Machine Learning

A. G. Schwing & M. Telgarsky

University of Illinois at Urbana-Champaign, 2018

L22: Autoregressive Methods (RNNs/LSTMs/GRUs)

Goals of this lecture

Goals of this lecture

- Getting to know Recurrent Neural Nets (RNNs)

Goals of this lecture

- Getting to know Recurrent Neural Nets (RNNs)
- Getting to know Long short term memory (LSTM)

Goals of this lecture

- Getting to know Recurrent Neural Nets (RNNs)
- Getting to know Long short term memory (LSTM)
- Getting to know Gated recurrent unit (GRU)

Goals of this lecture

- Getting to know Recurrent Neural Nets (RNNs)
- Getting to know Long short term memory (LSTM)
- Getting to know Gated recurrent unit (GRU)
- Seeing how to apply them

Goals of this lecture

- Getting to know Recurrent Neural Nets (RNNs)
- Getting to know Long short term memory (LSTM)
- Getting to know Gated recurrent unit (GRU)
- Seeing how to apply them

Reading Material

- Goodfellow et al.; Deep Learning; Chapter 10
- Papers cited on the slides

Pixel Recurrent Neural Networks

Recap: Our models so far

Recap: Our models so far

- Discriminative

Recap: Our models so far

- Discriminative

$$p(\boldsymbol{y}|x)$$

Recap: Our models so far

- Discriminative

$$p(\boldsymbol{y}|x)$$

- Generative

Recap: Our models so far

- Discriminative

$$p(\boldsymbol{y}|x)$$

- Generative

$$p(x)$$

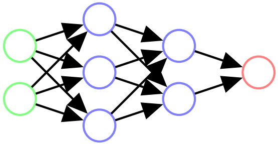
Recap: Our models so far

- Discriminative

$$p(\boldsymbol{y}|x)$$

- Generative

$$p(x)$$



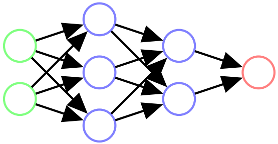
Recap: Our models so far

- Discriminative

$$p(\boldsymbol{y}|x)$$

- Generative

$$p(x)$$



What's missing?

More flexibility regarding inputs and outputs:

More flexibility regarding inputs and outputs:

- Sequences of inputs

More flexibility regarding inputs and outputs:

- Sequences of inputs
- Sequences of outputs

More flexibility regarding inputs and outputs:

- Sequences of inputs
- Sequences of outputs

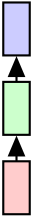
Length of sequences may vary

More flexibility regarding inputs and outputs:

- Sequences of inputs
- Sequences of outputs

Length of sequences may vary

one to one

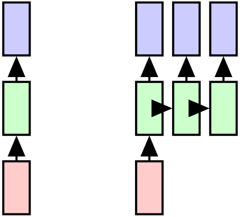


More flexibility regarding inputs and outputs:

- Sequences of inputs
- Sequences of outputs

Length of sequences may vary

one to one one to many

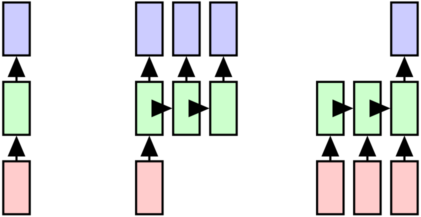


More flexibility regarding inputs and outputs:

- Sequences of inputs
- Sequences of outputs

Length of sequences may vary

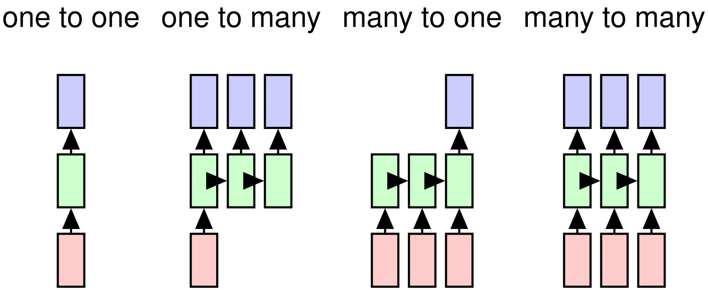
one to one one to many many to one



More flexibility regarding inputs and outputs:

- Sequences of inputs
- Sequences of outputs

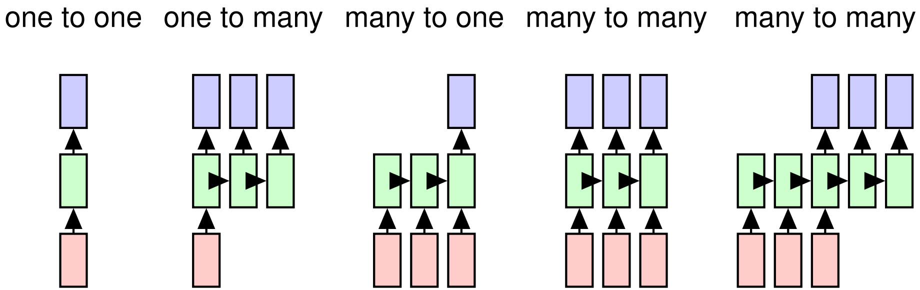
Length of sequences may vary



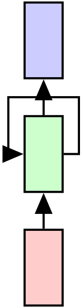
More flexibility regarding inputs and outputs:

