

Neural Machine Translation to the rescue!

- Store a sentence table implicitly.
- Simple and coherent.

But we need to understand Recurrent Neural Networks first!

Outline

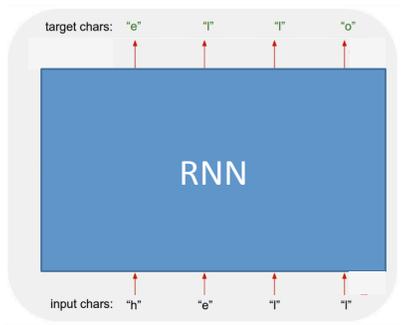
- Recurrent Neural Networks (RNNs)
- NMT basics (Sutskever et al., 2014)
- Attention mechanism (Bahdanau et al., 2015)

Recurrent Neural Networks (RNNs)

Character-level language model example

Vocabulary: [h,e,l,o]

Example training sequence: "hello"



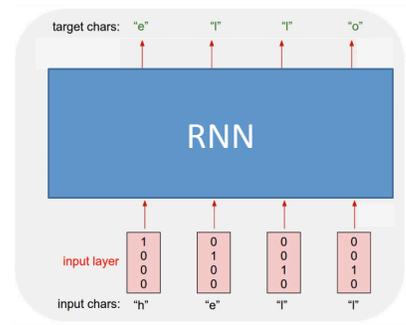
(Picture adapted from Andrej Karparthy)

RNN – Input Layer

Character-level language model example

Vocabulary: [h,e,l,o]

Example training sequence: "hello"



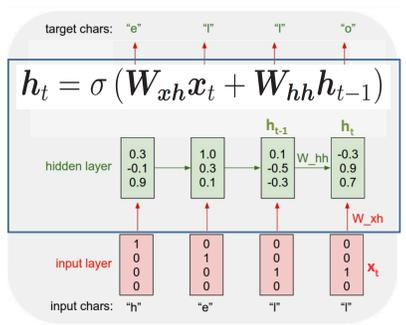
(Picture adapted from Andrej Karparthy)

RNN – Hidden Layer

Character-level language model example

Vocabulary: [h,e,l,o]

Example training sequence: "hello"



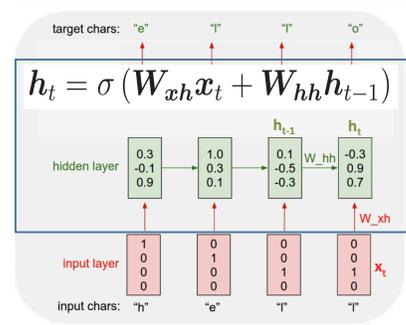
(Picture adapted from Andrej Karparthy)

RNN – Hidden Layer

Character-level language model example

Vocabulary: [h,e,l,o]

Example training sequence: "hello"



RNNs to represent sequences!

Outline

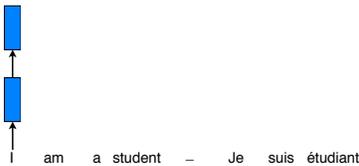
- Recurrent Neural Networks (RNNs)
- NMT basics (Sutskever et al., 2014)
 - Encoder-Decoder.
 - Training vs. Testing.
 - Backpropagation.
 - More about RNNs.
- Attention mechanism (Bahdanau et al., 2015)

Neural Machine Translation (NMT)

I am a student - Je suis étudiant

- Model $P(\text{target} \mid \text{source})$ directly.

Neural Machine Translation (NMT)



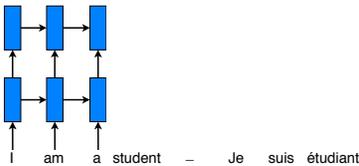
- RNNs trained **end-to-end** (Sutskever et al., 2014).

Neural Machine Translation (NMT)



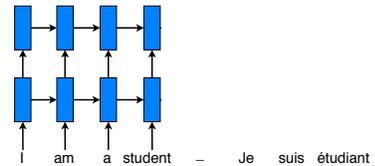
- RNNs trained **end-to-end** (Sutskever et al., 2014).

Neural Machine Translation (NMT)



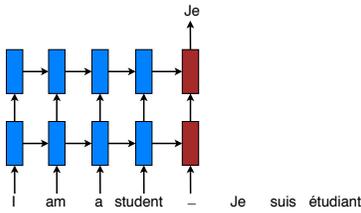
- RNNs trained **end-to-end** (Sutskever et al., 2014).

Neural Machine Translation (NMT)



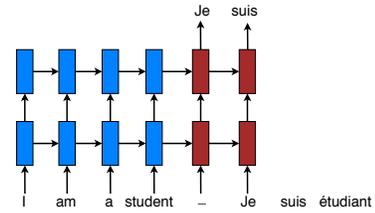
- RNNs trained **end-to-end** (Sutskever et al., 2014).

Neural Machine Translation (NMT)



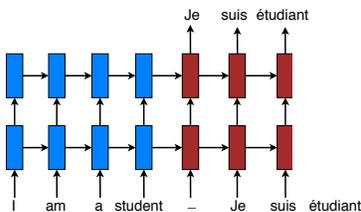
- RNNs trained **end-to-end** (Sutskever et al., 2014).

Neural Machine Translation (NMT)



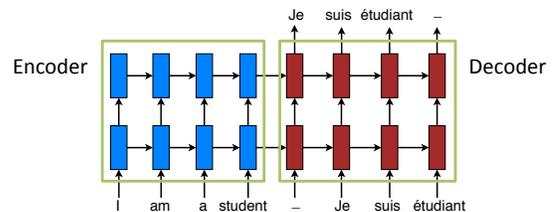
- RNNs trained **end-to-end** (Sutskever et al., 2014).

Neural Machine Translation (NMT)



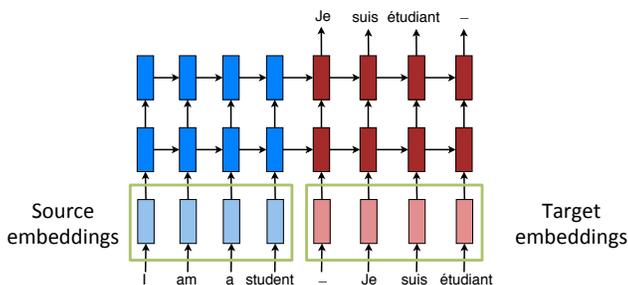
- RNNs trained **end-to-end** (Sutskever et al., 2014).

