Natural Language Processing with Deep Learning

CS224N/Ling284

Christopher Manning and Richard Socher

Lecture 11: Further topics in Neural Machine Translation and Recurrent Models

Lecture Plan: Going forwards and backwards

1. A final look at gated recurrent units like GRUs/LSTMs
2. *Research highlight: Lip reading sentences in the wild*
3. Machine translation evaluation
4. Dealing with the large output vocabulary
5. Sub-word and character-based models

Reminders/comments:

Midterm being returned ☺

Assignment 3 is looming ☹

Learn up on GPUs, Azure, Docker

Final project discussions – come meet with us!

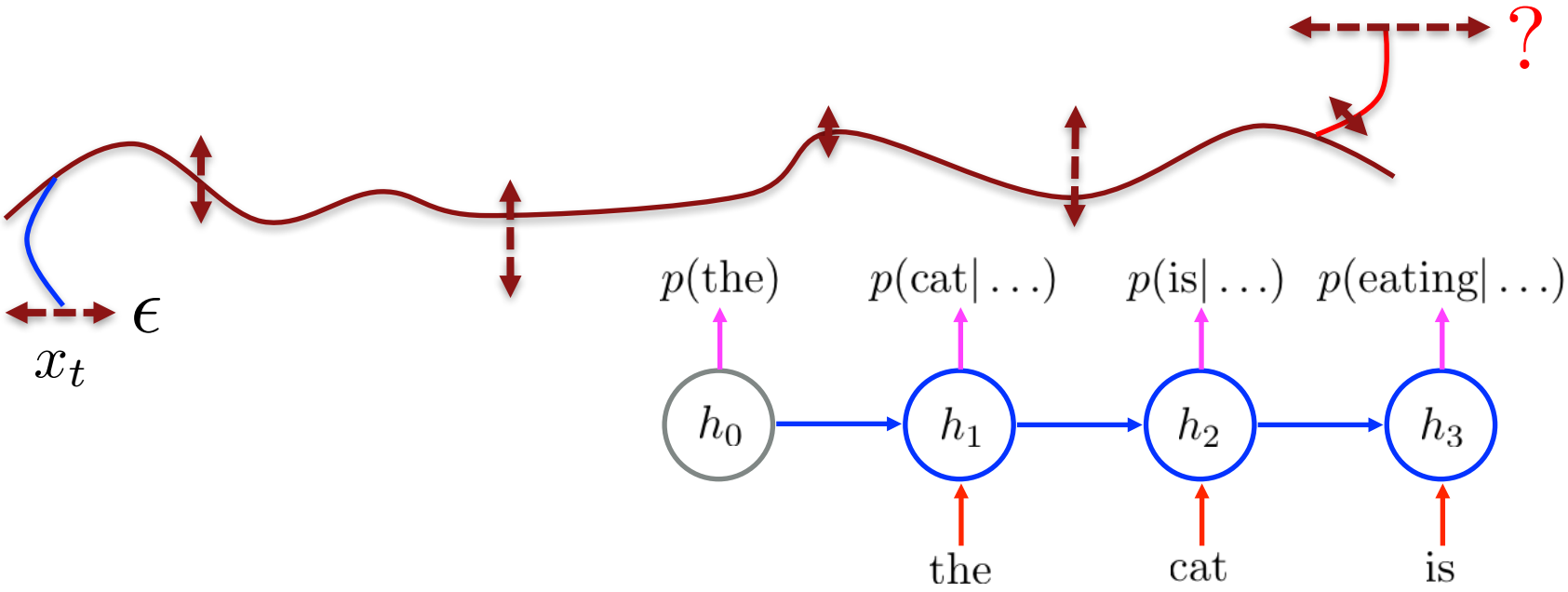
1. How Gated Units Fix Things – Backpropagation through Time

Intuitively, what happens with RNNs?

1. Measure the influence of the past on the future

$$\frac{\partial \log p(x_{t+n}|x_{<t+n})}{\partial h_t} = \frac{\partial \log p(x_{t+n}|x_{<t+n})}{\partial g} \frac{\partial g}{\partial h_{t+n}} \frac{\partial h_{t+n}}{\partial h_{t+n-1}} \cdots \frac{\partial h_{t+1}}{\partial h_t}$$

2. How does the perturbation at t affect $p(x_{t+n}|x_{<t+n})$?



Backpropagation through Time

Vanishing gradient is super-problematic

- When we only observe

$$\left\| \frac{\partial h_{t+N}}{\partial h_t} \right\| = \left\| \prod_{n=1}^{N} U^\top \operatorname{diag}\left(\frac{\partial \tanh(a_{t+n})}{\partial a_{t+n}} \right) \right\| \to 0 \ ,$$

- We cannot tell whether
  1. No dependency between *t* and *t+n* in data, or
  2. Wrong configuration of parameters (the vanishing gradient condition):

$$e_{\max}(U) < \frac{1}{\max \tanh'(x)}$$

Gated Recurrent Unit

- Is the problem with the naïve transition function?

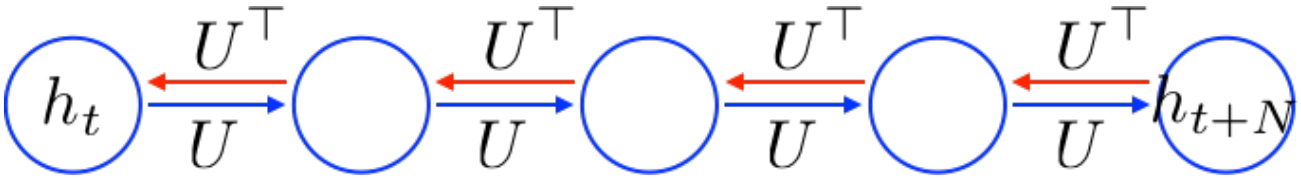
$$f(h_{t-1}, x_t) = \tanh(W\,[x_t] + U h_{t-1} + b)$$

- With it, the temporal derivative is

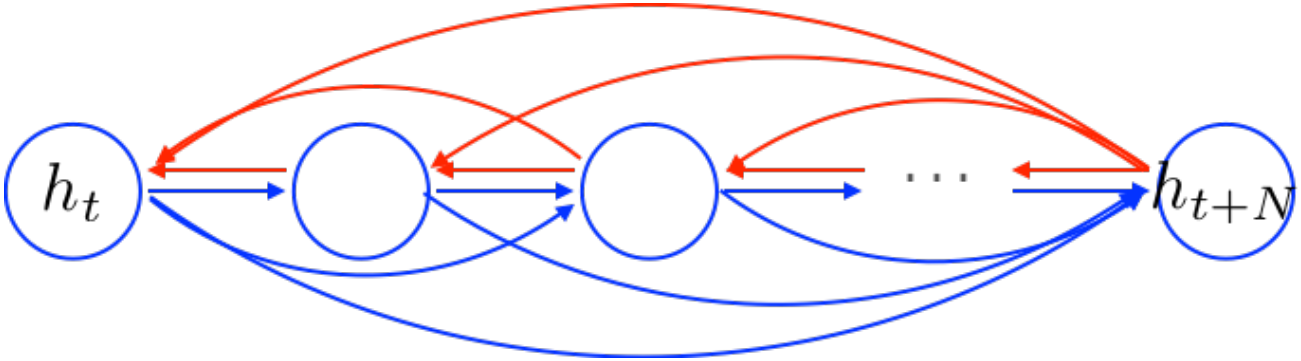
$$\frac{\partial h_{t+1}}{\partial h_t} = U^\top \frac{\partial \tanh(a)}{\partial a}$$

Gated Recurrent Unit

- It implies that the error must backpropagate through all the intermediate nodes:

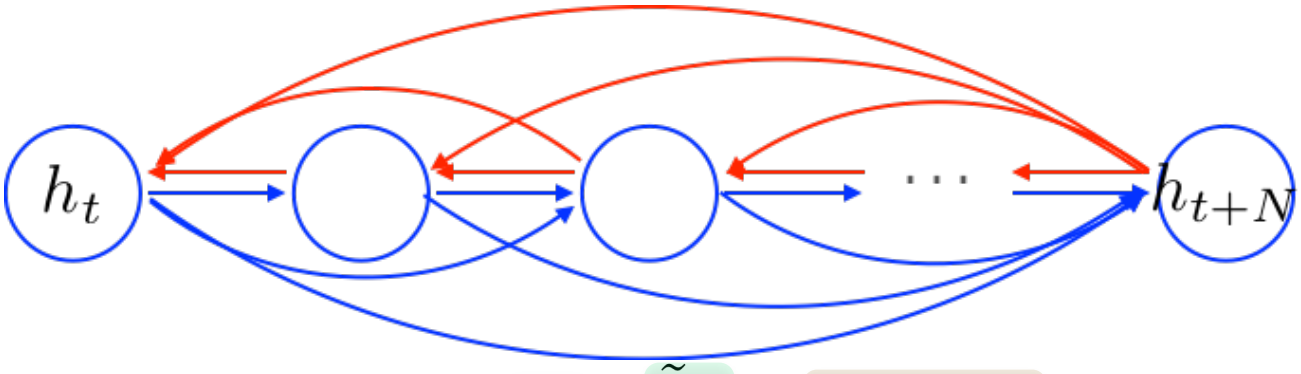


- Perhaps we can create shortcut connections.



Gated Recurrent Unit

- Perhaps we can create *adaptive* shortcut connections.



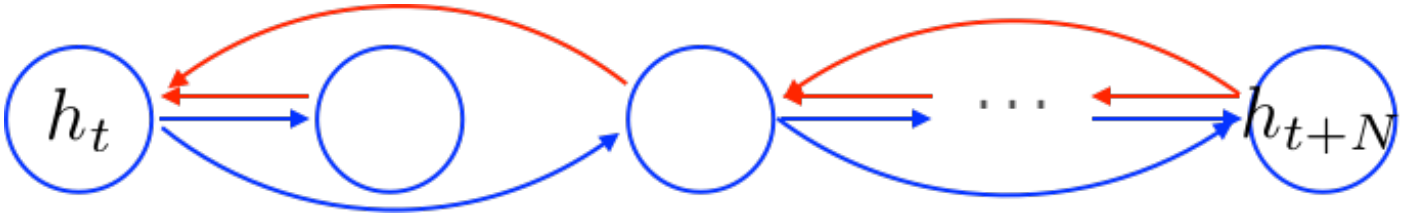
$$f(h_{t-1}, x_t) = u_t \odot \tilde{h}_t + (1 - u_t) \odot h_{t-1}$$

- Candidate Update $\tilde{h}_t = \tanh(W\,[x_t] + U h_{t-1} + b)$
- Update gate $u_t = \sigma(W_u\,[x_t] + U_u h_{t-1} + b_u)$

$\odot$: element-wise multiplication

Gated Recurrent Unit

- Let the net prune unnecessary connections *adaptively*.

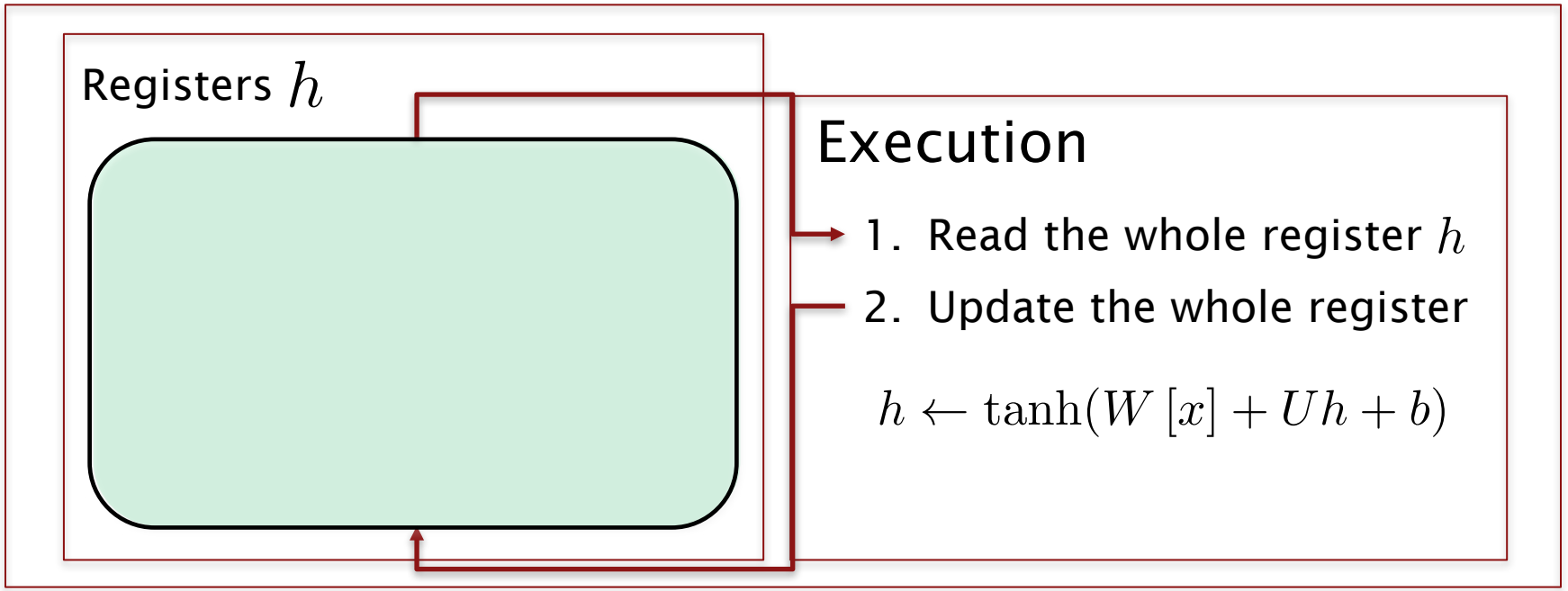


$$f(h_{t-1}, x_t) = u_t \odot \tilde{h}_t + (1 - u_t) \odot h_{t-1}$$

- Candidate Update $\tilde{h}_t = \tanh(W\,[x_t] + U(r_t \odot h_{t-1}) + b)$
- Reset gate $r_t = \sigma(W_r\,[x_t] + U_r h_{t-1} + b_r)$
- Update gate $u_t = \sigma(W_u\,[x_t] + U_u h_{t-1} + b_u)$

