Natural Language Processing with Deep Learning

CS224N/Ling284

Christopher Manning and Richard Socher

Lecture 10: Machine Translation and Models with Attention

Lecture Plan:
Going forwards and backwards

1. Translation, Machine Translation, Neural Machine Translation
2. Research highlight: Google's new NMT
3. Sequence models with **attention**
4. Sequence model decoders

Reminders/comments:

Midterm is over and graded (99%) ☺

Assignment 3 is looming ☹

Learn up on GPUs, Azure, Docker

Final project discussions – come meet with us!

1. Machine Translation

The classic test of language understanding!

Both language analysis & generation

Big MT needs … for humanity … and commerce

Translation is a US$40 billion a year industry

Huge in Europe, growing in Asia

Large social/government/military as well as commercial needs

The need for machine translation

Huge commercial use

- Google translates over 100 billion words a day
- Facebook in 2016 rolled out new homegrown MT
  - "When we turned [MT] off for some people, they went nuts!"
- eBay uses MT to enable cross-border trade

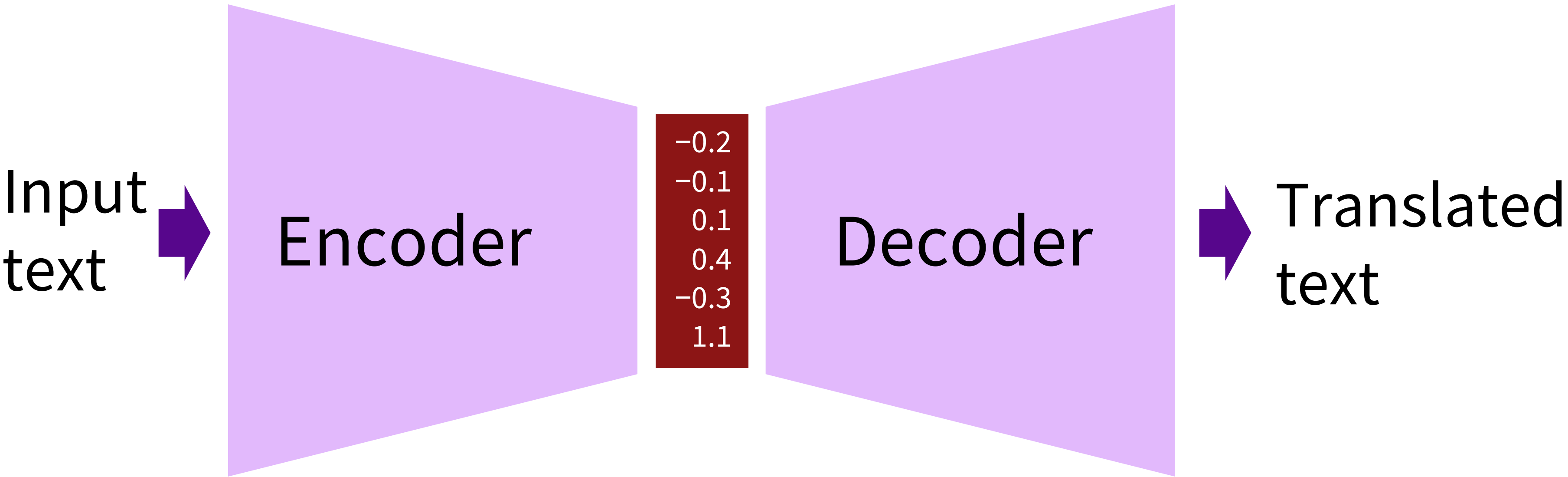
http://www.commonsenseadvisory.com/AbstractView.aspx?ArticleID=36540
https://googleblog.blogspot.com/2016/04/ten-years-of-google-translate.html
https://techcrunch.com/2016/05/23/facebook-translation/

What is Neural MT (NMT)?

Neural Machine Translation is the approach of modeling the entire MT process via one big artificial neural network*

*But sometimes we compromise this goal a little

Neural encoder-decoder architectures

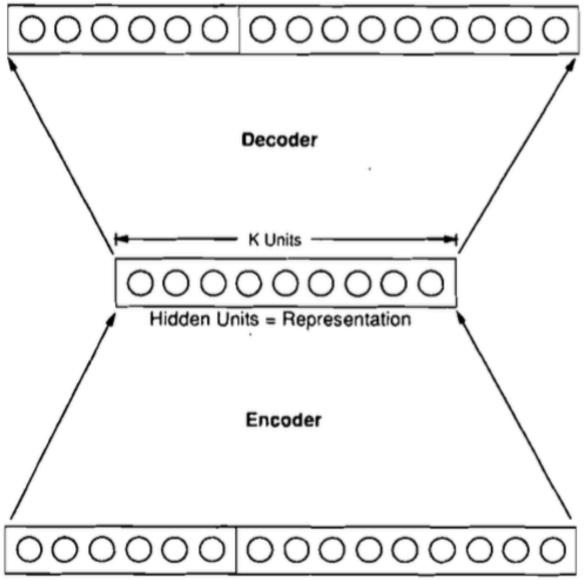


Neural MT: The Bronze Age

[Allen 1987 IEEE 1st ICNN]

3310 En-Es pairs constructed on 31 En, 40 Es words, max 10/11 word sentence; 33 used as test set

The grandfather offered the little girl a book ➔
El abuelo le ofrecio un libro a la nina pequena



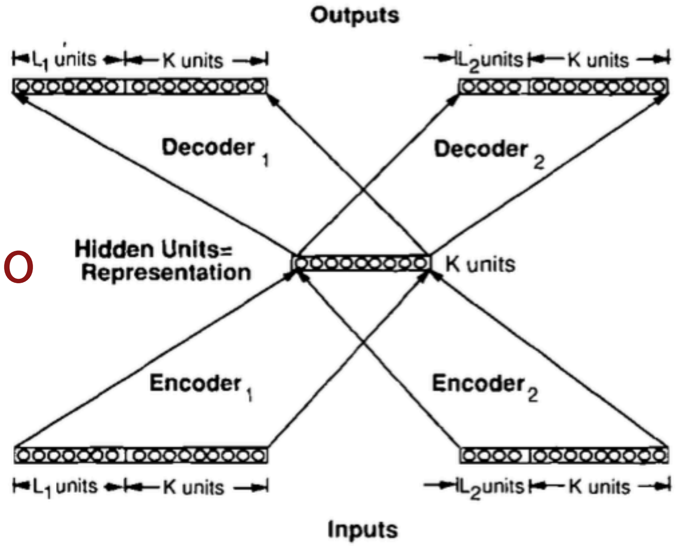
Binary encoding of words – 50 inputs, 66 outputs; 1 or 3 hidden 150-unit layers. Ave WER: 1.3 words

Neural MT: The Bronze Age

[Chrisman 1992 *Connection Science*]

Dual-ported RAAM architecture [Pollack 1990 *Artificial Intelligence*] applied to corpus of 216 parallel pairs of simple En-Es sentences:

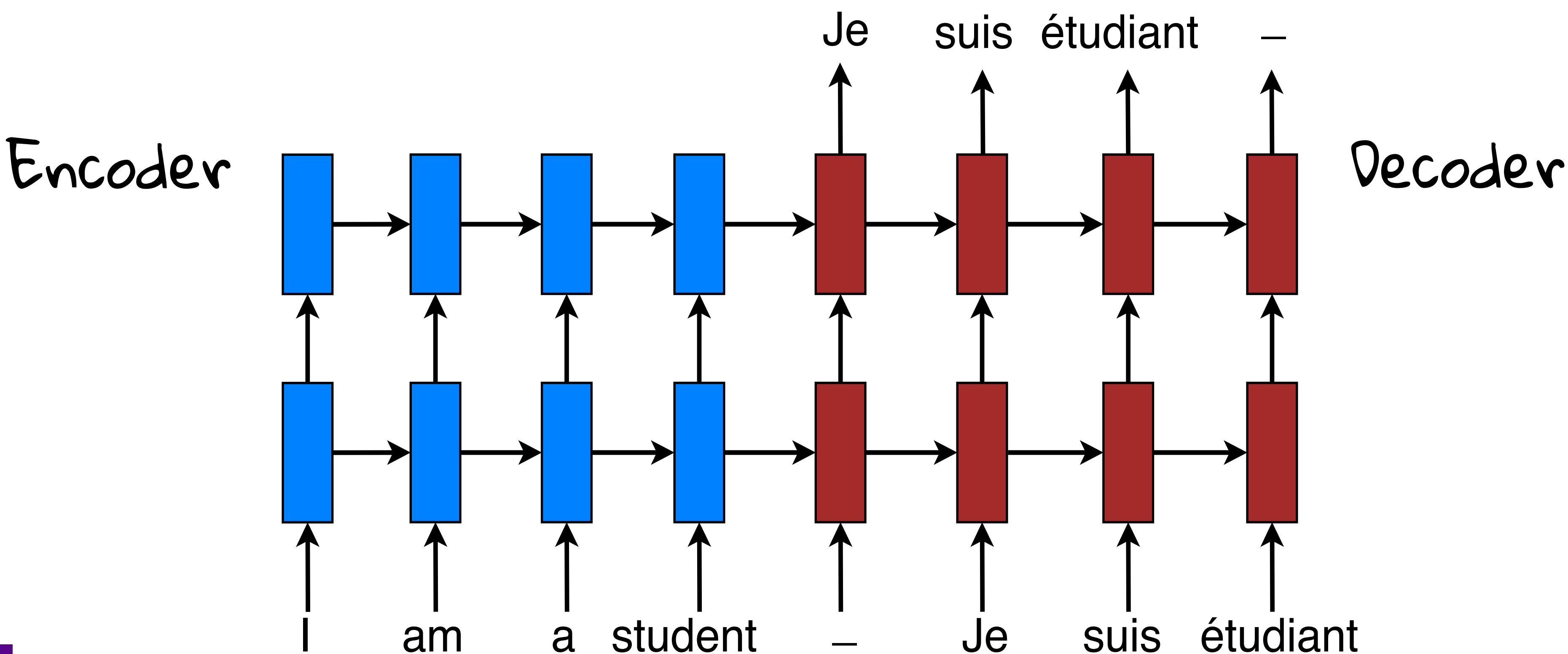
You are not angry ↔ Usted no esta furioso



Split 50/50 as train/test, 75% of sentences correctly translated!

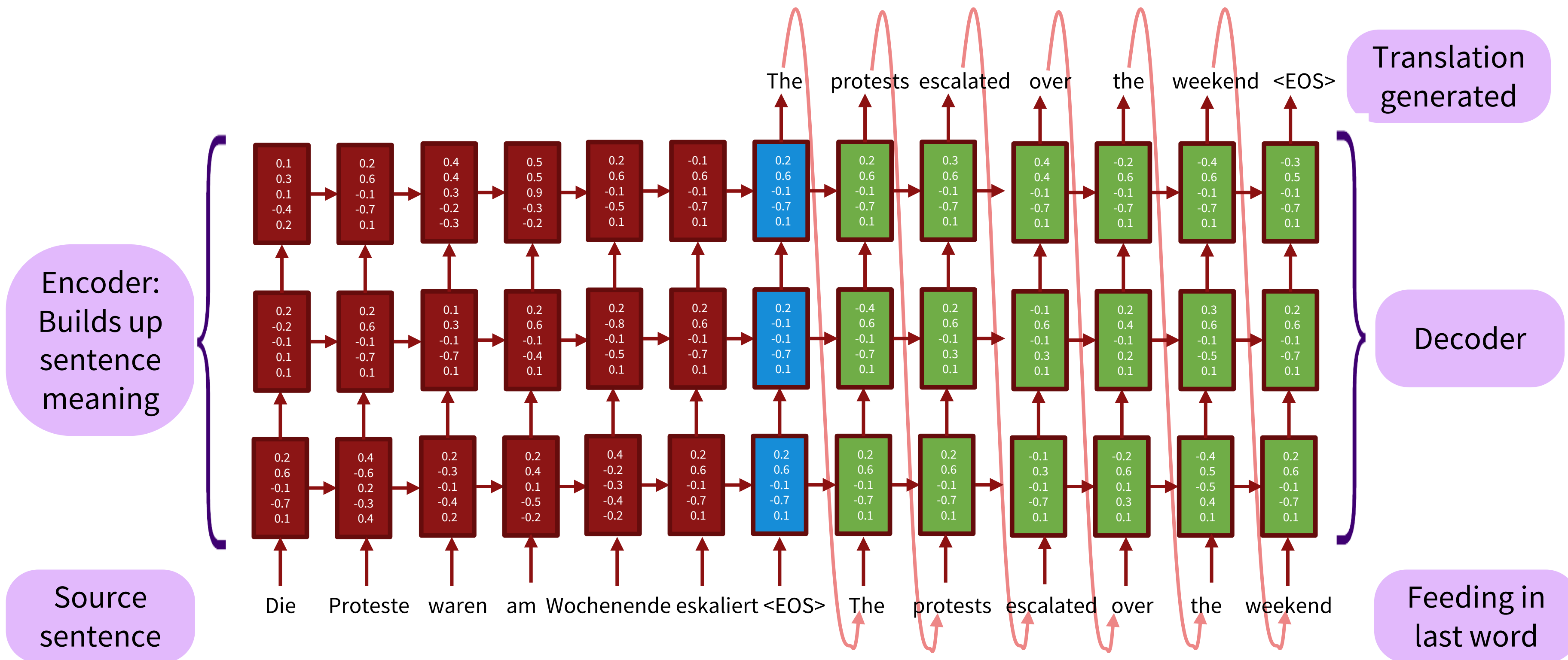
Modern Sequence Models for NMT

[Sutskever et al. 2014, cf. Bahdanau et al. 2014, et seq.]