- Sequences of inputs
- Sequences of outputs

Length of sequences may vary

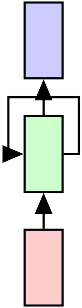


Recurrent Neural Nets (RNNs)



Recurrent Neural Nets (RNNs)

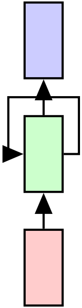
- input depends on previous output



Recurrent Neural Nets (RNNs)

- input depends on previous output

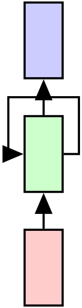
$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$



Recurrent Neural Nets (RNNs)

- input depends on previous output

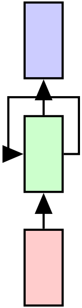
$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$



Recurrent Neural Nets (RNNs)

- input depends on previous output

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$

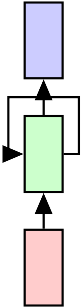


Applications:

Recurrent Neural Nets (RNNs)

- input depends on previous output

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$



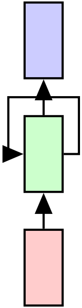
Applications:

- Natural language processing

Recurrent Neural Nets (RNNs)

- input depends on previous output

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$



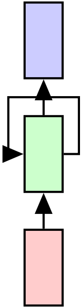
Applications:

- Natural language processing
- Speech recognition

Recurrent Neural Nets (RNNs)

- input depends on previous output

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$



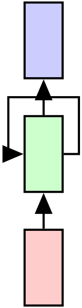
Applications:

- Natural language processing
- Speech recognition
- Image processing

Recurrent Neural Nets (RNNs)

- input depends on previous output

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$

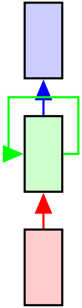


Applications:

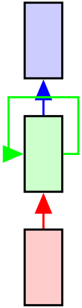
- Natural language processing
- Speech recognition
- Image processing
- Video processing

Important concept: Parameter sharing

Important concept: Parameter sharing

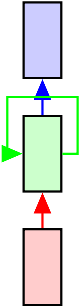


Important concept: Parameter sharing



$$
\begin{aligned}
h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\
y^{(t)} &= g(h^{(t)})
\end{aligned}
$$

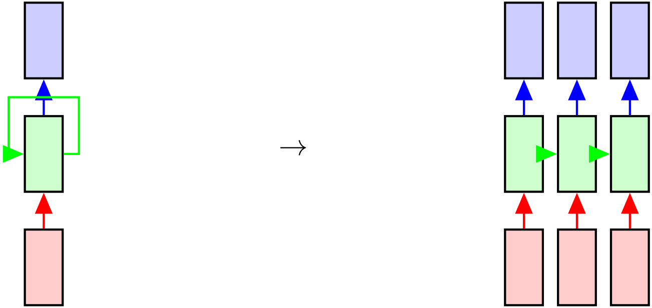
Important concept: Parameter sharing



$\rightarrow$

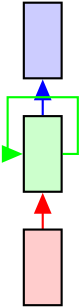
$$
\begin{aligned}
h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\
y^{(t)} &= g(h^{(t)})
\end{aligned}
$$

Important concept: Parameter sharing

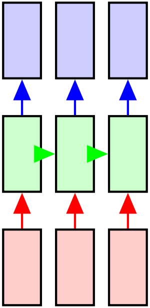


$$
\begin{aligned}
h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\
y^{(t)} &= g(h^{(t)})
\end{aligned}
$$

Important concept: Parameter sharing



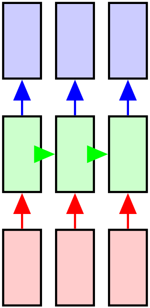
$\rightarrow$



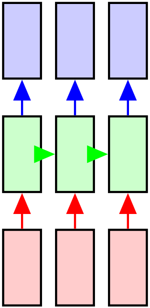
$$
\begin{aligned}
h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\
y^{(t)} &= g(h^{(t)})
\end{aligned}
$$

unfolded/unrolled network
performs identical operations
easier to understand

General structure for recurrence:

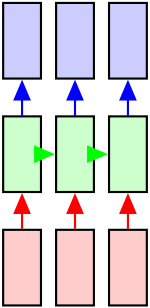


General structure for recurrence:



Mathematical description in general:

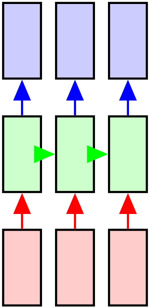
General structure for recurrence:



Mathematical description in general:

$$\begin{aligned} h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\ y^{(t)} &= g(h^{(t)}) \end{aligned}$$

General structure for recurrence:



Mathematical description in general:

$$\begin{aligned} h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\ y^{(t)} &= g(h^{(t)}) \end{aligned}$$

Note that f and g are independent of time

What are f and g?

What are f and g?