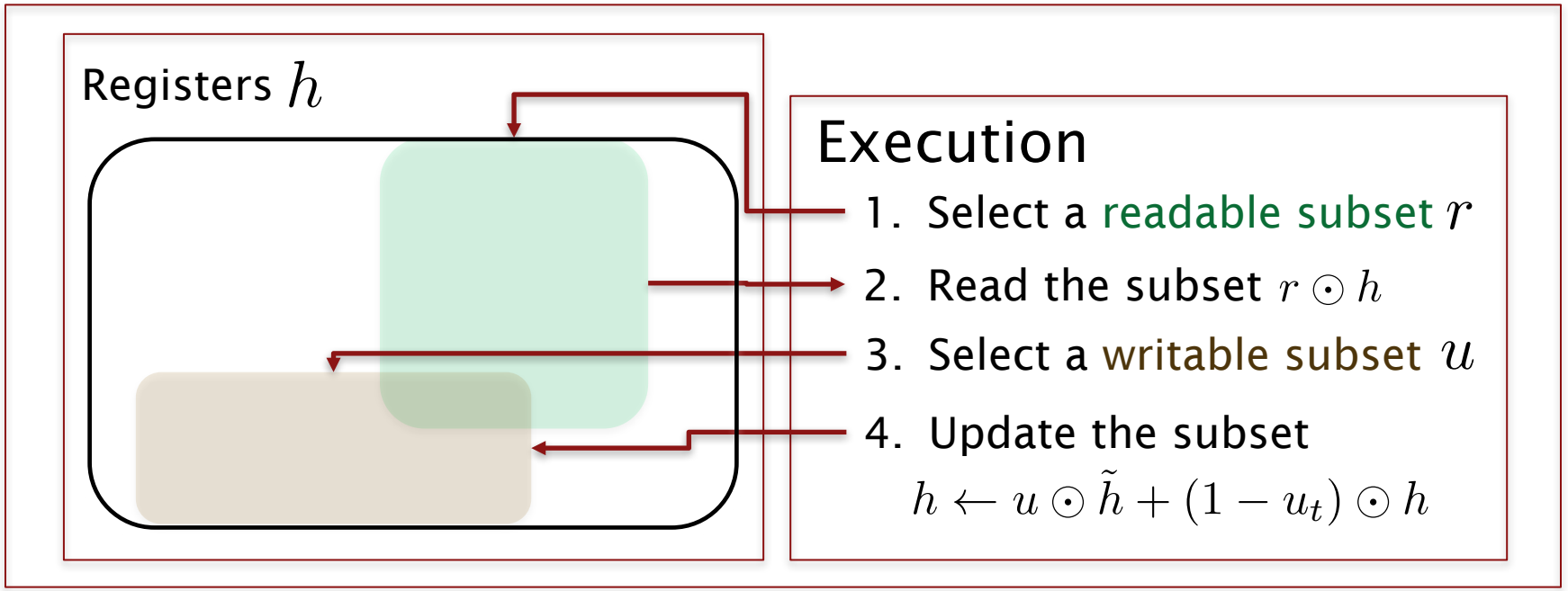
Gated Recurrent Unit

tanh-RNN ….



Gated Recurrent Unit

GRU …



Gated recurrent units are much more realistic!

Note that there is some overlap in ideas with attention

Gated Recurrent Unit

Two most widely used gated recurrent units

Gated Recurrent Unit

[Cho et al., EMNLP2014; Chung, Gulcehre, Cho, Bengio, DLUFL2014]

$$h_t = u_t \odot \tilde{h}_t + (1 - u_t) \odot h_{t-1}$$

$$\tilde{h} = \tanh(W\,[x_t] + U(r_t \odot h_{t-1}) + b)$$

$$u_t = \sigma(W_u\,[x_t] + U_u h_{t-1} + b_u)$$

$$r_t = \sigma(W_r\,[x_t] + U_r h_{t-1} + b_r)$$

Long Short-Term Memory

[Hochreiter & Schmidhuber, NC1999; Gers, Thesis2001]

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

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

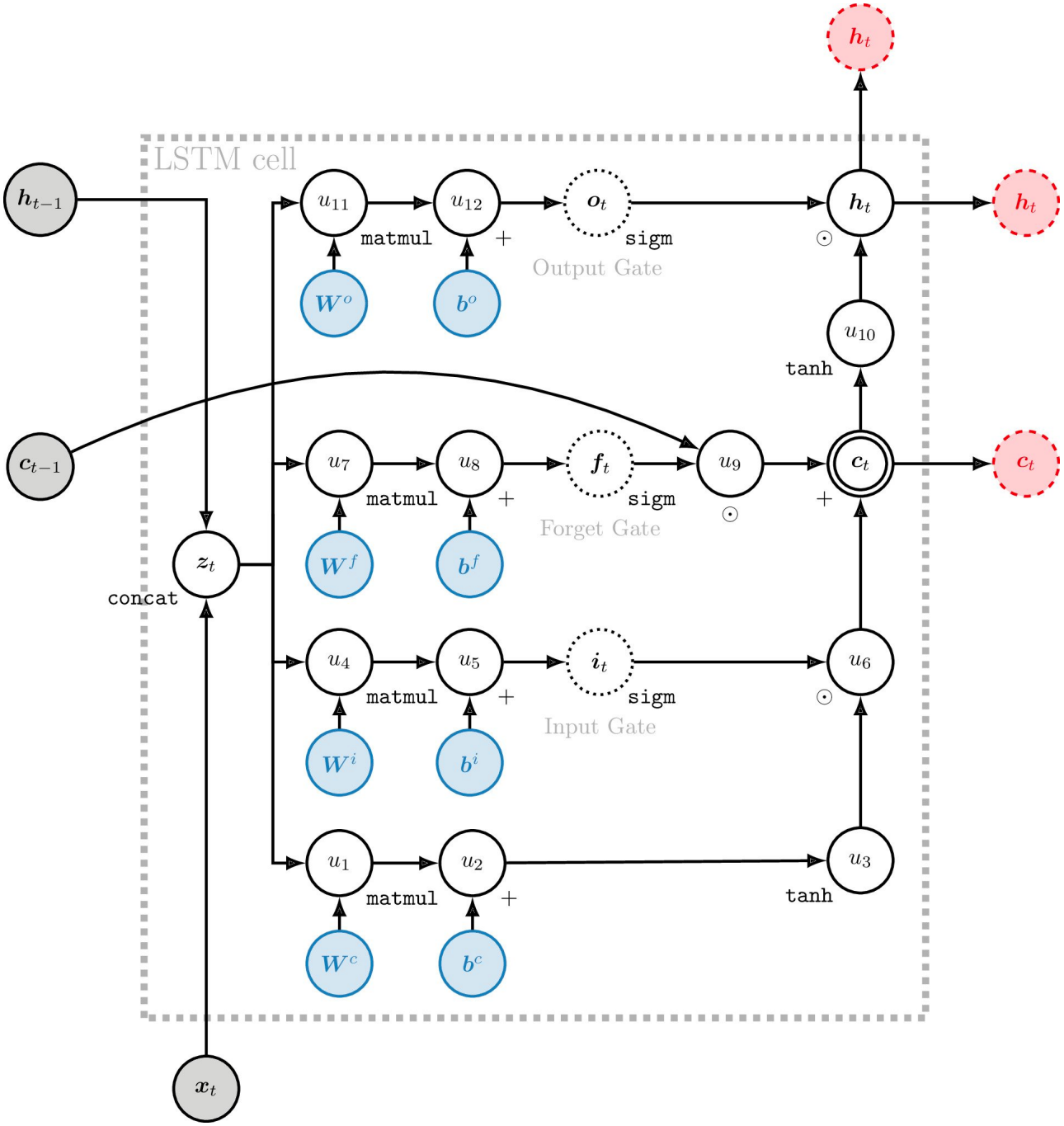
$$\tilde{c}_t = \tanh(W_c\,[x_t] + U_c h_{t-1} + b_c)$$

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

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

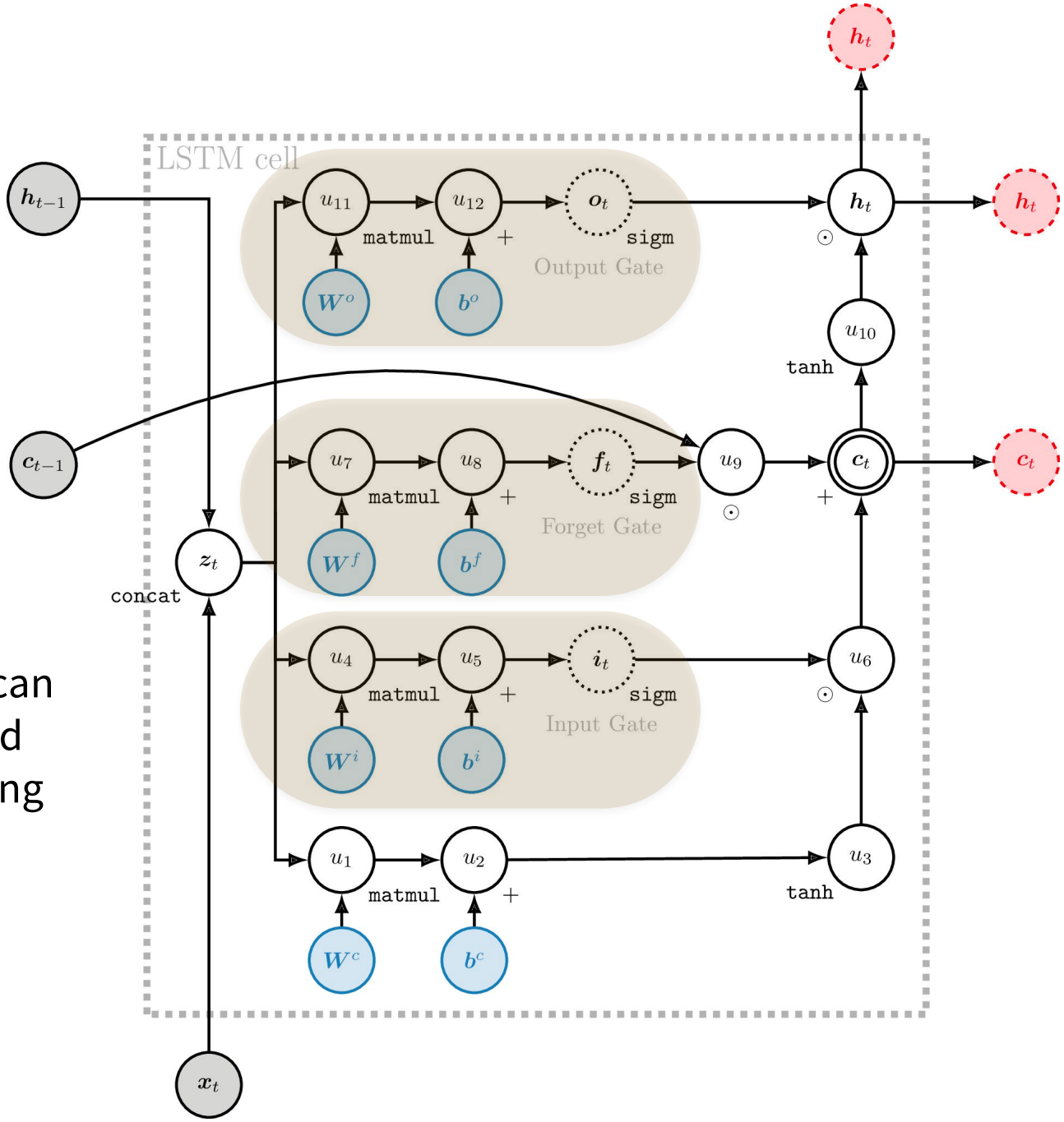
$$f_t = \sigma(W_f\,[x_t] + U_f h_{t-1} + b_f)$$