Modern Sequence Models for NMT

[Sutskever et al. 2014, cf. Bahdanau et al. 2014, et seq.]



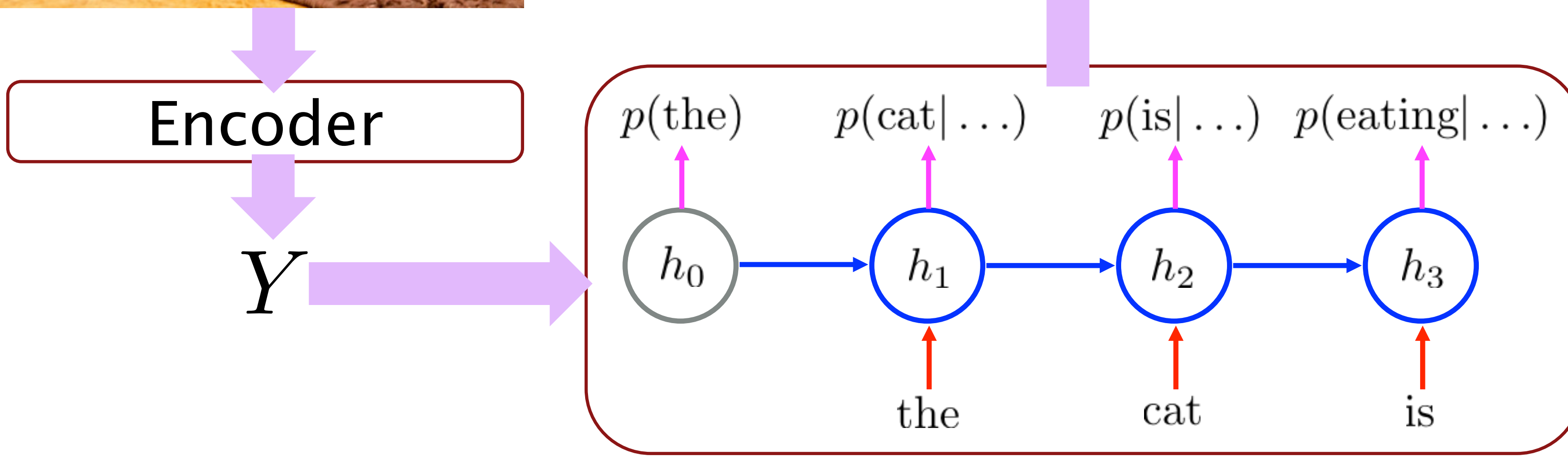
A deep recurrent neural network

Conditional Recurrent Language Model

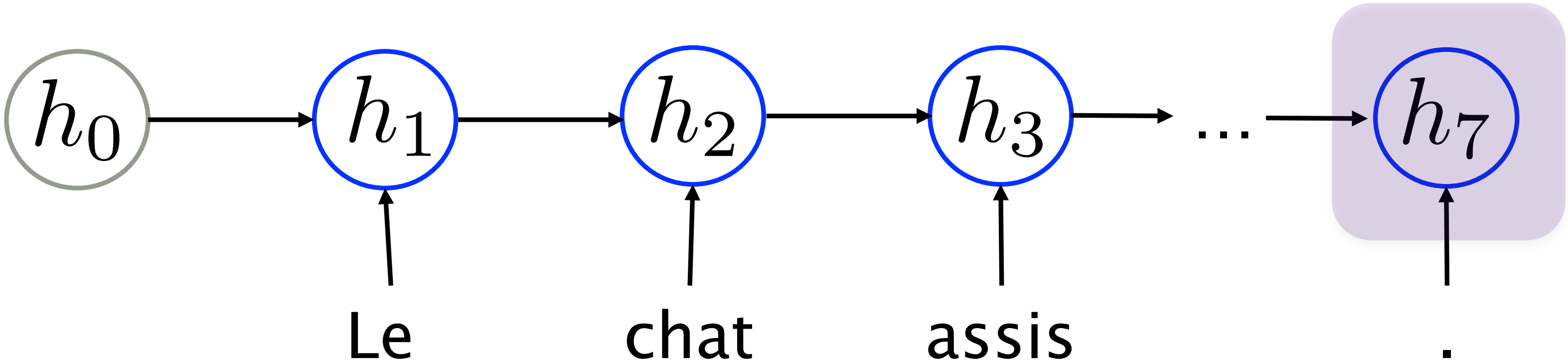
Le chat assis sur le tapis.

?

The cat sat on the mat.

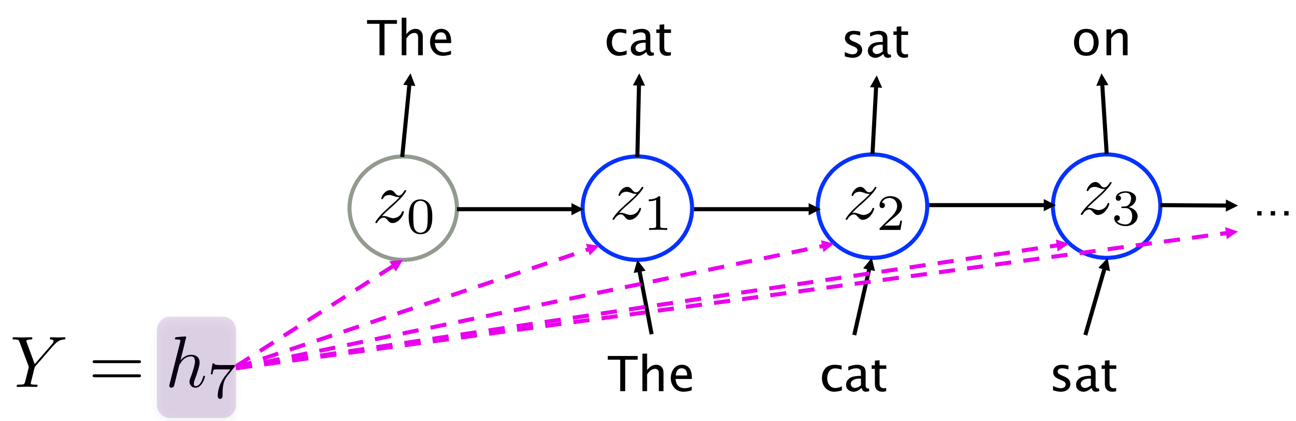


Recurrent Neural Network Encoder



- Read a source sentence one symbol at a time.
- The last hidden state Y summarizes the entire source sentence.
- Any recurrent activation function can be used:
    - Hyperbolic tangent $\tanh$
    - Gated recurrent unit [Cho et al., 2014]
    - Long short-term memory [Sutskever et al., 2014]
    - Convolutional network [Kalchbrenner & Blunsom, 2013]

Decoder: Recurrent Language Model

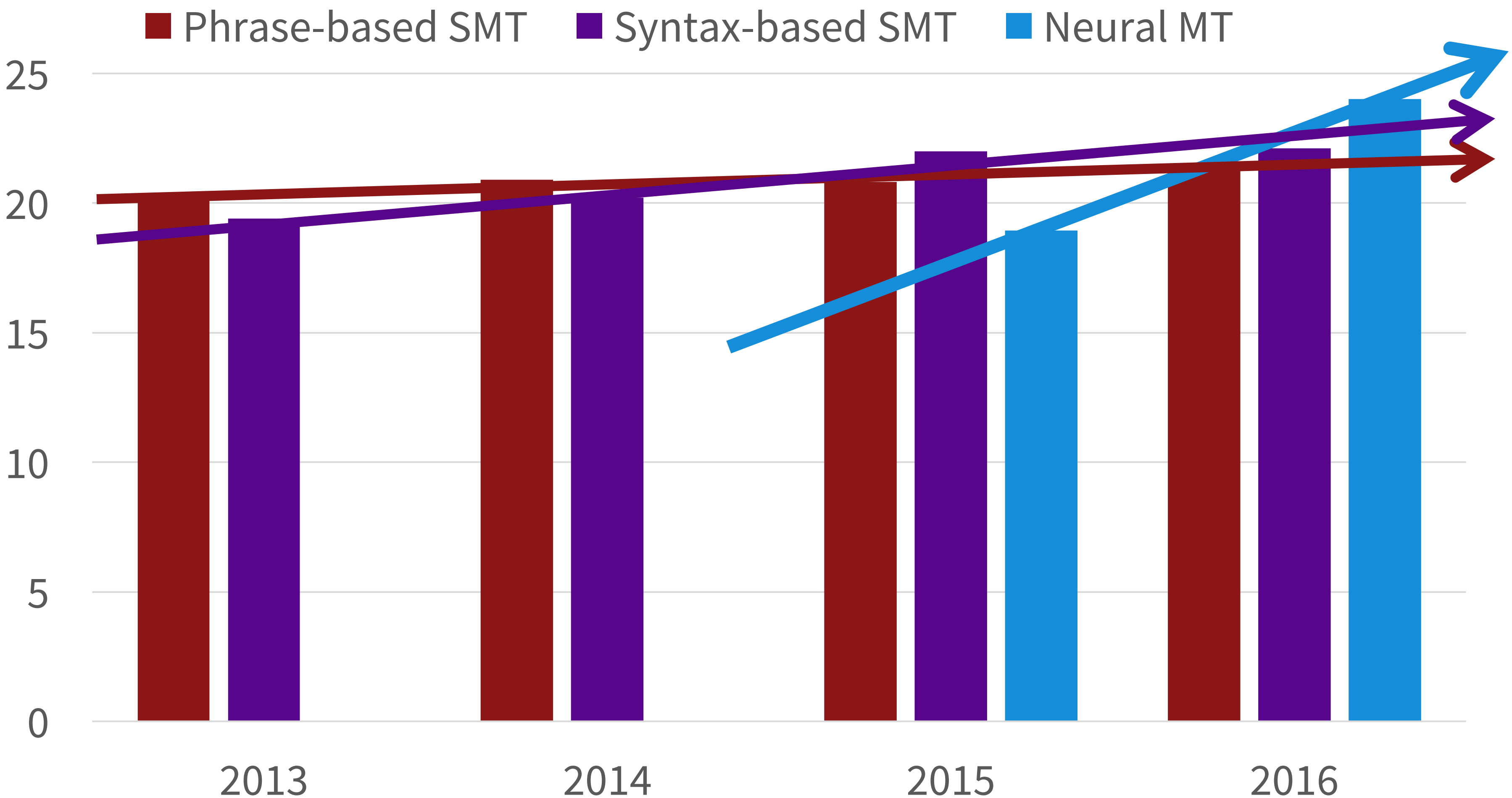


- Usual recurrent language model, except
  1. Transition $z_t = f(z_{t-1}, x_t, Y)$
  2. Backpropagation $\sum_t \partial z_t / \partial Y$
- Same learning strategy as usual: MLE with SGD

$$\mathcal{L}(\theta, D) = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T^n} \log p(x_t^n | x_1^n, \ldots, x_{t-1}^n, Y)$$

Progress in Machine Translation

[Edinburgh En-De WMT newstest2013 Cased BLEU; NMT 2015 from U. Montréal]



From [Sennrich 2016, http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf]

Neural MT went from a fringe research activity in 2014 to the widely-adopted leading way to do MT in 2016.

