# LSTM Networks

## CS109B Data Science 2

Pavlos Protopapas, Mark Glickman, and Chris Tanner

# Outline

Recap

GRU++

Long short-term memory LSTM

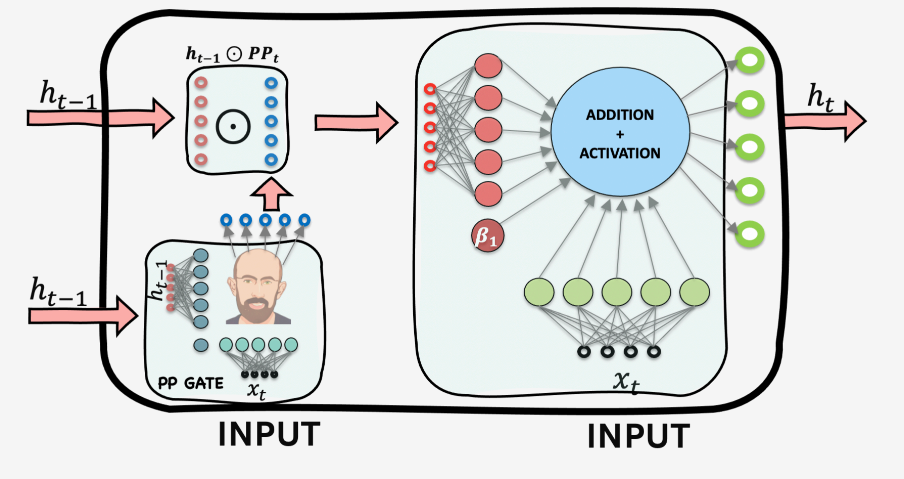
LSTM applications

$$\mathbf{h}_t = \tanh\left(\mathbf{V}\mathbf{X}_t + \mathbf{U}\mathbf{h}_{t-1} + \beta_1\right)$$

$$\mathbf{h}_t = \tanh\left(\mathbf{V}\mathbf{X}_t + \mathbf{U}\mathbf{h}_{t-1} + \beta_1\right)$$

$$\mathbf{h}_t = \tanh\left(\mathbf{V}\mathbf{X}_t + \mathbf{U}\left[\mathbf{PP}_t \odot \mathbf{h}_{t-1}\right] + \beta_1\right)$$

$$\mathbf{PP}_t = \sigma\left(\mathbf{V}_{pp}\mathbf{X}_t + \mathbf{U}_{pp}\mathbf{h}_{t-1} + \beta_{pp}\right)$$

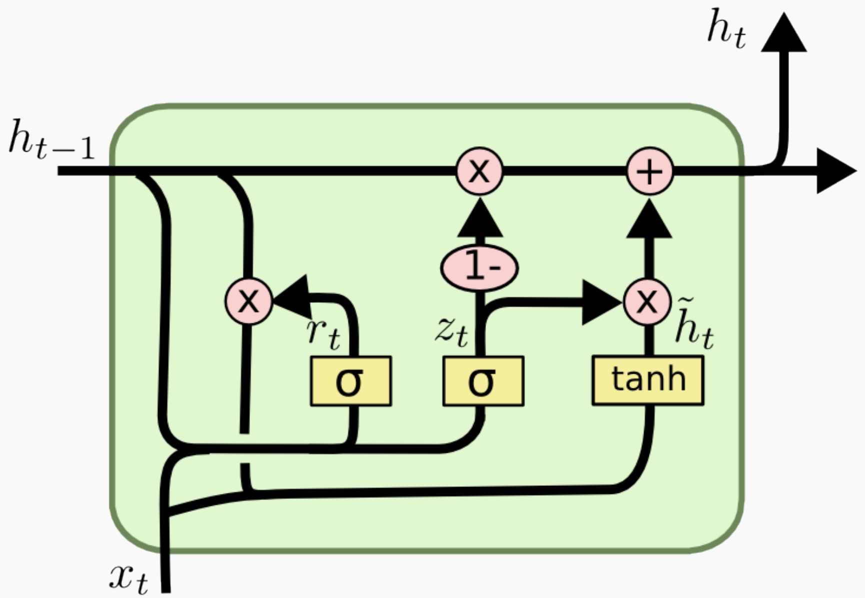


# Gated Recurrent Unit (GRU)

$$\tilde{\mathbf{h}}_t = \tanh\left(\mathbf{V}\mathbf{X}_t + \mathbf{U}\left[\mathbf{R}_t \odot \mathbf{h}_{t-1}\right] + \beta_1\right)$$

$$\mathbf{h}_t = \mathbf{Z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{Z}_t) \odot \tilde{\mathbf{h}}_\mathbf{t}$$

Reset Gate (equivalent to PP gate)




$$\mathbf{R}_t = \sigma\left(\mathbf{V}_R\mathbf{X}_t + \mathbf{U}_R\mathbf{h}_{t-1} + \beta_R\right)$$

$$\mathbf{Z}_t = \sigma\left(\mathbf{V}_Z\mathbf{X}_t + \mathbf{U}_Z\mathbf{h}_{t-1} + \beta_Z\right)$$

Update Gate

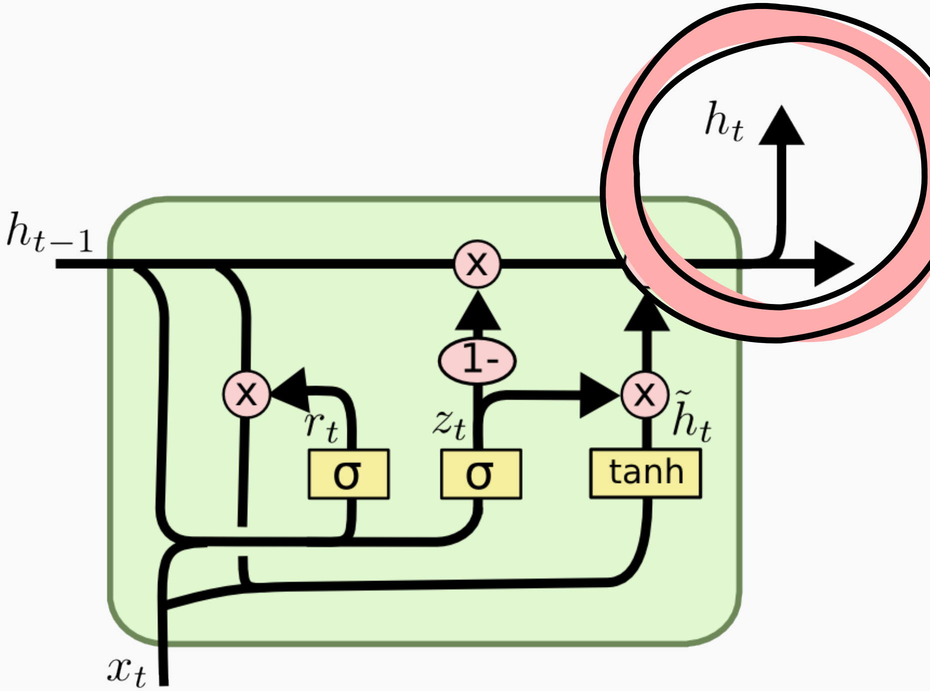
# GRU

## GRU STRENGTHS?

- Current input can affect how much of the past information to consider
- The update gate solves the vanishing gradient problem
- Hidden state more robust to outlier inputs because of the update gate

## GRU ISSUES?

- The same hidden state is used for memory and output
- With only two gates, performance suffers on longer sequences



# GRU++

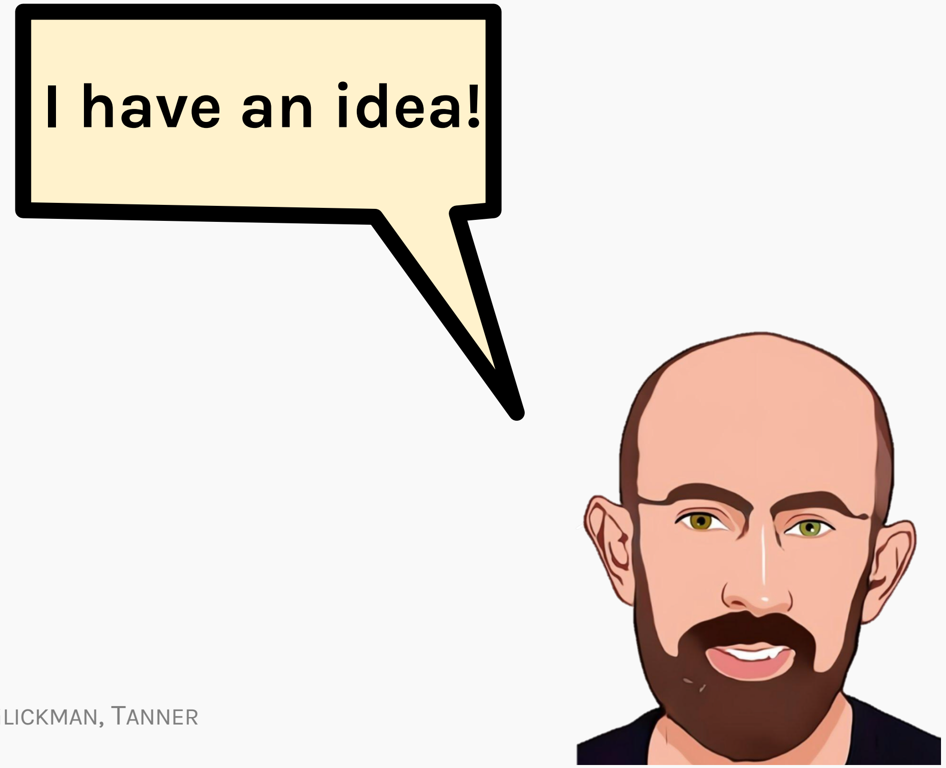
$$\tilde{h}_t = \tanh(X_t V + (R_t \odot h_{t-1})U + \beta_1)$$

