Recurrent Neural Networks

Mass Street University
PER EDUCATIONEM PROGRESSUS

Shortcomings of Neural Networks

- Neural Networks struggle to deal with spatial and sequential data.
- They have no memory of the past. Only impressions on the network that develop an intuition.
- Fixed inputs and outputs. Weights are not shared and can not store information.

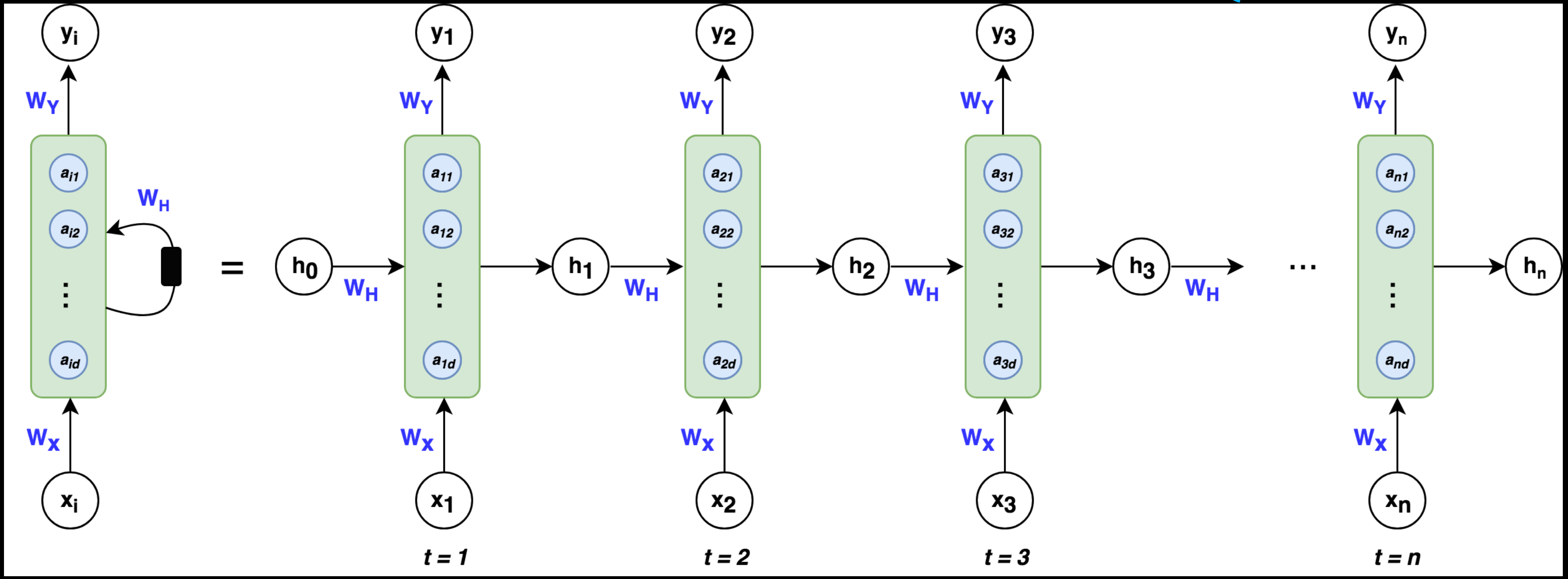
Recurrent Neural Networks

- RNNs are networks that reuse previous weights and information to continuously process information.
- This allows it to process sequential data and work with data that needs long term information, such as text data.

How Recurrent Neural Networks Work

- They have three unique parts. Inputs, Outputs, and State Vectors.
- The state vector and memory loop control what information is stored and the current information that is remembered.