Neural Machine Translation (NMT)



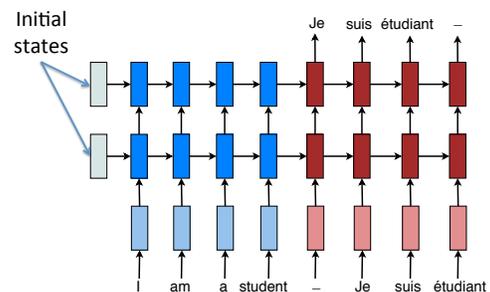
- RNNs trained **end-to-end** (Sutskever et al., 2014).
- **Encoder-decoder** approach.

Word Embeddings



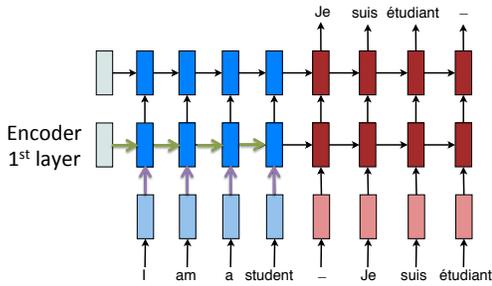
- **Randomly initialized**, one for each language.
 - Learnable parameters.

Recurrent Connections



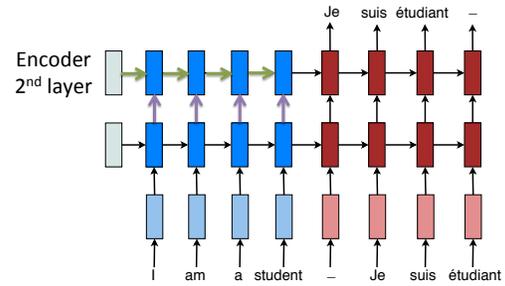
- Often set to 0.

Recurrent Connections



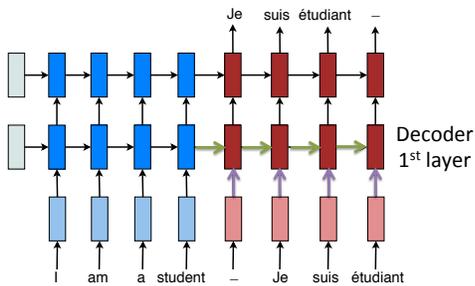
- Different across **layers** and **encoder / decoder**.

Recurrent Connections



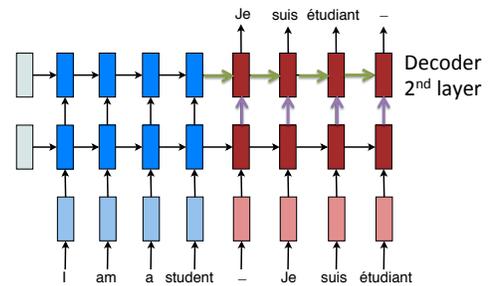
- Different across **layers** and **encoder / decoder**.

Recurrent Connections



- Different across **layers** and **encoder / decoder**.

Recurrent Connections



- Different across **layers** and **encoder / decoder**.

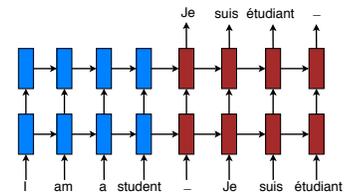
Outline

- Recurrent Neural Networks (RNNs)
- NMT basics (Sutskever et al., 2014)
 - Encoder-Decoder.
 - **Training vs. Testing.**
 - Backpropagation.
 - More about RNNs.
- Attention mechanism (Bahdanau et al., 2015)

Training vs. Testing

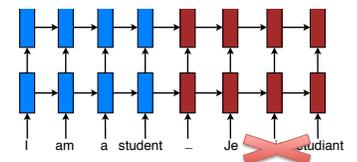
• Training

- **Correct translations** are available.

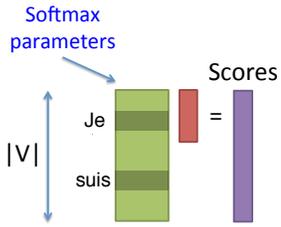
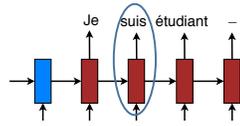


• Testing

- **Only source sentences** are given.

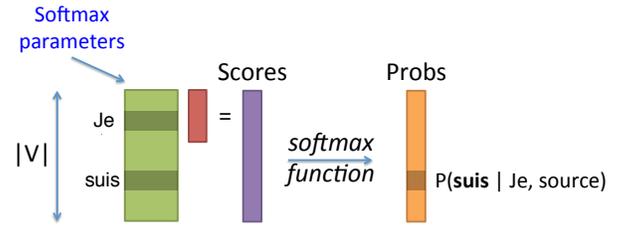
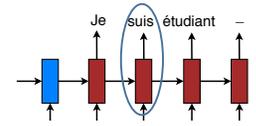


Training – Softmax



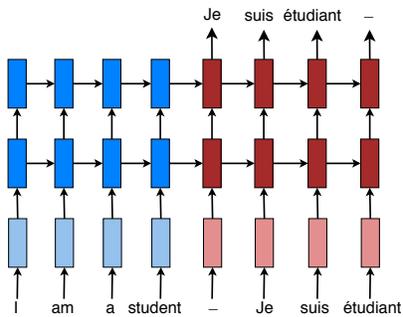
- Hidden states \mapsto scores.

Training – Softmax



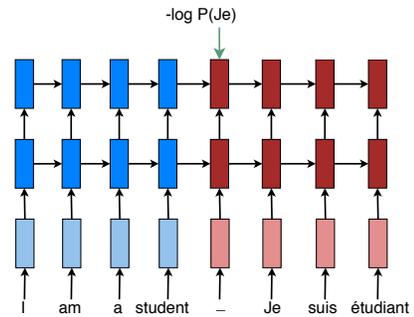
- Scores \mapsto probabilities.

Training Loss



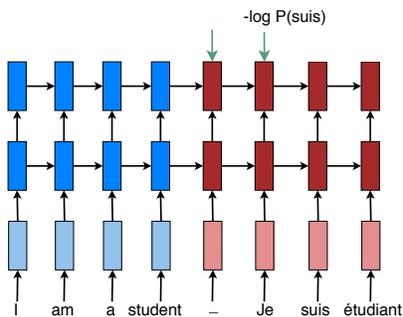
- Maximize $P(\text{target} \mid \text{source})$:
– Decompose into individual word predictions.

