Georgia Tech College of Computing

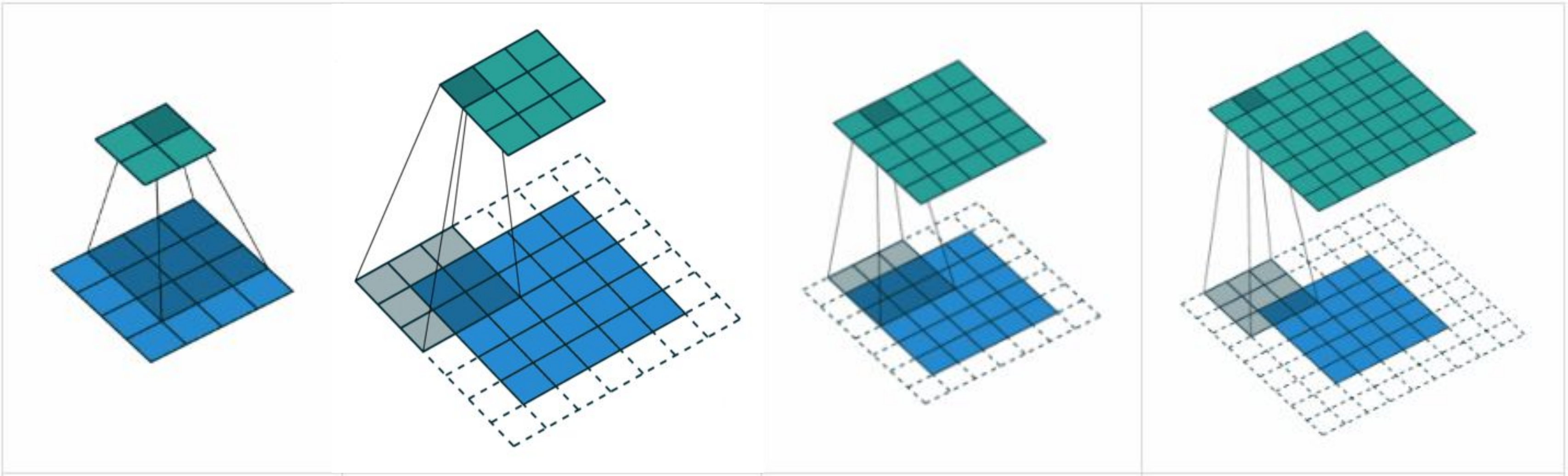
# CS4641 Spring 2025

# Recurrent Neural Networks

Bo Dai
School of CSE, Georgia Tech
bodai@cc.gatech.edu

Based on Fei-Fei Li & Ehsan Adeli's Slides

# Convolution Layer

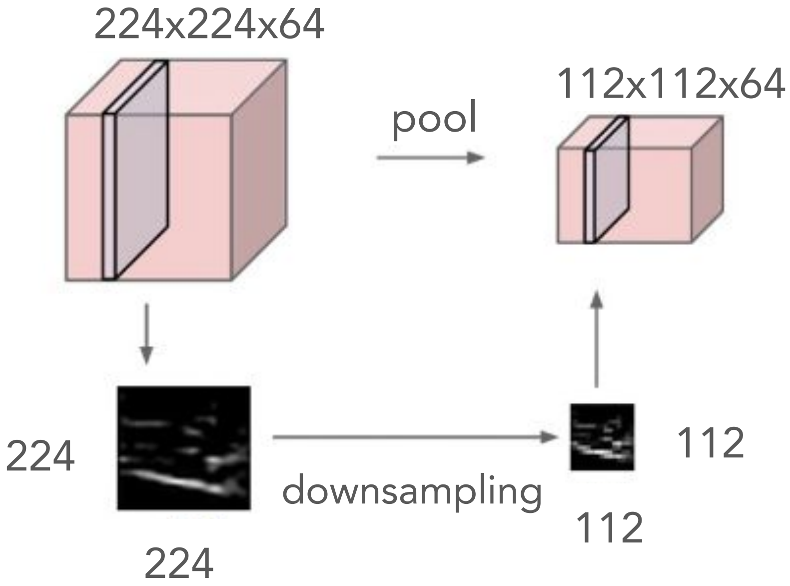


padding = 0, stride = 1 | padding = 1, stride = 2 | padding = 1, stride = 1 | padding = 2, stride = 1

$$W_{out} = \frac{W - F + 2P}{S} + 1$$

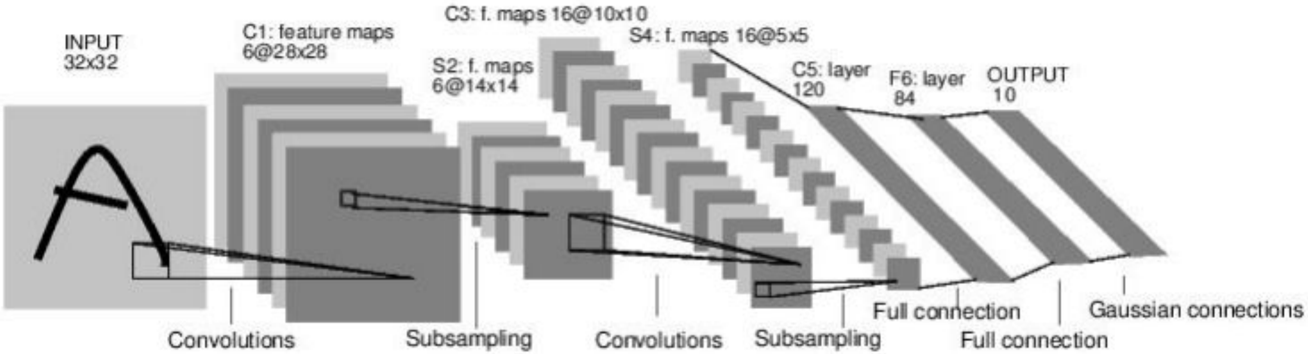
Animation from [Hochschule der Medien]

# Pooling (Subsampling)



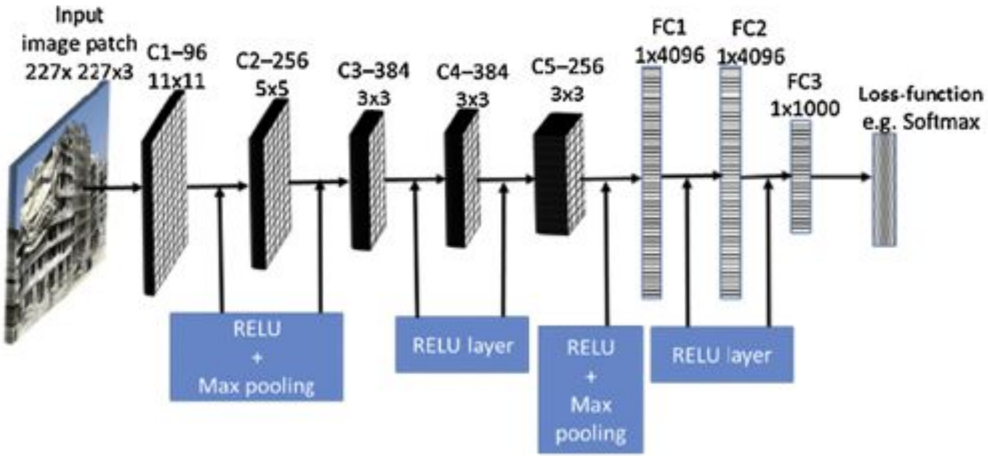

- Pooling layers simplify / subsample / compress the information in the output from the convolutional layer
- Reduce parameters

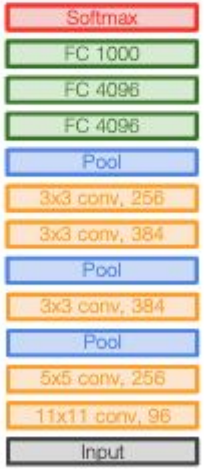
# Put Everything Together




*[LeNet-5, LeCun 1980]*

# AlexNet (2012)

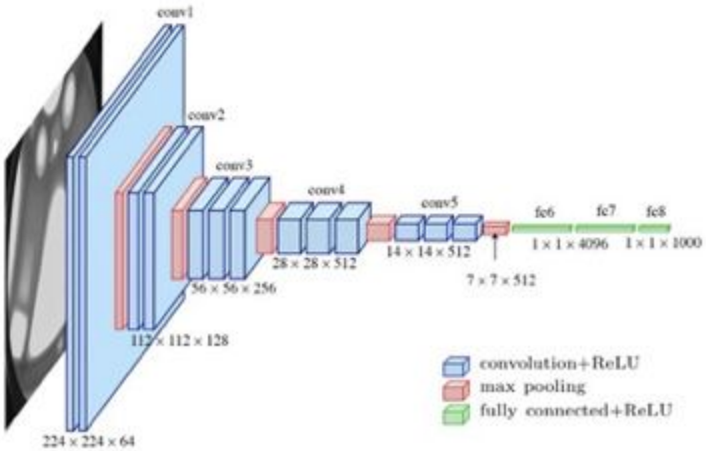



AlexNet

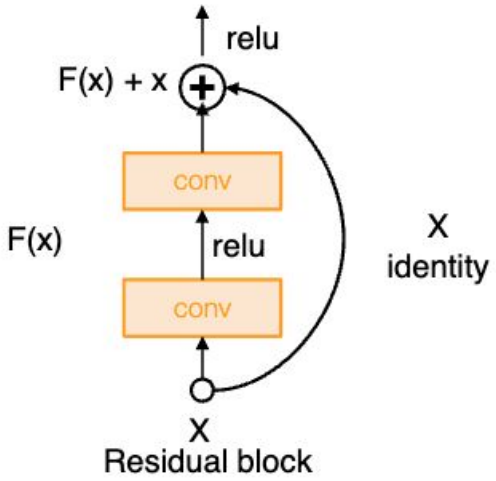
# VGGNet (2014)

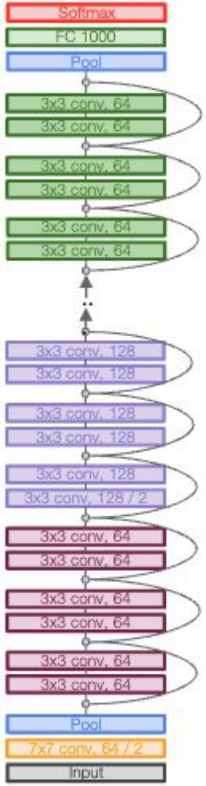
- Very Deep CNN
- With only 3*3 conv filters
    - Fewer parameters, deeper nonlinear layers




# ResNet (2015)

- Very Deep CNN with residual connections
  - 152-layer model for ImageNet
  - ILSVRC’15 classification winner (3.57% top 5 error)
  - Swept all classification and detection competitions in ILSVRC’15 and COCO’15!