The LSTM

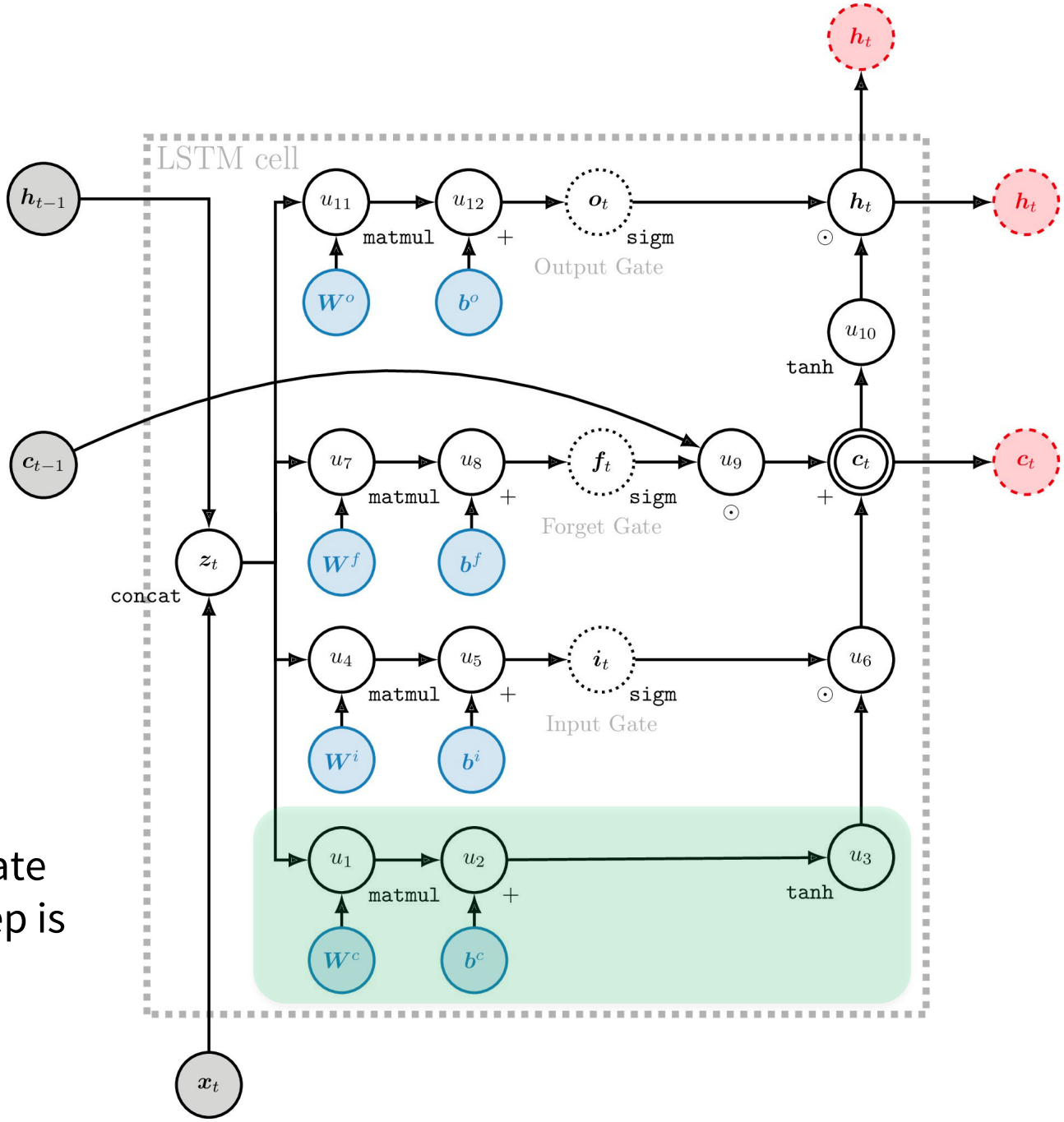

LSTM cell
h_{t-1}
c_{t-1}
x_t
z_t
concat
u_{11}
u_{12}
o_t
matmul
sigm
Output Gate
W^o
b^o
u_7
u_8
f_t
Forget Gate
W^f
b^f
u_4
u_5
i_t
Input Gate
W^i
b^i
u_1
u_2
u_3
tanh
W^c
b^c
u_9
u_{10}
h_t
c_t

The LSTM

The LSTM gates all operations so stuff can be forgotten/ignored rather than it all being crammed on top of everything else

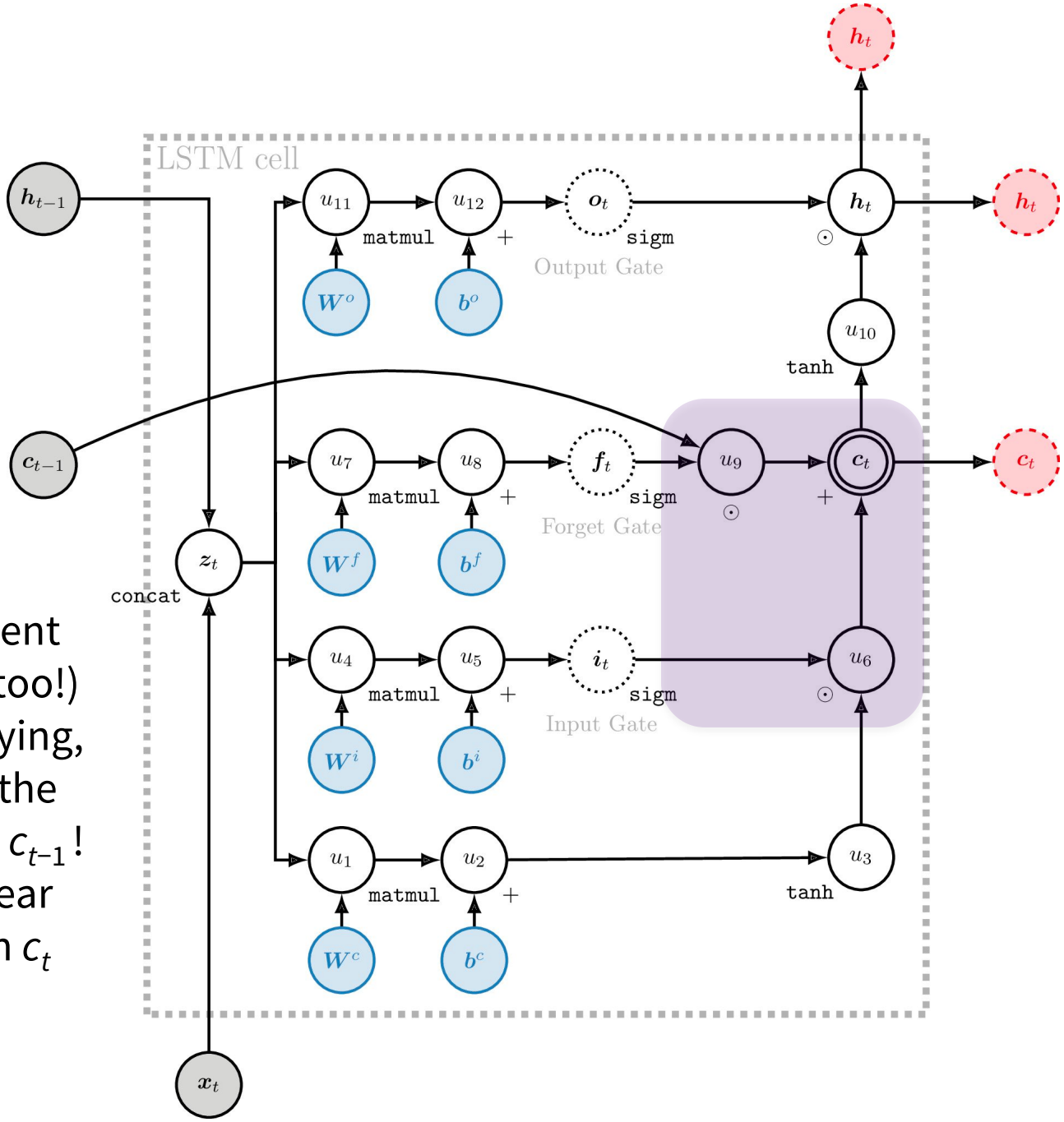


The LSTM



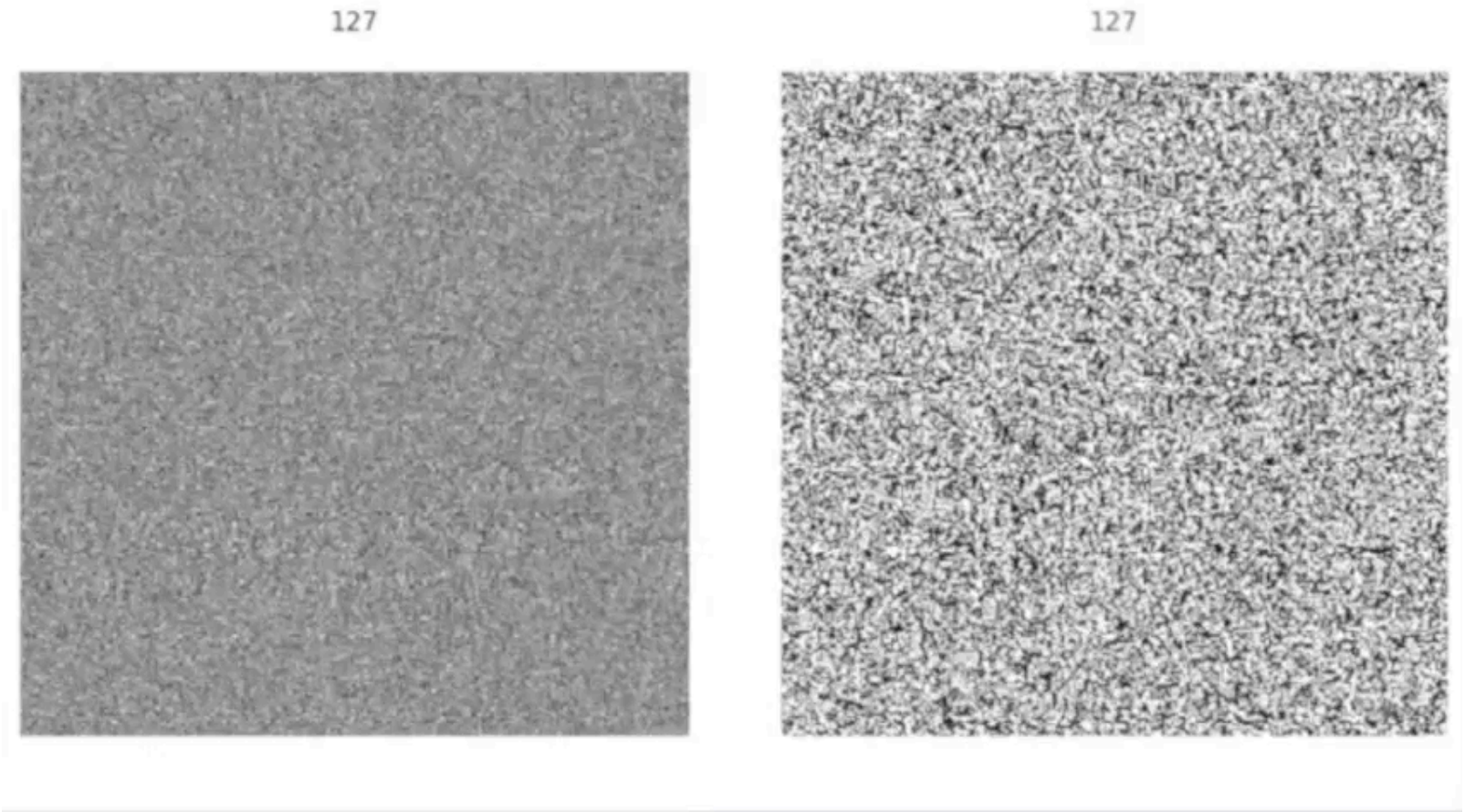
The non-linear update for the next time step is just like an RNN

The LSTM



This part is the the secret! (Of other recent things like ResNets too!) Rather than multiplying, we get c_t by adding the non-linear stuff and c_{t-1}! There is a direct, linear connection between c_t and c_{t-1}.

As a result, LSTM have a long memory



Training a (gated) RNN

1. Use an LSTM or GRU: *it makes your life so much simpler!*
2. Initialize recurrent matrices to be orthogonal
3. Initialize other matrices with a sensible (**small!**) scale
4. Initialize forget gate bias to 1: *default to remembering*
5. Use adaptive learning rate algorithms: *Adam, AdaDelta, …*
6. Clip the norm of the gradient: *1–5 seems to be a reasonable threshold when used together with Adam or AdaDelta.*
7. Either only dropout vertically or learn how to do it right
8. *Be patient!*

[Saxe et al., ICLR2014;
Ba, Kingma, ICLR2015;
Zeiler, arXiv2012;
Pascanu et al., ICML2013]

Ensembles

- Train 8–10 nets and average their predictions
- It’s easy to do and usually gives good gains!

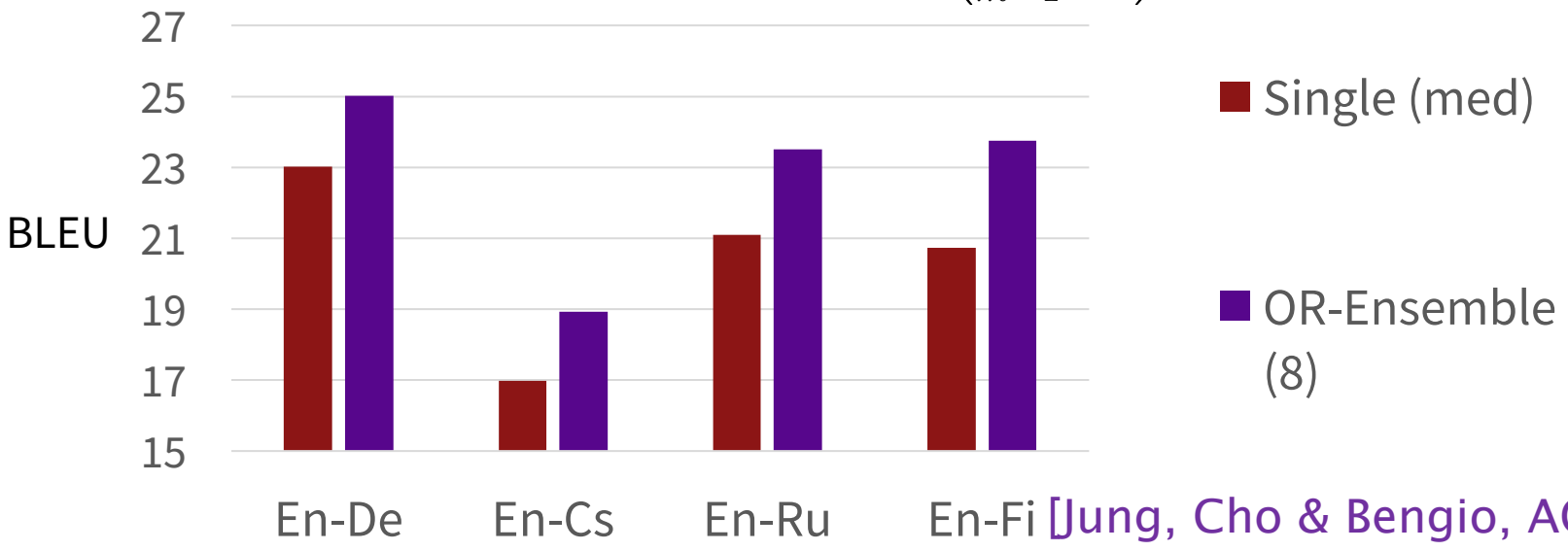
Ensemble of Conditional Recurrent LM

- Step-wise Ensemble: $p(x_t^{\text{ens}}|x_{<t}^{\text{ens}}, Y) = \oplus_{m=1}^{M} p(x_t^m|x_{<t}^m, Y)$
- Ensemble operator $\bigoplus$ implementations
  1. Majority voting scheme (OR):

$$\oplus_{m=1}^{M} p^{\text{ens}} = \frac{1}{M}\sum_{m=1}^{M} p^m$$

  2. Consensus building scheme (AND):

$$\oplus_{m=1}^{M} p^{\text{ens}} = \left(\prod_{m=1}^{M} p^m\right)^{1/M}$$



[Jung, Cho & Bengio, ACL2016]

Lip Reading Sentences in the Wild

Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman
Presented by: Michael Fang

Task: Lip Reading

+

→ The cat sat...


