Real-Time Translation
Learning to Decode

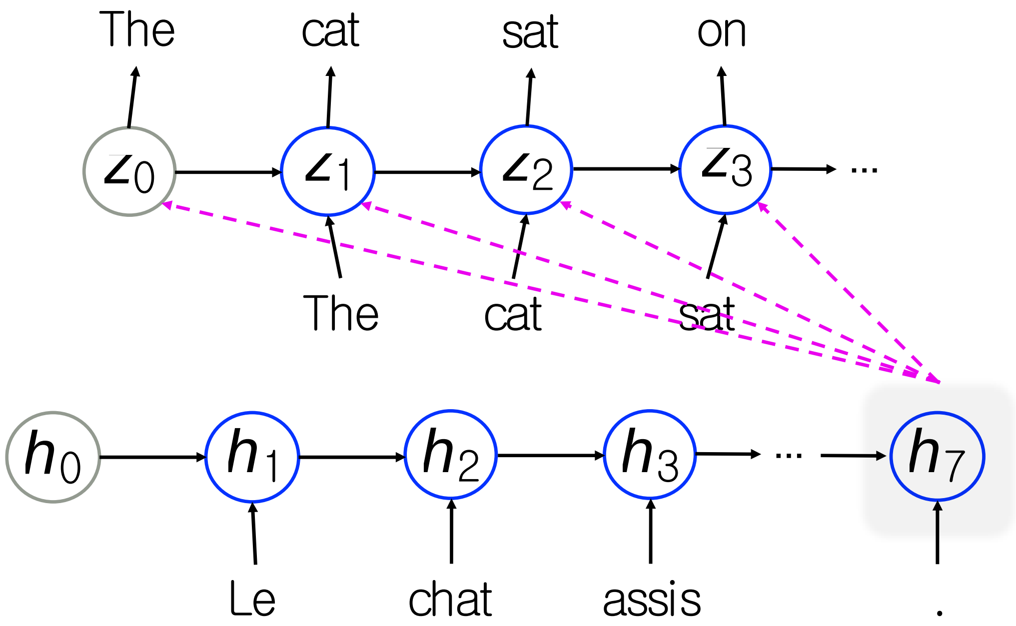
Kyunghyun Cho
New York University
Courant Institute (Computer Science) and Center for Data Science
Facebook AI Research

Jiatao Gu, Victor Li, Graham Neubig, Kyunghyun Cho.
Learning to translate in real-time with neural machine translation. EACL'17.