Any differentiable function can be used

What are f and g?

Any differentiable function can be used

Useful functions:

What are f and g?

Any differentiable function can be used

Useful functions:

- Original recurrent nets

What are f and g?

Any differentiable function can be used

Useful functions:

- Original recurrent nets
- LSTM nets

What are f and g?

Any differentiable function can be used

Useful functions:

- Original recurrent nets
- LSTM nets
- GRU nets

Original recurrent nets (Elman network):

(Jordan network is slightly different)

Generally:

Specifically:

Original recurrent nets (Elman network):

(Jordan network is slightly different)

Generally:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$

Specifically:

Original recurrent nets (Elman network):

(Jordan network is slightly different)

Generally:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$

Specifically:

$$h^{(t)} = \sigma_h(W_{hx}x^{(t)} + W_{hh}h^{(t-1)} + w_{hb})$$

Original recurrent nets (Elman network):

(Jordan network is slightly different)

Generally:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$

$$y^{(t)} = g(h^{(t)})$$

Specifically:

$$h^{(t)} = \sigma_h(W_{hx}x^{(t)} + W_{hh}h^{(t-1)} + w_{hb})$$

Original recurrent nets (Elman network):

(Jordan network is slightly different)

Generally:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$

$$y^{(t)} = g(h^{(t)})$$

Specifically:

$$h^{(t)} = \sigma_h(W_{hx}x^{(t)} + W_{hh}h^{(t-1)} + w_{hb})$$

$$y^{(t)} = \sigma_y(W_{yh}h^{(t)} + w_{yb})$$

Original recurrent nets (Elman network):

(Jordan network is slightly different)

Generally:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$

$$y^{(t)} = g(h^{(t)})$$

Specifically:

$$h^{(t)} = \sigma_h(W_{hx}x^{(t)} + W_{hh}h^{(t-1)} + w_{hb})$$

$$y^{(t)} = \sigma_y(W_{yh}h^{(t)} + w_{yb})$$

What is σ_h and σ_y?

Original recurrent nets (Elman network):

(Jordan network is slightly different)

Generally:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$

$$y^{(t)} = g(h^{(t)})$$

Specifically:

$$h^{(t)} = \sigma_h(W_{hx}x^{(t)} + W_{hh}h^{(t-1)} + w_{hb})$$

$$y^{(t)} = \sigma_y(W_{yh}h^{(t)} + w_{yb})$$

What is σ_h and σ_y?

Activation functions:

Original recurrent nets (Elman network):

(Jordan network is slightly different)

Generally:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}, \boldsymbol{w})$$

$$y^{(t)} = g(h^{(t)})$$

Specifically:

$$h^{(t)} = \sigma_h(W_{hx}x^{(t)} + W_{hh}h^{(t-1)} + w_{hb})$$

$$y^{(t)} = \sigma_y(W_{yh}h^{(t)} + w_{yb})$$

What is σ_h and σ_y?

Activation functions: tanh, sigmoid

Affine transformations and point-wise non-linearity

What are the problems?

Problems with classical recurrent neural nets:

Problems with classical recurrent neural nets:

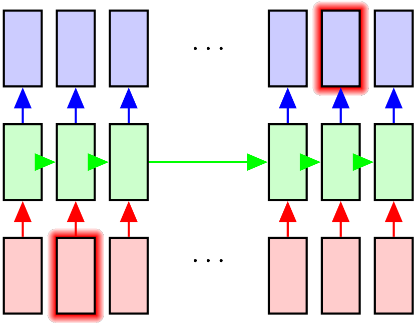
- Vanishing gradients

Problems with classical recurrent neural nets:

- Vanishing gradients
- Long term dependency

Problems with classical recurrent neural nets:

- Vanishing gradients
- Long term dependency



Long short term memory (LSTM)

Long short term memory (LSTM)

- Particular functional relation

Long short term memory (LSTM)

- Particular functional relation
- Shown to better capture long-term dependencies

Long short term memory (LSTM)

- Particular functional relation
- Shown to better capture long-term dependencies
- Shown to address the vanishing gradient problem

Long short term memory (LSTM)

- Particular functional relation
- Shown to better capture long-term dependencies
- Shown to address the vanishing gradient problem

Generally:

$$
\begin{aligned}
h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\
y^{(t)} &= g(h^{(t)})
\end{aligned}
$$

Long short term memory (LSTM)

- Particular functional relation
- Shown to better capture long-term dependencies
- Shown to address the vanishing gradient problem

Generally:

$$\begin{aligned} h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\ y^{(t)} &= g(h^{(t)}) \end{aligned}$$

Specifically: ($\circ$ denotes Hadamard product; σ is activation function)

$$i^{(t)} = \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) \qquad \text{Input gate}$$

Long short term memory (LSTM)

- Particular functional relation
- Shown to better capture long-term dependencies
- Shown to address the vanishing gradient problem

Generally:

$$\begin{aligned} h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\ y^{(t)} &= g(h^{(t)}) \end{aligned}$$

Specifically: ($\circ$ denotes Hadamard product; σ is activation function)

$$\begin{aligned} i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate} \\ f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate} \end{aligned}$$

