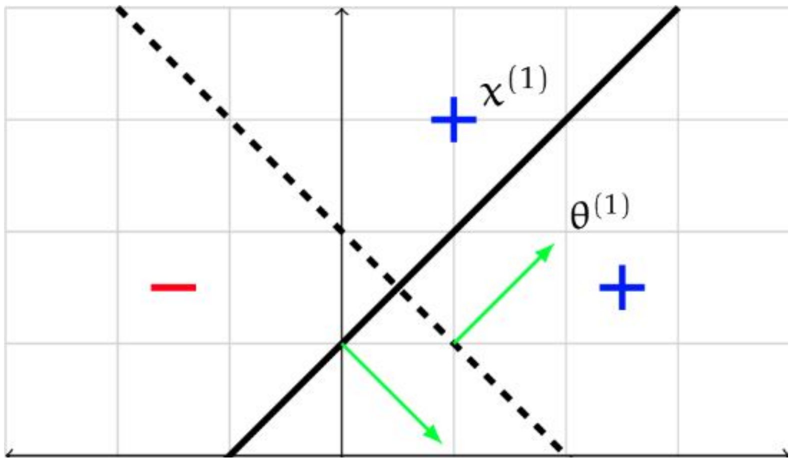
Introduction to Machine Learning



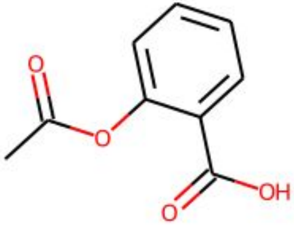
Recurrent Neural Networks

Neural Network Architectures

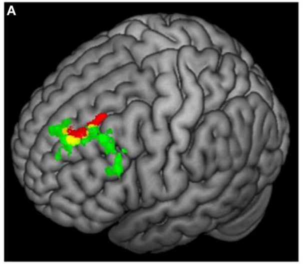
Inject structural knowledge about the input domain into our neural network.

Image

Audio



Molecules



fMRI

If the neural network is “aware” of input domain structure, then it can learn faster and generalize better.

Neural Network Architectures

- Convolution Neural Networks: for processing data where there is **locality** and **translational invariance**

- Recurrent Neural network architecture tailored for processing sequential data

    - Language
    - Audio
    - Time series data

Sequence Classification

1. Input consists of “(“ and “)”

2. Detect whether “(( ))” occurs in the sequence

Sequence Classification

Input	Output
( )	0
) ((	0
( ) (( ))	1
(( ))) ((	1
((((( )	0

Sequence Classification

1-D convolutions to the rescue!

One-hot representation: (= $\begin{bmatrix}1\\0\end{bmatrix}$,) = $\begin{bmatrix}0\\1\end{bmatrix}$

(	)	(	(	)	)
1	0	1	1	0	0
0	1	0	0	1	1

Sequence Classification

1-D convolutions to the rescue!

One-hot representation: (=

1
0

,) =

0
1

( ) (( ))

1	0	1	1	0	0
0	1	0	0	1	1

1	1	0	0
0	0	1	1

Convolutional filter

Sequence Classification

1-D convolutions to the rescue!

One-hot representation: (= $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$,) = $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$

(	)	(	(	)	)
1	0	1	1	0	0
0	1	0	0	1	1

1	1	0	0
0	0	1	1

$\boxed{1}$

Sequence Classification

1-D convolutions to the rescue!

One-hot representation: (= $\begin{bmatrix}1\\0\end{bmatrix}$,) = $\begin{bmatrix}0\\1\end{bmatrix}$

(	)	(	(	)	)
1	0	1	1	0	0
0	1	0	0	1	1

1	1	0	0
0	0	1	1

1	1

Sequence Classification

1-D convolutions to the rescue!

One-hot representation: (=

1	
0	

,) =

0	
1	

( ) (( ))

1	0	1	1	0	0
0	1	0	0	1	1

1	1	0	0
0	0	1	1

1	1	4

Sequence Classification

1-D convolutions to the rescue!

One-hot representation: (= $\begin{bmatrix}1\\0\end{bmatrix}$,) = $\begin{bmatrix}0\\1\end{bmatrix}$

( ) (( ))

1	0	1	1	0	0
0	1	0	0	1	1

1	1	0	0
0	0	1	1

1	1	4

$\xrightarrow{\text{Max-pooling over time}}$

4

Sequence Classification

1-D convolutions to the rescue!

One-hot representation: (=

1
0

,) =

0
1

( ) (( ))

1	0	1	1	0	0
0	1	0	0	1	1

1	1	0	0
0	0	1	1

1	1	4

4

"1" if output after max pooling >= 4

Sequence Classification

Can deal with varying sequence lengths

(	(	)	(	(	(	)	)	)	(
1	1	0	1	1	1	0	0	0	1
0	0	1	0	0	0	1	1	1	0

1	1	0	0
0	0	1	1

Sequence Classification

Can deal with varying sequence lengths

(	(	)	(	(	(	)	)	)	(
1	1	0	1	1	1	0	0	0	1
0	0	1	0	0	0	1	1	1	0

1	1	0	0
0	0	1	1

3	1	1	3	4	3	1

Sequence Classification

Can deal with varying sequence lengths

(	(	)	(	(	(	)	)	)	(
1	1	0	1	1	1	0	0	0	1
0	0	1	0	0	0	1	1	1	0

1	1	0	0
0	0	1	1

3	1	1	3	4	3	1

4

Max-pooling over time

Sequence Classification

Can deal with varying sequence lengths

(	(	)	(	(	(	)	)	)	(
1	1	0	1	1	1	0	0	0	1
0	0	1	0	0	0	1	1	1	0

1D convolutions are great for detecting **local** patterns that are **translation invariant**

1	1	0	0
0	0	1	1

3	1	1	3	4	3	1

4

Max-pooling over time

Harder Sequence Classification

Bounded Parenthese Problem:

1. Input: consists of “(” and “)”
2. Every string has to have an equal number of “(” and “)”
3. Every string has to have a prefix where there are at least as many “(” as “)”

Harder Sequence Classification

Bounded Parenthese Problem:

1. Input: consists of “(” and “)”
2. Every string has to have an equal number of “(” and “)”
3. Every string has to have a prefix where there are at least as many “(” as “)”

Input	Output
( )	1
((( )))	1
( ) (( ))	1
(( ))) ( )	0
( )) (( )	0
(( )) ( )	1

Harder Sequence Classification

- We need to detect **global** vs **local** patterns

(((((((((( )))))))))

- Things are **not** translation invariant

( ) (( ))   (( ))) (

- Deeper convolution layers may work, but doesn’t feel like the right architecture.

State Machines

$$(\mathcal{S}, \mathcal{X}, \mathcal{Y}, s_0, f, g)$$

- $\mathcal{S}$ is a finite or infinite set of possible states;
- $\mathcal{X}$ is a finite or infinite set of possible inputs;
- $\mathcal{Y}$ is a finite or infinite set of possible outputs;
- $s_0 \in \mathcal{S}$ is the initial state of the machine;
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ is a *transition function*, which takes an input and a previous state and produces a next state;
- $g : \mathcal{S} \to \mathcal{Y}$ is an *output function*, which takes a state and produces an output.