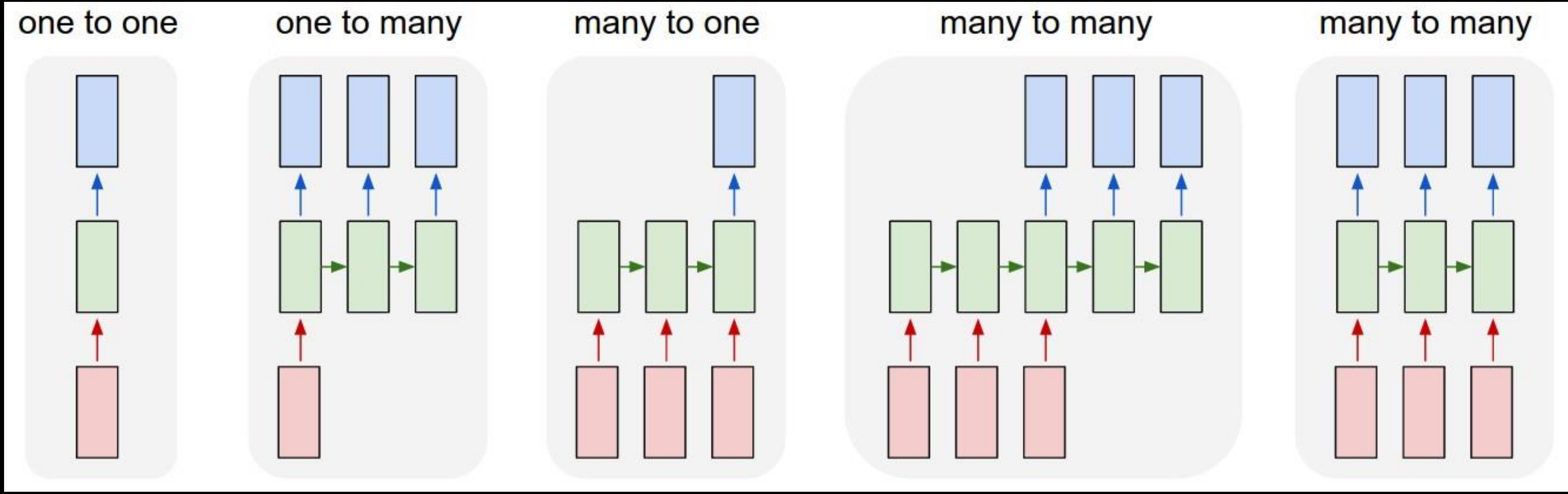
$$H = \tanh(Wh * H + Wx * X)$$



Mass Street University
PER EDUCATIONEM PROGRESSUS

RNN Uses

- Sequence data including One to One, One to Many, Many to One, and Many to Many.



Mass Street University
PER EDUCATIONEM PROGRESSUS

RNN uses

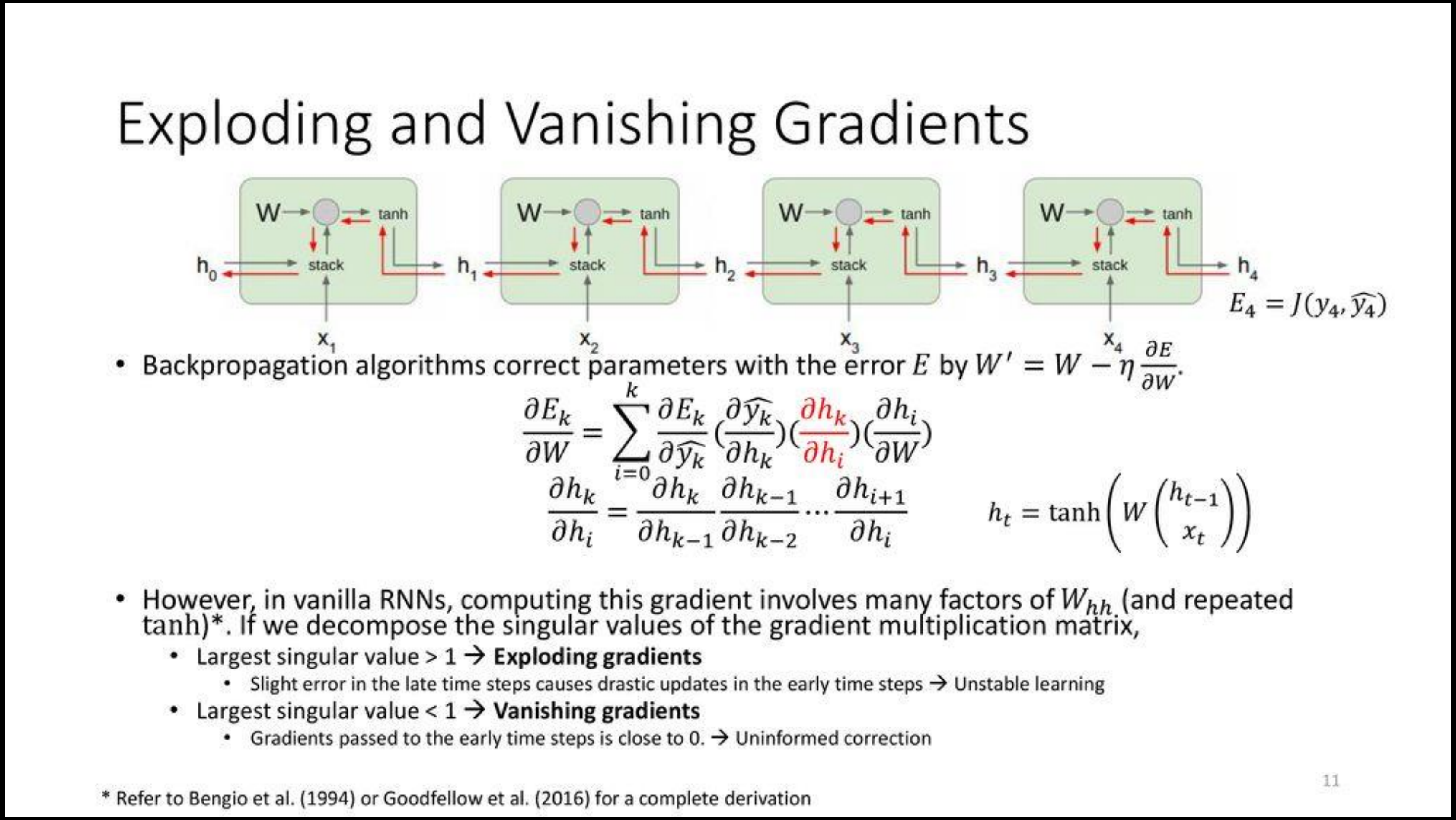
- Sequence data
- One to many operations
  - Single input many outputs
- Many to one operations
  - Single output classification on the sequence
- Many to many operations
  - Synched – an output at every step
  - Un-synched – delay between first input and output

