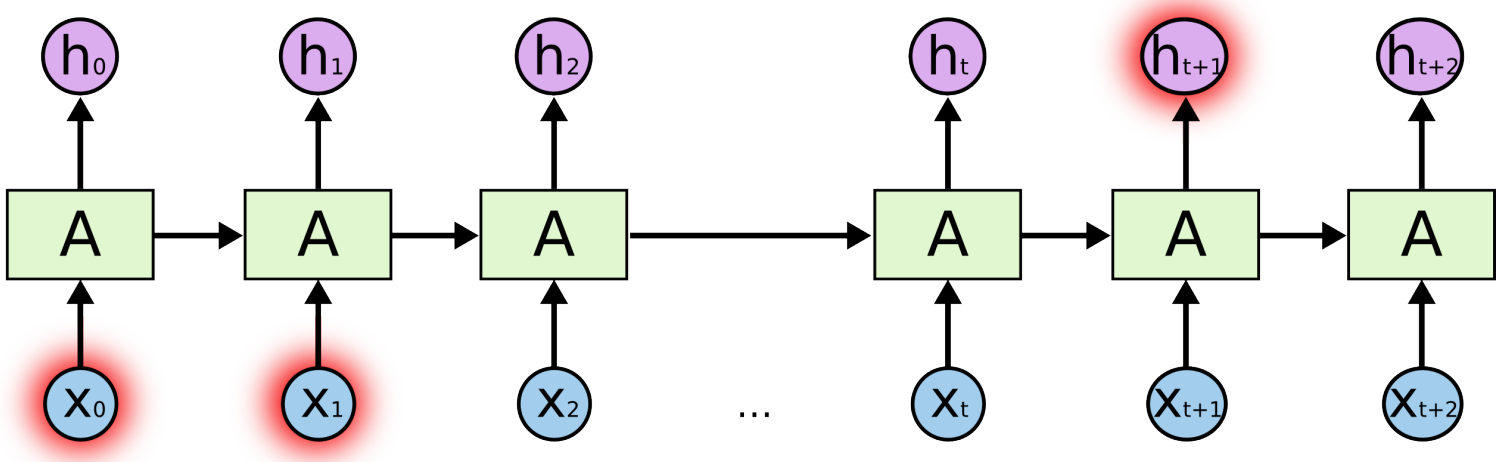
LSTMs, GRUs and Applications

IE643 - Lectures 22 & 23

October 22 & Nov 1, 2024.

Recurrent Neural Network - Drawbacks

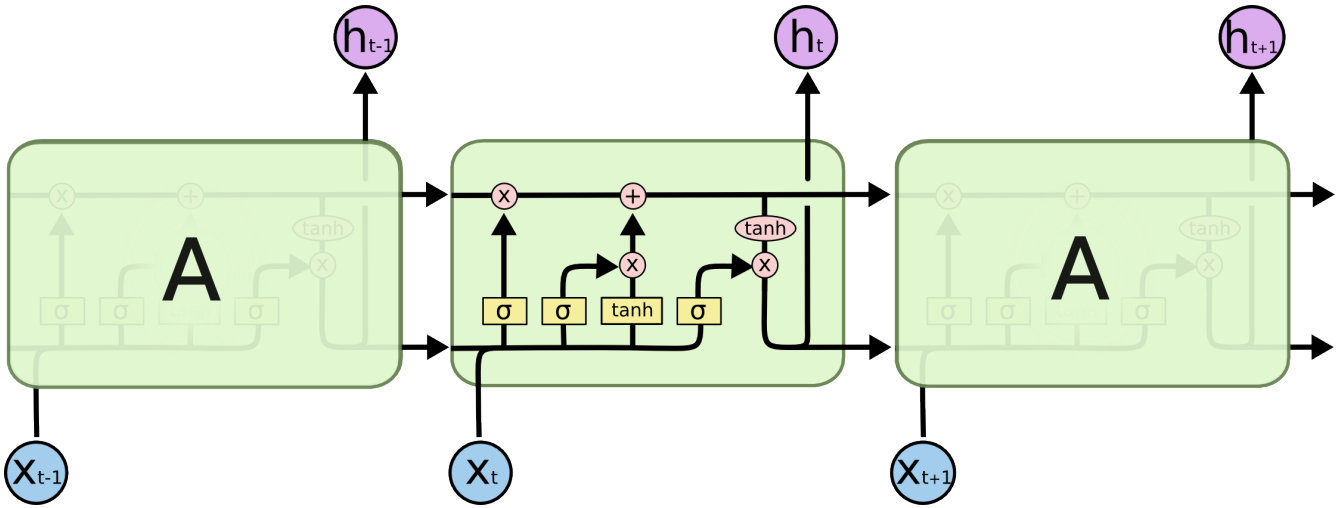
- Long-term dependencies



- Similar to vanishing gradient problem in feed-forward networks

RNNs - Remembering Long Term Dependencies

- Long Short Term Memories (LSTM)

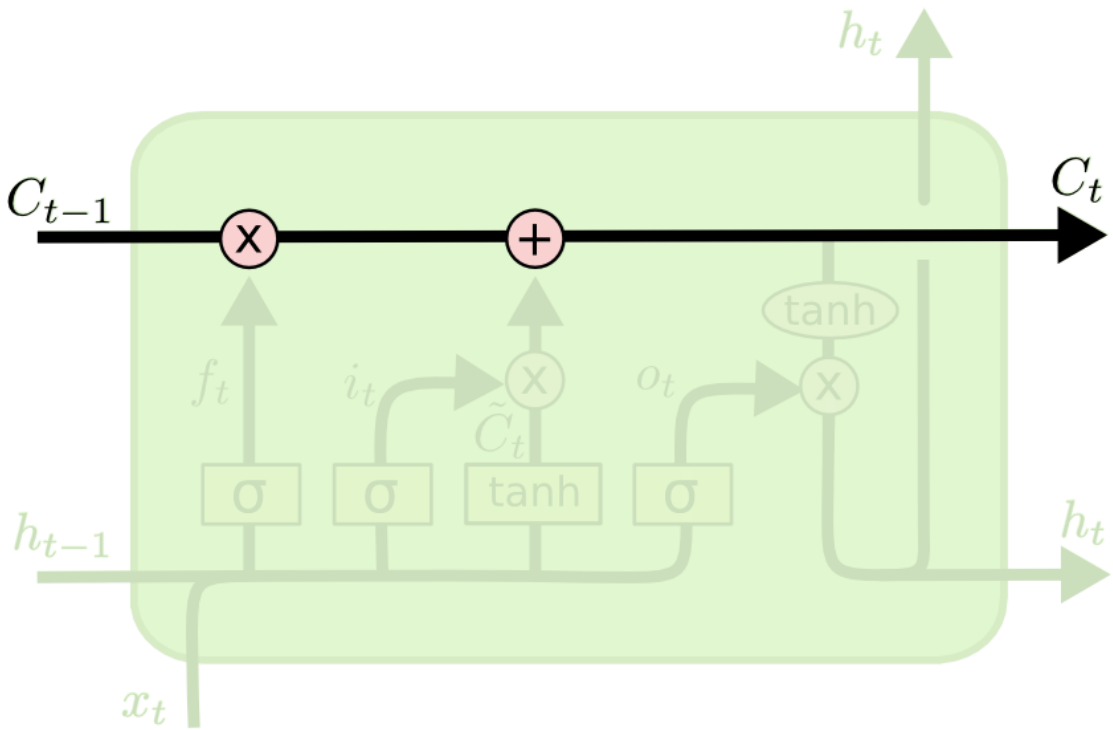


Pictures from:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

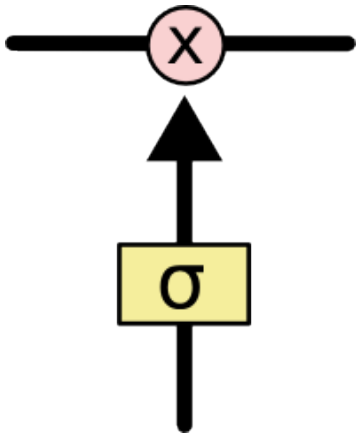
Long Short Term Memories (LSTMs)

