Long Short Term Memory (LSTM)

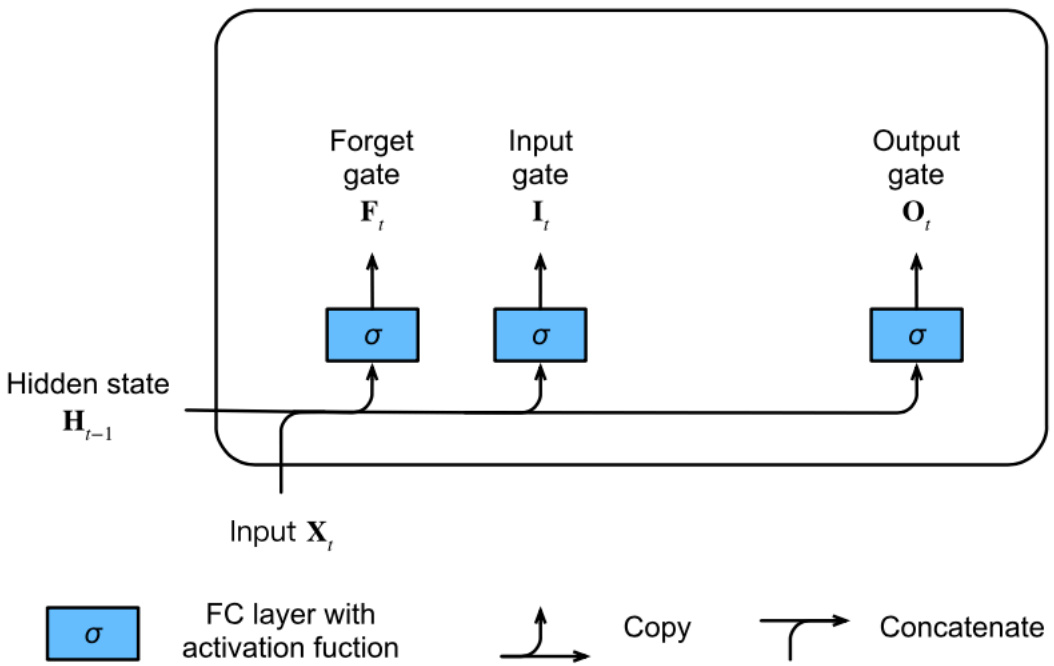
- The challenge to address long-term information preservation and short-term input skipping in latent variable models has existed for a long time.
- One of the earliest approaches to address this was the LSTM
    - More complex than GRU but predates GRU by almost two decades.

Long Short Term Memory (LSTM)

- Main feature is a new structure called the (memory) *cell* which
    - has the same shape as the hidden state.
    - is a fancy version of a hidden state, engineered to record additional information.
- To control a memory cell 3 gates are used:
    - The *output* gate is used to read out the entries from the cell.
    - The *input* gate is used to read data into the cell.
    - The *forget* gate is used to reset the contents of the cell.

Input Gates, Forget Gates, and Output Gates

- Similarly to GRU gates, the following figure shows how the values of input, forget and output gates are computed a each time step:
  - They are functions of $\mathbf{X}_t$ and $\mathbf{H}_{t-1}$.
  - The function output is given by a fully connected layer with a sigmoid as its activation function.



- Assume, for a given time step t, the minibatch input is $\mathbf{X}_t \in \mathbb{R}^{n \times d}$ (number of examples: n, number of inputs: d) and the hidden state of the last time step is $\mathbf{H}_{t-1} \in \mathbb{R}^{n \times h}$ (number of hidden states: h).

- Then, the input gate $\mathbf{I}_t \in \mathbb{R}^{n \times h}$, the forget gate $\mathbf{F}_t \in \mathbb{R}^{n \times h}$ and the output gate $\mathbf{O}_t \in \mathbb{R}^{n \times h}$ are computed as follows:

$$
\begin{aligned}
\mathbf{I}_t &= \sigma(\mathbf{X}_t \mathbf{W}_{xi} + \mathbf{H}_{t-1} \mathbf{W}_{hi} + \mathbf{b}_i),\\
\mathbf{F}_t &= \sigma(\mathbf{X}_t \mathbf{W}_{xf} + \mathbf{H}_{t-1} \mathbf{W}_{hf} + \mathbf{b}_f),\\
\mathbf{O}_t &= \sigma(\mathbf{X}_t \mathbf{W}_{xo} + \mathbf{H}_{t-1} \mathbf{W}_{ho} + \mathbf{b}_o),
\end{aligned}
$$

