Deep RNNs & Applications to Biology

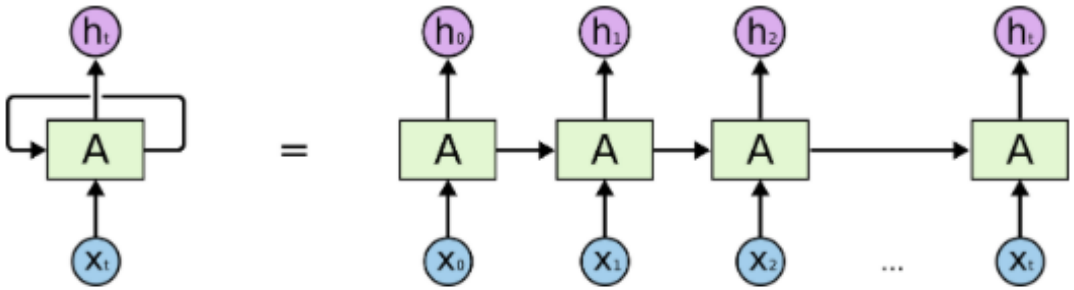
Motivation

- Vanilla (e.g., fully connected) neural networks
  - great for discovering complex relationships in data and making predictions using them
- Convolutional neural network
  - great for discovering complex *spacial* relationships in data (edges, etc.) and making predictions using them
- Autoencoders
  - great for compressing data and pretraining

... but what about *sequential* data?

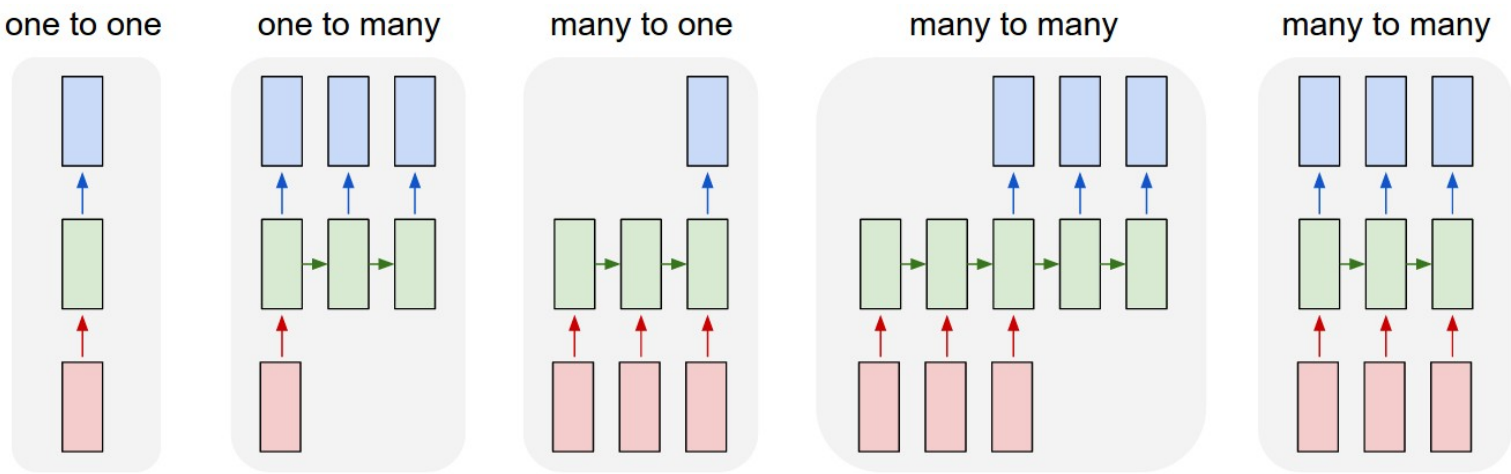
... how about sequential data *of a variable length*?

