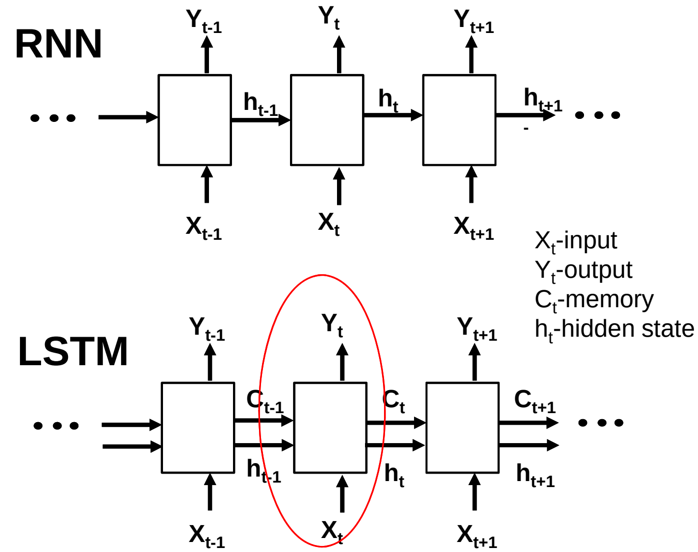
Long Short Term Memory (LSTM)

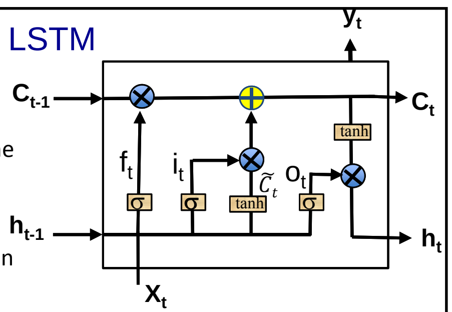
- Current state of the art
- Allows an RNN to remember things for a long time (like hundreds of time steps)
- Contains a specially designed memory cell, using logistic and linear units with multiplicative interactions
- The memory cell consists of several binary gates that control the information in the cell.
- Learns when to keep and forget the past state

See the link below for a good tutorial on LTSM
http://colah.github.io/posts/2015-08-Understanding-LSTMs/



LSTM

- The key idea is gating, which is used to decide what information to keep and to forget by varying the weight associated with it.
- Gate is implemented through a sigmoid function $\sigma: x \to R^1 \epsilon (0,1)$

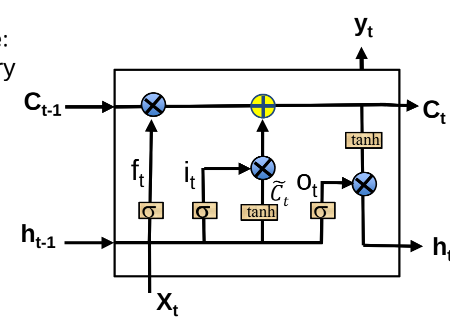


- Three types of binary gates are created
  1) Forget gate: $f_t=\sigma(W^{hf}h_{t-1}+W^{xf}X_t+W^f_0)$
     - decides what information to forget in the past memory
  2) Memory gate: $i_t=\sigma(W^{hi}h_{t-1}+W^{xi}X_t+W^i_0)$
     - decides what to store in current memory
  3) Output gate: $o_t=\sigma(W^{ho}h_{t-1}+W^{xo}X_t+W^o_0)$
     - decides what to output

LSTM

Gating operating is performed via pointwise operation
- Pointwise addition ⊕
- Pointwise multiplicaiton ⊗

Information to generate:
- Intermediate memory
- Current memory C_t
- Hidden state h_t
- Output: y_t



幻灯片 63

QJ1 Qiang Ji, 3/26/2018

LSTM

Information generation via gating

1) Intermediate memory content generation

$\tilde{C}_t$=tanh($W^{hc}h_{t-1}$+$W^{xc}X_t$+W^c_0)

