Recurrent Neural Networks

Isabel Agostino and Bryan Pan

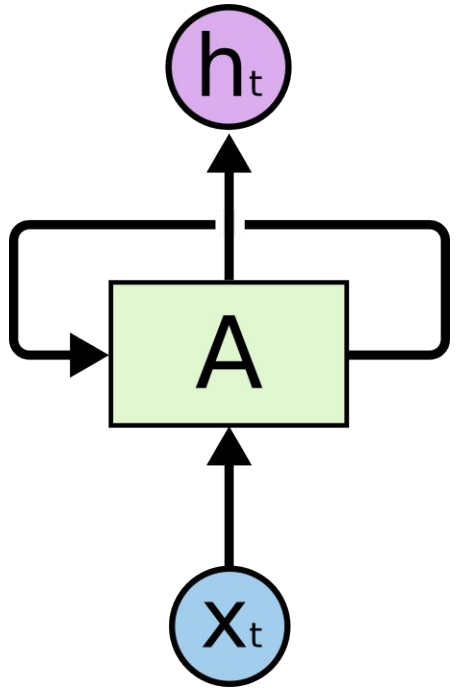
What if we have sequential or history-dependent information?

Motivation

- Sequential data examples:
  - Text, speech, time-series data
- Standard neural networks pass information forward in time
  - Vanishing gradient problem
  - Unable to process sequential data
- We need to introduce a structure to move history or past information into the future

Main Idea: include history/past information

- Recurrent neurons – hidden state with past information
  - Feed the hidden state into the next step of the NN
- Shared weights used across time

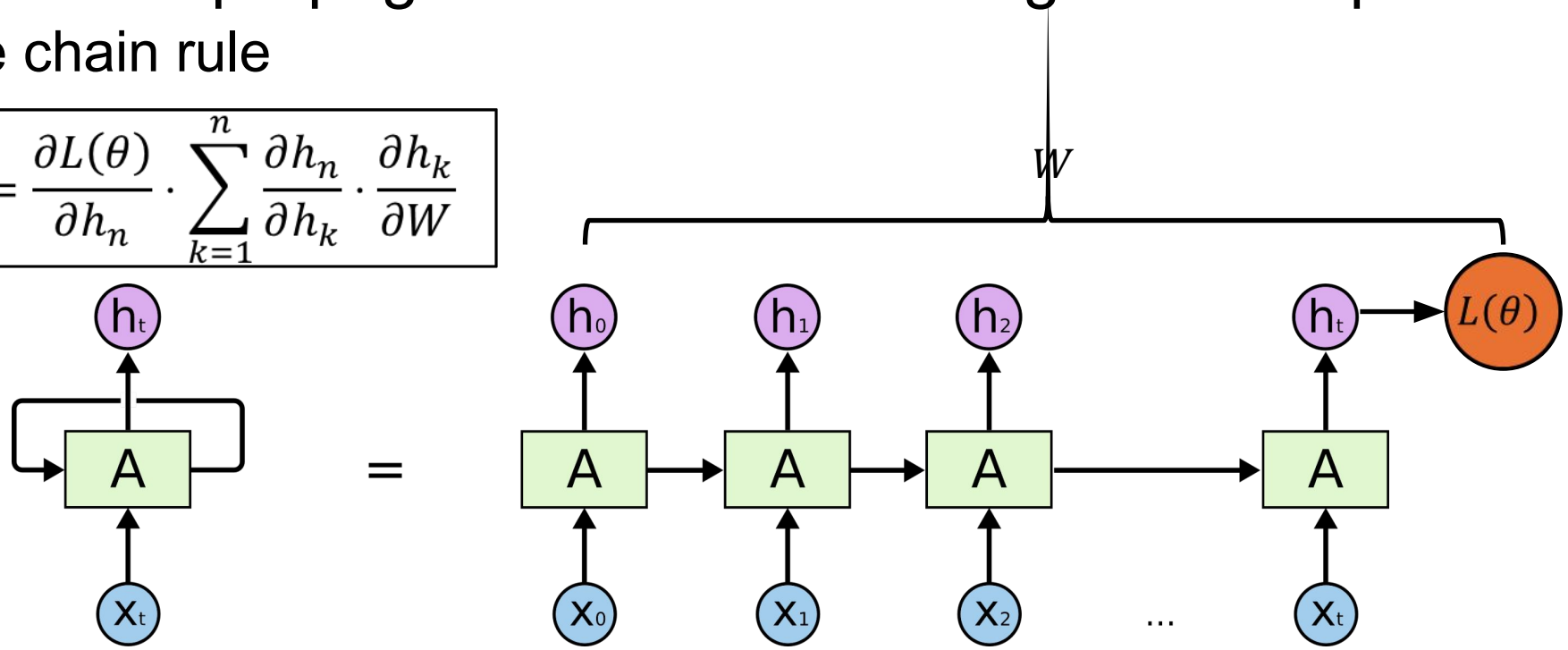


Olah, Christopher. 2015. “Understanding LSTM Networks.” Colah’s Blog.

Unrolling & Back Propagation through Time

- Recursion → Can unroll and compute iteratively
- Loss function depends on hidden states
- BPTT: gradients propagated backward through time steps
  - Iterative chain rule

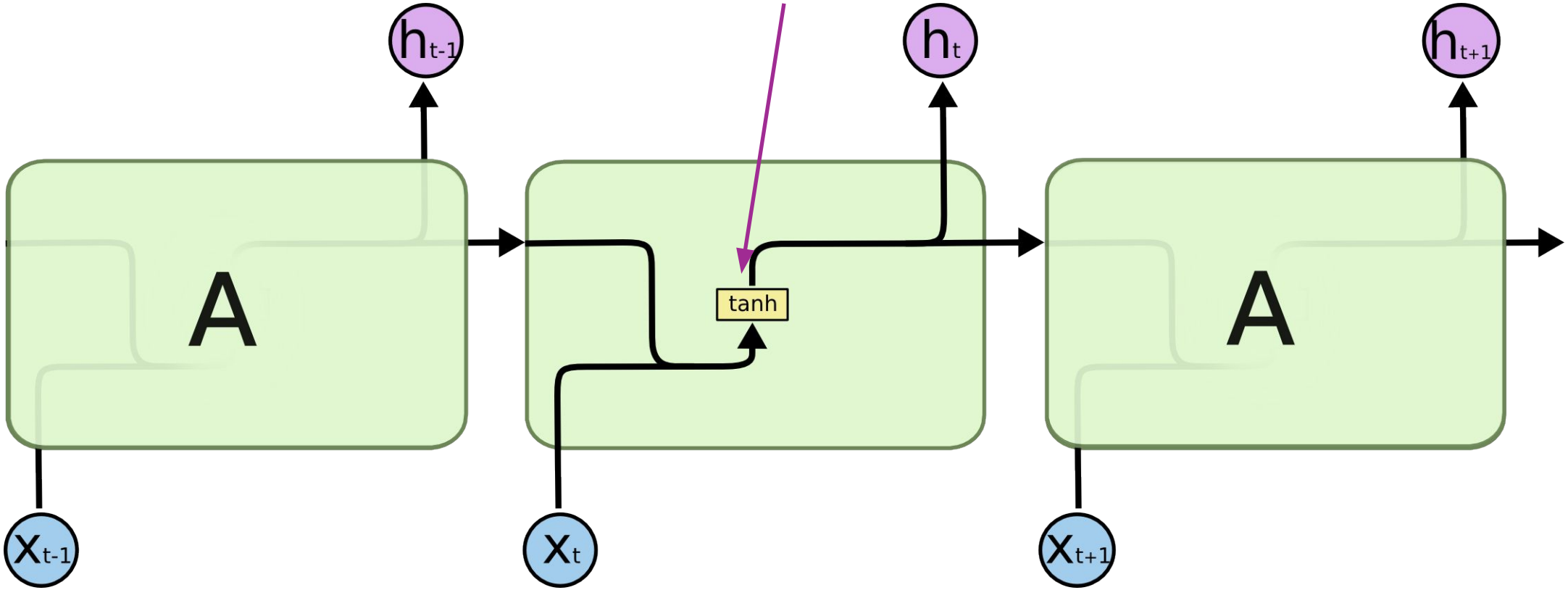
$$\frac{\partial L(\theta)}{\partial W} = \frac{\partial L(\theta)}{\partial h_n} \cdot \sum_{k=1}^{n} \frac{\partial h_n}{\partial h_k} \cdot \frac{\partial h_k}{\partial W}$$



Olah, Christopher. 2015. “Understanding LSTM Networks.” Colah’s Blog.