Cont.

- RNNs can process non-sequential data to keep spatial information intact like a CNN.
- Applications include document generation, text translation, chat bots, law searches for legal firms, lead generation, recommendation machines, and series forecasting.

Limitations

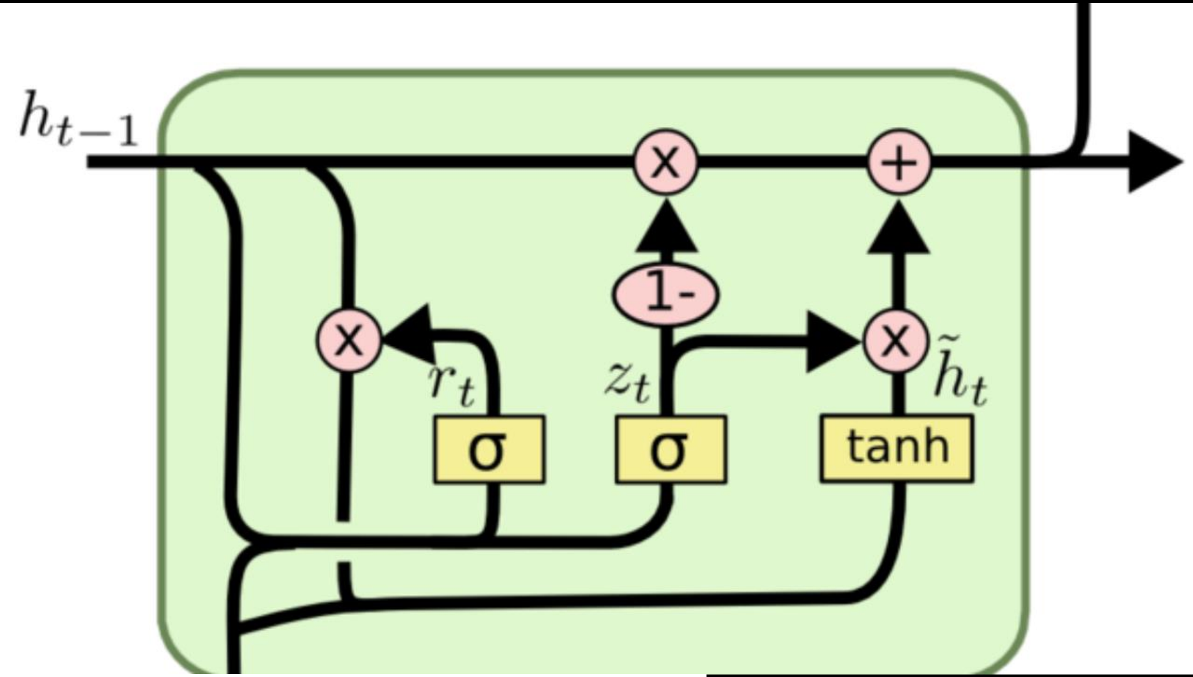
- Exploding and Vanishing Gradients.
- The length of the sequence (amount of memory) we can process.



