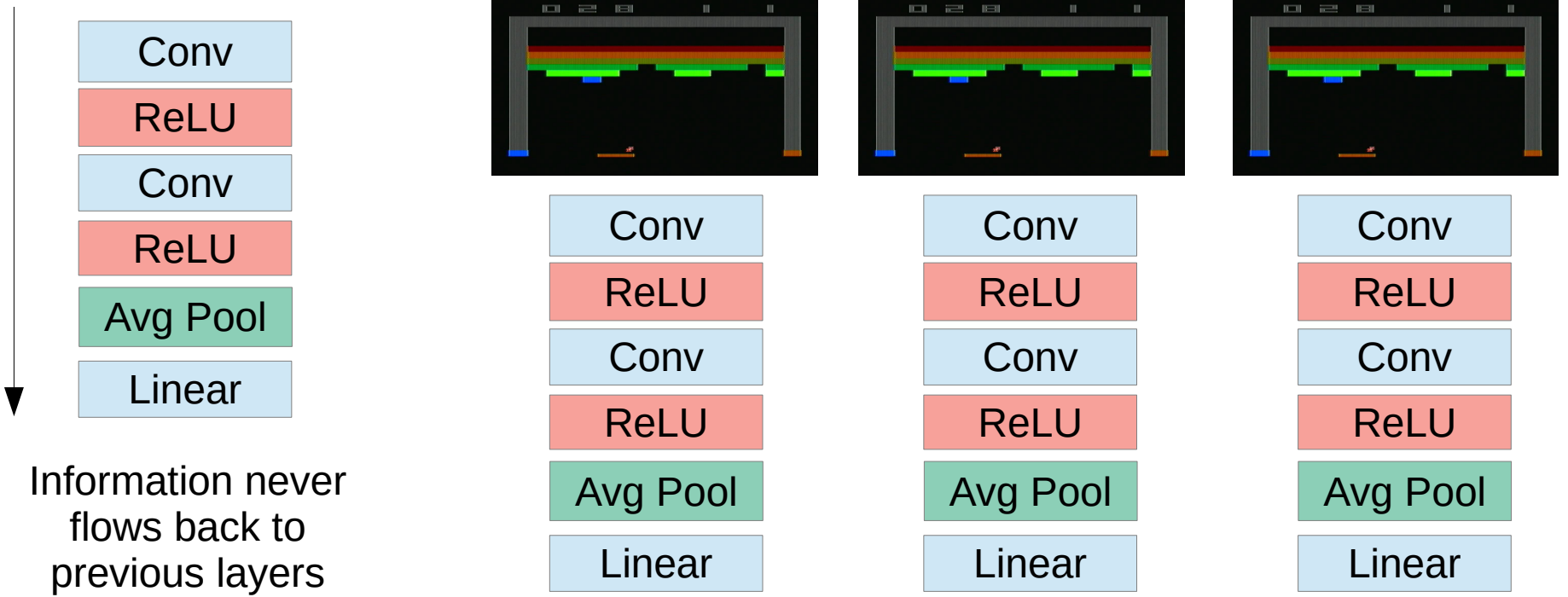
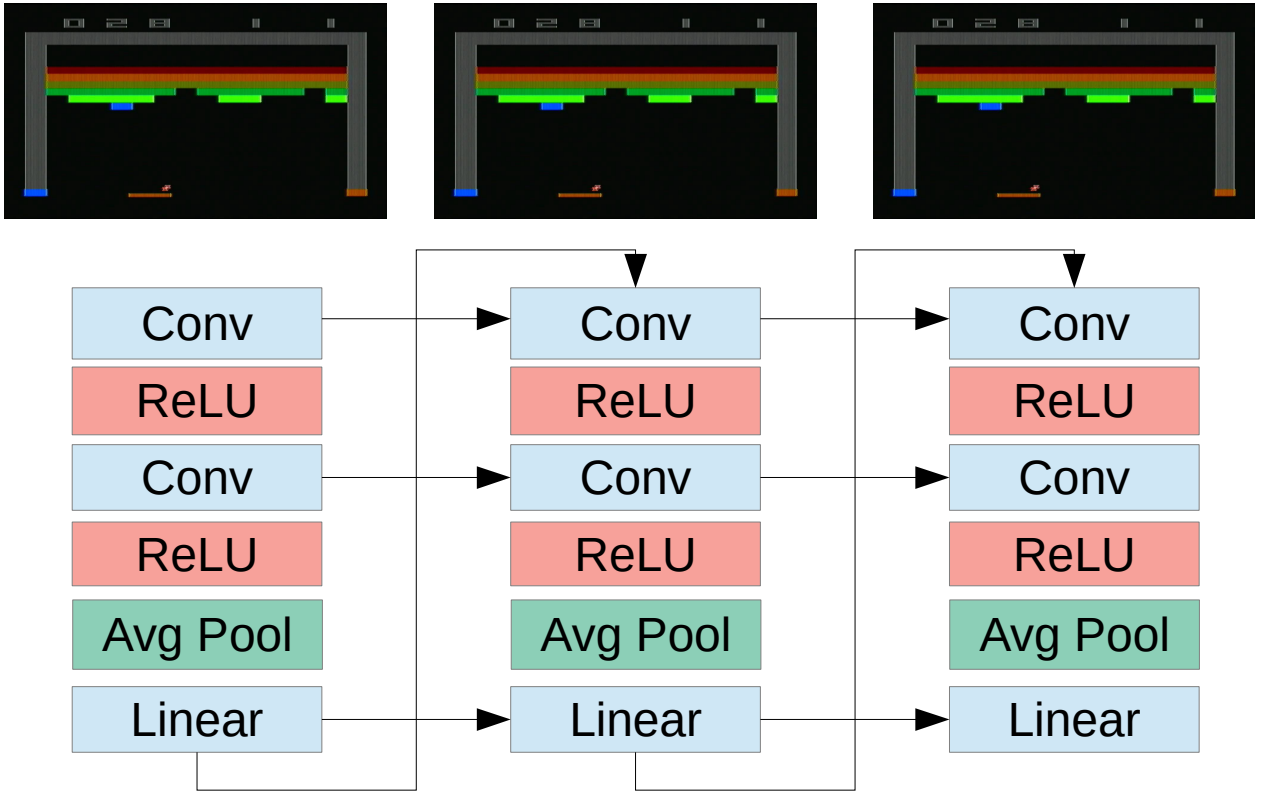
Sequence Models

