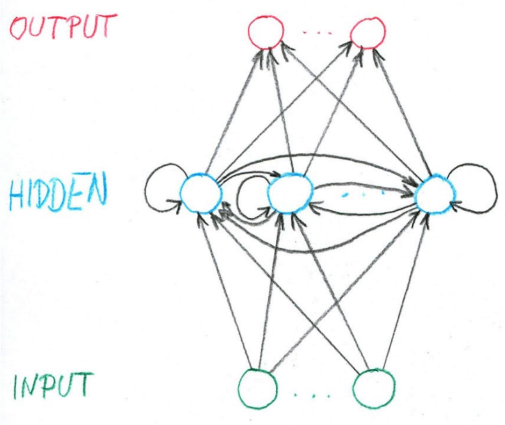
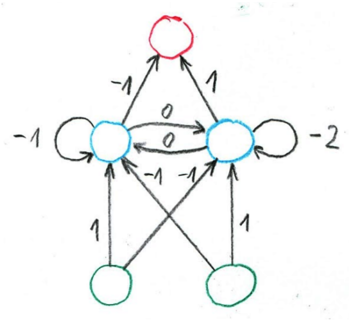
Recurrent Neural Networks - LSTM

RNN



- Input:
  $\vec{x} = (x_1, \ldots, x_M)$
- Hidden:
  $\vec{h} = (h_1, \ldots, h_H)$
- Output:
  $\vec{y} = (y_1, \ldots, y_N)$

RNN example



Activation function:

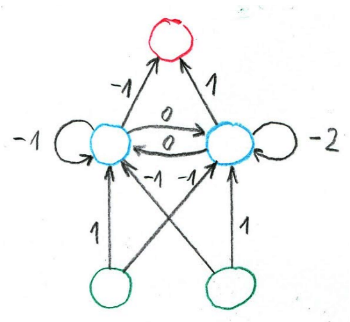
$$\sigma(\xi) = \begin{cases} 1 & \xi \geq 0 \\ 0 & \xi < 0 \end{cases}$$

y		1	0	1	
h	$(0,0)$	$(1,1)$	$(1,0)$	$(0,1)$	$\cdots$
x		$(0,0)$	$(1,0)$	$(1,1)$	

$-1 \cdot 1 + 0 \cdot 1 + 1 \cdot 1 + (-1) \cdot 0 = 0$

$0 \cdot 1 + (-2) \cdot 1 + (-1) \cdot 1 + 1 \cdot 0 = -2$

RNN example

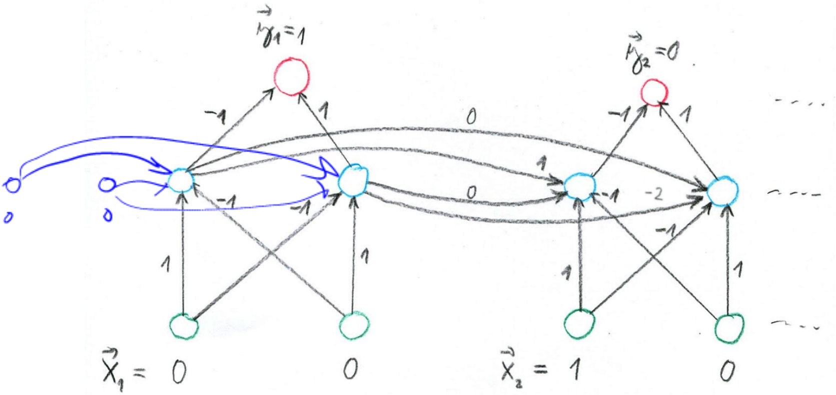


Activation function:

$$\sigma(\xi) = \begin{cases} 1 & \xi \geq 0 \\ 0 & \xi < 0 \end{cases}$$

y		$\vec{y}_1 = 1$	$\vec{y}_2 = 0$	$\vec{y}_3 = 1$	
h	$\vec{h}_0 = (0,0)$	$\vec{h}_1 = (1,1)$	$\vec{h}_2 = (1,0)$	$\vec{h}_3 = (0,1)$	$\cdots$
x		$\vec{x}_1 = (0,0)$	$\vec{x}_2 = (1,0)$	$\vec{x}_3 = (1,1)$	