Amazing!

Four big wins of Neural MT

1. **End-to-end training**
   All parameters are simultaneously optimized to minimize a loss function on the network's output
2. **Distributed representations share strength**
   Better exploitation of word and phrase similarities
3. **Better exploitation of context**
   NMT can use a much bigger context – both source and partial target text – to translate more accurately
4. **More fluent text generation**
   Deep learning text generation is much higher quality

What wasn't on that list?

1. Black box component models for reordering, transliteration, etc.
2. Explicit use of syntactic or semantic structures
3. Explicit use of discourse structure, anaphora, etc.

Statistical/Neural Machine Translation

A **marvelous** use of **big data** but....

1519年600名西班牙人在墨西哥登陆，去征服几百万人口的阿兹特克帝国，初次交锋他们损兵三分之二。

In 1519, six hundred Spaniards landed in Mexico to conquer the Aztec Empire with a population of a few million. They lost two thirds of their soldiers in the first clash.

translate.google.com (2009): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of soldiers against their loss.

translate.google.com (2011): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the initial loss of soldiers, two thirds of their encounters.

translate.google.com (2013): 1519 600 Spaniards landed in Mexico to conquer the Aztec empire, hundreds of millions of people, the initial confrontation loss of soldiers two-thirds.

translate.google.com (2014/15/16): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of the loss of soldiers they clash.

translate.google.com (2017): In 1519, 600 Spaniards landed in Mexico, to conquer the millions of people of the Aztec empire, the first confrontation they killed two-thirds.

Adoption!!! NMT aggressively rolled out by industry!

2016/02, Microsoft launches deep neural network MT running offline on Android/iOS. [Link to blog]

2016/08, Systran launches purely NMT model [Link to press release]

2016/09, Google launches NMT [Link to blog post]

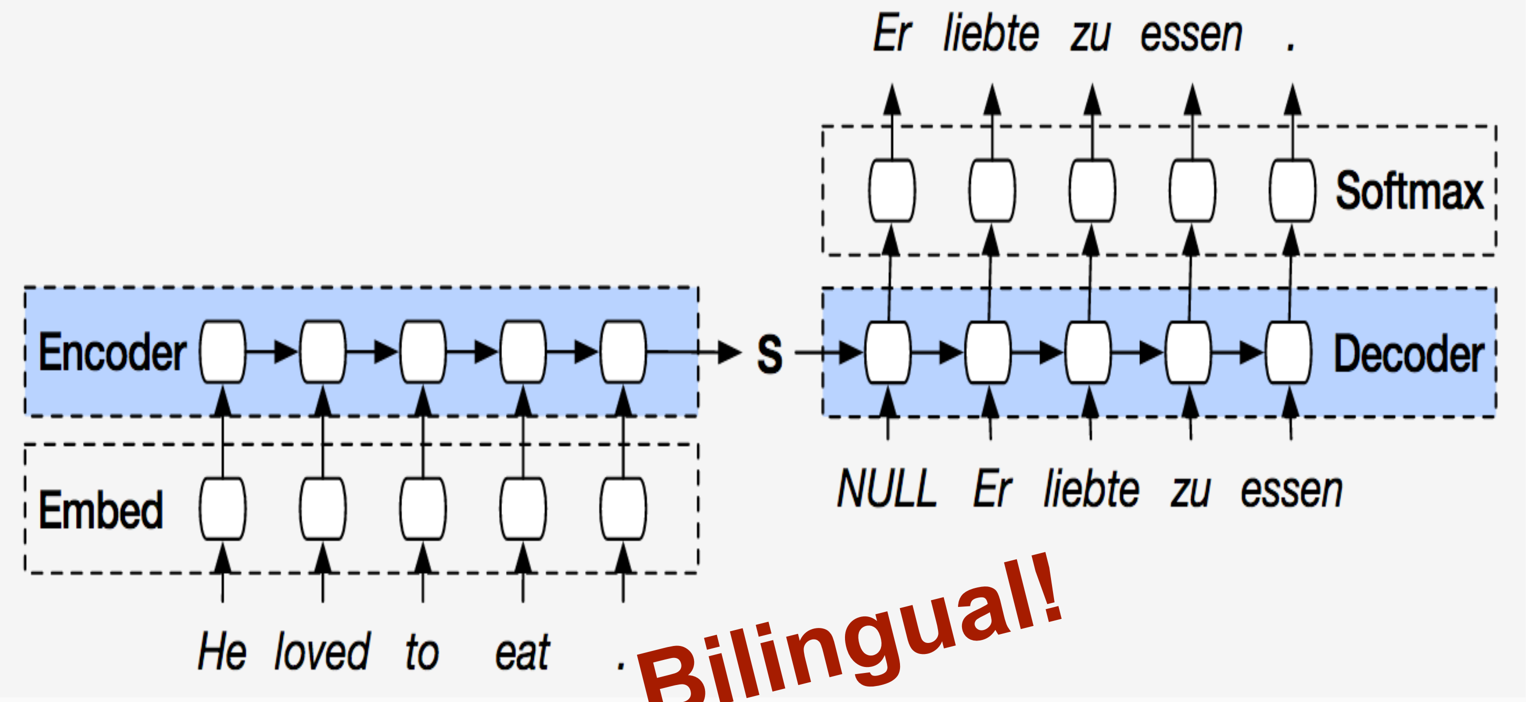
- With much more hype and gross overclaims of equaling human translation quality
- Great New York Times Magazine feature
- Paper on the research: https://arxiv.org/abs/1611.04558

Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat,Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, Jeffrey Dean

Presented by: Emma Peng

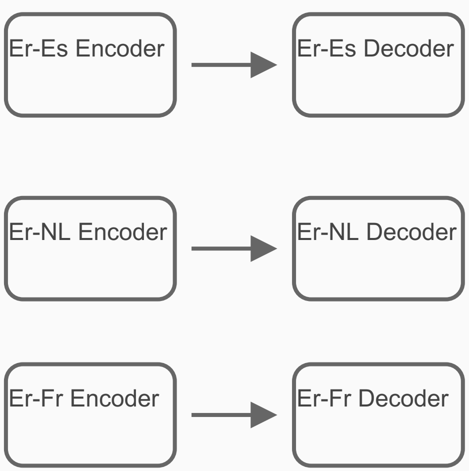
State-of-the-art:
Neural Machine Translation (NMT)



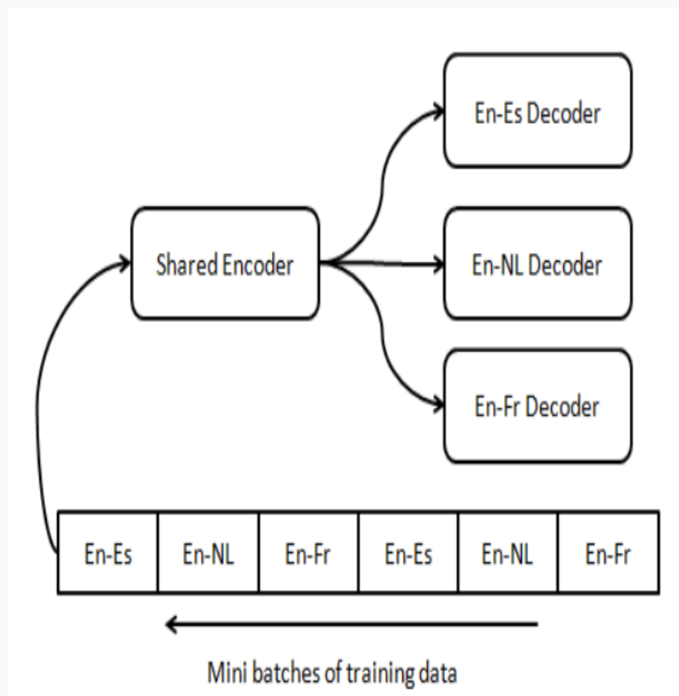
Bilingual!

Multilingual NMT? Previously...

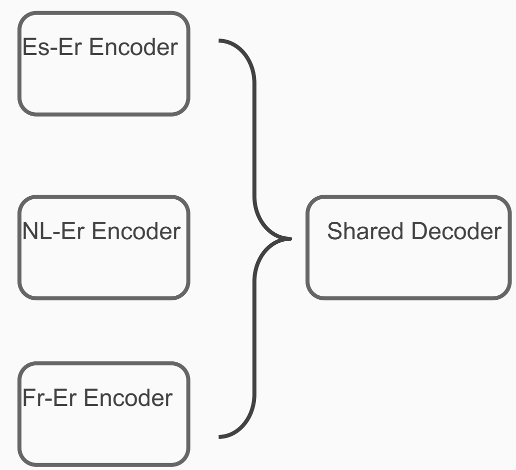
Multiple Encoders → Multiple Decoders [1]



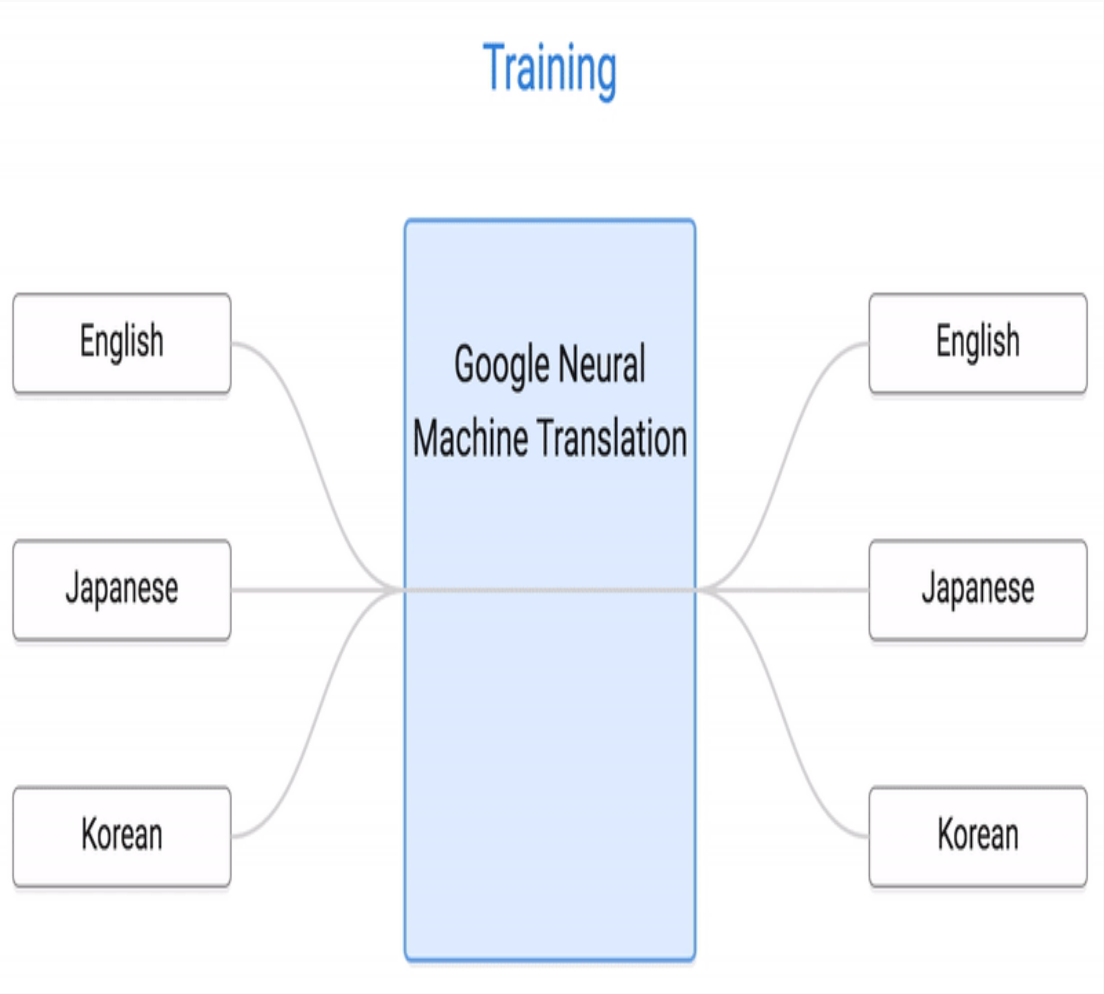
Shared Encoder → Multiple Decoder [2]