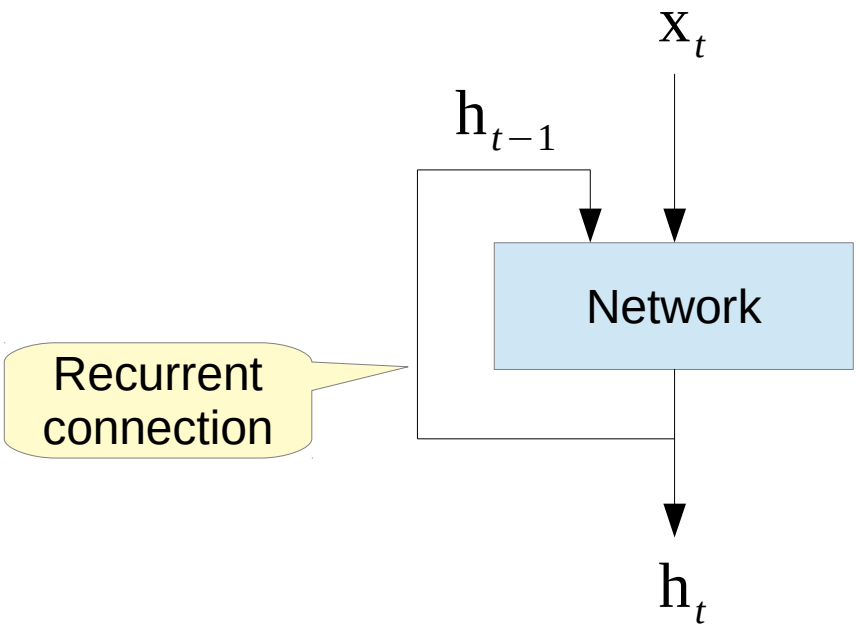
Feed-Forward Networks



Memory


Conv
ReLU
Conv
ReLU
Avg Pool
Linear
Conv
ReLU
Conv
ReLU
Avg Pool
Linear
Conv
ReLU
Conv
ReLU
Avg Pool
Linear

Recurrent Neural Network (RNN)