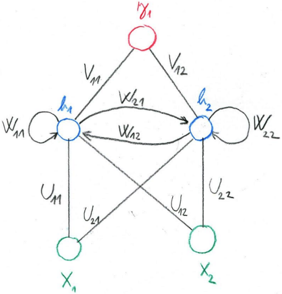
RNN example



y		$\vec{y}_1 = 1$	$\vec{y}_2 = 0$	$\vec{y}_3 = 1$	
h	$\vec{h}_0 = (0,0)$	$\vec{h}_1 = (1,1)$	$\vec{h}_2 = (1,0)$	$\vec{h}_3 = (0,1)$	$\cdots$
x		$\vec{x}_1 = (0,0)$	$\vec{x}_2 = (1,0)$	$\vec{x}_3 = (1,1)$	

RNN – formally

- M inputs: $\vec{x} = (x_1, \dots, x_M)$
- H hidden neurons: $\vec{h} = (h_1, \dots, h_H)$
- N output neurons: $\vec{y} = (y_1, \dots, y_N)$
- Weights:
  - $U_{kk'}$ from input $x_{k'}$ to hidden h_k
  - $W_{kk'}$ from hidden $h_{k'}$ to hidden h_k
  - $V_{kk'}$ from hidden $h_{k'}$ to output y_k



RNN – formally

- Input sequence: $\mathbf{x} = \vec{x}_1, \ldots, \vec{x}_T$

$$\vec{x}_t = (x_{t1}, \ldots, x_{tM})$$

RNN – formally

- Input sequence: $\mathbf{x} = \vec{x}_1, \ldots, \vec{x}_T$

$$\vec{x}_t = (x_{t1}, \ldots, x_{tM})$$

- Hidden sequence: $\mathbf{h} = \vec{h}_0, \vec{h}_1, \ldots, \vec{h}_T$

$$\vec{h}_t = (h_{t1}, \ldots, h_{tH})$$

We have $\vec{h}_0 = (0, \ldots, 0)$ and

$$\vec{h}_{tk} = \sigma\left(\sum_{k'=1}^{M} U_{kk'} x_{tk'} + \sum_{k'=1}^{H} W_{kk'} h_{(t-1)k'}\right)$$

RNN – formally

- Input sequence: $\mathbf{x} = \vec{x}_1, \ldots, \vec{x}_T$

  $$\vec{x}_t = (x_{t1}, \ldots, x_{tM})$$

- Hidden sequence: $\mathbf{h} = \vec{h}_0, \vec{h}_1, \ldots, \vec{h}_T$

  $$\vec{h}_t = (h_{t1}, \ldots, h_{tH})$$

  We have $\vec{h}_0 = (0, \ldots, 0)$ and

  $$\vec{h}_{tk} = \sigma\left(\sum_{k'=1}^{M} U_{kk'} x_{tk'} + \sum_{k'=1}^{H} W_{kk'} h_{(t-1)k'}\right)$$

- Output sequence: $\mathbf{y} = \vec{y}_1, \ldots, \vec{y}_T$

  $$\vec{y}_t = (y_{t1}, \ldots, y_{tN})$$

  where $y_{tk} = \sigma\left(\sum_{k'=1}^{H} V_{kk'} h_{tk'}\right)$.

RNN – in matrix form

- Input sequence: $\mathbf{x} = \vec{x}_1, \ldots, \vec{x}_T$

RNN – in matrix form

- Input sequence: $\mathbf{x} = \vec{x}_1, \ldots, \vec{x}_T$
- Hidden sequence: $\mathbf{h} = \vec{h}_0, \vec{h}_1, \ldots, \vec{h}_T$ where

$$\vec{h}_0 = (0, \ldots, 0)$$

and

$$\vec{h}_t = \sigma(U\vec{x}_t + W\vec{h}_{t-1})$$

RNN – in matrix form

- Input sequence: $\mathbf{x} = \vec{x}_1, \ldots, \vec{x}_T$
- Hidden sequence: $\mathbf{h} = \vec{h}_0, \vec{h}_1, \ldots, \vec{h}_T$ where

  $$\vec{h}_0 = (0, \ldots, 0)$$

  and

  $$\vec{h}_t = \sigma(U\vec{x}_t + W\vec{h}_{t-1})$$

- Output sequence: $\mathbf{y} = \vec{y}_1, \ldots, \vec{y}_T$ where

  $$y_t = \sigma(Vh_t)$$

RNN – Comments

- $\vec{h}_t$ is the memory of the network, captures what happened in all previous steps (with decaying quality).
- RNN **shares weights** U, V, W along the sequence.
  Note the similarity to convolutional networks where the weights were shared spatially over images, here they are shared temporally over sequences.
- RNN can deal with **sequences of variable length.**
  Compare with MLP which accepts only fixed-dimension vectors on input.