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

# GRU++

$$\tilde{h}_t = \tanh(X_t V + (R_t \odot h_{t-1})U + \beta_1)$$

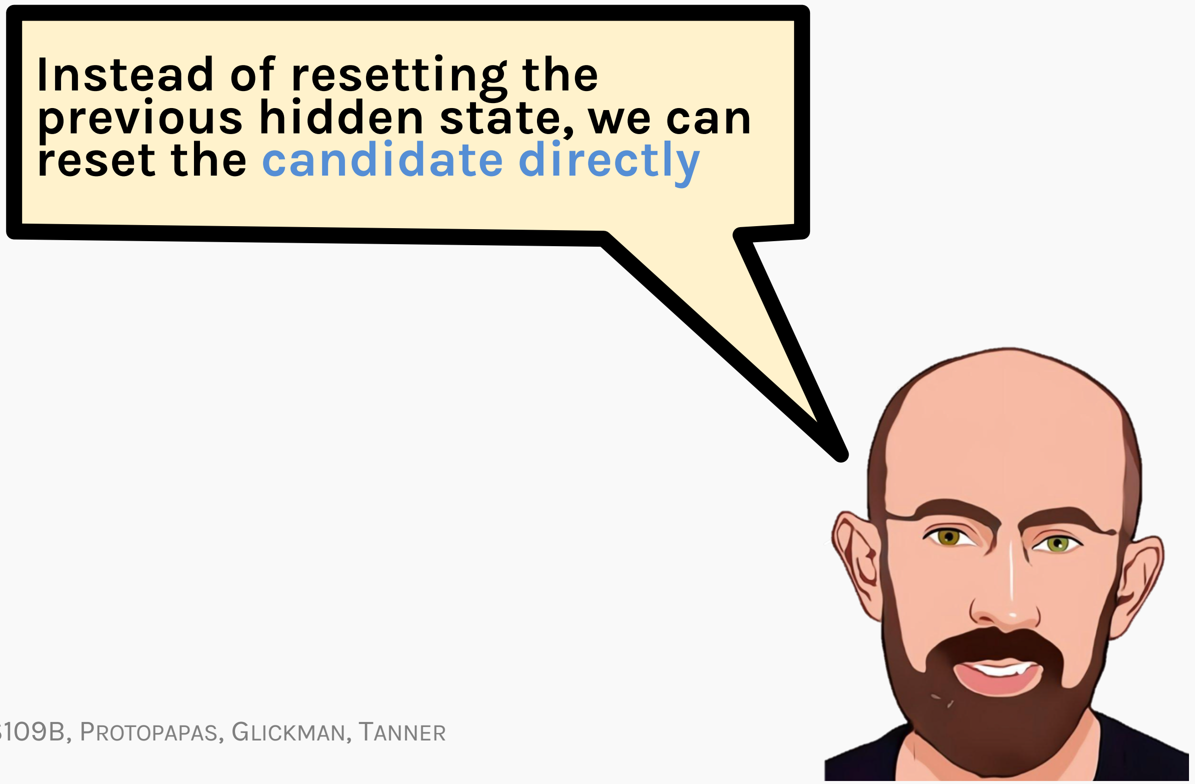
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t$$



# GRU++

$$\tilde{h}_t = \tanh(X_t V + \quad h_{t-1}\, U + \beta_1)$$

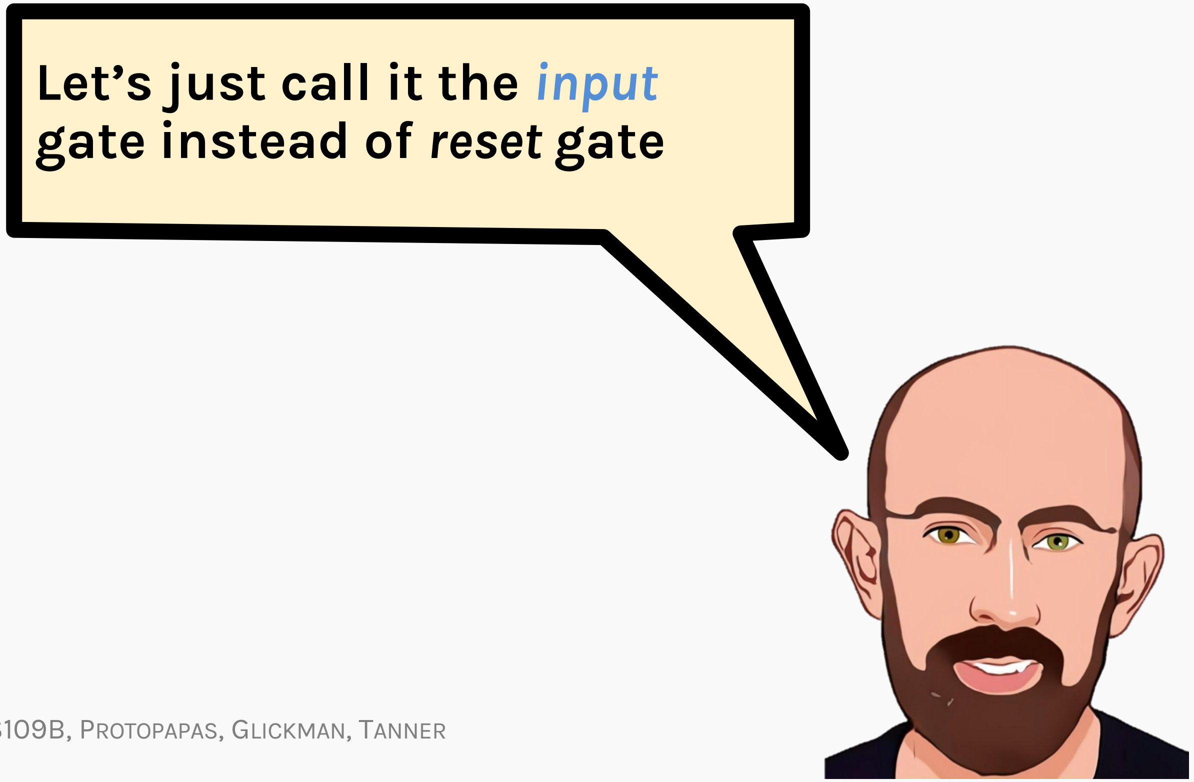
$$\tilde{h}_t = \boxed{R_t \odot \tilde{h}_t}$$