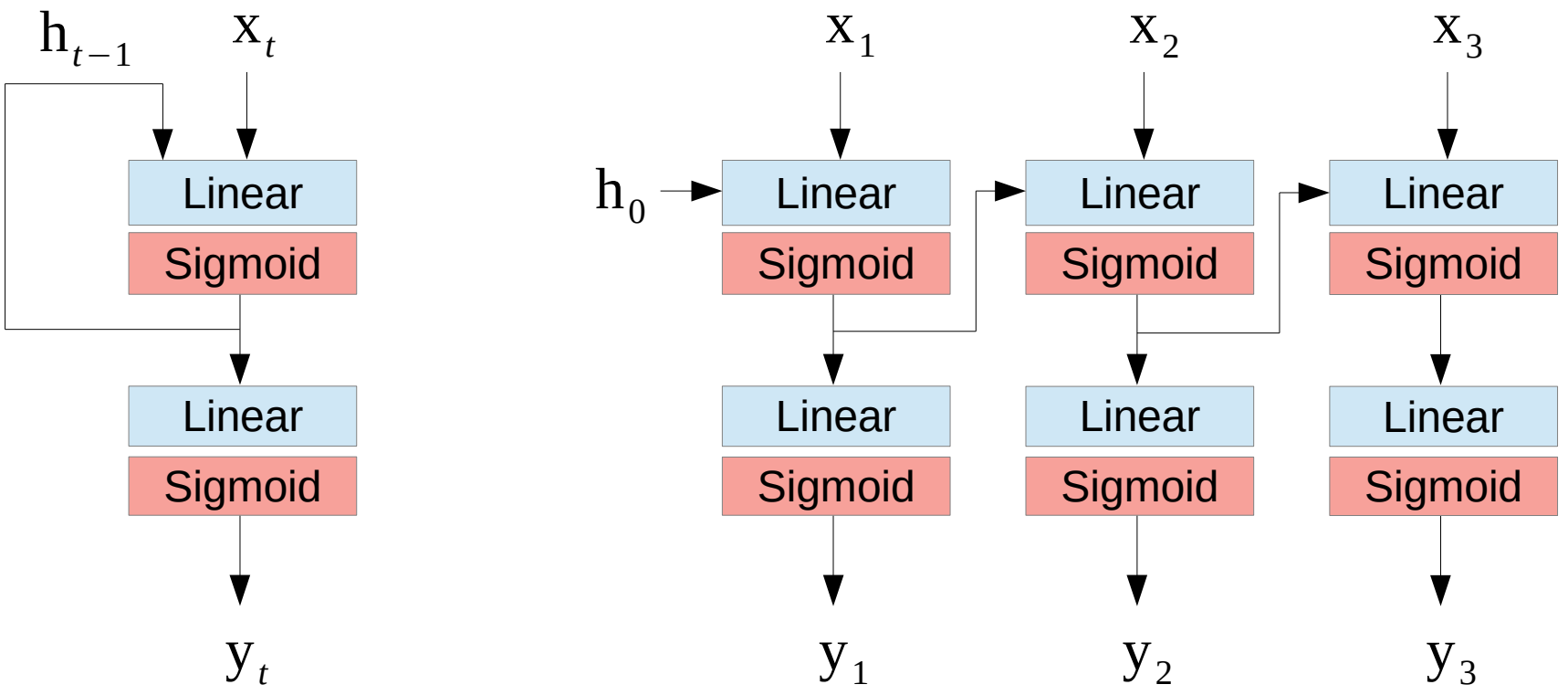
$$h_t = f_h(x_t, h_{t-1}, \theta_h)$$

Initial state: h_0

Zero

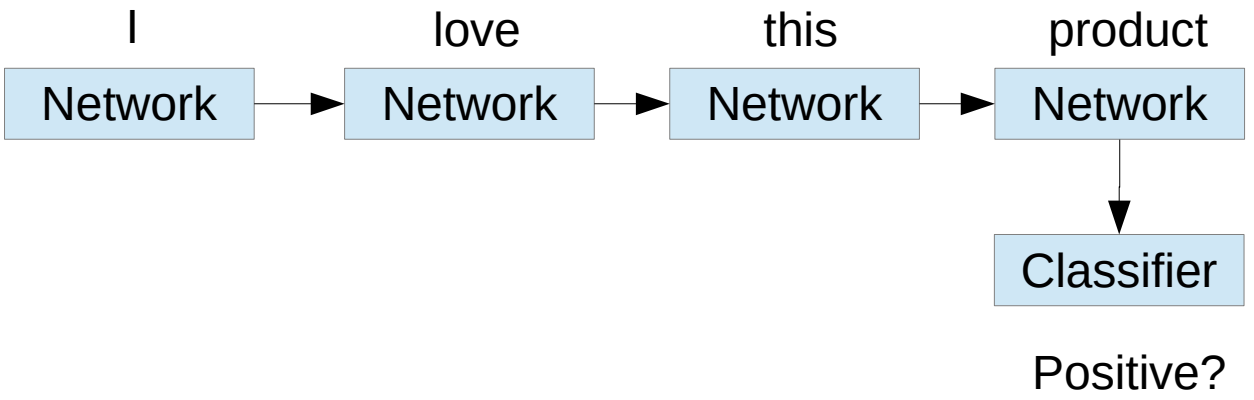
Learned

Unrolling RNNs



Applications

- Natural language processing
  - Either the input or output (or both) is a sequence of words
- Example: Is a review positive or negative?