Basic Structure



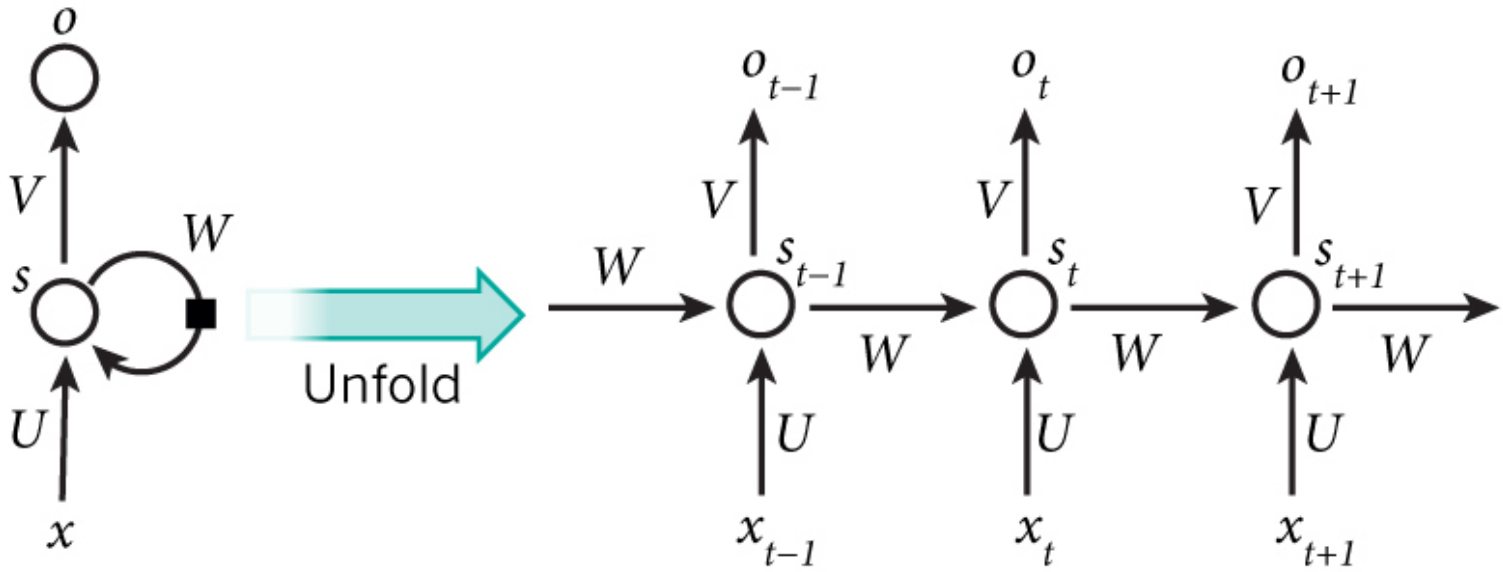
Source: *Cameron Godbout*

Various Forms



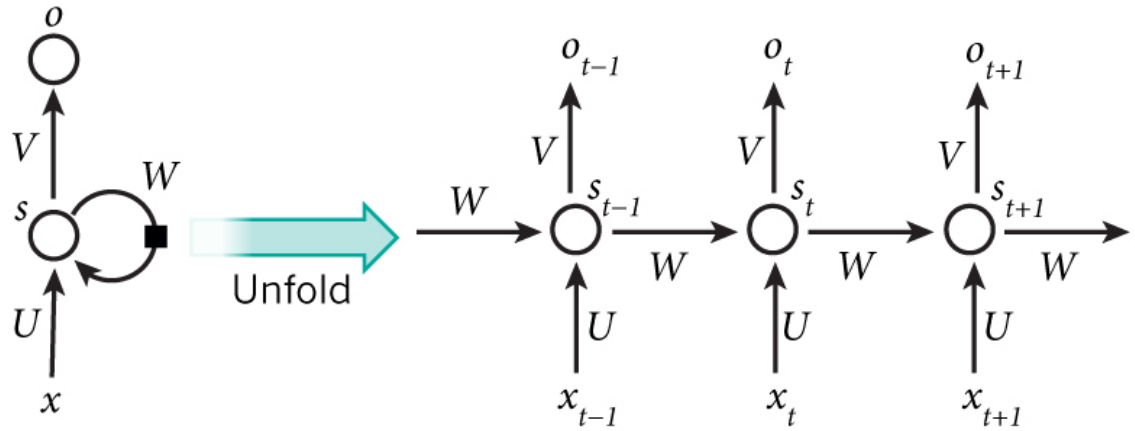
Source: *Andrej Karpathy*

Unfolded Structure



Source: *Cameron Godbout*

More Details



Source: *Nature via WildML*

- $\mathbf{x}_t$: input at t
- $\mathbf{o}_t$: output at t
- $\mathbf{s}_t$: "state" at t
- $\mathbf{W}$: transition matrix
- $\mathbf{U}$: input matrix
- $\mathbf{V}$: output matrix