# GRU++

$$\tilde{h}_t = \tanh(X_t V + \quad h_{t-1}\, U + \beta_1)$$

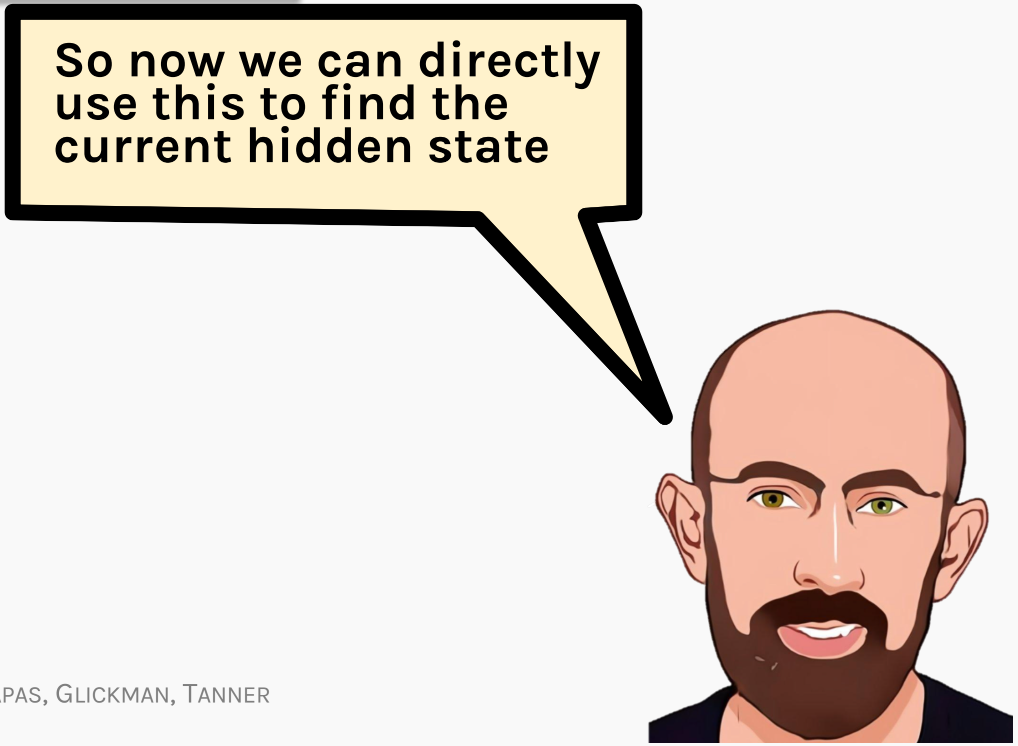
$$\tilde{h}_t = \boxed{i_t \odot \tilde{h}}$$



# GRU++

$$h_t = f_t \odot h_{t-1} + i_t \odot \tilde{h}_t$$

We change the notation of $Z_t$ to $f_t$



# GRU++

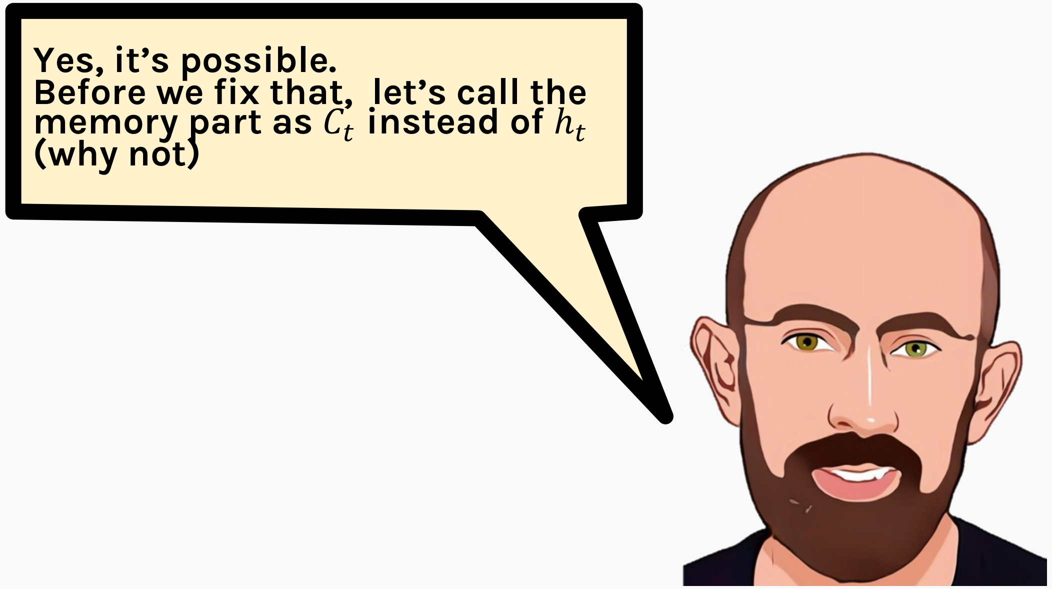
$$h_t = \boxed{f_t \odot h_{t-1} + i_t \odot \widetilde{h}_t}$$

← Can become unbounded!

**But isn't it possible for $h_t$ to become unbounded over time?**

# GRU++

$$c_t = \mathrm{f_t} \odot c_{t-1} + i_t \odot \tilde{c}_t$$



# GRU++

$$c_t = \mathrm{f}_\mathrm{t} \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$h_t = \tanh(c_t)$$

# GRU++

Note: We have two memories!
$h_t$ is bounded
$c_t$ is not bounded

$$c_t = \mathrm{f_t} \odot c_{t-1} + i_t \odot \tilde{c}_t$$

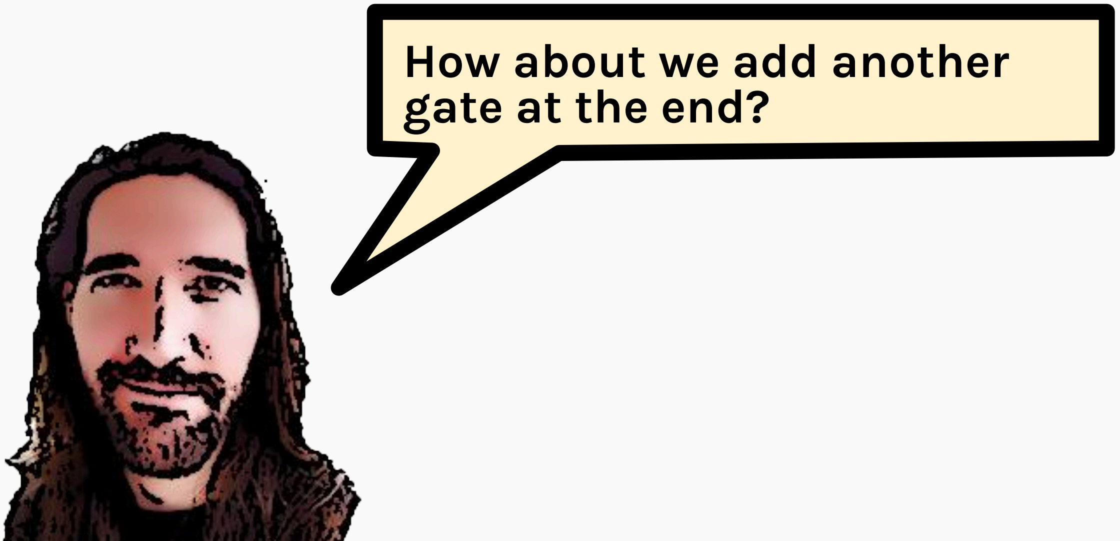
$$h_t = \tanh(c_t)$$

And use a nice activation function such as hyperbolic tan on it

# GRU++

$$c_t = \mathrm{f_t} \odot c_{t-1} + i_t \odot \tilde{c}_t$$

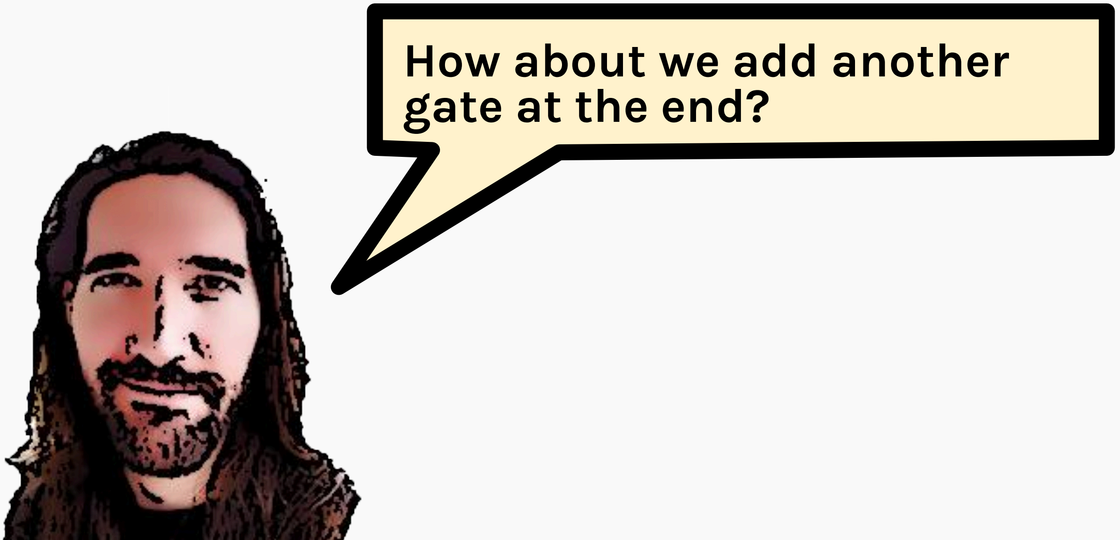
$$h_t = \tanh(c_t)$$



# GRU++

$$c_t = \mathrm{f_t} \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$h_t = \tanh(c_t)$$