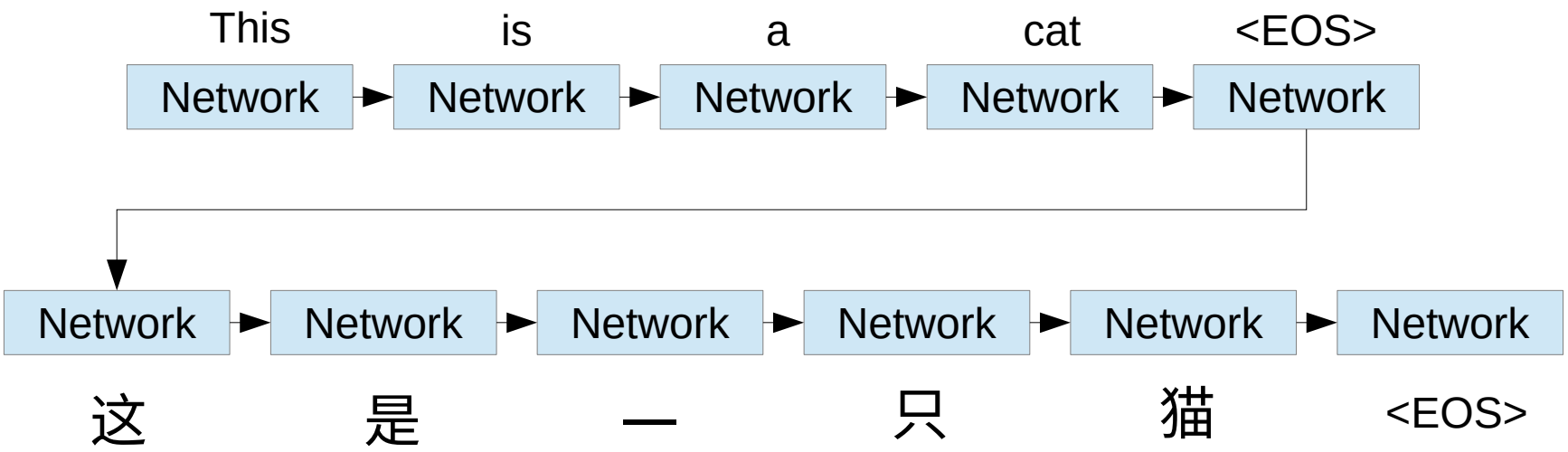
Example: Language Generation

Preprocessing

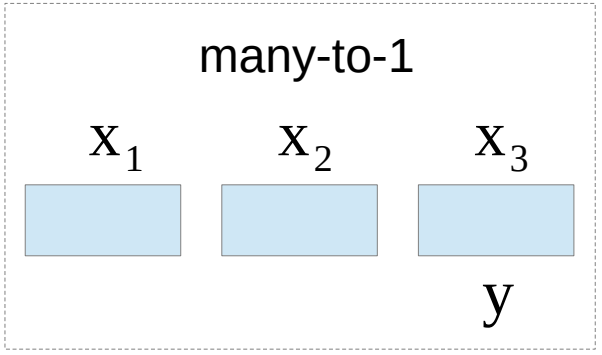
Network → Network → Network → Network

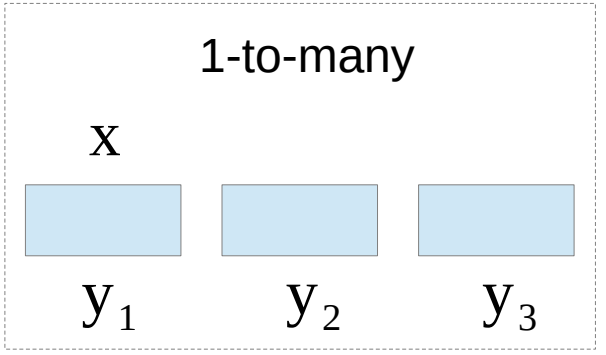
This is a game

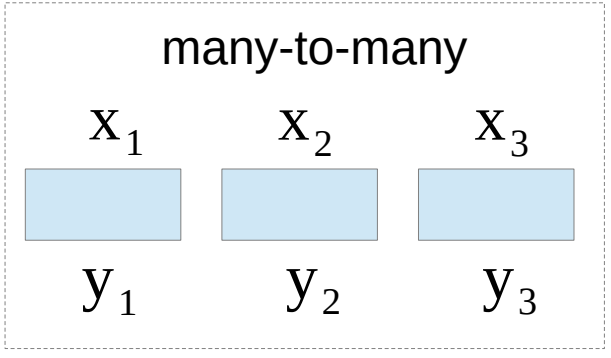
Example: Translation

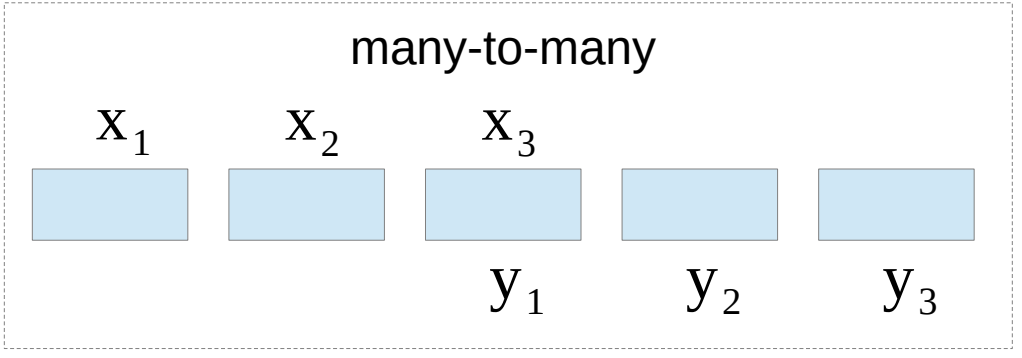


Kinds of RNN


many-to-1
x_1
x_2