RNN – training

Training set

$$\mathcal{T} = \left\{(\mathbf{x}_1, \mathbf{d}_1), \ldots, (\mathbf{x}_p, \mathbf{y}_p)\right\}$$

here

- each $\mathbf{x}_\ell = \vec{x}_{\ell 1}, \ldots, \vec{x}_{\ell T_\ell}$ is an input sequence,
- each $\mathbf{d}_\ell = \vec{d}_{\ell 1}, \ldots, \vec{d}_{\ell T_\ell}$ is an expected output sequence.

Here each $\vec{x}_{\ell t} = (x_{\ell t 1}, \ldots, x_{\ell t M})$ is an input vector and each $\vec{d}_{\ell t} = (d_{\ell t 1}, \ldots, d_{\ell t N})$ is an expected output vector.

Error function

In what follows I will consider a training set with a **single element** $(\mathbf{x}, \mathbf{d})$. I.e. drop the index ℓ and have

- $\mathbf{x} = \vec{x}_1, \ldots, \vec{x}_T$ where $\vec{x}_t = (x_{t1}, \ldots, x_{tM})$
- $\mathbf{d} = \vec{d}_1, \ldots, \vec{d}_T$ where $\vec{d}_t = (d_{t1}, \ldots, d_{tN})$

The squared error of $(\mathbf{x}, \mathbf{d})$ is defined by

$$E_{(\mathbf{x},\mathbf{d})} = \sum_{t=1}^{T} \sum_{k=1}^{N} \frac{1}{2}(y_{tk} - d_{tk})^2$$

Recall that we have a sequence of network outputs $\mathbf{y} = \vec{y}_1, \ldots, \vec{y}_T$ and thus y_{tk} is the k-th component of $\vec{y}_t$

Gradient descent (single training example)

Consider a single training example $(\mathbf{x}, \mathbf{d})$.

The algorithm computes a sequence of weight matrices as follows:

Gradient descent (single training example)

Consider a single training example $(\mathbf{x}, \mathbf{d})$.

The algorithm computes a sequence of weight matrices as follows:

- Initialize all weights randomly close to 0.

Gradient descent (single training example)

Consider a single training example $(\mathbf{x}, \mathbf{d})$.

The algorithm computes a sequence of weight matrices as follows:

- Initialize all weights randomly close to 0.
- In the step $\ell + 1$ (here $\ell = 0, 1, 2, \ldots$) compute "new" weights $U^{(\ell+1)}, V^{(\ell+1)}, W^{(\ell+1)}$ from the "old" weights $U^{(\ell)}, V^{(\ell)}, W^{(\ell)}$ as follows:

$$U_{kk'}^{(\ell+1)} = U_{kk'}^{(\ell)} - \varepsilon(\ell) \cdot \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta U_{kk'}}$$

$$V_{kk'}^{(\ell+1)} = V_{kk'}^{(\ell)} - \varepsilon(\ell) \cdot \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta V_{kk'}}$$

$$W_{kk'}^{(\ell+1)} = W_{kk'}^{(\ell)} - \varepsilon(\ell) \cdot \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta W_{kk'}}$$

Gradient descent (single training example)

Consider a single training example $(\mathbf{x}, \mathbf{d})$.

The algorithm computes a sequence of weight matrices as follows:

- Initialize all weights randomly close to 0.
- In the step $\ell + 1$ (here $\ell = 0, 1, 2, \ldots$) compute "new" weights $U^{(\ell+1)}, V^{(\ell+1)}, W^{(\ell+1)}$ from the "old" weights $U^{(\ell)}, V^{(\ell)}, W^{(\ell)}$ as follows:

$$U_{kk'}^{(\ell+1)} = U_{kk'}^{(\ell)} - \varepsilon(\ell) \cdot \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta U_{kk'}}$$

$$V_{kk'}^{(\ell+1)} = V_{kk'}^{(\ell)} - \varepsilon(\ell) \cdot \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta V_{kk'}}$$

$$W_{kk'}^{(\ell+1)} = W_{kk'}^{(\ell)} - \varepsilon(\ell) \cdot \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta W_{kk'}}$$

The above is THE learning algorithm that modifies weights!

Backpropagation

Computes the derivatives of E, no weights are modified!