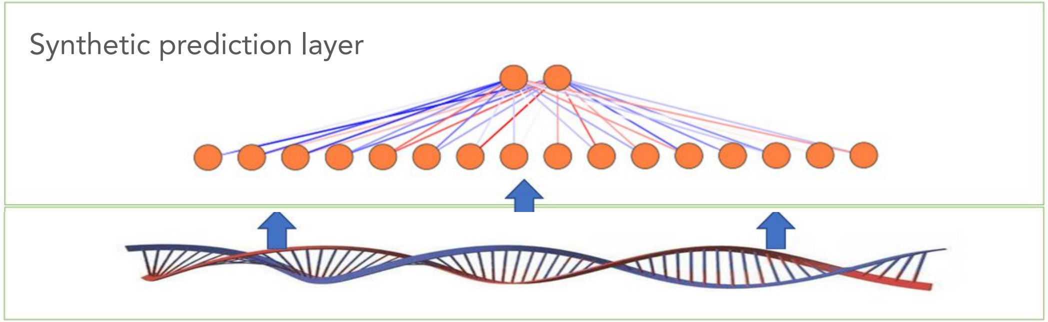



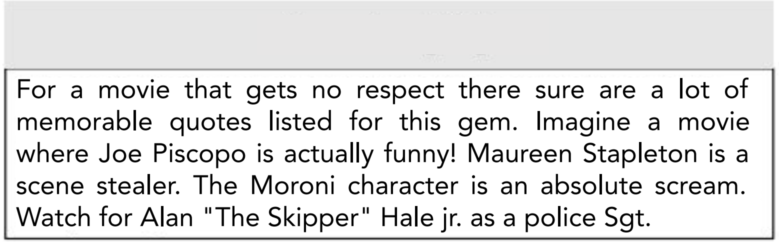
# Sequence Prediction

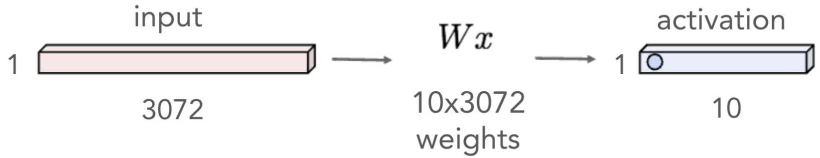
**`texts[0]`**

For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

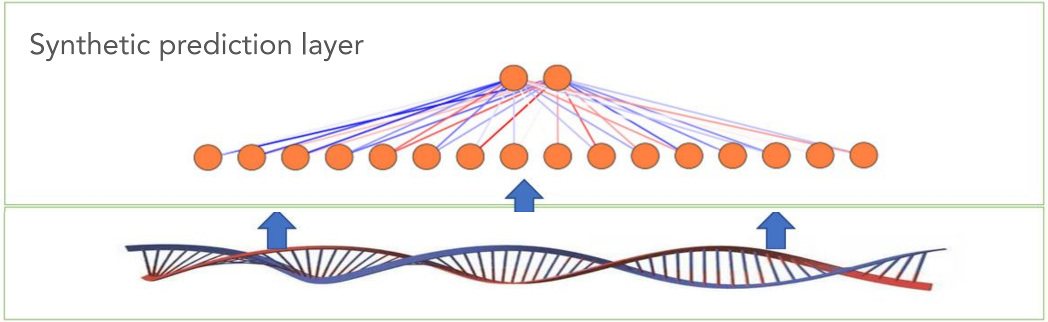
# Sequence Prediction



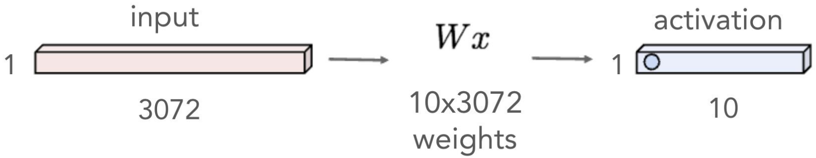



# Sequence Prediction



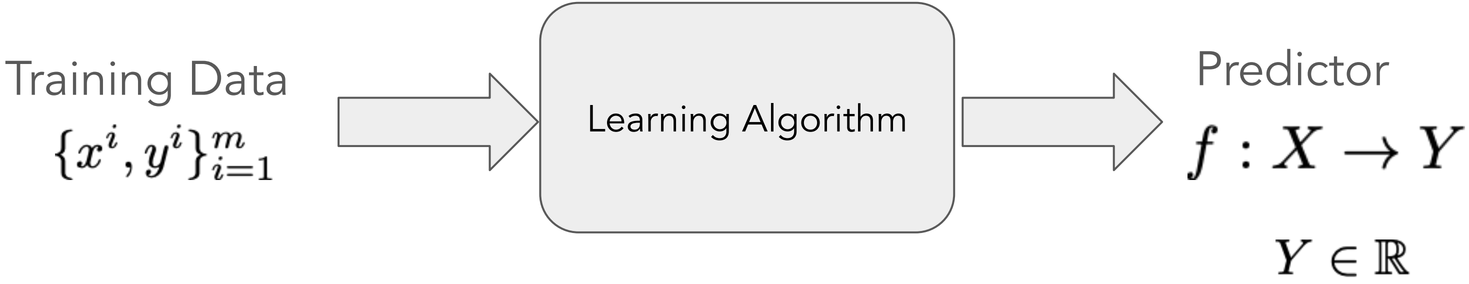

I saw the movie with two grown children. Although it was not as clever as Shrek, I thought it was rather good. In a movie theatre surrounded by children who were on spring break, there was not a sound so I know the children all liked it. There parents also seemed engaged. The death and apparent death of characters brought about the appropriate gasps and comments. Hopefully people realize this movie was made for kids.

For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.




What if the length of sequences varies?

# Sequential Regression Algorithms




## Linear Regression Pipeline

1. Build probabilistic models:
   Gaussian Distribution + RNN
2. Derive loss function: MLE and MAP
3. Select optimizer: (Stochastic) GD

# Sequential Binary Classification Algorithms

Training Data
$\{x^i, y^i\}_{i=1}^m$

Learning Algorithm

Predictor
$f : X \to Y$

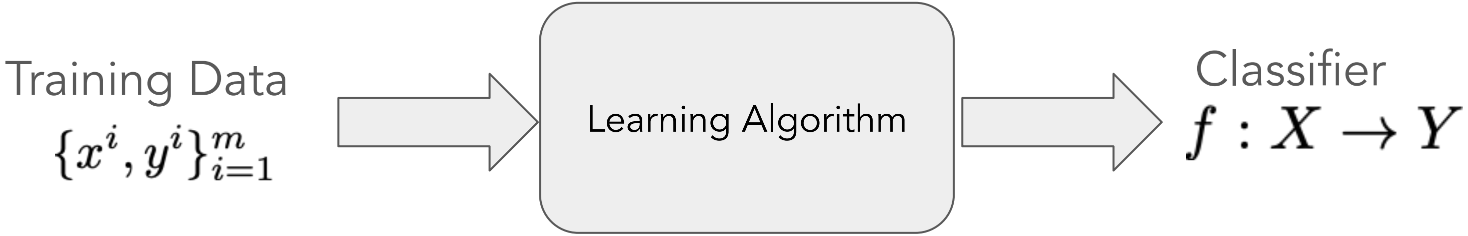
Binary Classification

$Y \in \{0, 1\}$

## Binary Logistic Regression Pipeline

1. Build probabilistic models:
   Bernoulli Distribution + RNN
2. Derive loss function: MLE and MAP
3. Select optimizer: (Stochastic) Gradient Descent

# Sequential Multiclass Logistic Regression Algorithms

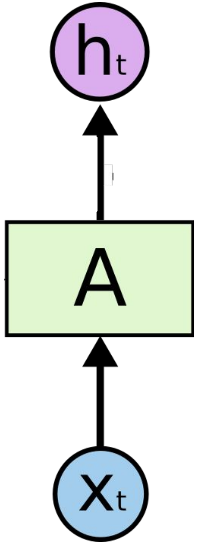



Multiclass Classification $Y \in \{0, 1, \ldots, k\}$

## Multiclass Logistic Regression Pipeline

1. Build probabilistic models:
   Categorical Distribution + RNN
2. Derive loss function: MLE and MAP
3. Select optimizer: (Stochastic) Gradient Descent