- $\mathbf{s}_t = f_s(\mathbf{x}_t, \mathbf{s}_{t-1}) = \phi_s(\mathbf{W}^T \mathbf{s}_{t-1} + \mathbf{U}^T \mathbf{x}_t)$
- $\mathbf{y}_t = f_o(\mathbf{s}_t, \mathbf{x}_t) = \phi_o(\mathbf{V}^T \mathbf{h}_t)$
- ϕ_x and ϕ_o: element-wise nonlinear functions

Training

- Can construct cost function given training sequences
- Stochastic gradient descent
- Backpropagation through time (BPTT)
  - Backpropagation (chain rule, etc.) on unrolled model
- NoBackTrack (Ollivier *et al.*, 2015)

NoBackTrack

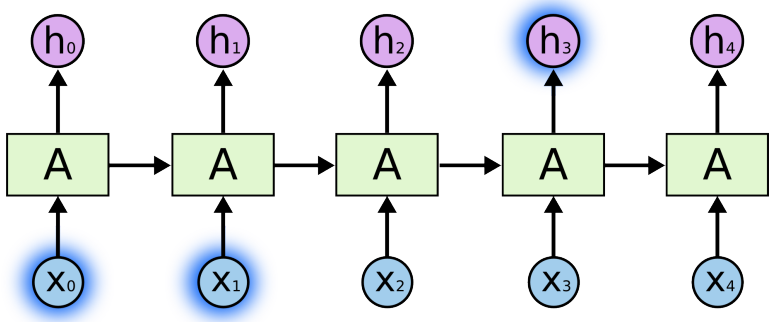
"Training recurrent networks online without backtracking" (Ollivier *et al.*, 2015)

- Alternative to BPTT
- Maintains only unbiased estimate of full gradient
- "Kalman-like filter"
- Potentially better than truncated BPTT in some situations

Challenges in Training

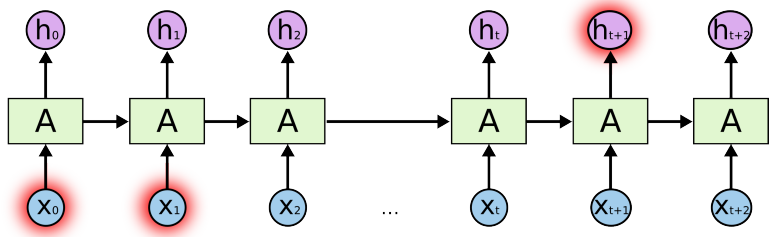
"On the difficulty of training recurrent neural networks" (Pascanu *et al.*, 2013), building on Bengio *et al.* (1994)

- exploding gradients $\rightarrow$ unpredictability
- vanishing gradients $\rightarrow$ lack of ability to "remember" earlier inputs



Source: *Christopher Olah*

vs



Source: *Christopher Olah*

Potential Solutions

- L1/L2 penalties
- Teacher forcing
- **LSTM** (more on this later)
- Structural damping
- Echo State Networks
- Scaling down the gradient
- Vanishing gradient regularization

Motivation

"How to Construct Deep Recurrent Neural Networks"
(Pascanu *et al.*, 2014)

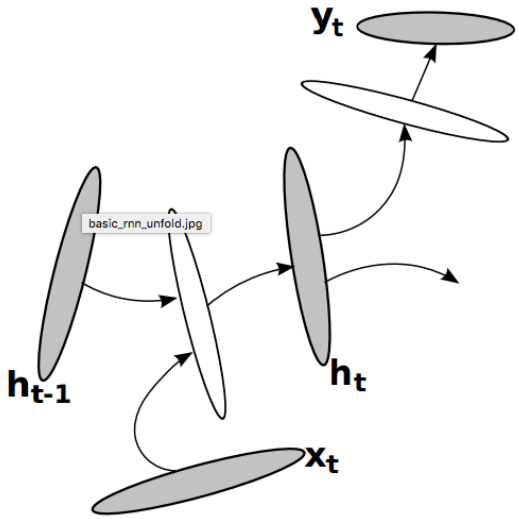
- Why would we want to make RNNs deep?
  - Simple: depth helps in other sorts of networks
- What does depth in the context of RNNs even mean?
  - Already deep?