Long short term memory (LSTM)

- Particular functional relation
- Shown to better capture long-term dependencies
- Shown to address the vanishing gradient problem

Generally:

$$
\begin{aligned}
h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\
y^{(t)} &= g(h^{(t)})
\end{aligned}
$$

Specifically: ($\circ$ denotes Hadamard product; σ is activation function)

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate} \\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate} \\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate}
\end{aligned}
$$

Long short term memory (LSTM)

- Particular functional relation
- Shown to better capture long-term dependencies
- Shown to address the vanishing gradient problem

Generally:

$$\begin{aligned} h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\ y^{(t)} &= g(h^{(t)}) \end{aligned}$$

Specifically: ($\circ$ denotes Hadamard product; σ is activation function)

$$\begin{aligned} i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate} \\ f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate} \\ o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate} \\ \tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell} \end{aligned}$$

Long short term memory (LSTM)

- Particular functional relation
- Shown to better capture long-term dependencies
- Shown to address the vanishing gradient problem

Generally:

$$
\begin{aligned}
h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\
y^{(t)} &= g(h^{(t)})
\end{aligned}
$$

Specifically: ($\circ$ denotes Hadamard product; σ is activation function)

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate} \\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate} \\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate} \\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell} \\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell}
\end{aligned}
$$

Long short term memory (LSTM)

- Particular functional relation
- Shown to better capture long-term dependencies
- Shown to address the vanishing gradient problem

Generally:

$$\begin{aligned} h^{(t)} &= f(h^{(t-1)}, x^{(t)}, \boldsymbol{w}) \\ y^{(t)} &= g(h^{(t)}) \end{aligned}$$

Specifically: ($\circ$ denotes Hadamard product; σ is activation function)

$$\begin{aligned} i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate} \\ f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate} \\ o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate} \\ \tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell} \\ c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell} \\ h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)}) \end{aligned}$$

Equations:

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate} \\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate} \\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate} \\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell} \\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell} \\
h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)})
\end{aligned}
$$

Intuition:

Equations:

$$
\begin{array}{rcll}
i^{(t)} & = & \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) & \text{Input gate} \\
f^{(t)} & = & \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) & \text{Forget gate} \\
o^{(t)} & = & \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) & \text{Output/Exposure gate} \\
\tilde{c}^{(t)} & = & \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) & \text{New memory cell} \\
c^{(t)} & = & f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} & \text{Final memory cell} \\
h^{(t)} & = & o^{(t)} \circ \sigma_h(c^{(t)}) &
\end{array}
$$

Intuition:

- $i^{(t)}$:

Equations:

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate}\\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate}\\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate}\\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell}\\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell}\\
h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)})
\end{aligned}
$$

Intuition:

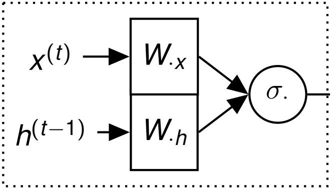
- $i^{(t)}$: Does $x^{(t)}$ matter?

Equations:

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate}\\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate}\\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate}\\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell}\\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell}\\
h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)})
\end{aligned}
$$

Intuition:

- $i^{(t)}$: Does $x^{(t)}$ matter?

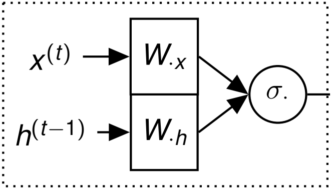


Equations:

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate}\\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate}\\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate}\\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell}\\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell}\\
h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)})
\end{aligned}
$$

Intuition:

- $i^{(t)}$: Does $x^{(t)}$ matter?
- $f^{(t)}$:

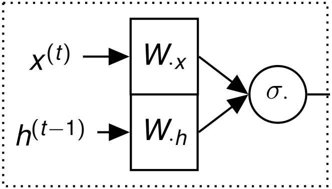


Equations:

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate}\\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate}\\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate}\\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell}\\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell}\\
h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)})
\end{aligned}
$$

Intuition:

- $i^{(t)}$: Does $x^{(t)}$ matter?
- $f^{(t)}$: Should $c^{(t-1)}$ be forgotten?

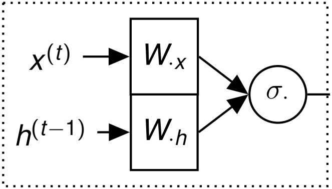


Equations:

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate}\\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate}\\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate}\\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell}\\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell}\\
h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)})
\end{aligned}
$$

Intuition:

- $i^{(t)}$: Does $x^{(t)}$ matter?
- $f^{(t)}$: Should $c^{(t-1)}$ be forgotten?
- $o^{(t)}$:

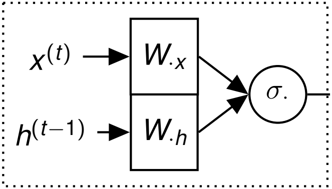


Equations:

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate}\\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate}\\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate}\\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell}\\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell}\\
h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)})
\end{aligned}
$$

Intuition:

- $i^{(t)}$: Does $x^{(t)}$ matter?
- $f^{(t)}$: Should $c^{(t-1)}$ be forgotten?
- $o^{(t)}$: How much $c^{(t)}$ should be exposed?