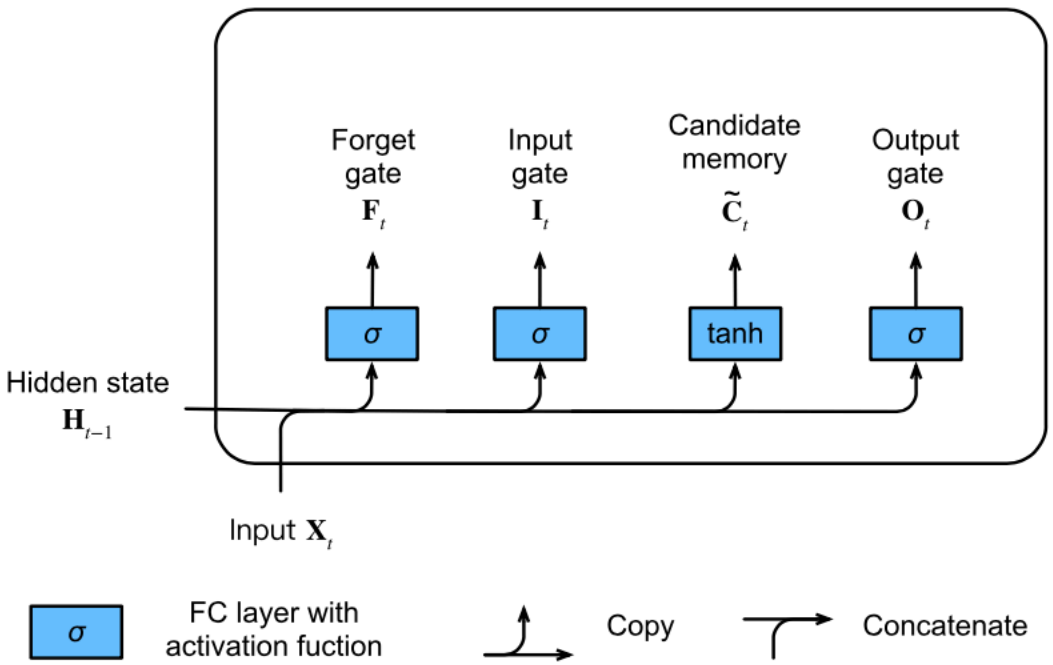
- $\mathbf{W}_{xi}, \mathbf{W}_{xf}, \mathbf{W}_{xo} \in \mathbb{R}^{d \times h}$ and $\mathbf{W}_{hi}, \mathbf{W}_{hf}, \mathbf{W}_{ho} \in \mathbb{R}^{h \times h}$ are weight parameters and $\mathbf{b}_i, \mathbf{b}_f, \mathbf{b}_o \in \mathbb{R}^{1 \times h}$ are bias parameters.

Candidate Memory Cell

- Akin to candidate hidden state in GRU.
- The *candidate* memory cell $\tilde{\mathbf{C}}_t \in \mathbb{R}^{n \times h}$ at time step t is computed from:

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{X}_t \mathbf{W}_{xc} + \mathbf{H}_{t-1} \mathbf{W}_{hc} + \mathbf{b}_c)$$

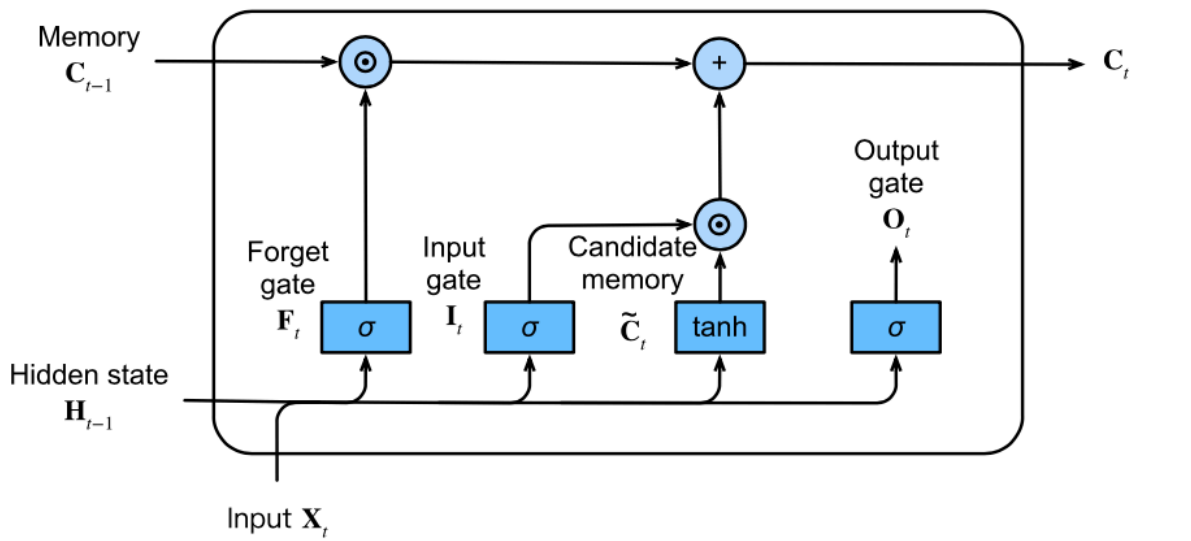
where $\mathbf{W}_{xc} \in \mathbb{R}^{d \times h}$ and $\mathbf{W}_{hc} \in \mathbb{R}^{h \times h}$ are weight parameters and $\mathbf{b}_c \in \mathbb{R}^{1 \times h}$ is a bias parameter.