1. Deep Input-to-Hidden Function

- Akin to preprocessing inputs (feature extraction)
- Probably would help, but not used in experiments of Pascanu *et al.* (2014)
- Need not be trained at same time
  - Chen and Deng (2013): speech recognition

2. Deep Hidden-to-Output Function

- "This allows the hidden state of the model to be more compact and may result in the model being able to summarize the history of previous inputs more efficiently" (Pascanu *et al.*, 2014).
- Here, standard feedforward layers

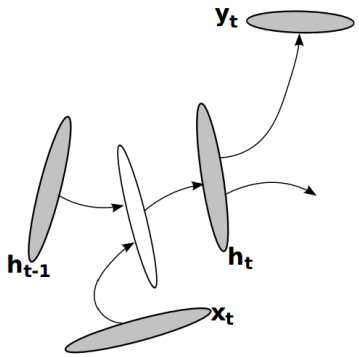


(c) DOT-RNN

Source: *Pascanu et al. (2014)*

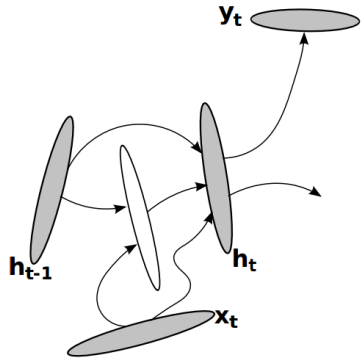
- $\mathbf{y}_t = f_o(\mathbf{h}_t) = \phi_o(\mathbf{V}_L^T \phi_{L-1}(\mathbf{V}_{L-1}^T \phi_{L-2}(\cdots \phi_1(\mathbf{V}_1^T \mathbf{h}_{t-1} + \mathbf{U}^T \mathbf{x}_t))))$
  - "ϕ_l and $\mathbf{V}_l$ are the element-wise nonlinear function and weight matrix for the l-th layer" (Pascanu *et al.*, 2014)