Multiple Encoders → Shared Decoder [3]

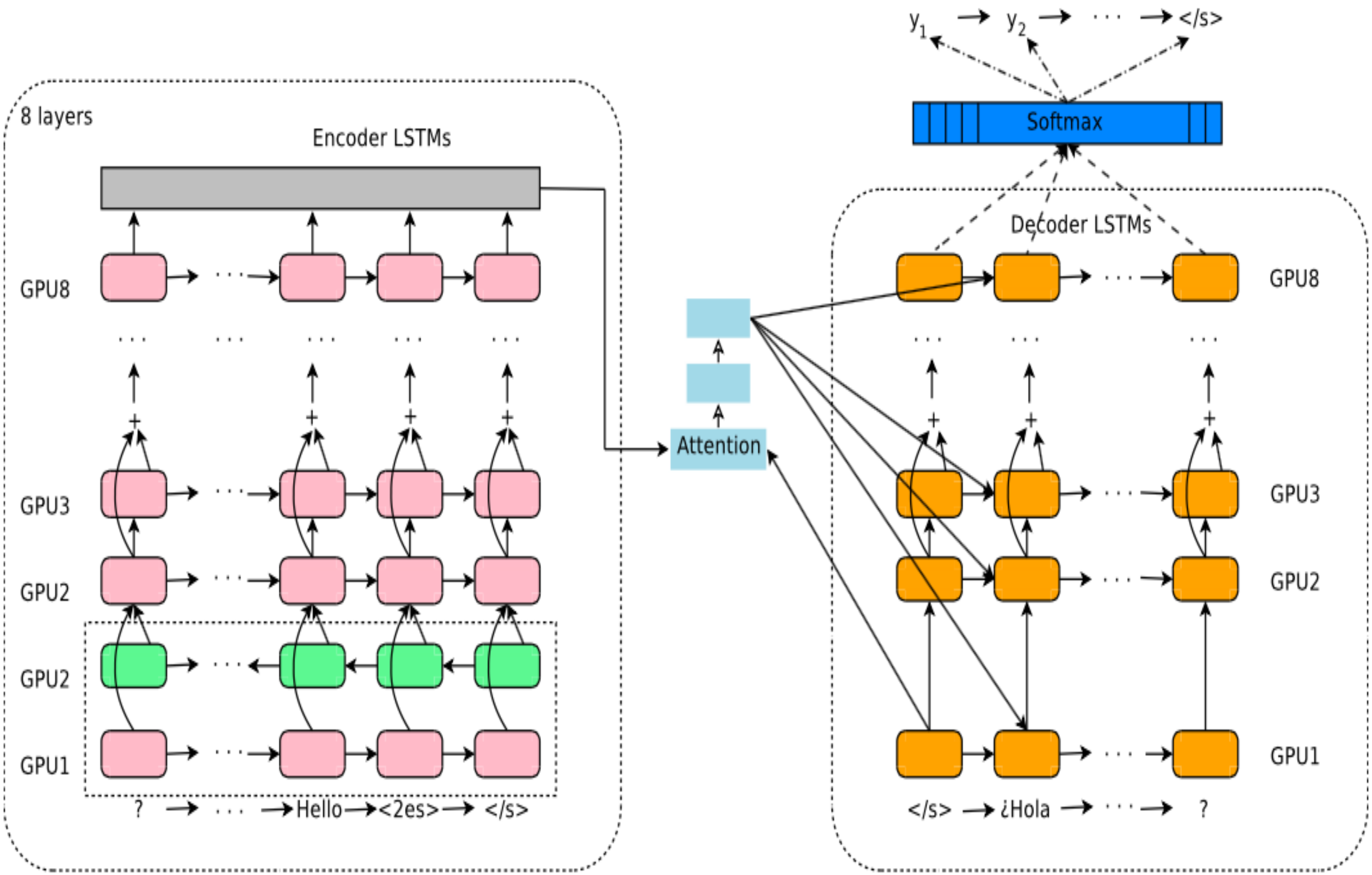


Google's Multilingual NMT System Benefits



- **Simplicity**: single model
- **Low-resource language improvements**
- **Zero-shot translation**

Google's Multilingual NMT System Architecture



Google's Multilingual NMT System Architecture

Artificial token at the beginning of the input sentence to indicate the target language

```
Hello, how are you? -> ¿Hola como estás?
```

Add <2es> to indicate that Spanish is the target language

```
<2es> Hello, how are you? -> ¿Hola como estás?
```

Google's Multilingual NMT System Experiments

- WMT'14:
  - Comparable performance: English → French
  - State-of-the-art: English → German, French → English
- WMT'15:
  - State-of-the-art: German → English

Google's Multilingual NMT System Zero-Shot Translation

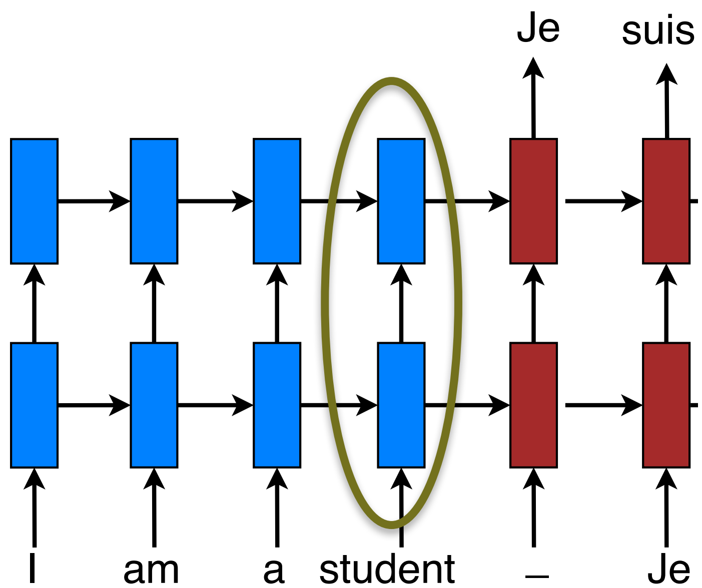
Table 5: Portuguese→Spanish BLEU scores using various models.

	Model	BLEU
(a)	PBMT bridged	28.99
(b)	NMT bridged	30.91
(c)	NMT Pt→Es	31.50
(d)	Model 1 (Pt→En, En→Es)	21.62
(e)	Model 2 (En↔{Es, Pt})	24.75
(f)	Model 2 + incremental training	31.77

- **Train**:
  - Portuguese → English, English → Spanish (Model 1)
  - Or, English ⟷ {Portuguese, Spanish} (Model 2)
- **Test**:
  - Portuguese → Spanish **Zero-Shot!**

Thank you!

3. Introducing Attention: Vanilla seq2seq & long sentences

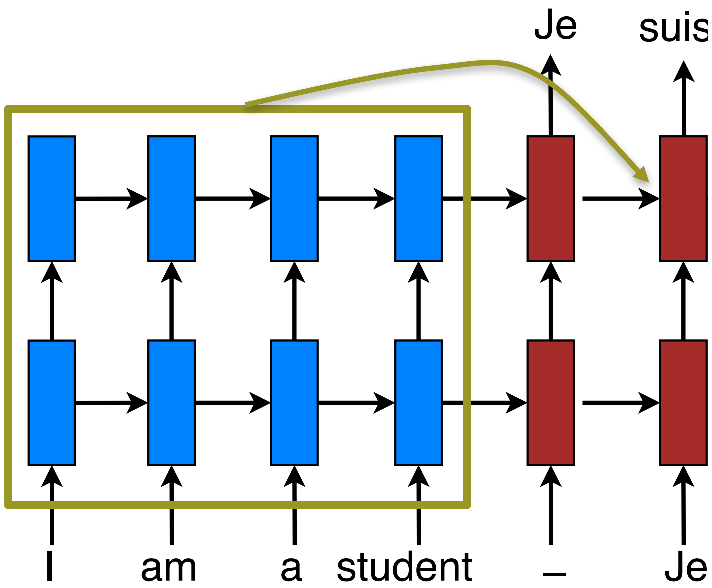


Problem: fixed-dimensional representation Y

Attention Mechanism

Started in computer vision! [Larochelle & Hinton, 2010], [Denil, Bazzani, Larochelle, Freitas, 2012]

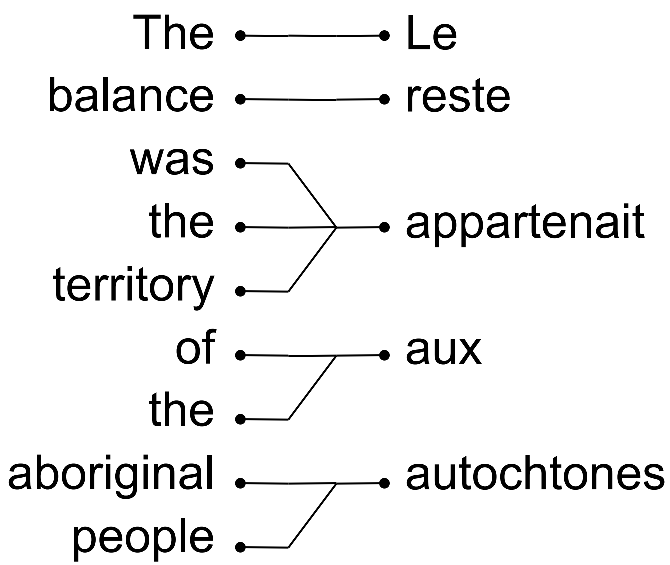
Pool of source states

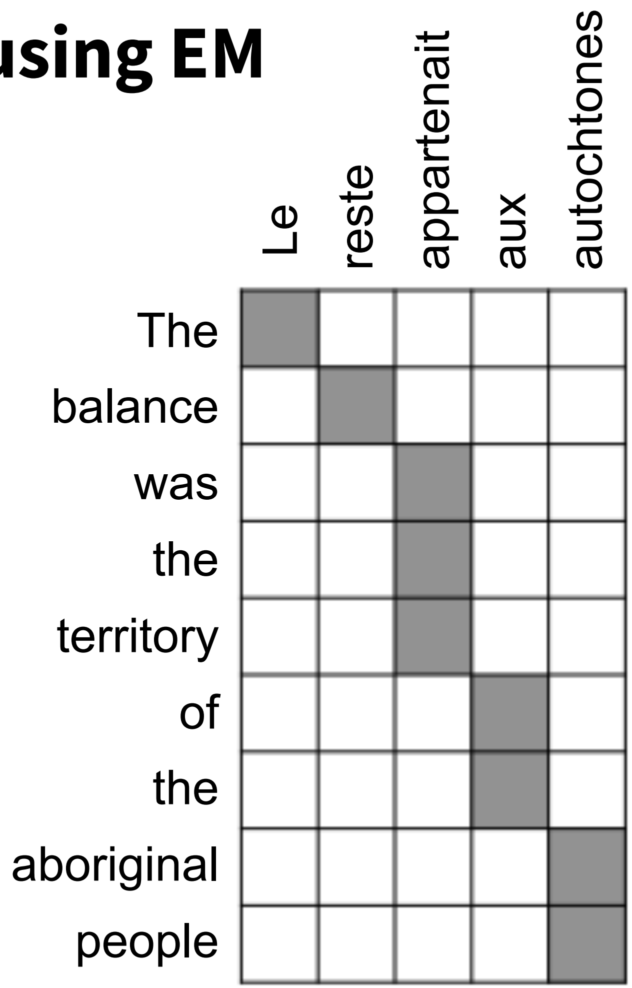


- Solution: random access memory
  - Retrieve as needed.