State Machines

$$(\mathcal{S}, \mathcal{X}, \mathcal{Y}, s_0, f, g)$$

Initial state s_0

$$s_t = f(s_{t-1}, x_t)$$

Update state with current input

$$y_t = g(s_t)$$

(Optional) Produce an output

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f(W^{sx} x_t + W^{ss} s_{t-1})$$

s_t is a two dimensional vector

$$s_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Count of “(“ minus count of “)”

Minimum of the above metric across time steps

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f(W^{sx} x_t + W^{ss} s_{t-1})$$

s_t is a two dimensional vector

$$s_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Count of “(“ minus count of “)”

Minimum of the above metric across time steps

Claim: if s_t is the zero vector after processing all the inputs, then it is a bounded parenthese string

State Machine for Bounded Parenthese

$$s_t = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Count of “(“ minus count of “)”

Minimum of the above metric across time steps

((( )))

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

( ) ( )

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

State Machine for Bounded Parenthese

$$s_t = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Count of "(" minus count of ")"

Minimum of the above metric across time steps

((( )))

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

( ) ( )

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

( )) (( )

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx}x_t + W^{ss}s_{t-1})$$

$$W^{sx} = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \qquad W^{ss} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$f_1\left(\begin{bmatrix} a \\ b \end{bmatrix}\right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

((( )))

x_t

s_t

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

((( )))

x_t

s_t $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

		(	(	(	)	)	)
x_t		$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$					
s_t	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$						

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

		(	(	(	)	)	)
x_t		$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$					
s_t	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$					

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

((( )))

x_t $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$

s_t $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

		(	(	(	)	)	)
x_t		$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$				
s_t	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 0 \end{bmatrix}$				

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{lccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & & & \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & & & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & & & \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 3 \\ 0 \end{bmatrix} & & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{lccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & & \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 3 \\ 0 \end{bmatrix} & & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

		(	(	(	)	)	)
x_t		$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$		
s_t	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 3 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 0 \end{bmatrix}$		

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 3 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

		(	(	(	)	)	)
x_t		$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
s_t	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 3 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$

		(	)	)	(	(	)
x_t							
s_t	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$						

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 3 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & & & & & \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & & & & & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 3 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & & & & & \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & & & & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 3 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & & & & \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & & & & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 3 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & & & & \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & & & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} \\
s_t & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}2\\0\end{bmatrix} & \begin{bmatrix}3\\0\end{bmatrix} & \begin{bmatrix}2\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\0\end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & & & \\
s_t & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\0\end{bmatrix} & & & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx}x_t + W^{ss}s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 3 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & & & \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} -1 \\ -1 \end{bmatrix} & & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx}x_t + W^{ss}s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} \\
s_t & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}2\\0\end{bmatrix} & \begin{bmatrix}3\\0\end{bmatrix} & \begin{bmatrix}2\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\0\end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & & \\
s_t & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}-1\\-1\end{bmatrix} & & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\Big(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \Big) \qquad f_1\Big(\begin{bmatrix} a \\ b \end{bmatrix} \Big) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} \\
s_t & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}2\\0\end{bmatrix} & \begin{bmatrix}3\\0\end{bmatrix} & \begin{bmatrix}2\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\0\end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & & \\
s_t & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}-1\\-1\end{bmatrix} & \begin{bmatrix}0\\-1\end{bmatrix} & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{ccccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 3 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} -1 \\ -1 \end{bmatrix} & \begin{bmatrix} 0 \\ -1 \end{bmatrix} & &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{ccccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} \\
s_t & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}2\\0\end{bmatrix} & \begin{bmatrix}3\\0\end{bmatrix} & \begin{bmatrix}2\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\0\end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}0\\1\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \\
s_t & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}0\\0\end{bmatrix} & \begin{bmatrix}-1\\-1\end{bmatrix} & \begin{bmatrix}0\\-1\end{bmatrix} & \begin{bmatrix}1\\-1\end{bmatrix} &
\end{array}$$

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

		(	(	(	)	)	)
x_t		$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\end{bmatrix}$	$\begin{bmatrix}0\\1\end{bmatrix}$	$\begin{bmatrix}0\\1\end{bmatrix}$
s_t	$\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}2\\0\end{bmatrix}$	$\begin{bmatrix}3\\0\end{bmatrix}$	$\begin{bmatrix}2\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\end{bmatrix}$

		(	)	)	(	(	)
x_t		$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\end{bmatrix}$	$\begin{bmatrix}0\\1\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\end{bmatrix}$
s_t	$\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{bmatrix}-1\\-1\end{bmatrix}$	$\begin{bmatrix}0\\-1\end{bmatrix}$	$\begin{bmatrix}1\\-1\end{bmatrix}$	

State Machine for Bounded Parenthese

$$s_t = f(s_{t-1}, x_t) = f_1(W^{sx} x_t + W^{ss} s_{t-1})$$

$$s_t = f_1\left(\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} s_{t-1} \right) \qquad f_1\left(\begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ \min(a, b) \end{bmatrix}$$

$$\begin{array}{cccccccc}
 & & (& (& (&) &) &) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 3 \\ 0 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{array}$$

$$\begin{array}{cccccccc}
 & & (&) &) & (& (&) \\
x_t & & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
s_t & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} -1 \\ -1 \end{bmatrix} & \begin{bmatrix} 0 \\ -1 \end{bmatrix} & \begin{bmatrix} 1 \\ -1 \end{bmatrix} & \begin{bmatrix} 0 \\ -1 \end{bmatrix}
\end{array}$$

Recurrent Neural Networks

- State machine with learnable parameters

$$s_t = f_1\left(W^{sx}x_t + W^{ss}s_{t-1} + W_0^{ss}\right)$$

$$y_t = f_2\left(W^{o}s_t + W_0^{o}\right)$$

$x_t : \ell \times 1$

$s_t : m \times 1$

$y_t : v \times 1$

$W^{sx} : m \times \ell$

$W^{ss} : m \times m$

$W_0^{ss} : m \times 1$

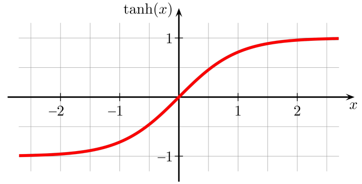
$W^{o} : v \times m$

$W_0^{o} : v \times 1$

Recurrent Neural Networks

- State machine with learnable parameters

$$s_t = f_1 (W^{sx} x_t + W^{ss} s_{t-1} + W_0^{ss})$$
$$y_t = f_2 (W^{o} s_t + W_0^{o})$$



$x_t : \ell \times 1$

$s_t : m \times 1$

$y_t : v \times 1$

$W^{sx} : m \times \ell$

$W^{ss} : m \times m$

$W_0^{ss} : m \times 1$

$W^{o} : v \times m$

$W_0^{o} : v \times 1$

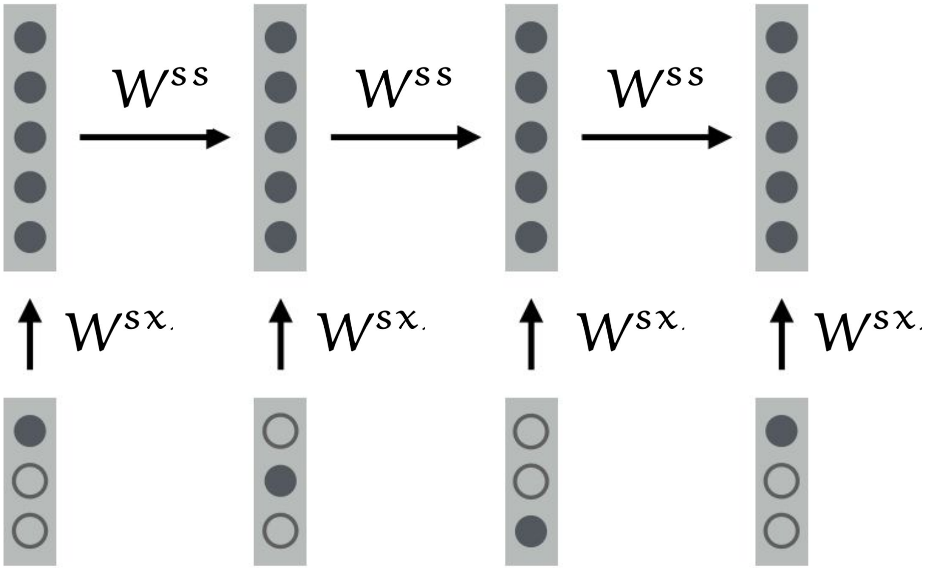
f1: non-linear function (e.g., tanh)

f2: depends on output (e.g., softmax if predicting something at each time step)