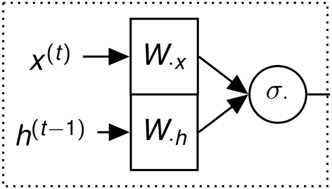


Equations:

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate}\\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate}\\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate}\\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell}\\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell}\\
h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)})
\end{aligned}
$$

Intuition:

- $i^{(t)}$: Does $x^{(t)}$ matter?
- $f^{(t)}$: Should $c^{(t-1)}$ be forgotten?
- $o^{(t)}$: How much $c^{(t)}$ should be exposed?
- $\tilde{c}^{(t)}$:

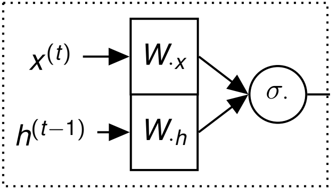


Equations:

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate}\\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate}\\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate}\\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell}\\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell}\\
h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)})
\end{aligned}
$$

Intuition:

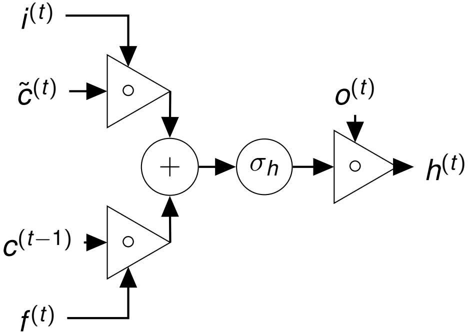
- $i^{(t)}$: Does $x^{(t)}$ matter?
- $f^{(t)}$: Should $c^{(t-1)}$ be forgotten?
- $o^{(t)}$: How much $c^{(t)}$ should be exposed?
- $\tilde{c}^{(t)}$: Compute new memory



Equations:

$$
\begin{aligned}
i^{(t)} &= \sigma_i(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + w_{bi}) && \text{Input gate}\\
f^{(t)} &= \sigma_f(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + w_{bf}) && \text{Forget gate}\\
o^{(t)} &= \sigma_o(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + w_{bo}) && \text{Output/Exposure gate}\\
\tilde{c}^{(t)} &= \sigma_c(W_{cx}x^{(t)} + W_{ch}h^{(t-1)} + w_{bc}) && \text{New memory cell}\\
c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} && \text{Final memory cell}\\
h^{(t)} &= o^{(t)} \circ \sigma_h(c^{(t)})
\end{aligned}
$$

Intuition:



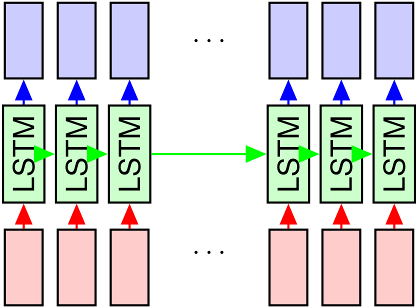
Long short term memory (LSTM):

Long short term memory (LSTM):

- Can be interpreted as a block in a neural net

Long short term memory (LSTM):

- Can be interpreted as a block in a neural net



Gated recurrent unit (GRU):

Gated recurrent unit (GRU):

- Performance similar to LSTM

Gated recurrent unit (GRU):

- Performance similar to LSTM
- Fewer parameters compared to LSTM (no output gate)

Gated recurrent unit (GRU):

- Performance similar to LSTM
- Fewer parameters compared to LSTM (no output gate)

Equations: ($\circ$ denotes Hadamard product)

Gated recurrent unit (GRU):

- Performance similar to LSTM
- Fewer parameters compared to LSTM (no output gate)

Equations: ($\circ$ denotes Hadamard product)

$$z^{(t)} = \sigma_z(W_{zx}x^{(t)} + W_{zh}h^{(t-1)} + w_{bz}) \qquad \text{Update gate}$$

Gated recurrent unit (GRU):

- Performance similar to LSTM
- Fewer parameters compared to LSTM (no output gate)

Equations: ($\circ$ denotes Hadamard product)

$$
\begin{aligned}
z^{(t)} &= \sigma_z(W_{zx}x^{(t)} + W_{zh}h^{(t-1)} + w_{bz}) && \text{Update gate} \\
r^{(t)} &= \sigma_r(W_{rx}x^{(t)} + W_{rh}h^{(t-1)} + w_{br}) && \text{Reset gate}
\end{aligned}
$$

Gated recurrent unit (GRU):

- Performance similar to LSTM
- Fewer parameters compared to LSTM (no output gate)