Word alignments

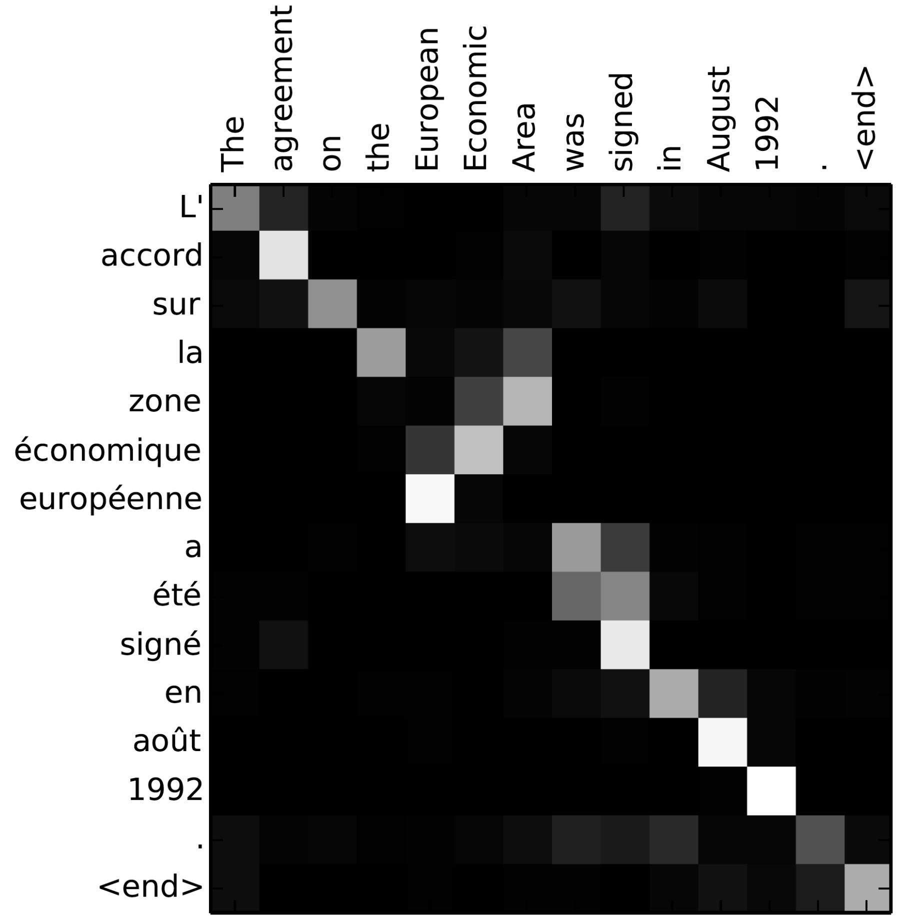
Phrase-based SMT aligned words in a preprocessing-step, usually using EM





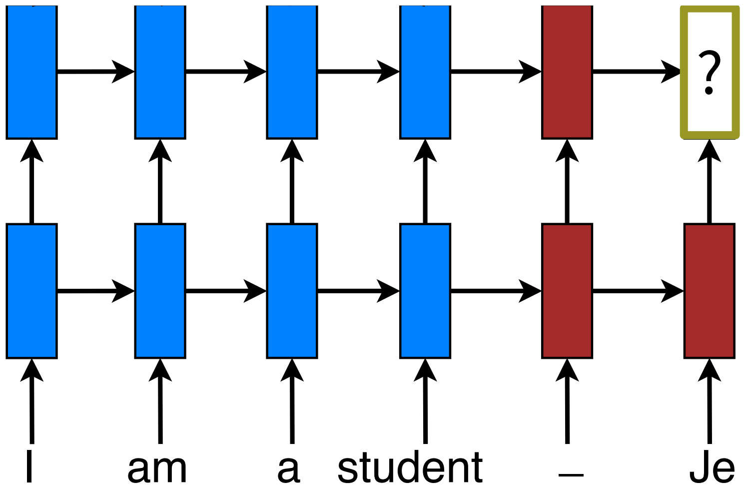
Learning both translation & alignment





Dzmitry Bahdanau, KyungHuyn Cho, and Yoshua Bengio. ***Neural Machine Translation by Jointly Learning to Translate and Align****. ICLR'15.*

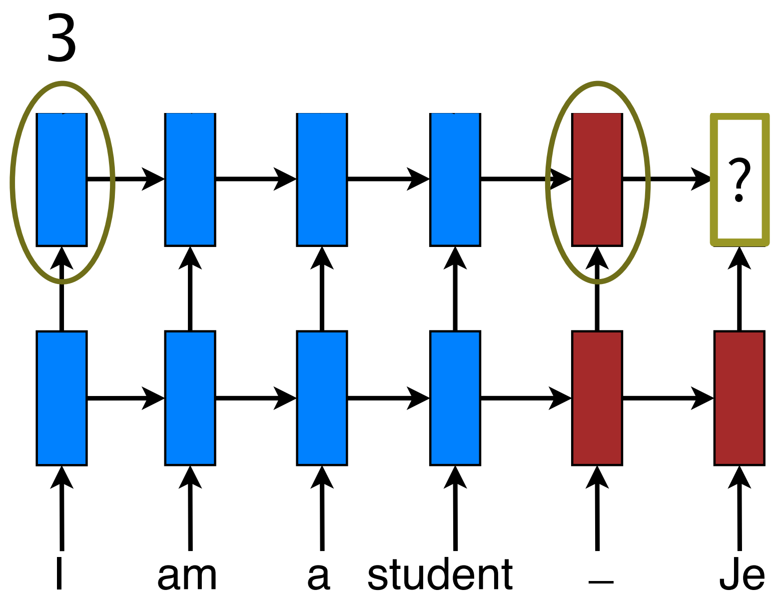
Attention Mechanism



Simplified version of (Bahdanau et al., 2015)

Attention Mechanism – *Scoring*

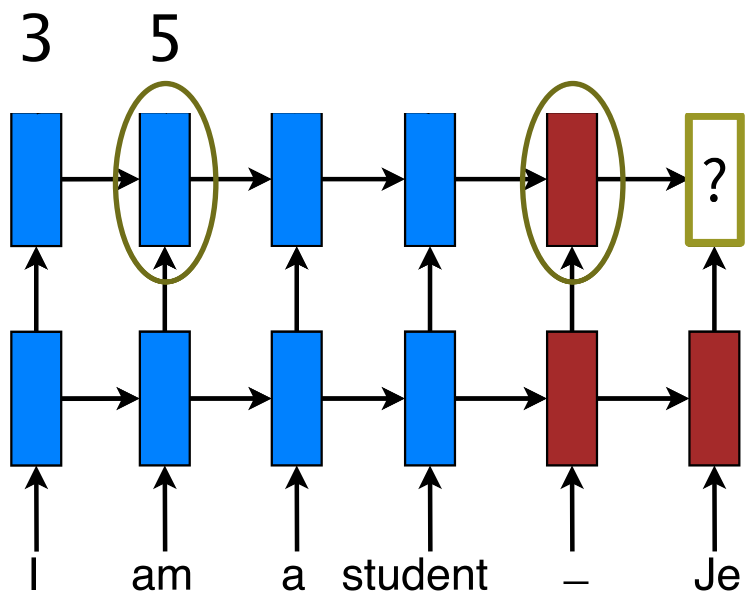
$$\text{score}(\boldsymbol{h}_{t-1}, \bar{\boldsymbol{h}}_s)$$



- Compare target and source hidden states.

Attention Mechanism – *Scoring*

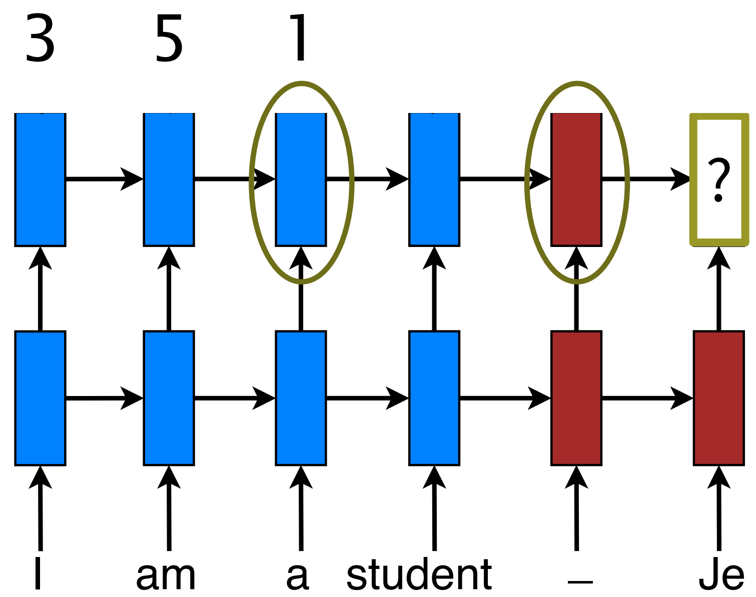
$$\mathrm{score}(\boldsymbol{h}_{t-1}, \bar{\boldsymbol{h}}_s)$$



- Compare target and source hidden states.

Attention Mechanism – *Scoring*

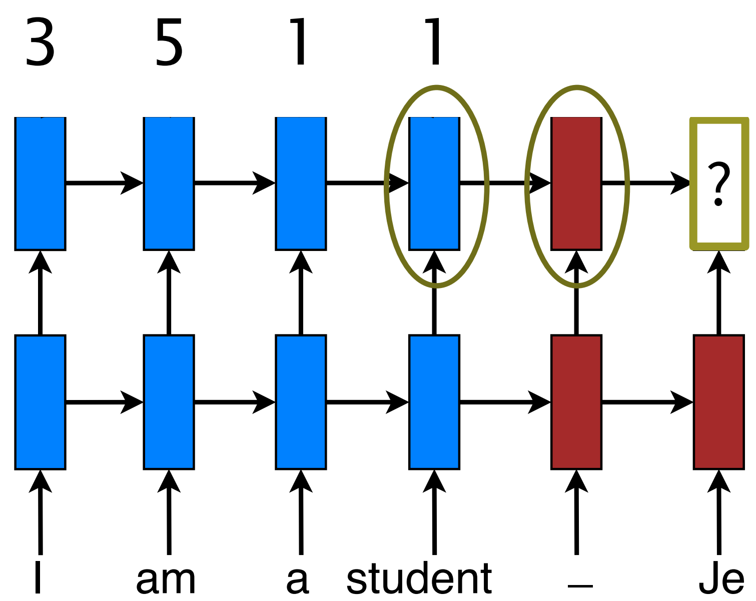
$$\mathrm{score}(\boldsymbol{h}_{t-1}, \bar{\boldsymbol{h}}_s)$$



- Compare target and source hidden states.

Attention Mechanism – *Scoring*

$$\mathrm{score}(\boldsymbol{h}_{t-1}, \bar{\boldsymbol{h}}_s)$$

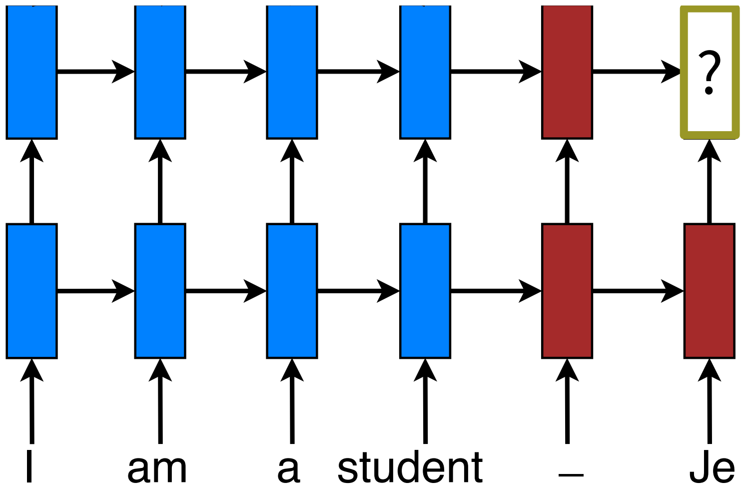


- Compare target and source hidden states.

Attention Mechanism – *Normalization*

$$a_t(s) = \frac{e^{\mathrm{score}(s)}}{\sum_{s'} e^{\mathrm{score}(s')}}$$

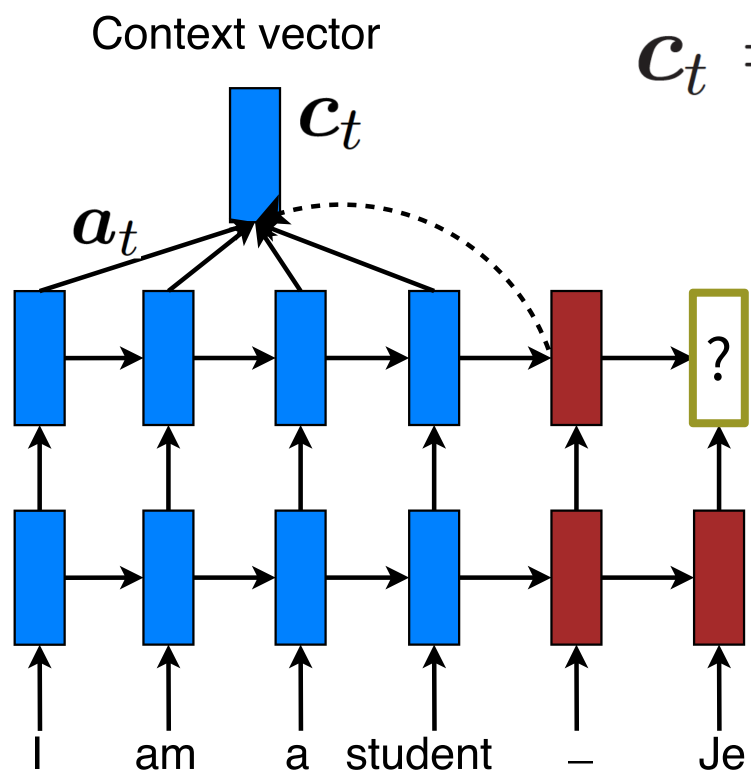
a_t 0.3 0.5 0.1 0.1



- Convert into alignment weights.