Training Loss



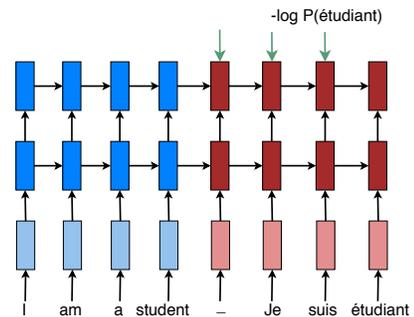
- Sum of all individual losses

Training Loss



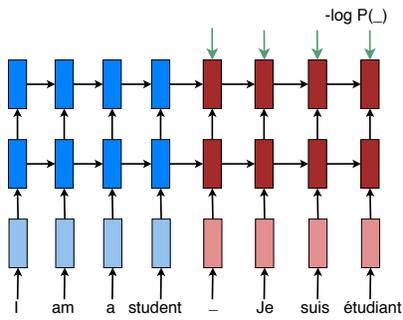
- Sum of all individual losses

Training Loss



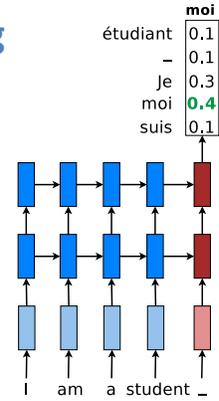
- Sum of all individual losses

Training Loss



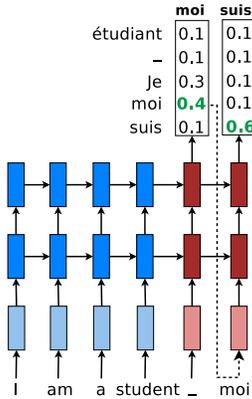
- Sum of all individual losses

Testing



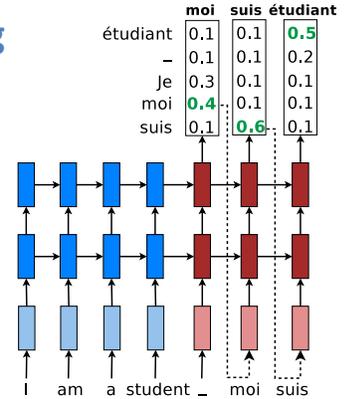
- Feed the **most likely** word

Testing



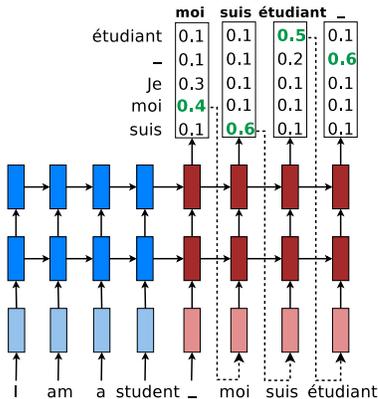
- Feed the **most likely** word

Testing



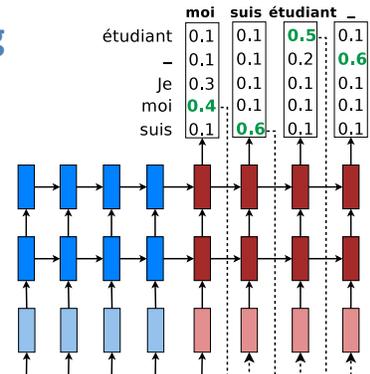
- Feed the **most likely** word

Testing



- Feed the **most likely** word

Testing

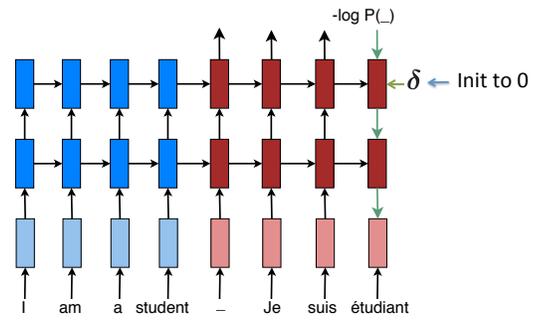


NMT beam-search decoders are much simpler!

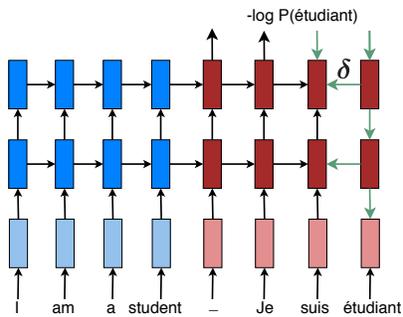
Outline

- Recurrent Neural Networks (RNNs)
- NMT basics (Sutskever et al., 2014)
 - Encoder-Decoder.
 - Training vs. Testing.
 - **Backpropagation.**
 - More about RNNs.
- Attention mechanism (Bahdanau et al., 2015)