# GRU++

$$c_t = \mathrm{f_t} \odot c_{t-1} + i_t \odot \tilde{c}_t$$

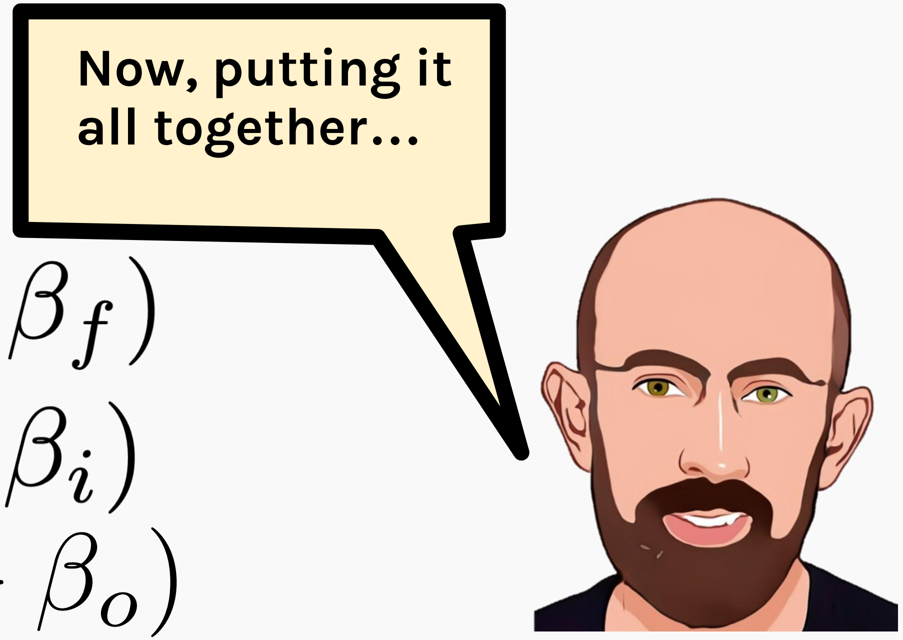
$$h_t = \mathrm{o_t} \odot \tanh(c_t)$$

$$\tilde{c}_t = \tanh(X_t V + \quad h_{t-1} U + \beta_1)$$

$$c_t = \mathrm{f}_\mathrm{t} \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$h_t = \mathrm{o}_\mathrm{t} \odot \tanh(c_t)$$

$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f \mathbf{X}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \beta_f\right)$$
$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i \mathbf{X}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \beta_i\right)$$
$$\mathbf{o}_t = \sigma\left(\mathbf{V}_o \mathbf{X}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \beta_o\right)$$



# GRU++ → LSTM

## Where...

$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f \mathbf{X}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \beta_f\right)$$

$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i \mathbf{X}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \beta_i\right)$$

$$\mathbf{o}_t = \sigma\left(\mathbf{V}_o \mathbf{X}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \beta_o\right)$$

# LSTM

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{X}_t \mathbf{V} + \mathbf{h}_{t-1}\mathbf{U} + \beta_1\right)$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{\mathbf{c}}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)$$


$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f \mathbf{X}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \beta_f\right)$$

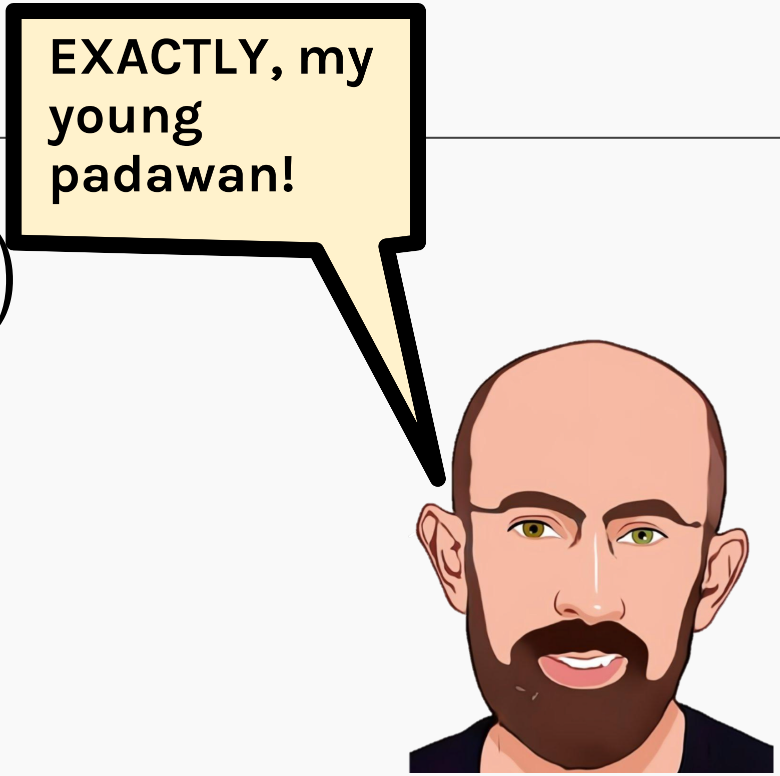
$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i \mathbf{X}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \beta_i\right)$$

$$\mathbf{o}_t = \sigma\left(\mathbf{V}_o \mathbf{X}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \beta_o\right)$$

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{X}_t\mathbf{V} + \mathbf{h}_{t-1}\mathbf{U} + \beta_1\right)$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{\mathbf{c}}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)$$





$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f\mathbf{X}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \beta_f\right)$$

$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i\mathbf{X}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \beta_i\right)$$

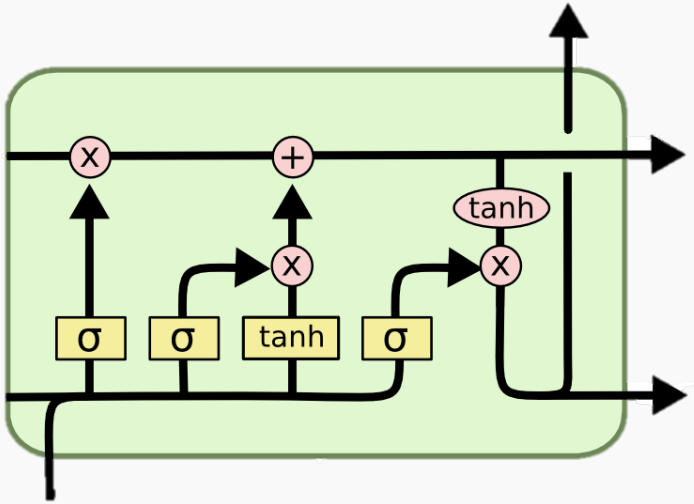
$$\mathbf{o}_t = \sigma\left(\mathbf{V}_o\mathbf{X}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \beta_o\right)$$

# LSTM

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{X}_t \mathbf{V} + \mathbf{h}_{t-1} \mathbf{U} + \beta_1\right)$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{\mathbf{c}}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)$$




$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f \mathbf{X}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \beta_f\right)$$

$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i \mathbf{X}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \beta_i\right)$$

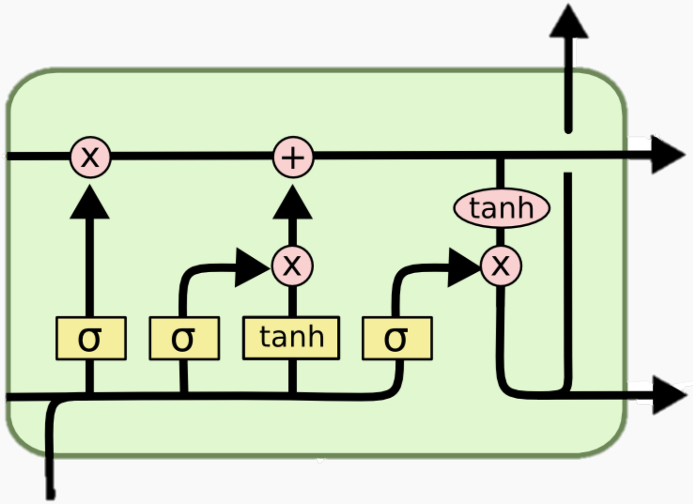
$$\mathbf{o}_t = \sigma\left(\mathbf{V}_o \mathbf{X}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \beta_o\right)$$

# LSTM

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{X}_t\mathbf{V} + \mathbf{h}_{t-1}\mathbf{U} + \beta_1\right)$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{\mathbf{c}}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)$$




$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f\mathbf{X}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \beta_f\right)$$