“Vanilla” RNN

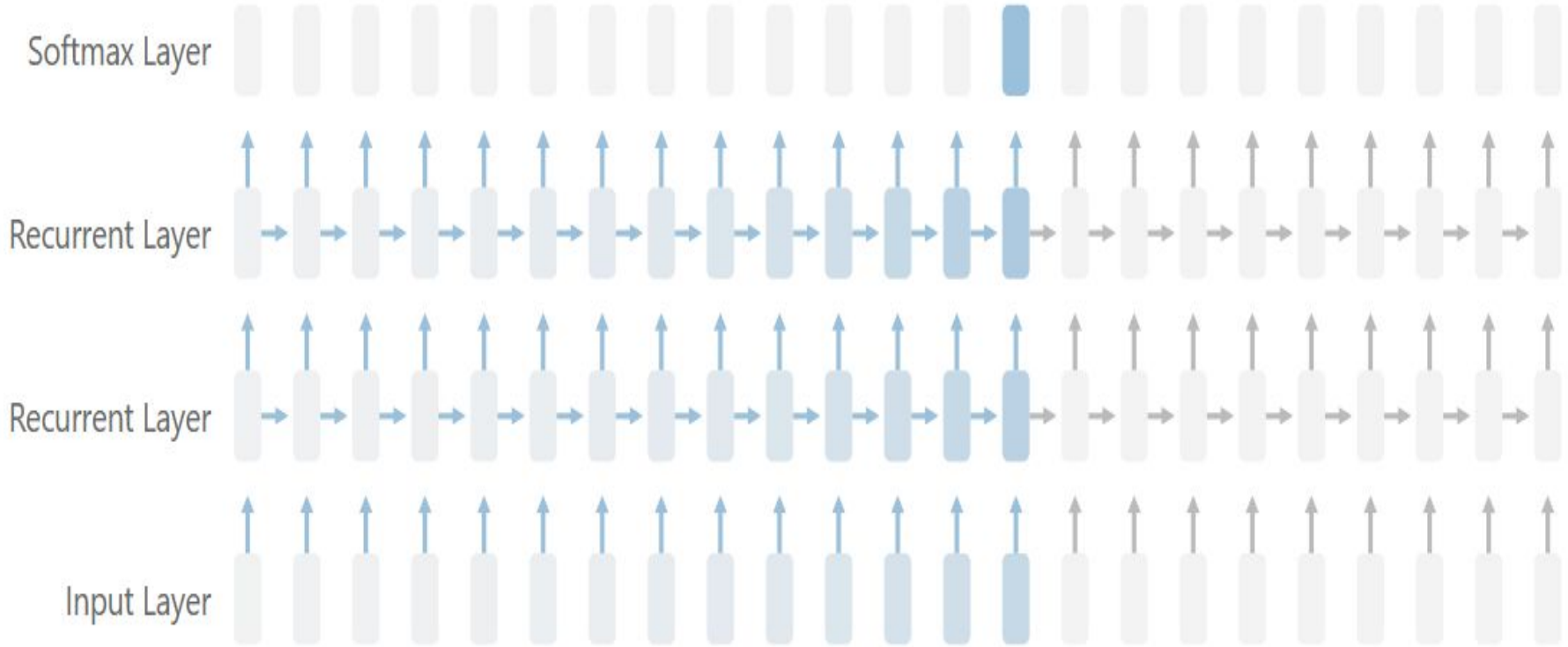
Activation function: $h_t = \tanh(W_h \cdot h_{t-1} + W_x \cdot x_t)$



Olah, Christopher. 2015. “Understanding LSTM Networks.” Colah’s Blog.

Issue: Vanishing Gradient

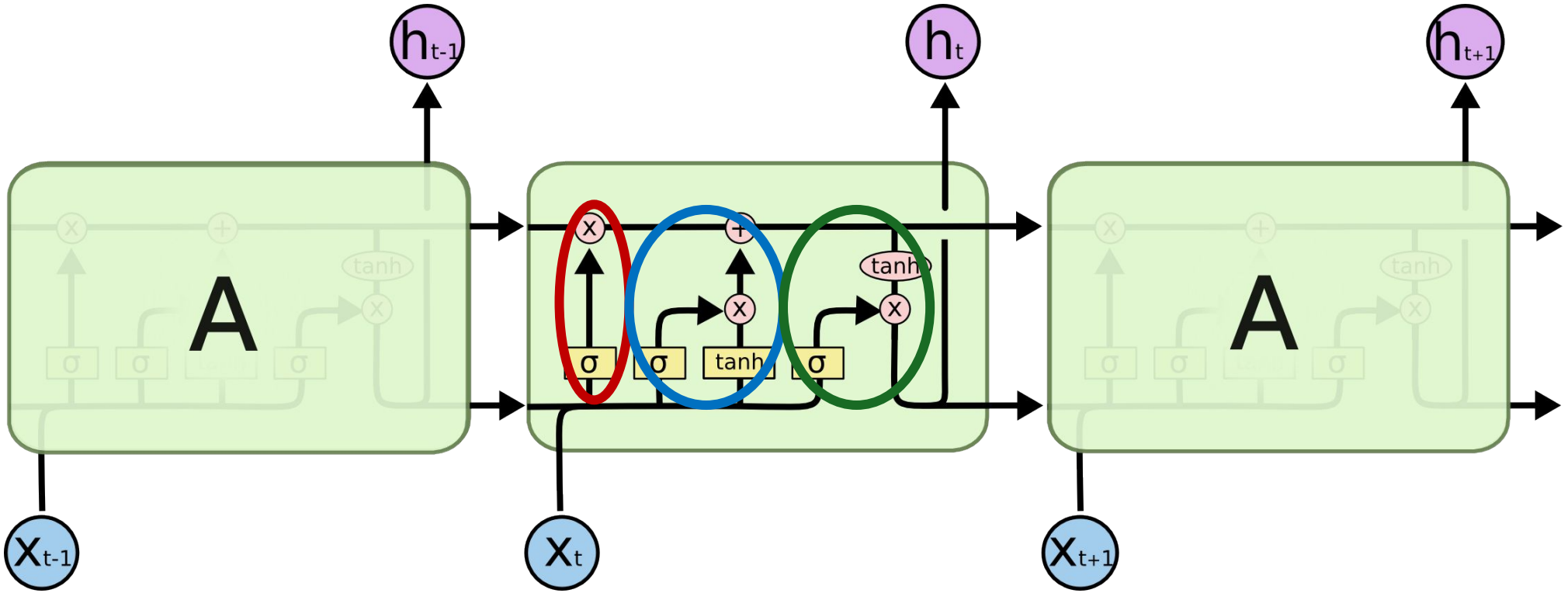
- $\frac{\partial L(\theta)}{\partial W} = \frac{\partial L(\theta)}{\partial h_n} \cdot \sum_{k=1}^{n} \frac{\partial h_n}{\partial h_k} \cdot \frac{\partial h_k}{\partial W}$ vanishes over many steps
  - Makes it difficult to process long sequences



Madsen, Andreas. 2019. “Visualizing Memorization in RNNs.” *Distill* 4 (3): e16

Long Short-Term Memory (LSTM)

- Each cell has 3 gates:
  - Forget gate: What past information should be removed?
  - Input gate: How much new information should be added?
  - Output gate: What information should be output at the current step?

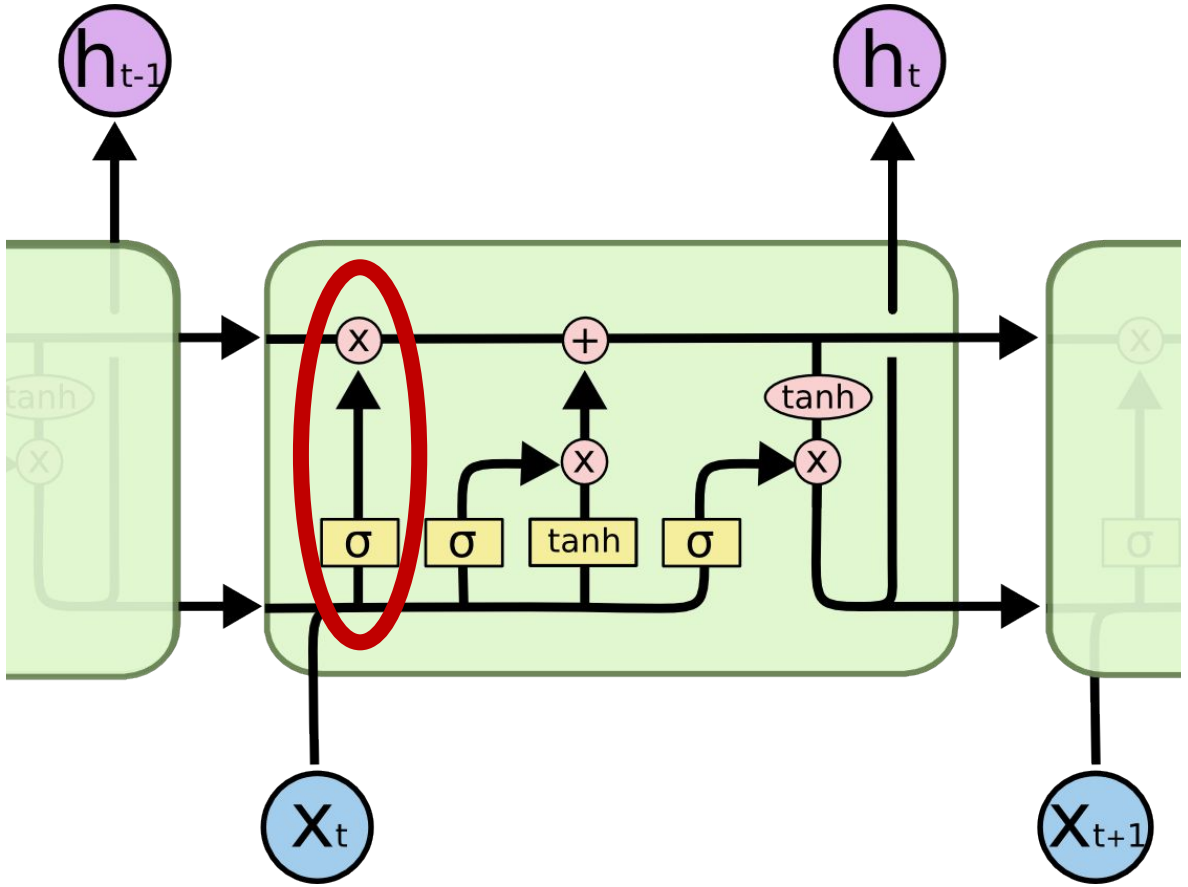


Olah, Christopher. 2015. “Understanding LSTM Networks.” Colah’s Blog.