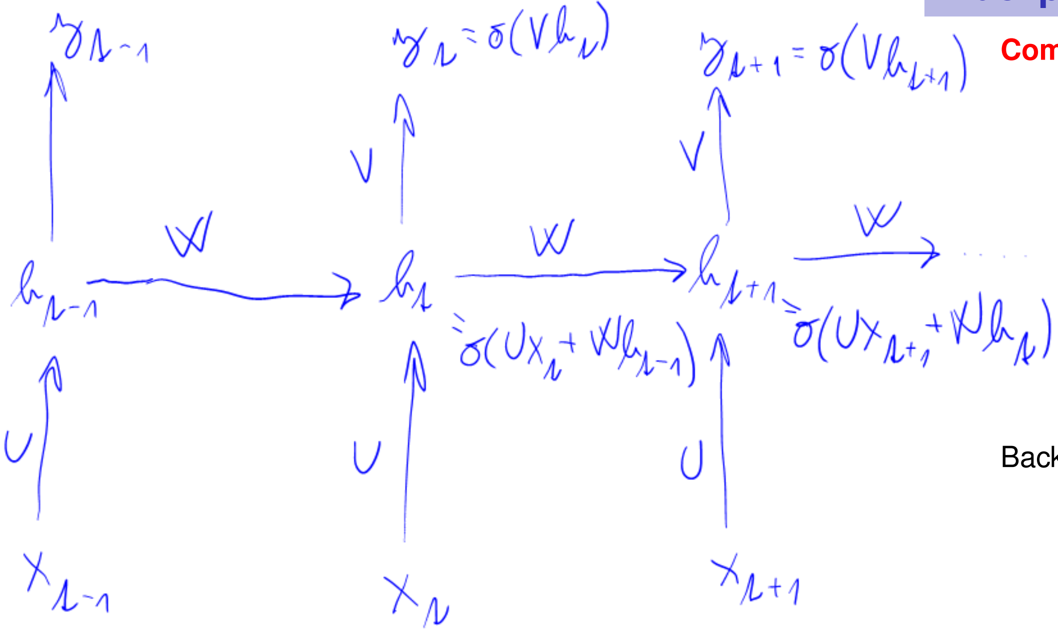
Backpropagation

Computes the derivatives of E, no weights are modified!

$$\frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta U_{kk'}} = \sum_{t=1}^{T} \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta h_{tk}} \cdot \sigma' \cdot x_{tk'} \qquad k' = 1, \ldots, M$$

$$\frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta V_{kk'}} = \sum_{t=1}^{T} \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta y_{tk}} \cdot \sigma' \cdot h_{tk'} \qquad k' = 1, \ldots, H$$

$$\frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta W_{kk'}} = \sum_{t=1}^{T} \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta h_{tk}} \cdot \sigma' \cdot h_{(t-1)k'} \qquad k' = 1, \ldots, H$$



Computes the derivatives of E, no weights are modified!

$$\frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta U_{kk'}} = \sum_{t=1}^{T} \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta h_{tk}} \cdot \sigma' \cdot x_{tk'} \qquad k' = 1,\ldots,M$$

$$\frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta V_{kk'}} = \sum_{t=1}^{T} \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta y_{tk}} \cdot \sigma' \cdot h_{tk'} \qquad k' = 1,\ldots,H$$

$$\frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta W_{kk'}} = \sum_{t=1}^{T} \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta h_{tk}} \cdot \sigma' \cdot h_{(t-1)k'} \qquad k' = 1,\ldots,H$$

Backpropagation:

$$\frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta y_{tk}} = y_{tk} - d_{tk} \qquad \text{(assuming squared error)}$$

$$\frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta h_{tk}} = \sum_{k'=1}^{N} \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta y_{tk'}} \cdot \sigma' \cdot V_{k'k} + \sum_{k'=1}^{H} \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta h_{(t+1)k'}} \cdot \sigma' \cdot W_{k'k}$$

Long-term dependencies

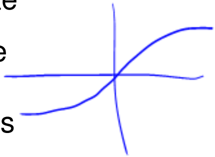
$$\frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta h_{tk}} = \sum_{k'=1}^{N} \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta y_{tk'}} \cdot \sigma' \cdot V_{k'k} + \sum_{k'=1}^{H} \frac{\delta E_{(\mathbf{x},\mathbf{d})}}{\delta h_{(t+1)k'}} \cdot \sigma' \cdot W_{k'k}$$

- Unless $\sum_{k'=1}^{H} \sigma' \cdot W_{k'k} \approx 1$, the gradient either vanishes, or explodes.
- For a large T (long-term dependency), the gradient "deeper" in the past tends to be too small (large).
- A solution: LSTM