Mass Street University
PER EDUCATIONEM PROGRESSUS

Gated Recurrent Networks.

- GRUs use two types of gates: Update and Reset gates. These gates are vectors that are multiplied by the input vectors.
- The update gate (Zt) keeps information and the reset gate (Rt) forgets information.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

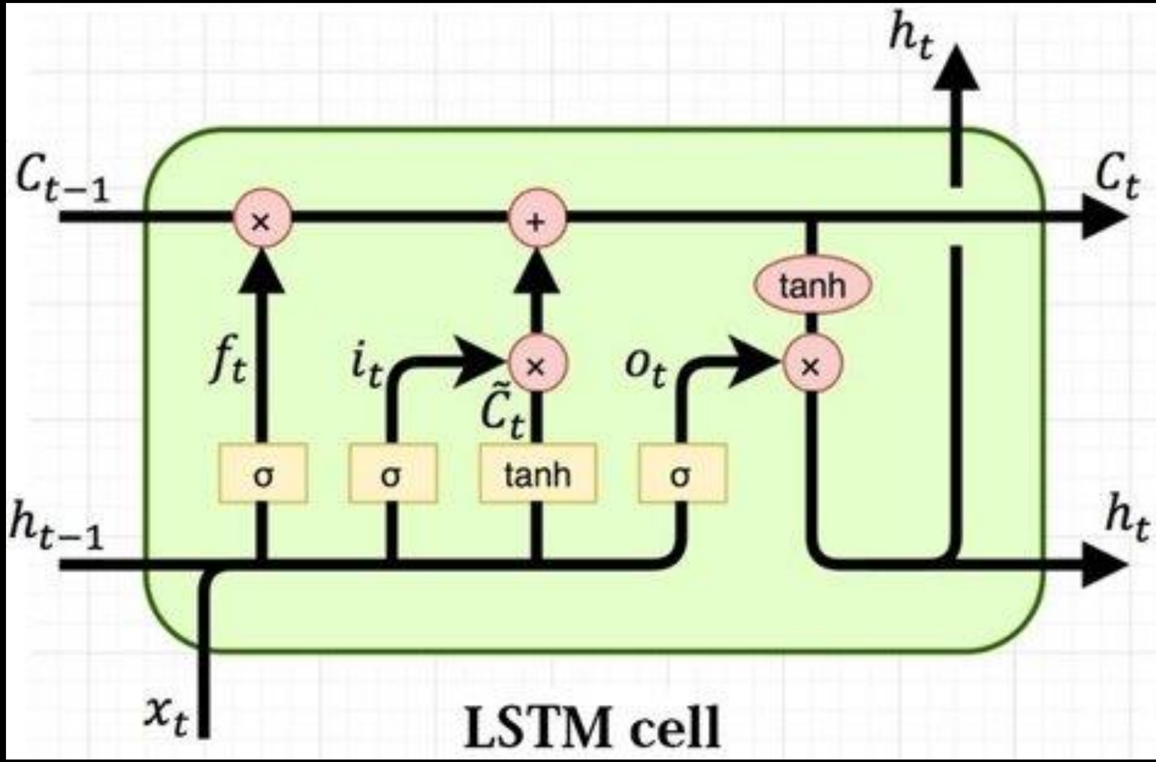
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

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

Long Short Term Memory Networks

- LSTMs use three different gates: Forget Gate, Input Gate, Output Gate. It also has a cell state that retains the past values.
- The gates regulate the flow of information in the RNN much like the GRU.