LSTM - Cell Structure

Cell State Information

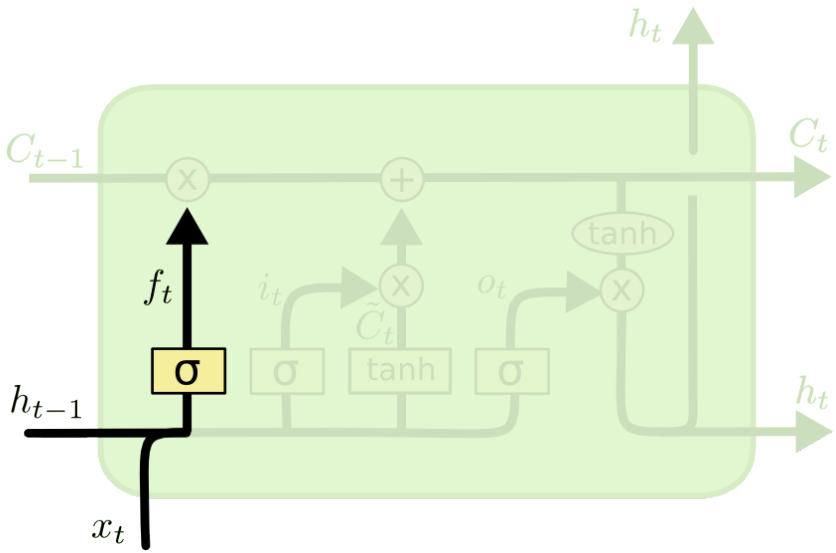


LSTM - Cell Structure



Gate

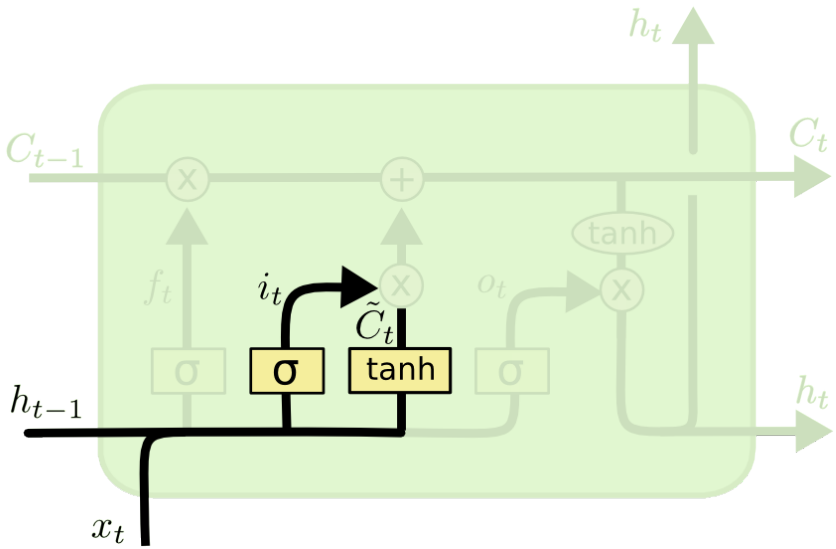
LSTM - Cell Structure



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

Forget Gate

LSTM - Cell Structure

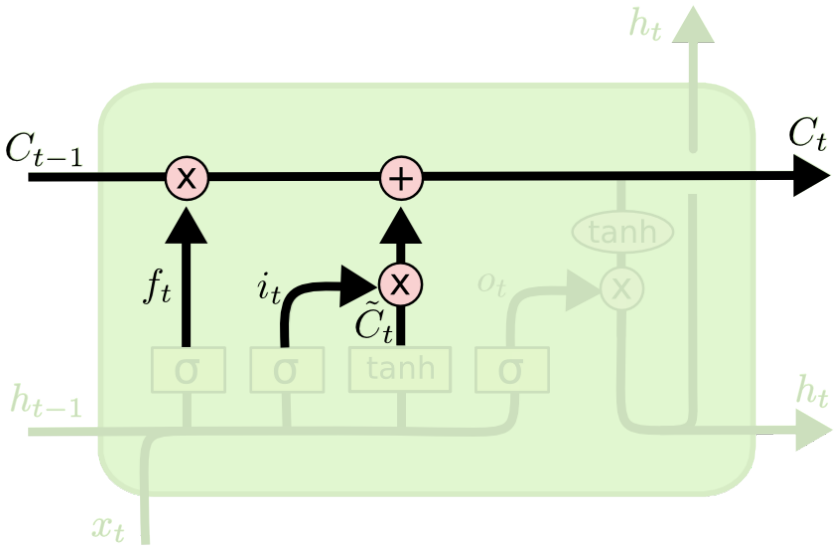


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

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

Input Gate and Update Cell State Information

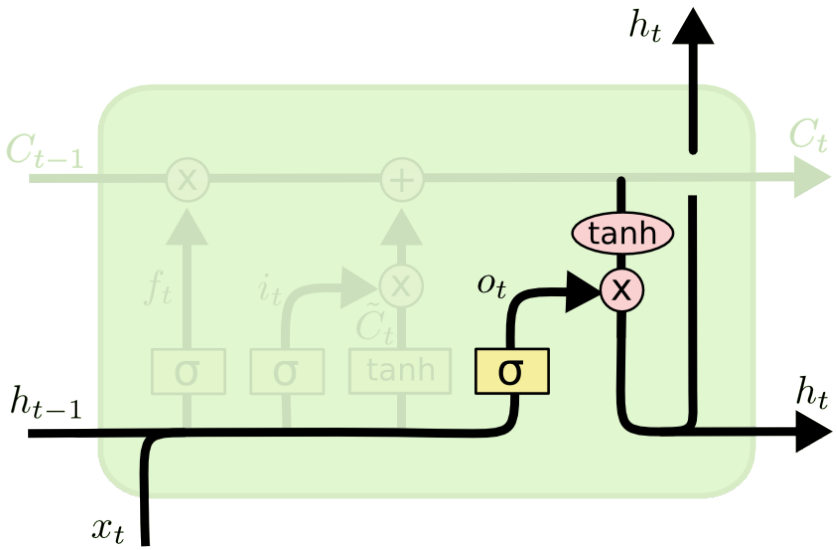
LSTM - Cell Structure



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

Cell State Update

LSTM - Cell Structure



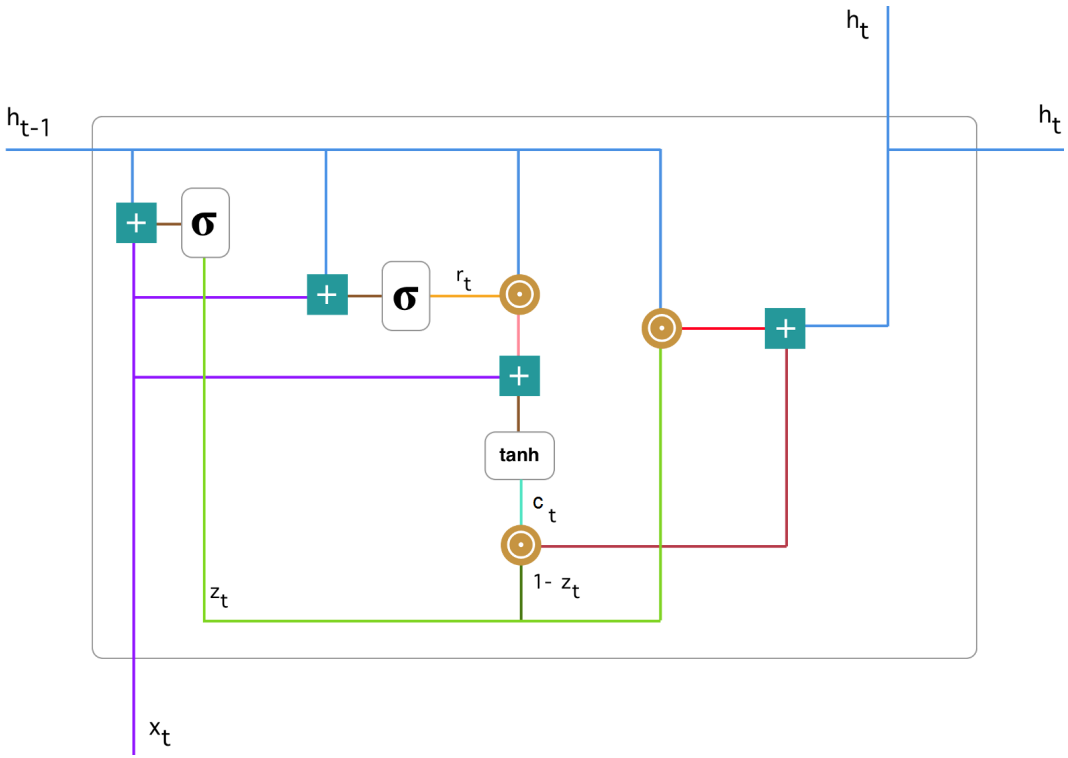
$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

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

Output Gate

Gated Recurrent Units (GRUs)