2) Current memory content generation via gating

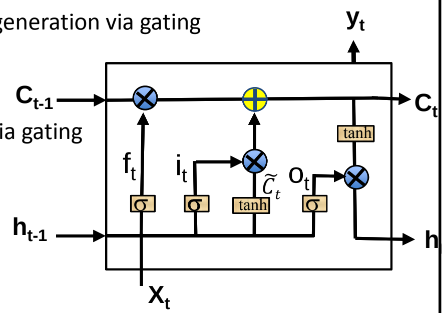
C_t=f_t⊗C_{t-1}⊕i_t ⊗ $\tilde{C}_t$

3) Current state generation via gating

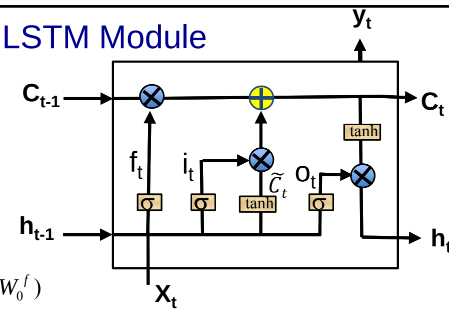
h_t=o_t ⊗ tanh(C_t)

4) Output generation

y_t=softmax($W^y h_t$+W^y_0)



A LSTM Module



$$f_t = \sigma(W^{fh}h_{t-1} + W^{xf}X_t + W_0^f)$$
$$i_t = \sigma(W^{hi}h_{t-1} + W^{xi}X_t + W_0^i)$$
$$\widetilde{C}_t = \tanh(W^{Ch}h_{t-1} + W^{cx}X_t + W_0^C)$$
$$C_t = f_t \otimes C_{t-1} \oplus i_t \otimes \widetilde{C}_t$$

The pointwise multiplication decides what to keep in current memory C

$$o_t = \sigma(W^{ho}h_{t-1} + W^{xo}X_t + W_0^o)$$
$$h_t = o_t \otimes \tanh(C_t)$$

The pointwise multiplication decides what to store in current hidden state

$$y_t = soft\max(W^y h_t + W_0^y)$$

where ⊕ and ⊗ represents pointwise multiplication and addition

LSTM Example

Assume two LSTM modules at time t=1 and t=2

At t=1

- C_0 and H_0 are initialized to 0.5
- f_1=σ($W^{hf}H_0$ +$W^{xh}X_1$+W^f_0)
- i_1=σ($W^{hi}H_0$ +$W^{xi}X_1$+W^i_0)
- o_1=σ($W^{ho}H_0$ +$W^{xo}X_1$+W^o_0)
- $\tilde{C}_1$=tanh($W^{hc}H_0$+$W^{xc}X_1$+W^c_0)
- C_1=f_1⊗C_0⊕i_t ⊗ $\tilde{C}_1$
- H_1=o_1 ⊗ tanh(C_1)
- y_1=softmax($W^y H_1$+W^y_0)

LSTM Example

At t=2

- f_2=σ($W^{hf}H_1$ +$W^{xh}X_2$+W^f_0)
- i_2=σ($W^{hi}H_1$ +$W^{xi}X_2$+W^i_0)
- o_2=σ($W^{ho}H_1$ +$W^{xo}X_2$+W^o_0)
- $\tilde{C}_2$=tanh($W^{hc}H_1$+$W^{xc}X_2$+W^c_0)
- C_2=f_2⊗C_1⊕i_2 ⊗ $\tilde{C}_1$
- H_2=o_2 ⊗ tanh(C_2)
- y_2=softmax($W^y H_2$+W^y_0)

LSTM Learning

Back propagation can be used to train LSTM.

For each LSTM module

- Compute the gradients for $\nabla \mathbf{y}_t$, $\nabla \mathbf{h}_t$ and $\nabla \mathbf{C}_t$
- Compute the gradients for weights associated with $\mathbf{y}_t$, $\mathbf{h}_t$, and C_t
- Aggregate the weight gradients for all modules over time
- Update the weights using the aggregated gradients