Long Short-Term Memory (LSTM)

Forget gate:

- Sigmoid computes "keep score" value for h_{t-1} and x_t
- Multiply into cell state

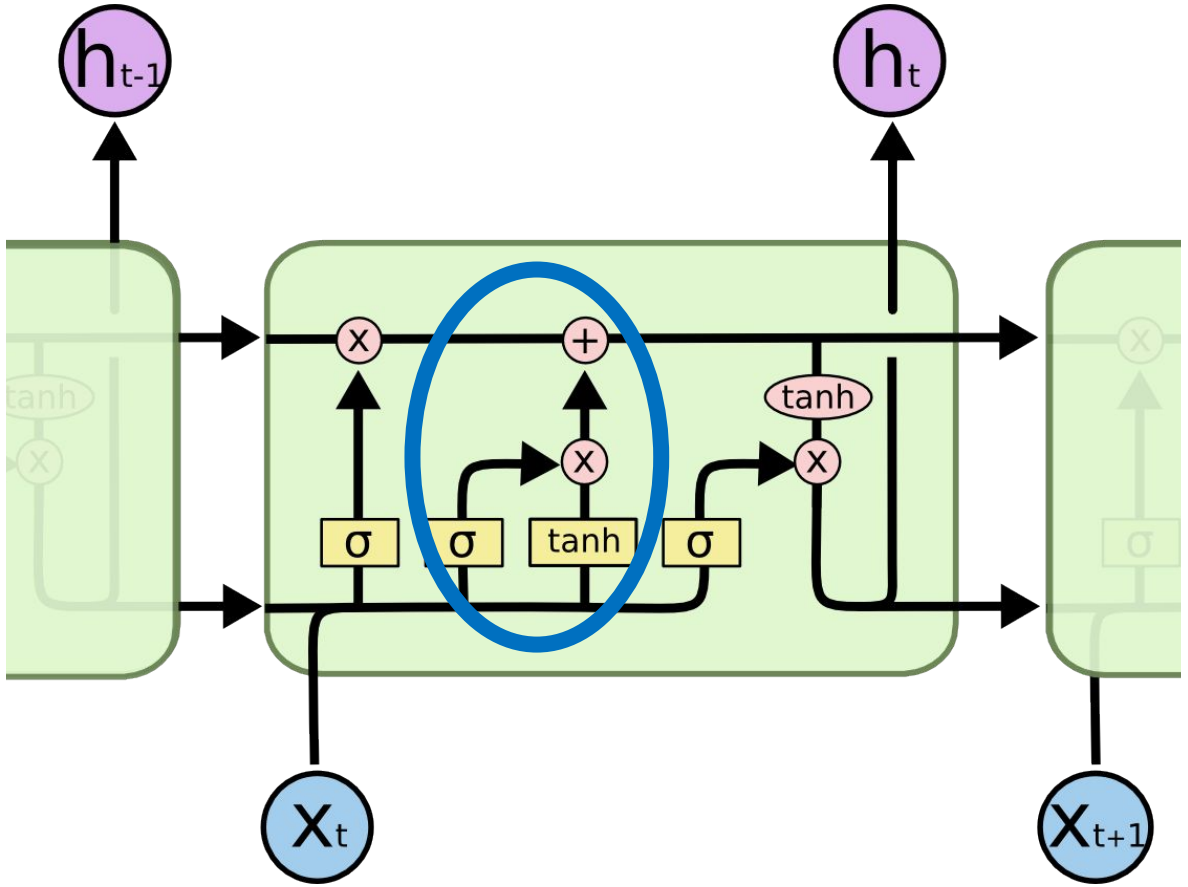


Olah, Christopher. 2015. "Understanding LSTM Networks." Colah's Blog.

Long Short-Term Memory (LSTM)

Input gate:

- Sigmoid decides values to update
- Tanh creates new candidates
- Add into cell state

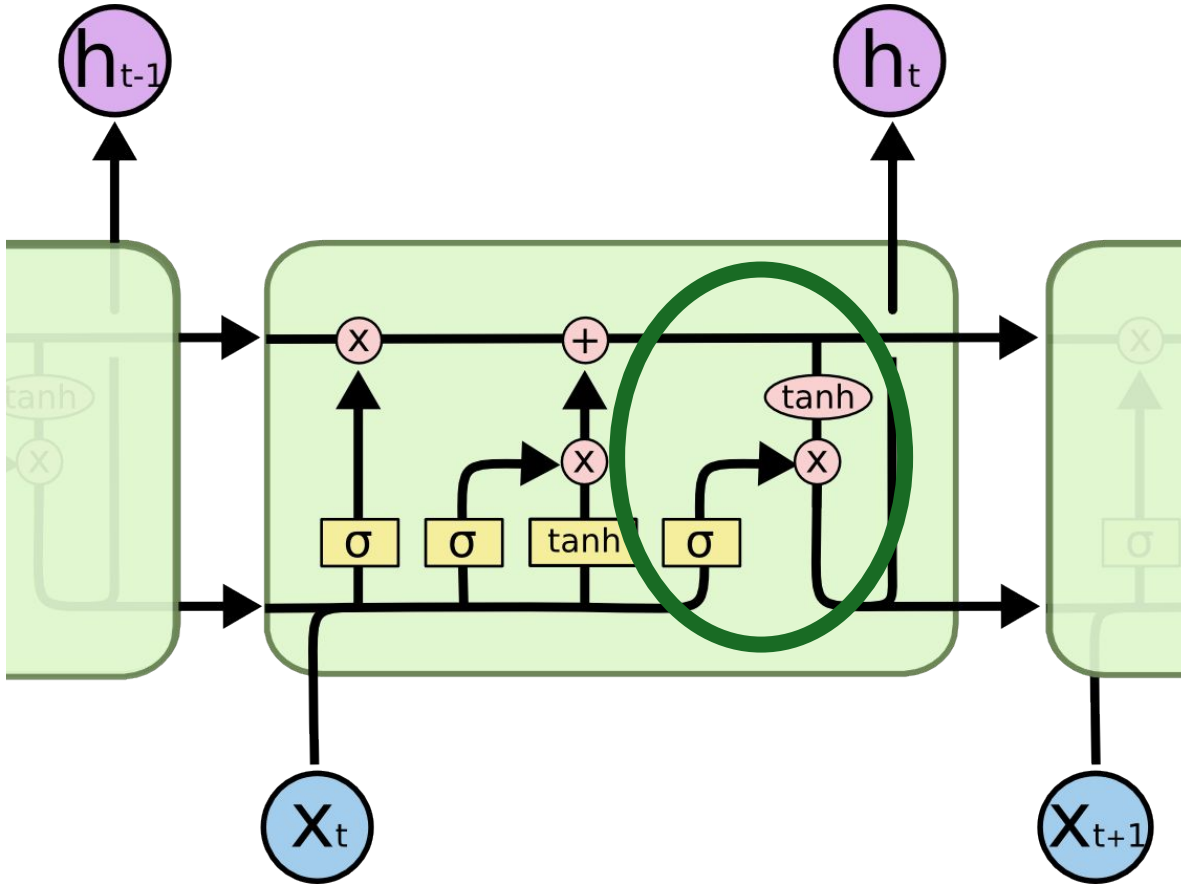


Olah, Christopher. 2015. “Understanding LSTM Networks.” Colah’s Blog.

Long Short-Term Memory (LSTM)

Output gate:

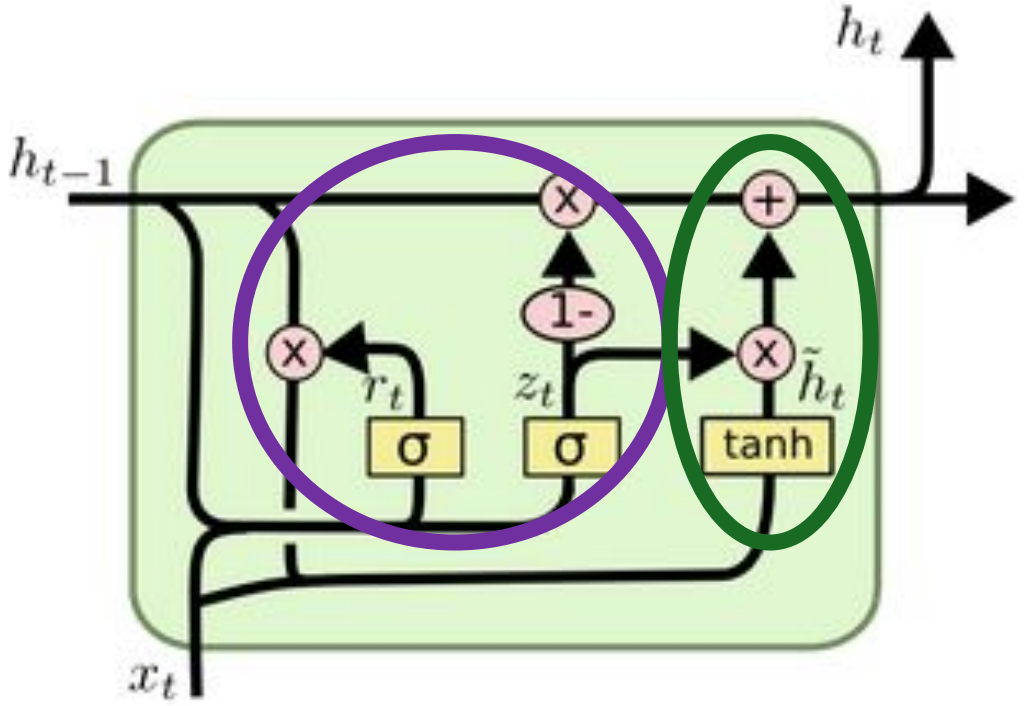
- Sigmoid computes state to output
- Apply tanh to cell state
- Multiply by sigmoid



Olah, Christopher. 2015. “Understanding LSTM Networks.” Colah’s Blog.