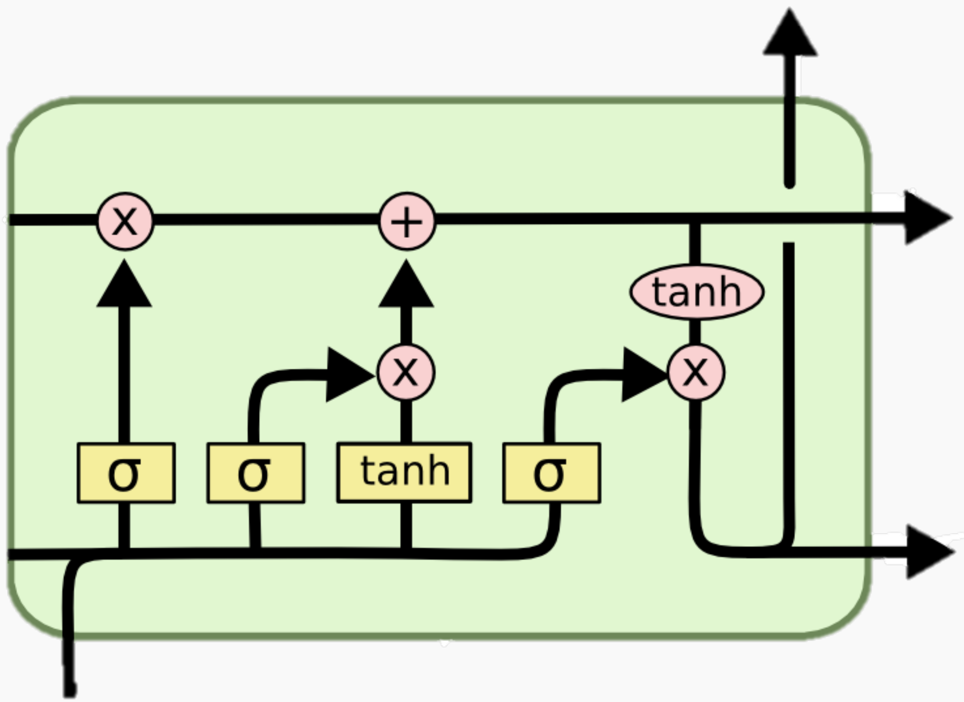
$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i\mathbf{X}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \beta_i\right)$$

$$\mathbf{o}_t = \sigma\left(\mathbf{V}_o\mathbf{X}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \beta_o\right)$$

# LSTM

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{X}_t\mathbf{V} + \mathbf{h}_{t-1}\mathbf{U} + \beta_1\right)$$
$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{\mathbf{c}}_t$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)$$

$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f\mathbf{X}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \beta_f\right)$$
$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i\mathbf{X}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \beta_i\right)$$
$$\mathbf{o}_t = \sigma\left(\mathbf{V}_o\mathbf{X}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \beta_o\right)$$



# LSTM

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{X}_t \mathbf{V} + \mathbf{h}_{t-1}\mathbf{U} + \beta_1\right)$$

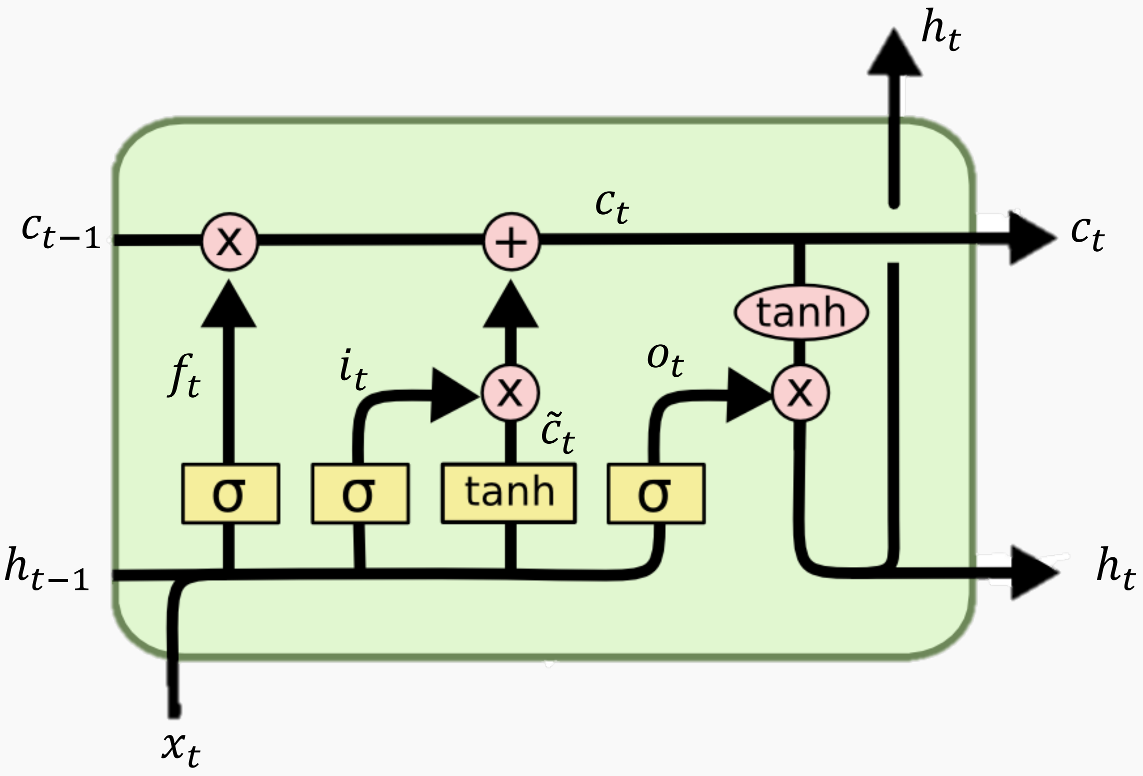
$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{\mathbf{c}}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)$$

$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f \mathbf{X}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \beta_f\right)$$

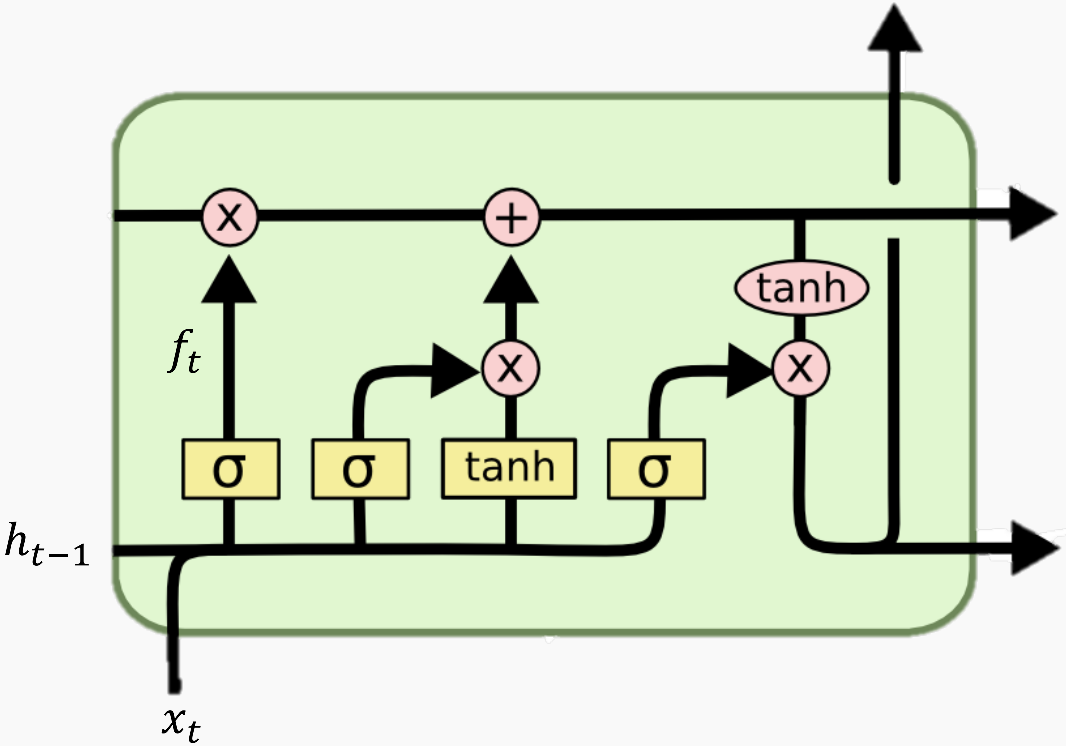
$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i \mathbf{X}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \beta_i\right)$$

$$\mathbf{o}_t = \sigma\left(\mathbf{V}_o \mathbf{X}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \beta_o\right)$$