x_3
y

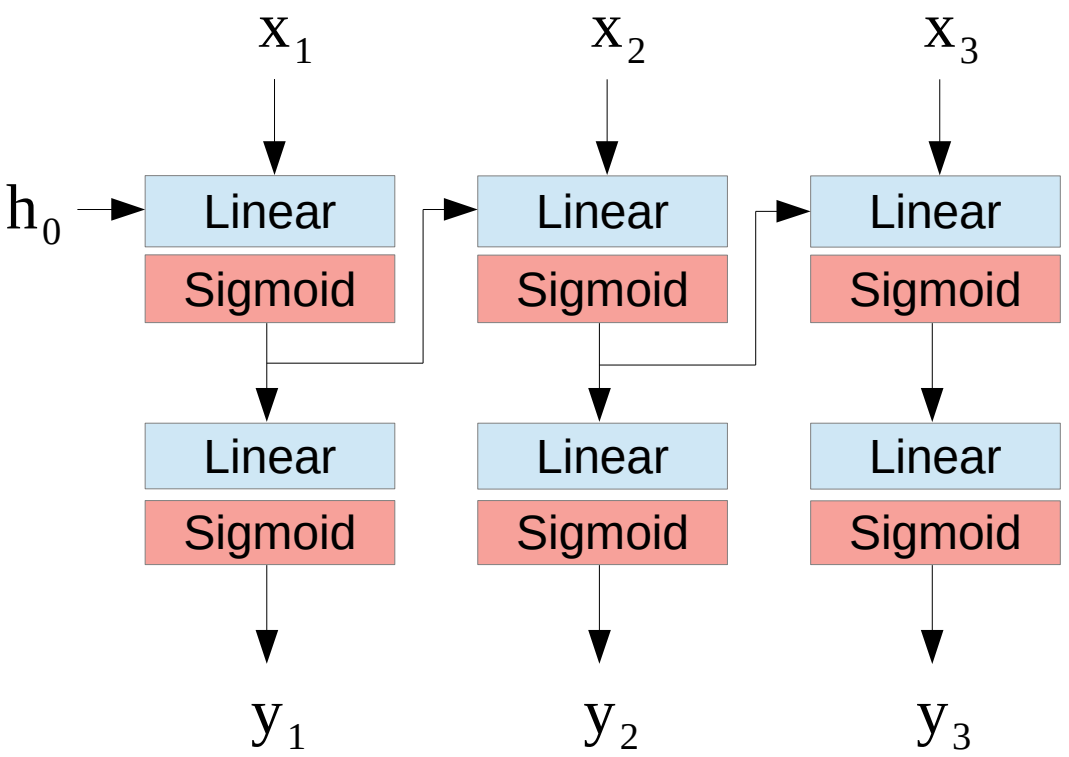

1-to-many
x
y_1
y_2
y_3


many-to-many
x_1
x_2
x_3
y_1
y_2
y_3

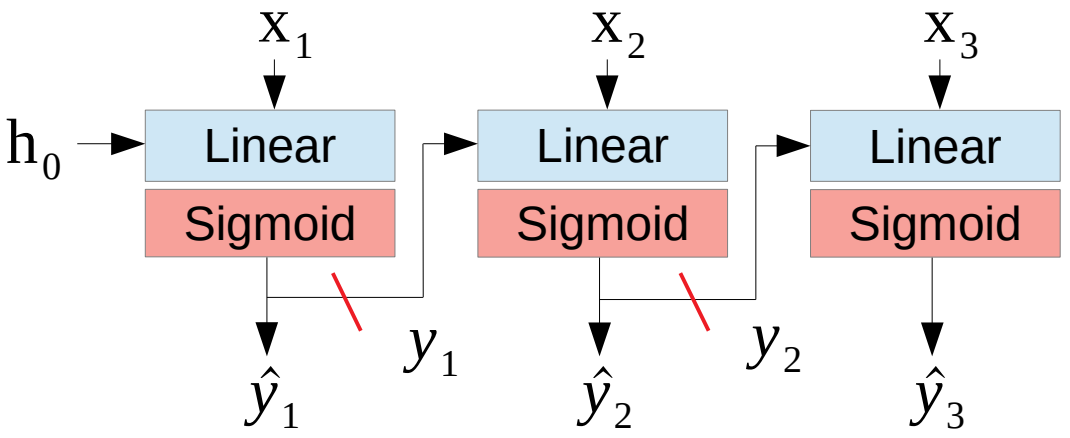

many-to-many
x_1
x_2
x_3
y_1
y_2
y_3

Training RNNs

- Unrolled RNNs have a feed-forward computation graph
- Regular backprop handles shared weights
- Long sequences leads to vanishing/exploding gradients