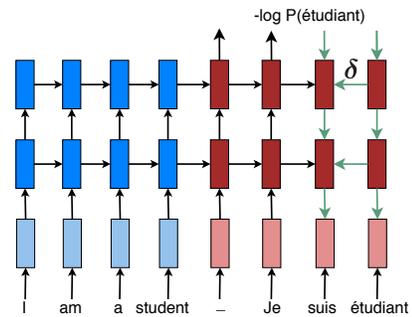
Backpropagation Through Time



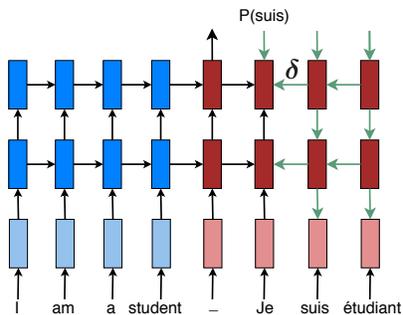
Backpropagation Through Time



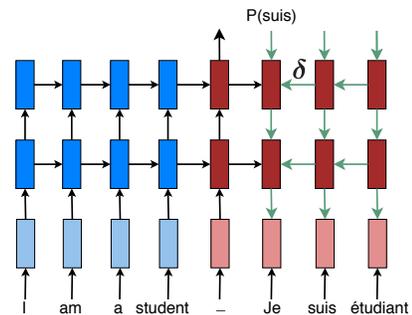
Backpropagation Through Time



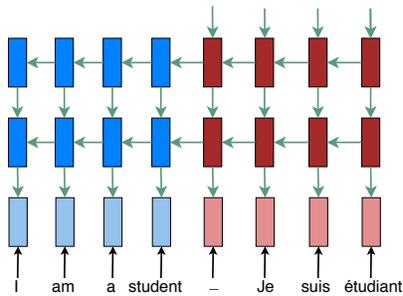
Backpropagation Through Time



Backpropagation Through Time



Backpropagation Through Time

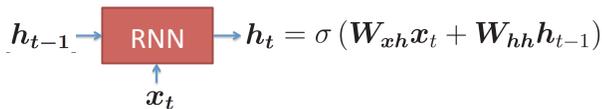
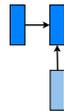


RNN gradients are accumulated.

Outline

- Recurrent Neural Networks (RNNs)
- NMT basics (Sutskever et al., 2014)
 - Encoder-Decoder.
 - Training vs. Testing.
 - Backpropagation.
 - [More about RNNs.](#)
- Attention mechanism (Bahdanau et al., 2015)

Recurrent types – vanilla RNN



Vanishing gradient problem!

Vanishing gradients

$$\mathbf{h}_t = \sigma(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1})$$

$$\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} = \text{diag}(\sigma'(\dots)) \mathbf{W}_{hh}^\top \quad \text{Chain Rule}$$

$$\left\| \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} \right\| \leq \|\text{diag}(\sigma'(\dots))\| \|\mathbf{W}_{hh}^\top\| \quad \text{Bound Rules}$$

$$\leq \gamma \lambda_1$$

$$\text{Bound } \|\text{diag}(\sigma'(\dots))\| \quad \text{Largest singular value } \mathbf{W}_{hh}^\top$$

(Pascanu et al., 2013)

Vanishing gradients

$$\mathbf{h}_t = \sigma(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1})$$

$$\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} = \text{diag}(\sigma'(\dots)) \mathbf{W}_{hh}^\top \quad \text{Chain Rule}$$

$$\left\| \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} \right\| \leq \gamma \lambda_1 \quad \text{Bound Rules}$$

$$\left\| \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-k}} \right\| \leq (\gamma \lambda_1)^k \quad \text{Chain Rule}$$

(Pascanu et al., 2013)

Vanishing gradients

$$\mathbf{h}_t = \sigma(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1})$$

$$\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} = \text{diag}(\sigma'(\dots)) \mathbf{W}_{hh}^\top \quad \text{Chain Rule}$$

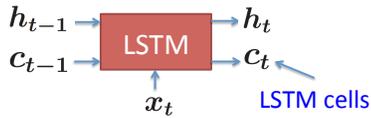
$$\left\| \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} \right\| \leq \gamma \lambda_1 \quad \text{Bound Rules}$$

$$\left\| \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-k}} \right\| \leq (\gamma \lambda_1)^k \rightarrow 0 \text{ if } \lambda_1 < \frac{1}{\gamma} \quad \text{Sufficient Cond}$$

(Pascanu et al., 2013)

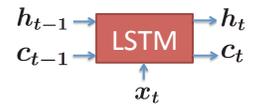
Recurrent types – LSTM

C'mon, it's been around for 20 years!



- Long-Short Term Memory (LSTM)
 - (Hochreiter & Schmidhuber, 1997)
- LSTM cells are **additively** updated
 - Make backprop through time easier.

Building LSTM



$$\begin{pmatrix} \hat{h}_t \\ c_t \end{pmatrix} = \begin{pmatrix} \text{tanh} \\ \text{sigm} \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$

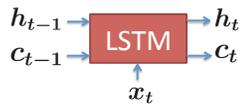
$$c_t = c_{t-1} + \hat{h}_t$$

$$h_t = c_t$$

Nice gradients!
 $\frac{\partial c_t}{\partial c_{t-1}} = I$

- A naïve version.

Building LSTM



Input gates

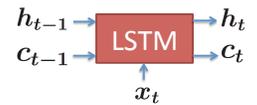
$$\begin{pmatrix} i_t \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{tanh} \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$

$$c_t = c_{t-1} + i_t \circ \hat{h}_t$$

$$h_t = c_t$$

- Add **input** gates: control input signal.

Building LSTM



Forget gates

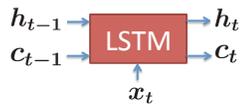
$$\begin{pmatrix} f_t \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{tanh} \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{h}_t$$

$$h_t = c_t$$

- Add **forget** gates: control memory.

Building LSTM



Output gates

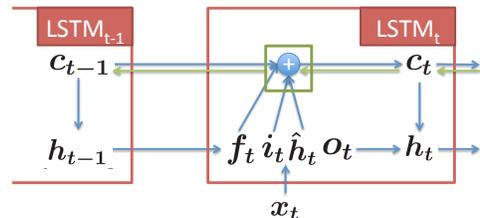
$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{h}_t$$

$$h_t = o_t \circ \tanh(c_t)$$

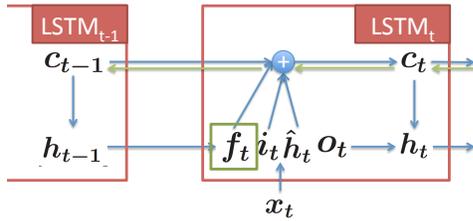
- Add **output** gates: extract information.
- (Zaremba et al., 2014).

Why LSTM works?



- The **additive** operation is the key!
- **Backpropagation path** through the cell is effective.

Why LSTM works?



- The **additive** operation is the key!
- **Backpropagation path** through the cell is effective.

Forget gates are important!

Other RNN units

- (Graves, 2013): **revived LSTM**.
 - Direct connections between cells and gates.
- **Gated Recurrent Unit (GRU)** – (Cho et al., 2014a)
 - No cells, same additive idea.
- **LSTM vs. GRU**: mixed results (Chung et al., 2015).