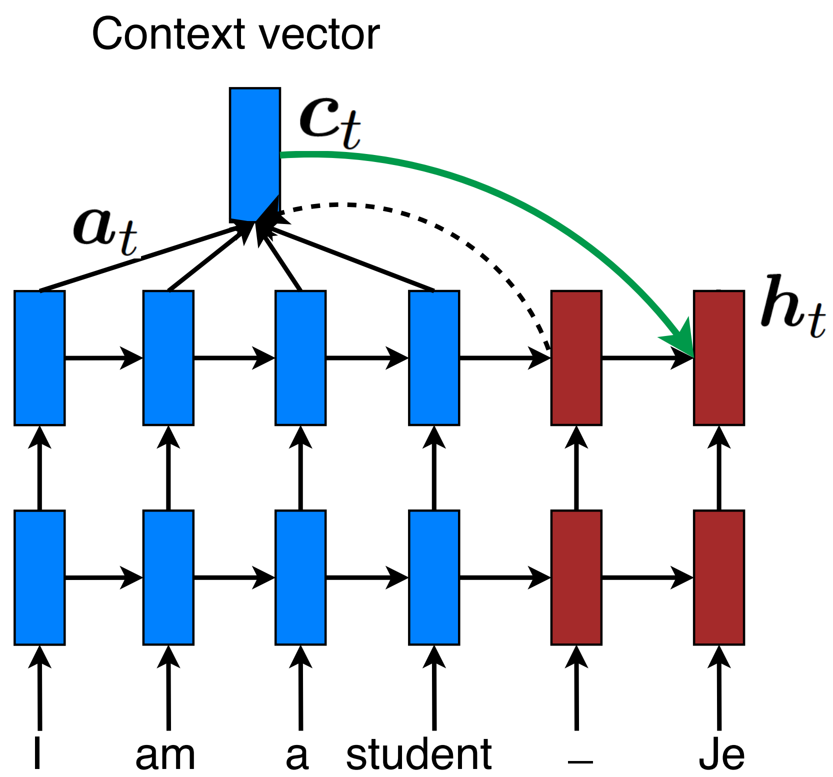
Attention Mechanism – *Context*

$$c_t = \sum_s a_t(s)\bar{h}_s$$



- Build context vector: weighted average.

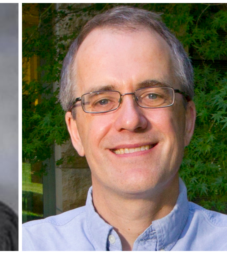
Attention Mechanism – *Hidden State*



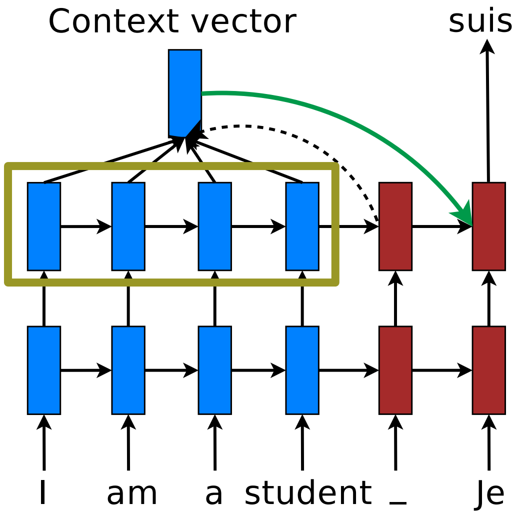
- Compute the next hidden state.

Attention Mechanisms+





- Simplified mechanism & more functions:



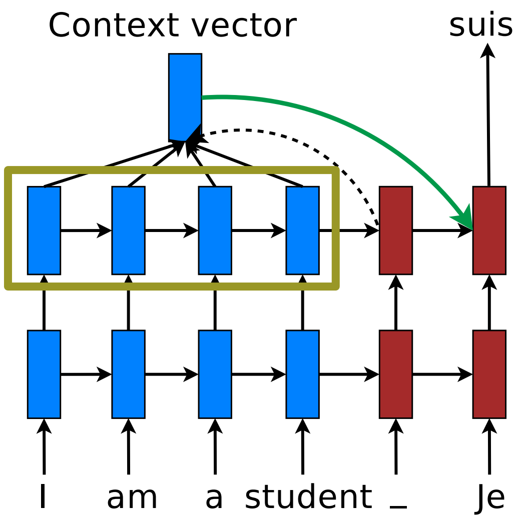
$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s \\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) \end{cases}$$

Thang Luong, Hieu Pham, and Chris Manning. ***Effective Approaches to Attention-based Neural Machine Translation****. EMNLP'15.*

Attention Mechanisms+



- Simplified mechanism & more functions:



Bilinear form: well-adopted.

$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s \\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) \end{cases}$$

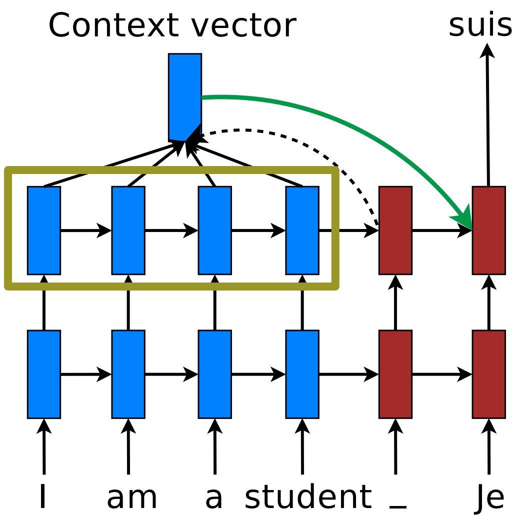
GitHub, Inc. [US] https://github.com/harvardnlp/seq2seq-attn

Sequence-to-Sequence Learning with Attentional Neural Networks

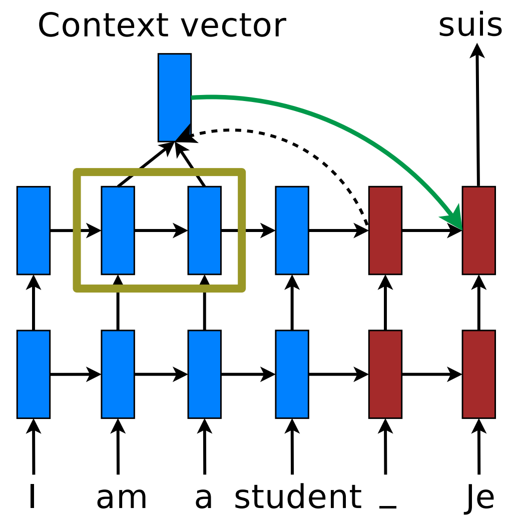
The attention model is from Effective Approaches to Attention-based Neural Machine Translation, Luong et al. EMNLP 2015. We use the *global-general-attention* model with the *input-feeding* approach from the paper. Input-feeding is optional and can be turned off.

Global vs. Local

- Avoid focusing on everything at each time



Global: **all** source states.

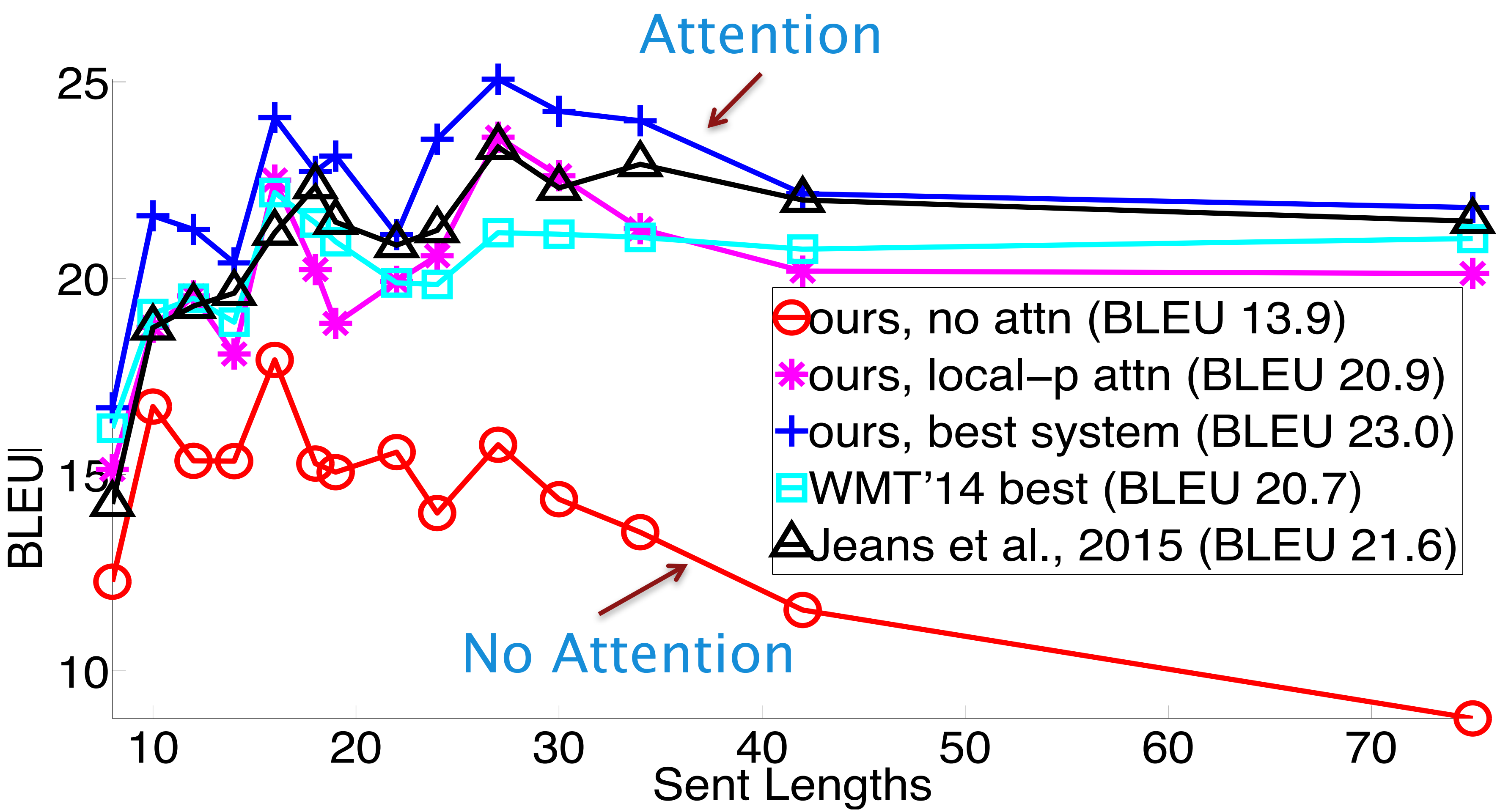


Local: subset of source states.

Potential for long sequences!

Thang Luong, Hieu Pham, and Chris Manning. ***Effective Approaches to Attention-based Neural Machine Translation****. EMNLP'15.*

Better Translation of Long Sentences



Sample English-German translations

source	Orlando Bloom and *Miranda Kerr* still love each other
human	Orlando Bloom und **Miranda Kerr** lieben sich noch immer
+attn	Orlando Bloom und **Miranda Kerr** lieben einander noch immer .
base	Orlando Bloom und **Lucas Miranda** lieben einander noch immer .

- Translates names correctly.