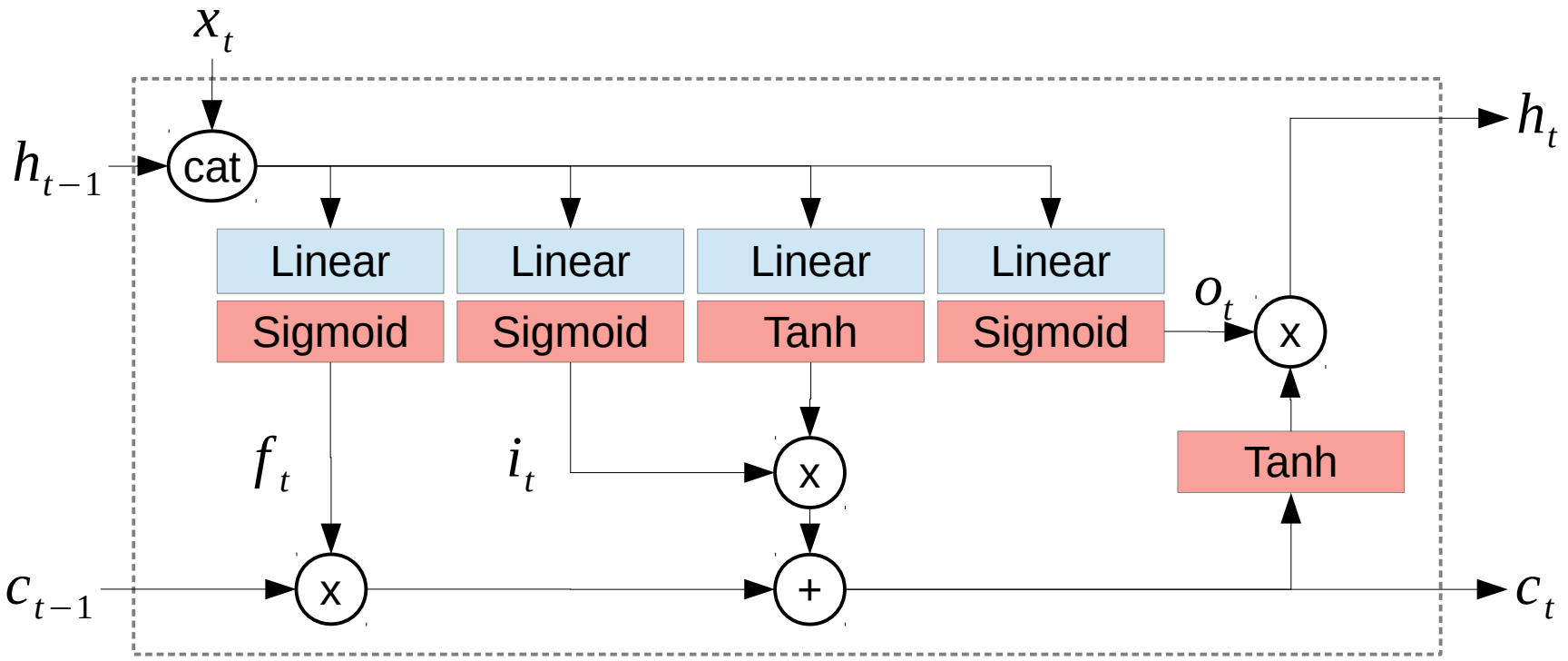
Vanishing/Exploding Gradients



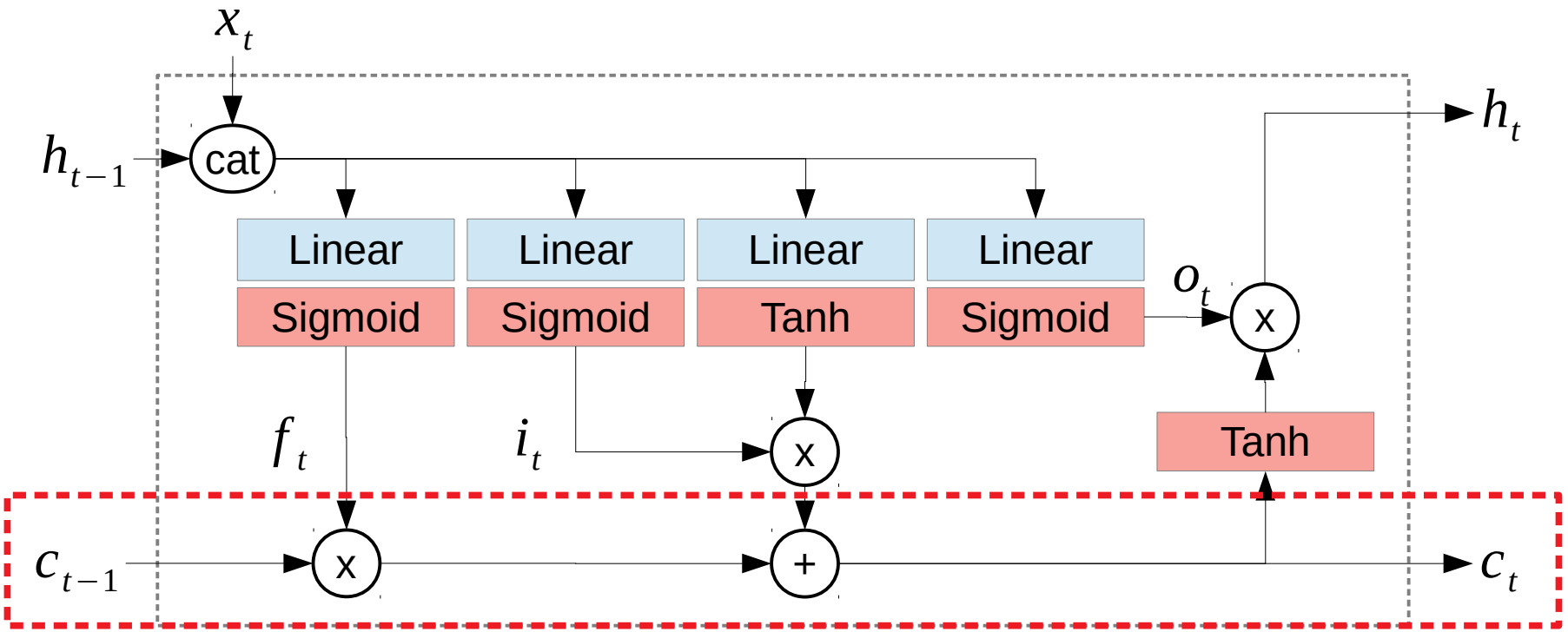
Exploding gradients: $\nabla L(\theta) = \min\left(1, \frac{\epsilon}{|\nabla L(\theta)|}\right) \nabla L(\theta)$

Vanishing gradients: must rely on network structure

Long Short-Term Memory (LSTM)



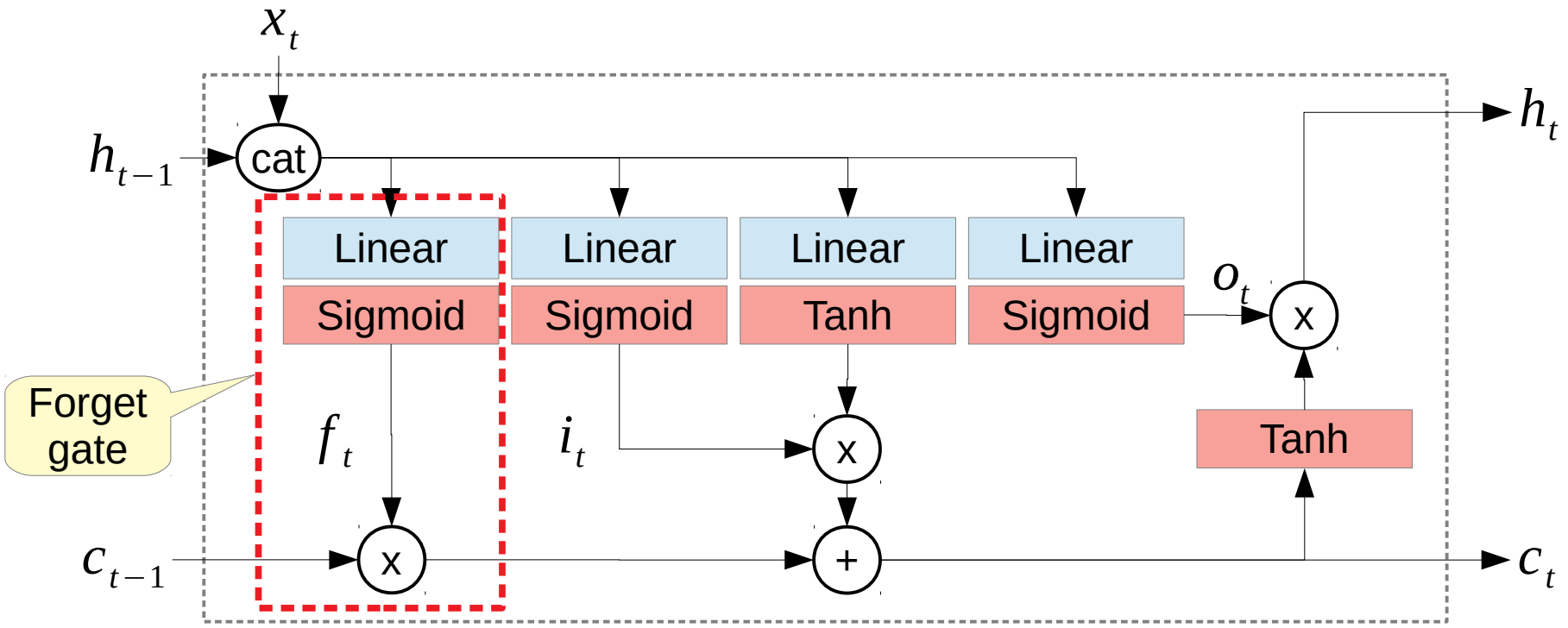
Long Short-Term Memory (LSTM)