LSTM

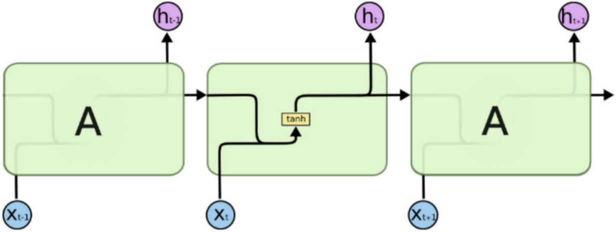
$(2,5) \circ (3,1) = (2 \cdot 3, 5 \cdot 1) = (6,5)$

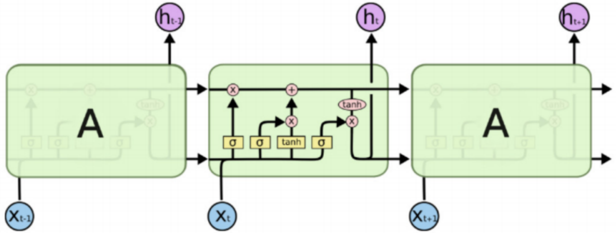
$\vec{h}_t = \vec{o}_t \circ \sigma_h(\vec{C}_t)$	output
$\vec{C}_t = \vec{f}_t \circ \vec{C}_{t-1} + \vec{i}_t \circ \tilde{C}_t$	memory
$\tilde{C}_t = \sigma_h(W_C \cdot \vec{h}_{t-1} + U_C \cdot \vec{x}_t)$	new memory contents
$\vec{o}_t = \sigma_g(W_o \cdot \vec{h}_{t-1} + U_o \cdot \vec{x}_t)$	output gate
$\vec{f}_t = \sigma_g(W_f \cdot \vec{h}_{t-1} + U_f \cdot \vec{x}_t)$	forget gate
$\vec{i}_t = \sigma_g(W_i \cdot \vec{h}_{t-1} + U_i \cdot \vec{x}_t)$	input gate



- $\circ$ is the component-wise product of vectors
- $\cdot$ is the matrix-vector product
- σ_h hyperbolic tangents (applied component-wise)
- σ_g logistic sigmoid (aplied component-wise)

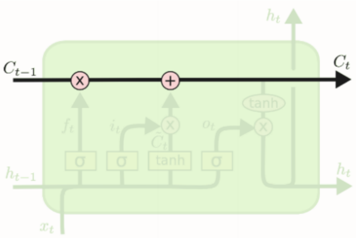
RNN vs LSTM





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

LSTM



$$\vec{h}_t = \vec{o}_t \circ \sigma_h(\vec{C}_t)$$

$$\Rightarrow \vec{C}_t = \vec{f}_t \circ \vec{C}_{t-1} + \vec{i}_t \circ \tilde{C}_t$$

$$\tilde{C}_t = \sigma_h(W_C \cdot \vec{h}_{t-1} + U_C \cdot \vec{x}_t)$$

$$\vec{o}_t = \sigma_g(W_o \cdot \vec{h}_{t-1} + U_o \cdot \vec{x}_t)$$

$$\vec{f}_t = \sigma_g(W_f \cdot \vec{h}_{t-1} + U_f \cdot \vec{x}_t)$$

$$\vec{i}_t = \sigma_g(W_i \cdot \vec{h}_{t-1} + U_i \cdot \vec{x}_t)$$

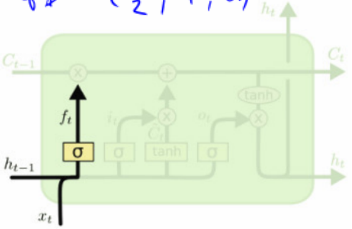
Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM

$5/2 \quad 6 \quad 0.005 \ldots$

$C_{t-1} = (5, 6, 5)$

$f_t = (\frac{1}{2}, 1, 0.000001)$



$$\vec{h}_t = \vec{o}_t \circ \sigma_h(\vec{C}_t)$$

$$\vec{C}_t = \vec{f}_t \circ \vec{C}_{t-1} + \vec{i}_t \circ \tilde{C}_t$$

$$\tilde{C}_t = \sigma_h(W_C \cdot \vec{h}_{t-1} + U_C \cdot \vec{x}_t)$$

$$\vec{o}_t = \sigma_g(W_o \cdot \vec{h}_{t-1} + U_o \cdot \vec{x}_t)$$

$$\Rightarrow \vec{f}_t = \sigma_g(W_f \cdot \vec{h}_{t-1} + U_f \cdot \vec{x}_t)$$

$$\vec{i}_t = \sigma_g(W_i \cdot \vec{h}_{t-1} + U_i \cdot \vec{x}_t)$$

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

LSTM

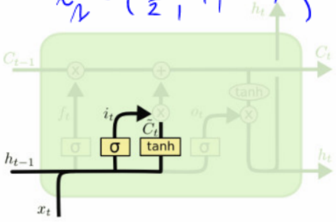
$\vec{f}_n \circ \vec{C}_{n-1} = (5, 6, 4)$