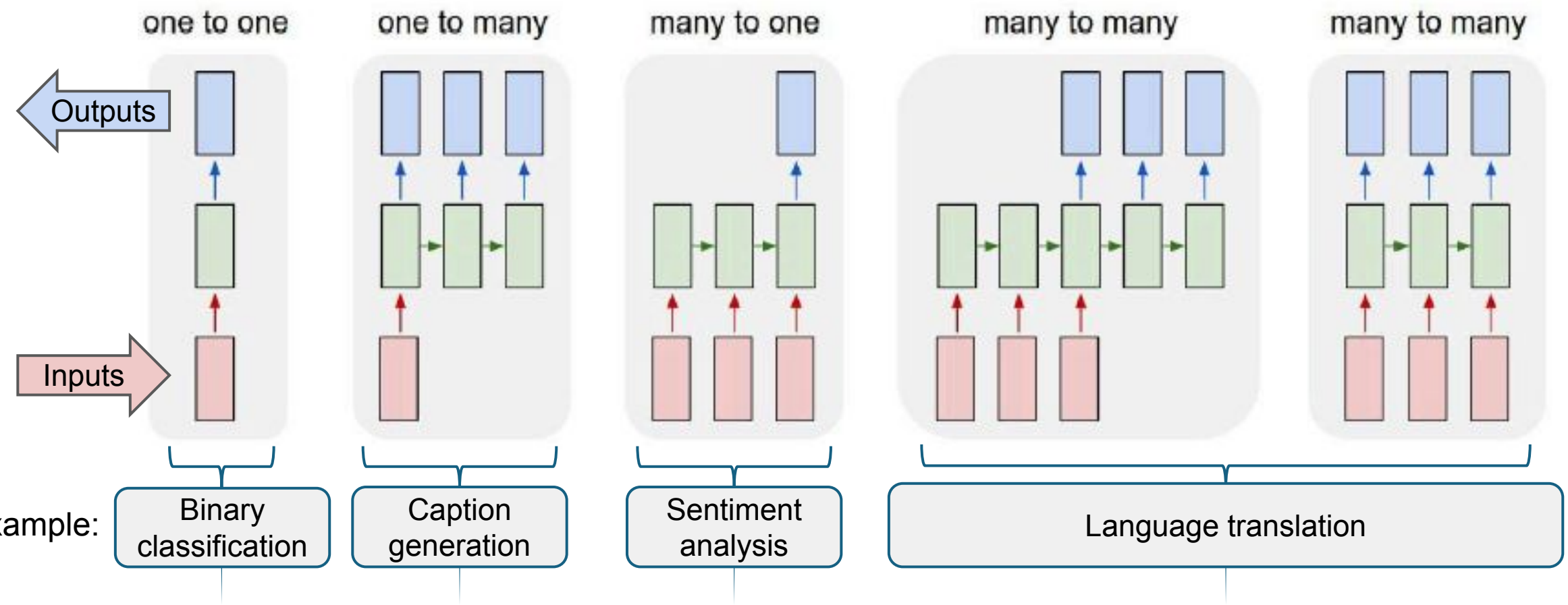
Gated Recurrent Unit (GRU)

- Merges cell state and hidden state
- Update gate: Input gate+ forget gate
- Simpler than standard LSTM
- Very popular



Olah, Christopher. 2015. “Understanding LSTM Networks.” Colah’s Blog.

Architectures



“Introduction to Recurrent Neural Networks - GeeksforGeeks.” 2025.

Application: Machine Translation

Thought question: How much history do we need to translate text?

Example:

she → èl she → ella

Example:

I eat sweet bread

Yo como pan dulce

난 달달한 빵 먹어

- Goal: find a translation f given a source sentence e maximizing

$$Pr(f|e) \propto Pr(e|f)Pr(f)$$

- Input & output: lengths vary, complex relationships
  - Train an RNN to predict the next symbol in a sequence

BLEU Scores

- **Bi**L**ingual** **E**valuation **U**nderstudy - score for translation quality
- Varies based on languages, domain, complexity, etc.

General Interpretation:

- **0-30**: Poor to understandable translation
- **30-40**: Understandable to good translation
- **40-50**: High-quality translation
- **50-60**: Very high quality, adequate, fluent translation
- **>60**: Better than human

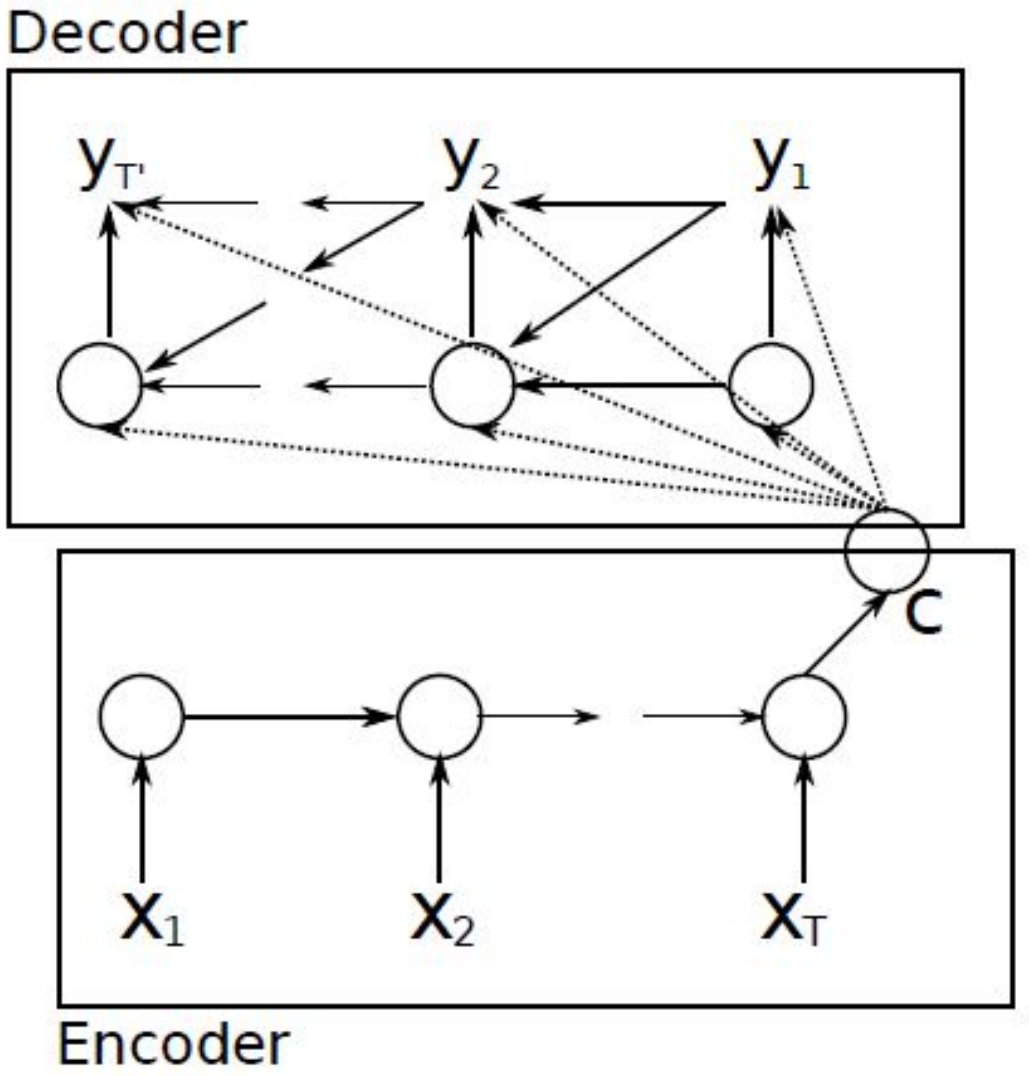
Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

Cho et. al (2014)