Decoding
from a recurrent language model

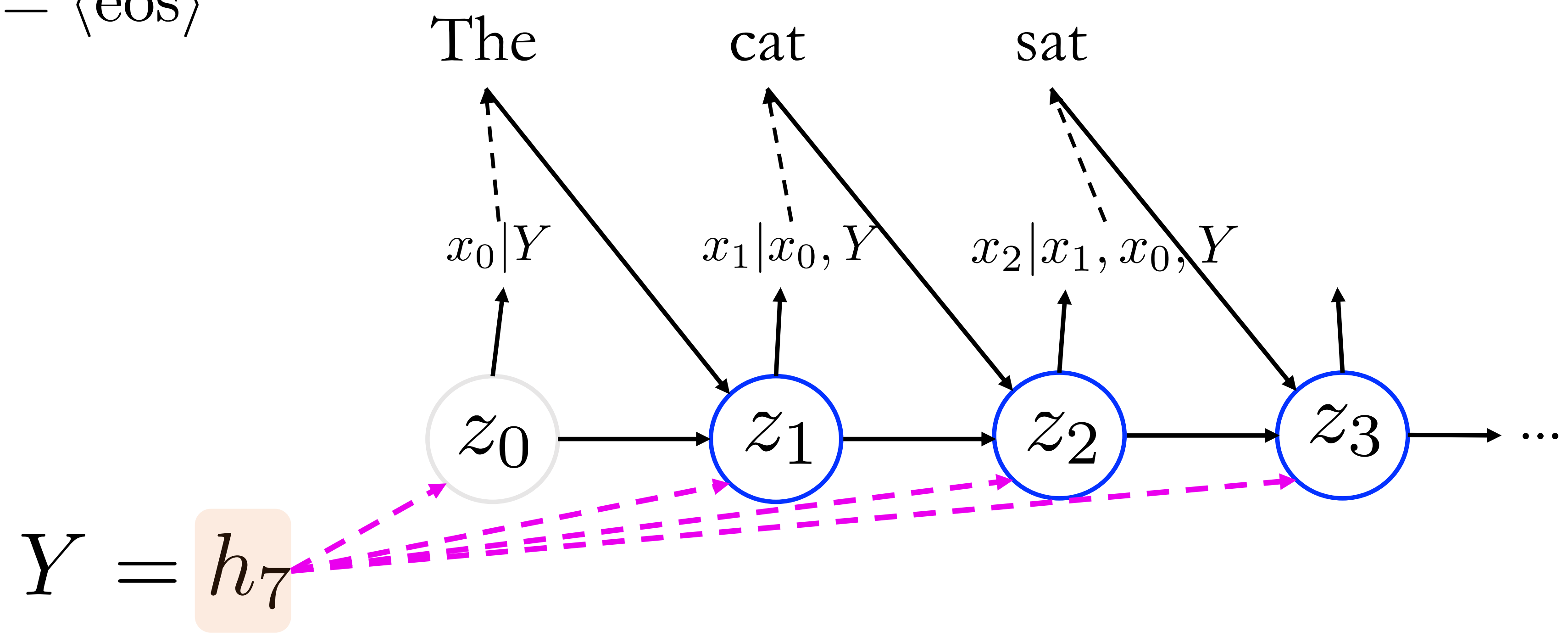
Decoding (0) – Exhaustive Search

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



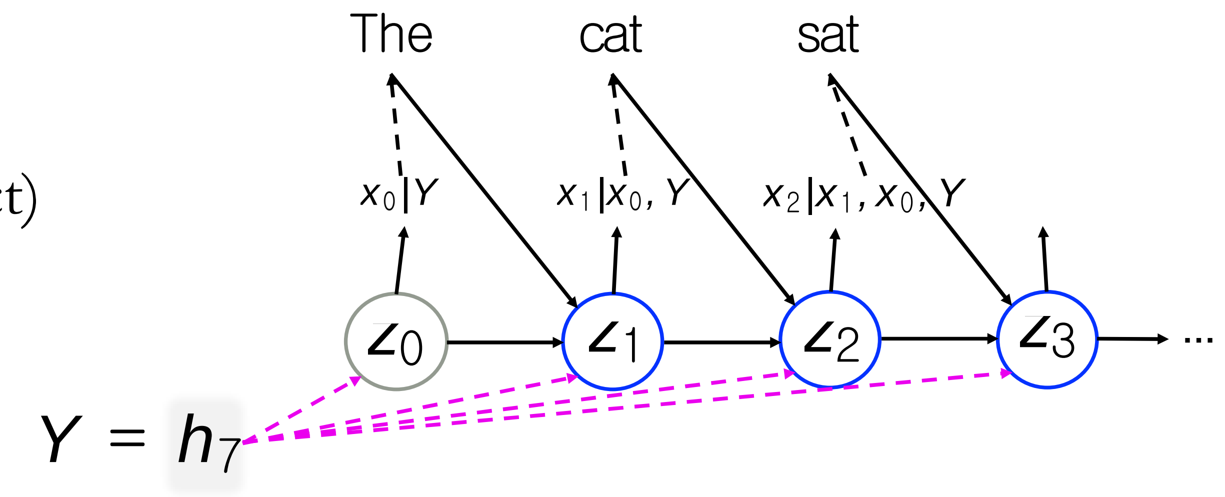
Decoding (1) – Ancestral Sampling

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



Decoding (1) – Ancestral Sampling

- Efficient, unbiased 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 this procedure for N times: $\left\{ \tilde{X}^1, \ldots, \tilde{X}^N \right\}$
- Choose $\arg\max_{\tilde{X}^n} \log p(\tilde{X}^n | Y)$
- Pros:
  1. Unbiased (asymptotically exact)
- Cons:
  1. High variance
  2. Pretty inefficient