$\tilde{C}_n = (2, 4, 6)$

$\vec{i}_n = (\frac{1}{2}, 1, 0)$

$(1, 4, 0)$

$(5, 6, 4) + (1, 4, 0) = (6, 10, 4) = \vec{C}_n$



$$\vec{h}_t = \vec{o}_t \circ \sigma_h(\vec{C}_t)$$

$$\vec{C}_t = \vec{f}_t \circ \vec{C}_{t-1} + \vec{i}_t \circ \tilde{C}_t$$

$$\Rightarrow \tilde{C}_t = \sigma_h(W_C \cdot \vec{h}_{t-1} + U_C \cdot \vec{x}_t)$$

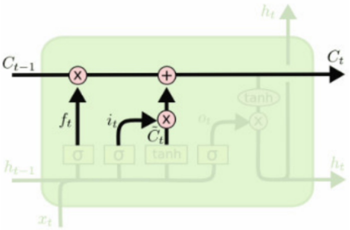
$$\vec{o}_t = \sigma_g(W_o \cdot \vec{h}_{t-1} + U_o \cdot \vec{x}_t)$$

$$\vec{f}_t = \sigma_g(W_f \cdot \vec{h}_{t-1} + U_f \cdot \vec{x}_t)$$

$$\Rightarrow \vec{i}_t = \sigma_g(W_i \cdot \vec{h}_{t-1} + U_i \cdot \vec{x}_t)$$

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

LSTM



$$\vec{h}_t = \vec{o}_t \circ \sigma_h(\vec{C}_t)$$

$$\Rightarrow \vec{C}_t = \vec{f}_t \circ \vec{C}_{t-1} + \vec{i}_t \circ \tilde{C}_t$$

$$\Rightarrow \tilde{C}_t = \sigma_h(W_C \cdot \vec{h}_{t-1} + U_C \cdot \vec{x}_t)$$

$$\vec{o}_t = \sigma_g(W_o \cdot \vec{h}_{t-1} + U_o \cdot \vec{x}_t)$$

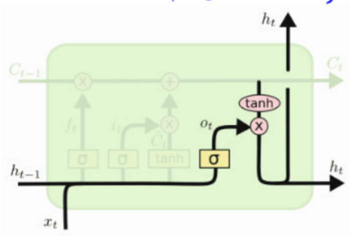
$$\Rightarrow \vec{f}_t = \sigma_g(W_f \cdot \vec{h}_{t-1} + U_f \cdot \vec{x}_t)$$

$$\Rightarrow \vec{i}_t = \sigma_g(W_i \cdot \vec{h}_{t-1} + U_i \cdot \vec{x}_t)$$

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

LSTM

$$\vec{C}_t = (1, 2, 3)$$
$$\vec{o}_t = (0, \tfrac{1}{2}, 1)$$
$$\Big\} = (0, 1, 3) = \vec{h}_t$$



$$\Rightarrow \vec{h}_t = \vec{o}_t \circ \sigma_h(\vec{C}_t)$$
$$\vec{C}_t = \vec{f}_t \circ \vec{C}_{t-1} + \vec{i}_t \circ \tilde{C}_t$$
$$\tilde{C}_t = \sigma_h(W_C \cdot \vec{h}_{t-1} + U_C \cdot \vec{x}_t)$$

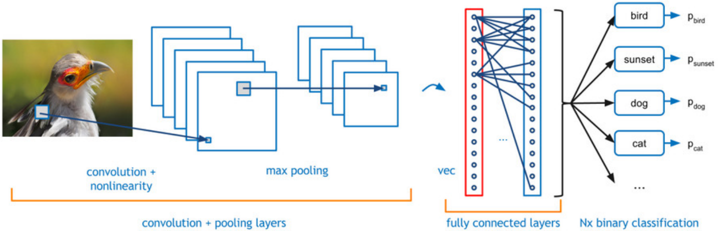
$$\Rightarrow \vec{o}_t = \sigma_g(W_o \cdot \vec{h}_{t-1} + U_o \cdot \vec{x}_t)$$
$$\vec{f}_t = \sigma_g(W_f \cdot \vec{h}_{t-1} + U_f \cdot \vec{x}_t)$$
$$\vec{i}_t = \sigma_g(W_i \cdot \vec{h}_{t-1} + U_i \cdot \vec{x}_t)$$

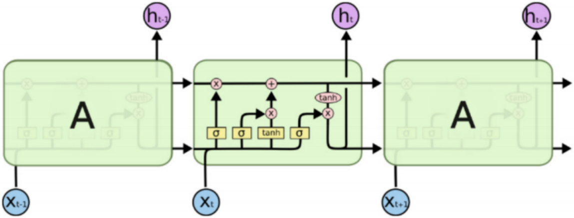
Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM – summary

- LSTM (almost) solves the vanishing gradient problem w.r.t. the "internal" state of the network.
- Learns to control its own memory (via forget gate).
- Revolution in machine translation and text processing.