Main Points

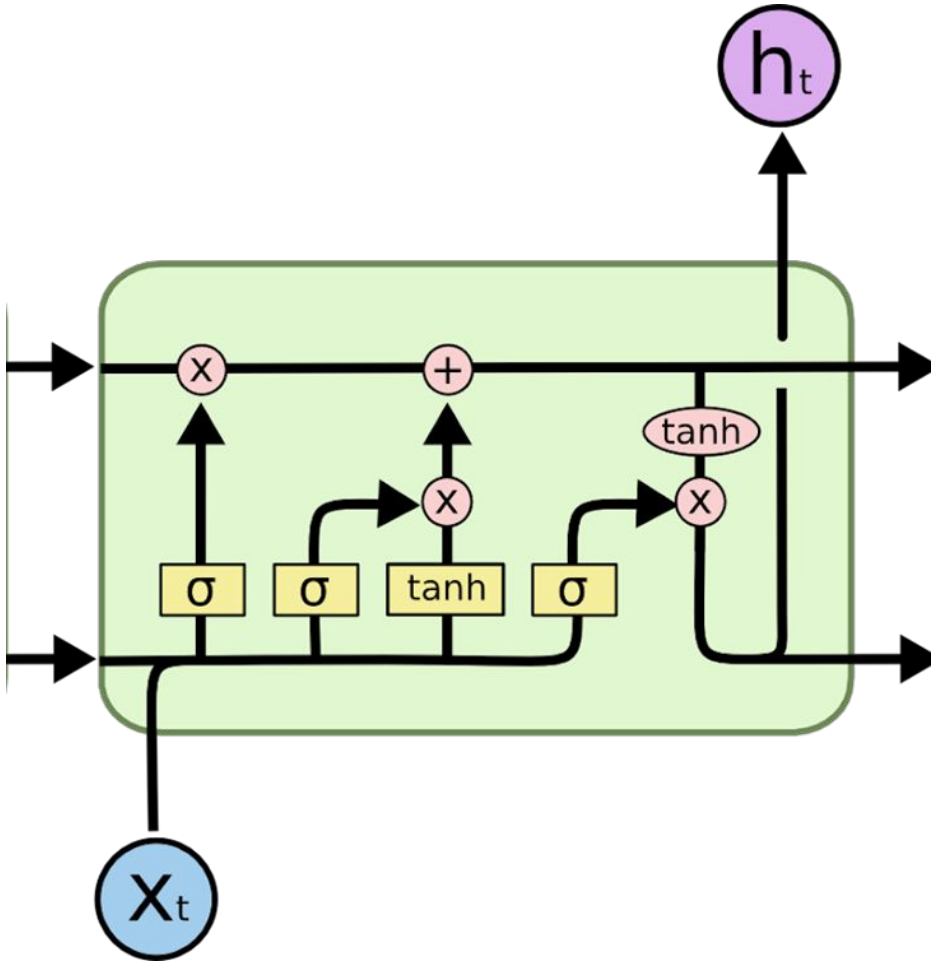
- New architecture: RNN Encoder-Decoder
  - Two RNN in sequence (one encodes, the other decodes)
- Introduction of GRU
- Train networks jointly to maximize conditional log-likelihood:

$$\max_{\theta} \frac{1}{\theta} \sum_{n=1}^{N} \log p_\theta \left(y_n | x_n\right)$$

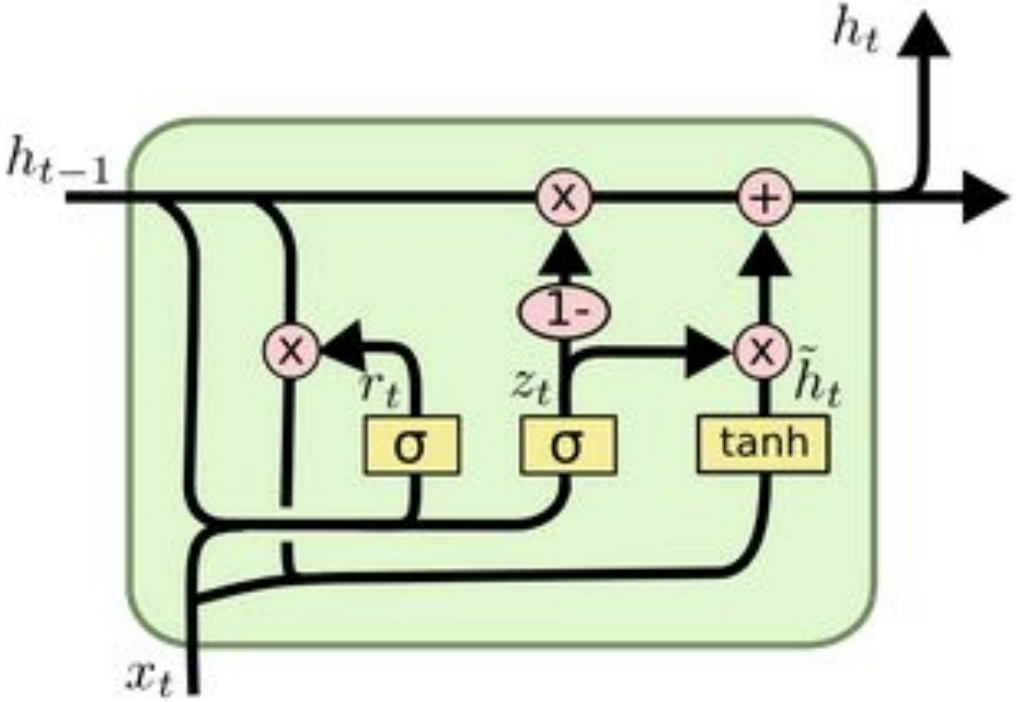


Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, et al. 2014. “Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation.” arXiv:1406.1078.

LSTM



GRU



Olah, Christopher. 2015. “Understanding LSTM Networks.” Colah’s Blog.

Technical Details

- Compared RNN Encoder-Decoder with traditional collaborative small language model (CSLM)

Training data:

- 418M words for CSLM
- 348M words for encoder-decoder
- Limited to 15,000 most common words in each language (~93% data)

Structure of Encoder-Decoder:

- 1000 hidden units
- Adadelta and stochastic gradient descent

Source	Translation Model	RNN Encoder–Decoder
at the end of the	[a la fin de la] [f la fin des années] [être supprimés à la fin de la]	[à la fin du] [à la fin des] [à la fin de la]
for the first time	[r © pour la premrëre fois] [été donnés pour la première fois] [été commémorée pour la première fois]	[pour la première fois] [pour la première fois ,] [pour la première fois que]
in the United States and	[? aux ?tats-Unis et] [été ouvertes aux États-Unis et] [été constatées aux États-Unis et]	[aux Etats-Unis et] [des Etats-Unis et] [des États-Unis et]
, as well as	[?s , qu'] [?s , ainsi que] [?re aussi bien que]	[, ainsi qu'] [, ainsi que] [, ainsi que les]
one of the most	[?t ?l' un des plus] [?l' un des plus] [être retenue comme un de ses plus]	[l' un des] [le] [un des]

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, et al. 2014. "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation." arXiv:1406.1078.