MANCHESTER CITY TRAVEL TO PARIS IN THE CHAMPIONS LEAGUE THIS EVENING

Outline

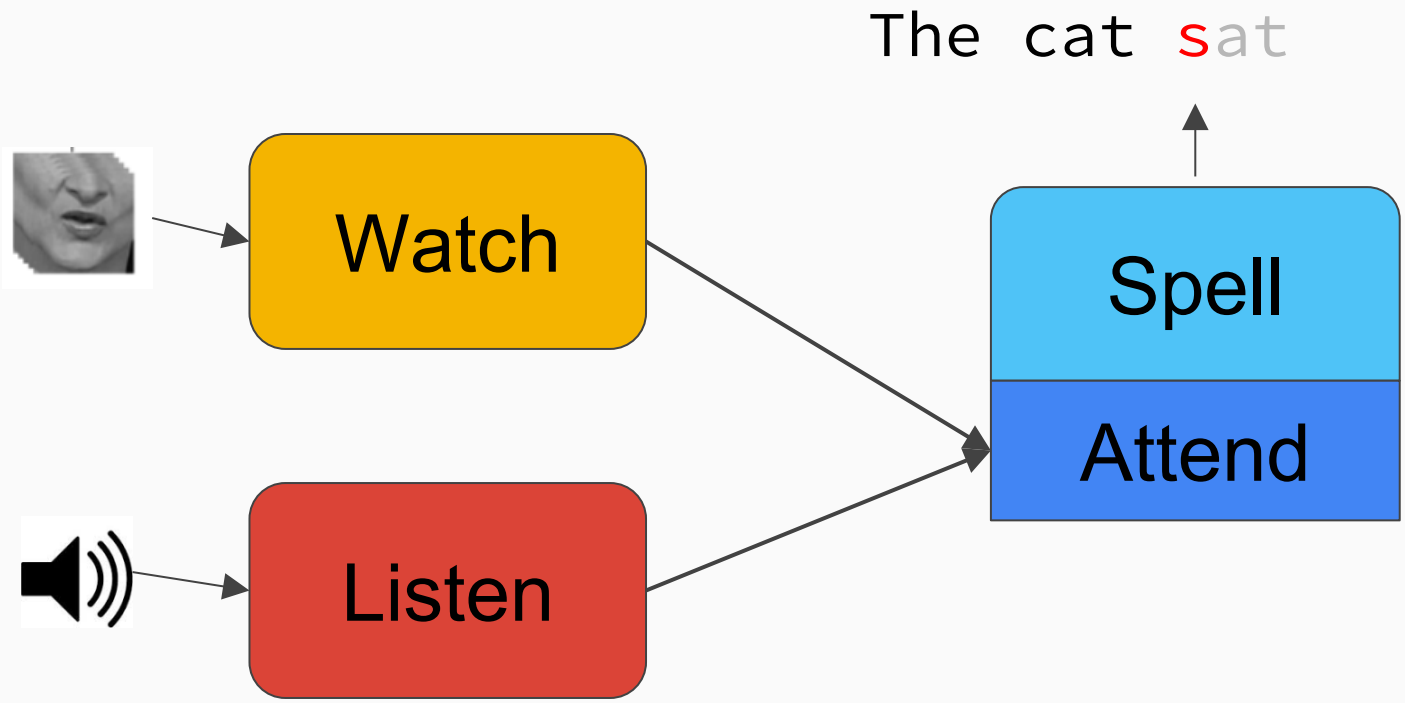
Model Architecture: Watch, Listen, Attend and Spell

Training Strategies

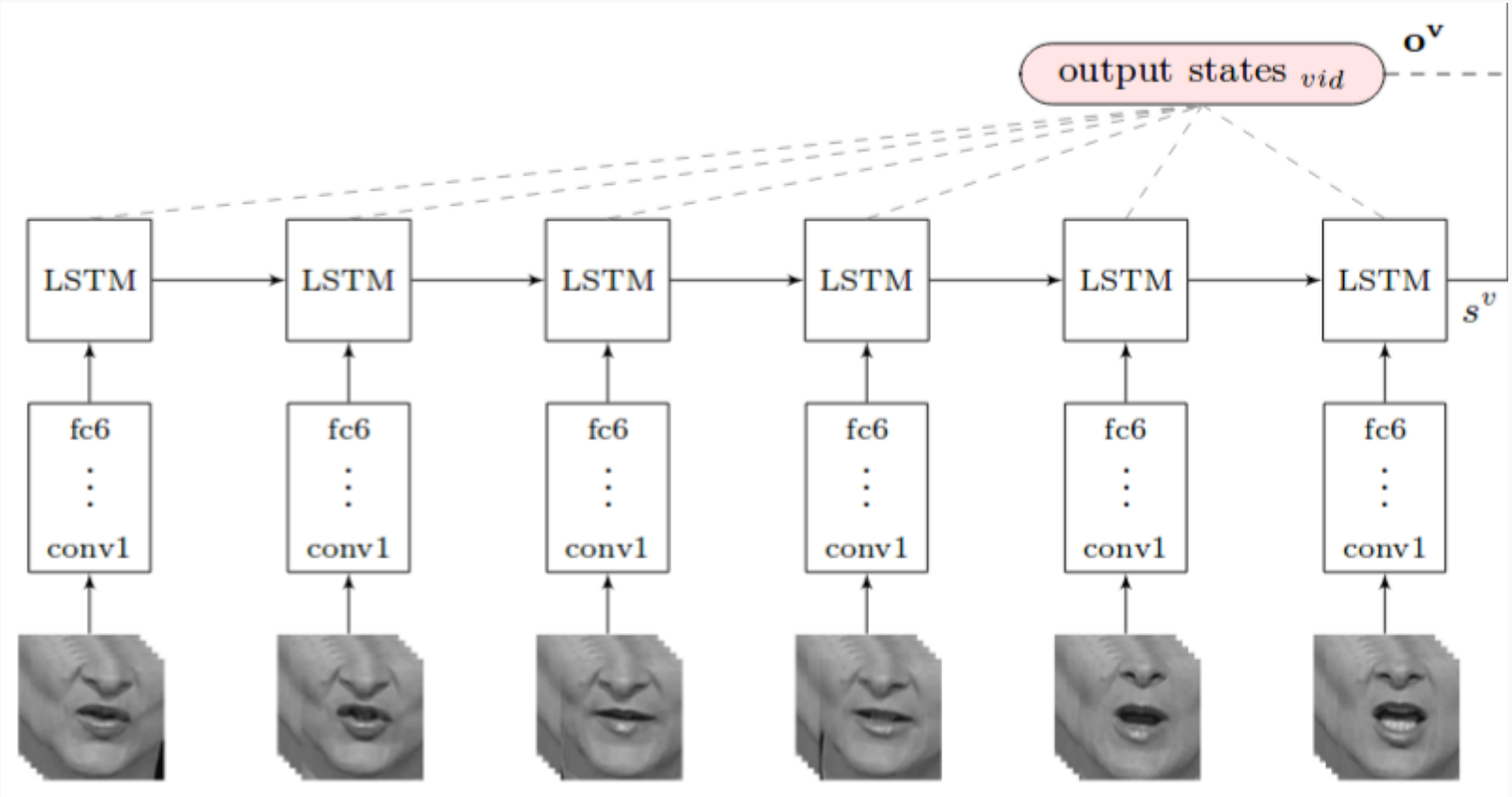
Dataset

(Professional-Surpassing!) Results

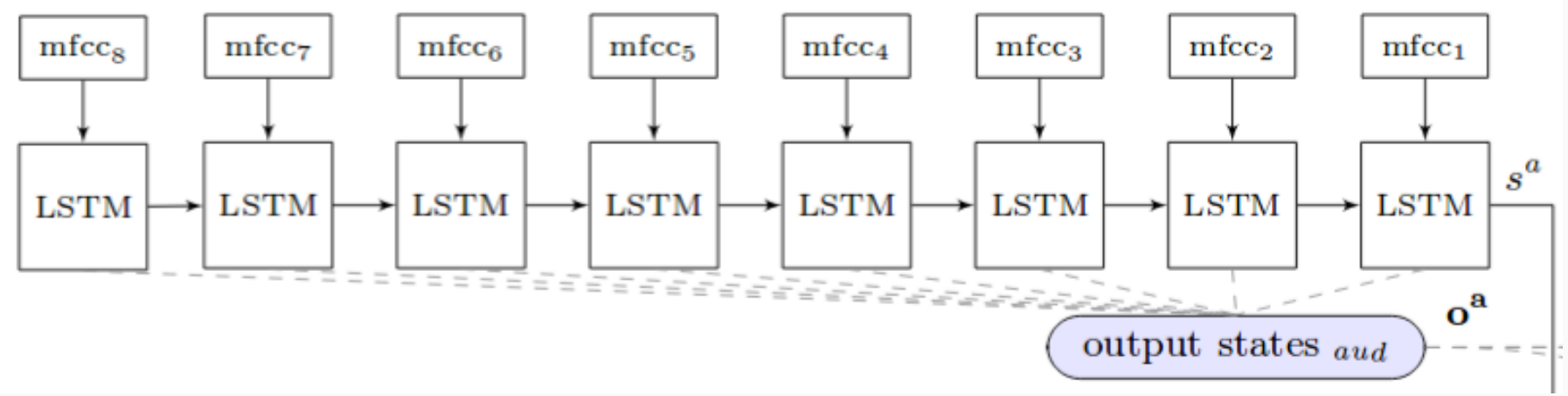
Architecture



Watch

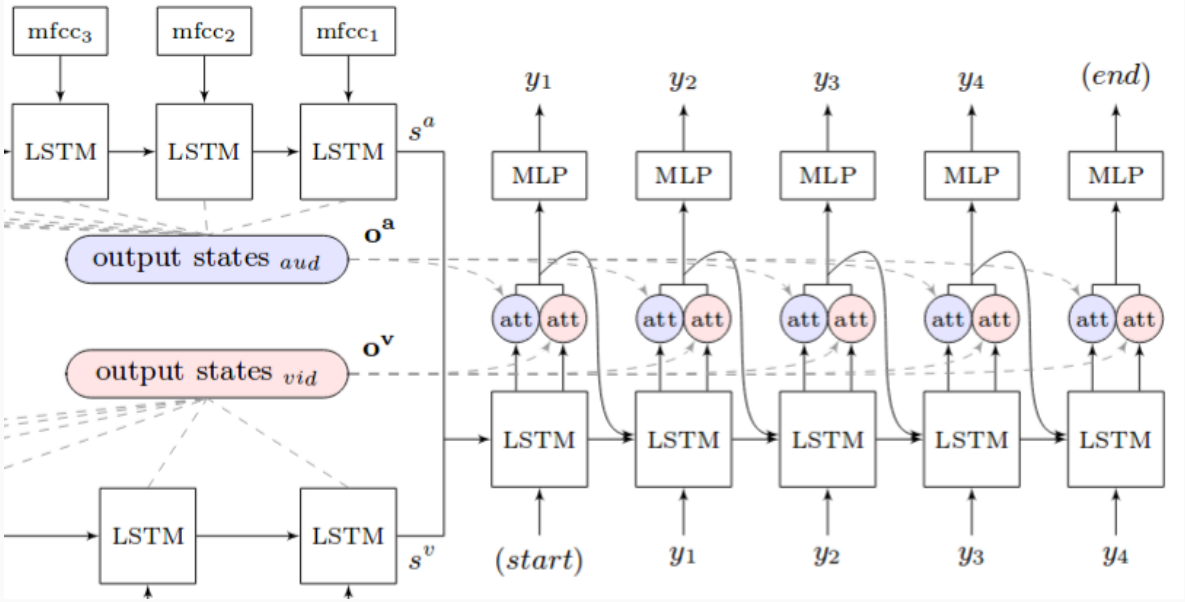

output states $_{vid}$
$\mathbf{o^v}$
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
s^v
fc6
conv1
fc6
conv1
fc6
conv1
fc6
conv1
fc6
conv1
fc6
conv1

Listen


mfcc8
mfcc7
mfcc6
mfcc5
mfcc4
mfcc3
mfcc2
mfcc1
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
s^a
$\mathbf{o^a}$
output states aud

Attend and Spell

$$h_k^d, o_k^d = \texttt{LSTM}(h_{k-1}^d, y_{k-1}, c_{k-1}^v, c_{k-1}^a)$$
$$c_k^v = \mathbf{o}^v \cdot \texttt{Attention}^{\texttt{v}}(h_k^d, \mathbf{o}^v)$$
$$c_k^a = \mathbf{o}^a \cdot \texttt{Attention}^{\texttt{a}}(h_k^d, \mathbf{o}^a)$$
$$P(y_i|\mathbf{x}^v, \mathbf{x}^a, y_{<i}) = \texttt{softmax}(\texttt{MLP}(o_k^d, c_k^v, c_k^a))$$