English – French WMT Results

Systems	BLEU
SOTA in WMT'14 (Durrani et al., 2014)	37.0
<i>Standard MT + neural components</i>	
Schwenk (2014) – neural language model	33.3
Cho et al. (2014) – phrase table neural features	34.5

English – French WMT Results

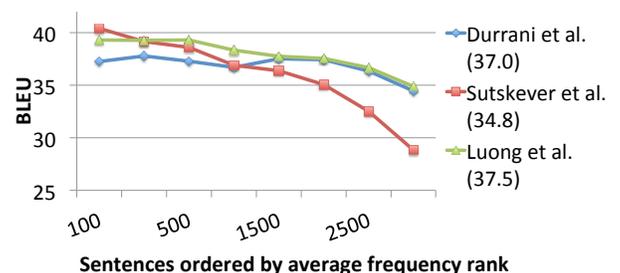
Systems	BLEU
SOTA in WMT'14 (Durrani et al., 2014)	37.0
<i>Standard MT + neural components</i>	
Schwenk (2014) – neural language model	33.3
Cho et al. (2014) – phrase table neural features	34.5
<i>NMT</i>	
Sutskever et al. (2014) – ensemble LSTMs	34.8

English – French WMT Results

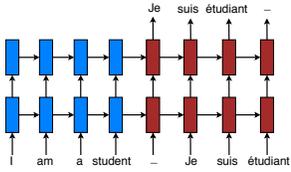
Systems	BLEU
SOTA in WMT'14 (Durrani et al., 2014)	37.0
<i>Standard MT + neural components</i>	
Schwenk (2014) – neural language model	33.3
Cho et al. (2014) – phrase table neural features	34.5
<i>NMT</i>	
Sutskever et al. (2014) – ensemble LSTMs	34.8
Luong et al. (2015a) – ensemble LSTMs + rare word	37.5

New SOTA!

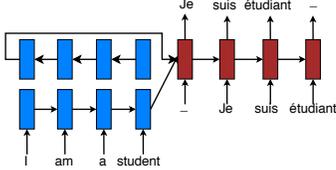
Effects of Translating Rare Words



Deep RNNs (Sutskever et al., 2014)



Bidirectional RNNs (Bahdanau et al., 2015)



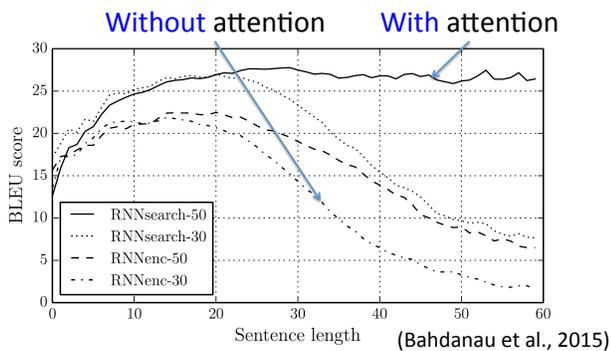
Summary

- Generalize well.
- Small memory.
- Simple decoder.

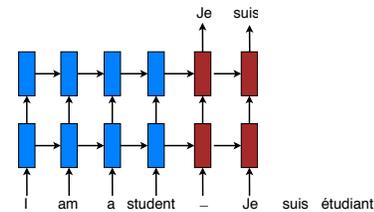
Outline

- Recurrent Neural Networks (RNNs)
- NMT basics (Sutskever et al., 2014)
- Attention mechanism (Bahdanau et al., 2015)

Sentence Length Problem

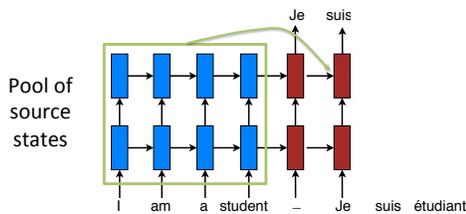


Why?



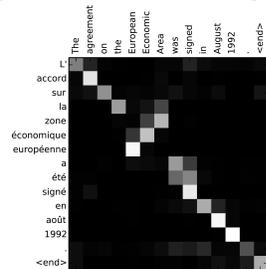
- A **fixed-dimensional** source vector.
- **Problem:** Markovian process.

Attention Mechanism



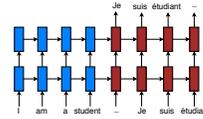
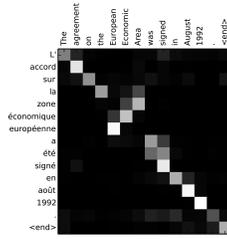
- **Solution:** random access memory
 - Retrieve as needed.
 - cf. Neural Turing Machine (Graves et al., 2014).

Alignments as a by-product

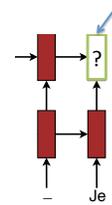


- **Recent innovation** in deep learning:
 - Control problem (Mnih et al., 14)
 - Speech recognition (Chorowski et al., 15)
 - Image caption generation (Xu et al., 15)

Simplified Attention
(Bahdanau et al., 2015)

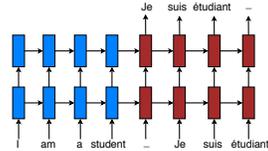


What's next?

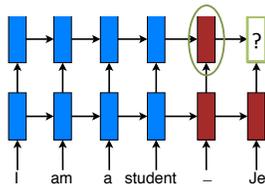


+

Deep LSTM
(Sutskever et al., 2014)



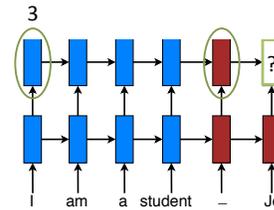
Attention Mechanism – Scoring



- Compare target and source hidden states.

Attention Mechanism – Scoring

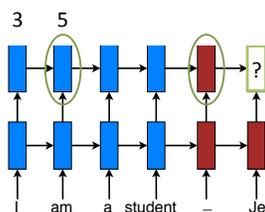
$$\text{score}(h_t, \bar{h}_s)$$



- Compare target and source hidden states.

Attention Mechanism – Scoring

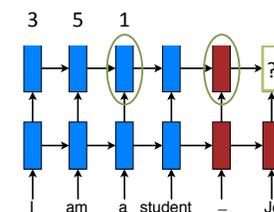
$$\text{score}(h_t, \bar{h}_s)$$



- Compare target and source hidden states.