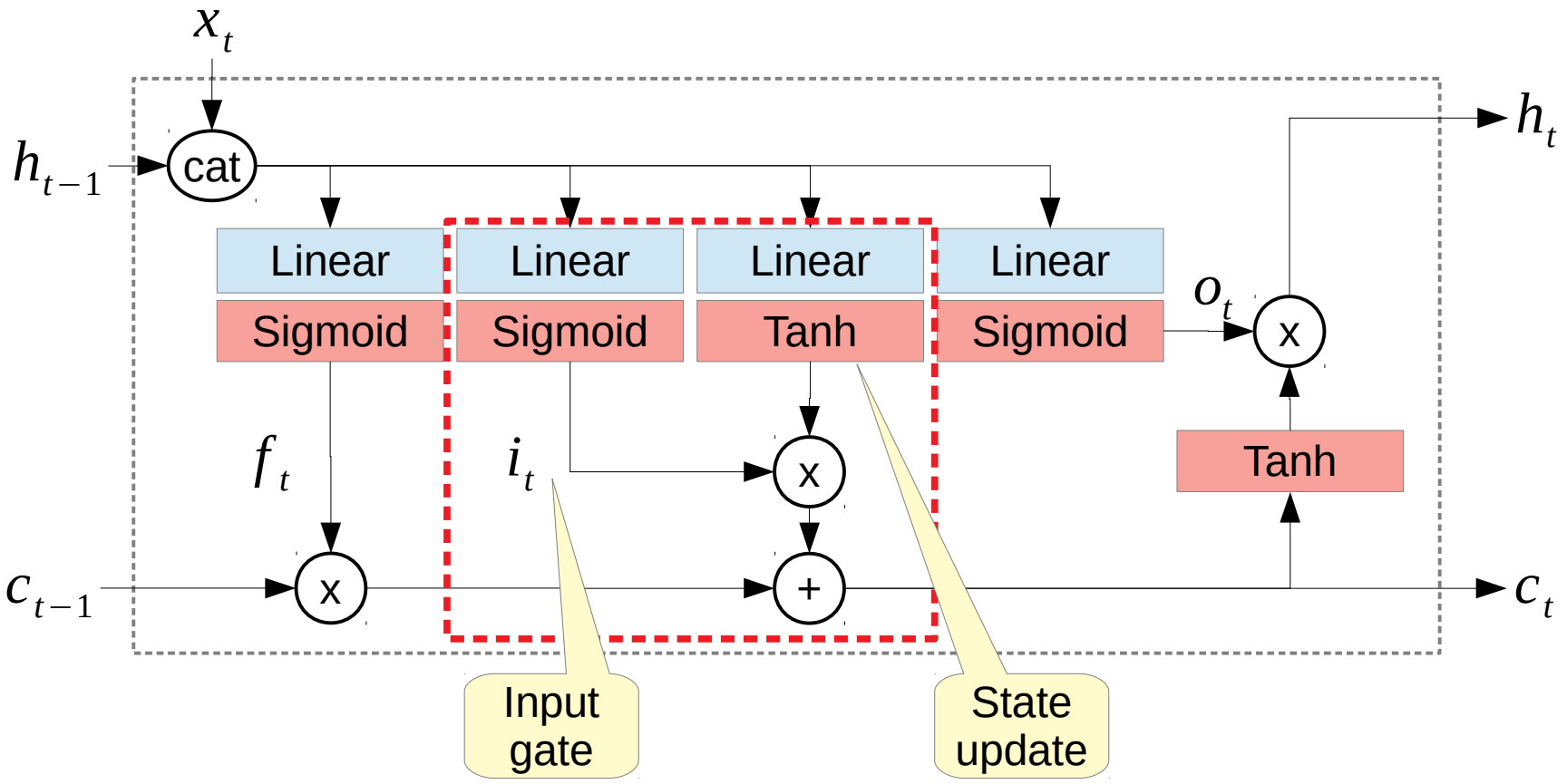
Long-term memory

Similar to residual connections

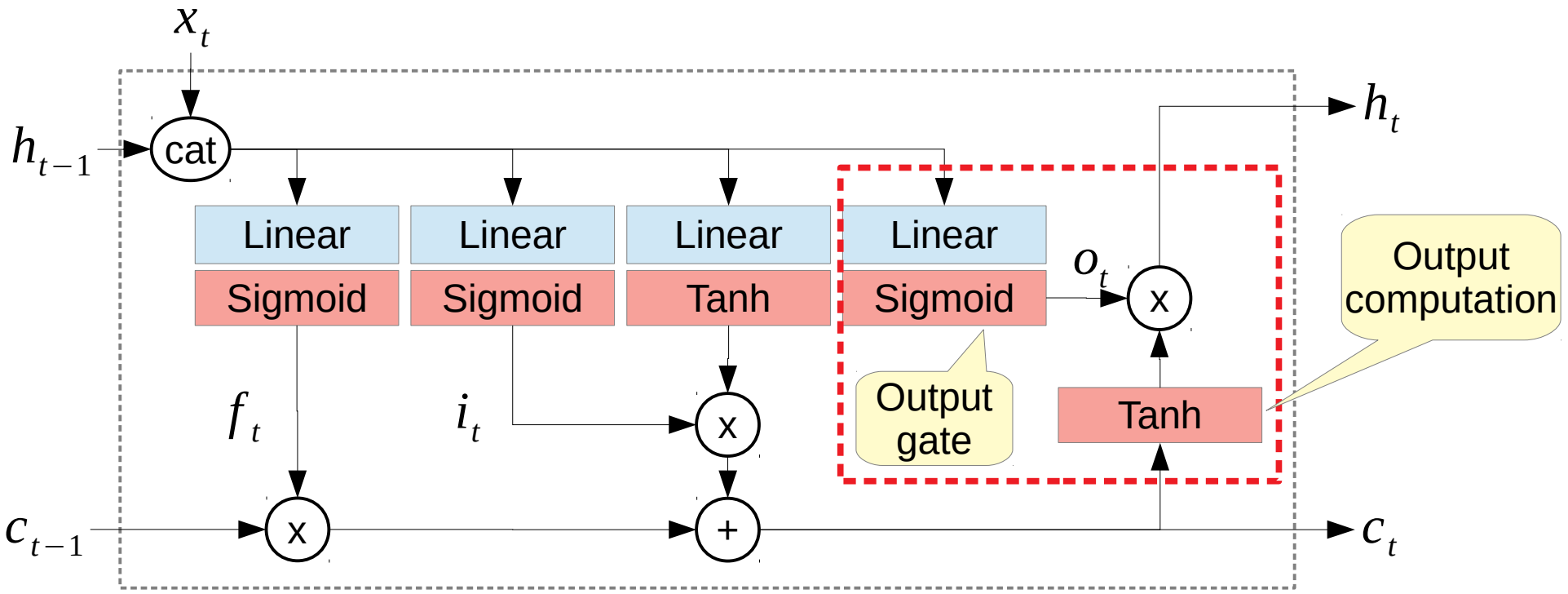
Long Short-Term Memory (LSTM)



Long Short-Term Memory (LSTM)