Curriculum Learning

Slowly increase the length of training sequences

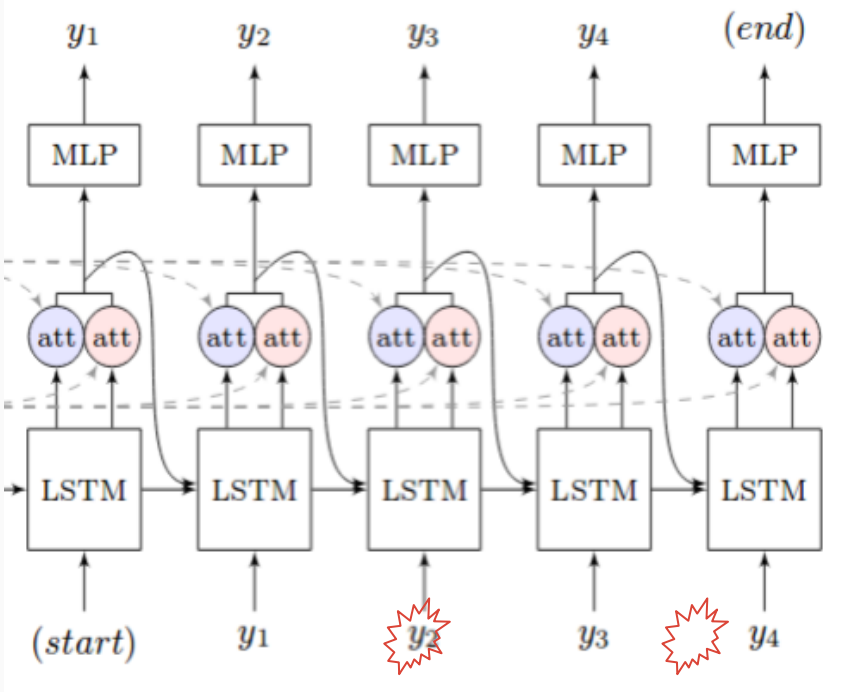
Converges training faster, decreases overfitting

```
The
The cat
The cat sat
The cat sat on
The cat sat on the
The cat sat on the mat
```

Scheduled Sampling

Randomly sample from previous prediction instead of ground truth during training

Makes training scenario more similar to testing



Dataset

Channel	Series name	# hours	# sent.
BBC 1 HD	News†	1,584	50,493
BBC 1 HD	Breakfast	1,997	29,862
BBC 1 HD	Newsnight	590	17,004
BBC 2 HD	World News	194	3,504
BBC 2 HD	Question Time	323	11,695
BBC 4 HD	World Today	272	5,558
All		4,960	118,116

Results

Method	SNR	CER	WER	BLEU†
Lips only				
Professional‡	-	58.7%	73.8%	23.8
WAS	-	59.9%	76.5%	35.6
WAS+CL	-	47.1%	61.1%	46.9
WAS+CL+SS	-	42.4%	58.1%	50.0
WAS+CL+SS+BS	-	39.5%	50.2%	54.9
Audio only				
Google Speech API	clean	17.6%	22.6%	78.4
Kaldi SGMM+MMI⋆	clean	9.7%	16.8%	83.6
LAS+CL+SS+BS	clean	10.4%	17.7%	84.0
LAS+CL+SS+BS	10dB	26.2%	37.6%	66.4
LAS+CL+SS+BS	0dB	50.3%	62.9%	44.6
Audio and lips				
WLAS+CL+SS+BS	clean	7.9%	13.9%	87.4
WLAS+CL+SS+BS	10dB	17.6%	27.6%	75.3
WLAS+CL+SS+BS	0dB	29.8%	42.0%	63.1

3. MT Evaluation

- Manual (the best!?):
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - **Adequacy and Fluency** (5 or 7 point scales)
  - Error categorization
  - **Comparative ranking of translations**

- Testing in an application that uses MT as one sub-component
  - E.g., question answering from foreign language documents
    - May not test many aspects of the translation (e.g., cross-lingual IR)

- Automatic metric:
  - WER (word error rate) – why problematic?
  - **BLEU (Bilingual Evaluation Understudy)**

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is between 0 & 1)
  - What percent of machine n-grams can be found in the reference translation?
    - An n-gram is a sequence of n words
  - For each n-gram size, not allowed to match identical portion of reference translation more than once (two MT words *airport* are only correct if two reference words *airport;* can't cheat by typing out "the the the the the")
- Brevity Penalty
  - Can't just type out single word "the" (precision 1.0!)
- It was thought hard to "game" the metric (i.e., to find a way to change MT output so that BLEU goes up, but quality doesn't)

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- BLEU is a weighted geometric mean of n-gram precision (is translation in reference?), with a brevity penalty factor added.
- BLEU4 counts n-grams ≤ length k = 4

p_n = # matched n-gram / # MT n-gram
w_n = weight, e.g., w_1=1, w_2=½, w_3=¼, w_4=⅛
BP = exp(min(0, 1 − (len_{ref}/len_{MT})))

$$BLEU = BP * \prod_{n=1}^{N} p_n^{w_n}$$

log BLEU = exp (1.0 * log p_1 + 0.5 * log p_2 +
0.25 * log p_3 + 0.125 * log p_4 –
max(words-in-ref / words-in-MT – 1, 0)

Note: only works at corpus level (zeroes kill it); there's a smoothed variant for sentence-level

BLEU in Action

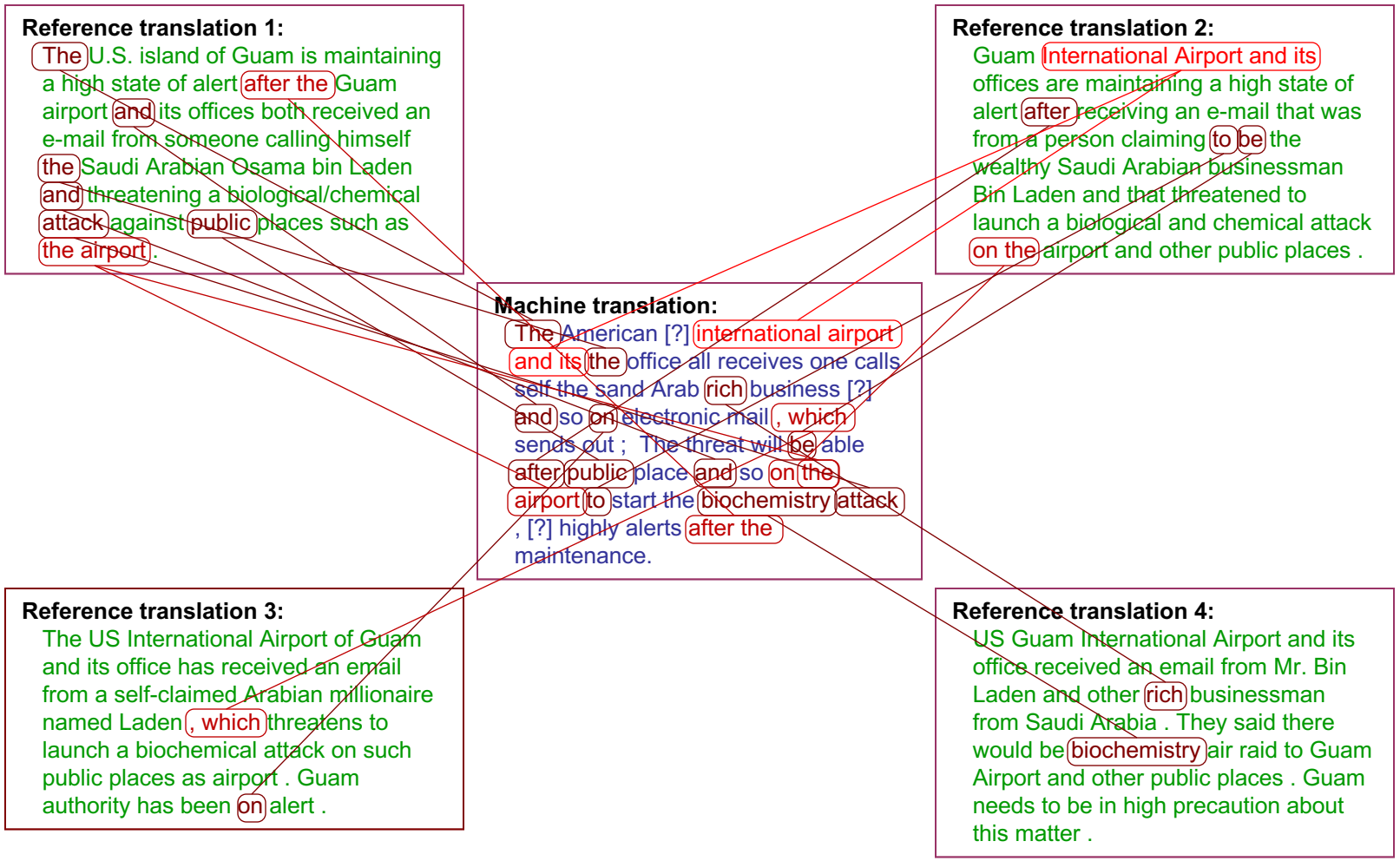
枪手被警方击毙。 (Foreign Original)

the gunman was shot to death by the police . (Reference Translation)

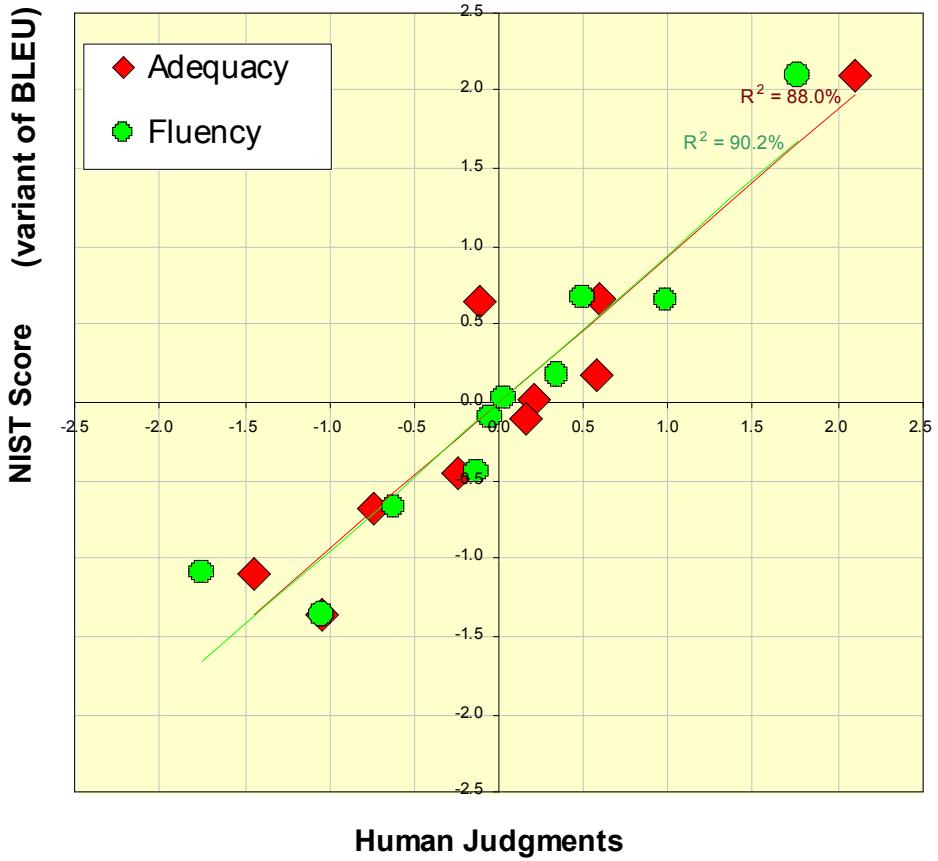
the gunman was police kill .	#1
wounded police jaya of	#2
the gunman was shot dead by the police .	#3
the gunman arrested by police kill .	#4
the gunmen were killed .	#5
the gunman was shot to death by the police .	#6
gunmen were killed by police ?SUB>0 ?SUB>0	#7
al by the police .	#8
the ringer is killed by the police .	#9
police killed the gunman .	#10

green = 4-gram match (good!)
red = word not matched (bad!)