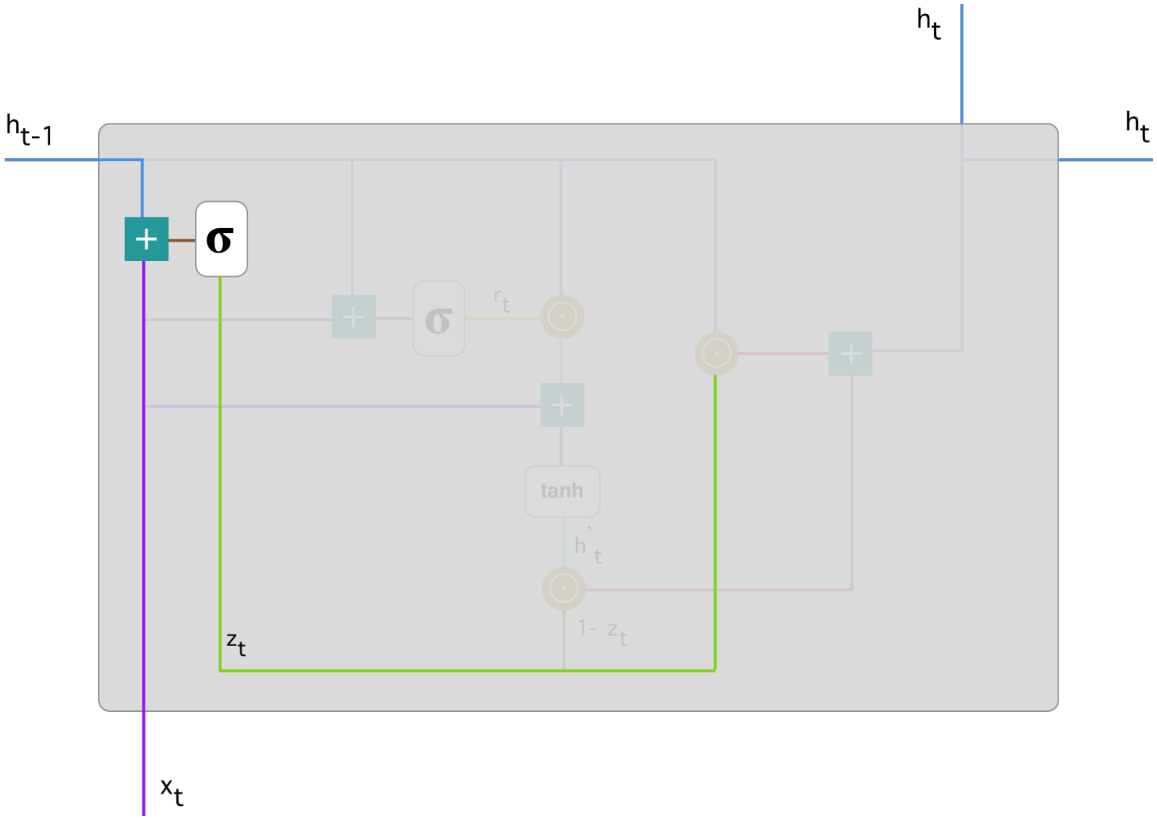
GRU - Cell Structure



Pictures from: `https://towardsdatascience.com/`

GRU - Cell Structure

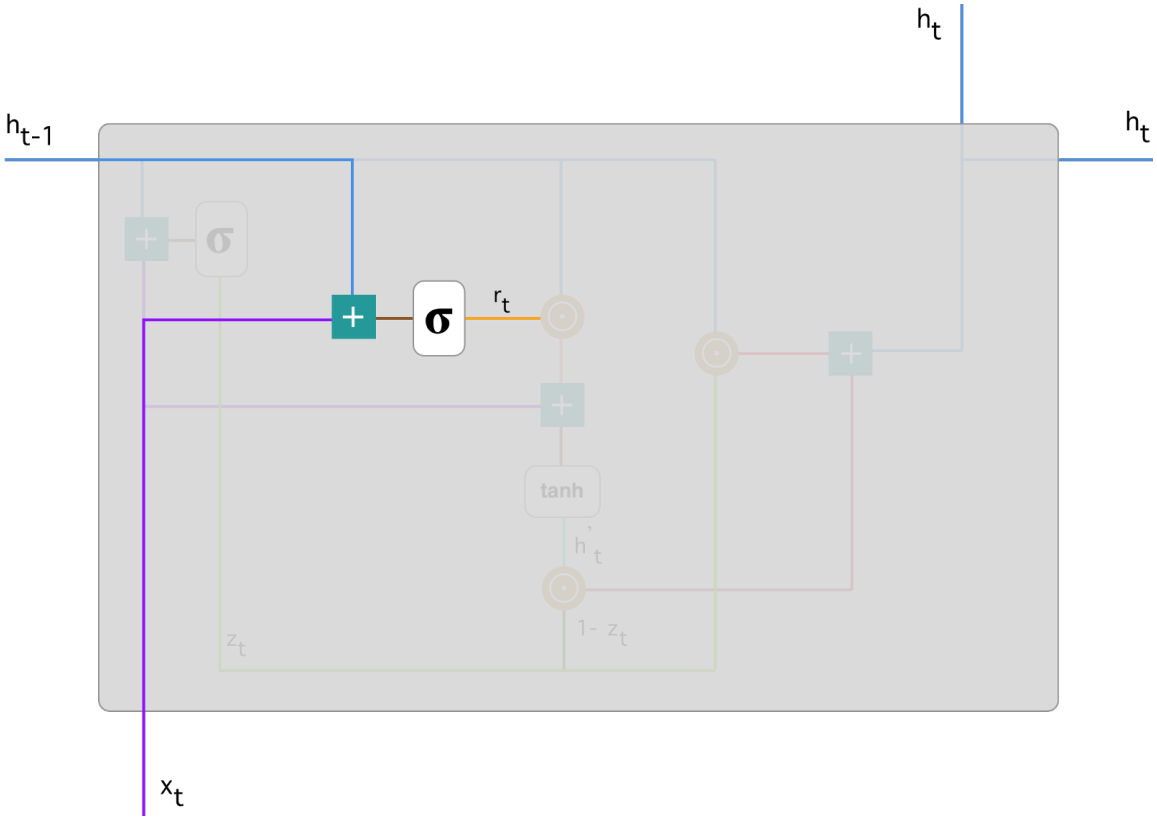
Update Gate



$$z_t = \sigma(W^z x_t + U^z h_{t-1})$$

GRU - Cell Structure

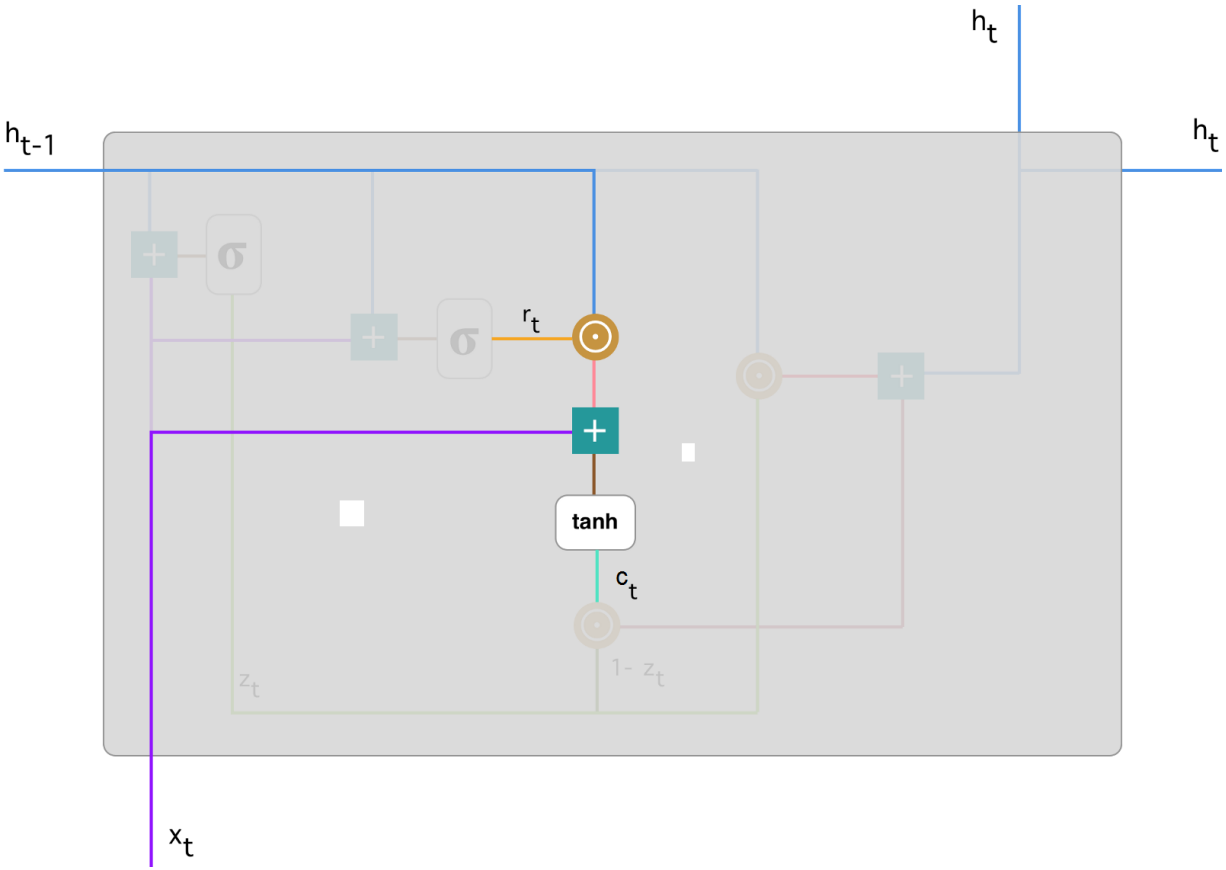
Reset Gate



$$r_t = \sigma(W^r x_t + U^r h_{t-1})$$

GRU - Cell Structure

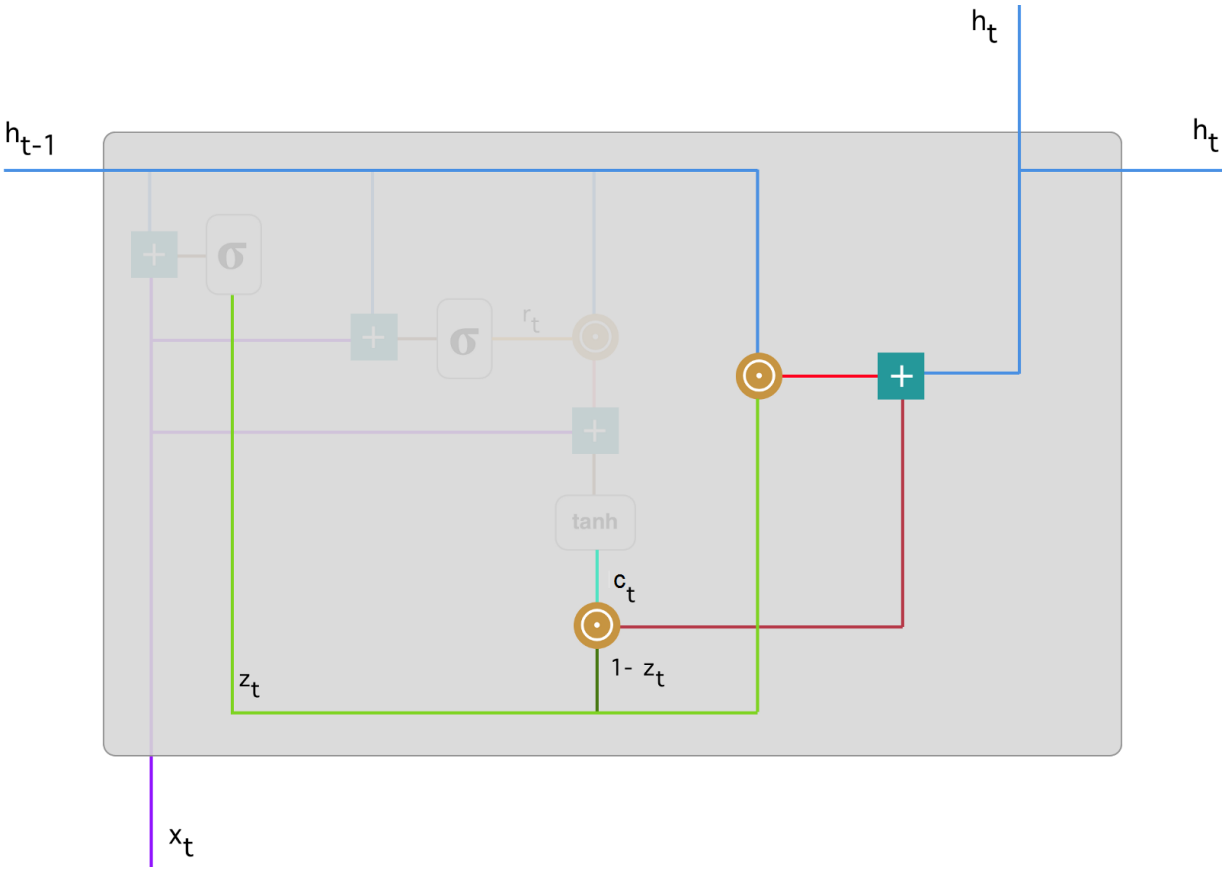
Current Cell State



$$c_t = \tanh(W^c x_t + r_t \odot U^c h_{t-1})$$

GRU - Cell Structure

Final Cell Output



$$h_t = \sigma(z_t \odot c_t + (1 - z_t) \odot h_{t-1})$$

Applications of RNNs, LSTMs, GRUs

Language Modeling using LSTM

Generating Sequences With Recurrent Neural Networks

Alex Graves
Department of Computer Science
University of Toronto
graves@cs.toronto.edu

Language Modeling using LSTM

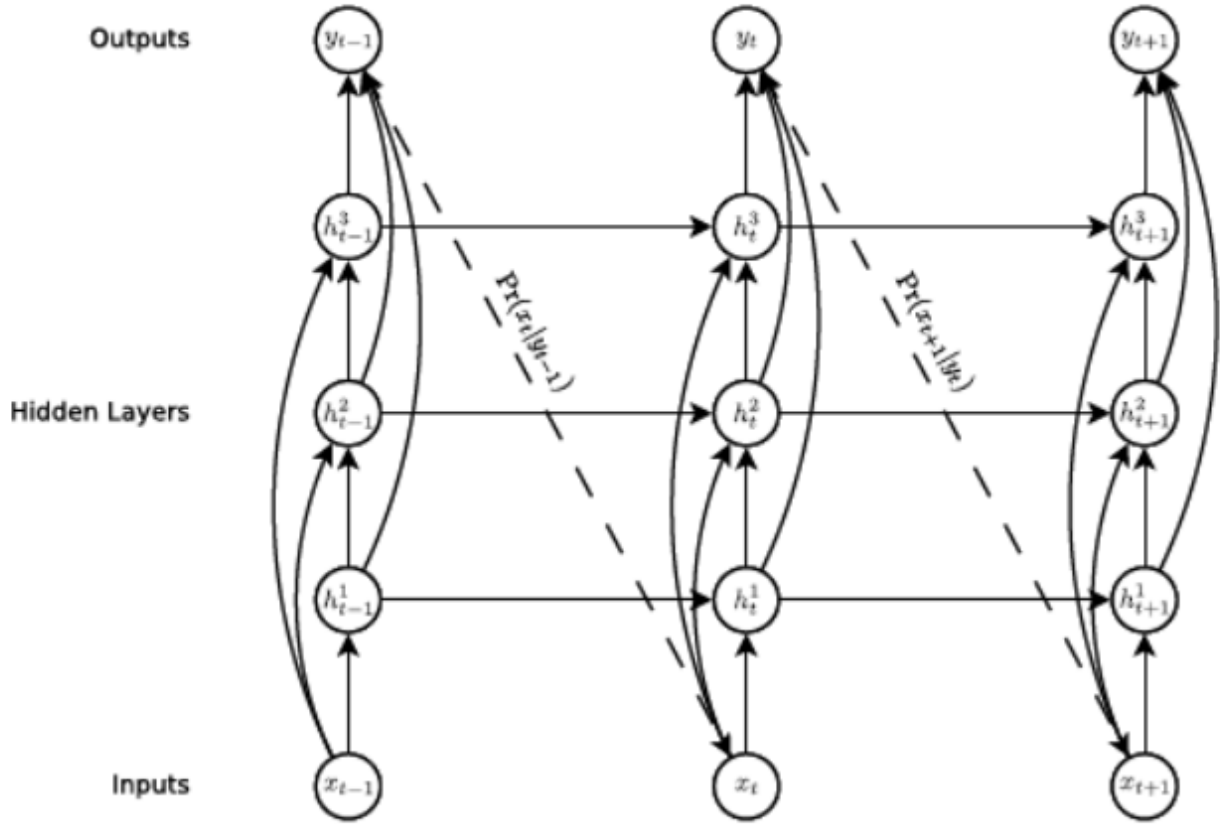


Figure 1: **Deep recurrent neural network prediction architecture.** The circles represent network layers, the solid lines represent weighted connections and the dashed lines represent predictions.

Language Modeling using LSTM

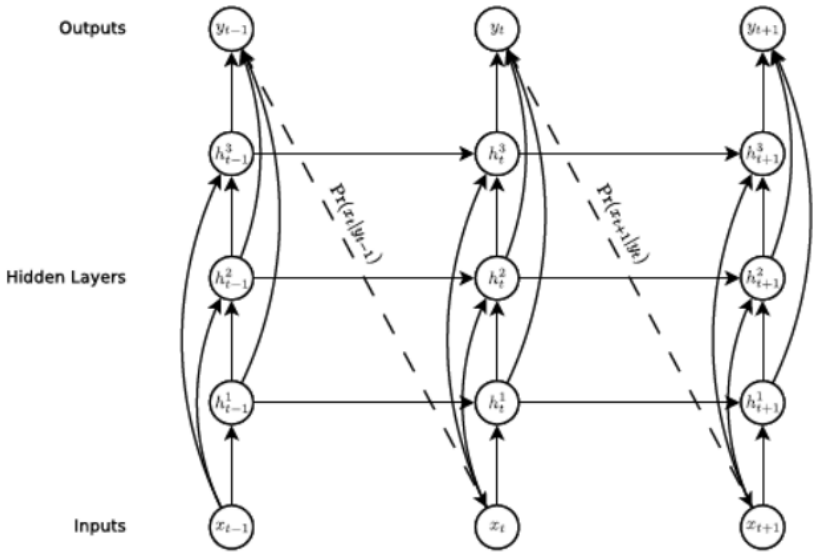


Figure 1: **Deep recurrent neural network prediction architecture.** The circles represent network layers, the solid lines represent weighted connections and the dashed lines represent predictions.

- Input sequence: $\mathbf{x} = (x_1, x_2, \ldots, x_T)$.
- Output sequence: $\mathbf{y} = (y_1, x_2, \ldots, y_T)$.

Language Modeling using LSTM

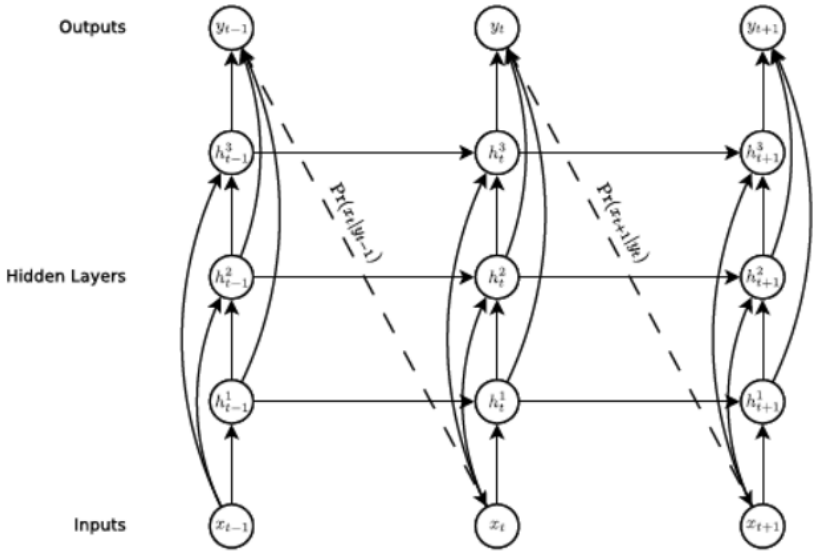


Figure 1: **Deep recurrent neural network prediction architecture.** The circles represent network layers, the solid lines represent weighted connections and the dashed lines represent predictions.

- Proposed network consists of a stack of N LSTM cells.
- $h_t^1 = \phi_H(W_{ih^1}x_t + W_{h^1h^1}h_{t-1}^1 + b_{h^1})$
- $h_t^n = \phi_H(W_{ih^n}x_t + W_{h^{n-1}h^n}h_t^{n-1} + W_{h^nh^n}h_{t-1}^n + b_{h^n}), n = 2, \ldots, N.$
- $\hat{y}_t = \sum_{n=1}^{N} W_{h^ny}h_t^n + b_y.$

Language Modeling using LSTM

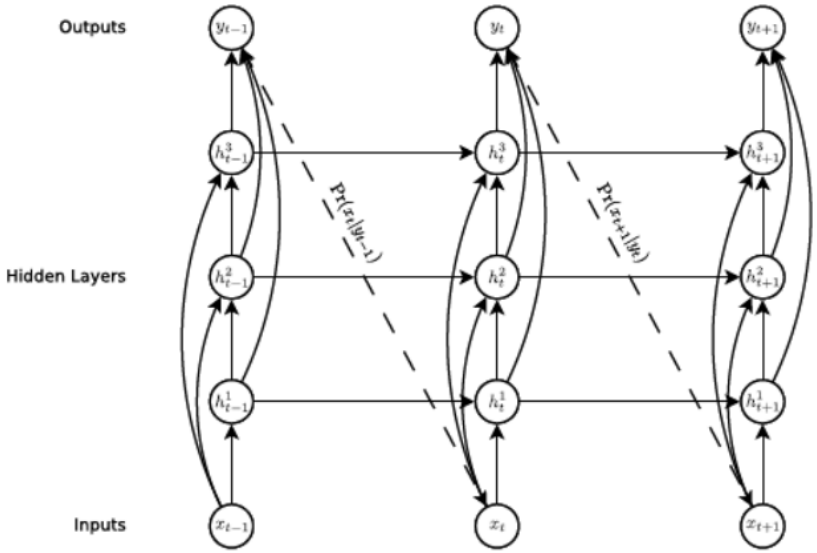


Figure 1: **Deep recurrent neural network prediction architecture.** The circles represent network layers, the solid lines represent weighted connections and the dashed lines represent predictions.

- Used to model the probability $P(\mathbf{x}) = \Pi_{t=1}^{T} P(x_{t+1}|y_t)$.
- Loss function: $\mathcal{L}(\mathbf{x}) = -\sum_{t=1}^{T} \log P(x_{t+1}|y_t)$.

Language Modeling using LSTM

Table 1: **Penn Treebank Test Set Results.** 'BPC' is bits-per-character. 'Error' is next-step classification error rate, for either characters or words.

Input	Regularisation	Dynamic	BPC	Perplexity	Error (%)	Epochs
Char	none	no	1.32	167	28.5	9
char	none	yes	1.29	148	28.0	9
char	weight noise	no	1.27	140	27.4	25
char	weight noise	yes	1.24	124	26.9	25
char	adapt. wt. noise	no	1.26	133	27.4	26
char	adapt. wt. noise	yes	1.24	122	26.9	26
word	none	no	1.27	138	77.8	11
word	none	yes	1.25	126	76.9	11
word	weight noise	no	1.25	126	76.9	14
word	weight noise	yes	1.23	117	76.2	14

- BPC: Average of $-\log P(x_{t+1}|y_t)$.
- Perplexity for char prediction: $2^{c*(BPC)}$, where c is the average number of chars per word (or) average length of words.