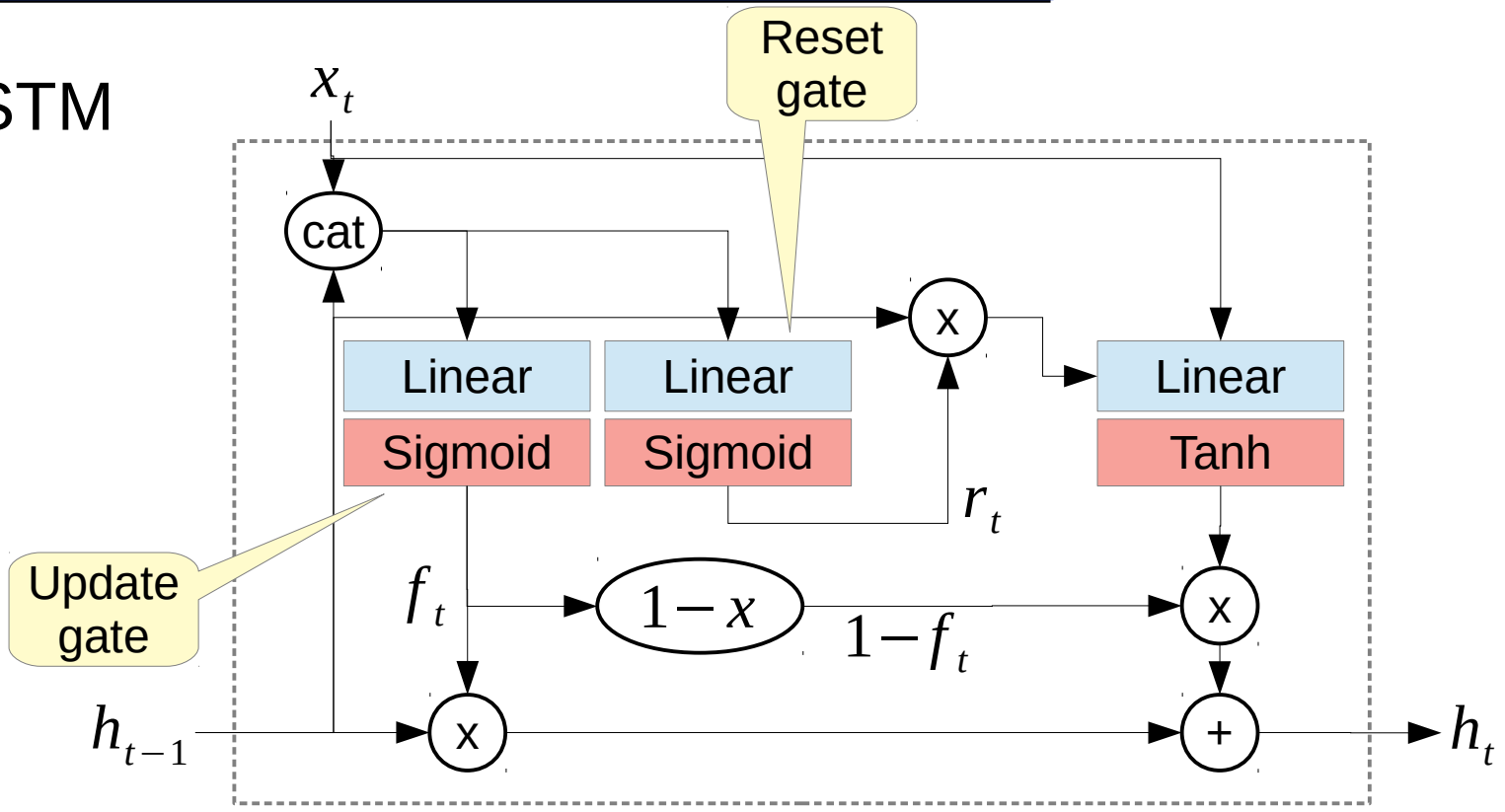


Long Short-Term Memory (LSTM)



Gated Recurrent Unit (GRU)

- Simplified LSTM
- Single state
- Fewer gates
- Similar performance



LSTM/GRU Networks