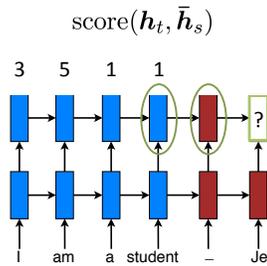
Attention Mechanism – Scoring

$$\text{score}(h_t, \bar{h}_s)$$



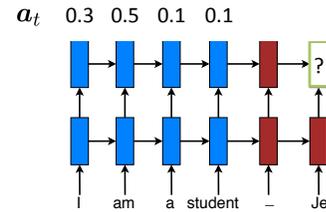
- Compare target and source hidden states.

Attention Mechanism – Scoring



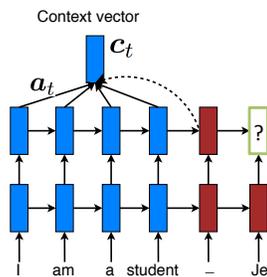
- Compare target and source hidden states.

Attention Mechanism – Normalization



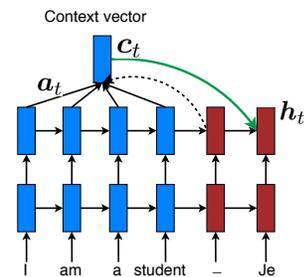
- Convert into alignment weights.

Attention Mechanism – Context vector



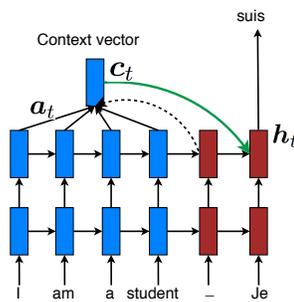
- Build context vector: weighted average.

Attention Mechanism – Hidden state



- Compute the next hidden state.

Attention Mechanism – Predict



- Predict the next word.

Attention Mechanism – Score Functions

$$\text{score}(h_t, \bar{h}_s) = \begin{cases} v_a^T \tanh(W_a[h_t; \bar{h}_s]) & \text{(Bahdanau et al., 2015)} \end{cases}$$

Attention Mechanism – Score Functions

$$\text{score}(\mathbf{h}_t, \bar{\mathbf{h}}_s) = \begin{cases} \mathbf{h}_t^\top \bar{\mathbf{h}}_s & \text{(Luong et al., 2015b)} \\ \mathbf{h}_t^\top \mathbf{W}_a \bar{\mathbf{h}}_s & \\ \mathbf{v}_a^\top \tanh(\mathbf{W}_a [\mathbf{h}_t; \bar{\mathbf{h}}_s]) & \text{(Bahdanau et al., 2015)} \end{cases}$$

- More **focused attention** (Luong et al., 2015b)
 - Focus on a subset of words each time.



English-German WMT Results

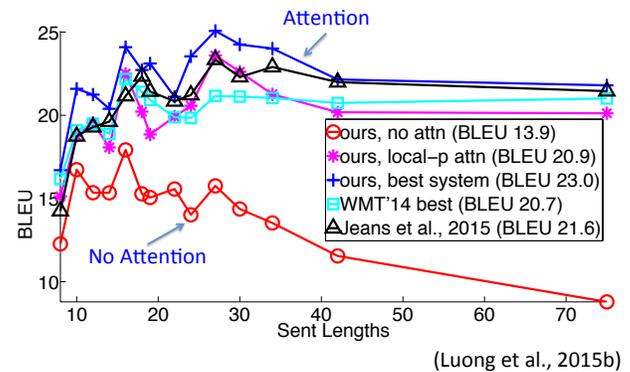
Systems	BLEU
SOTA in WMT'14 (Buck et al., 2014)	20.7
<i>NMT</i>	
Jean et al., (2015) – GRUs + attention	21.6

English-German WMT Results

Systems	BLEU
SOTA in WMT'14 (Buck et al., 2014)	20.7
<i>NMT</i>	
Jean et al., (2015) – GRUs + attention	21.6
Luong et al. (2015b) – LSTMs + improved attention	23.0 (+2.3)



Translate Long Sentences



Sample English-German translations

src	Orlando Bloom and Miranda Kerr still love each other
ref	Orlando Bloom und Miranda Kerr lieben sich noch immer
best	Orlando Bloom und Miranda Kerr lieben einander noch immer .
base	Orlando Bloom und Lucas Miranda lieben einander noch immer .

- Translate names correctly.

(Luong et al., 2015b)

Sample English-German translations

src	" We ' re pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security , " said Roger Dow , CEO of the U.S. Travel Association .
ref	" Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht im Widerspruch zur Sicherheit steht " , sagte Roger Dow , CEO der U.S. Travel Association .
best	" Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit unvereinbar ist " , sagte Roger Dow , CEO der US - die .
base	" Wir freuen uns über die <unk> , dass ein <unk> <unk> mit Sicherheit nicht vereinbar ist mit Sicherheit und Sicherheit " , sagte Roger Cameron , CEO der US - <unk> .

- Translate a **doubly-negated phrase** correctly

(Luong et al., 2015b)

Sample English-German translations

src	"We're pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security," said Roger Dow, CEO of the U.S. Travel Association.
ref	"Wir freuen uns, dass die FAA erkennt, dass ein angenehmes Passagiererlebnis nicht im Widerspruch zur Sicherheit steht ", sagte Roger Dow, CEO der U.S. Travel Association.
best	"Wir freuen uns, dass die FAA anerkennt, dass ein angenehmes ist nicht mit Sicherheit und Sicherheit unvereinbar ist", sagte Roger Dow, CEO der US - die .
base	"Wir freuen uns über die <unk>, dass ein <unk> <unk> mit Sicherheit nicht vereinbar ist mit Sicherheit und Sicherheit", sagte Roger Cameron, CEO der US - <unk> .

- Translate a **doubly-negated phrase** correctly

(Luong et al., 2015b)

Sample English-German translations

src	"We're pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security," said Roger Dow, CEO of the U.S. Travel Association.
ref	"Wir freuen uns, dass die FAA erkennt, dass ein angenehmes Passagiererlebnis nicht im Widerspruch zur Sicherheit steht ", sagte Roger Dow, CEO der U.S. Travel Association.
best	"Wir freuen uns, dass die FAA anerkennt, dass ein angenehmes ist nicht mit Sicherheit und Sicherheit unvereinbar ist", sagte Roger Dow, CEO der US - die .
base	"Wir freuen uns über die <unk>, dass ein <unk> <unk> mit Sicherheit nicht vereinbar ist mit Sicherheit und Sicherheit", sagte Roger Cameron, CEO der US - <unk> .