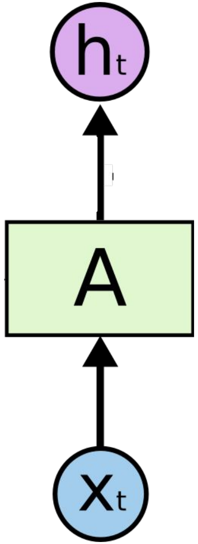
# Recurrent Neural Network



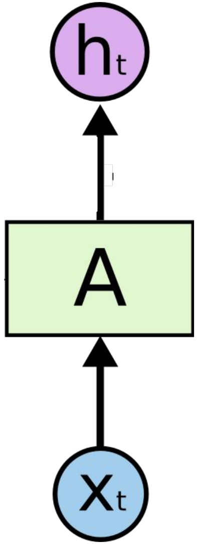

For

For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

# Recurrent Neural Network



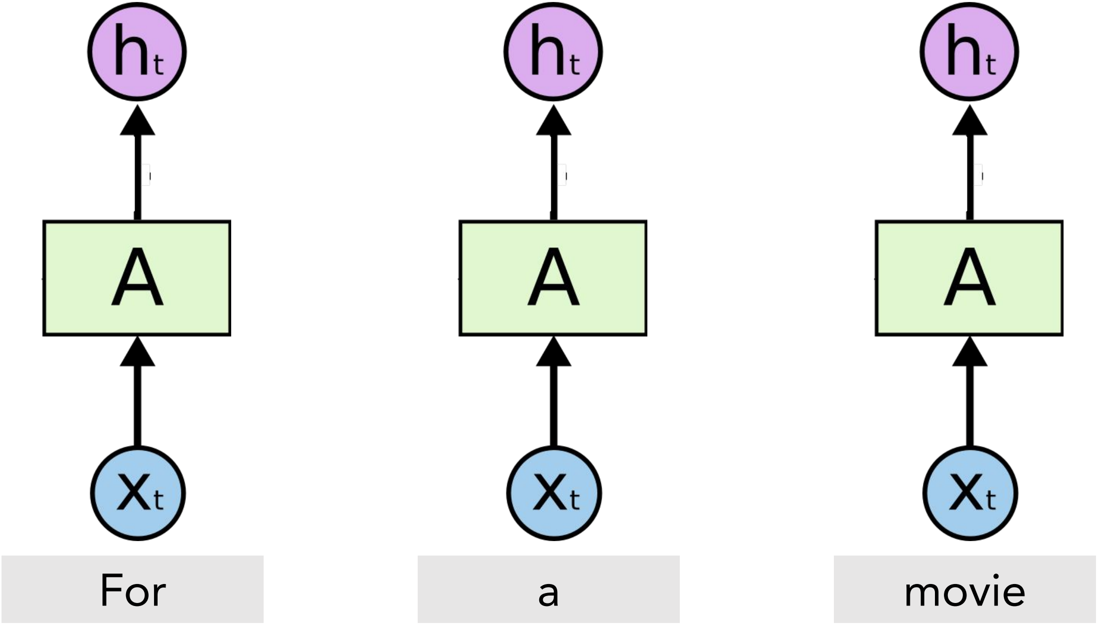

For




a

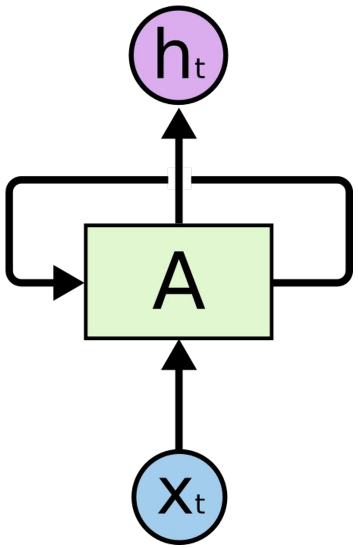
For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