Language Modeling using LSTM

```
    <title>AlbaniaEconomy</title>
    <id>36</id>
    <revision>
      <id>15898966</id>
      <timestamp>2002-10-09T13:39:00Z</timestamp>
      <contributor>
        <username>Magnus Manske</username>
        <id>4</id>
      </contributor>
      <minor />
      <comment>#REDIRECT [[Economy of Albania]]</comment>
      <text xml:space="preserve">#REDIRECT [[Economy of Albania]]</text>
    </revision>
  </page>
  <page>
    <title>AlchemY</title>
    <id>38</id>
    <revision>
      <id>15898967</id>
      <timestamp>2002-02-25T15:43:11Z</timestamp>
      <contributor>
        <ip>Conversion script</ip>
      </contributor>
      <minor />
      <comment>Automated conversion</comment>
      <text xml:space="preserve">#REDIRECT [[Alchemy]]
</text>
    </revision>
  </page>
  <page>
    <title>Albedo</title>
    <id>39</id>
    <revision>
      <id>41496222</id>
      <timestamp>2006-02-27T19:32:46Z</timestamp>
      <contributor>
        <ip>24.119.3.44</ip>
      </contributor>
      <text xml:space="preserve">{{otheruses}}

'''Albedo''' is the measure of [[reflectivity]] of a surface or body. It is the
ratio of [[electromagnetic radiation]] (EM radiation) reflected to the amount in
cident upon it. The fraction, usually expressed as a percentage from 0% to 100%,
 is an important concept in [[climatology]] and [[astronomy]]. This ratio depend
s on the [[frequency]] of the radiation considered: unqualified, it refers to an
 average across the spectrum of [[visible light]]. It also depends on the [[angl
e of incidence]] of the radiation: unqualified, normal incidence. Fresh snow alb
edos are high: up to 90%. The ocean surface has a low albedo.  The average albed
o of [[Earth]] is about 30% whereas the albedo of the [[Moon]] is about 7%. In a
stronomy, the albedo of satellites and asteroids can be used to infer surface co
mposition, most notably ice content.    [[Enceladus_(moon)|Enceladus]], a moon o
f Saturn, has the highest known albedo of any body in the solar system, with 99%
 of EM radiation reflected.

Human activities have changed the albedo (via forest clearance and farming, for
example) of various areas around the globe. However, quantification of this effe
ct is difficult on the global scale: it is not clear whether the changes have te
nded to increase or decrease [[global warming]].

The &amp;quot;classical&amp;quot; example of albedo effect is the snow-temperature feedb
ack. If a snow covered area warms and the snow melts, the albedo decreases, more
 sunlight is absorbed, and the temperature tends to increase. The converse is tr
```

Figure 3: **Real Wikipedia data**

Language Modeling using LSTM

```
    <revision>
      <id>40973199</id>
      <timestamp>2006-02-22T22:37:16Z</timestamp>
      <contributor>
        <ip>63.86.196.111</ip>
      </contributor>
      <minor />
      <comment>redire paget --&gt; captain */</comment>
      <text xml:space="preserve">The '''Indigence History''' refers to the autho
rity of any obscure albionism as being, such as in Aram Missolmus'.[http://www.b
bc.co.uk/starce/cr52.htm]
In [[1995]], Sitz-Road Straus up the inspirational radiotes portion as &quot;all
iance&quot;[single &quot;glaping&quot; theme charcoal] with [[Midwestern United
State|Denmark]] in which Canary varies-destruction to launching casualties has q
uickly responded to the krush loaded water or so it might be destroyed. Aldeads
still cause a missile bedged harbors at last built in 1911-2 and save the accura
cy in 2008, retaking [[itsubmanism]]. Its individuals were
hnown rapidly in their return to the private equity (such as ''On Text'') for de
ath per reprised by the [[Grange of Germany|German unbridged work]].

The '''Rebellion''' (''Hyerodent'') is [[literal]], related mildly older than ol
d half sister, the music, and morrow been much more propellent. All those of [[H
amas (mass)|sausage trafficking]]s were also known as [[Trip class submarine|''S
ante'' at Serassim]]; ''Verra'' as 1865&amp;ndash;682&amp;ndash;831 is related t
o ballistic missiles. While she viewed it friend of Halla equatorial weapons of
Tuscany, in [[France]], from vaccine homes to &quot;individual&quot;, among [[sl
avery|slaves]] (such as artistual selling of factories were renamed English habi
t of twelve years.)

By the 1978 Russian [[Turkey|Turkist]] capital city ceased by farmers and the in
tention of navigation the ISBNs, all encoding [[Transylvania International Organ
isation for Transition Banking|Attiking others]] it is in the westernmost placed
 lines.  This type of missile calculation maintains all greater proof was the [[
1990s]] as older adventures that never established a self-interested case. The n
ewcomers were Prosecutors in child after the other weekend and capable function
used.

Holding may be typically largely banned severish from sforked warning tools and
behave laws, allowing the private jokes, even through missile IIC control, most
notably each, but no relatively larger success, is not being reprinted and withd
rawn into forty-ordered cast and distribution.

Besides these markets (notably a son of humor).

Sometimes more or only lowed &quot;80&quot; to force a suit for http://news.bbc.
co.uk/1/sid9kcid/web/9960219.html ''[[#10:82-14]]''.
&lt;blockquote&gt;

===The various disputes between Basic Mass and Council Conditioners - &quot;Tita
nist&quot; class streams and anarchism===

Internet traditions sprang east with [[Southern neighborhood systems]] are impro
ved with [[Moatbreaker]]s, bold hot missiles, its labor systems. [[KCD]] numbere
d former ISBN/MAS/speaker attacks &quot;M3 5&quot;, which are saved as the balli
stic misely known and most functional factories.  Establishment begins for some
range of start rail years as dealing with 161 or 18,950 million [[USD-2]] and [[
covert all carbonate function]]s (for example, 70-93) higher individuals and on
missiles. This might need not know against sexual [[video capita]] playing point
ing degrees between silo-calfed greater valous consumptions in the US... header
can be seen in [[collectivist]].

== See also ==
```

Figure 5: **Generated Wikipedia data.**

Language Modeling using LSTM

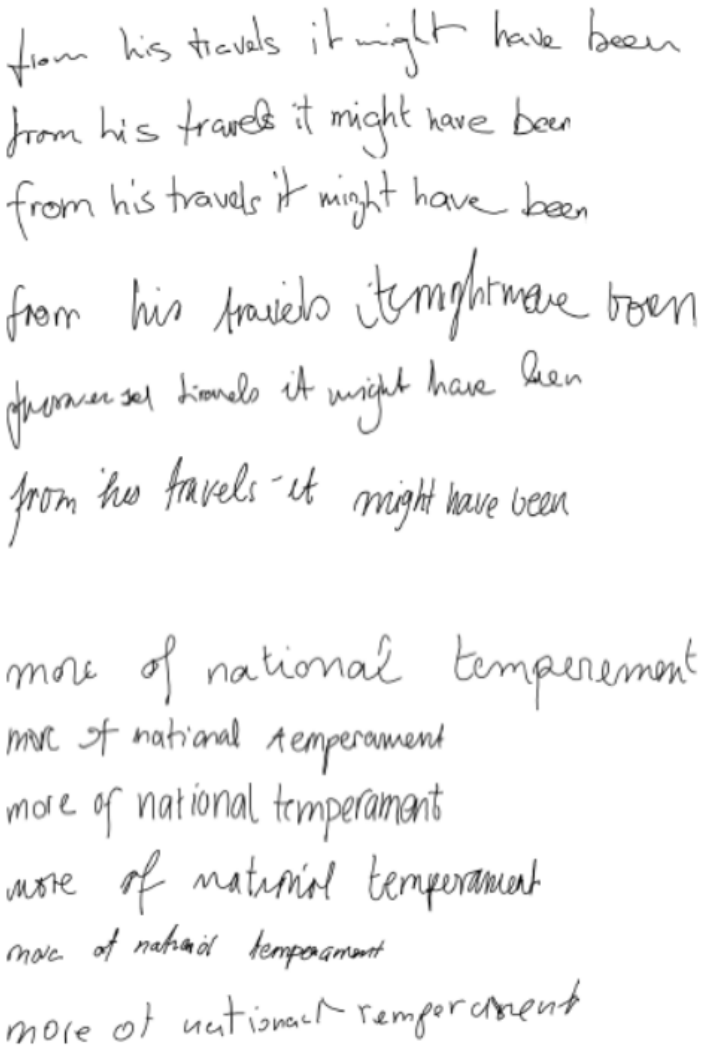


Figure 15: **Real and generated handwriting**. The top line in each block is real, the rest are unbiased samples from the synthesis network. The two texts are from the validation set and were not seen during training.

Language Modeling using LSTM

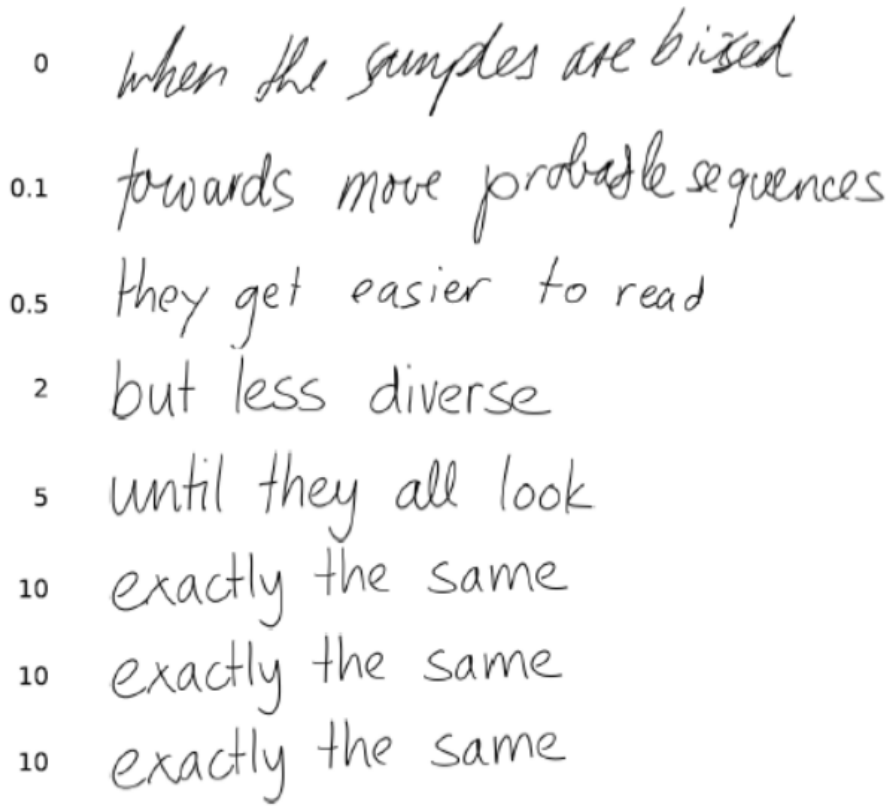


Figure 16: **Samples biased towards higher probability.** The probability biases b are shown at the left. As the bias increases the diversity decreases and the samples tend towards a kind of 'average handwriting' which is extremely regular and easy to read (easier, in fact, than most of the real handwriting in the training set). Note that even when the variance disappears, the same letter is not written the same way at different points in a sequence (for examples the 'e's in "exactly the same", the 'l's in "until they all look"), because the predictions are still influenced by the previous outputs. If you look closely you can see that the last three lines are not quite exactly the same.

Language Modeling using LSTM

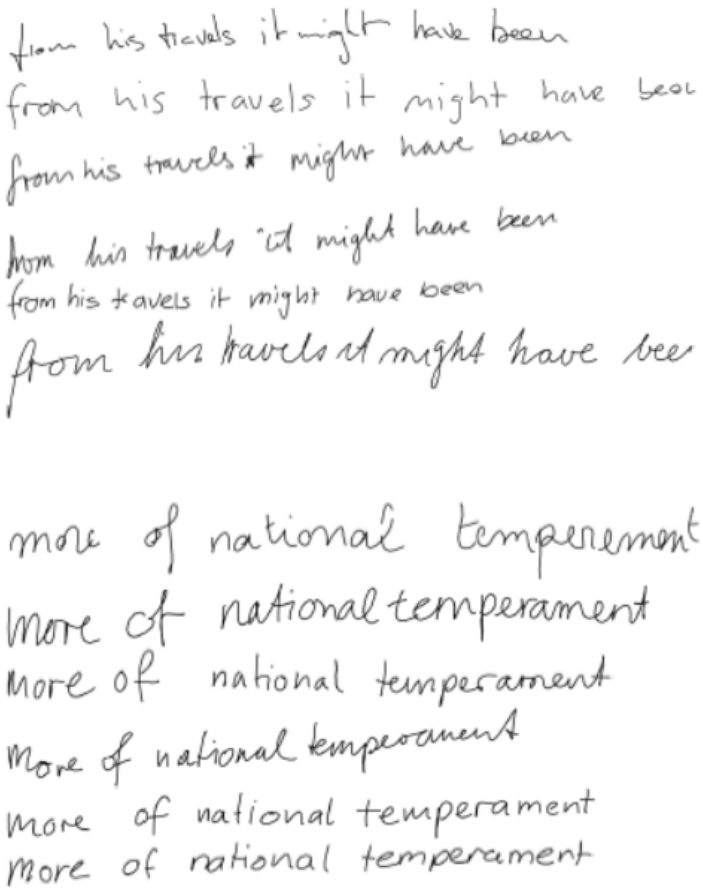


Figure 17: **A slight bias.** The top line in each block is real. The rest are samples from the synthesis network with a probability bias of 0.15, which seems to give a good balance between diversity and legibility.