Multiple Reference Translations



Initial results showed that BLEU predicts human judgments well

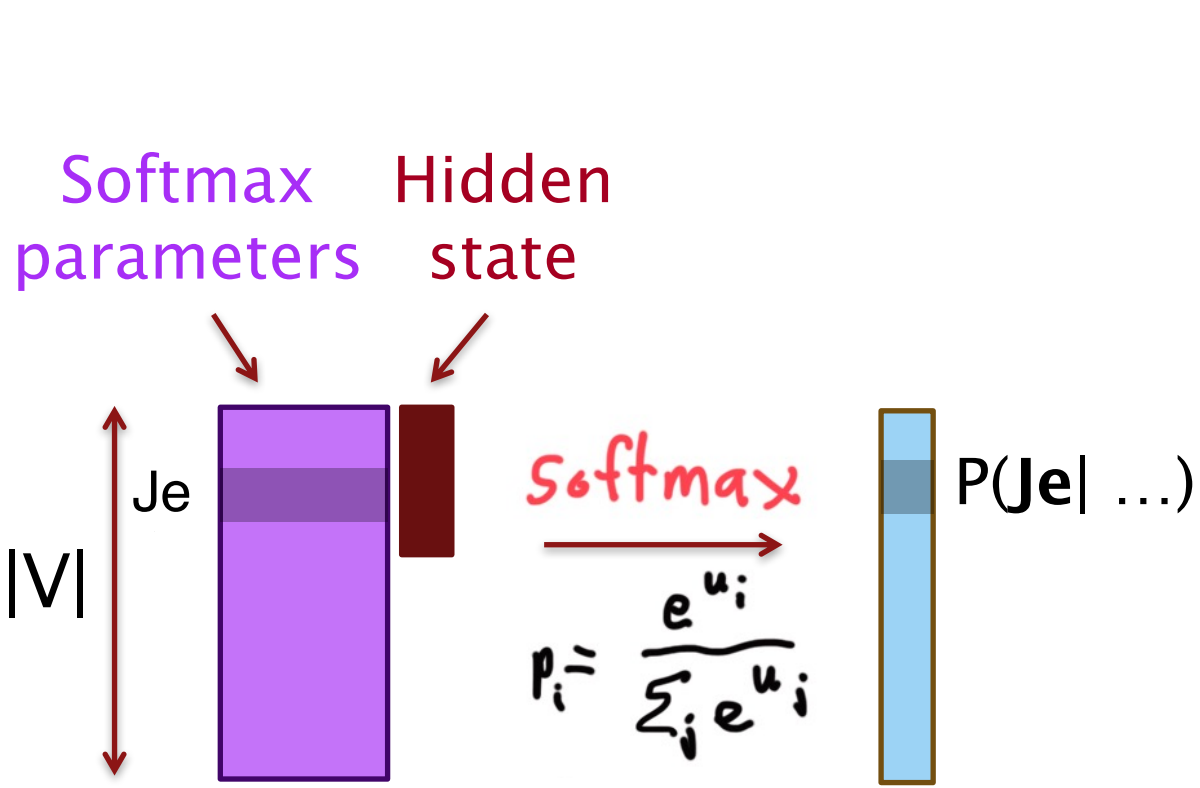


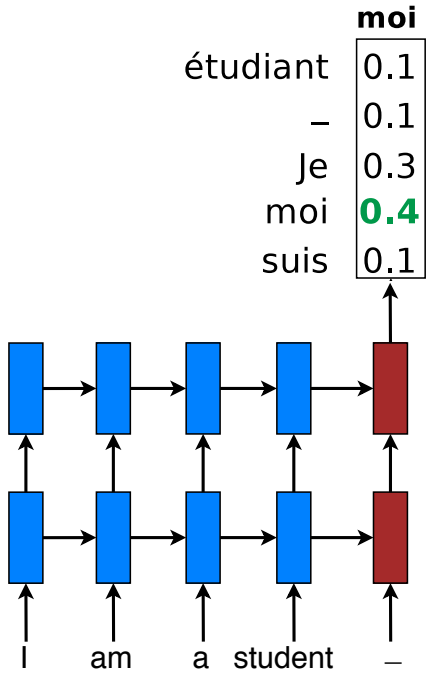
Slide from G. Doddington (NIST)

Automatic evaluation of MT

- People started optimizing their systems to maximize BLEU score
  - BLEU scores improved rapidly
  - The correlation between BLEU and human judgments of quality went way, way down
  - MT BLEU scores now approach those of human translations but their true quality remains far below human translations
- Coming up with automatic MT evaluations has become its own research field
  - There are many proposals: TER, METEOR, MaxSim, SEPIA, our own RTE-MT
  - TERpA is a representative good one that handles some word choice variation.
- MT research **requires** *some* automatic metric to allow a rapid development and evaluation cycle.

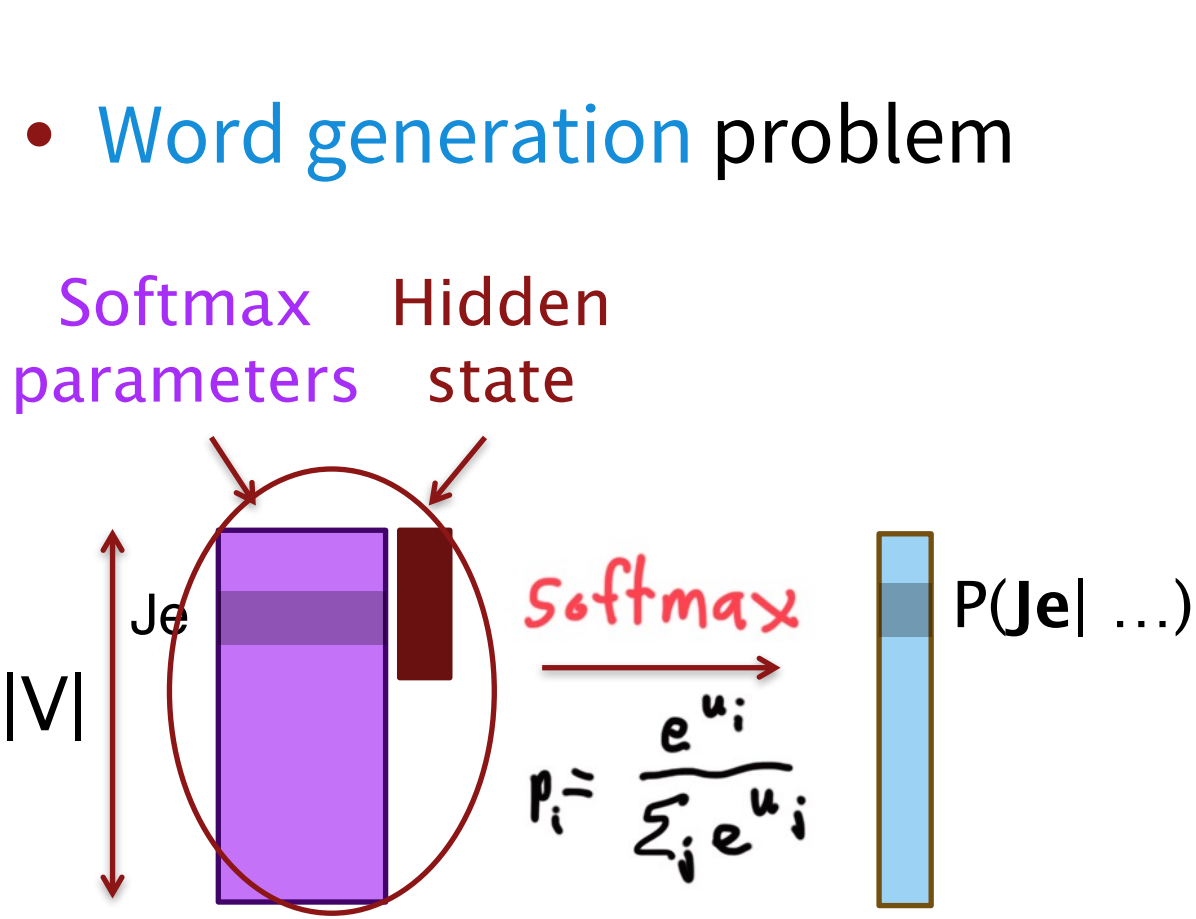
4. The word generation problem: dealing with a large output vocab

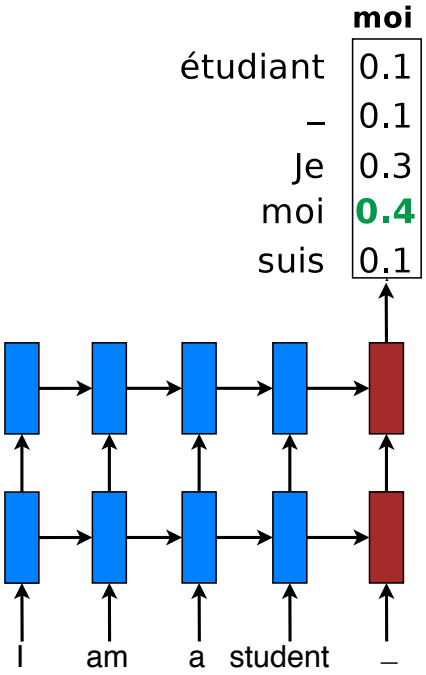




The word generation problem

- Word generation problem

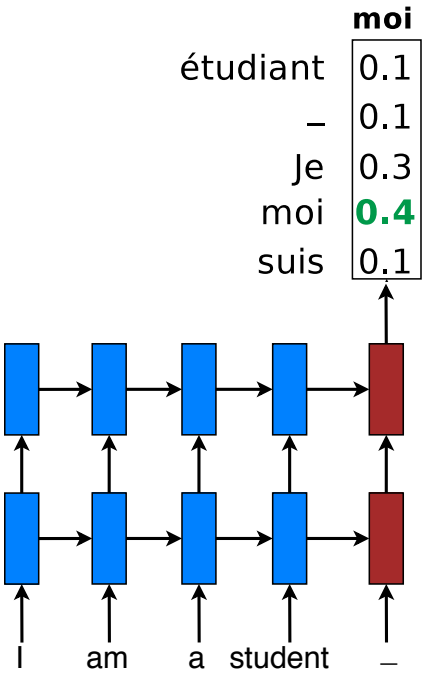




Softmax computation is expensive.

The word generation problem

- Word generation problem
  - If vocabs are modest, e.g., 50K



The ecotax portico in Pont-de-Buis
Le portique écotaxe de Pont-de-Buis

The <unk> portico in <unk>
Le <unk> <unk> de <unk>

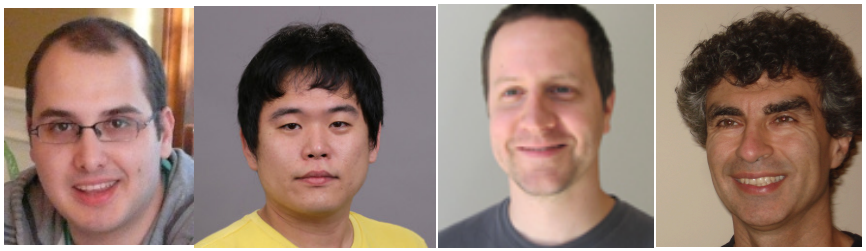
First thought: scale the softmax

- Lots of ideas from the neural LM literature!
- *Hierarchical models*: tree-structured vocabulary
  - [Morin & Bengio, AISTATS'05], [Mnih & Hinton, NIPS'09].
  - Complex, sensitive to tree structures.
- *Noise-contrastive estimation*: binary classification
  - [Mnih & Teh, ICML'12], [Vaswani et al., EMNLP'13].
  - Different noise samples per training example.*

Not GPU-friendly

*We'll mention a simple fix for this!

Large-vocab NMT



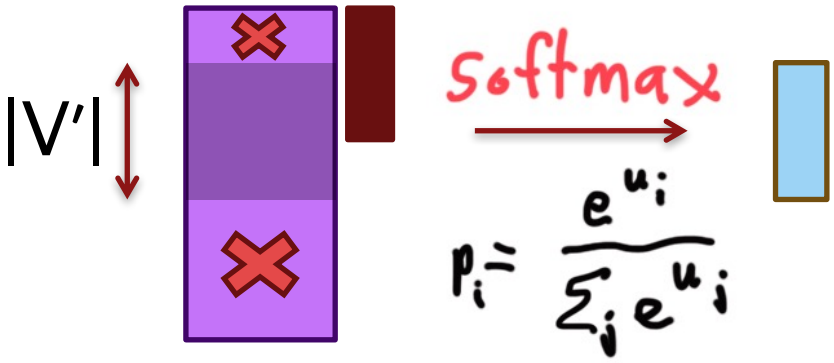
- GPU-friendly.
- *Training*: a subset of the vocabulary at a time.
- *Testing*: smart on the set of possible translations.

Fast at both train & test time.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, Yoshua Bengio. ***On Using Very Large Target Vocabulary for Neural Machine Translation****. ACL'15.*

Training

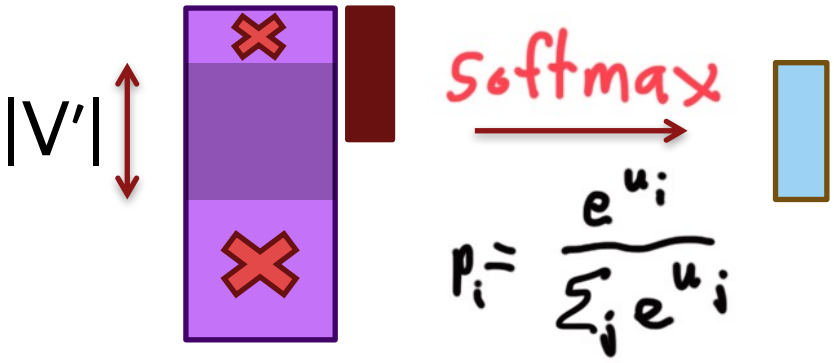
- Each time train on a smaller vocab $V' \ll V$



How do we select V'?

Training

- Each time train on a smaller vocab $V' \ll V$



- Partition training data in subsets:
  - Each subset has τ distinct target words, $|V'| = \tau$.

Training – *Segment data*

- Sequentially select examples: $|V'| = 5$.

she loves cats
he likes dogs
cats have tails
dogs have tails
dogs chase cats
she loves dogs
cats hate dogs

$V' = \{\text{she, loves, cats, he, likes}\}$

Training – *Segment data*

- Sequentially select examples: $|V'| = 5$.

she loves cats
he likes dogs
cats have tails
dogs have tails
dogs chase cats
she loves dogs
cats hate dogs

$V' = \{cats, have, tails, dogs, chase\}$

Training – *Segment data*

- Sequentially select examples: $|V'| = 5$.

she loves cats
he likes dogs
cats have tails
dogs have tails
dogs chase cats
she loves dogs
cats hate dogs

V' = {she, loves, dogs, cats, hate}

- *Practice*: $|V| = 500K$, $|V'| = 30K$ or $50K$.