Equations: ($\circ$ denotes Hadamard product)

$$\begin{aligned}
z^{(t)} &= \sigma_z(W_{zx}x^{(t)} + W_{zh}h^{(t-1)} + w_{bz}) && \text{Update gate} \\
r^{(t)} &= \sigma_r(W_{rx}x^{(t)} + W_{rh}h^{(t-1)} + w_{br}) && \text{Reset gate} \\
\tilde{h}^{(t)} &= \sigma_h(W_{hx}x^{(t)} + W_{rwh}(r^{(t)} \circ h^{(t-1)}) + w_{bh}) && \text{New memory cell}
\end{aligned}$$

Gated recurrent unit (GRU):

- Performance similar to LSTM
- Fewer parameters compared to LSTM (no output gate)

Equations: ($\circ$ denotes Hadamard product)

$$
\begin{aligned}
z^{(t)} &= \sigma_z(W_{zx}x^{(t)} + W_{zh}h^{(t-1)} + w_{bz}) && \text{Update gate}\\
r^{(t)} &= \sigma_r(W_{rx}x^{(t)} + W_{rh}h^{(t-1)} + w_{br}) && \text{Reset gate}\\
\tilde{h}^{(t)} &= \sigma_h(W_{hx}x^{(t)} + W_{rwh}(r^{(t)} \circ h^{(t-1)}) + w_{bh}) && \text{New memory cell}\\
h^{(t)} &= (1 - z^{(t)}) \circ \tilde{h}^{(t)} + z^{(t)} \circ h^{(t-1)} && \text{Hidden state}
\end{aligned}
$$

Gated recurrent unit (GRU):

- Performance similar to LSTM
- Fewer parameters compared to LSTM (no output gate)

Equations: ($\circ$ denotes Hadamard product)

$$
\begin{aligned}
z^{(t)} &= \sigma_z(W_{zx}x^{(t)} + W_{zh}h^{(t-1)} + w_{bz}) && \text{Update gate} \\
r^{(t)} &= \sigma_r(W_{rx}x^{(t)} + W_{rh}h^{(t-1)} + w_{br}) && \text{Reset gate} \\
\tilde{h}^{(t)} &= \sigma_h(W_{hx}x^{(t)} + W_{rwh}(r^{(t)} \circ h^{(t-1)}) + w_{bh}) && \text{New memory cell} \\
h^{(t)} &= (1 - z^{(t)}) \circ \tilde{h}^{(t)} + z^{(t)} \circ h^{(t-1)} && \text{Hidden state}
\end{aligned}
$$

Can again be interpreted as a block in the computation graph

Equations: ($\circ$ denotes Hadamard product)

$$\begin{aligned}
z^{(t)} &= \sigma_z(W_{zx}x^{(t)} + W_{zh}h^{(t-1)} + w_{bz}) && \text{Update gate}\\
r^{(t)} &= \sigma_r(W_{rx}x^{(t)} + W_{rh}h^{(t-1)} + w_{br}) && \text{Reset gate}\\
\tilde{h}^{(t)} &= \sigma_h(W_{hx}x^{(t)} + W_{rwh}(r^{(t)} \circ h^{(t-1)}) + w_{bh}) && \text{New memory cell}\\
h^{(t)} &= (1 - z^{(t)}) \circ \tilde{h}^{(t)} + z^{(t)} \circ h^{(t-1)} && \text{Hidden state}
\end{aligned}$$

Intuition:

Equations: ($\circ$ denotes Hadamard product)

$$\begin{aligned}
z^{(t)} &= \sigma_z(W_{zx}x^{(t)} + W_{zh}h^{(t-1)} + w_{bz}) && \text{Update gate} \\
r^{(t)} &= \sigma_r(W_{rx}x^{(t)} + W_{rh}h^{(t-1)} + w_{br}) && \text{Reset gate} \\
\tilde{h}^{(t)} &= \sigma_h(W_{hx}x^{(t)} + W_{rwh}(r^{(t)} \circ h^{(t-1)}) + w_{bh}) && \text{New memory cell} \\
h^{(t)} &= (1 - z^{(t)}) \circ \tilde{h}^{(t)} + z^{(t)} \circ h^{(t-1)} && \text{Hidden state}
\end{aligned}$$

Intuition:

- $r^{(t)}$:

Equations: ($\circ$ denotes Hadamard product)

$$\begin{aligned}
z^{(t)} &= \sigma_z(W_{zx}x^{(t)} + W_{zh}h^{(t-1)} + w_{bz}) && \text{Update gate}\\
r^{(t)} &= \sigma_r(W_{rx}x^{(t)} + W_{rh}h^{(t-1)} + w_{br}) && \text{Reset gate}\\
\tilde{h}^{(t)} &= \sigma_h(W_{hx}x^{(t)} + W_{rwh}(r^{(t)} \circ h^{(t-1)}) + w_{bh}) && \text{New memory cell}\\
h^{(t)} &= (1 - z^{(t)}) \circ \tilde{h}^{(t)} + z^{(t)} \circ h^{(t-1)} && \text{Hidden state}
\end{aligned}$$

Intuition:

- $r^{(t)}$: Include $h^{(t-1)}$ in new memory?