Sample English-German translations

source	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 .
human	Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht **im Wider- spruch zur Sicherheit steht** , sagte Roger Dow , CEO der U.S. Travel Association .
+attn	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 u ̈ber die <unk> , dass ein <unk> <unk> mit Sicherheit nicht **vereinbar** ist mit Sicherheit und Sicherheit , sagte Roger Cameron , CEO der US - <unk> .

- Translates a doubly-negated phrase correctly.

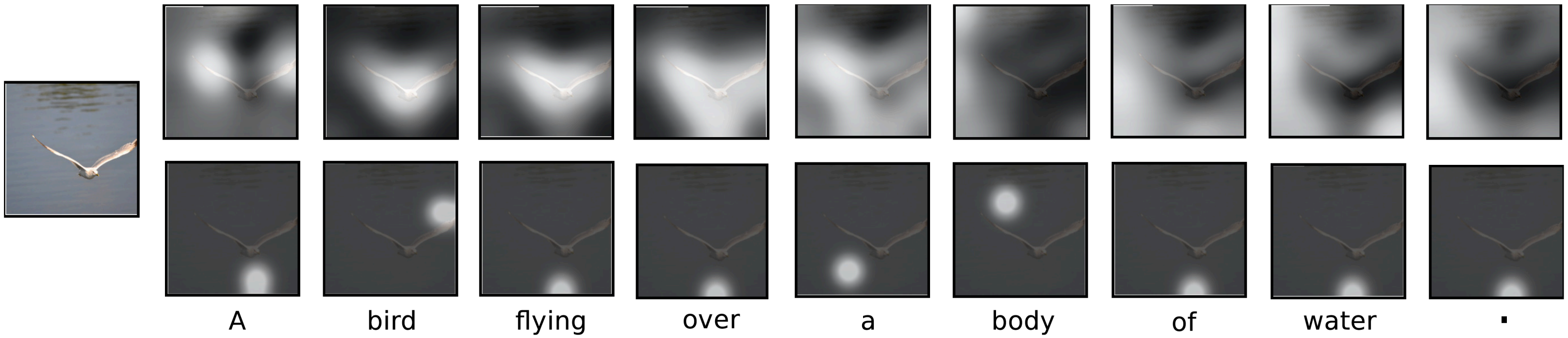
Sample English-German translations

source	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 .
human	Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht **im Wider- spruch zur Sicherheit steht** , sagte Roger Dow , CEO der U.S. Travel Association .
+attn	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 u ̈ber die <unk> , dass ein <unk> <unk> mit Sicherheit nicht **vereinbar** ist mit Sicherheit und Sicherheit , sagte Roger Cameron , CEO der US - <unk> .

- Translates a doubly-negated phrase correctly.

More Attention! *The idea of coverage*

- Caption generation



How to not miss an important image patch?

Xu, Ba, Kiros, Cho, Courville, Salakhutdinov, Zemel, Bengio. ***Show, Attend and Tell: Neural Image Caption Generation with Visual Attention****. ICML'15*

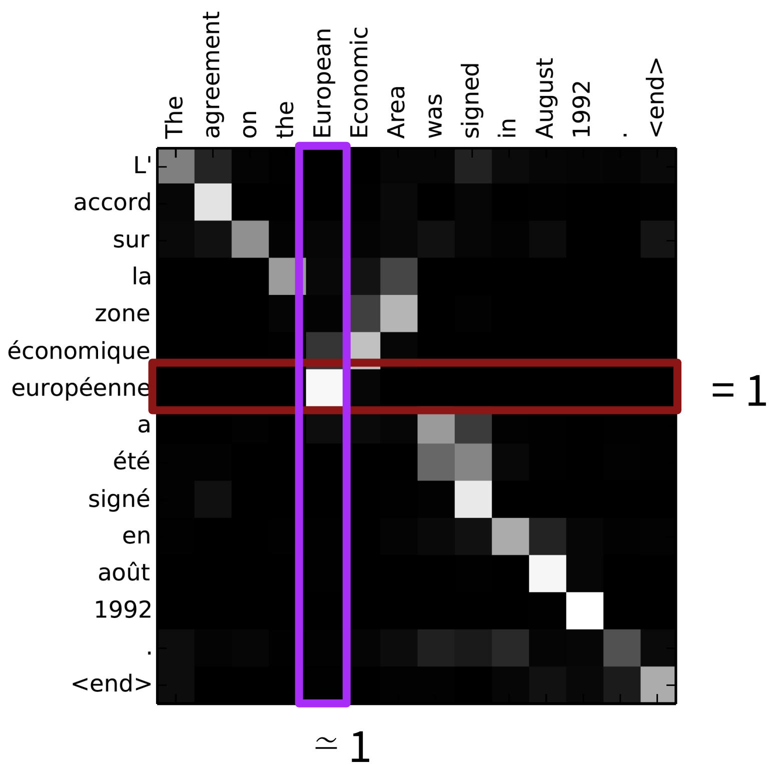
Doubly attention

$$-\log(P(\mathbf{y}|\mathbf{x})) + \lambda \sum_{i}^{L}(1 - \sum_{t}^{C} \alpha_{ti})^2$$

Per image patch

Sum across caption words

- Sum to 1 in both dimensions



Coverage set exists long time ago in SMT!

Xu, Ba, Kiros, Cho, Courville, Salakhutdinov, Zemel, Bengio. ***Show, Attend and Tell: Neural Image Caption Generation with Visual Attention****. ICML'15*

Extending attention with linguistic ideas previously in alignment models

- [Tu, Lu, Liu, Liu, Li, ACL'16]: NMT model with coverage-based attention
- [Cohn, Hoang, Vymolova, Yao, Dyer, Haffari, NAACL'16]: More substantive models of attention using: position (IBM2) + Markov (HMM) + fertility (IBM3-5) + alignment symmetry (BerkeleyAligner)

$$-\log(P(\mathbf{y}|\mathbf{x})) + \lambda \sum_{i}^{L} (1 - \sum_{t}^{C} \alpha_{ti})^2$$

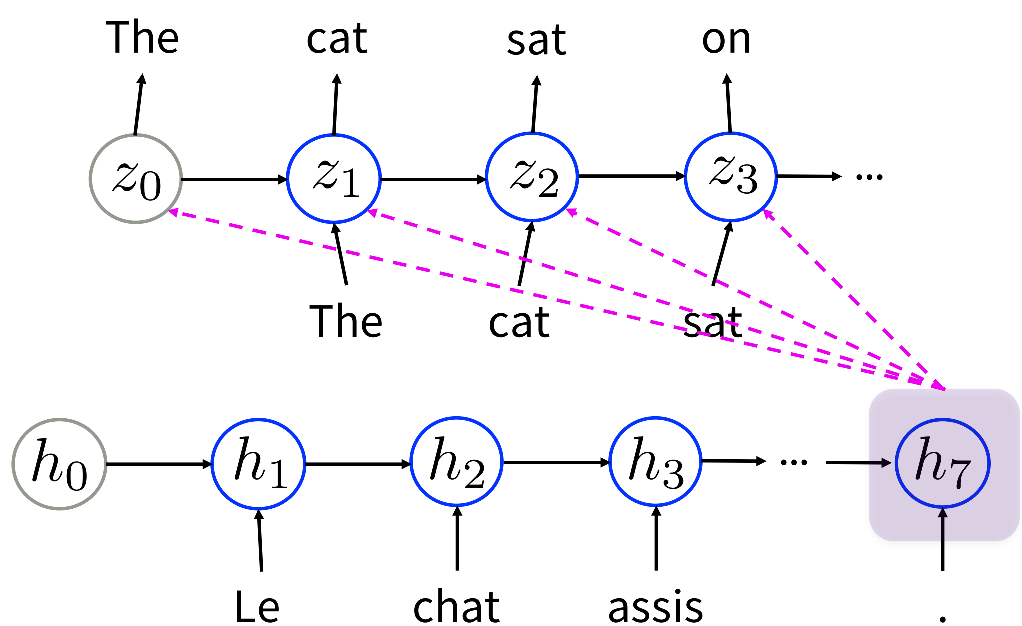
Per source word

Source word fertility

4. Sequence Model Decoders: Decoding (0) – Exhaustive Search

- Simple and exact decoding algorithm
- Score each and every possible translation
- Pick the best one

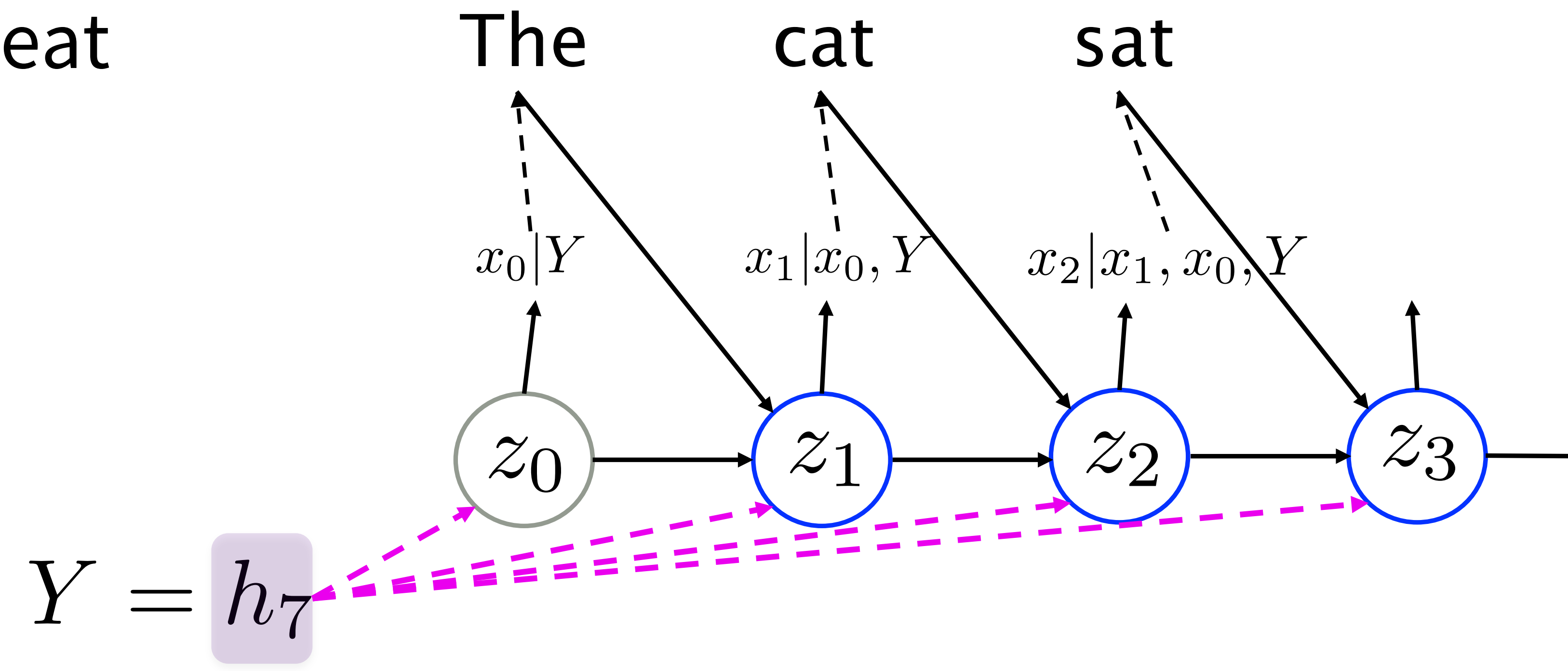
DO NOT EVEN THINK of TRYING IT OUT!*



* Perhaps with quantum computer and quantum annealing?