Recurrent Neural Networks

$$s_t = f_1\left(W^{sx}x_t + W^{ss}s_{t-1} + W_0^{ss}\right)$$

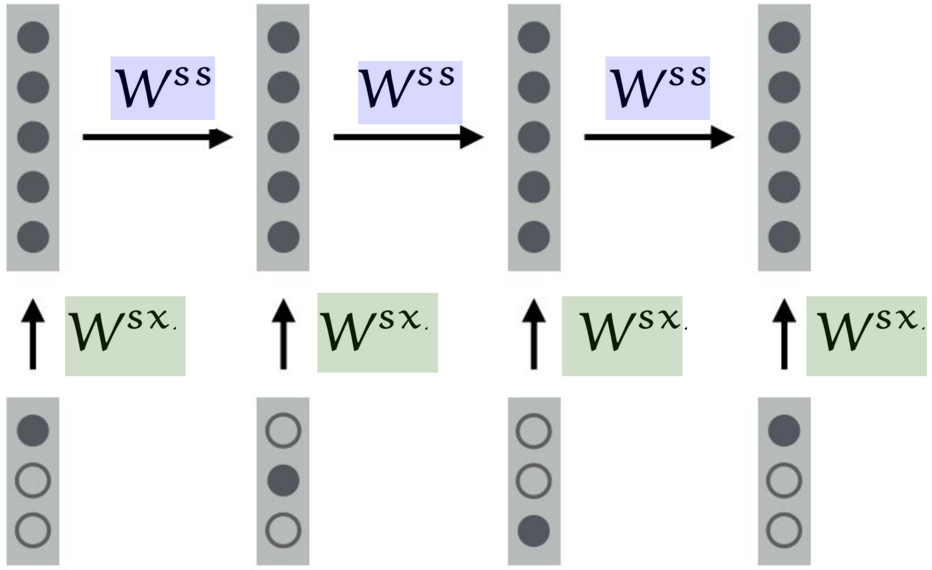
- Hidden state is a function of previous hidden state and current input.



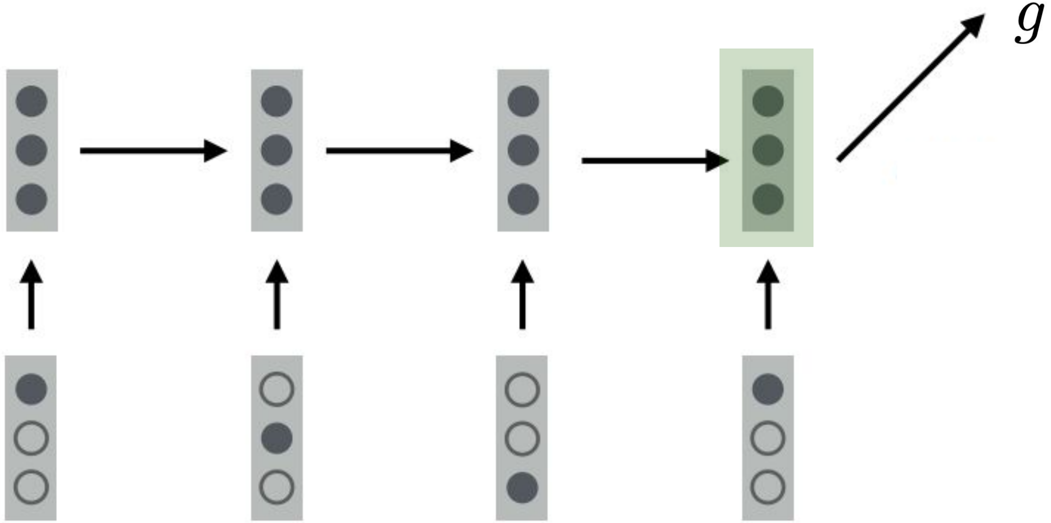
Recurrent Neural Networks

$$s_t = f_1\left(W^{sx}x_t + W^{ss}s_{t-1} + W_0^{ss}\right)$$

- Hidden state is a function of previous hidden state and current input.
- Same weights at each state ⇒ parameter sharing!



RNNs for Sequence Classification

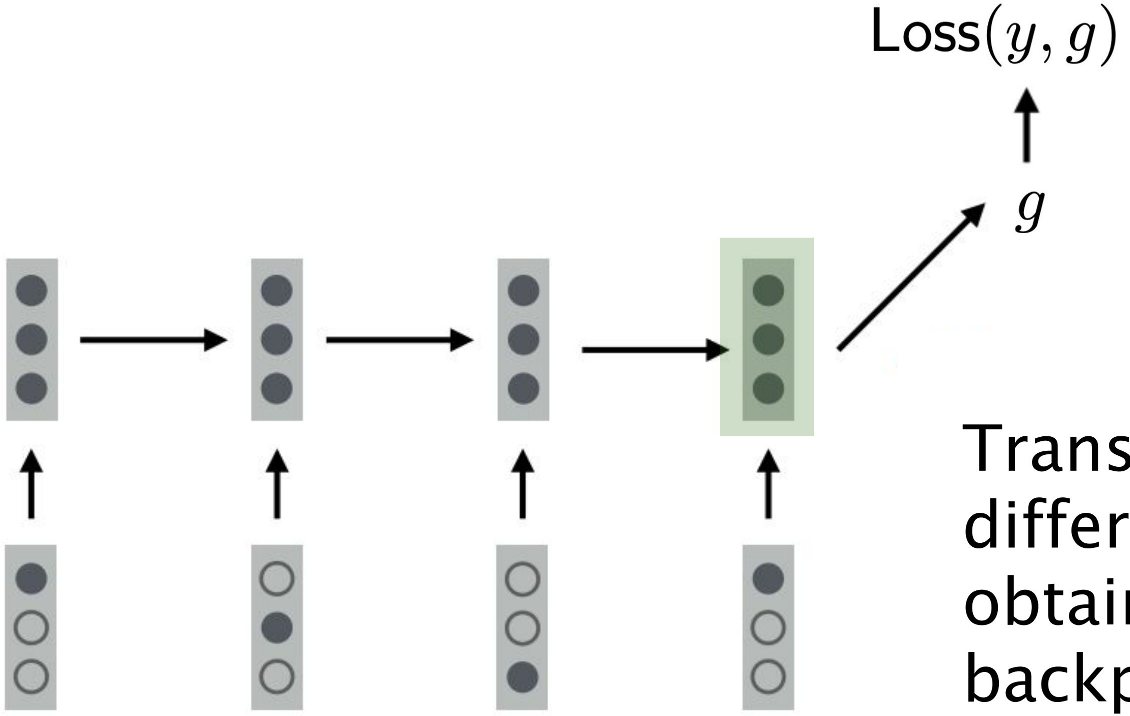


$$g = w^\top h_T + b$$

$$g = \sigma(w^\top h_T + b)$$

$$g = \mathrm{softmax}(W h_T + b)$$

RNNs for Sequence Classification



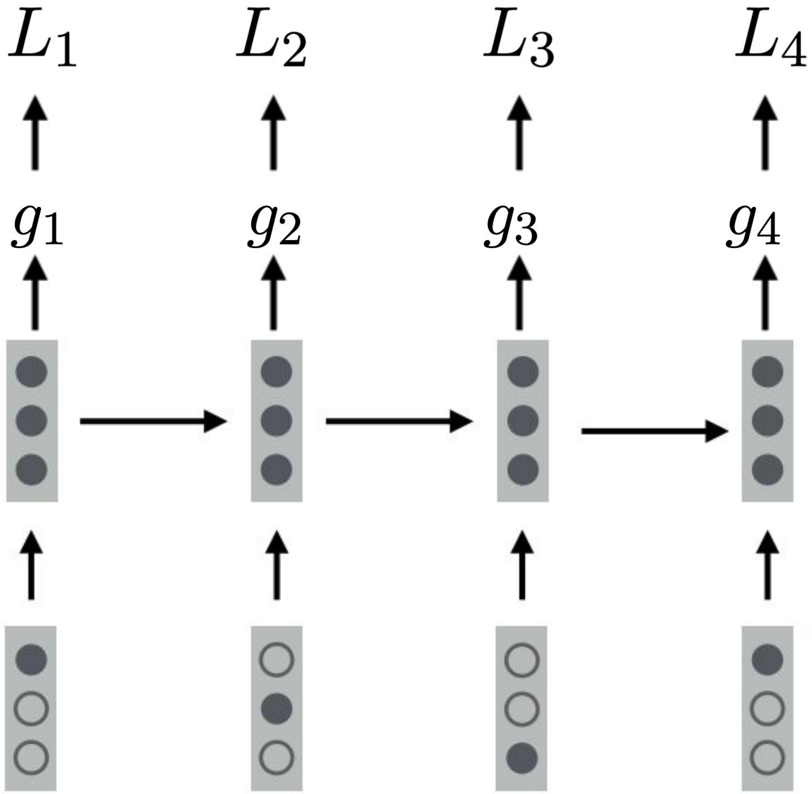
$$g = w^\top h_T + b$$

$$g = \sigma(w^\top h_T + b)$$

$$g = \text{softmax}(W h_T + b)$$

Transition function is differentiable ⇒ can obtain gradients with backpropagation.