Equations: ($\circ$ denotes Hadamard product)

$$\begin{aligned}
z^{(t)} &= \sigma_z(W_{zx}x^{(t)} + W_{zh}h^{(t-1)} + w_{bz}) && \text{Update gate}\\
r^{(t)} &= \sigma_r(W_{rx}x^{(t)} + W_{rh}h^{(t-1)} + w_{br}) && \text{Reset gate}\\
\tilde{h}^{(t)} &= \sigma_h(W_{hx}x^{(t)} + W_{rwh}(r^{(t)} \circ h^{(t-1)}) + w_{bh}) && \text{New memory cell}\\
h^{(t)} &= (1 - z^{(t)}) \circ \tilde{h}^{(t)} + z^{(t)} \circ h^{(t-1)} && \text{Hidden state}
\end{aligned}$$

Intuition:

- $r^{(t)}$: Include $h^{(t-1)}$ in new memory?
- $z^{(t)}$:

Equations: ($\circ$ denotes Hadamard product)

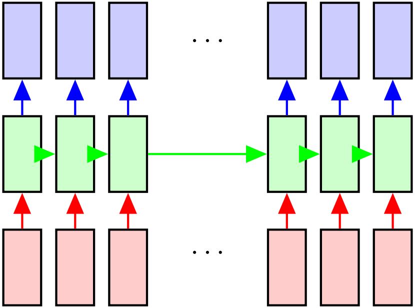
$$
\begin{aligned}
z^{(t)} &= \sigma_z(W_{zx}x^{(t)} + W_{zh}h^{(t-1)} + w_{bz}) && \text{Update gate} \\
r^{(t)} &= \sigma_r(W_{rx}x^{(t)} + W_{rh}h^{(t-1)} + w_{br}) && \text{Reset gate} \\
\tilde{h}^{(t)} &= \sigma_h(W_{hx}x^{(t)} + W_{rwh}(r^{(t)} \circ h^{(t-1)}) + w_{bh}) && \text{New memory cell} \\
h^{(t)} &= (1 - z^{(t)}) \circ \tilde{h}^{(t)} + z^{(t)} \circ h^{(t-1)} && \text{Hidden state}
\end{aligned}
$$

Intuition:

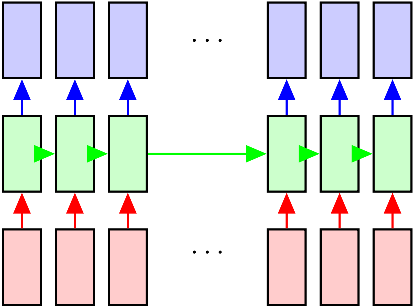
- $r^{(t)}$: Include $h^{(t-1)}$ in new memory?
- $z^{(t)}$: How much $h^{(t-1)}$ in next state?

Recurrent nets generally:

Recurrent nets generally:

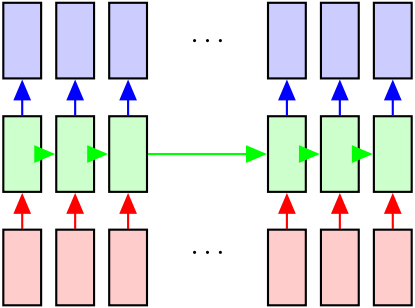


Recurrent nets generally:



Other variants:

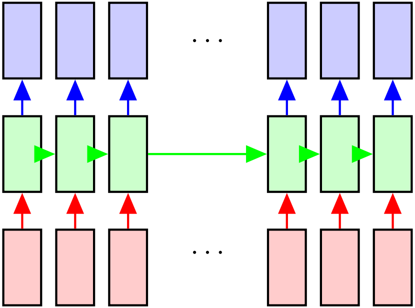
Recurrent nets generally:



Other variants:

- Bi-directional LSTMs [Schuster&Paliwal (1997), Graves&Schmidhuber (2005)]

Recurrent nets generally:



Other variants:

- Bi-directional LSTMs [Schuster&Paliwal (1997), Graves&Schmidhuber (2005)]
- Continuous time RNNs

How do we learn the parameters in the network?

How do we learn the parameters in the network?

$$p(y_1, \ldots, y_T) = \prod_{i=1}^{T} p(y_i | y_1, \ldots y_{i-1})$$

How do we learn the parameters in the network?

$$p(y_1, \ldots, y_T) = \prod_{i=1}^{T} p(y_i | y_1, \ldots y_{i-1})$$

Maximum likelihood specifies loss function

How do we learn the parameters in the network?

$$p(y_1, \ldots, y_T) = \prod_{i=1}^{T} p(y_i | y_1, \ldots y_{i-1})$$

Maximum likelihood specifies loss function

Relation to structured models?

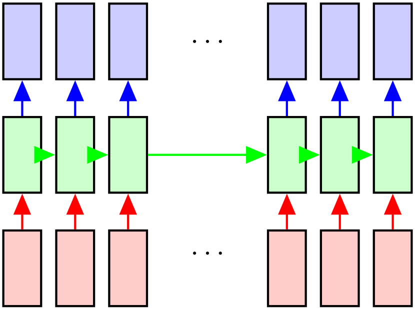
Training via gradient descent:

Training via gradient descent:

- How?
- What order?
- What information do we need to store?