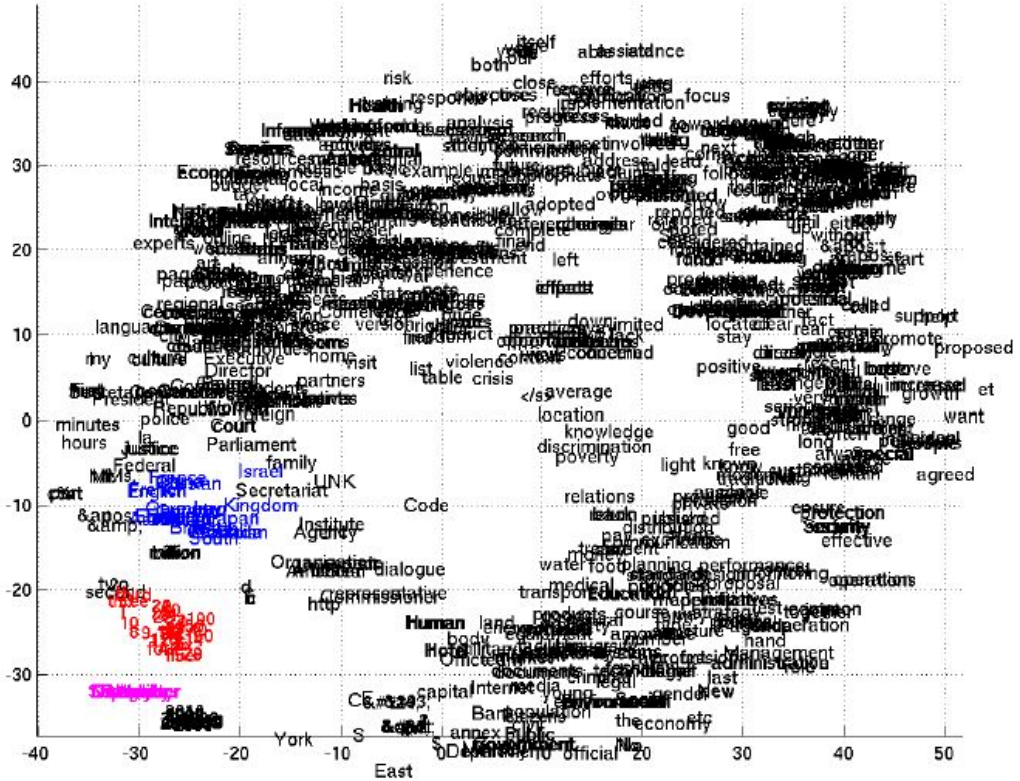
Findings

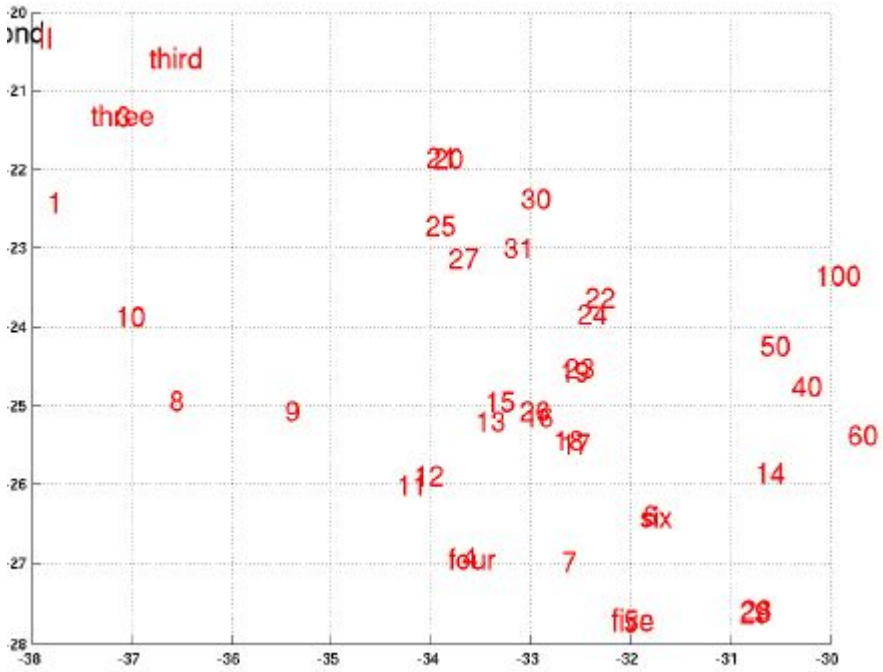
- Encoder-Decoder:
  - Translations close to literal/actual translations
  - Prefers shorter phrases
  - “preserves both the semantic and syntactic structure of the phrase”
- Best results with both:
  - Two models not too correlated

Models	BLEU	
	dev	test
Baseline	30.64	33.30
RNN	31.20	33.87
CSLM + RNN	31.48	34.64
CSLM + RNN + WP	31.50	34.54

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, et al. 2014. “Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation.” arXiv:1406.1078.

Word Embeddings





Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, et al. 2014. “Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation.” arXiv:1406.1078.

Sequence to Sequence Learning with Neural Networks

Sutskever, Vinyals, Le (2014)

Main Ideas

- Encoder-Decoder architecture: 2 LSTM in sequence
- Deep LSTMs outperform shallow LSTM
- Reversed source sentence order → performs better on long sentences
  - Introduce short-term dependencies
  - Minimal lag time reduced, but average lag same (BPTT easier)

Example Sentence	Lag(I)	Lag(eat)	Lag(sweet)	Lag(bread)	Average Lag	Minimum Lag
I eat sweet bread. → Yo como pan dulce.	4	4	5	3	4	3
Bread sweet eat I. → Yo como pan dulce.	1	3	6	6	4	1

Technical Details

Training data:

- 348M French words and 304M English words
- Input vocab: 160K words → Output vocab: 80,000 words

Structure:

- LSTM with 4 layers, 1000 cells/layer
  - “each additional layer reduced perplexity by nearly 10%”
- 1000 dimensional word embeddings

Results

- Reversing sentences:
  - Test perplexity: 5.8 to 4.7 (good)
  - Test BLEU: 25.9 to 30.96 (good)
- Increasing depth increased score

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**

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. “Sequence to Sequence Learning with Neural Networks.” arXiv:1409.3215.

Results

Comparison to other models:

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	~45

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. "Sequence to Sequence Learning with Neural Networks." arXiv:1409.3215.

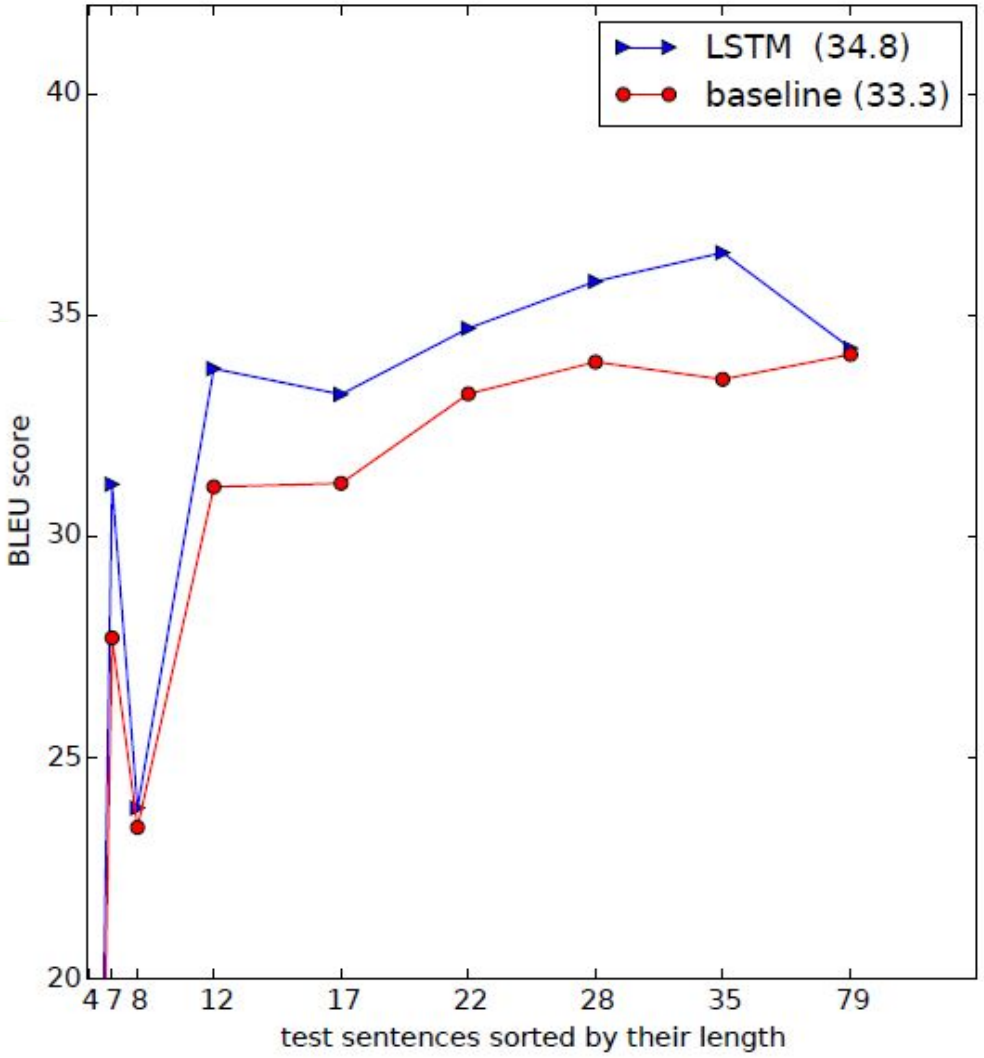
Performance on Long Sentences

RNN:

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 .

Actual:

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 .



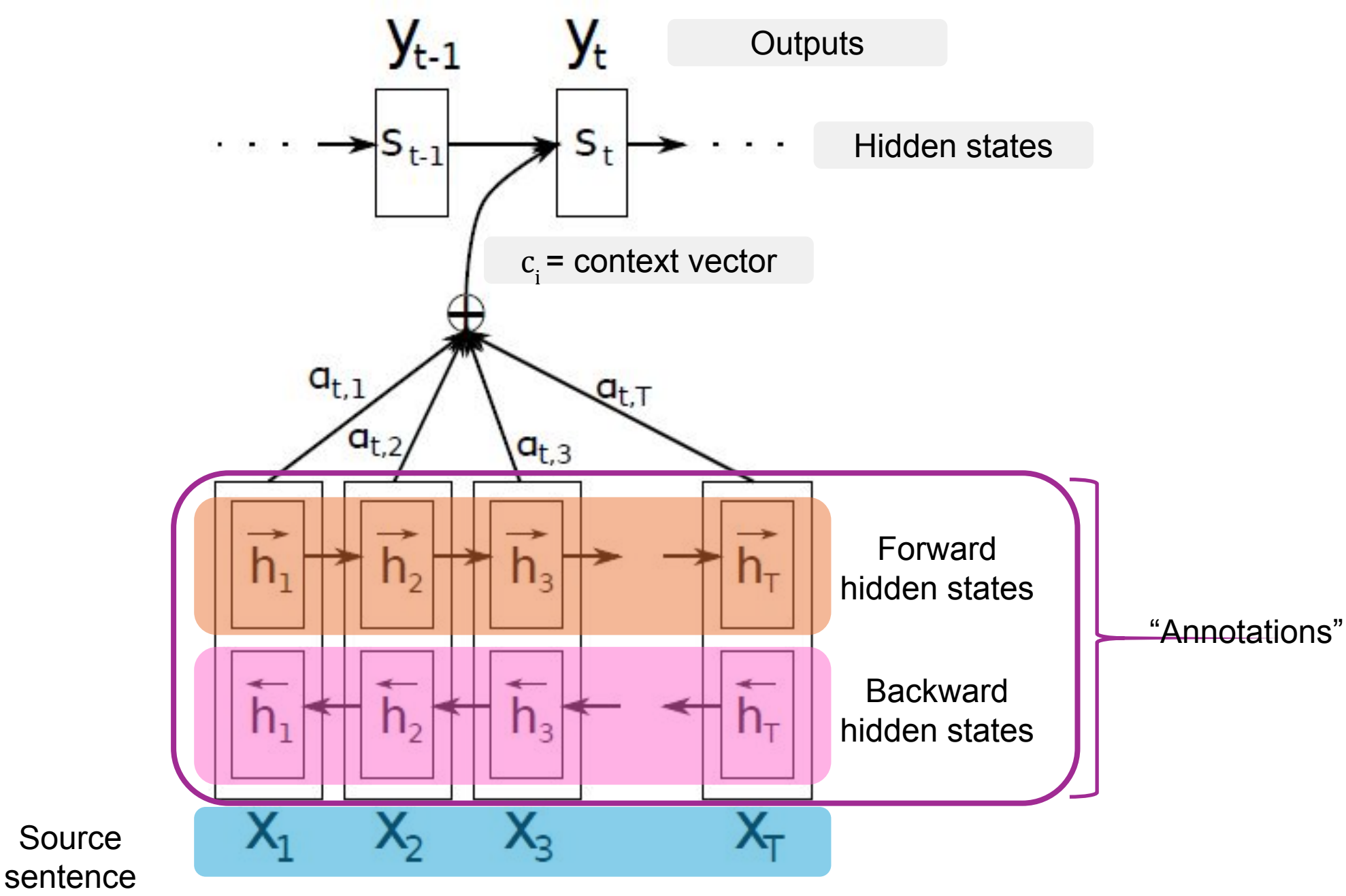
Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. "Sequence to Sequence Learning with Neural Networks." arXiv:1409.3215.