3. Deep Hidden-to-Hidden Transition



(b) DT-RNN

Source: *Pascanu et al. (2014)*



(b*) DT(S)-RNN

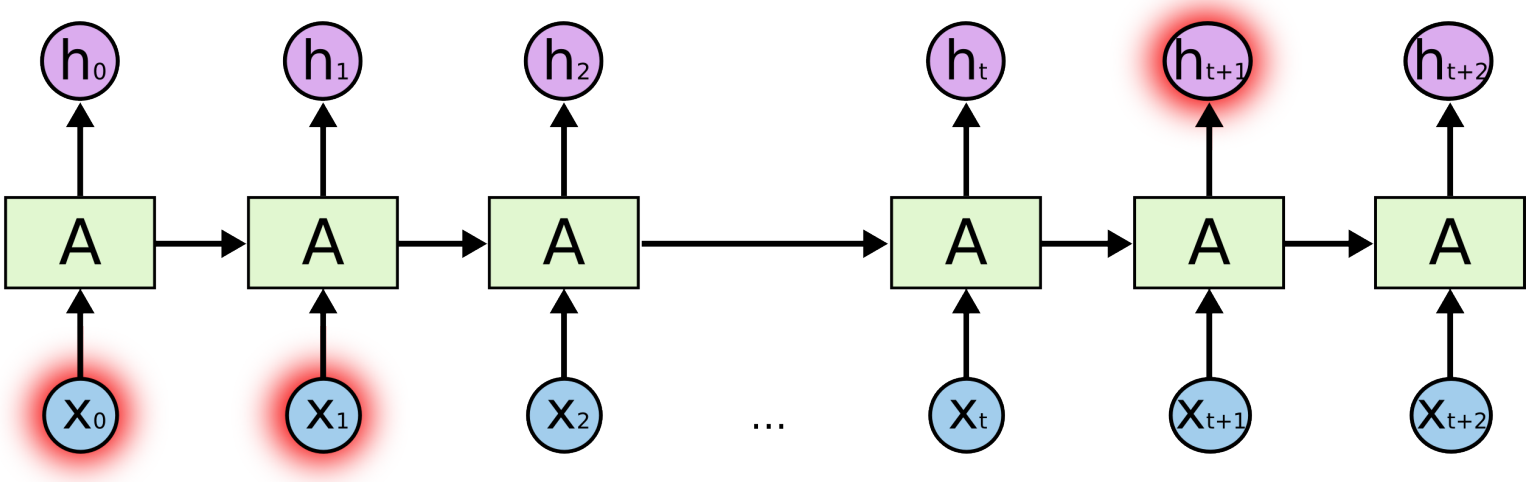
Source: *Pascanu et al. (2014)*

- Replace "shallow operation" with "highly nonlinear transition...modeled by an MLP" (Pascanu *et al.*, 2014)
- Issue: training (vanishing gradients)
  - Solution: shortcut connections
- $\mathbf{h}_t = f_h(\mathbf{x}_t, \mathbf{h}_{t-1}) = \phi_h(\mathbf{W}_L^T \phi_{L-1}(\mathbf{W}_{L-1}^T \phi_{L-2}(\cdots \phi_1(\mathbf{W}_1^T \mathbf{h}_{t-1} + \mathbf{U}^T \mathbf{x}_t))))$
  - "ϕ_l and $\mathbf{W}_l$ are the element-wise nonlinear function and weight matrix for the l-th layer" (Pascanu *et al.*, 2014)

Motivation

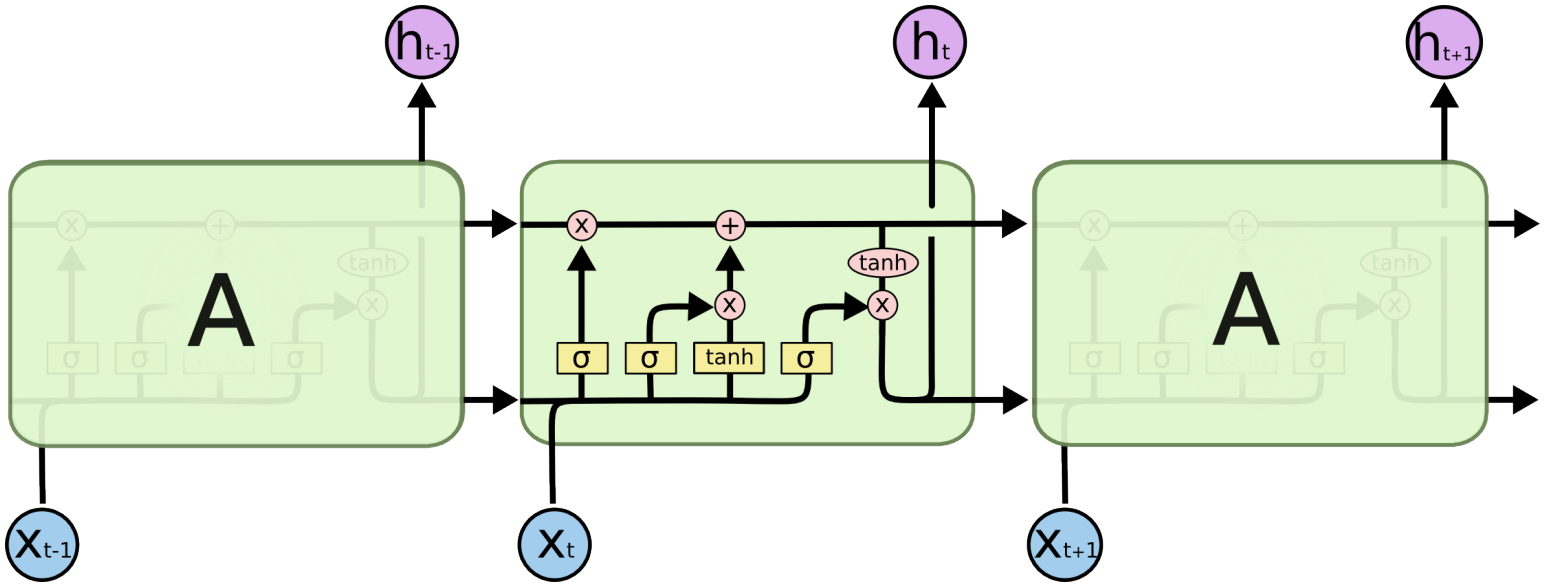
"Long Short-Term Memory" (Hochreiter and Schmidhuber, 1997)

- Recall difficulties discussed earlier: exploding and vanishing gradients
- Can we modify cell structure to avoid these problems?