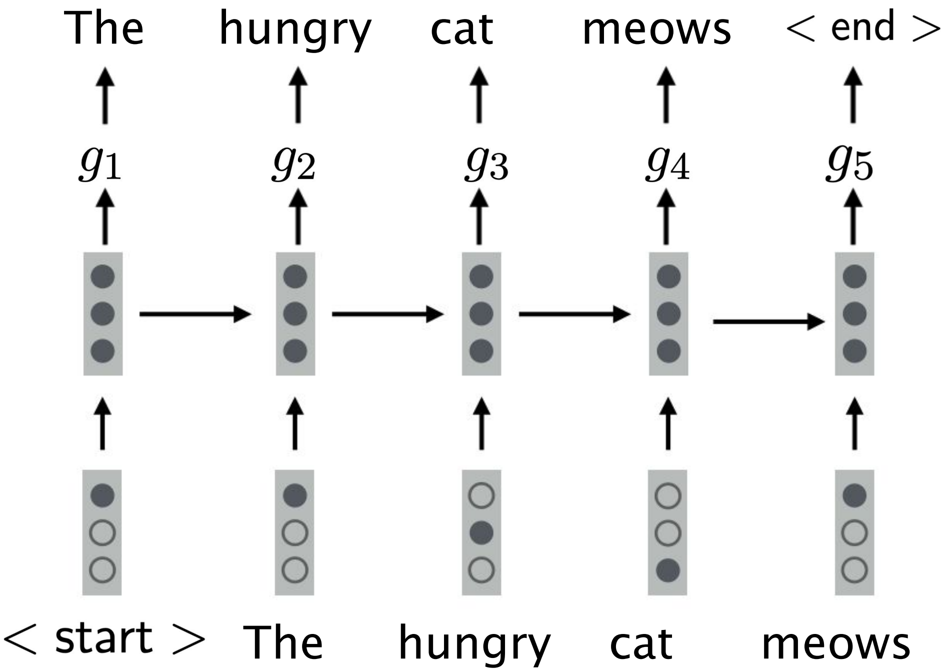
RNNs for Sequence Tagging



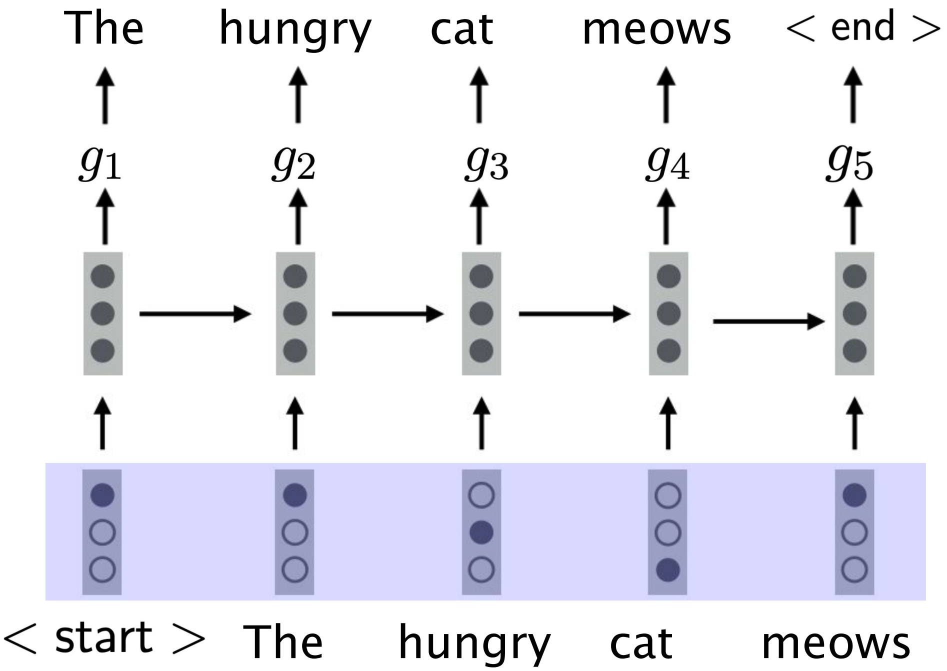
$$L_t = \mathsf{Loss}(g_t, y_t)$$

$$L = \sum_{t=1}^{T} \mathsf{Loss}(g_t, y_t)$$

RNNs for Language Modeling



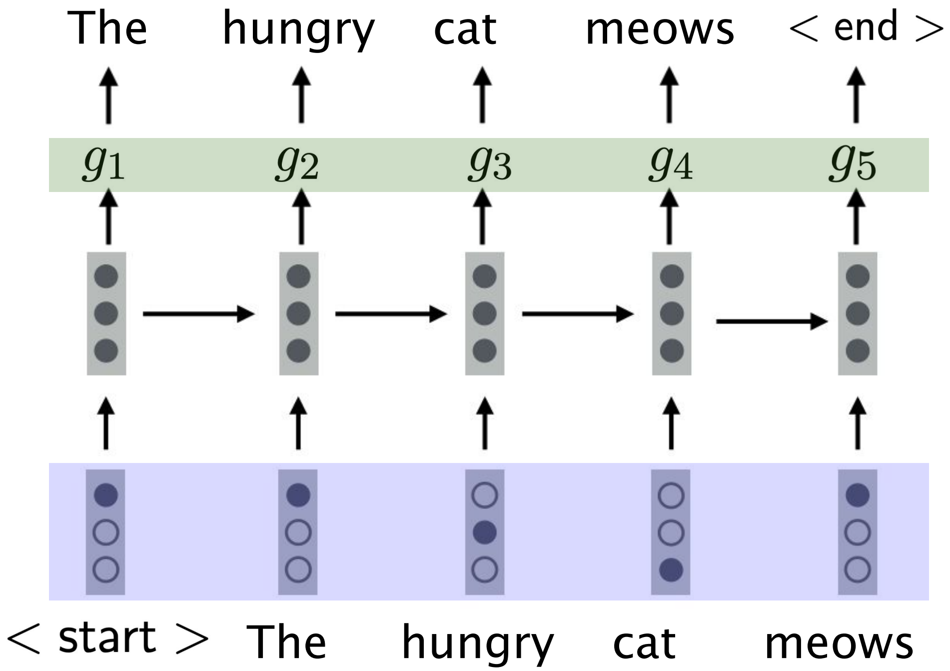
RNNs for Language Modeling



$$s_t = \text{tanh}(W^{sx}x_t + W^{ss}s_{t-1} + W_0^{ss})$$

One-hot vector with dimension
= Vocab size
(10K-100K)