LSTM Learning

Forward propagation

For t=1 to T

- $f_t=\sigma(W^{hf}H_{t-1} + W^{xh}X_t + W^f_0)$
- $i_t=\sigma(W^{hi}H_{t-1} + W^{xi}X_t + W^i_0)$
- $o_t=\sigma(W^{ho}H_{t-1} + W^{xo}X_t + W^o_0)$
- $\tilde{C}_t=\tanh(W^{hc}H_{t-1} + W^{xc}X_t + W^c_0)$
- $C_t=f_t \otimes C_{t-1} \oplus i_t \otimes \tilde{C}_t$
- $H_t=o_t \otimes \tanh(C_t)$
- $\hat{y}_t=\mathrm{softmax}(W^yH_t + W^y_0)$

LSTM Learning

Backpropagation

- Compute the output gradients

$$\nabla Y_t = \frac{\partial \sum_{t=1}^{T} L(\mathbf{Y}_t, \hat{\mathbf{Y}}_t)}{\partial \hat{\mathbf{Y}}_t}, \nabla H_t = \frac{\partial \sum_{t=1}^{T} L(\mathbf{Y}_t, \hat{\mathbf{Y}}_t)}{\partial H_t},$$

- Compute weight gradients

$$\text{Given } \nabla Y_t, \nabla W^y = \frac{\partial \hat{y}_t}{\partial W^y}\nabla Y_t, \nabla W_0{}^y = \frac{\partial \hat{y}_t}{\partial W_0{}^y}\nabla Y_t \qquad y_t = soft\max(W^y h_t + W_0^y)$$

$$\text{Given } \nabla H_t, \nabla o_t = \frac{\partial H_t}{\partial o_t}\nabla H_t, \nabla C_t = \frac{\partial H_t}{\partial C_t}\nabla H_t \qquad h_t = o_t \otimes \tanh(C_t)$$

$$\text{Given } \nabla o_t, \nabla W^{ho} = \frac{\partial O_t}{\partial W^{ho}}\nabla o_t, \nabla W^{xo} = \frac{\partial O_t}{\partial W^{xo}}\nabla o, \nabla W_0{}^o = \frac{\partial O_t}{\partial W_0{}^o}\nabla o_t \qquad o_t = \sigma(W^{ho}h_{t-1} + W^{xo}X_t + W_0^o)$$

$$\text{Given } \nabla C_t, \nabla f_t = \frac{\partial C_t}{\partial f_t}\nabla C_t, \nabla i_t = \frac{\partial C_t}{\partial i_t}\nabla C_t, \nabla \widetilde{C}_t = \frac{\partial C_t}{\partial \widetilde{C}_t}\nabla C_t \qquad C_t = f_t \otimes C_{t-1} \oplus i_t \otimes \widetilde{C}_t$$

$$\text{Given } \nabla f_t, \nabla W^{hf} = \frac{\partial f_t}{\partial W^{hf}}\nabla f_t, \nabla W^{xf} = \frac{\partial f_t}{\partial W^{xf}}\nabla f, \nabla W_0{}^f = \frac{\partial f_t}{\partial W_0{}^f}\nabla f_t \qquad f_t = \sigma(W^{fh}h_{t-1} + W^{xf}X_t + W_0^f)$$

$$\text{Given } \nabla i_t, \nabla W^{hi} = \frac{\partial i_t}{\partial W^{hi}}\nabla i_t, \nabla W^{xi} = \frac{\partial i_t}{\partial W^{xi}}\nabla i_t, \nabla W_0{}^i = \frac{\partial i_t}{\partial W_0{}^i}\nabla i_t \qquad i_t = \sigma(W^{hi}h_{t-1} + W^{xi}X_t + W_0^i)$$

$$\text{Given } \nabla \widetilde{C}_t, \nabla W^{Ch} = \frac{\partial \widetilde{C}_t}{\partial W^{Ch}}\nabla \widetilde{C}_t, \nabla W^{Cx} = \frac{\partial \widetilde{C}_t}{\partial W^{Cx}}\nabla \widetilde{C}_t, \nabla W_0{}^C = \frac{\partial \widetilde{C}_t}{\partial W_0{}^C}\nabla \widetilde{C}_t \qquad \widetilde{C}_t = \tanh(W^{Ch}h_{t-1} + W^{cx}X_t + W_0^C)$$