# Recurrent Neural Network



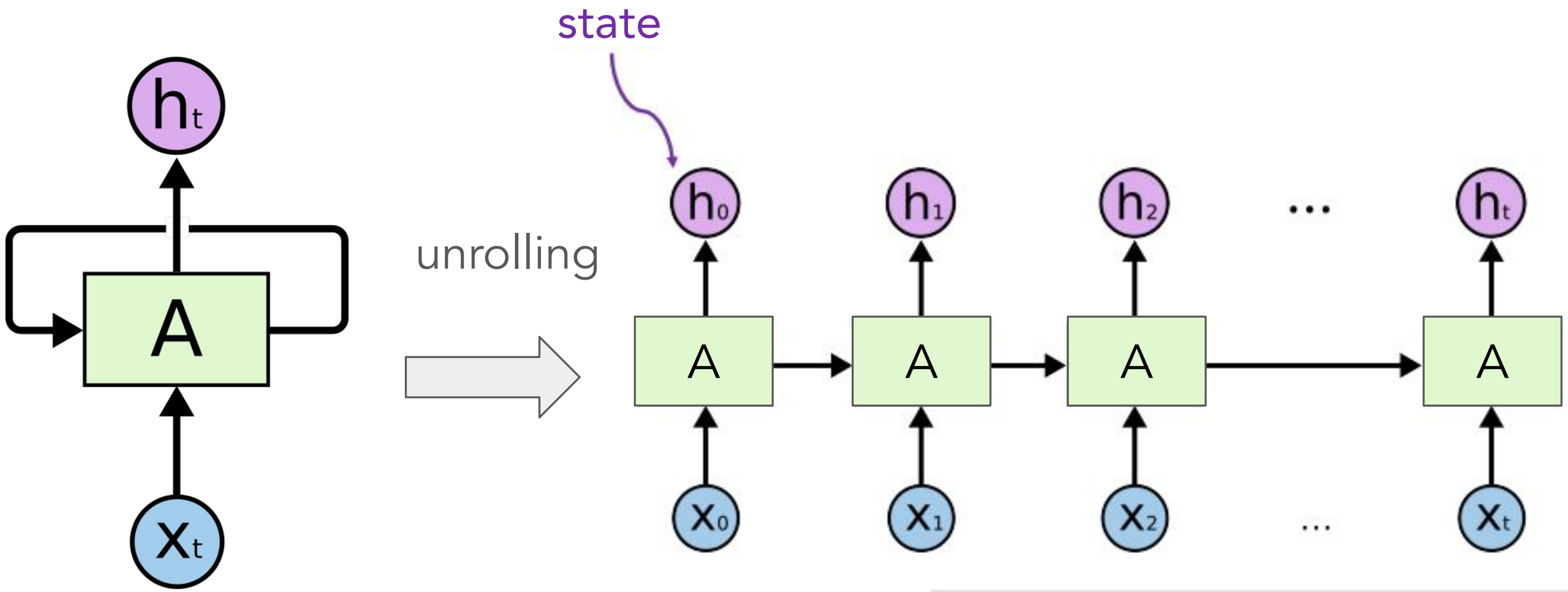

For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

# Recurrent Neural Network



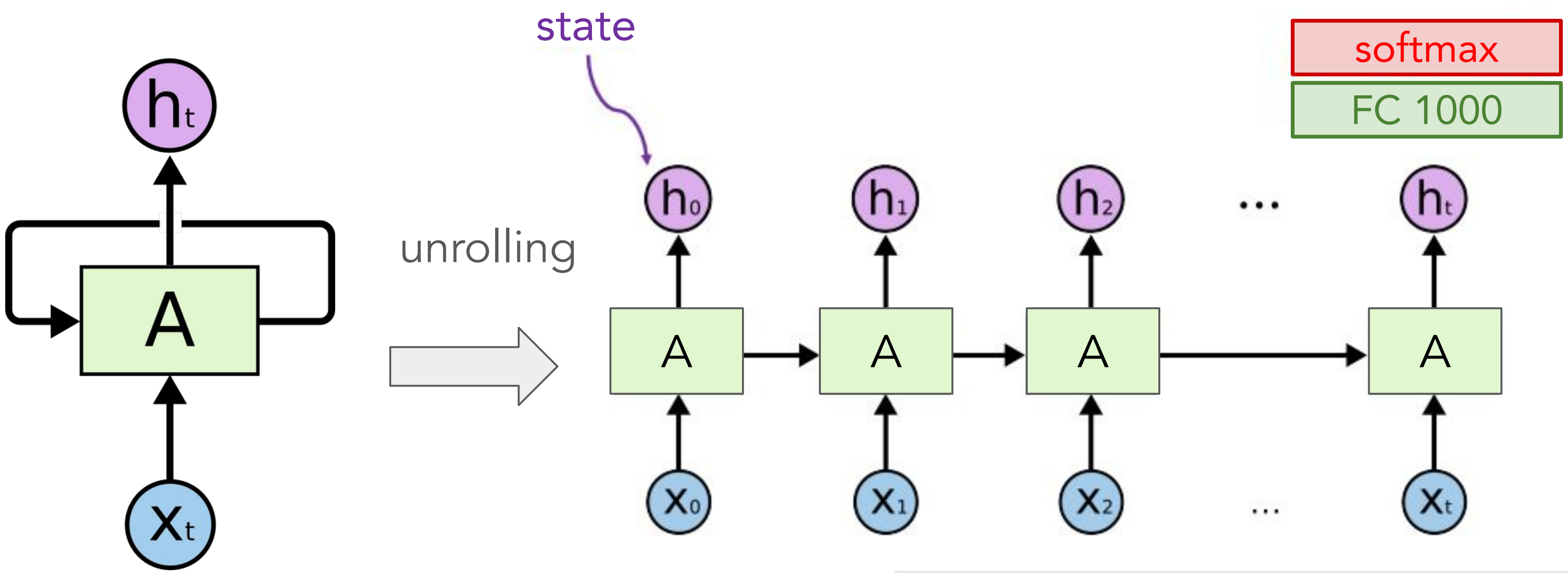

For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

# Recurrent Neural Network



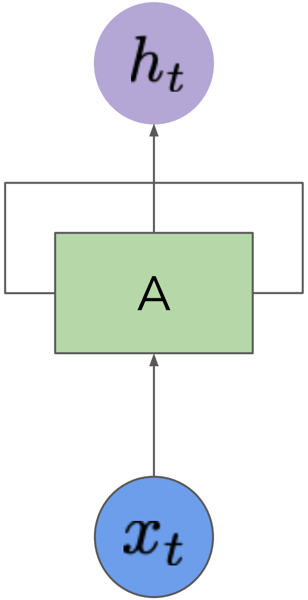

For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

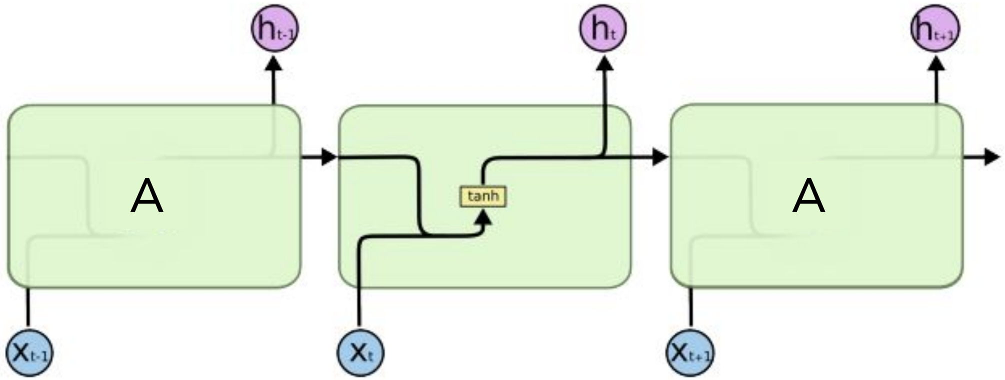
# Recurrent Neural Network

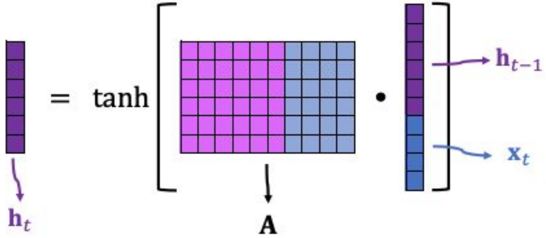



For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

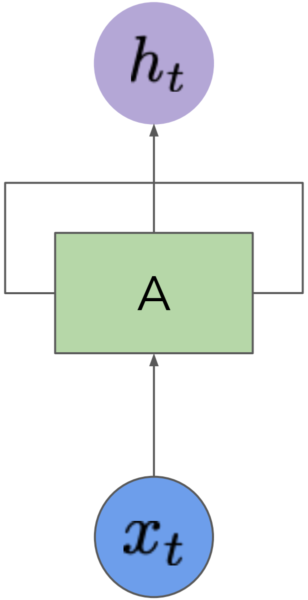
# Simple RNN Cell

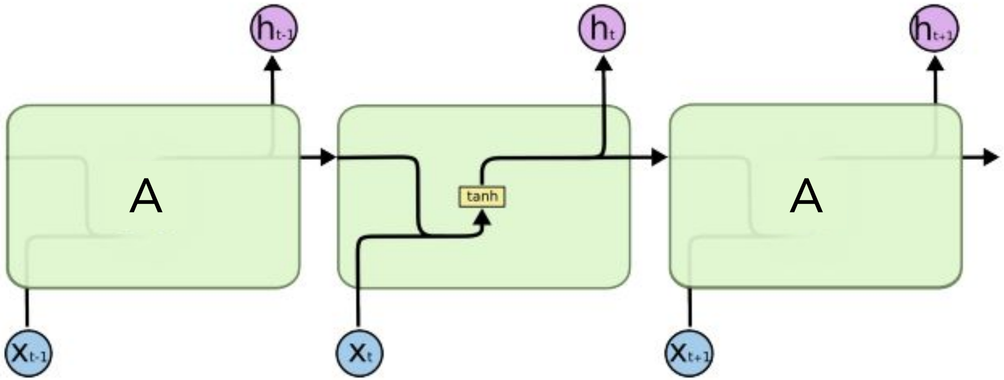


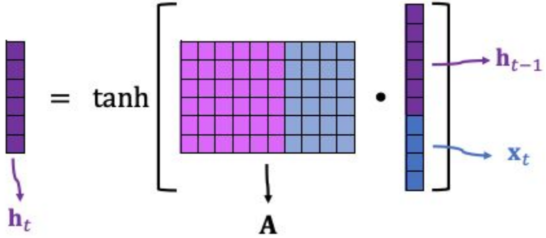



# Simple RNN Cell

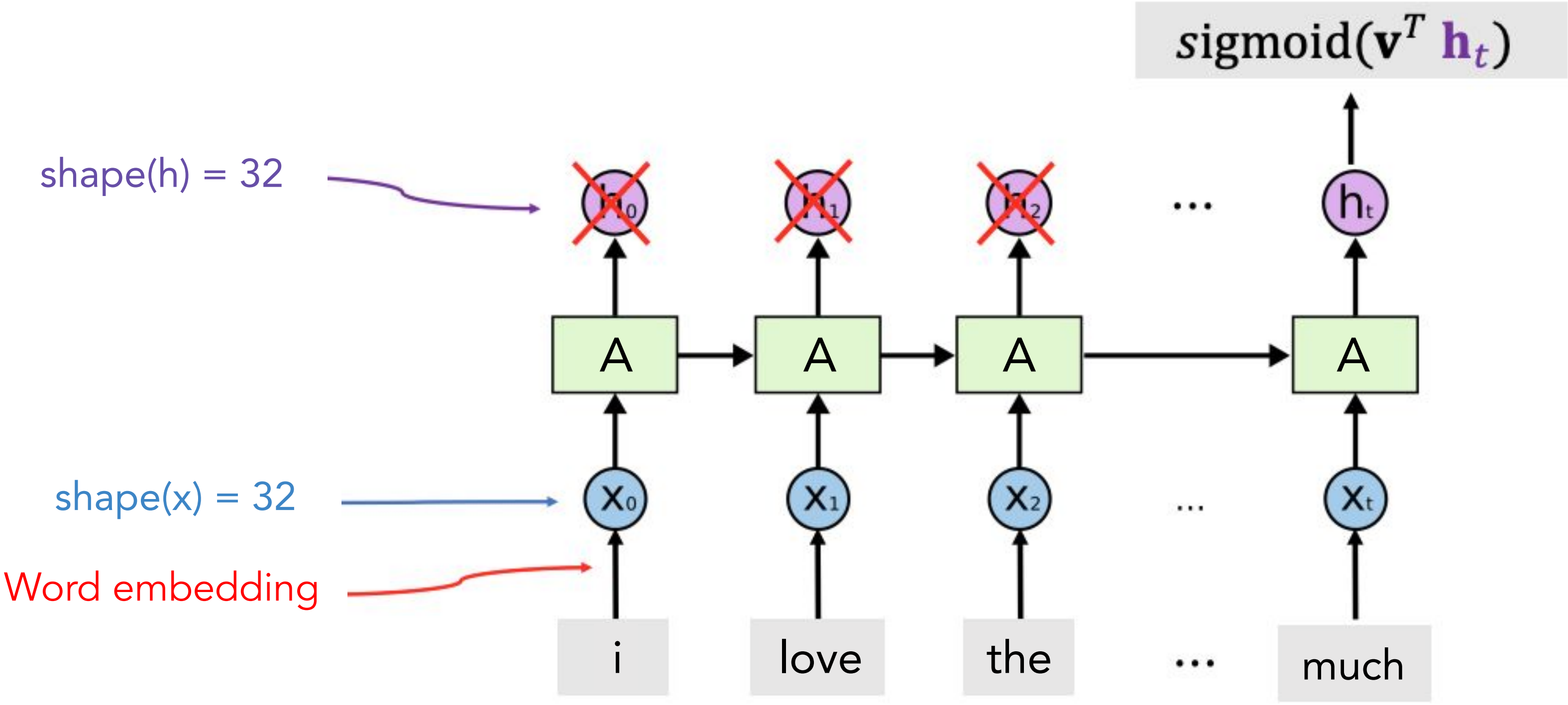


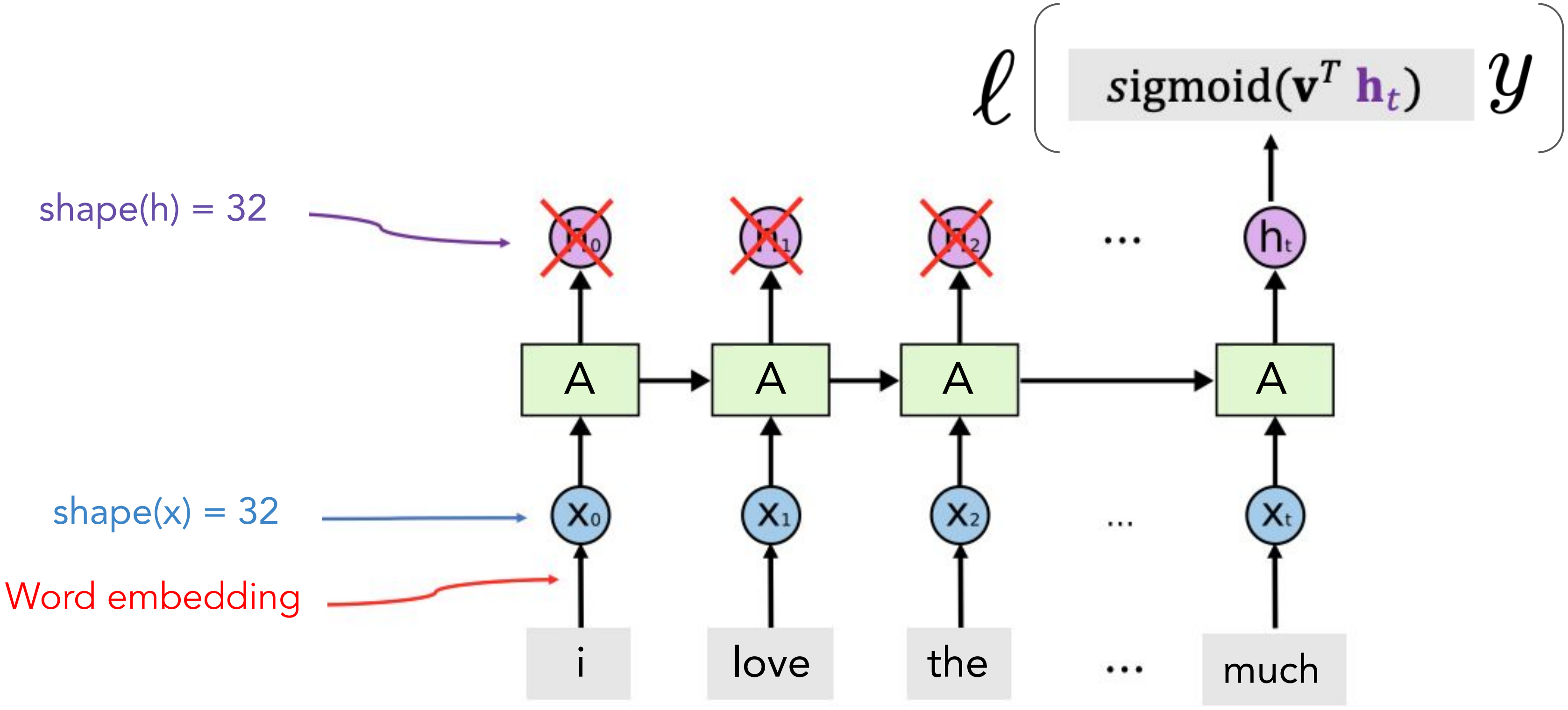

| | | |
|---|---|---|
| Binary step | | $\begin{cases}0 & \text{if } x<0\\ 1 & \text{if } x\ge 0\end{cases}$ |