RNNs for Language Modeling



Distribution over words in the vocab

$$g_t = \text{softmax}(W^0 s_t + W_1^0)$$

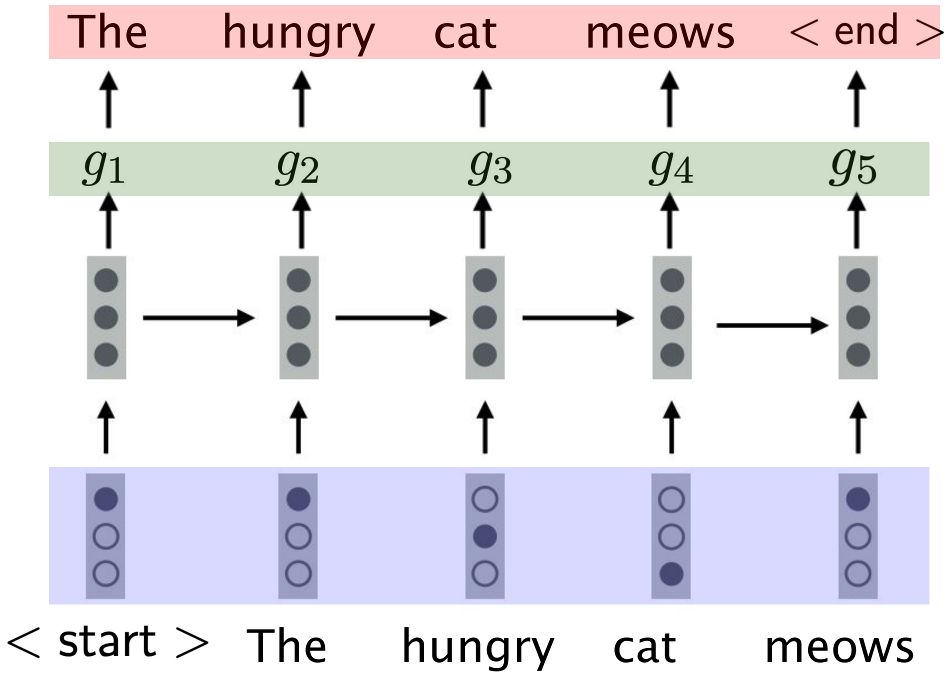
$$s_t = \tanh(W^{sx} x_t + W^{ss} s_{t-1} + W_0^{ss})$$

One-hot vector with dimension = Vocab size (10K-100K)

RNNs for Language Modeling

$$L = \sum_{t=1}^{T} \text{Loss}(g_t, x_{t+1})$$

Total loss = sum over multiclass negative log likelihood



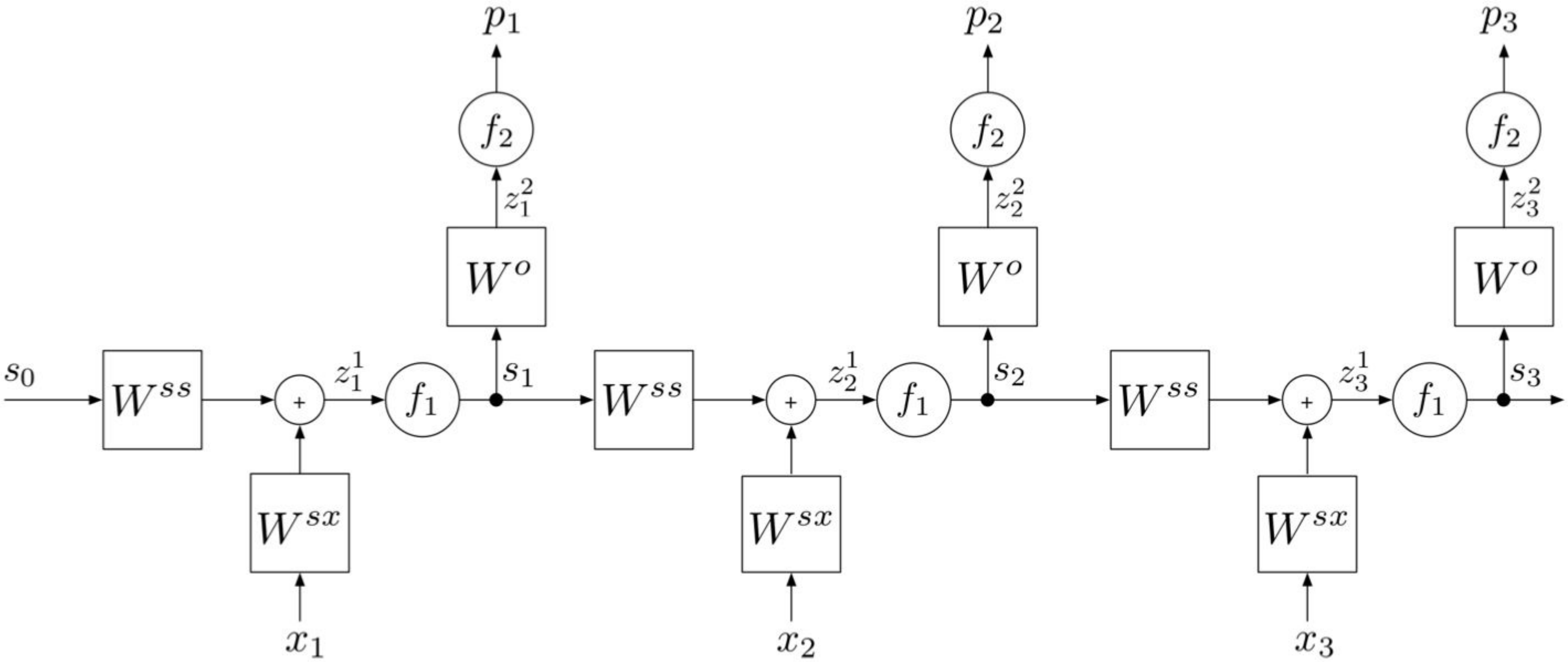
Distribution over words in the vocab

$$g_t = \text{softmax}(W^0 s_t + W_1^0)$$

$$s_t = \tanh(W^{sx} x_t + W^{ss} s_{t-1} + W_0^{ss})$$

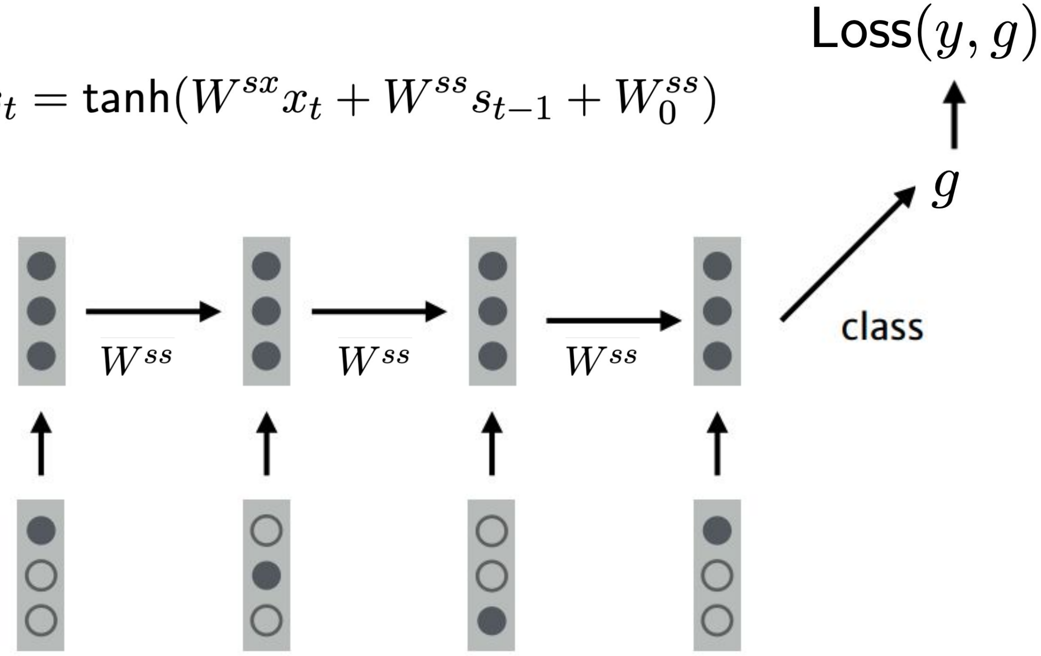
One-hot vector with dimension = Vocab size (10K-100K)