Source: *Christopher Olah*

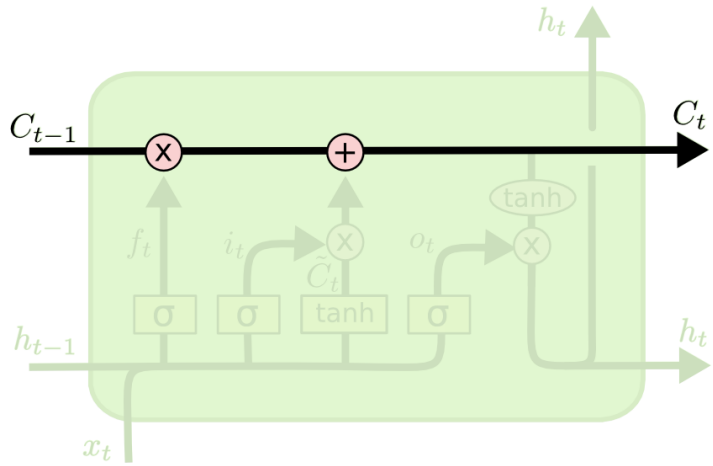
Better



Source: *Christopher Olah*

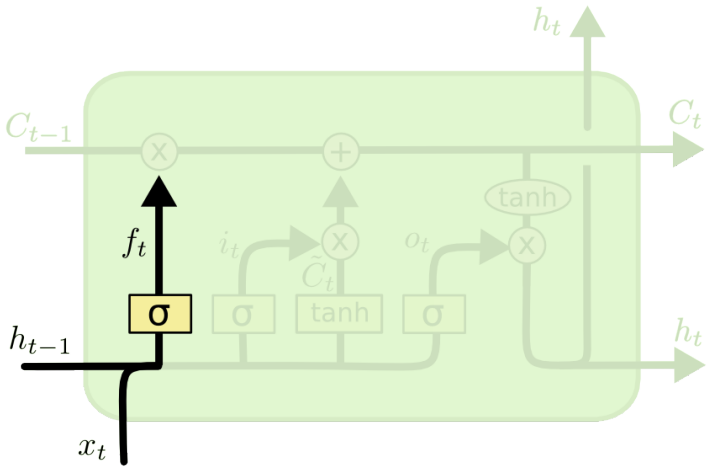
Big Picture

Traditional RNN has cell state that depends only on previous state and input.
LSTM also has cell state that it passes along, but it doesn't need to change every time, and in addition its previous output is passed on.