- **Fail** to translate "passenger experience".

(Luong et al., 2015b)

Sample German-English translations

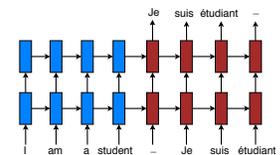
src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke, in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird, sind viele Menschen der Ansicht, das Projekt Europa sei zu weit gegangen.
ref	The usterity imposed by Berlin and the European Central Bank, coupled with the straitjacket imposed on national economies through adherence to the common currency, has led many people to think Project Europe has gone too far.
best	Because of the strict usterity measures imposed by Berlin and the European Central Bank in connection with the straitjacket in which the respective national economy is forced to adhere to the common currency, many people believe that the European project has gone too far.
base	Because of the pressure imposed by the European Central Bank and the Federal Central Bank with the strict austerity imposed on the national economy in the face of the single currency, many people believe that the European project has gone too far.

- Translate well long sentences. (Luong et al., 2015b)



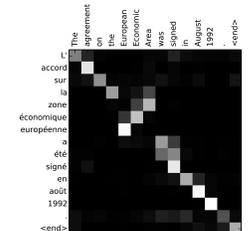
Summary

Deep LSTM
(Sutskever et al., 2014)



+

Simplified Attention
(Bahdanau et al., 2015)



References (1)

- [Bahdanau et al., 2015] Neural Translation by Jointly Learning to Align and Translate. <http://arxiv.org/pdf/1409.0473.pdf>
- [Cho et al., 2014a] Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. <http://aclweb.org/anthology/D/D14/D14-1179.pdf>
- [Cho et al., 2014b] On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. <http://www.aclweb.org/anthology/W14-4012>
- [Chorowski et al., 2015] Attention-Based Models for Speech Recognition. <http://arxiv.org/pdf/1506.07503v1.pdf>
- [Chung et al., 2015] Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. <http://arxiv.org/pdf/1412.3555.pdf>
- [Graves, 2013] Generating Sequences With Recurrent Neural Networks. <http://arxiv.org/pdf/1308.0850v5.pdf>
- [Graves, 2014] Neural Turing Machine. <http://arxiv.org/pdf/1410.5401v2.pdf>
- [Hochreiter & Schmidhuber, 1997] Long Short-term Memory. http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf

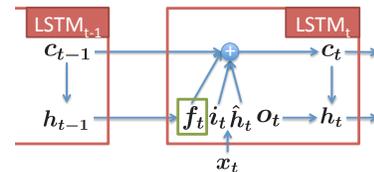
References (2)

- [Kalchbrenner & Blunsom, 2013] Recurrent Continuous Translation Models. http://nal.co/papers/KalchbrennerBlunsom_EMNLP13
- [Luong et al., 2015a] Addressing the Rare Word Problem in Neural Machine Translation. <http://www.aclweb.org/anthology/P15-1002>
- [Luong et al., 2015b] Effective Approaches to Attention-based Neural Machine Translation. <https://aclweb.org/anthology/D/D15/D15-1166.pdf>
- [Mnih et al., 2014] Recurrent Models of Visual Attention. <http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention.pdf>
- [Pascanu et al., 2013] On the difficulty of training Recurrent Neural Networks. <http://arxiv.org/pdf/1211.5063v2.pdf>
- [Xu et al., 2015] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. <http://jmlr.org/proceedings/papers/v37/xuc15.pdf>
- [Sutskever et al., 2014] Sequence to Sequence Learning with Neural Networks. <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>
- [Zaremba et al., 2015] Recurrent Neural Network Regularization. <http://arxiv.org/pdf/1409.2329.pdf>

Encoder-decoder Summary

	Encoder	Decoder
(Sutskever et al., 2014) (Luong et al., 2015a) (Luong et al., 2015b)	Deep LSTM	Deep LSTM
(Cho et al., 2014a) (Bahdanau et al., 2015) (Jean et al., 2015)	(Bidirectional) GRU	GRU
(Kalchbrenner & Blunsom, 2013)	CNN	(Inverse CNN) RNN
(Cho et al., 2014b)	Gated Recursive CNN	GRU

Important LSTM components?

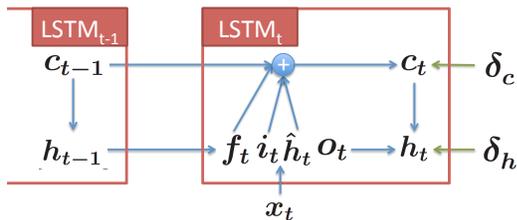


- (Jozefowicz et al., 2015): **forget** gate bias of 1.

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{h}_t$$
- (Greff et al., 2015): **forget** gates & **output** acts.

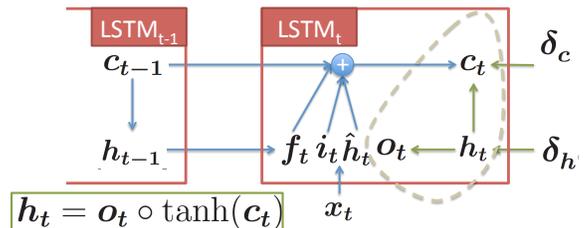
$$h_t = o_t \circ \tanh(c_t)$$

LSTM Backpropagation

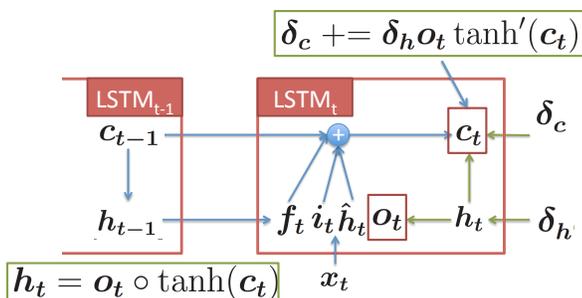


- Deltas sent back from the top layers.