# LSTM

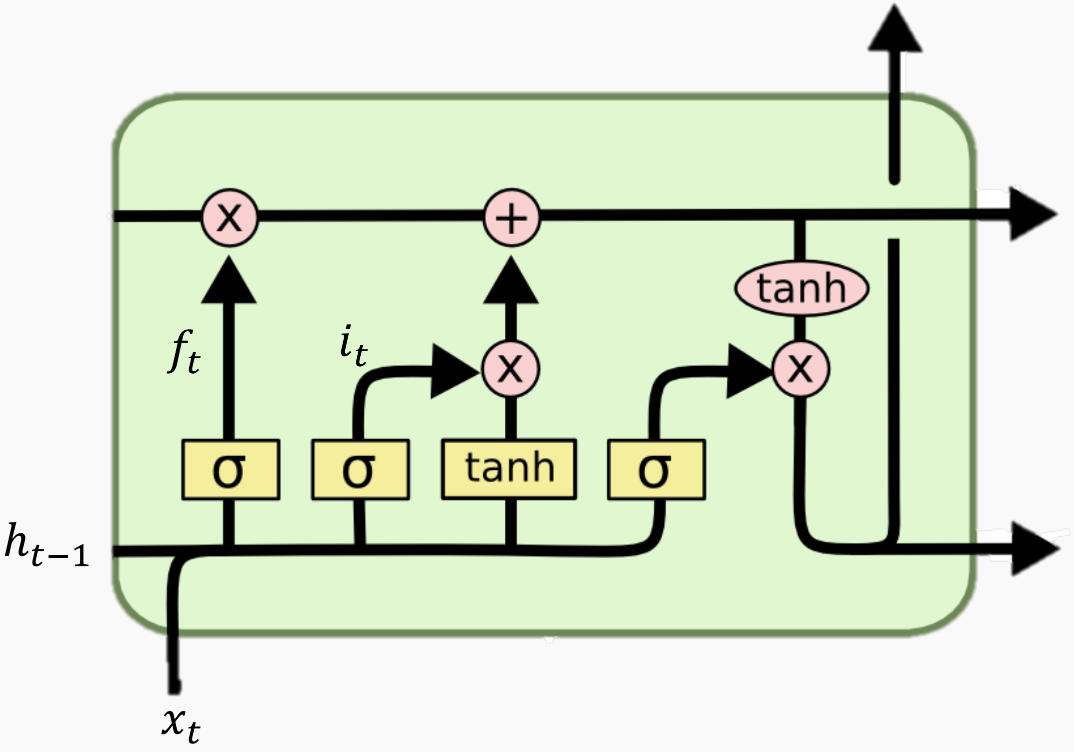
$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f \mathbf{X}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \beta_f\right)$$



# LSTM

$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f\mathbf{X}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \beta_f\right)$$

$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i\mathbf{X}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \beta_i\right)$$

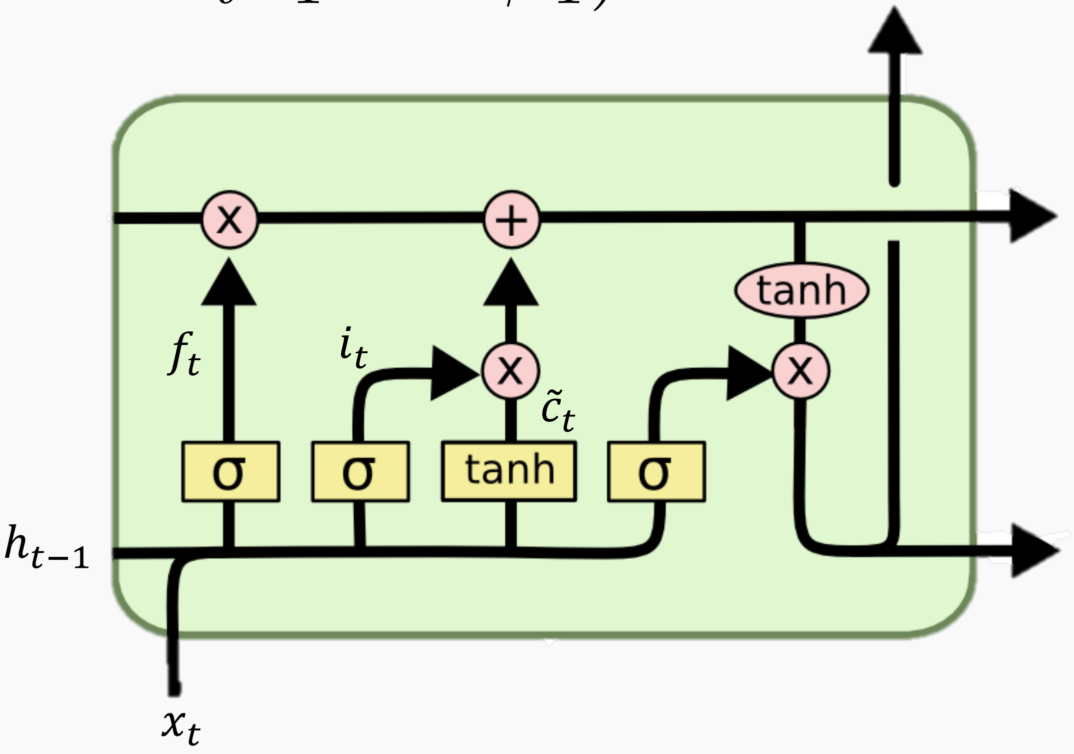


# LSTM

$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f \mathbf{X}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \beta_f\right)$$

$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i \mathbf{X}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \beta_i\right)$$

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{X}_t \mathbf{V} + \mathbf{h}_{t-1} \mathbf{U} + \beta_1\right)$$



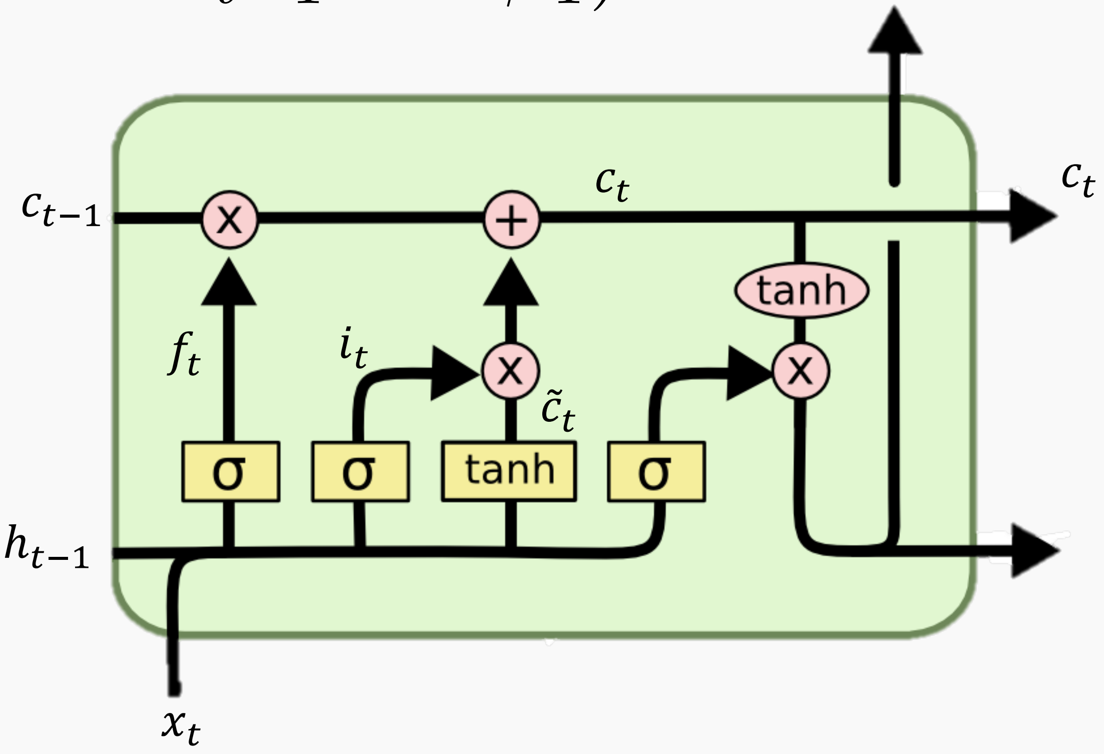
# LSTM

$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f \mathbf{X}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \beta_f\right)$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{\mathbf{c}}_t$$

$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i \mathbf{X}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \beta_i\right)$$

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{X}_t \mathbf{V} + \mathbf{h}_{t-1} \mathbf{U} + \beta_1\right)$$