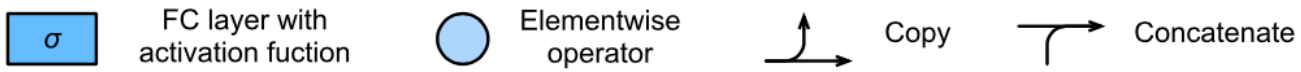


Memory Cell

- In GRUs, we had a single mechanism to govern input and forgetting.
- In LSTMs we have two parameters:
  - input gate $\mathbf{I}_t$ defines how much we take new data into account from $\tilde{\mathbf{C}}_t$
  - forget gate $\mathbf{F}_t$ defines how much of the old memory cell content $\mathbf{C}_{t-1} \in \mathbb{R}^{n \times h}$ is retained.
- Put together we have the following update equation:

$$\mathbf{C}_t = \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t.$$

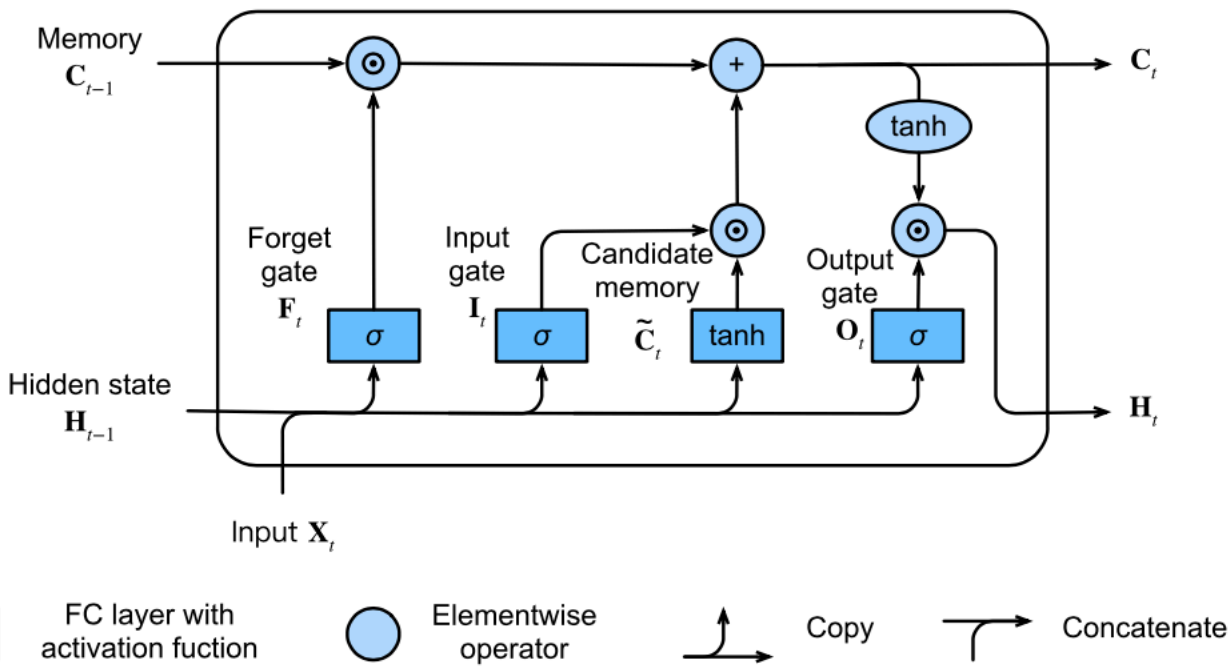




Hidden States

- Last, we need to define how to compute the hidden state $\mathbf{H}_t \in \mathbb{R}^{n \times h}$.
- In LSTM it is simply a gated version of the $\tanh$ of the memory cell.
- This is where the output gate comes into play:

$$\mathbf{H}_t = \mathbf{O}_t \odot \tanh(\mathbf{C}_t).$$



Summary

- LSTMs have three types of gates: input gates, forget gates, and output gates which control the flow of information.
- The hidden layer output of LSTM includes hidden states and memory cells. Only hidden states are passed into the output layer. Memory cells are entirely internal.
- LSTMs can cope with vanishing and exploding gradients.