LSTM Backpropagation – Context

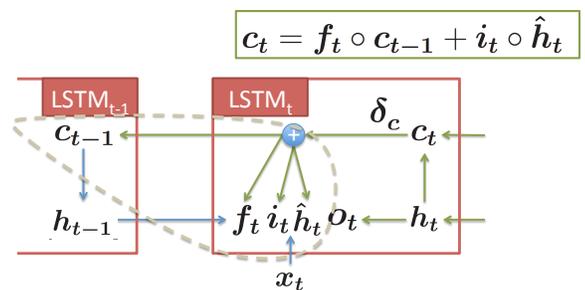


LSTM Backpropagation – Context

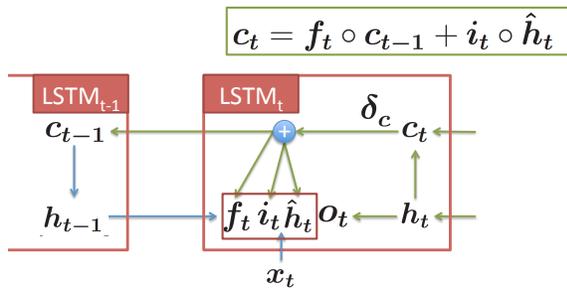


- Complete **context vector** gradient.

LSTM Backpropagation – Context

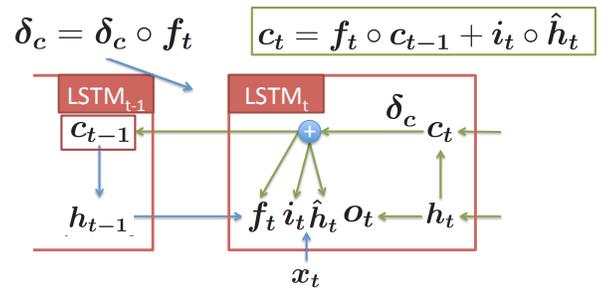


LSTM Backpropagation – Context



- First, use δ_c to compute gradients for $f_t i_t \hat{h}_t$.

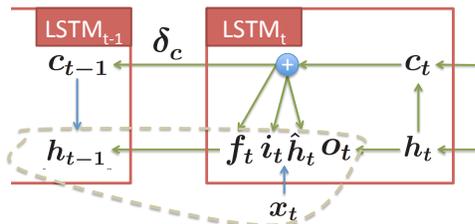
LSTM Backpropagation – Context



- Then, update δ_c .

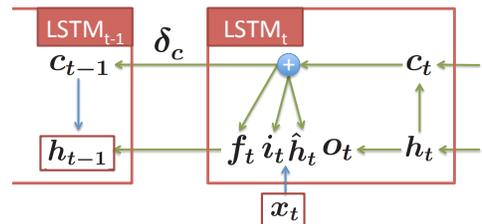
LSTM Backprop

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$



LSTM Backprop

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$

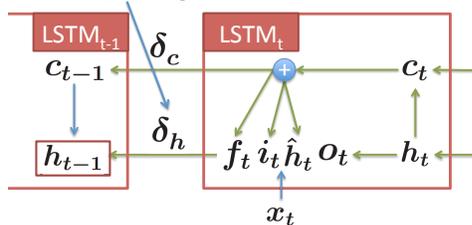


- Compute gradients for $T_{4n \times 2n}$, x_t , h_{t-1} .

LSTM Backprop

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$

δ_h += upper grad



- Add gradients from the [loss / upper](#) layers.

Summary

- LSTM backpropagation is nasty. 😞

- But it will be much easier if:

- Know your matrix calculus!
- Pay attention to δ_c and δ_h .

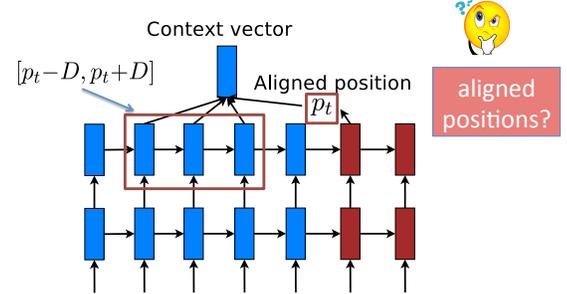


Other Attention Functions

- Content-based: $\mathbf{a}_t = \text{Attend}(\mathbf{h}_{t-1}, \bar{\mathbf{h}}_{1\dots S})$
- Location-based: $\mathbf{a}_t = \text{Attend}(\mathbf{h}_{t-1}, \mathbf{a}_{t-1})$
– (Graves, 2013): hand-writing synthesis model.
- Hybrid: $\mathbf{a}_t = \text{Attend}(\mathbf{h}_{t-1}, \mathbf{a}_{t-1}, \bar{\mathbf{h}}_{1\dots S})$
– (Chorowski et al., 2015) for speech recognition.

Local Attention (Luong et al., 2015b)

- More focused attention.
– Potentially useful for longer text sequences.



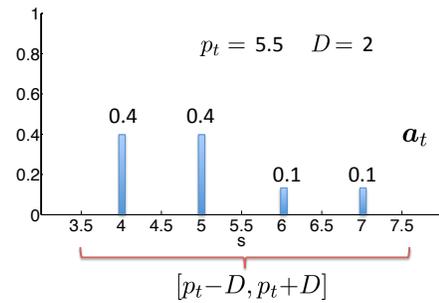
$$p_t = S \cdot \text{sigmoid}(\mathbf{v}_p^\top \tanh(\mathbf{W}_p \mathbf{h}_t))$$

Real value in $[0, S]$ Source sentence

How do we learn to the position parameters?

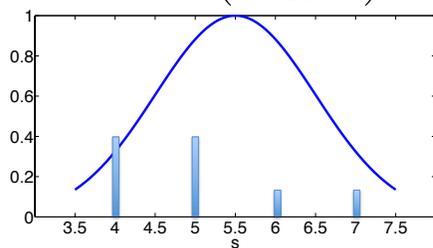
$$p_t = S \cdot \text{sigmoid}(\mathbf{v}_p^\top \tanh(\mathbf{W}_p \mathbf{h}_t))$$

Alignment Weights



Truncated Gaussian

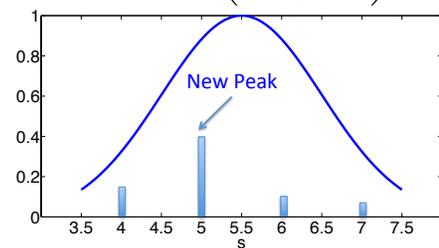
$$\exp\left(-\frac{(s-p_t)^2}{2\sigma^2}\right)$$



- Favor points close to the center.

Scaled Alignment Weights

$$\mathbf{a}_t(s) \exp\left(-\frac{(s-p_t)^2}{2\sigma^2}\right)$$



Differentiable almost everywhere!