Testing – *Select candidate words*

- **K** most frequent words: unigram prob.

de,
,
la
.
et
des
les
…

Testing – *Select candidate words*

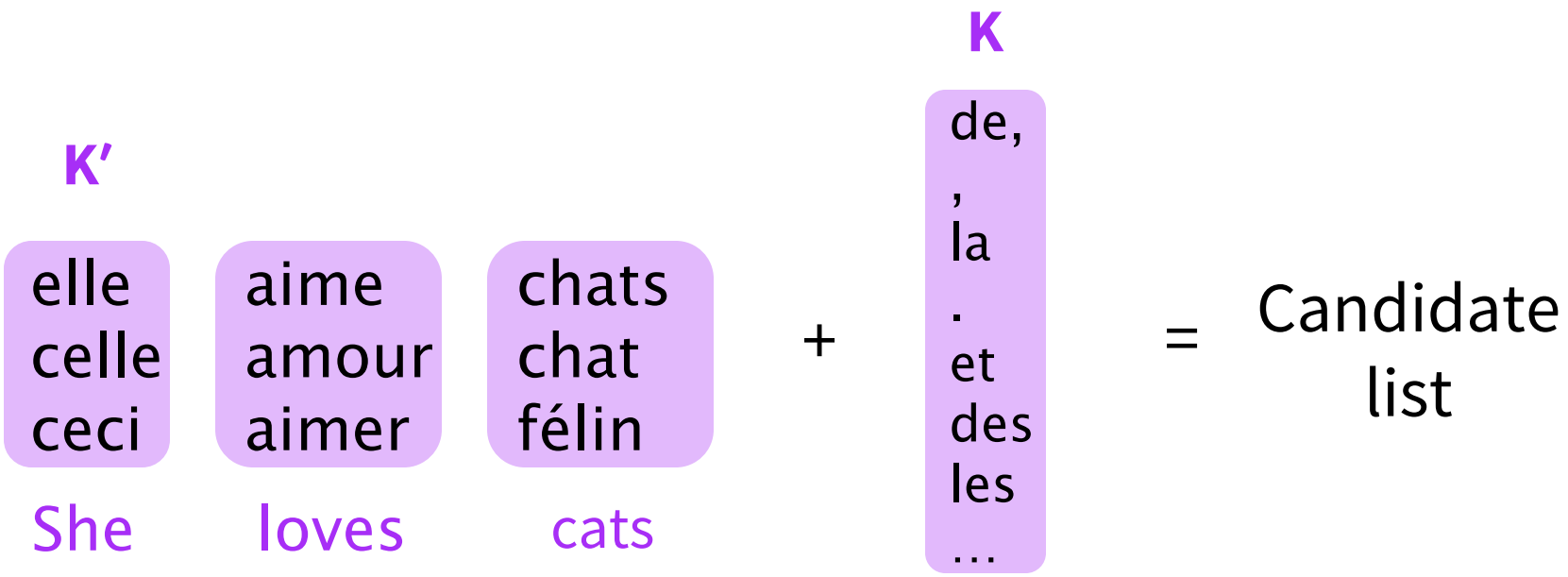
- **K** most frequent words: unigram prob.

de,
,
la
.
et
des
les
…

- Candidate target words
  - **K′** choices per source word. K′ = 3.

She	loves	cats
elle	aime	chats
celle	amour	chat
ceci	aimer	félin

Testing – *Select candidate words*



- Produce translations within the candidate list
- *Practice*: K′ = 10 or 20, K = 15k, 30k, or 50k.

More on large-vocab techniques

- "BlackOut: Speeding up Recurrent Neural Network Language Models with very Large Vocabularies" – [Ji, Vishwanathan, Satish, Anderson, Dubey, ICLR'16].
  - Good survey over many techniques.

- "Simple, Fast Noise Contrastive Estimation for Large RNN Vocabularies" – [Zoph, Vaswani, May, Knight, NAACL'16].
  - Use the same samples per minibatch. GPU efficient.

2nd thought on word generation

- Scaling softmax is insufficient:
  - New names, new numbers, etc., at test time.

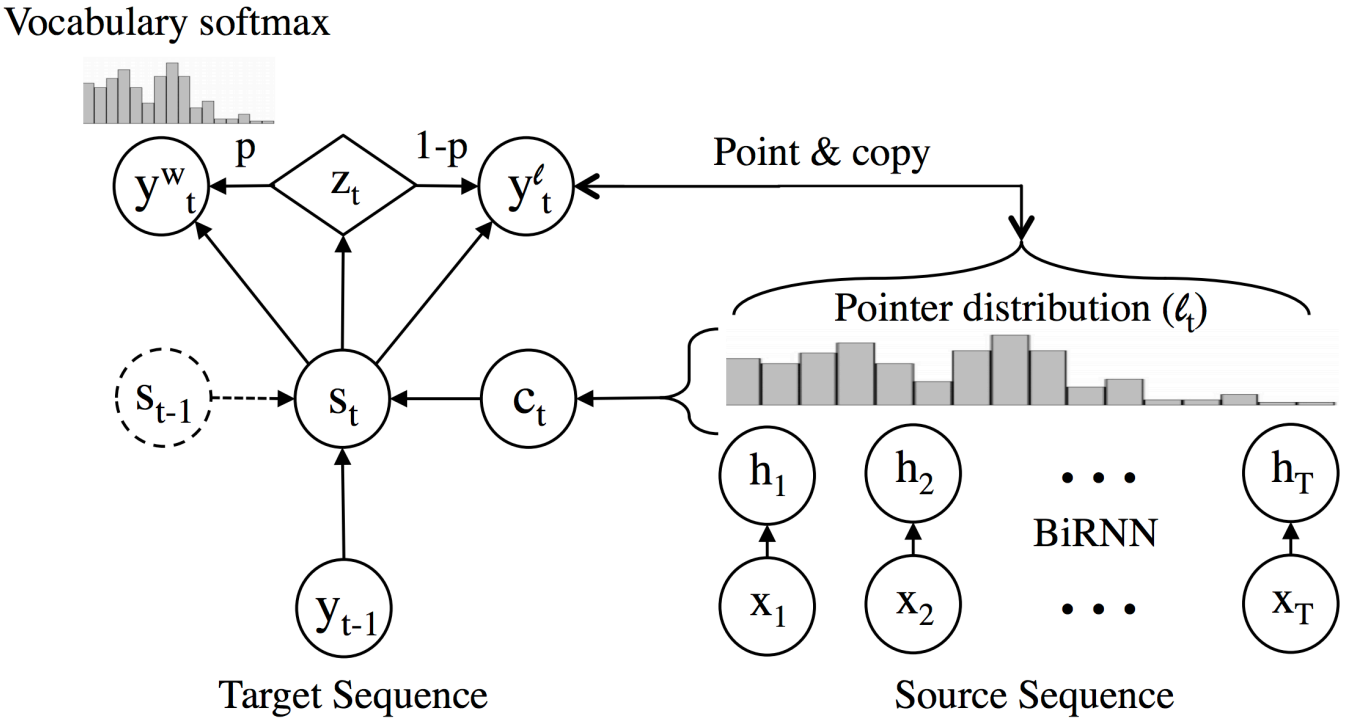
- But previous MT models can copy words.

Models with a pointer/copying

- Recall the Pointer Sentinel Mixture Models (Merity et al. 2017) that Richard mentioned

Copying/pointer networks

- Gulcehre, Ahn, Nallapati, Zhou, Bengio (2016) Pointing the Unknown Words



Copying/pointer networks

Table 5: Europarl Dataset (EN-FR)

	BLEU-4
NMT	20.19
NMT + PS	**23.76**

- Caution from Google NMT paper: In principle can train a "copy model" but this approach is both unreliable at scale – the attention mechanism is unstable when the network is deep – and copying may not always be the best strategy for rare words – sometimes transliteration is more appropriate

Extending NMT to more languages

- "Copy" mechanisms are **not sufficient**.
  - Transliteration: Christopher ↦ Kryštof
  - Multi-word alignment: Solar system ↦ Sonnensystem

- Need to handle **large, open vocabulary**
  - Rich morphology: nejneobhospodařovávatelnějšímu
    ("to the worst farmable one")
  - Informal spelling: goooooood morning !!!!!

Be able to operate at sub-word levels.

Sub-word NMT: two trends

- Same seq2seq architecture:
  - Use smaller units.
  - [Sennrich, Haddow, Birch, ACL'16a], [Chung, Cho, Bengio, ACL'16].
- Hybrid architectures:
  - RNN for *words* + something else for *characters*.
  - [Costa-Jussà & Fonollosa, ACL'16], [Luong & Manning, ACL'16].

Byte Pair Encoding

- A compression algorithm:
  - Most frequent byte pair ↦ a new byte.

Replace bytes with character ngrams

Rico Sennrich, Barry Haddow, and Alexandra Birch. ***Neural Machine Translation of Rare Words with Subword Units****. ACL 2016.*

Byte Pair Encoding

- A word segmentation algorithm:
  - Start with a vocabulary of characters.
  - Most frequent ngram pairs ↦ a new ngram.

Byte Pair Encoding

- A word segmentation algorithm:
  - Start with a vocabulary of characters.
  - Most frequent ngram pairs ↦ a new ngram.

Dictionary

5 l o w
2 l o w e r
6 n e w e s t
3 w i d e s t

Vocabulary

l, o, w, e, r, n, w, s, t, i, d

Start with all characters in vocab

(Example from Sennrich)

Byte Pair Encoding

- A word segmentation algorithm:
  - Start with a vocabulary of characters.
  - Most frequent ngram pairs ↦ a new ngram.

Dictionary

5 l o w
2 l o w e r
6 n e w **es** t
3 w i d **es** t

Vocabulary

l, o, w, e, r, n, w, s, t, i, d, **es**

Add a pair (e, s) with freq 9

(Example from Sennrich)

Byte Pair Encoding

- A word segmentation algorithm:
  - Start with a vocabulary of characters.
  - Most frequent ngram pairs ↦ a new ngram.

Dictionary

5 l o w
2 l o w e r
6 n e w **est**
3 w i d **est**

Vocabulary

l, o, w, e, r, n, w, s, t, i, d, es, **est**

Add a pair (es, t) with freq 9

(Example from Sennrich)

Byte Pair Encoding

- A word segmentation algorithm:
  - Start with a vocabulary of characters.
  - Most frequent ngram pairs ↦ a new ngram.

Dictionary

5 **lo** w
2 **lo** w e r
6 n e w est
3 w i d est

Vocabulary

l, o, w, e, r, n, w, s, t, i, d, es, est, **lo**

Add a pair (l, o) with freq 7

(Example from Sennrich)

Byte Pair Encoding

- A word segmentation algorithm:
  - Start with a vocabulary of characters.
  - Most frequent ngram pairs ↦ a new ngram.
- Automatically decide vocabs for NMT

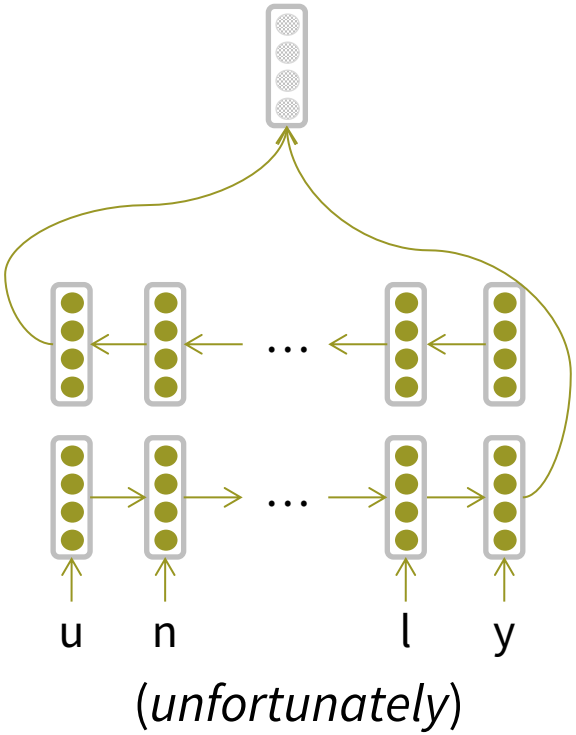
Top places in WMT 2016!

https://github.com/rsennrich/nematus

Wordpiece model

- GNMT uses a variant of this, the wordpiece model, which is generally similar but uses a greedy approximation to maximizing language model log likelihood to choose the pieces

Character-based LSTM

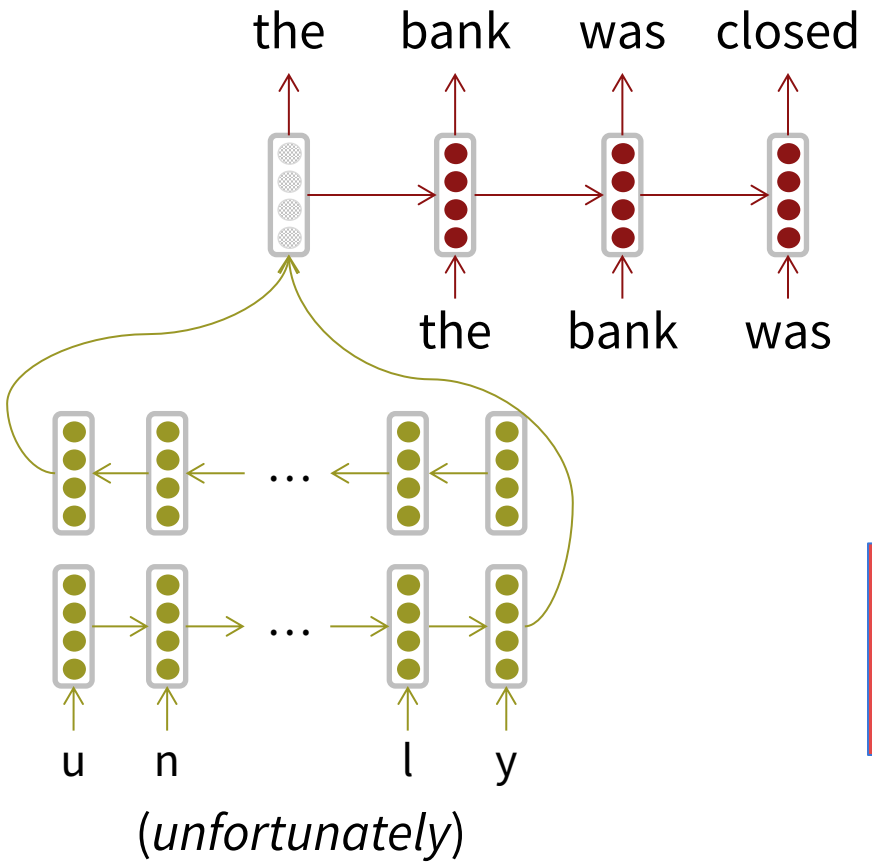


(*unfortunately*)