Neural Machine Translation by Jointly Learning to Align and Translate

Bahdanau, Cho, Bengio (2016)

Main Ideas

- Issue with encoder-decoder: encoder must output fixed-length vector
  - Difficult to capture meaning of long sentences
- Instead: maps input to vector sequence and selects subsets for decoding
- Uses a bidirectional RNN
  - “emulates searching through a source sentence during decoding a translation”
- Probability explicitly conditioned on context vector and alignment
  - Usually considered latent variables in an RNN



Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2016. "Neural Machine Translation by Jointly Learning to Align and Translate."

Technical Details

- Same training data as first paper!

Training data:

- 418M words for CSLM
- 348M words for encoder-decoder
- Limited to 15,000 most common words in each language (~93% data)

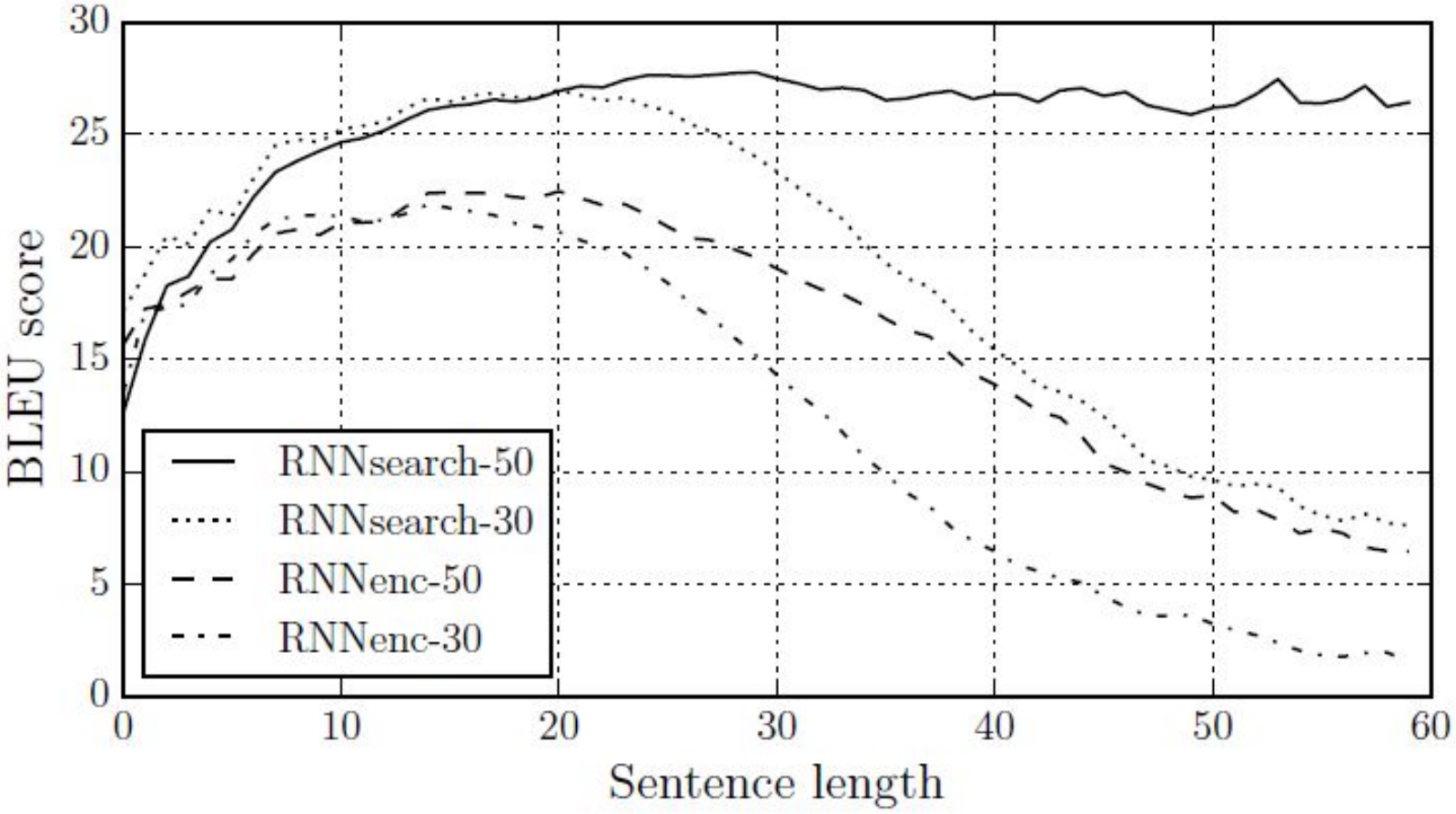
Structure:

- RNN forward and backward direction, 1000 hidden states each
- Adadelta and minibatch SGD

Performance vs. Plain Encoder-Decoder

Model	All	No UNK°
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50*	28.45	36.15
Moses	33.30	35.63

Their model performed as well as conventional translation system!



Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2016. “Neural Machine Translation by Jointly Learning to Align and Translate.” arXiv:1409.0473.

Shortcomings of RNNs

- BPTT can be computationally expensive and time-consuming to train
- Sensitive to hyperparameter choice
- Struggle with long-term dependencies
- Biased toward recent data
- No global context
- Hard to parallelize

Examples!

Toy Examples: Shakespeare

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

```
VIOLA:
Why, Salisbury must find his flesh and thought
That which I am not aps, not a man and in fire,
To show the reining of the raven and the wars
To grace my hand reproach within, and not a fair are hand,
That Caesar and my goodly father's world;
When I was heaven of presence and our fleets,
We spare with hours, but cut thy council I am great,
Murdered and by thy master's ready there
My power to give thee but so much as hell:
Some service in the noble bondman here,
Would show him to her wine.

KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.
```

Toy Examples: Wikipedia

```
Naturalism and decision for the majority of Arab countries' capitalide was grounded
by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated
with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal
in the [[Protestant Immineners]], which could be said to be directly in Cantonese
Communication, which followed a ceremony and set inspired prison, training. The
emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom
of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known
in western [[Scotland]], near Italy to the conquest of India with the conflict.
Copyright was the succession of independence in the slop of Syrian influence that
was a famous German movement based on a more popular servicious, non-doctrinal
and sexual power post. Many governments recognize the military housing of the
[[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]],
that is sympathetic to be to the [[Punjab Resolution]]
(PJS)[http://www.humah.yahoo.com/guardian.
cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery
was swear to advance to the resources for those Socialism's rule,
was starting to signing a major tripad of aid exile.]]
```

```
{ { cite journal | id=Cerling Nonforest Department|format=Newlymeslated|none } }
''www.e-complete''.

'''See also''': [[List of ethical consent processing]]

== See also ==
*[[Iender dome of the ED]]
*[[Anti-autism]]

===[[Religion|Religion]]===
*[[French Writings]]
*[[Maria]]
*[[Revelation]]
*[[Mount Agamul]]

== External links==
* [http://www.biblegateway.nih.gov/entrepre/ Website of the World Festival. The labour

==External links==
* [http://www.romanology.com/ Constitution of the Netherlands and Hispanic Competition
```

```
<page>
  <title>Antichrist</title>
  <id>865</id>
  <revision>
    <id>15900676</id>
    <timestamp>2002-08-03T18:14:12Z</timestamp>
    <contributor>
      <username>Paris</username>
      <id>23</id>
    </contributor>
    <minor />
    <comment>Automated conversion</comment>
    <text xml:space="preserve">#REDIRECT [[Christianity]]</text>
  </revision>
</page>
```

Toy Examples: Latex

For $\bigoplus_{n\equiv 1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the compariconly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, ?? and the fact that any U affine, see Morphisms, Lemma ??. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma ?? we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example ??. It may replace S by $X_{spaces,étale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma ??. Namely, by Lemma ?? we see that R is geometrically regular over S.