| Logistic, sigmoid, or soft step | | $\sigma(x) \doteq \frac{1}{1+e^{-x}}$ |
| Hyperbolic tangent (tanh) | | $\tanh(x) \doteq \frac{e^x-e^{-x}}{e^x+e^{-x}}$ |
| Rectified linear unit (ReLU)[13] | | $(x)^+ \doteq \begin{cases}0 & \text{if } x\le 0\\ x & \text{if } x>0\end{cases} = \max(0,x) = x\mathbf{1}_{x>0}$ |
| Gaussian Error Linear Unit (GELU)[5] | | $\frac{1}{2}x\left(1+\text{erf}\left(\frac{x}{\sqrt{2}}\right)\right) = x\Phi(x)$ where erf is the gaussian error function. |
| Softplus[14] | | $\ln(1+e^x)$ |
| Exponential linear unit (ELU)[15] | | $\begin{cases}\alpha\left(e^x-1\right) & \text{if } x\le 0\\ x & \text{if } x>0\end{cases}$ with parameter $\alpha$ |



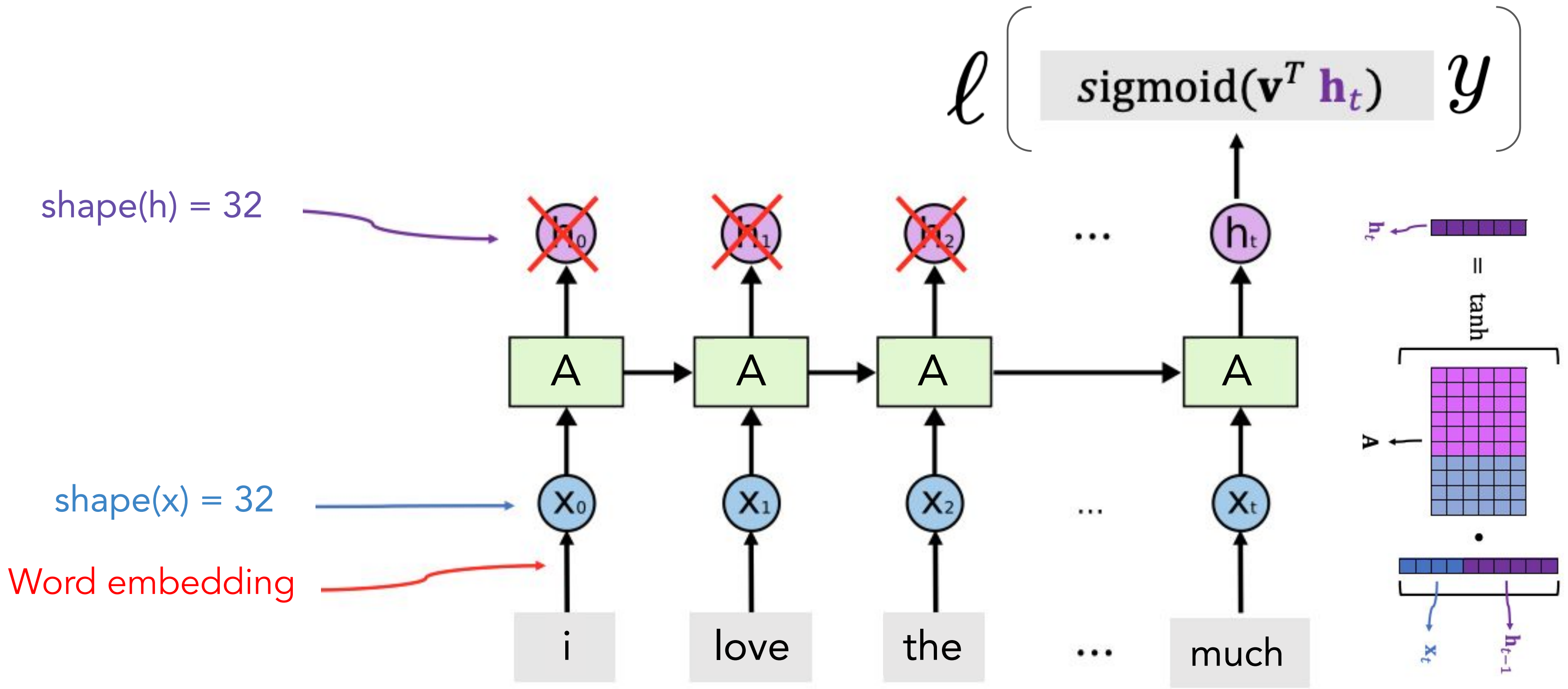
# Simple RNN for IMDB Review



# Backpropagation

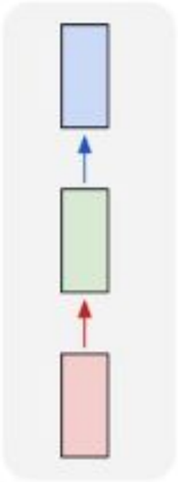


# Backpropagation



# More Usages of RNNs

one to one | one to many



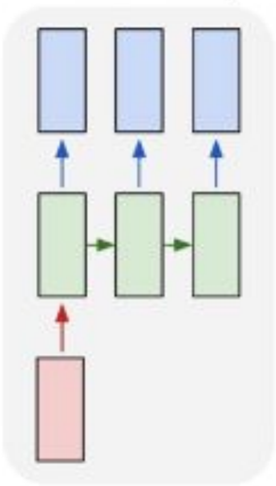
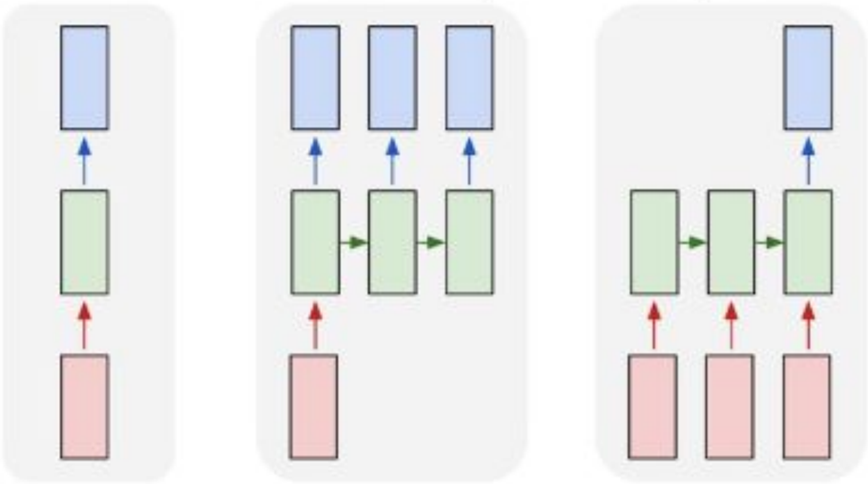