LSTM Learning

Weight updating

- Aggregate the gradients over time

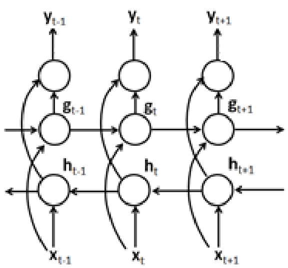
$$\nabla W = \sum_{t=1}^{T} \nabla W_t$$

- Update the gradients through gradient descnet

$$W(k) = W(k-1) - \eta \nabla W$$

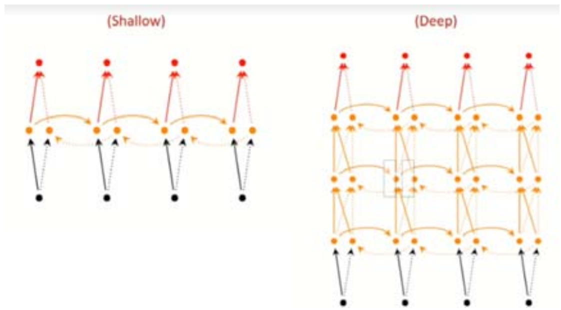
Variants of RNNs

- Bi-directional RNN



Variants of RNNs

- Deep RNN

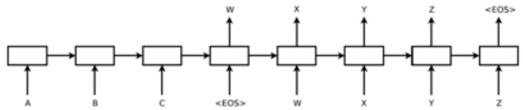


LSTM Applications

- Sequence to sequence translation
- Image captioning
- Video to sentence
- Binary addition
- Cursive handwriting recognition

Sequence-to-sequence language translation

Use one LSTM to read the input sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$, one time step at a time, and output another sequence $(\mathbf{y}_1, \ldots, \mathbf{y}_{T'})$



The model reads an input sentence "ABC" and produces "WXYZ" as the output sentence.

The model stops making predictions after outputting the end-of-sentence token.

Sutskever, Vinyals, and Le NIPS 2014

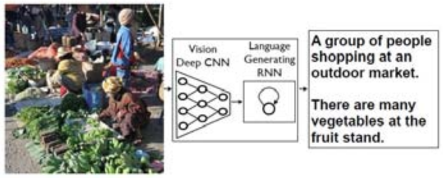
Sequence-to-sequence language translation

LSTM can correctly translate very long sentences

Type	Sentence
Our model	Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .
Truth	Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .
Our model	" Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air " , dit UNK .
Truth	" Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord " , a déclaré Rosenker .
Our model	Avec la crémation , il y a un " sentiment de violence contre le corps d' un être cher " , qui sera " réduit à une pile de cendres " en très peu de temps au lieu d' un processus de décomposition " qui accompagnera les étapes du deuil " .
Truth	Il y a , avec la crémation , " une violence faite au corps aimé " , qui va être " réduit à un tas de cendres " en très peu de temps , et non après un processus de décomposition , qui " accompagnerait les phases du deuil " .

Generate image caption

- Use a CNN as an image encoder and transform it to a fixed-length vector
- Input as a sequence of image patches
- It then uses a RNN to generate the target sequence

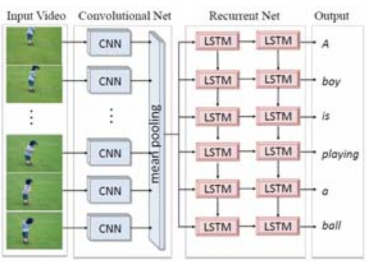


Vinyals et al. arXiv 2014

Translate videos to sentences

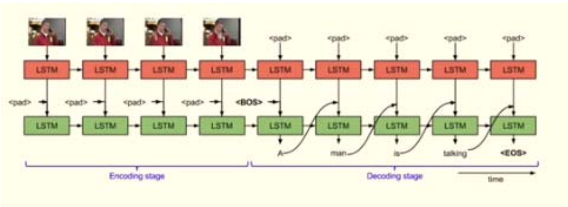
Each frame is modeled as CNN pre-trained on ImageNet

The meaning state and sequence of words is modeled by a RNN pre-trained on images with associated with sentence captions



Translate videos to sentences

Machine output: *A cat is playing with toy.*
Humans: *A Ferret and cat fighting with each other. / A cat and a ferret are playing. / kitten and a ferret are playfully wresting.*