# LSTM

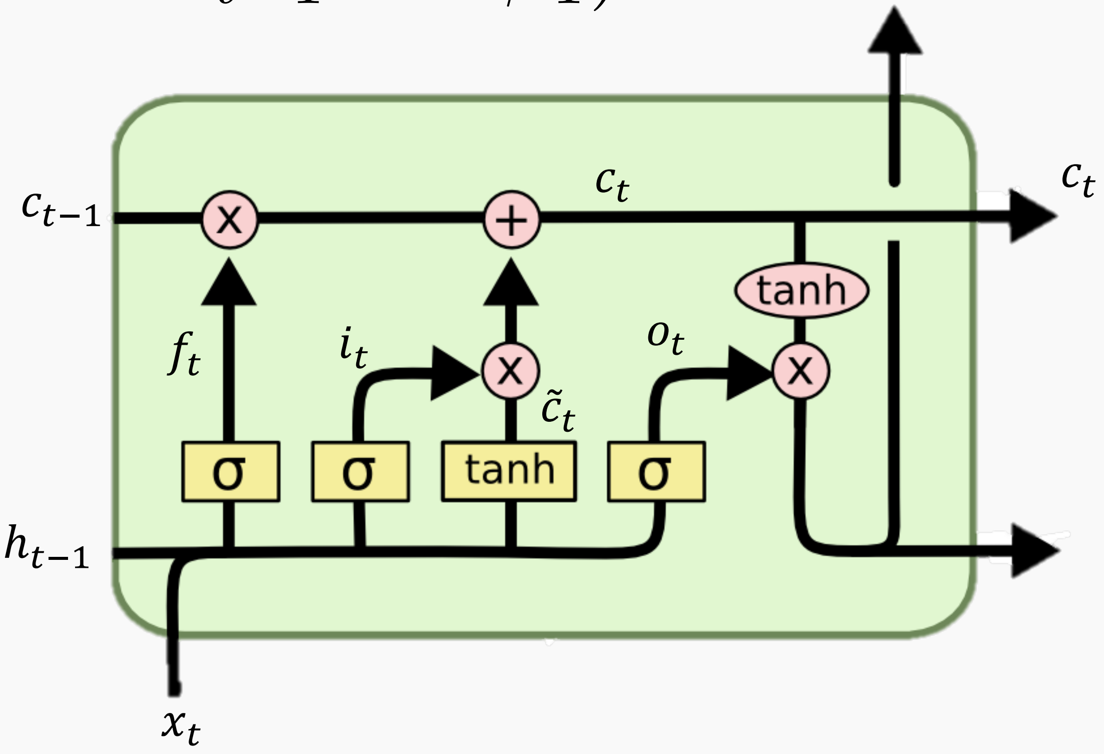
$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f \mathbf{X}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \beta_f\right)$$

$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i \mathbf{X}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \beta_i\right)$$

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{X}_t \mathbf{V} + \mathbf{h}_{t-1} \mathbf{U} + \beta_1\right)$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{\mathbf{c}}_t$$

$$\mathbf{o}_t = \sigma\left(\mathbf{V}_o \mathbf{X}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \beta_o\right)$$



# LSTM

$$\mathbf{f}_t = \sigma\left(\mathbf{V}_f\mathbf{X}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \beta_f\right)$$

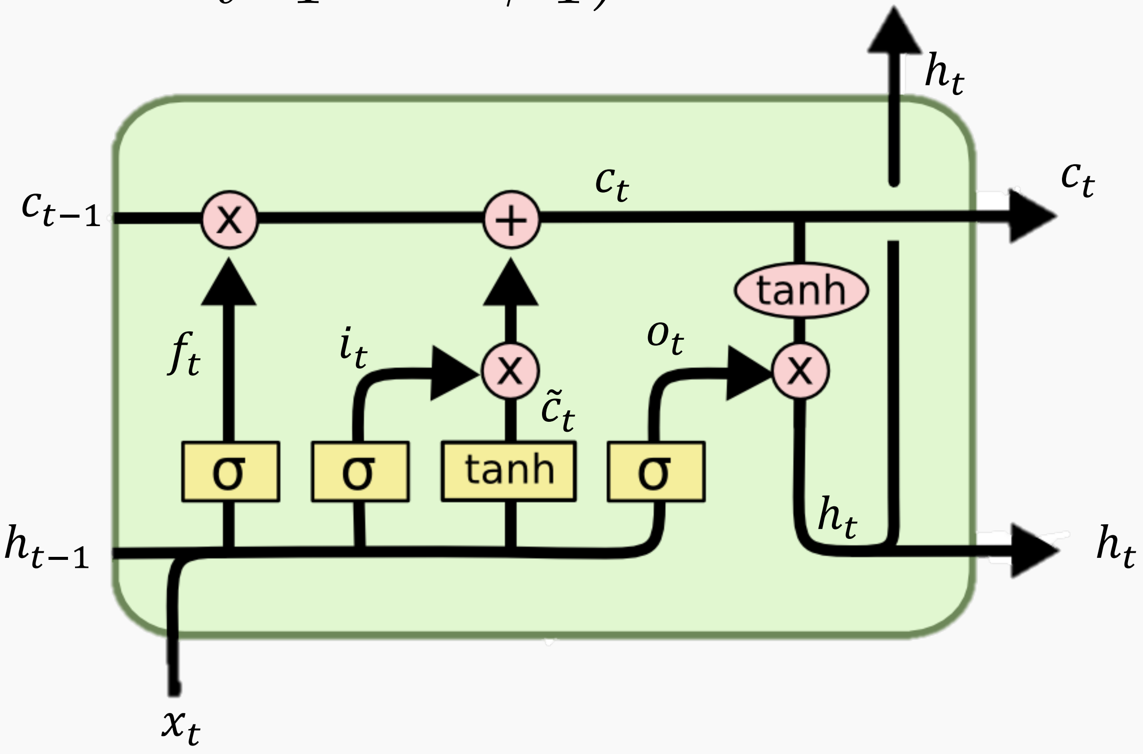
$$\mathbf{i}_t = \sigma\left(\mathbf{V}_i\mathbf{X}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \beta_i\right)$$

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{X}_t\mathbf{V} + \mathbf{h}_{t-1}\mathbf{U} + \beta_1\right)$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{\mathbf{c}}_t$$

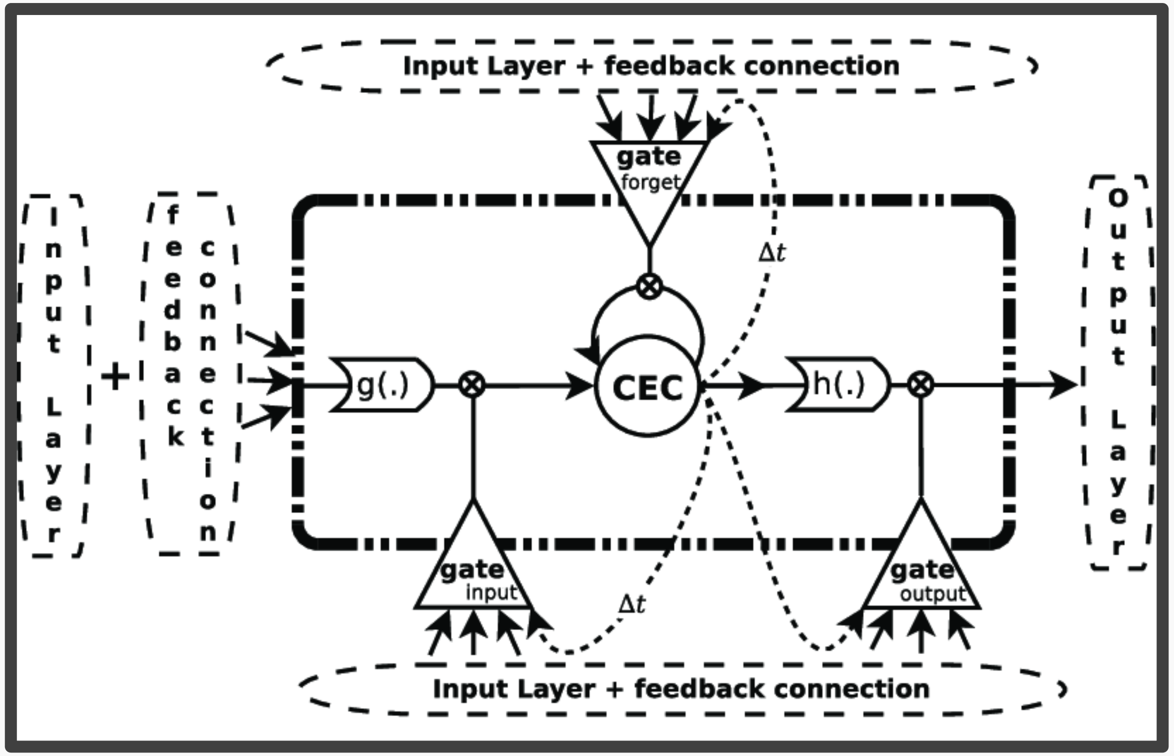
$$\mathbf{o}_t = \sigma\left(\mathbf{V}_o\mathbf{X}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \beta_o\right)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)$$