Lemma 0.1. *Assume (3) and (3) by the construction in the description.*

Suppose $X = \lim |X|$ *(by the formal open covering* X *and a single map* $\underline{Proj}_X(\mathcal{A}) = \mathrm{Spec}(B)$ *over* U *compatible with the complex*

$$Set(\mathcal{A}) = \Gamma(X, \mathcal{O}_{X,\mathcal{O}_X}).$$

When in this case of to show that $\mathcal{Q} \to \mathcal{C}_{Z/X}$ *is stable under the following result in the second conditions of (1), and (3). This finishes the proof. By Definition ?? (without element is when the closed subschemes are catenary. If* T *is surjective we may assume that* T *is connected with residue fields of* S*. Moreover there exists a closed subspace* $Z \subset X$ *of* X *where* U *in* X' *is proper (some defining as a closed subset of the uniqueness it suffices to check the fact that the following theorem*

(1) f *is locally of finite type. Since* $S = \mathrm{Spec}(R)$ *and* $Y = \mathrm{Spec}(R)$.

Proof. This is form all sheaves of sheaves on X. But given a scheme U and a surjective étale morphism $U \to X$. Let $U \cap U = \coprod_{i=1,\ldots,n} U_i$ be the scheme X over S at the schemes $X_i \to X$ and $U = \lim_i X_i$. □

The following lemma surjective restrocomposes of this implies that $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} = \mathcal{F}_{\mathcal{X},\ldots,0}$.

Lemma 0.2. *Let* X *be a locally Noetherian scheme over* S*,* $E = \mathcal{F}_{X/S}$*. Set* $\mathcal{I} = \mathcal{J}_1 \subset \mathcal{I}'_n$*. Since* $\mathcal{I}^n \subset \mathcal{I}^n$ *are nonzero over* $i_0 \leq \mathfrak{p}$ *is a subset of* $\mathcal{J}_{n,0} \circ \overline{A}_2$ *works.*

Lemma 0.3. *In Situation ??. Hence we may assume* $\mathfrak{q}' = 0$.

Proof. We will use the property we see that $\mathfrak{p}$ is the mext functor (??). On the other hand, by Lemma ?? we see that

$$D(\mathcal{O}_{X'}) = \mathcal{O}_X(D)$$

where K is an F-algebra where δ_{n+1} is a scheme over S. □

Proof. Omitted. □

Lemma 0.1. *Let* $\mathcal{C}$ *be a set of the construction.*

Let $\mathcal{C}$ *be a gerber covering. Let* $\mathcal{F}$ *be a quasi-coherent sheaves of* $\mathcal{O}$*-modules. We have to show that*

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

.

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{étale}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{morph_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \to \mathcal{F}$ of $\mathcal{O}$-modules. □

Lemma 0.2. *This is an integer* $\mathcal{Z}$ *is injective.*

Proof. See Spaces, Lemma ??. □

Lemma 0.3. *Let* S *be a scheme. Let* X *be a scheme and* X *is an affine open covering. Let* $\mathcal{U} \subset \mathcal{X}$ *be a canonical and locally of finite type. Let* X *be a scheme. Let* X *be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X *be a scheme. Let* X *be a scheme covering. Let*

$$b : X \to Y' \to Y \to Y \to Y' \times_X Y \to X.$$

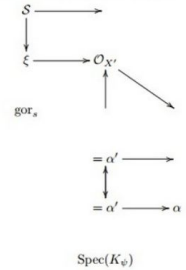
be a morphism of algebraic spaces over S *and* Y.

Proof. Let X be a nonzero scheme of X. Let X be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$-modules. The following are equivalent

(1) $\mathcal{F}$ is an algebraic space over S.
(2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram



is a limit. Then $\mathcal{G}$ is a finite type and assume S is a flat and $\mathcal{F}$ and $\mathcal{G}$ is a finite type $f_\bullet$. This is of finite type diagrams, and

- the composition of $\mathcal{G}$ is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

□

Proof. We have see that $X = \mathrm{Spec}(R)$ and $\mathcal{F}$ is a finite type representable by algebraic space. The property $\mathcal{F}$ is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U. □

Proof. This is clear that $\mathcal{G}$ is a finite presentation, see Lemmas ??.
A *reduced above* we conclude that U is an open covering of $\mathcal{C}$. The functor $\mathcal{F}$ is a "field

$$\mathcal{O}_{X,x} \longrightarrow \mathcal{F}_{\overline{x}} \quad \text{-1}(\mathcal{O}_{X_{étale}}) \longrightarrow \mathcal{O}_{X_t}^{-1}\mathcal{O}_{X_\lambda}(\mathcal{O}_{X_\eta}^{\overline{x}})$$

is an isomorphism of covering of $\mathcal{O}_{X_i}$. If $\mathcal{F}$ is the unique element of $\mathcal{F}$ such that X is an isomorphism.
The property $\mathcal{F}$ is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme $\mathcal{O}_X$-algebra with $\mathcal{F}$ are opens of finite type over S.
If $\mathcal{F}$ is a scheme theoretic image points. □

If $\mathcal{F}$ is a finite direct sum $\mathcal{O}_{X_\lambda}$ is a closed immersion, see Lemma ??. This is a sequence of $\mathcal{F}$ is a similar morphism.

Evolution Over Iterations: War and Peace

100:

```
tyntd-iafhatawiaoihrdemot  lytdws  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng
```

300:

```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
```

700:

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overelical and ofter.
```

1200:

```
"Kite vouch!" he repeated by her
door. "But I would be done and quarts, feeling, then, son is people...."
```

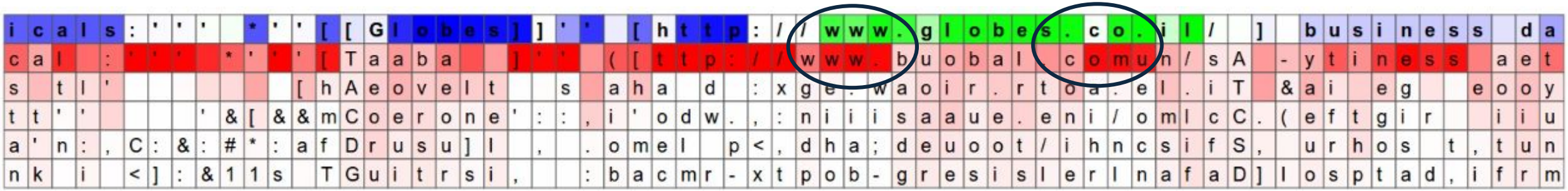
2000:

```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had oftened him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

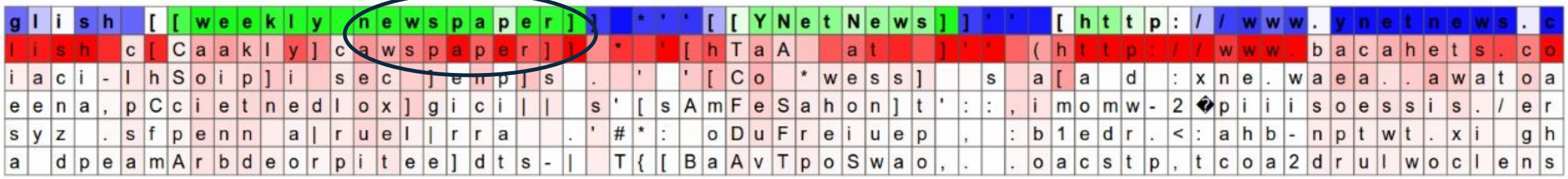
Karpathy, Andrej. 2015. “The Unreasonable Effectiveness of Recurrent Neural Networks.” Accessed October 6, 2025.

Neuron Firings

- First row: green = very excited and blue = very unexcited
- Below: 5 most likely next characters (red more probability)



Neuron excited about hyperlinks



Neuron excited about [[ ]] environment

Karpathy, Andrej. 2015. “The Unreasonable Effectiveness of Recurrent Neural Networks.” Accessed October 6, 2025.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2016. “Neural Machine Translation by Jointly Learning to Align and Translate.” arXiv:1409.0473. Preprint, arXiv, May 19. https://doi.org/10.48550/arXiv.1409.0473.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, et al. 2014. “Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation.” arXiv:1406.1078. Preprint, arXiv, September 3. https://doi.org/10.48550/arXiv.1406.1078.

“Introduction to Recurrent Neural Networks - GeeksforGeeks.” 2025. Accessed October 6, 2025. https://www.geeksforgeeks.org/machine-learning/introduction-to-recurrent-neural-network/.

Madsen, Andreas. 2019. “Visualizing Memorization in RNNs.” *Distill* 4 (3): e16. https://doi.org/10.23915/distill.00016.

Olah, Christopher. 2015. “Understanding LSTM Networks.” Colah’s Blog. Accessed October 6, 2025. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. “Sequence to Sequence Learning with Neural Networks.” arXiv:1409.3215. Preprint, arXiv, December 14. https://doi.org/10.48550/arXiv.1409.3215.

Karpathy, Andrej. 2015. “The Unreasonable Effectiveness of Recurrent Neural Networks.” Accessed October 6, 2025. https://karpathy.github.io/2015/05/21/rnn-effectiveness/.

Colab

Simple single layer RNN to predict temperature:

https://colab.research.google.com/drive/15NWGJ8gSKSVJ4TNh87sQMU1hBNQ8CkbT?usp=sharing