Convolutions & LSTM in action – cancer research





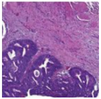
Colorectal cancer outcome prediction

The problem: Predict 5-year survival probability from an image of a small region of tumour tissue (1 mm diameter).

Colorectal cancer outcome prediction

The problem: Predict 5-year survival probability from an image of a small region of tumour tissue (1 mm diameter).

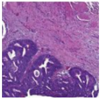
Input: Digitized haematoxylin-eosin-stained tumour tissue microarray samples.
Output: Estimated survival probability.



Colorectal cancer outcome prediction

The problem: Predict 5-year survival probability from an image of a small region of tumour tissue (1 mm diameter).

Input: Digitized haematoxylin-eosin-stained tumour tissue microarray samples.
Output: Estimated survival probability.



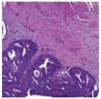
Data:

- Training set: 420 patients of Helsinki University Centre Hospital, diagnosed with colorectal cancer, underwent primary surgery.
- Test set: 182 patients
- Follow-up time and outcome known for each patient.

Colorectal cancer outcome prediction

The problem: Predict 5-year survival probability from an image of a small region of tumour tissue (1 mm diameter).

Input: Digitized haematoxylin-eosin-stained tumour tissue microarray samples.
Output: Estimated survival probability.



Data:

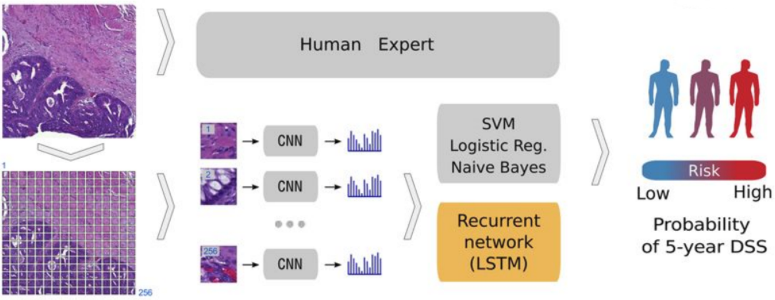
- Training set: 420 patients of Helsinki University Centre Hospital, diagnosed with colorectal cancer, underwent primary surgery.
- Test set: 182 patients
- Follow-up time and outcome known for each patient.

Human expert comparison:

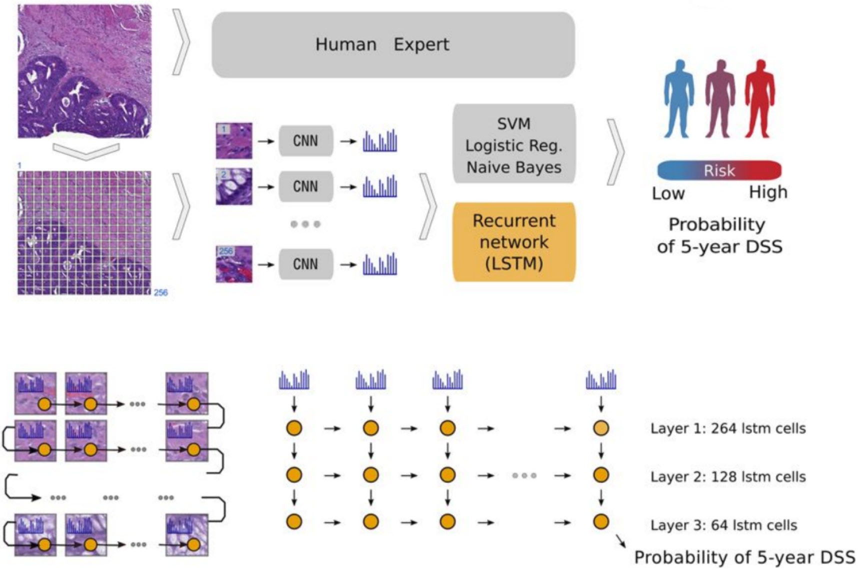
- Histological grade assessed at the time of diagnosis.
- Visual Risk Score: Three pathologists classified to high/low-risk categories (by majority vote).