Bi-LSTM builds word representations

Ling, Luís, Marujo, Astudillo, Amir, Dyer, Black, Trancoso. ***Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation****. EMNLP'15.*

Character-based LSTM



Recurrent Language Model

Bi-LSTM builds word representations

Ling, Luís, Marujo, Astudillo, Amir, Dyer, Black, Trancoso. ***Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation****. EMNLP'15.*

Hybrid NMT

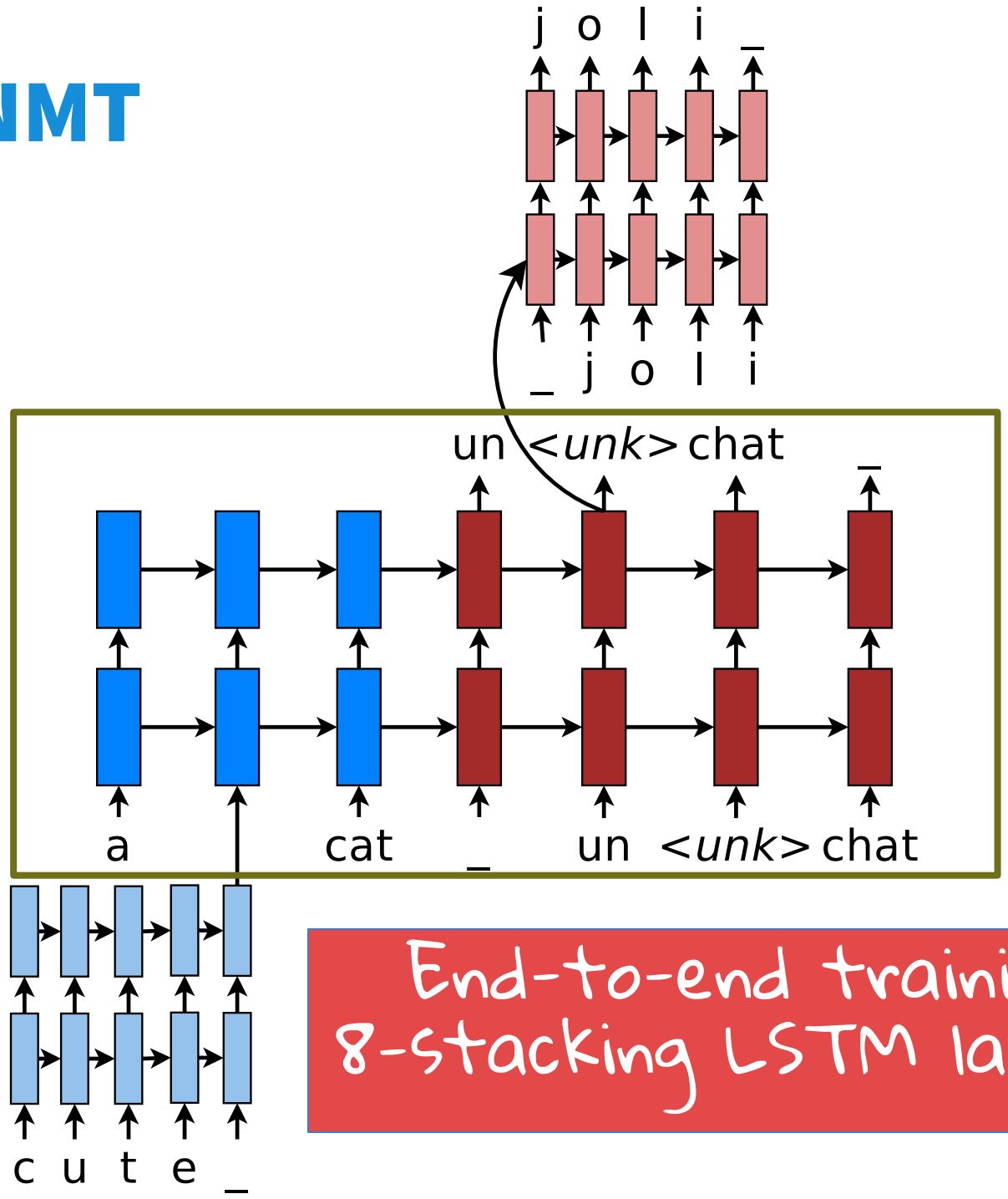

- A *best-of-both-worlds* architecture:
  - Translate mostly at the word level
  - Only go to the character level when needed.

- More than 2 BLEU improvement over a copy mechanism.

Thang Luong and Chris Manning. ***Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models****. ACL 2016.*

Hybrid NMT

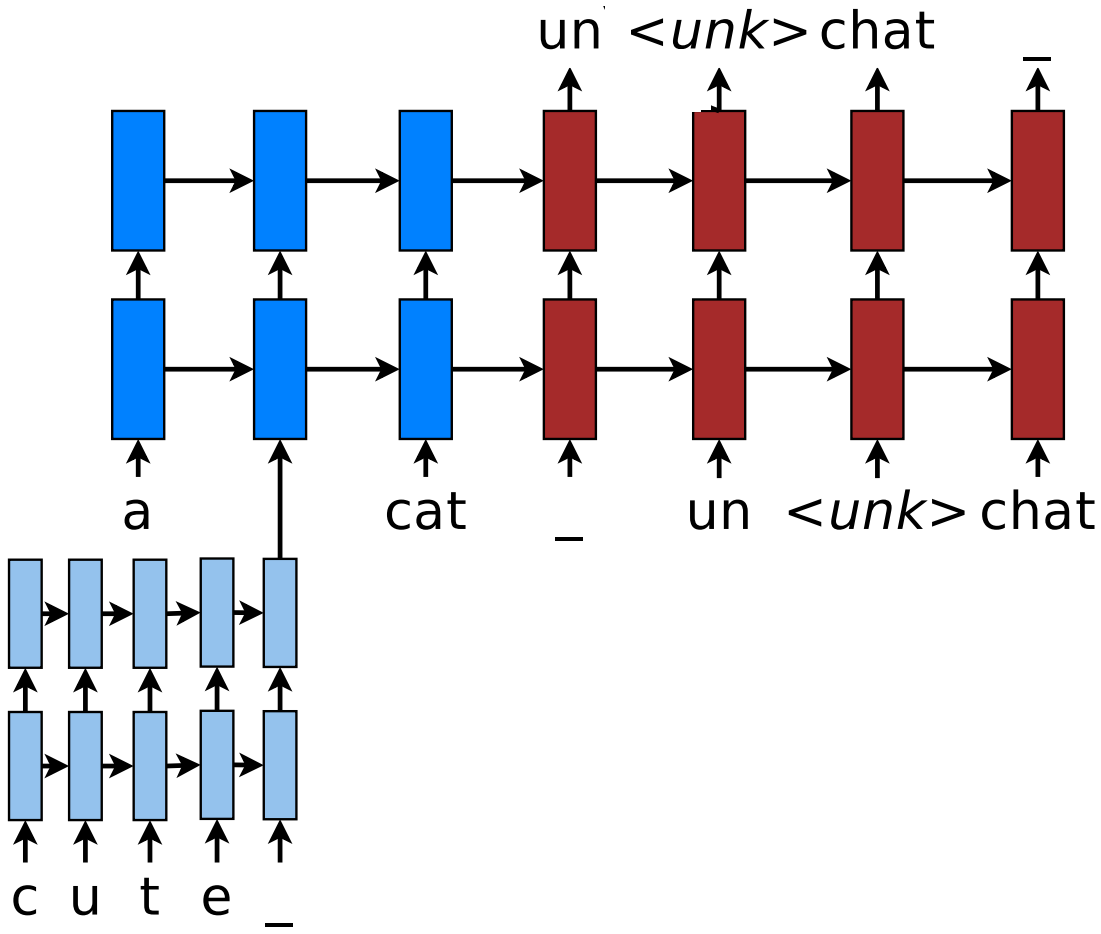
Word-level
(4 layers)



End-to-end training
8-stacking LSTM layers.

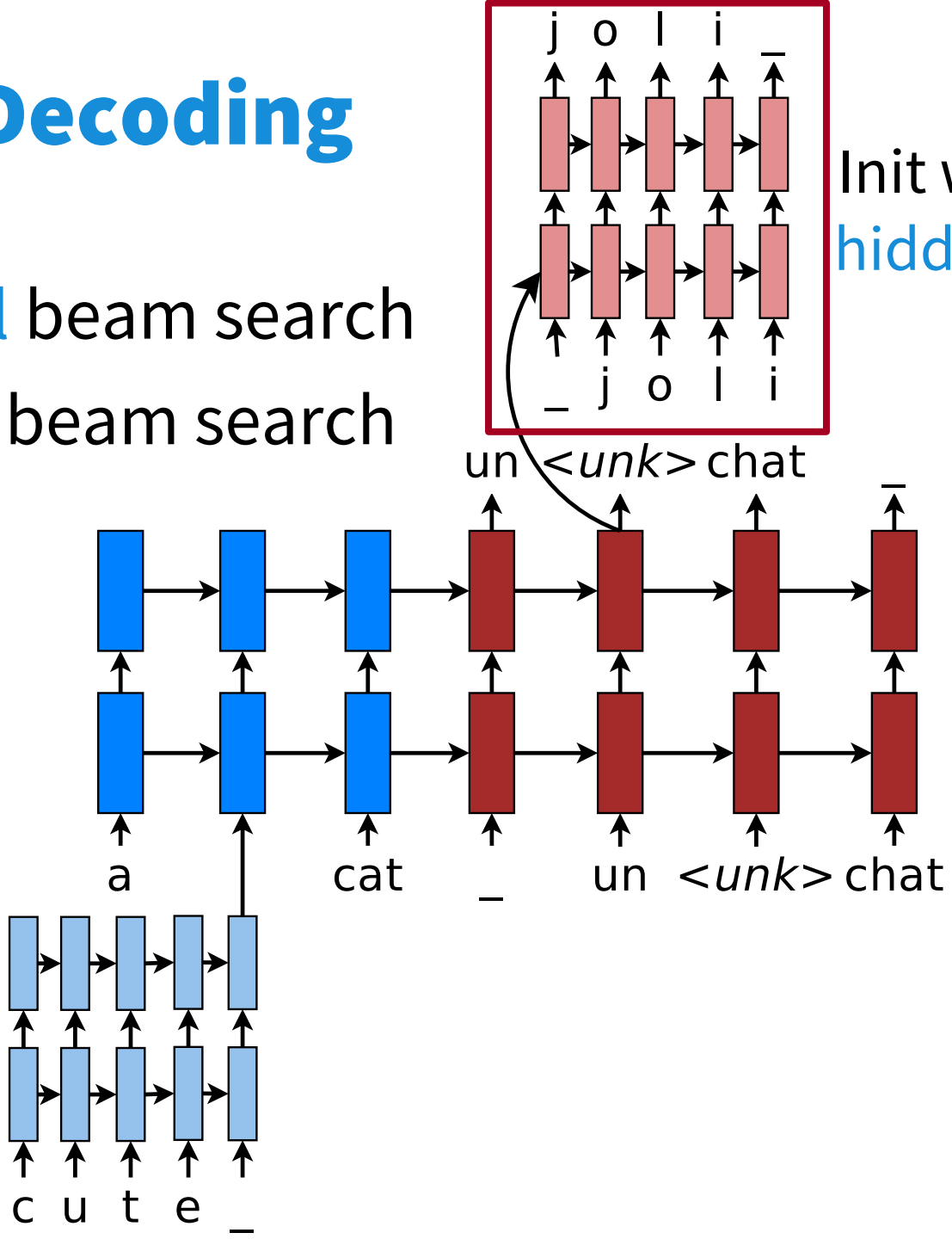
2-stage Decoding

- Word-level beam search



2-stage Decoding

- Word-level beam search
- Char-level beam search for <unk>.



English-Czech Results

- Train on WMT'15 data (12M sentence pairs)
  - newstest2015

Systems	BLEU
Winning WMT'15 (Bojar & Tamchyna, 2015)	18.8
Word-level NMT (Jean et al., 2015)	18.3

30x data
3 systems

Large vocab
+ copy mechanism

English-Czech Results

- Train on WMT'15 data (12M sentence pairs)
  - newstest2015

Systems	BLEU
Winning WMT'15 (Bojar & Tamchyna, 2015)	18.8
Word-level NMT (Jean et al., 2015)	18.3
Hybrid NMT (Luong & Manning, 2016)*	**20.7**

30x data
3 systems

Large vocab
+ copy mechanism

Sample English-Czech translations

source	Her ***11-year-old*** daughter , ***Shani Bart*** , said it felt a little bit ***weird***
human	Její **jedenáctiletá** dcera **Shani Bartová** prozradila , že je to trochu **zvláštní**
word	Její <unk> dcera <unk> <unk> řekla , že je to trochu divné
	Její **11-year-old** dcera **Shani ,** řekla , že je to trochu ***divné***
hybrid	Její <unk> dcera , <unk> <unk> , řekla , že je to <unk> <unk>
	Její **jedenáctiletá** dcera , **Graham** ***Bart*** , řekla , že cítí trochu ***divný***

- Word-based: identity copy fails.

Sample English-Czech translations

source	Her *11-year-old* daughter , *Shani Bart* , said it felt a little bit *weird*
human	Její **jedenáctiletá** dcera **Shani Bartová** prozradila , že je to trochu **zvláštní**
word	Její <unk> dcera <unk> <unk> řekla , že je to trochu divné
	Její **11-year-old** dcera **Shani ,** řekla , že je to trochu ***divné***
hybrid	Její <unk> dcera , <unk> <unk> , řekla , že je to <unk> <unk>
	Její **jedenáctiletá** dcera , **Graham** ***Bart*** , řekla , že cítí trochu ***divný***

- Hybrid: correct, ***11-year-old*** – **jedenáctiletá**.