RNNs for Language Modeling



RNN Training: Backpropagation Through Time

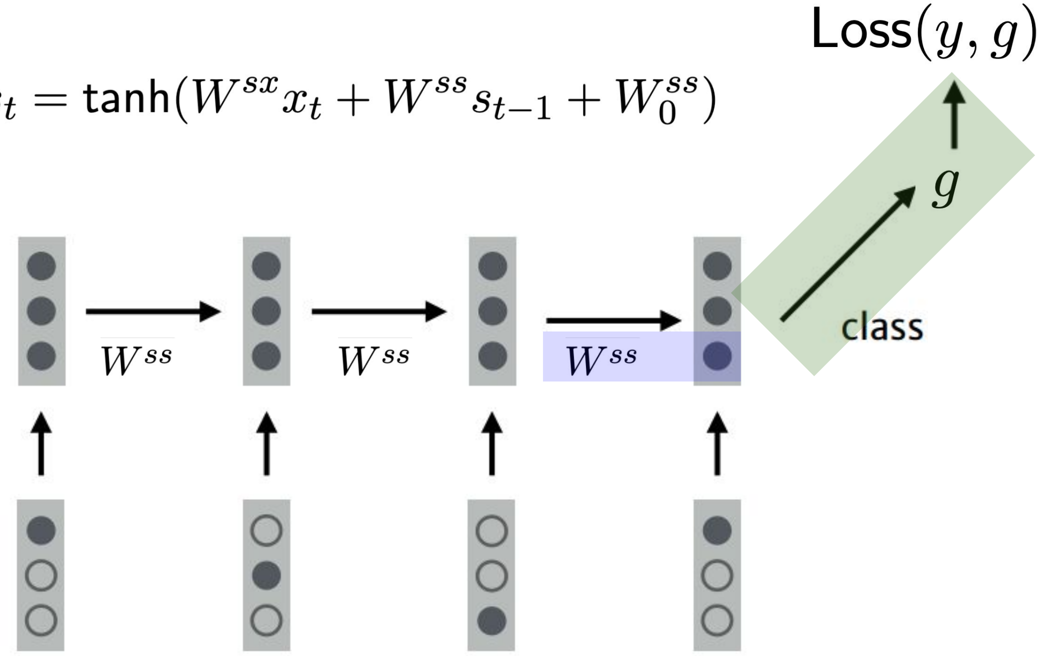
$$s_t = \mathsf{tanh}(W^{sx} x_t + W^{ss} s_{t-1} + W_0^{ss})$$



$$\frac{\partial L}{\partial W^{ss}} =$$

RNN Training: Backpropagation Through Time

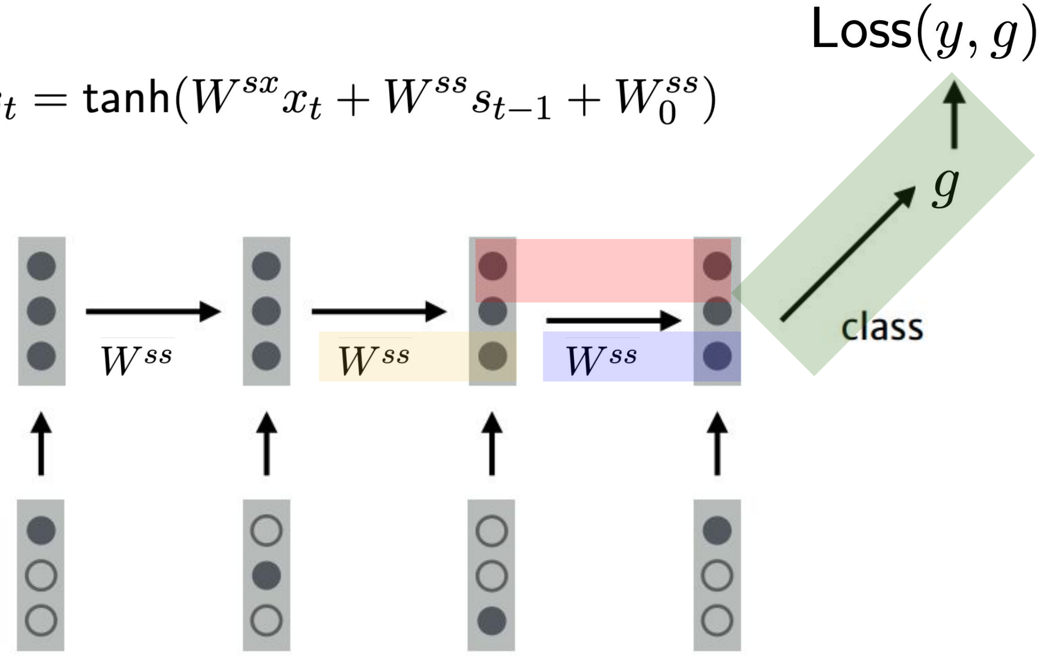
$$s_t = \mathsf{tanh}(W^{sx}x_t + W^{ss}s_{t-1} + W_0^{ss})$$



$$\frac{\partial L}{\partial W^{ss}} = \frac{\partial s_T}{\partial W^{ss}}\frac{\partial L}{\partial s_T}$$

RNN Training: Backpropagation Through Time

$$s_t = \mathrm{tanh}(W^{sx} x_t + W^{ss} s_{t-1} + W_0^{ss})$$

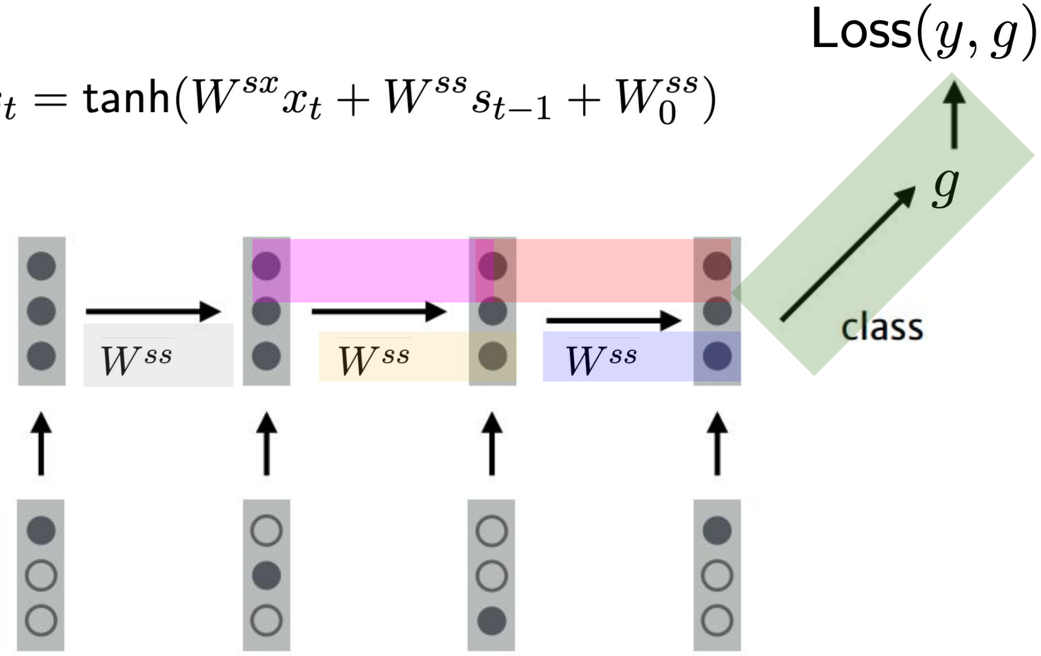


$$\frac{\partial L}{\partial W^{ss}} = \frac{\partial s_T}{\partial W^{ss}} \frac{\partial L}{\partial s_T}$$

$$+ \frac{\partial s_{T-1}}{\partial W^{ss}} \frac{\partial s_T}{\partial s_{T-1}} \frac{\partial L}{\partial s_T}$$