Source: *Christopher Olah*

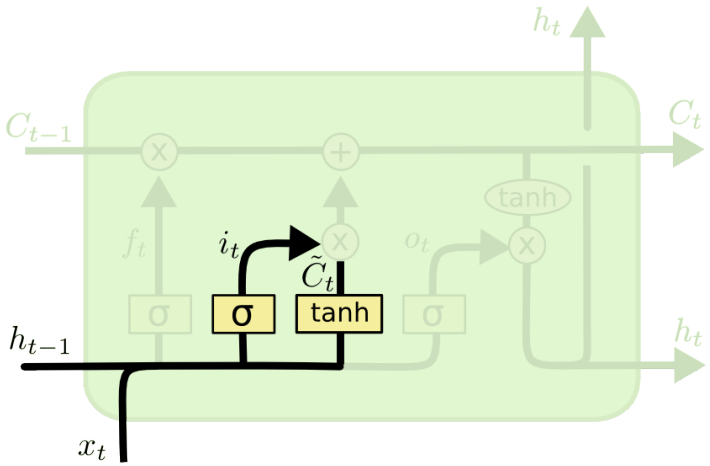
Forget Gate



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Source: *Christopher Olah*

Input Gates

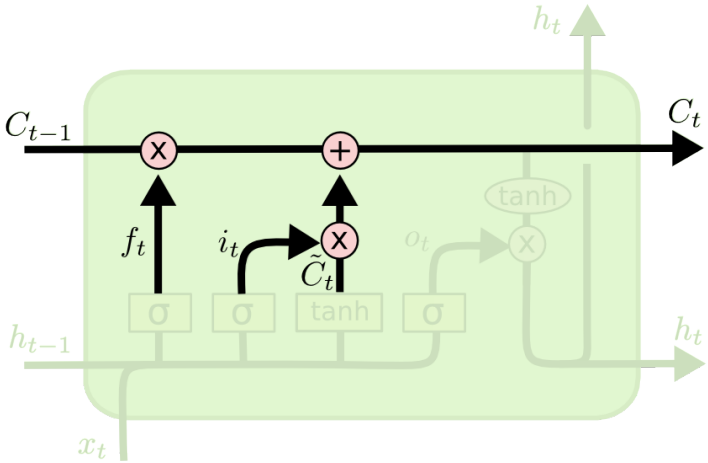


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Source: *Christopher Olah*

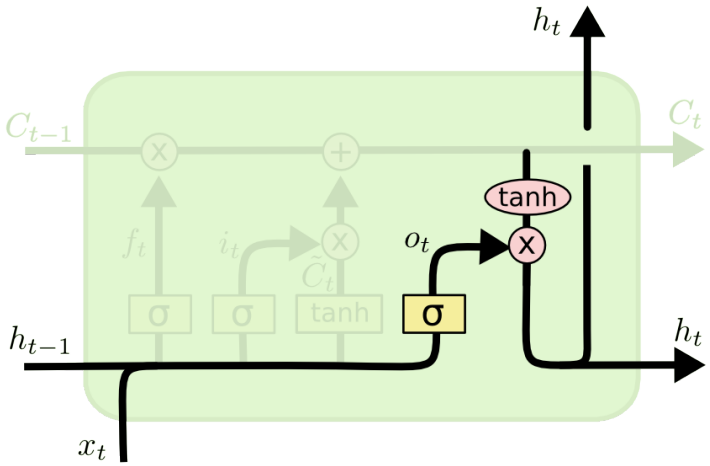
Combining



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Source: *Christopher Olah*

Output Gates



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

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

Source: *Christopher Olah*

Speech Recognition

"Speech Recognition with Deep Recurrent Neural Networks" (Graves *et al.*, 2013)

- TIMIT corpus for phoneme recognition
- Deep LSTM
  - Peephole
  - Bidirectional
- Training
  - Techniques for particular task
  - Regularization: early stopping, weight noise

Emotion Detection

"Continuous Emotion Detection Using EEG Signals and Facial Expressions" (Soleymani *et al.*, 2014)

- Interested in whether it possible to detect valence (positive emotion) from a viewer from both EEG signals and facial data
- MAHNOB-HCI database
  - Screened clips for participants
  - Annotated emotional responses
  - Recorded EEG signals and facial tracker
- Used several models, including LSTM
- Results
  - LSTM best predictor
  - Correlation between EEG and facial expressions
  - Detection with expressionless face somewhat successful
  - EEG signals don't much underperform facial expressions (surprising because facial expressions are ground truth)

Emotion Detection (continued)

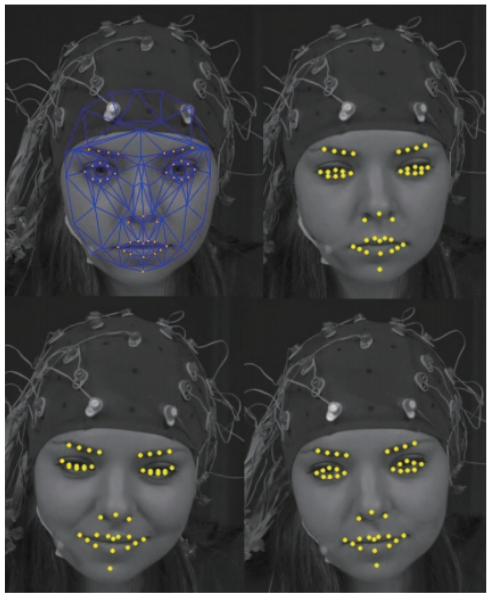


Fig. 1. Examples of the recorded camera view including tracked facial points. The top left image shows the active appearance model that is fit to the face.