Image Captioning
image -> sequence of words

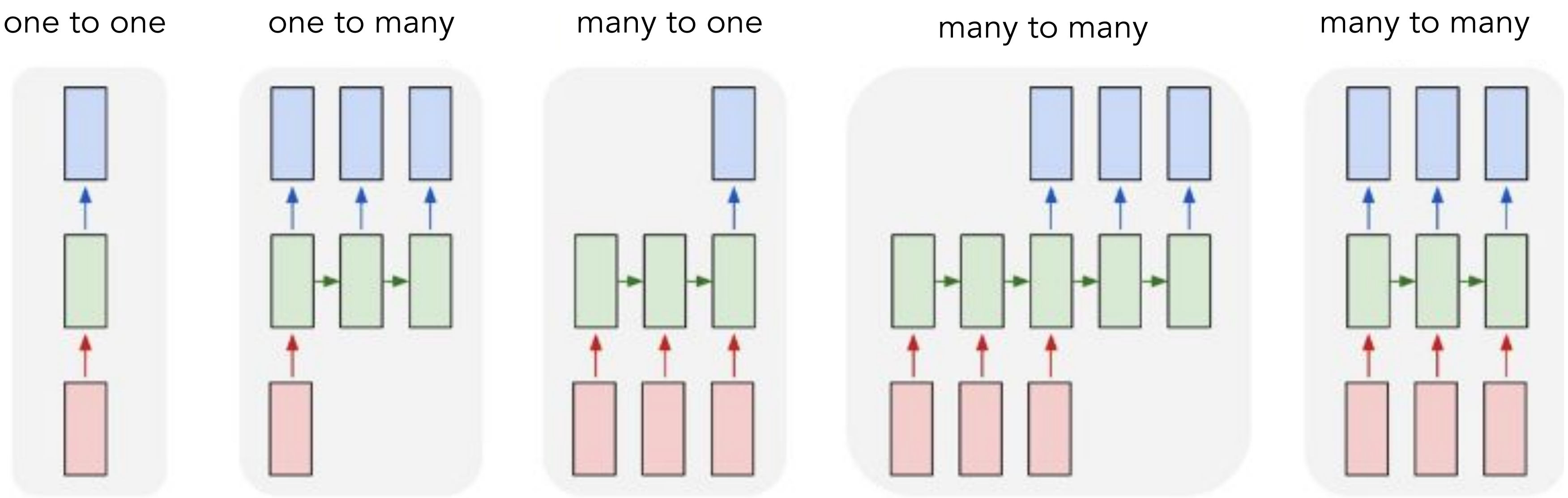
# More Usages of RNNs

one to one        one to many        many to one



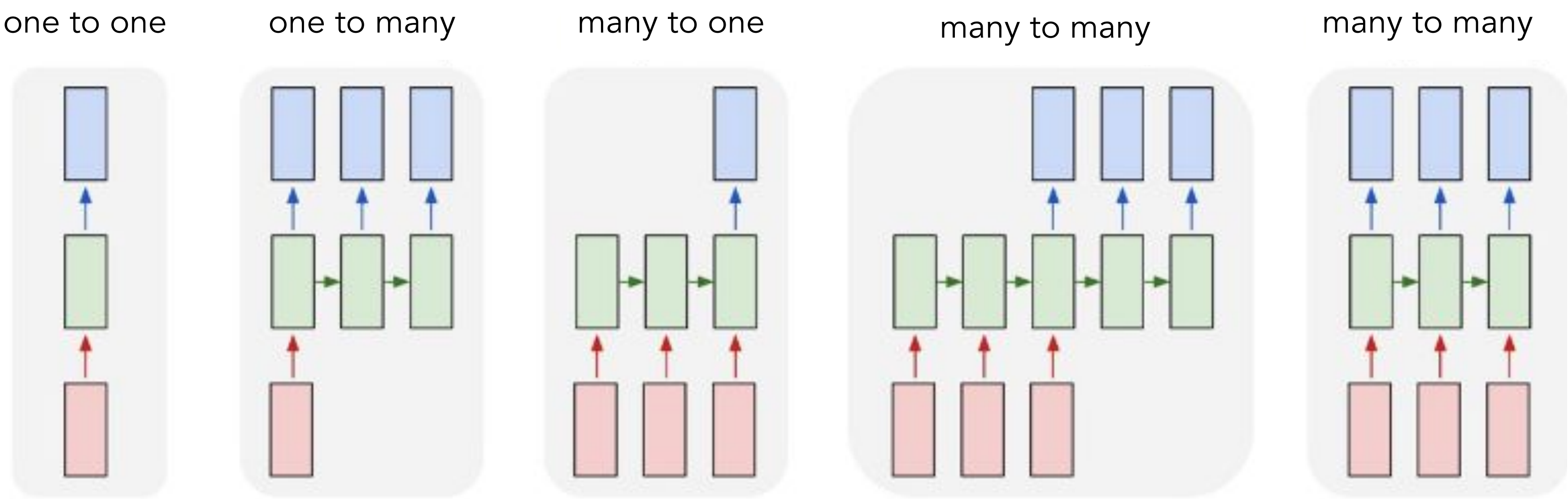
IMDB text review classification

# More Usages of RNNs



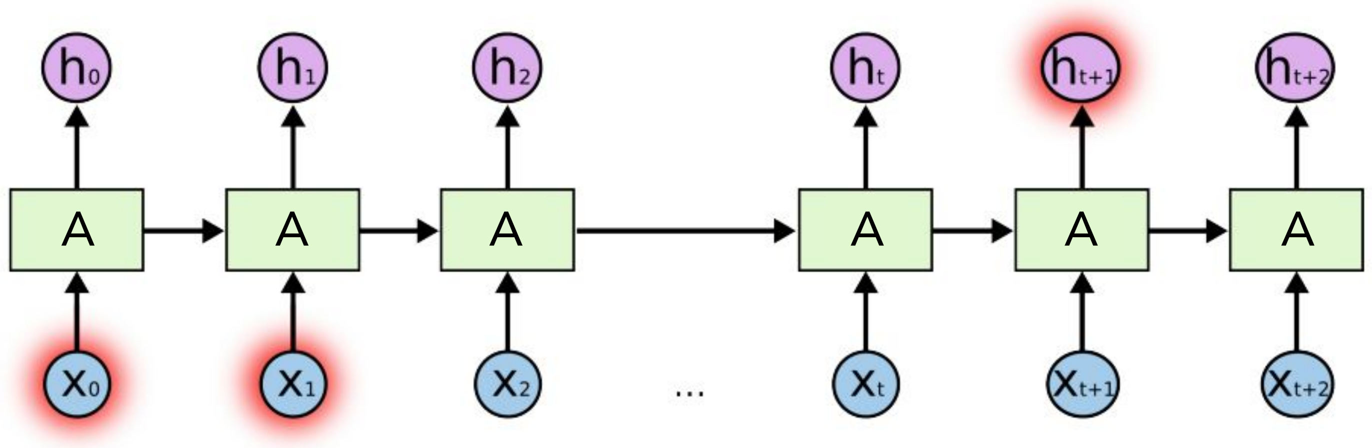

Translation

# Training of RNNs




Backpropagation Through Unrolling Steps

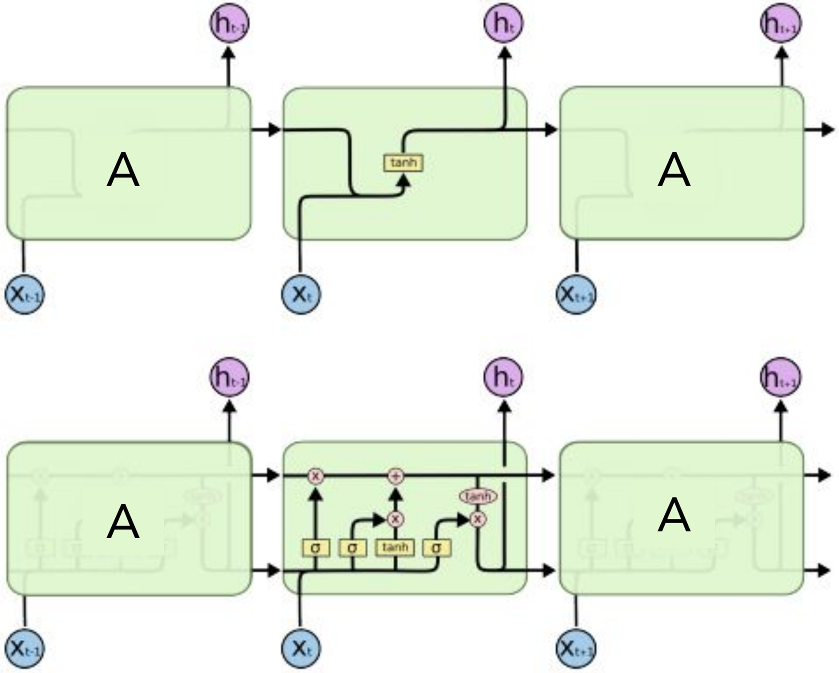
# Simple RNN is not good at long-term dependence




$\mathbf{h}_{100}$ is almost irrelevant to $\mathbf{x}_1$: $\frac{\partial \, \mathbf{h}_{100}}{\partial \, \mathbf{x}_1}$ is near zero.