RNN Training: Backpropagation Through Time

$$s_t = \mathsf{tanh}(W^{sx}x_t + W^{ss}s_{t-1} + W_0^{ss})$$



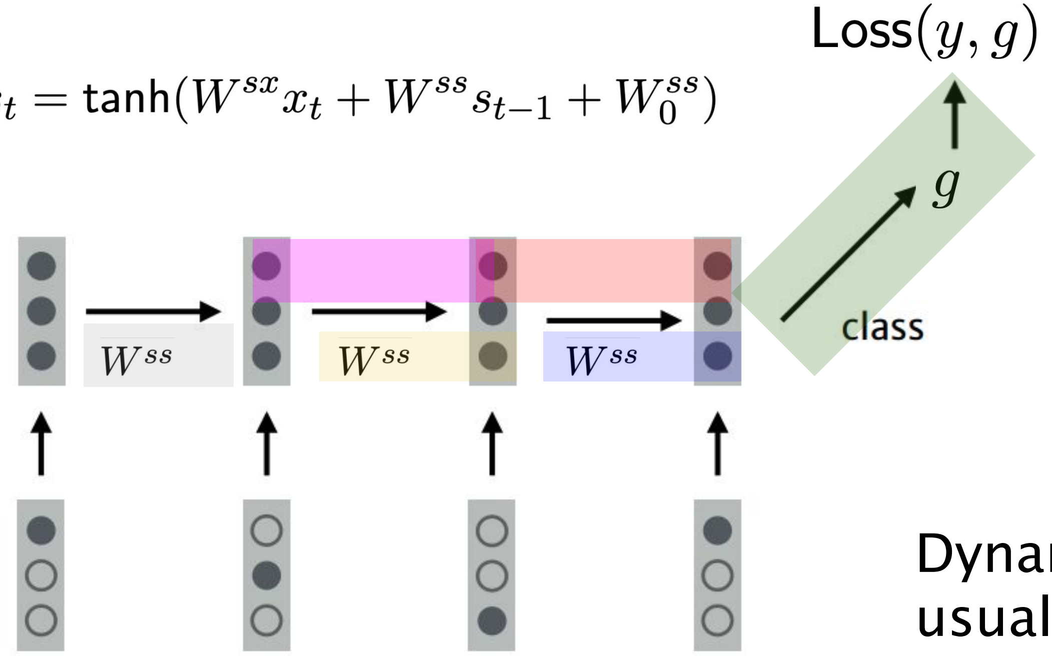
$$\frac{\partial L}{\partial W^{ss}} = \frac{\partial s_T}{\partial W^{ss}}\frac{\partial L}{\partial s_T}$$

$$+ \frac{\partial s_{T-1}}{\partial W^{ss}}\frac{\partial s_T}{\partial s_{T-1}}\frac{\partial L}{\partial s_T}$$

$$+ \frac{\partial s_{T-2}}{\partial W^{ss}}\frac{\partial s_{T-1}}{\partial s_{T-2}}\frac{\partial s_T}{\partial s_{T-1}}\frac{\partial L}{\partial s_T}$$

RNN Training: Backpropagation Through Time

$$s_t = \text{tanh}(W^{sx}x_t + W^{ss}s_{t-1} + W_0^{ss})$$

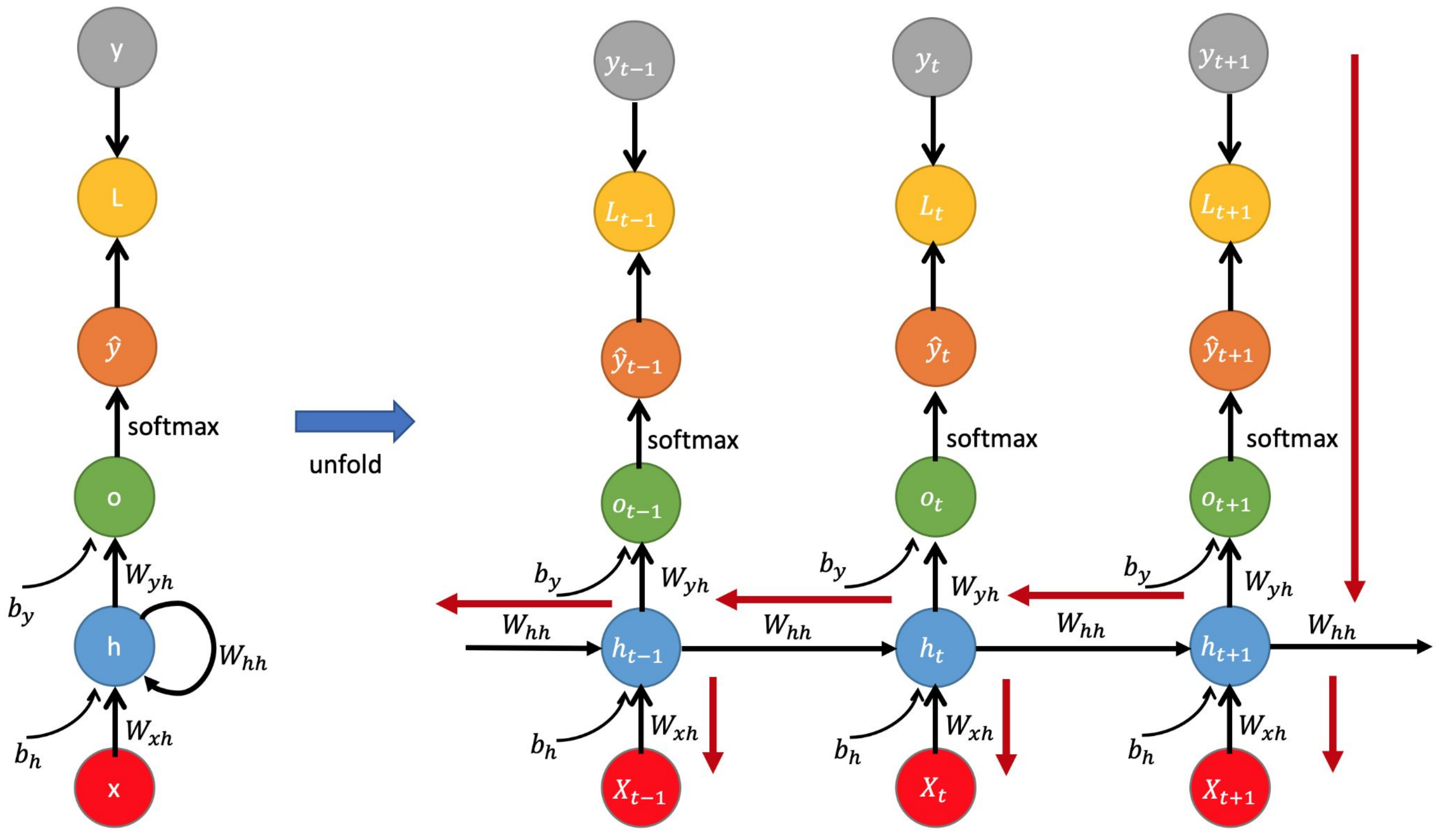


$$\frac{\partial L}{\partial W^{ss}} = \frac{\partial s_T}{\partial W^{ss}}\frac{\partial L}{\partial s_T}$$
$$+ \frac{\partial s_{T-1}}{\partial W^{ss}}\frac{\partial s_T}{\partial s_{T-1}}\frac{\partial L}{\partial s_T}$$
$$+ \frac{\partial s_{T-2}}{\partial W^{ss}}\frac{\partial s_{T-1}}{\partial s_{T-2}}\frac{\partial s_T}{\partial s_{T-1}}\frac{\partial L}{\partial s_T}$$