Source: *Soleymani et al. (2014)*

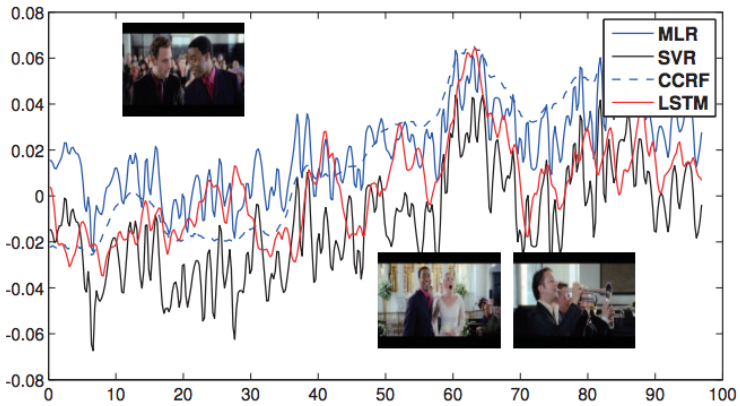


Fig. 3. The average valence curve, emotional pleasantness profile, resulted from the EEG signals of 10 participants who did not show any facial expressions while watching a scene from Love Actually. The joyful moments and a highlights are still detectable from the curve and its trend.

Source: *Soleymani et al. (2014)*

Epileptic Seizures Prediction

"Recurrent neural network based prediction of epileptic seizures in intra- and extracranial EEG" (Petrosian *et al.*, 2000)

- Interested not in seizure detection, but prediction
- Intracranial and extracranial detection
- Testing of variations on RNN
- Results
  - Better results when trained on preprocessed wavelet data
  - Intracranial data produces more stable results than extracranial data
  - Not totally successful, but promising results

Protein Secondary Structure Prediction

"Protein Secondary Structure Prediction with Long Short Term Memory Networks" (Sønderby and Winther, 2015)

- Updating of Baldi *et al.* (1999): LSTM, larger data, GPU
- Architecture
  - Bidirection LSTM
  - Feed-forward network between states (deep transition)
  - Feed-forward network combining bidirectional outputs (deep output function)
- Amino acid sequence $\rightarrow$ 8 secondary structure classes

Table 1. Description of protein secondary structure classes and class frequencies in the dataset. In the litterature the 8-class DSSP output is typically mapped to 3 classes. The 8 to 3 class mappings are included for reference.

8-class (Q8)	3 class (Q3)	Frequency	Name
H	H	0.34535	α-helix
E	E	0.21781	β-strand
L	C	0.19185	loop or irregular
T	C	0.11284	β-turn
S	C	0.08258	bend
G	H	0.03911	3_{10}-helix
B	E	0.01029	β-bridge
I	C	0.00018	π-helix

Source: *Sønderby and Winther (2015)*

deepMiRGene

"deepMiRGene: Deep Neural Network based Precursor microRNA Prediction" (Park *et al.*, 2016)

- Goal: microRNA (miRNA) identification
- Complex input data
  - Sequence
  - Palindromic secondary structure

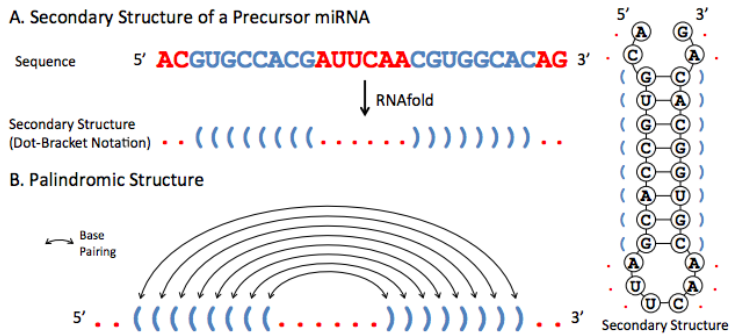


Figure 2. The secondary structure of a precursor miRNA and its panlindrome. (A) The left means the sequence of a precursor miRNA and the right represents the secondary structure of a given sequence. Dot-bracket notation (DBN), below the sequence, is the method for describing a secondary structure. Unpaired nucleotides are represented as "." and base-paired nuclotides are represented as opening "("s and closing ")"s. (B) A Palindrome in secondary structure. The forward strand (5' → 3') on the left side of the middle point and the backward strand (3' → 5') on the right side of the middel point match complementarily.

Source: *Park et al. (2016)*

Table 1. The number of datasets used in this study

Type	Human	Cross-species	New pre-miRNAs
Positive set	863	1677	690
Negative set	7422	8266	8246