Gradient Vanishing

# Long Short Term Memory (LSTM)



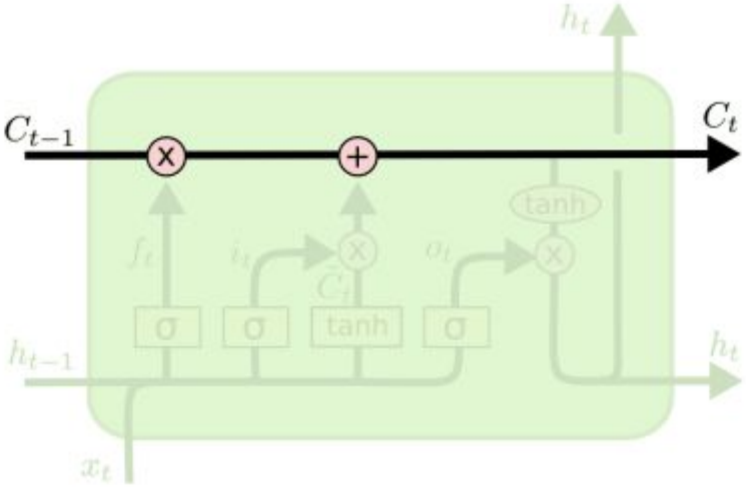

Simple RNN

LSTM

Figures from Christopher Olah's blog

# LSTM Cell: Conveyor Belt

The past information directly flows to the future.



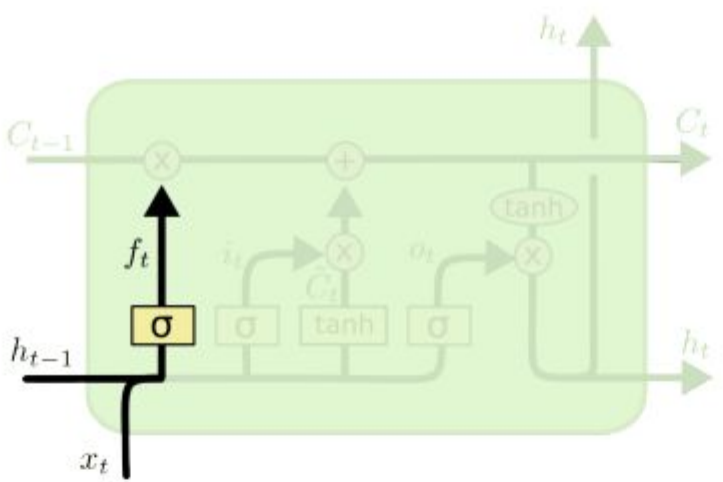

$$C_t = C_{t-1} \circ f_t + \hat{C}_t \circ i_t$$

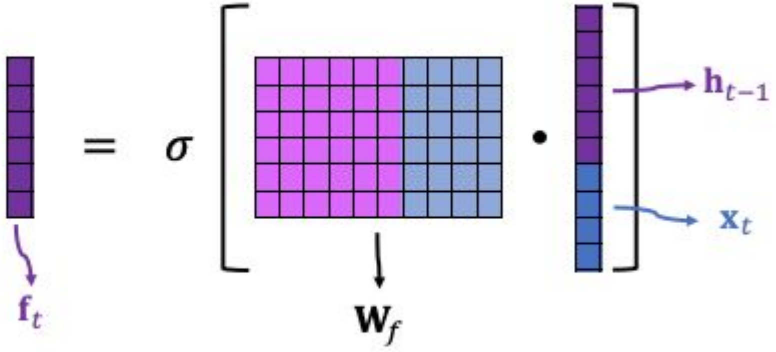
# LSTM Cell: Forget Gate

A value of *zero* means "let *nothing* through."

A value of *one* means "let *everything* through!"

$$f_t = \sigma(W_f x_t + U_f h_{t-1})$$

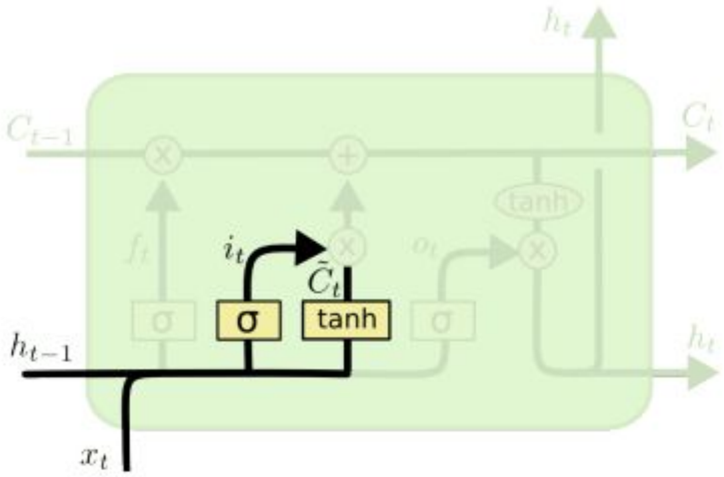


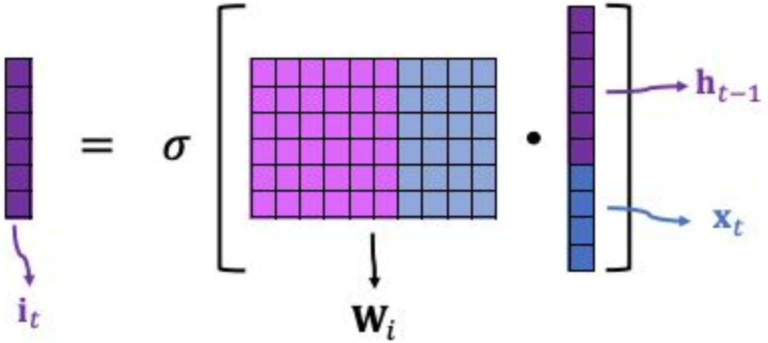

$$C_t = C_{t-1} \circ f_t + \hat{C}_t \circ i_t$$

# LSTM Cell: Input Gate

How much information current context provided.




$$i_t = \sigma(W_i x_t + U_i h_{t-1})$$

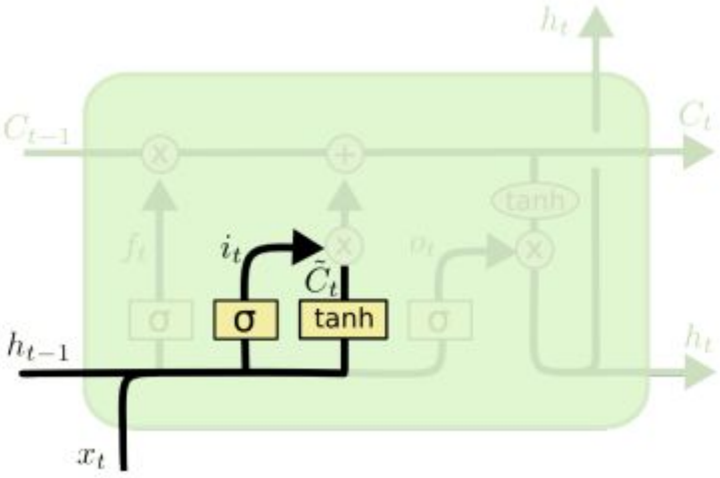


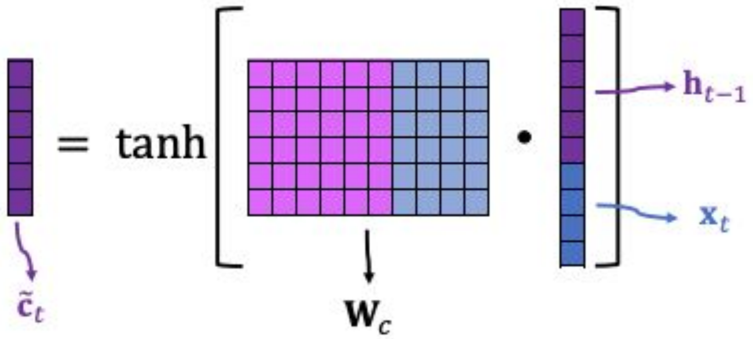

$$C_t = C_{t-1} \circ f_t + \hat{C}_t \circ i_t$$

# LSTM Cell: New Value

"local" context, only up to immediately preceding state

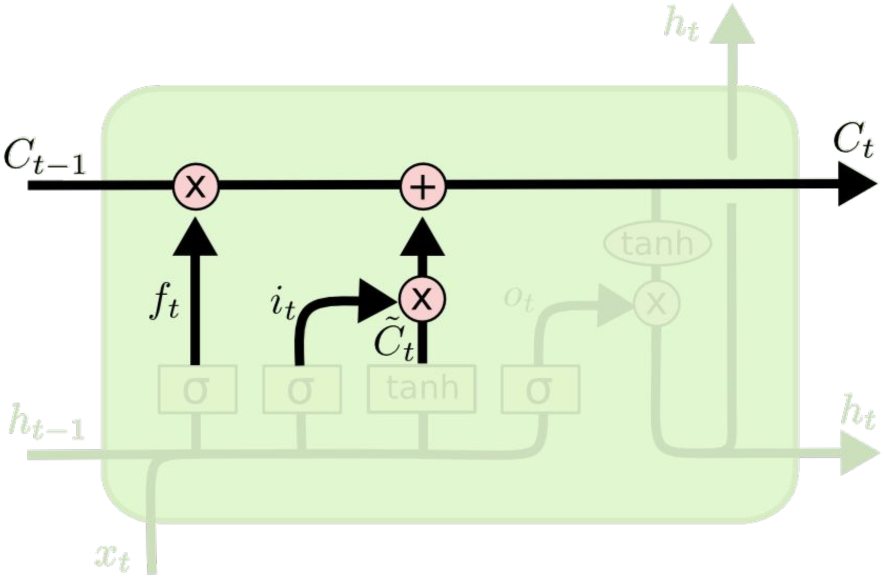
$$\hat{C}_t = \sigma(W_c x_t + U_c h_{t-1})$$




