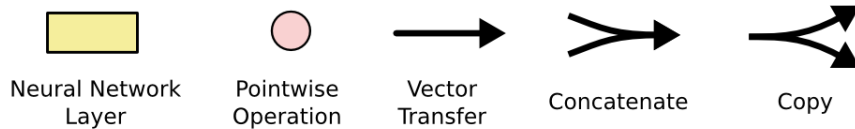
Idea: allow gate layers to look at the cell state

$$f_t = \sigma(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f)$$

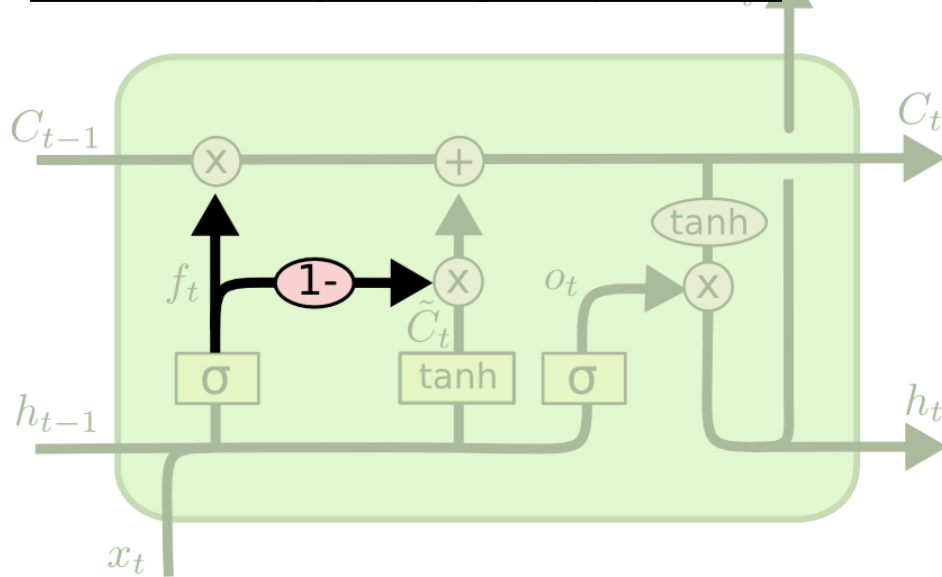
$$i_t = \sigma(W_i \cdot [C_{t-1}, h_{t-1}, x_t] + b_i)$$

$$o_t = \sigma(W_o \cdot [C_t, h_{t-1}, x_t] + b_o)$$

LSTM with Coupled Forget/Input Gates



LSTM with Coupled Forget/Input Gates



Idea: instead of separately deciding what to forget and what we should add new information to, we make those decisions together

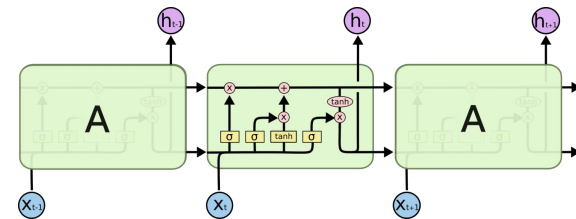
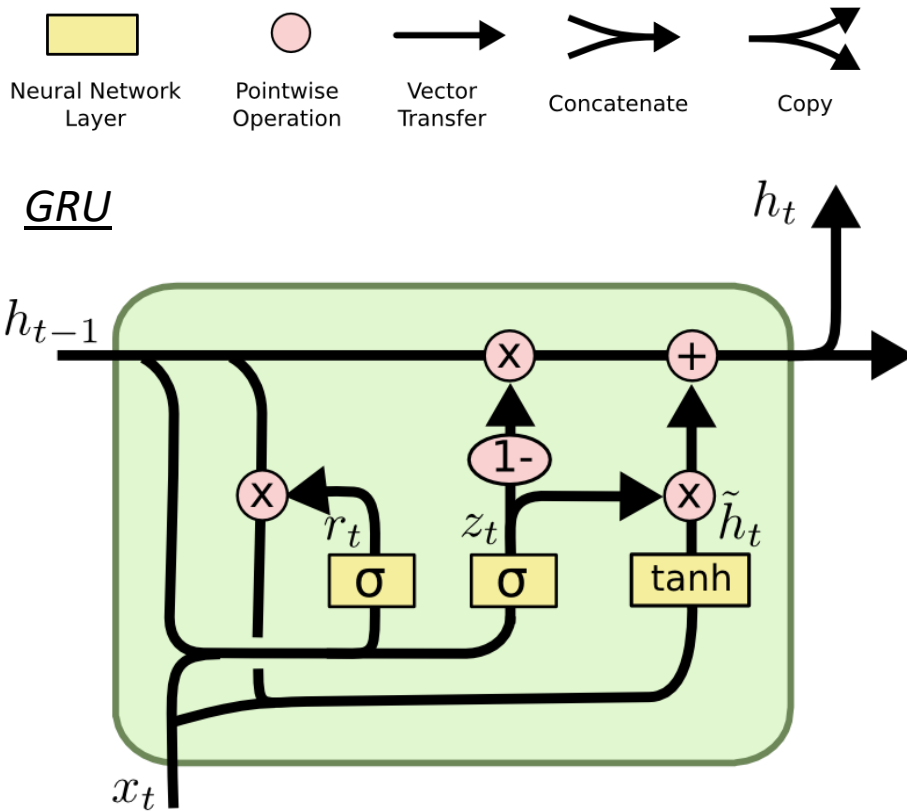
$$C_t = f_t * C_{t-1} + (1 - f_t) * \tilde{C}_t$$

We only forget when we're going to input something in its place, and vice versa.

Gated Recurrent Unit

Addressing Vanishing Gradient Problem

Gated Recurrent Unit (GRU)



Idea: combine the forget and input gates into a single “update gate”; merge the cell state and hidden state

$$\text{update gate: } z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

$$\text{reset gate: } r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

$r_t=0$: ignore previous memory and only stores the new word information

GRU is simpler and has less parameters than LSTM

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

Concluding Remarks

Gating mechanism for vanishing gradient problem

Gated RNN

- Long Short-Term Memory (LSTM)
- Peephole Connections
- Coupled Forget/Input Gates
- Gated Recurrent Unit (GRU)