# Long short-term memory (LSTM)

- First introduced in **1995** to counter the vanishing gradient problem.
- Training was done using a mixture of Real Time Recurrent Learning and Backpropagation Through Time.
- Underwent several major modifications including addition of *forget gate* , *peephole connections* etc.
- In the last two decades, several other variants were introduced with mixed results



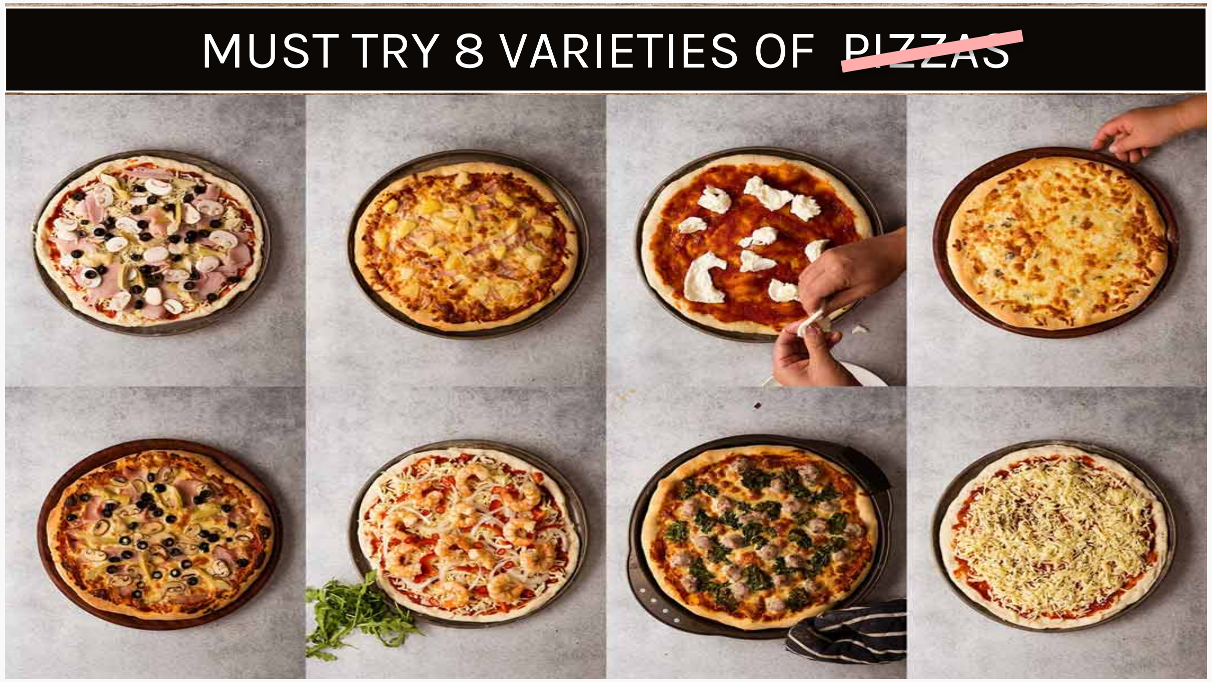

- *First prototype of LSTM cell*

# Long short-term memory (LSTM)

After its introduction in 1995, here are the eight popular types of LSTM variants other than the vanilla version:

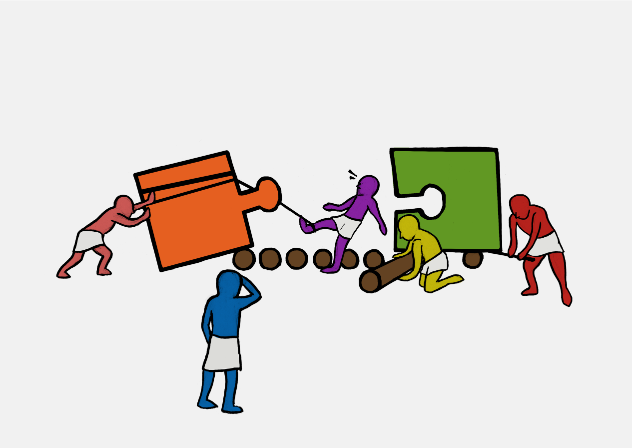
- **NIG – No input gate**
- **NFG - No forget gate**
- **NOG – No output gate**
- **NIAF – No input activation**
- **NOAF – No output activation**
- **CIFG – Coupled input/forget gate**
- **NP – No peepholes**
- **FGR – Full gate recurrence**

LSTMs

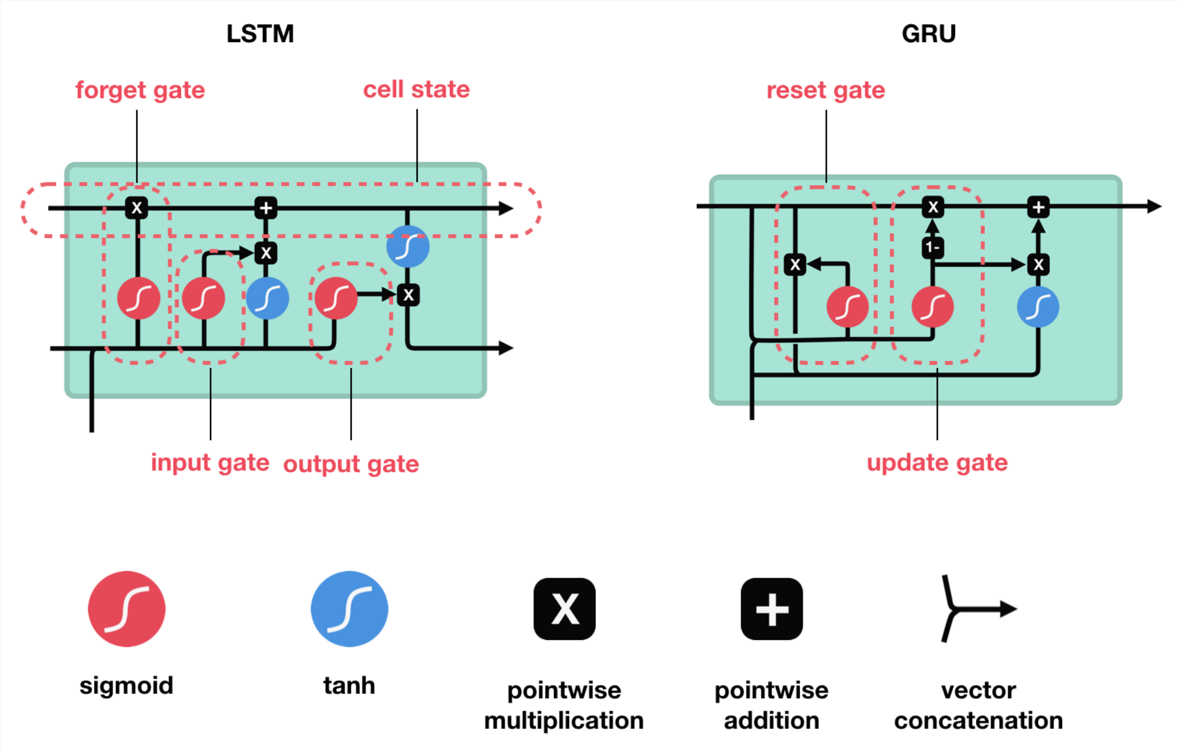



Please refer to the paper ***LSTM: A Search Space Odyssey*** for a thorough analysis of all the variants

# **Exercise:** LSTM v/s GRU

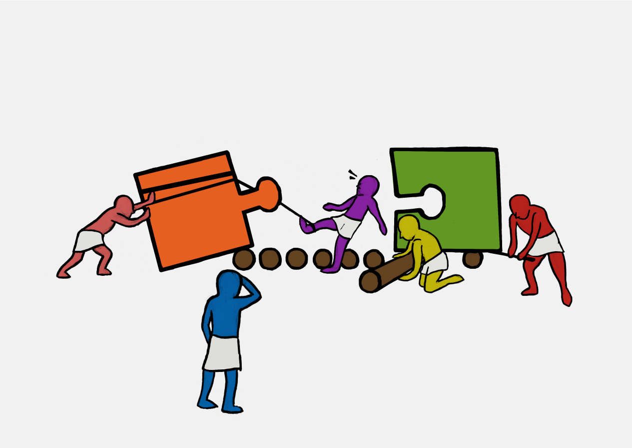


The goal of this exercise is to compare the performance between two popular gating methods, i.e LSTM and GRUs:



# **Exercise:** LSTM v/s GRU

As in previous exercises you need to add an embedding layer.



## EMBEDDING LAYER