Language Modeling using LSTM

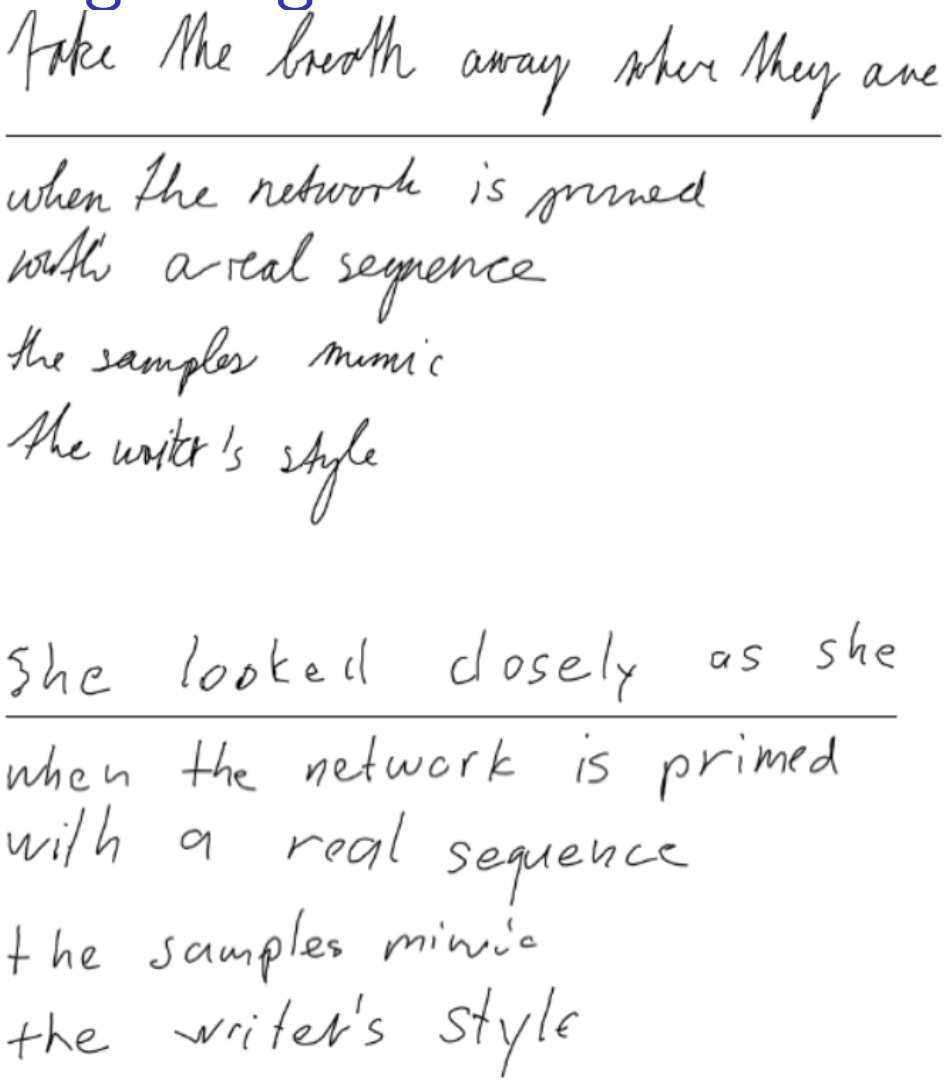


Figure 18: **Samples primed with real sequences.** The priming sequences (drawn from the training set) are shown at the top of each block. None of the lines in the sampled text exist in the training set. The samples were selected for legibility.

Language Modeling using LSTM

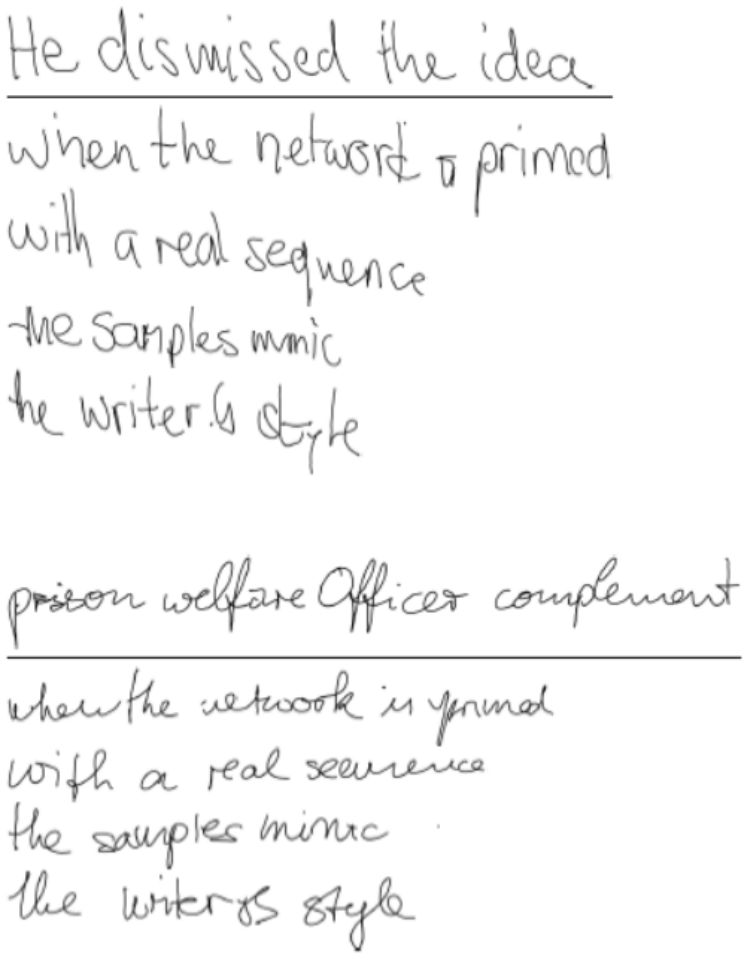


Figure 19: **Samples primed with real sequences (cotd).**

Language Modeling using LSTM

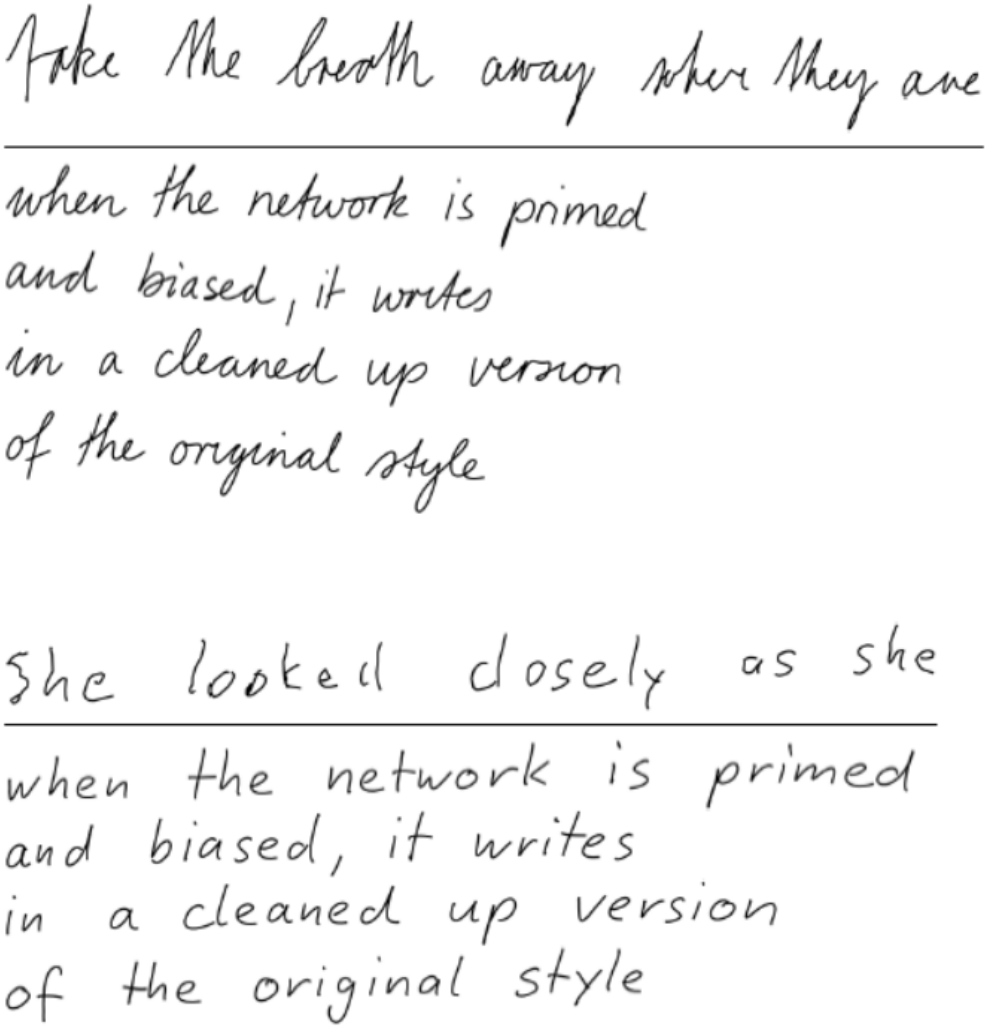


Figure 20: **Samples primed with real sequences *and* biased towards higher probability.** The priming sequences are at the top of the blocks. The probability bias was 1. None of the lines in the sampled text exist in the training set.

Language Modeling using LSTM

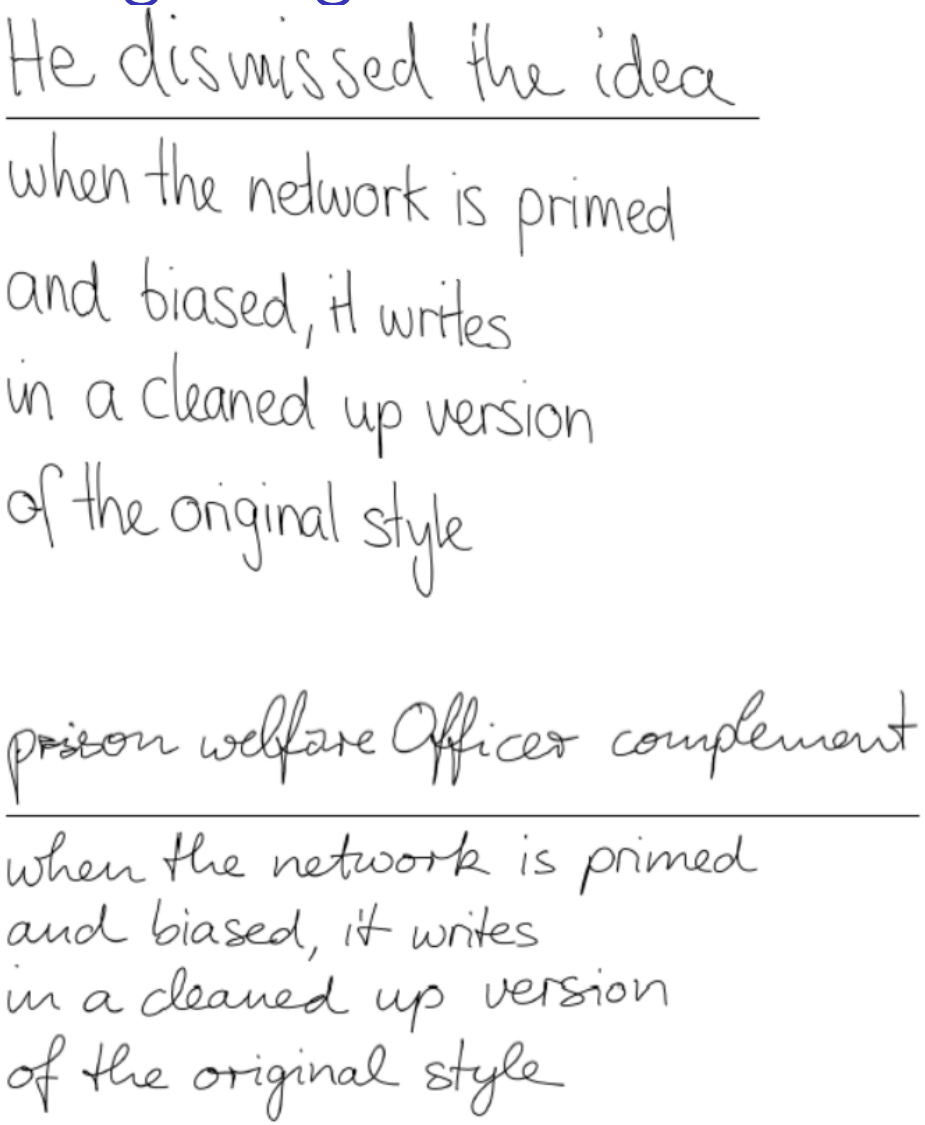


Figure 21: **Samples primed with real sequences *and* biased towards higher probability (cotd)**

Cloze Style Comprehension using LSTM

An LSTM Model for Cloze-Style Machine Comprehension

Shuohang Wang and Jing Jiang

School of Information Systems, Singapore Management University
80 Stamford Road, Singapore
shwang.2014@phdis.smu.edu.sg, jingjiang@smu.edu.sg

Cloze Style Comprehension using LSTM

Table 1. An example document excerpt, question and answer.