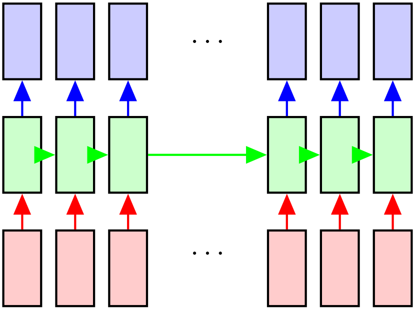
Training via gradient descent:

- How?
- What order?
- What information do we need to store?



Training via gradient descent:

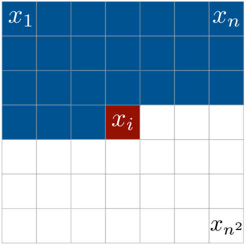
- How?
- What order?
- What information do we need to store?



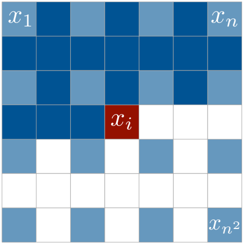
Backpropagation through time (BPTT)

Pixel Recurrent Neural Networks

PixelRNN model (Autoregressive model):



Context



Multi-scale context

Generative models overview:

Generative models overview:

- Variational Auto-encoders (VAEs):

Generative models overview:

- Variational Auto-encoders (VAEs):

- Generative Adversarial Nets (GANs):

Generative models overview:

- Variational Auto-encoders (VAEs):

- Generative Adversarial Nets (GANs):

- Autoregressive models (RNNs):

Generative models overview:

- Variational Auto-encoders (VAEs):
  - Pro:
  - Con:
- Generative Adversarial Nets (GANs):
  - Pro:
  - Con:
- Autoregressive models (RNNs):
  - Pro:
  - Con:

Generative models overview:

- Variational Auto-encoders (VAEs):
  - Pro: probabilistic graphical model interpretation
  - Con:
- Generative Adversarial Nets (GANs):
  - Pro:
  - Con:
- Autoregressive models (RNNs):
  - Pro:
  - Con:

Generative models overview:

- Variational Auto-encoders (VAEs):
  - Pro: probabilistic graphical model interpretation
  - Con: slightly blurry examples
- Generative Adversarial Nets (GANs):
  - Pro:
  - Con:
- Autoregressive models (RNNs):
  - Pro:
  - Con:

Generative models overview:

- Variational Auto-encoders (VAEs):
  - Pro: probabilistic graphical model interpretation
  - Con: slightly blurry examples

- Generative Adversarial Nets (GANs):
  - Pro: generate sharp images
  - Con:

- Autoregressive models (RNNs):
  - Pro:
  - Con:

Generative models overview:

- Variational Auto-encoders (VAEs):
  - Pro: probabilistic graphical model interpretation
  - Con: slightly blurry examples
- Generative Adversarial Nets (GANs):
  - Pro: generate sharp images
  - Con: difficult to optimize (unstable) [at least early variants]
- Autoregressive models (RNNs):
  - Pro:
  - Con:

Generative models overview:

- Variational Auto-encoders (VAEs):
  - Pro: probabilistic graphical model interpretation
  - Con: slightly blurry examples

- Generative Adversarial Nets (GANs):
  - Pro: generate sharp images
  - Con: difficult to optimize (unstable) [at least early variants]

- Autoregressive models (RNNs):
  - Pro: stable training & good likelihoods
  - Con:

Generative models overview:

- Variational Auto-encoders (VAEs):
  - Pro: probabilistic graphical model interpretation
  - Con: slightly blurry examples

- Generative Adversarial Nets (GANs):
  - Pro: generate sharp images
  - Con: difficult to optimize (unstable) [at least early variants]

- Autoregressive models (RNNs):
  - Pro: stable training & good likelihoods
  - Con: inefficient sampling & no low-dimensional codes

Generative models overview:

- Variational Auto-encoders (VAEs):
  - Pro: probabilistic graphical model interpretation
  - Con: slightly blurry examples

- Generative Adversarial Nets (GANs):
  - Pro: generate sharp images
  - Con: difficult to optimize (unstable) [at least early variants]

- Autoregressive models (RNNs):
  - Pro: stable training & good likelihoods
  - Con: inefficient sampling & no low-dimensional codes

Very active research area

Quiz:

Quiz:

- Describe the prediction process for an RNN?

Quiz:

- Describe the prediction process for an RNN?
- Describe the training process for RNNs?

Quiz:

- Describe the prediction process for an RNN?
- Describe the training process for RNNs?
- Contrast generative modeling techniques?

Important topics of this lecture

Important topics of this lecture

- Getting to know RNNs and its variants

Important topics of this lecture

- Getting to know RNNs and its variants
- Getting to know their use

Important topics of this lecture

- Getting to know RNNs and its variants
- Getting to know their use
- Contrasting RNNs to generative models

Important topics of this lecture

- Getting to know RNNs and its variants
- Getting to know their use
- Contrasting RNNs to generative models

Next up:

Reinforcement learning