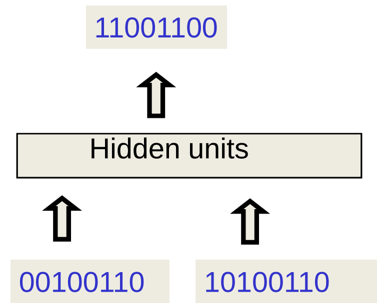
Venugopalan et al. arXiv 2014

Translate videos to sentences

FGM: factor graph model, using templates to generate sentences

YT: LSTM trained on the YouTube video dataset
YT_C: LSTM with pre-training on the Coco image dataset
YT_CF: LSTM with pre-training on the CoCo and Flicker image datasets
GT: Ground truth from human description

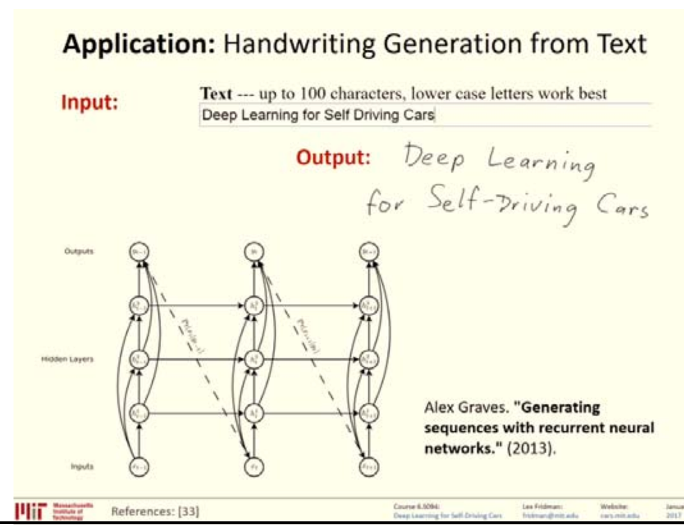
Binary Number Addition



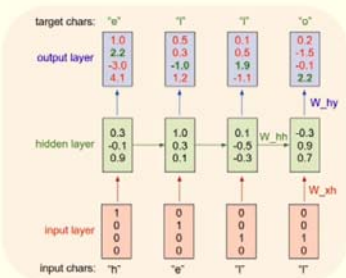
Reading cursive handwriting

- This is a natural task for an RNN.
- The input is a sequence of (x,y,p) coordinates of the tip of the pen, where p indicates whether the pen is up or down.
- The output is a sequence of characters.
- Graves & Schmidhuber (2009) showed that RNNs with LSTM are currently the best systems for reading cursive writing.
  - They used a sequence of small images as input rather than pen coordinates.

Slide from Geoffrey Hinton



Application: Character-Level Text Generation



Life Is About The Weather!
Life Is About The True Love Of Mr. Mom
Life Is About Kids
Life Is About An Eating Story
Life Is About The Truth Now

The meaning of life is literary recognition

The meaning of life is the tradition of the ancient human reproduction

Andrej Karpathy. "The Unreasonable Effectiveness of Recurrent Neural Networks." (2015).

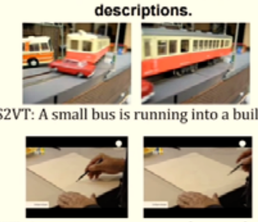
Application: Video Description Generation

Correct descriptions.

S2VT: A man is doing stunts on his bike.

S2VT: A herd of zebras are walking in a field.

Relevant but incorrect descriptions.



S2VT: A small bus is running into a building.

S2VT: A man is cutting a piece of a pair of a paper.

Venugopalan et al. "Sequence to sequence-video to text." 2015.

Code: https://vsubhashini.github.io/s2vt.html

Deep Learning News (Jan, 2017)

- Tweet sentiment analysis for stock prediction (4/13/17)
  - http://www.npr.org/2017/02/04/513469456/when-trump-tweets-this-bot-makes-money

- DeepBach
  - Bach-like music composition using deep learning
  - https://www.youtube.com/watch?v=QiBM7-5hA6o&app=desktop