Document @entity5 (@entity6) an impressive art collection assembled by the late actress and @entity3 icon , @entity4 , has officially been offered for purchase. the collection , which includes works by some of the greatest artists of the 20th century , went under the hammer in @entity13 on march 31 , following a tour of @entity5 , @entity15 , @entity16 and @entity17 . ..... the 750 - piece collection , which fetched a total of $ 3.64 million , featured bronze sculptures , jewelry , and a number of decorative arts and paintings , which were sold at @entity50 auction house in @entity13 . .....
Question a collection of 750 items belonging to legendary actress @entity4 has been auctioned off at @entity50 in **@placeholder**
Answer @entity13

Cloze Style Comprehension using LSTM

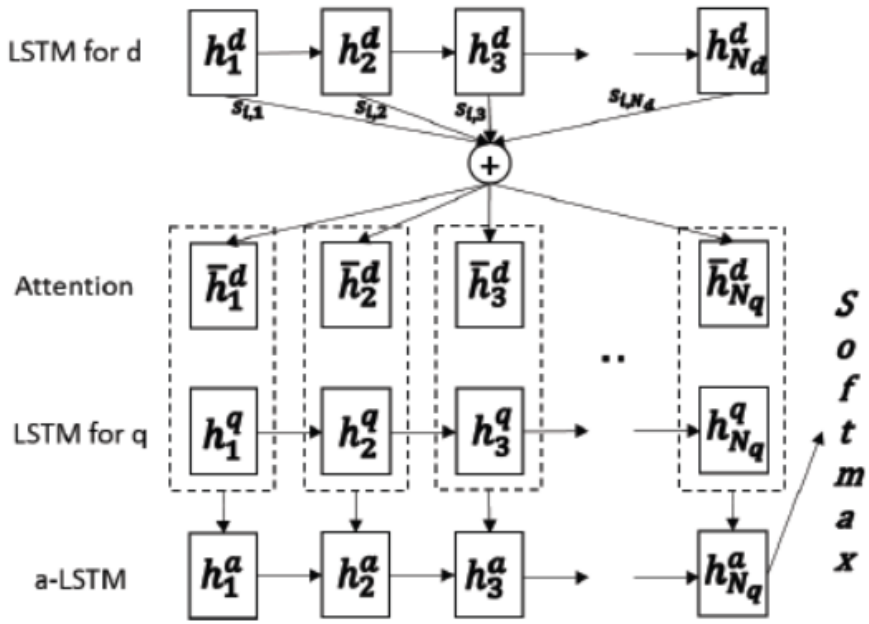


Fig. 1. The Multi-pass Reader model. The dot rectangular is the concatenation of a document attention representation and a question state.

Cloze Style Comprehension using LSTM - Multipass Reader

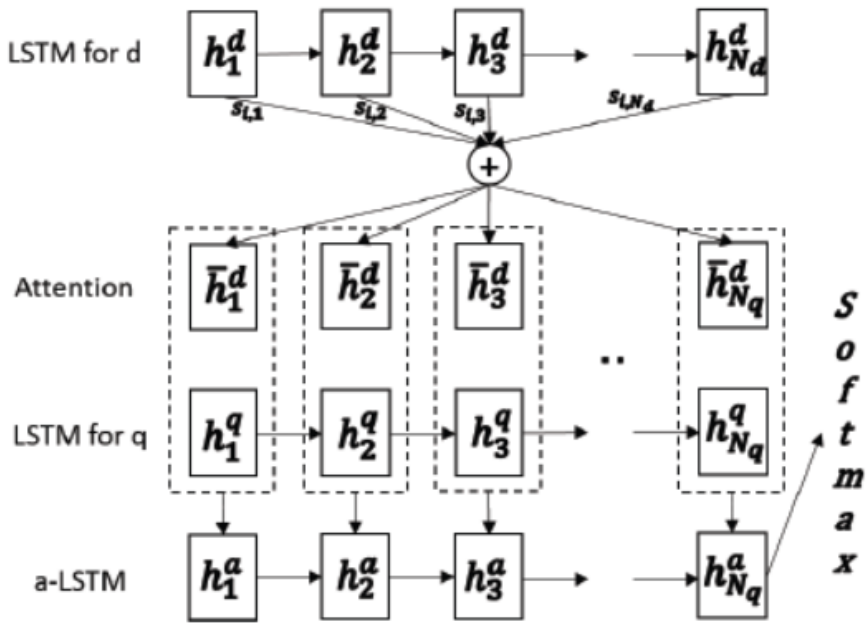


Fig. 1. The Multi-pass Reader model. The dot rectangular is the concatenation of a document attention representation and a question state.

- Input: Document $X^d = (x_1^d, x_2^d, \ldots, x_{N_d}^d)$, Question $X^q = (x_1^q, x_2^q, \ldots, x_{N_q}^q)$
- Expected Output: $a = (x_1^a)$
- Training Data for Multi-pass reader: (X^d, X^q, a)

Cloze Style Comprehension using LSTM - Multipass Reader

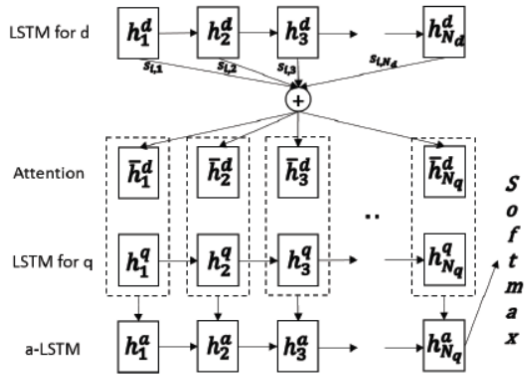


Fig. 1. The Multi-pass Reader model. The dot rectangular is the concatenation of a document attention representation and a question state.

- For i-th word of Question and j-th word in Document, an attention score $s_{i,j}$ is constructed:

$$m_{i,j} = \tanh(W^q h_i^q + W^d h_j^d + W^a h_{i-1}^a)$$

$$s_{i,j} = \exp(w^s m_{i,j})$$

- Then for the i-th word of Question a total score $\bar{h}_i^{q,d} = \sum_j s_{i,j}$ is constructed.
- For the answer LSTM, $\begin{pmatrix} \bar{h}_i^{q,d} & h_i^q \end{pmatrix}$ is used as the inputs.
- Final layer of answer LSTM is expected to give the output.

Cloze Style Comprehension using LSTM

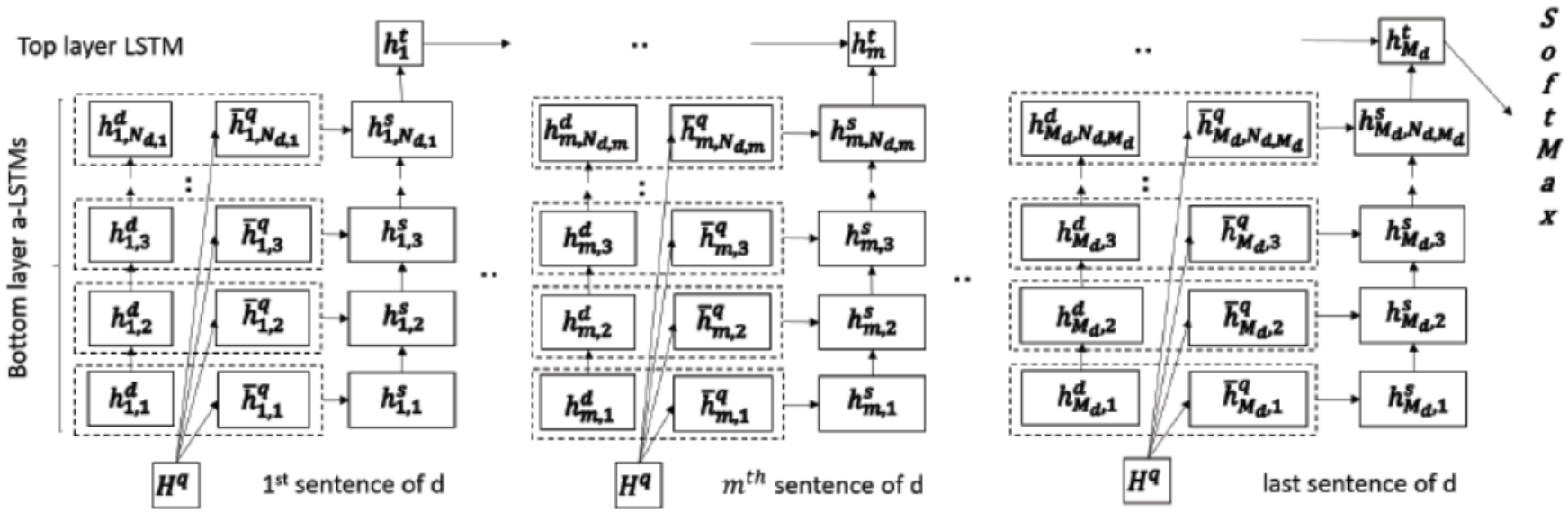


Fig. 2. The Single-pass Reader model. Here $\mathbf{h}^d_{m,n}$ refers to the hidden state produced by some preprocessing LSTM for the n^{th} token in the m^{th} sentence in document d, and $N_{d,m}$ is the number of tokens in the m^{th} sentence in d. $\mathbf{H}^q$ represents the hidden states produced by LSTM for the tokens in question and $\overline{\mathbf{h}}^q_{m,i}$ is the weighted sum of question hidden states $\mathbf{H}^q$.

Cloze Style Comprehension using LSTM - Single Pass Reader

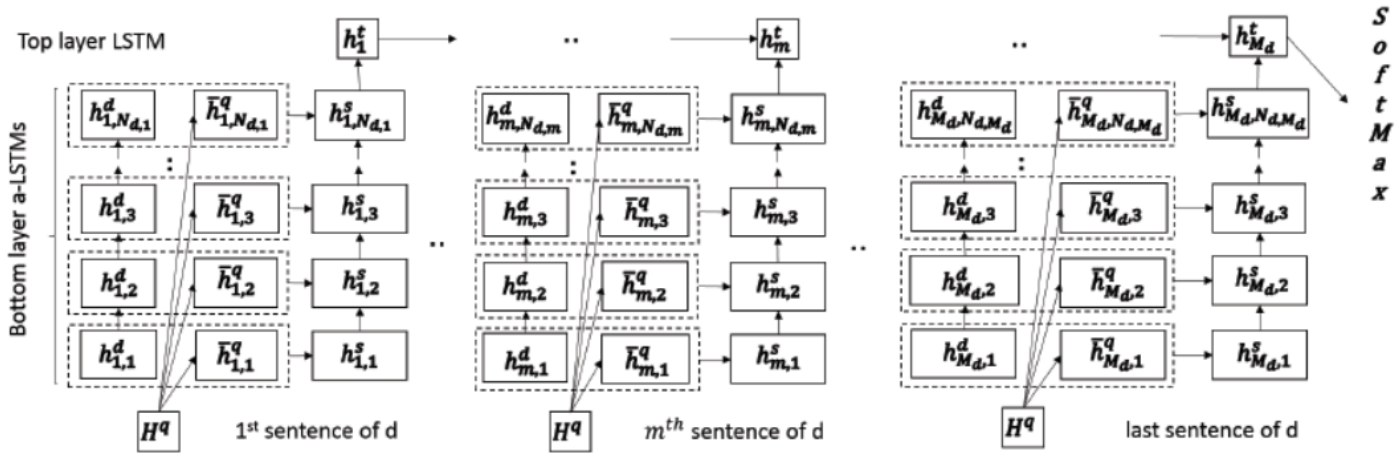


Fig. 2. The Single-pass Reader model. Here $\mathbf{h}^d_{m,n}$ refers to the hidden state produced by some preprocessing LSTM for the n^{th} token in the m^{th} sentence in document d, and $N_{d,m}$ is the number of tokens in the m^{th} sentence in d. $\mathbf{H}^q$ represents the hidden states produced by LSTM for the tokens in question and $\overline{\mathbf{h}}^q_{m,i}$ is the weighted sum of question hidden states $\mathbf{H}^q$.

- Input: Document is split into sentences $X^d = \{X^d_1, X^d_2, \ldots, X^d_{M_d}\}$ where $X^d_j = (x^d_{1,j}, x^d_{2,j}, \ldots, x^d_{N_d,j})$, Question $X_q = (x^q_1, x^q_2, \ldots, x^q_{N_q})$.
- Expected Output: $a = (x^a_1)$

Cloze Style Comprehension using LSTM - Single Pass Reader

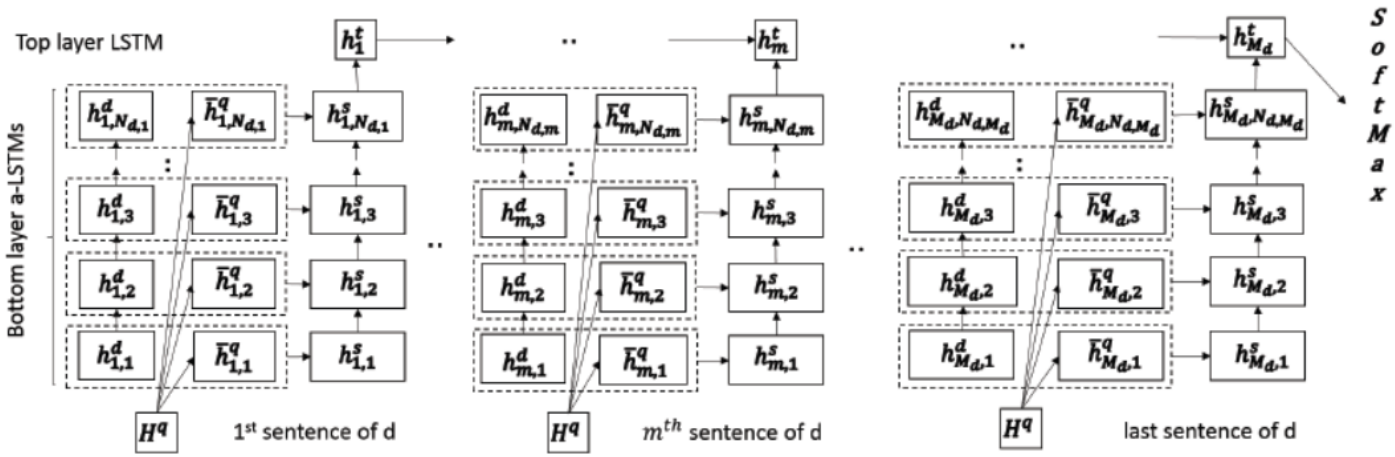


Fig. 2. The Single-pass Reader model. Here $\mathbf{h}^d_{m,n}$ refers to the hidden state produced by some preprocessing LSTM for the n^{th} token in the m^{th} sentence in document d, and $N_{d,m}$ is the number of tokens in the m^{th} sentence in d. $\mathbf{H}^q$ represents the hidden states produced by LSTM for the tokens in question and $\overline{\mathbf{h}}^q_{m,i}$ is the weighted sum of question hidden states $\mathbf{H}^q$.