Source: D. Bychkov et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific Reports, Nature, 2018.

Colorectal cancer outcome prediction



Colorectal cancer outcome prediction



Data & workflow

- Input images: 3500 px $\times$ 3500 px
  - Cut into tiles: 224 px $\times$ 224 px $\Rightarrow$ 256 tiles
- Each tile pased to a convolutional network (CNN)
  - Ouptut of CNN: 4096 dimensional vector.
- A "string" of 256 vectors (each of the dimension 4096) pased into a LSTM.
- LSTM outputs the probability of 5-year survival.

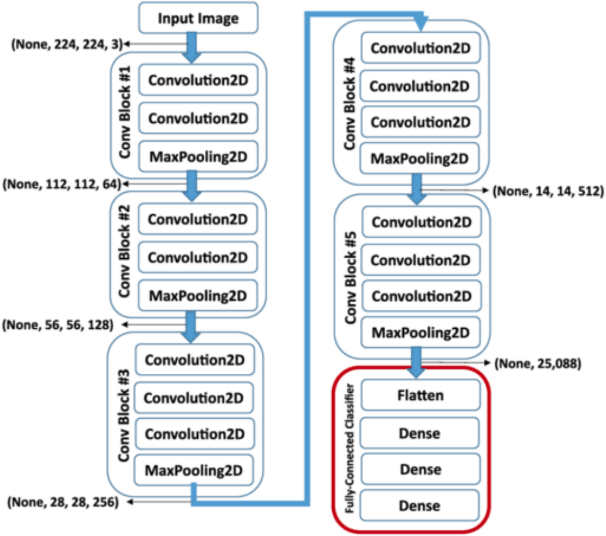
Data & workflow

- Input images: 3500 px $\times$ 3500 px
  - Cut into tiles: 224 px $\times$ 224 px $\Rightarrow$ 256 tiles
- Each tile pased to a convolutional network (CNN)
  - Ouptut of CNN: 4096 dimensional vector.
- A "string" of 256 vectors (each of the dimension 4096) pased into a LSTM.
- LSTM outputs the probability of 5-year survival.

The authors also tried to substitute the LSTM on top of CNN with

- logistic regression
- naive Bayes
- support vector machines

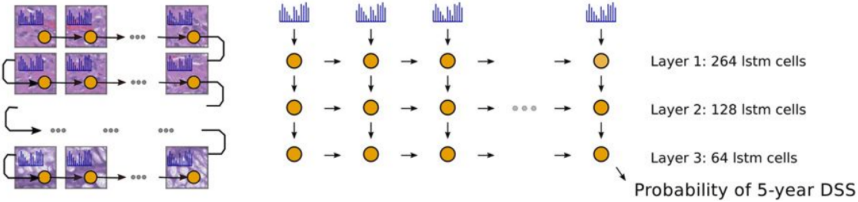
CNN architecture – VGG-16



(Pre)trained on ImageNet (cats, dogs, chairs, etc.)

LSTM architecture

- LSTM has three layers (264, 128, 64 cells)

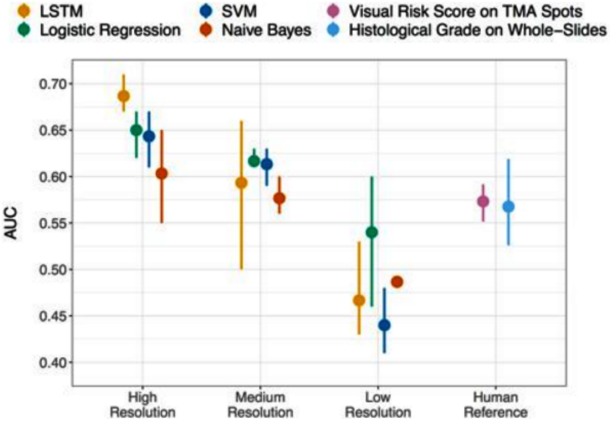


LSTM – training

- L1 regularization (0.005) at each hidden layer of LSTM
  i.e. 0.005 times the sum of absolute values of weights added to the error
- L2 regularization (0.005) at each hidden layer of LSTM
  i.e. 0.005 times the sum of squared values of weights added to the error
- Dropout 5% at the input and the last hidden layers of LSTM

- Datasets:
  - Training: 220 samples,
  - Validation 60 samples,
  - Test 140 samples.

Colorectal cancer outcome prediction



Source: D. Bychkov et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific Reports, Nature, 2018.