Decoding (1) – Ancestral Sampling

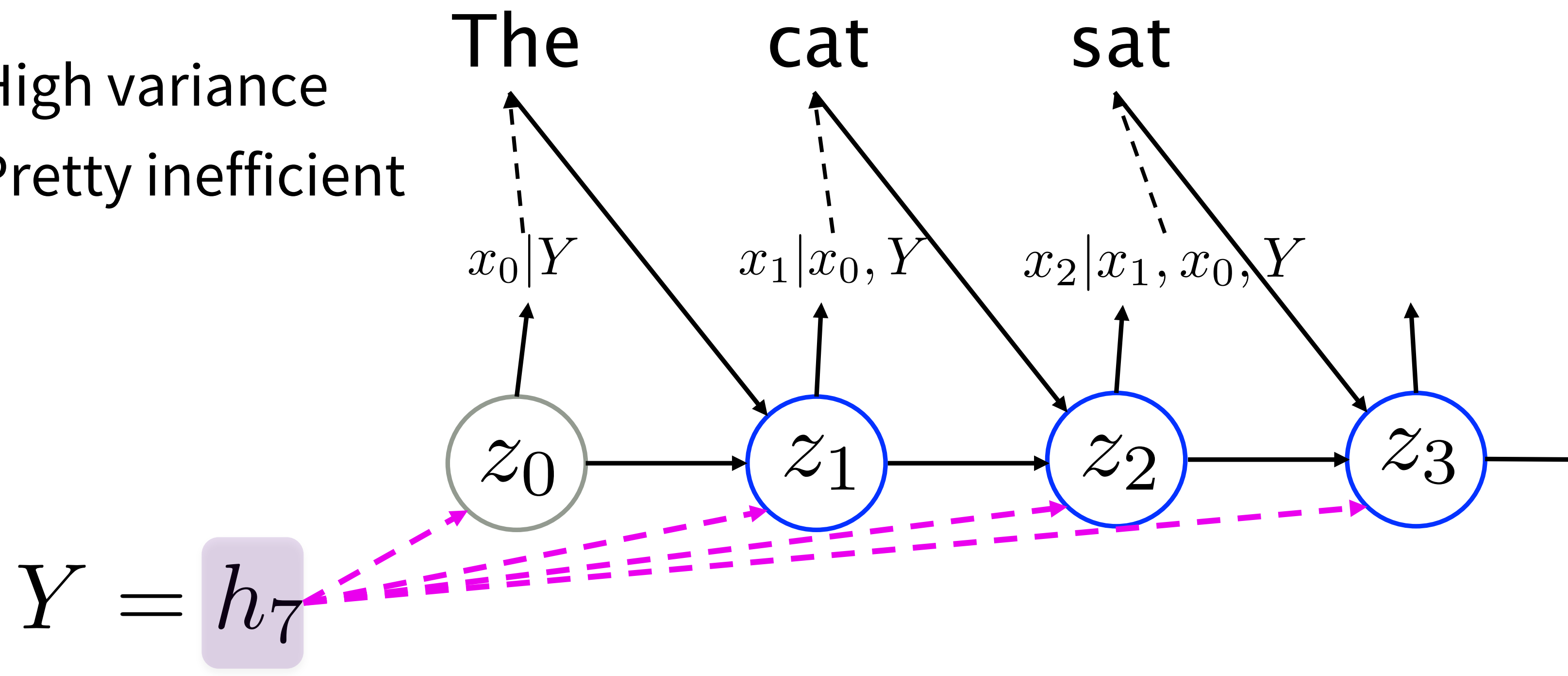
- One symbol at a time from $\tilde{x}_t \sim x_t | x_{t-1}, \ldots, x_1, Y$
- Until $\tilde{x}_t = \langle \text{eos} \rangle$
- Repeat



Decoding (1) – Ancestral Sampling

- Pros:
  1. Efficient and unbiased (asymptotically exact)

- Cons:
  1. High variance
  2. Pretty inefficient

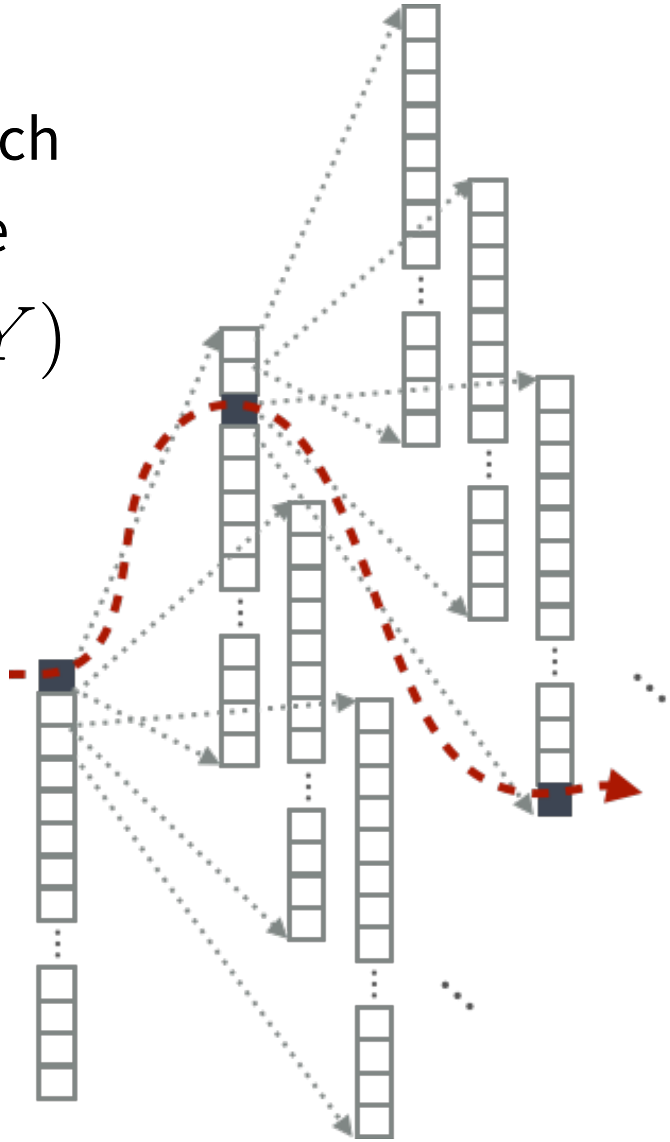


Decoding (2) – Greedy Search

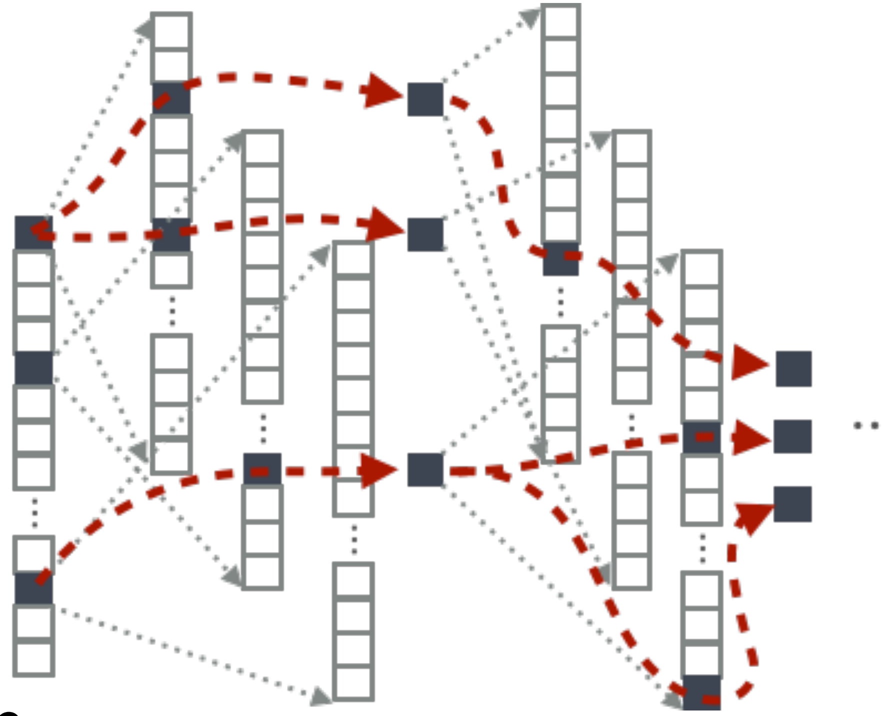
- Efficient, but heavily suboptimal search
- Pick the most likely symbol each time

$$\tilde{x}_t = \arg\max_x \log p(x|x_{<t}, Y)$$

- Until $\tilde{x}_t = \langle \text{eos} \rangle$
- Pros:
  1. Super-efficient
     - Both computation and memory
- Cons:
  1. Heavily suboptimal



Decoding (3) – Beam Search



- Pretty, but *not very* efficient
- Maintain *K* hypotheses at a time

$$\mathcal{H}_{t-1} = \left\{(\tilde{x}_1^1, \tilde{x}_2^1, \ldots, \tilde{x}_{t-1}^1), (\tilde{x}_1^2, \tilde{x}_2^2, \ldots, \tilde{x}_{t-1}^2), \ldots, (\tilde{x}_1^K, \tilde{x}_2^K, \ldots, \tilde{x}_{t-1}^K)\right\}$$

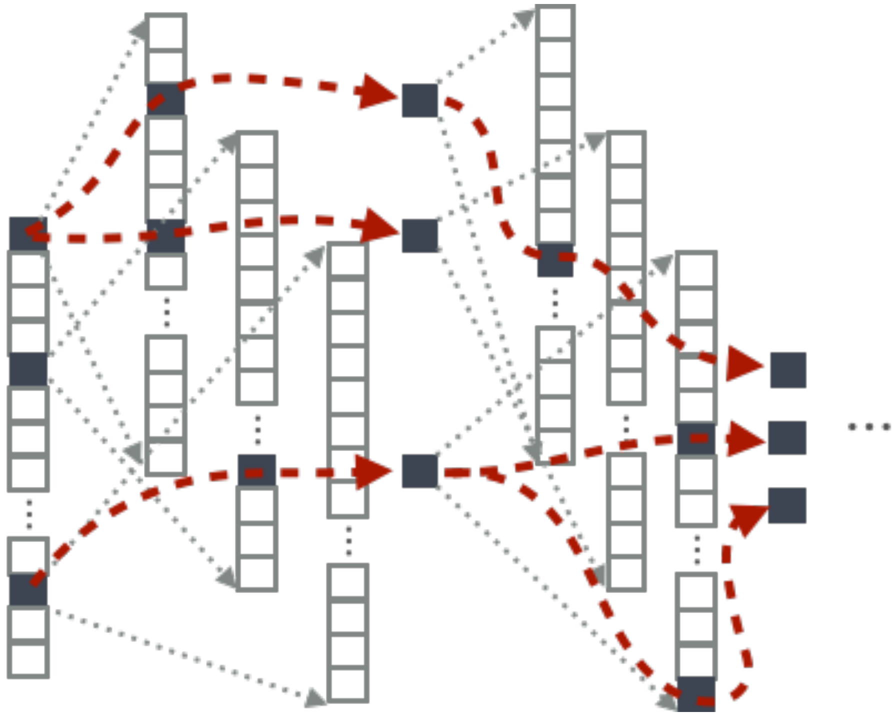
- Expand each hypothesis

$$\mathcal{H}_t^k = \left\{(\tilde{x}_1^k, \tilde{x}_2^k, \ldots, \tilde{x}_{t-1}^k, v_1), (\tilde{x}_1^k, \tilde{x}_2^k, \ldots, \tilde{x}_{t-1}^k, v_2), \ldots, (\tilde{x}_1^k, \tilde{x}_2^k, \ldots, \tilde{x}_{t-1}^k, v_{|V|})\right\}$$

- Pick top-K hypotheses from the union $\mathcal{H}_t = \cup_{k=1}^K \mathcal{B}_k$, where

$$\mathcal{B}_k = \arg\max_{\tilde{X} \in \mathcal{A}_k} \log p(\tilde{X}|Y), \ \mathcal{A}_k = \mathcal{A}_{k-1} - \mathcal{B}_{k-1}, \text{ and } \mathcal{A}_1 = \cup_{k'=1}^K \mathcal{H}_t^{k'}.$$

Decoding (3) – Beam Search



- Asymptotically exact, as $K \to \infty$
- But, not necessarily monotonic improvement w.r.t. K
- *K* should be selected to maximize the translation quality on a validation set.

Decoding

- En-Cz: 12m training sentence pairs

Strategy	# Chains	Valid Set		Test Set	
		NLL	BLEU	NLL	BLEU
Ancestral Sampling	50	22.98	15.64	26.25	16.76
Greedy Decoding	-	27.88	15.50	26.49	16.66
Beamsearch	5	20.18	17.03	22.81	18.56
Beamsearch	10	19.92	17.13	22.44	18.59

[Cho, arXiv 2016]

Decoding

- Greedy Search
  - Computationally efficient
  - Not great quality
- Beam Search
  - Computationally expensive
  - Not easy to parallelize
  - Much better quality

Beam search with a small beam is de facto standard in NMT

[Cho, arXiv 2016]