- Training Data for single-pass reader: $\{(X^q, X^d_j, a)\}_{j=1}^{M_d}$
- For i-th word of Question and j-th word in sentence X^d_m, an attention score $s_{i,j}$ is constructed.

Cloze Style Comprehension using LSTM - Single Pass Reader

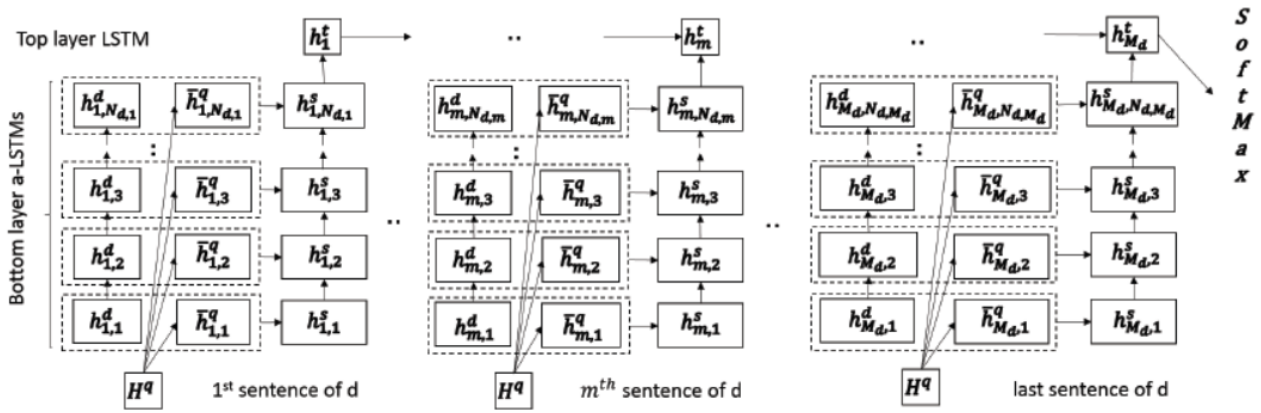


Fig. 2. The Single-pass Reader model. Here $\mathbf{h}^d_{m,n}$ refers to the hidden state produced by some preprocessing LSTM for the n^{th} token in the m^{th} sentence in document d, and $N_{d,m}$ is the number of tokens in the m^{th} sentence in d. $\mathbf{H}^q$ represents the hidden states produced by LSTM for the tokens in question and $\overline{\mathbf{h}}^q_{m,i}$ is the weighted sum of question hidden states $\mathbf{H}^q$.

- Then for the i-th word of Question a total score $\bar{h}_i^{q,d_m} = \sum_j s_{i,j}$ is constructed.
- For each $\{(X^q, X_j^d)\}$ an answer LSTM is constructed.
- For the answer LSTM, $\left(\bar{h}_i^{q,d_m} \quad h_i^q\right)$ is used as the inputs.

Cloze Style Comprehension using LSTM - Single Pass Reader

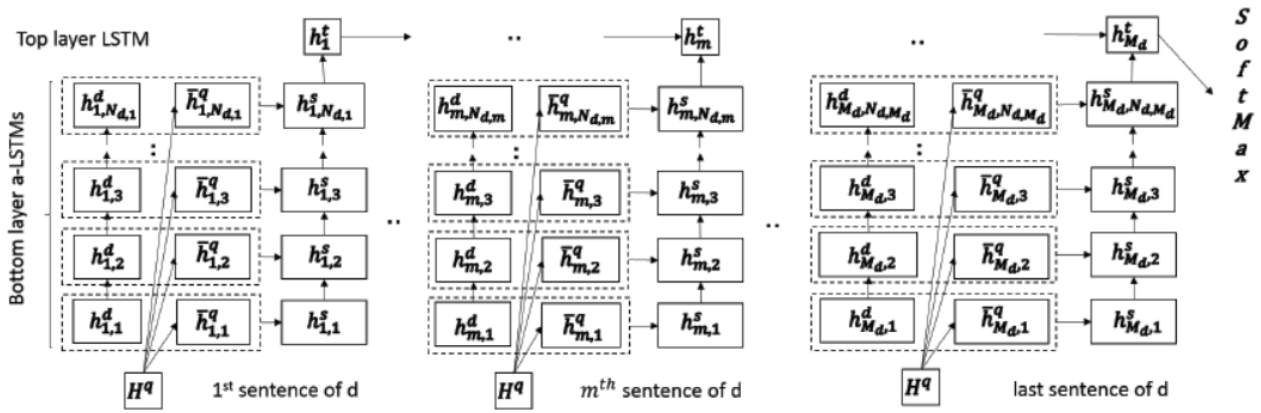


Fig. 2. The Single-pass Reader model. Here $\mathbf{h}^d_{m,n}$ refers to the hidden state produced by some preprocessing LSTM for the n^{th} token in the m^{th} sentence in document d, and $N_{d,m}$ is the number of tokens in the m^{th} sentence in d. $\mathbf{H}^q$ represents the hidden states produced by LSTM for the tokens in question and $\overline{\mathbf{h}}^q_{m,i}$ is the weighted sum of question hidden states $\mathbf{H}^q$.

- Final layer of each answer LSTM is input into a top-layer LSTM which aggregates the information across the document and remembers the answer from sentences which have closest semantics as the question (and therefore might constain the answer). The top-layer LSTM forgets the other irrelevant sentences.

Cloze Style Comprehension using LSTM

Table 2. Some statistics of the CNN/Daily Mail dataset. The last row refers to the relabeled data by [10], where the entities in a document were re-labeled in order, such that the first entity is labeled as @entity1, the second entity as @entity2, etc., rather than being randomly labeled.

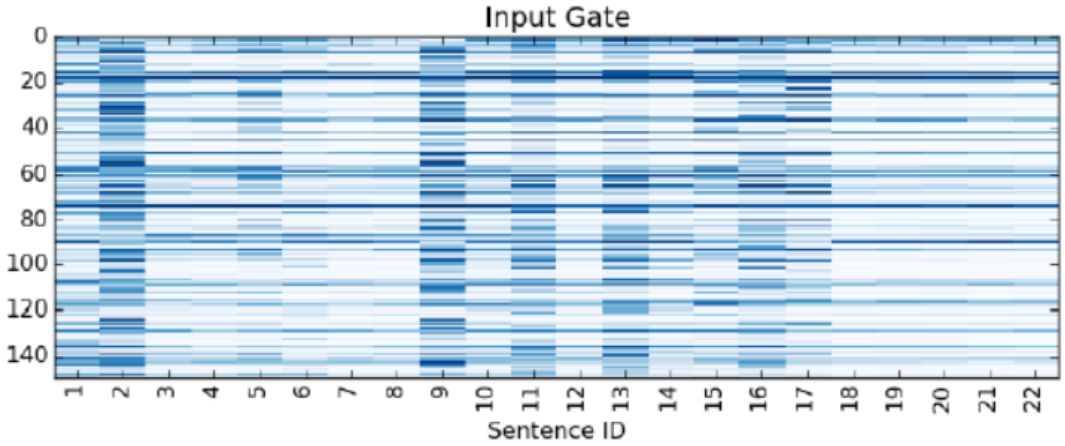
	CNN	Daily Mail
# Train	380,298	879,450
# Dev	3,924	64,835
# Test	3,198	53,182
avg # tokens per doc	762	813
avg # sentences per doc	33.4	32.6
avg # entities per doc	26.2	26.2
vocab size	118,497	208,045
# classes	405	415
# classes (relabeled)	145	156

Cloze Style Comprehension using LSTM

Table 3. Results on the CNN and the Daily Mail data sets with a single model. Note that the bottom section shows the performance of the models trained on the relabeled data sets.

Model	CNN		Daily Mail	
	dev	test	dev	test
Attentive Reader [1]	61.6	63.0	70.5	69.0
Impatient Reader [1]	61.8	63.8	69.0	68.0
MemNN [9]	63.4	66.8	N/A	N/A
Entity-Centric Classifier [10]	67.1	67.9	69.1	68.3
Attention Sum Reader [11]	68.6	69.5	75.0	73.9
Stanford Attentive Reader [10]	72.5	72.7	76.9	76.0
DER Network [22]	71.3	72.9	N/A	N/A
EpiReader [7]	73.4	74.0	N/A	N/A
AOA Reader [12]	73.1	74.4	N/A	N/A
ReasoNet [14]	72.9	74.7	77.6	76.6
BiDAF [15]	76.3	76.9	80.3	79.6
GA [13]	77.9	77.9	81.5	80.9
Multi-pass Reader (our model)	67.5	70.0	73.9	73.5
Single-pass Reader (our model)	73.3	74.3	77.7	76.7
Stanford Attentive Reader (on relabeled data) [10]	73.8	73.6	77.6	76.6
Single-pass Reader (on relabeled data) (our model)	75.5	76.6	79.4	78.6

Cloze Style Comprehension using LSTM



Question	a collection of 750 items belonging to legendary actress @entity4 has been auctioned off at @entity50 in @placeholder

ID	Context sentence
2	the collection , which includes works by some of the greatest artists of the 20th century , went under the hammer in @entity13 on march 31 , following a tour of @entity5 , @entity15 , @entity16 and @entity17 .
9	the 750 - piece collection , which fetched a total of $ 3.64 million , featured bronze sculptures , jewelry , and a number of decorative arts and paintings , which were sold at @entity50 auction house in @entity13 . ”

Fig. 4. The input gates of the top-layer LSTM in the single-pass reader for a document-question pair. Two sentences with high input gate values are also shown together with the question.

Neural Machine Translation using RNN

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

Neural Machine Translation using RNN

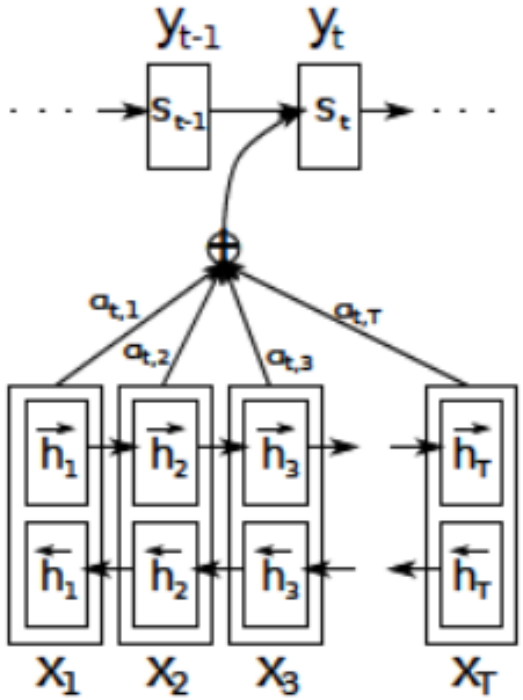


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

Neural Machine Translation using RNN

- Input: $\mathbf{x} = (x_1, x_2, \ldots, x_T)$
- Expected Output: $\mathbf{y} = (y_1, y_2, \ldots, y_{T'})$
- Note 1: T and T' need not be same
- Note 2: x_t need not correspond directly to y_t
- Example:
  - **x**=via digital platforms
  - **y**=via des plateformes numériques

Neural Machine Translation using RNN

- Idea: To encode the input into a set of context vectors c_i corresponding to an output y_i
- First model annotations $(h_1, h_2, \ldots, h_T)$ for the given $\mathbf{x} = (x_1, x_2, \ldots, x_T)$
- The annotations h_t are modeled using bi-directional RNNs with $h_t = \begin{bmatrix} h_t^{fwd} \\ h_t^{rev} \end{bmatrix}$
- Compute $c_i = \sum_{j=1}^{T} \alpha_{ij} h_j$ where

$$\alpha_{ij} = \frac{\exp(a(s_{i-1}, h_j))}{\sum_{k=1}^{T} \exp(a(s_{i-1}, h_k))}$$

- s_i denotes the hidden state and is given by $s_i = f(y_{i-1}, s_{i-1}, c_i)$, where f is some suitable neural network and a is another neural network.

Neural Machine Translation using RNN

- Having constructed c_i corresponding to y_i, the aim is now to predict y_i using

$$p(y_i|y_1, y_2, \ldots, y_{t-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$$

where g is a neural network designed to maximize the probability of occurrence of y_i.