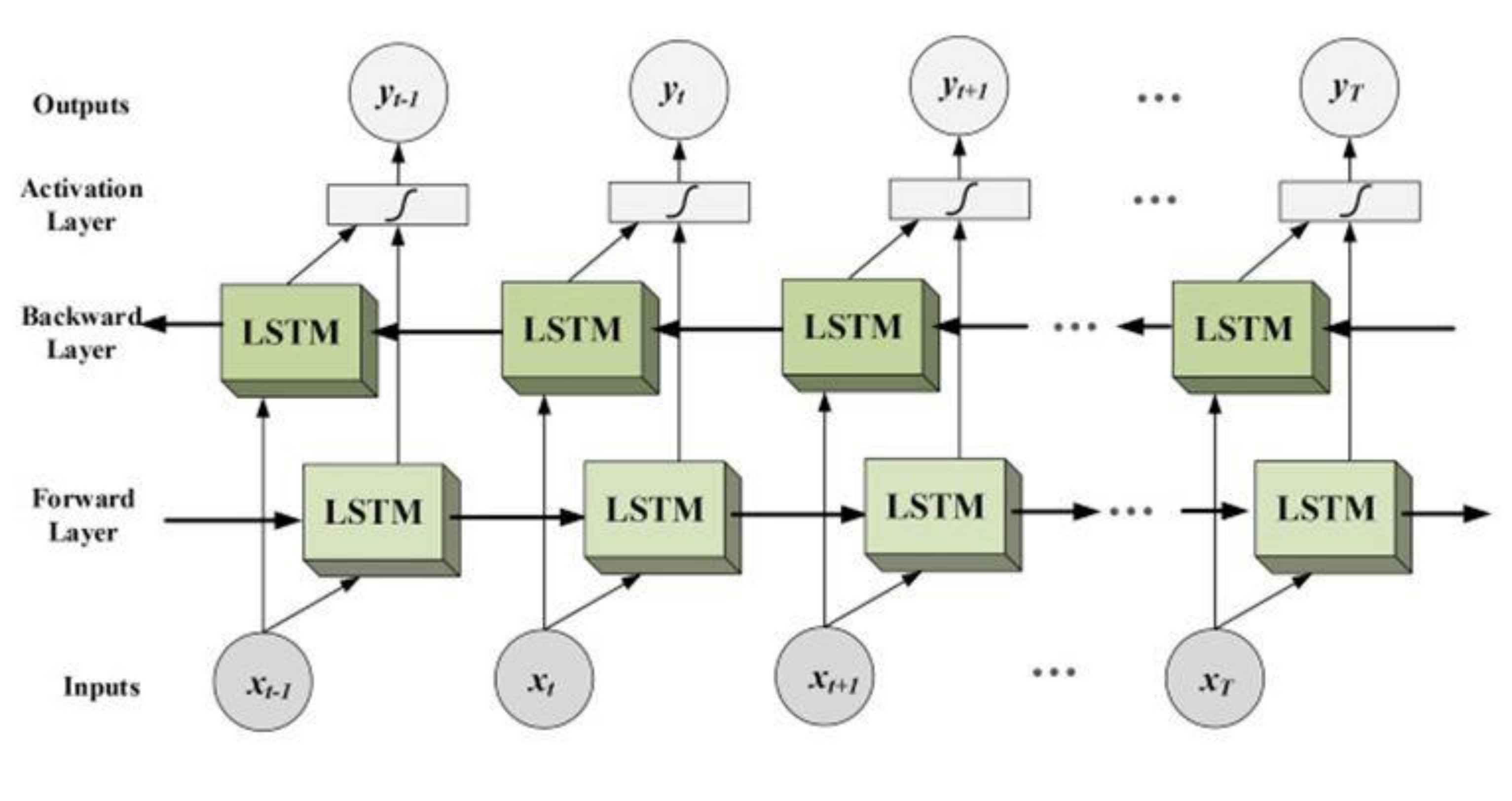
LSTM Cont.



$$i_t = \sigma\left(x_t U^i + h_{t-1} W^i\right)$$
$$f_t = \sigma\left(x_t U^f + h_{t-1} W^f\right)$$
$$o_t = \sigma\left(x_t U^o + h_{t-1} W^o\right)$$
$$\tilde{C}_t = \tanh\left(x_t U^g + h_{t-1} W^g\right)$$
$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right)$$
$$h_t = \tanh(C_t) * o_t$$

Bi-Directional RNNs

- They process information forward and backwards to better understand context and long-term dependencies.



Mass Street University
PER EDUCATIONEM PROGRESSUS