$$C_t = C_{t-1} \circ f_t + \hat{C}_t \circ i_t$$

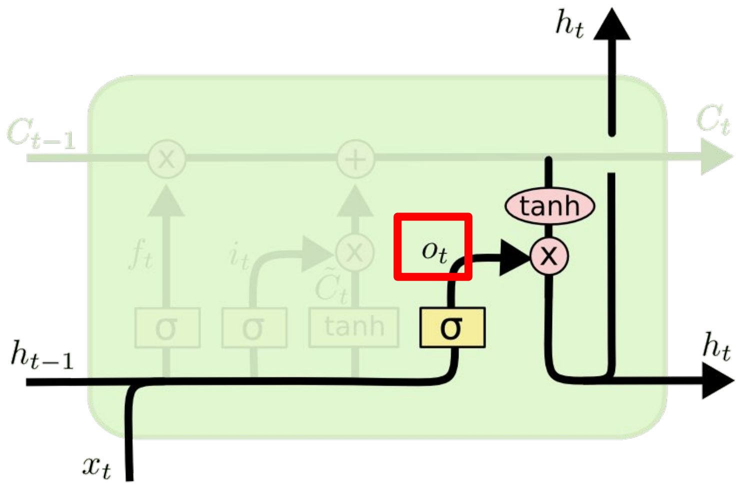
# LSTM Cell: Update Conveyor Belt

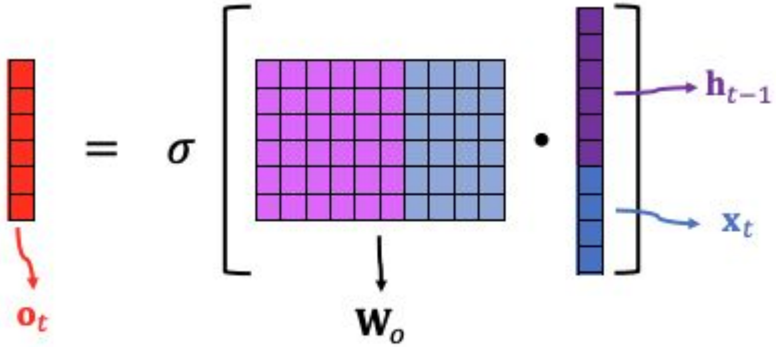



$$C_t = C_{t-1} \circ f_t + \hat{C}_t \circ i_t$$

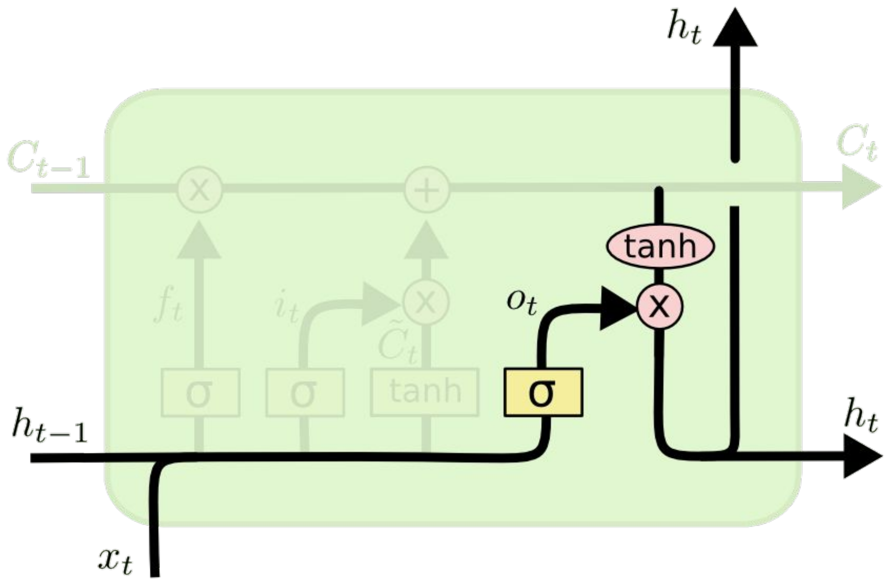
# LSTM Cell: Output Gate

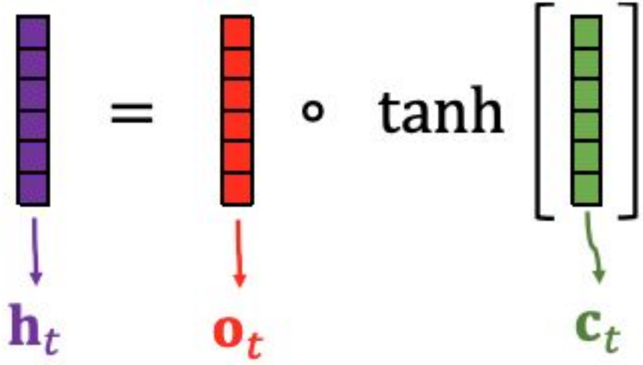
$$h_t = o_t * \tanh\left(C_t\right)$$



# LSTM Cell: Update State




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

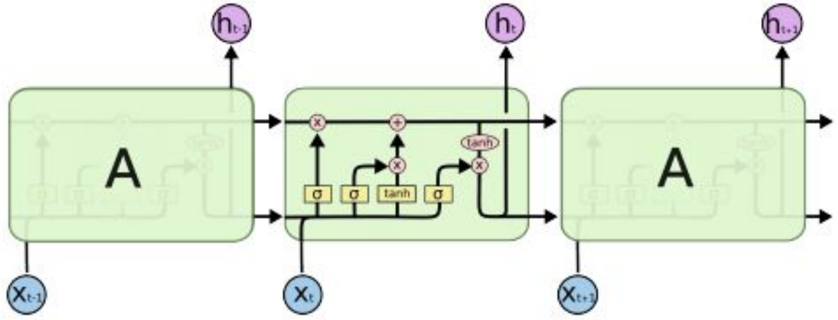
# Auto-differentiation Packages

PyTorch

JAX

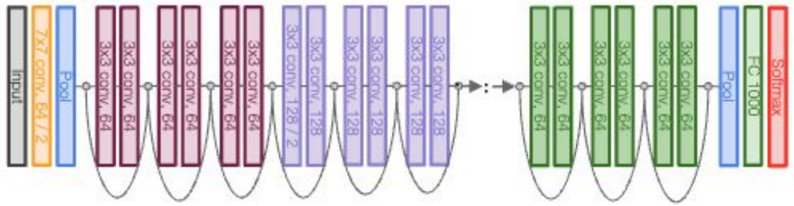
Tensorflow

# LSTM vs. ResNet




$$C_t = C_{t-1} \circ f_t + \hat{C}_t \circ i_t$$

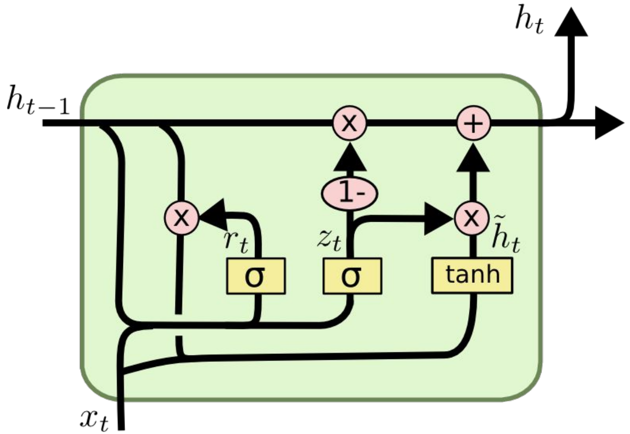
LSTM




Similar to ResNet

# Other Variants of RNN

## Gated Recurrent Unit




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

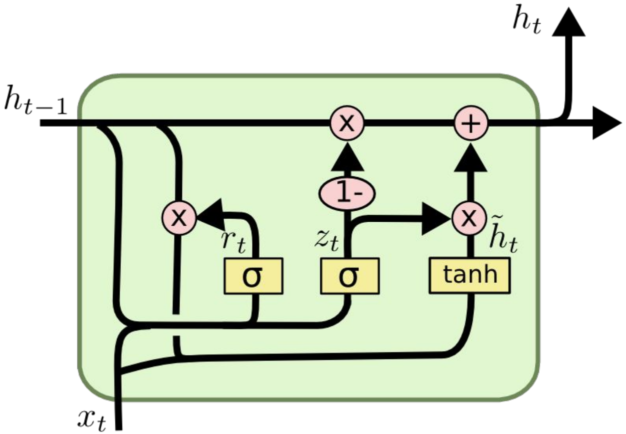
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

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

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

# Other Variants of RNN

## Gated Recurrent Unit




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

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

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

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

# Summary

- RNNs allow a lot of flexibility in architecture design

# Summary

- RNNs allow a lot of flexibility in architecture design
- Vanilla RNNs are simple but don't work very well

# Summary

- RNNs allow a lot of flexibility in architecture design
- Vanilla RNNs are simple but don't work very well
- Backward flow of gradients in RNN can explode or vanish. Exploding is controlled with gradient clipping. Vanishing is controlled with additive interactions (LSTM)

# Q&A