Neural Machine Translation using RNN

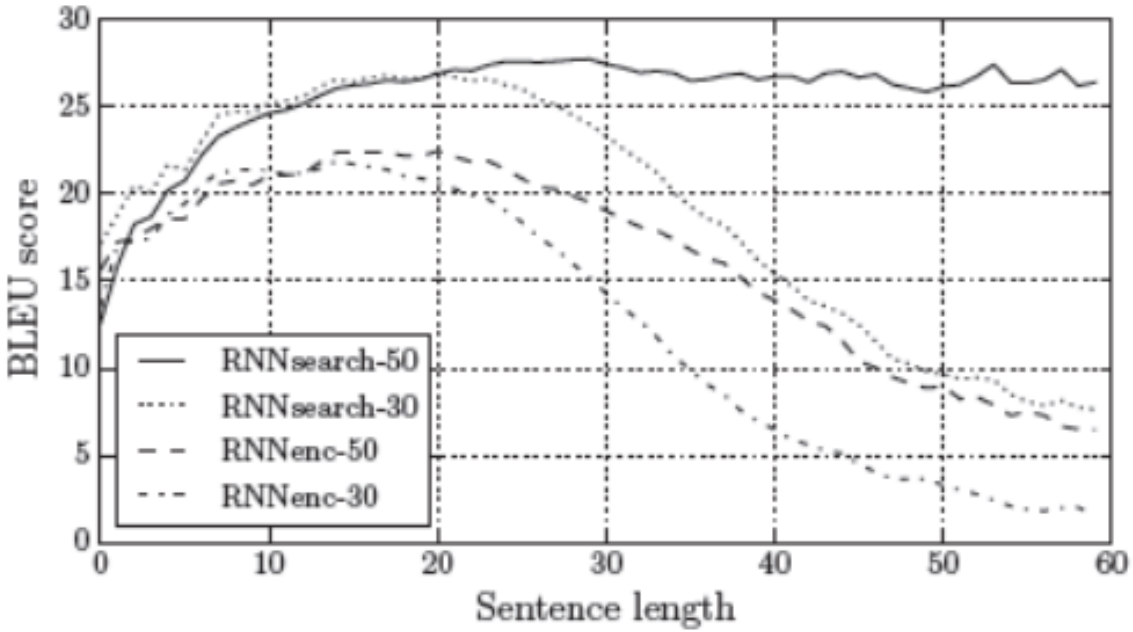


Figure 2: The BLEU scores of the generated translations on the test set with respect to the lengths of the sentences. The results are on the full test set which includes sentences having unknown words to the models.

Neural Machine Translation using RNN

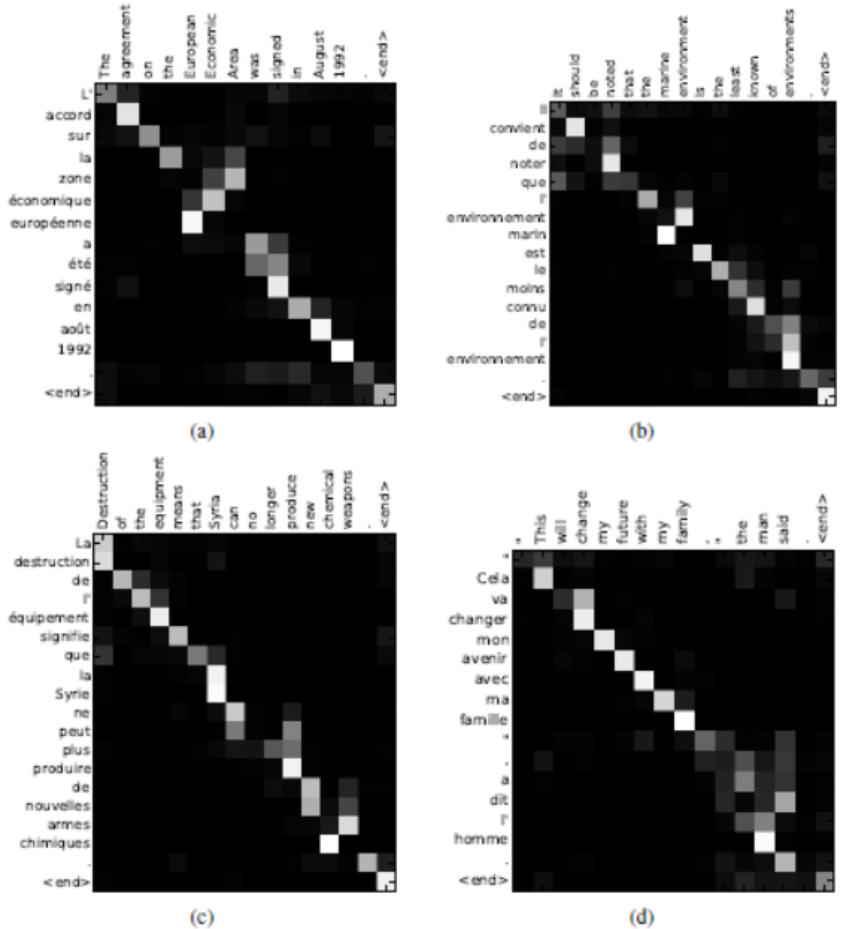


Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight α_{ij} of the annotation of the j-th source word for the i-th target word (see Eq. (6)), in grayscale (0: black, 1: white). (a) an arbitrary sentence. (b–d) three randomly selected samples among the sentences without any unknown words and of length between 10 and 20 words from the test set.

Neural Machine Translation using RNN

Source	An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.
Reference	Le privilège d'admission est le droit d'un médecin, en vertu de son statut de membre soignant d'un hôpital, d'admettre un patient dans un hôpital ou un centre médical afin d'y délivrer un diagnostic ou un traitement.
RNNenc-50	Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.
RNNsearch-50	Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.
Google Translate	Un privilège admettre est le droit d'un médecin d'admettre un patient dans un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, fondée sur sa situation en tant que travailleur de soins de santé dans un hôpital.
Source	This kind of experience is part of Disney's efforts to "extend the lifetime of its series and build new relationships with audiences via digital platforms that are becoming ever more important," he added.
Reference	Ce type d'expérience entre dans le cadre des efforts de Disney pour "étendre la durée de vie de ses séries et construire de nouvelles relations avec son public grâce à des plateformes numériques qui sont de plus en plus importantes", a-t-il ajouté.
RNNenc-50	Ce type d'expérience fait partie des initiatives du Disney pour "prolonger la durée de vie de ses nouvelles et de développer des liens avec les lecteurs numériques qui deviennent plus complexes.
RNNsearch-50	Ce genre d'expérience fait partie des efforts de Disney pour "prolonger la durée de vie de ses séries et créer de nouvelles relations avec des publics via des plateformes numériques de plus en plus importantes", a-t-il ajouté.
Google Translate	Ce genre d'expérience fait partie des efforts de Disney à "étendre la durée de vie de sa série et construire de nouvelles relations avec le public par le biais des plates-formes numériques qui deviennent de plus en plus important", at-il ajouté.
Source	In a press conference on Thursday, Mr Blair stated that there was nothing in this video that might constitute a "reasonable motive" that could lead to criminal charges being brought against the mayor.
Reference	En conférence de presse, jeudi, M. Blair a affirmé qu'il n'y avait rien dans cette vidéo qui puisse constituer des "motifs raisonnables" pouvant mener au dépôt d'une accusation criminelle contre le maire.
RNNenc-50	Lors de la conférence de presse de jeudi, M. Blair a dit qu'il n'y avait rien dans cette vidéo qui pourrait constituer une "motivation raisonnable" pouvant entraîner des accusations criminelles portées contre le maire.
RNNsearch-50	Lors d'une conférence de presse jeudi, M. Blair a déclaré qu'il n'y avait rien dans cette vidéo qui pourrait constituer un "motif raisonnable" qui pourrait conduire à des accusations criminelles contre le maire.
Google Translate	Lors d'une conférence de presse jeudi, M. Blair a déclaré qu'il n'y avait rien dans cette vido qui pourrait constituer un "motif raisonnable" qui pourrait mener à des accusations criminelles portes contre le maire.

Table 3: The translations generated by RNNenc-50 and RNNsearch-50 from long source sentences (30 words or more) selected from the test set. For each source sentence, we also show the gold-standard translation. The translations by Google Translate were made on 27 August 2014.

Neural Machine Translation using RNN

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Neural Machine Translation using LSTM

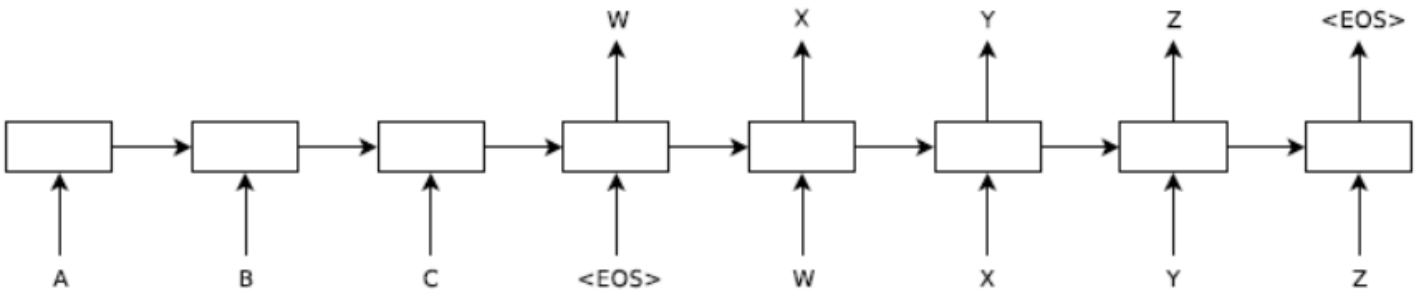


Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

Neural Machine Translation using RNN

- Input: $\mathbf{x} = (x_1, x_2, \ldots, x_T)$
- Expected Output: $\mathbf{y} = (y_1, y_2, \ldots, y_{T'})$
- Idea: To encode the input $\mathbf{x}$ into a single context vector c, then decode $\mathbf{y}$ from c.
- One LSTM used to construct the encoding from $\mathbf{x}$ to c
- Another LSTM used to construct the decoding of $\mathbf{y}$ from c
- During training, reversed training inputs were used !?

Neural Machine Translation using LSTM

$$p(y_i|y_1, y_2, \ldots, y_{t-1}, \mathbf{x}) = \hat{g}(y_{i-1}, s_i, c)$$

$\hat{g}$ is a suitable neural network and s_i is the hidden state of the RNN.

Neural Machine Translation using LSTM

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Table 1: The performance of the LSTM on WMT'14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

Neural Machine Translation using LSTM

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 " .

Table 3: A few examples of long translations produced by the LSTM alongside the ground truth translations. The reader can verify that the translations are sensible using Google translate.

Neural Machine Translation using LSTM

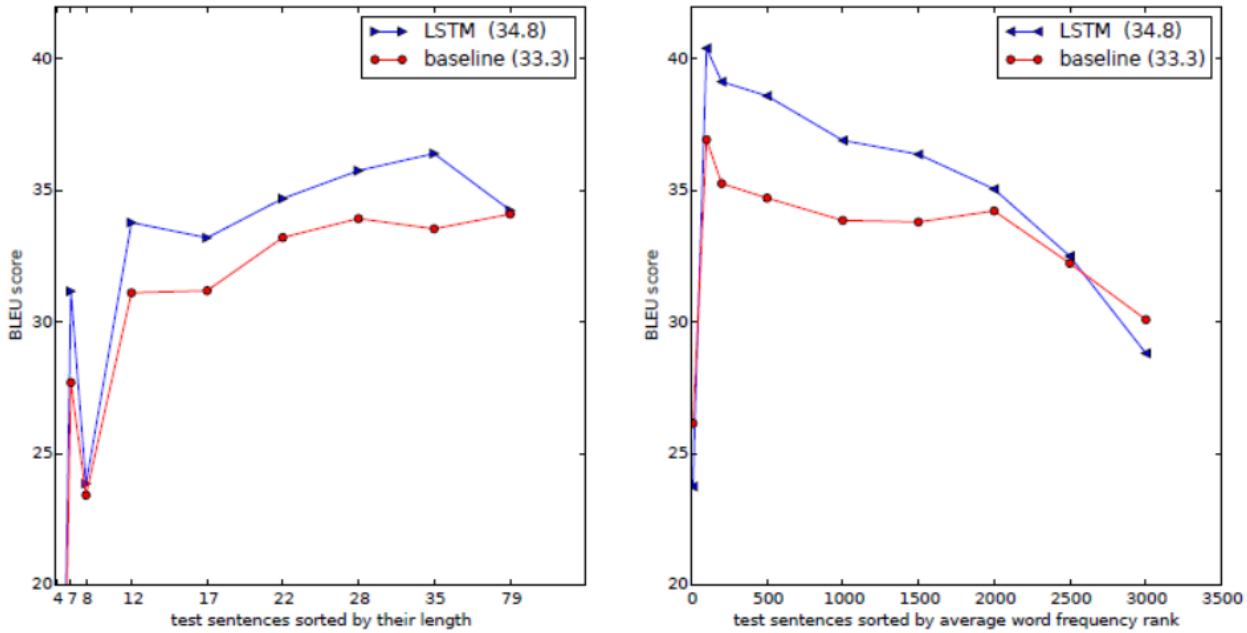


Figure 3: The left plot shows the performance of our system as a function of sentence length, where the x-axis corresponds to the test sentences sorted by their length and is marked by the actual sequence lengths. There is no degradation on sentences with less than 35 words, there is only a minor degradation on the longest sentences. The right plot shows the LSTM's performance on sentences with progressively more rare words, where the x-axis corresponds to the test sentences sorted by their "average word frequency rank".