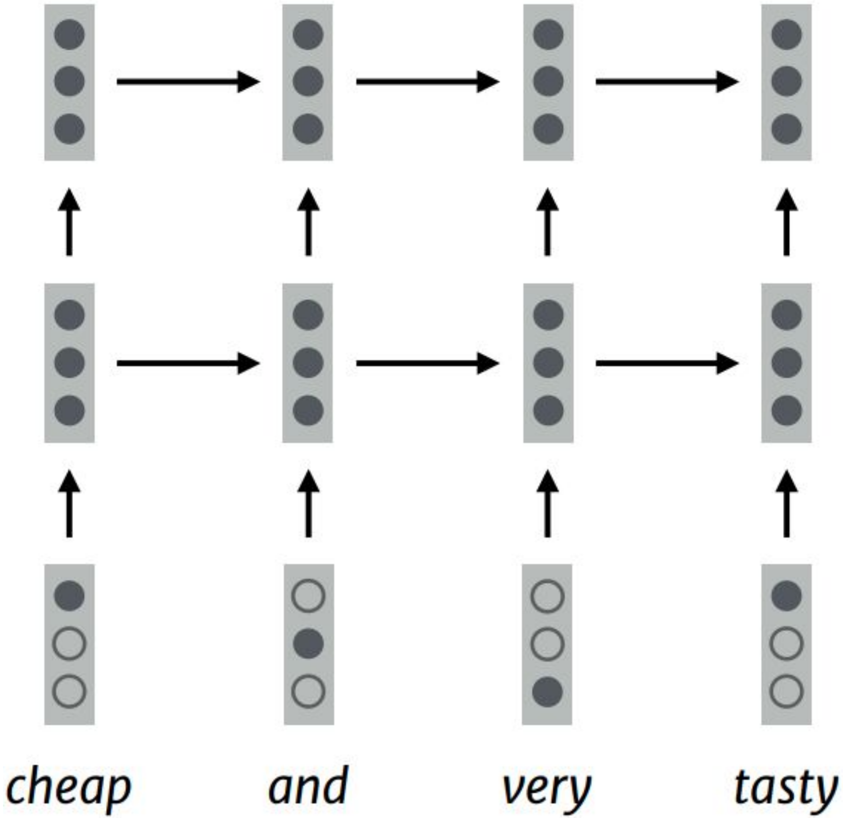
Dynamic programming (as usual) to calculate gradients

Intuition: like a regular neural network "unrolled" in time

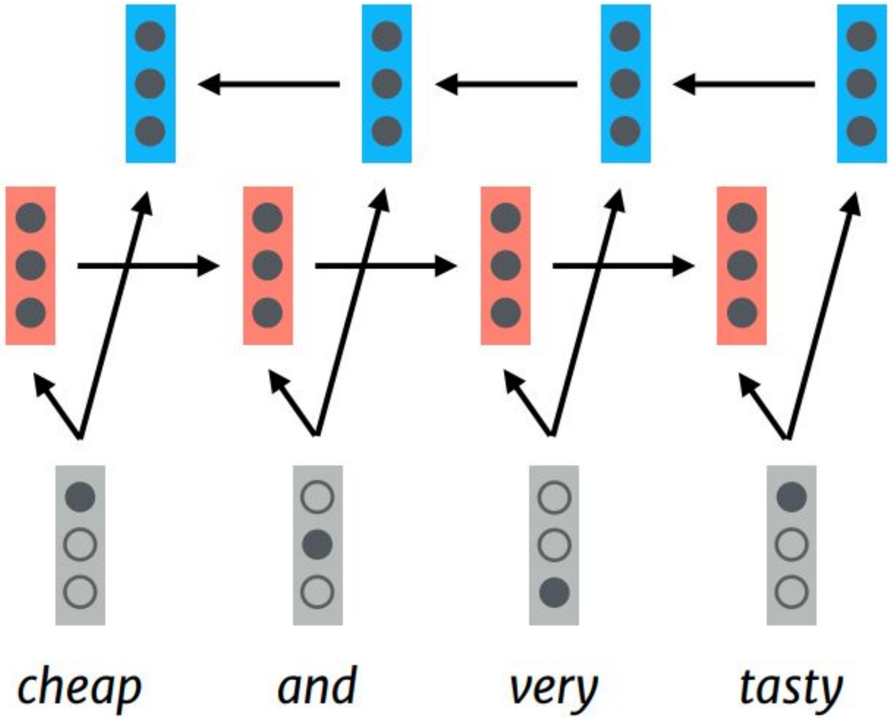
RNN Training: Backpropagation Through Time



Deeper RNNs

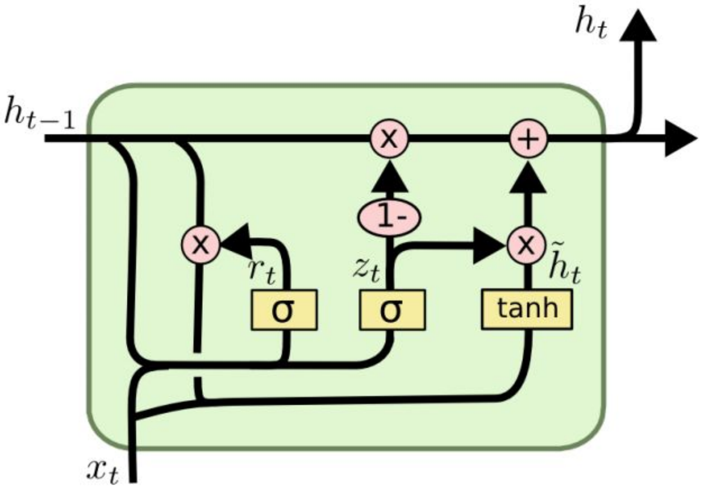


Bidirectional RNNs



Gated RNNs

Gated Recurrent Unit (GRU) [Chung et al. 2014, Cho et al. 2014]



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

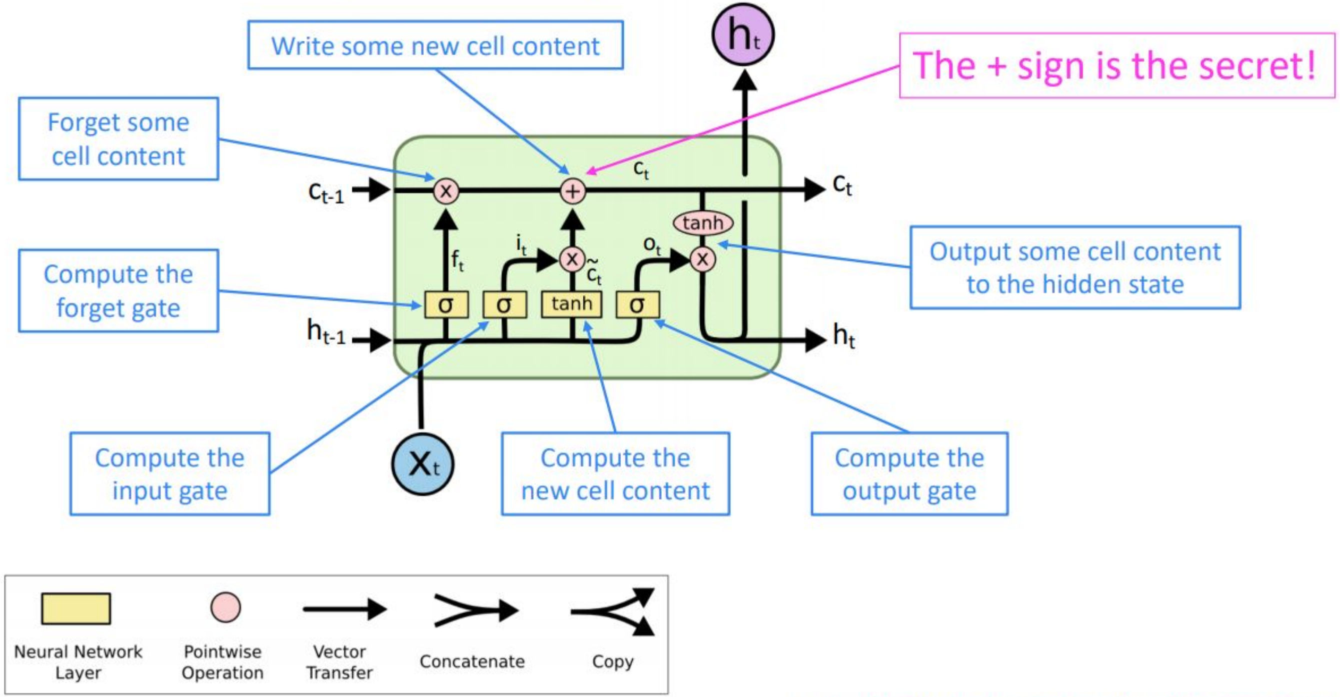
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

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

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

Gated RNNs

Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber 1997]



Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Summary

- Recurrent Neural Networks: tailored for processing sequential data

- RNN Applications:
    - Sequence Classification
    - Language Modeling (GPT3 is language model!)

- RNN Variants
    - Deeper / Bi-directional RNNs
    - Gated RNNs