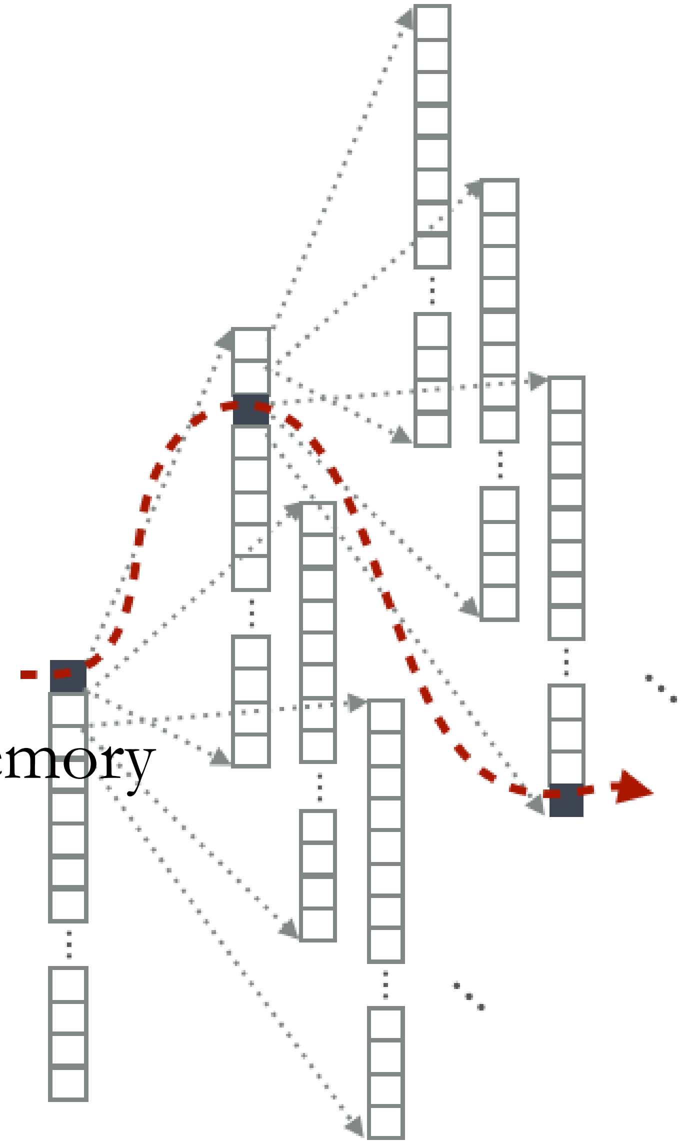
Decoding (2) – Greedy Search

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

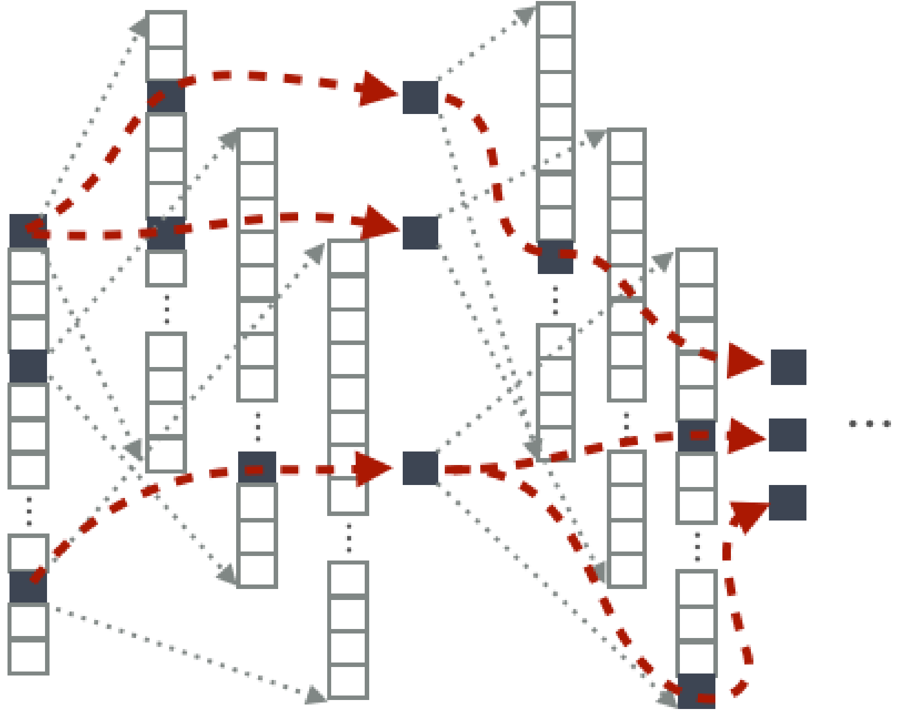
$$\tilde{x}_t = \arg\max_x \log p(x|x_{<t}, Y)$$

- 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 quite 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

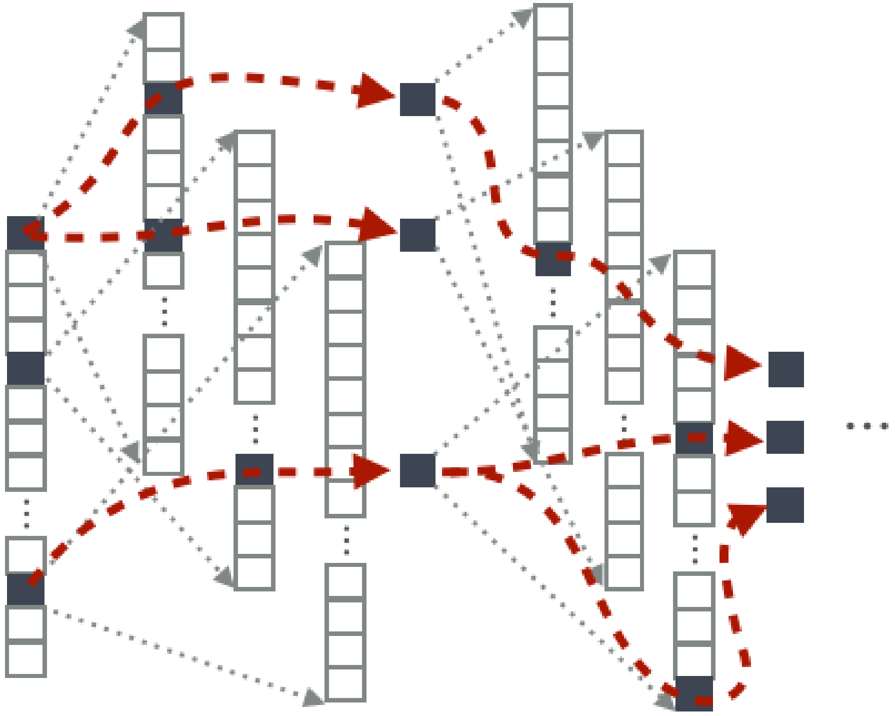
$$\mathcal{H}_t^k = \left\{(\tilde{x}_1^k, \tilde{x}_2^k, \ldots, \tilde{x}_{t-1}^k, v_1), (\tilde{x}_1^k, \tilde{x}_2^k, \ldots, \tilde{x}_{t-1}^k, v_2), \ldots, (\tilde{x}_1^k, \tilde{x}_2^k, \ldots, \tilde{x}_{t-1}^k, v_{|V|})\right\}$$

- 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 $K \to \infty$
- Not necessarily monotonic improvement w.r.t. K
- K is selected to maximize the translation quality on a **validation** set.



Decoding

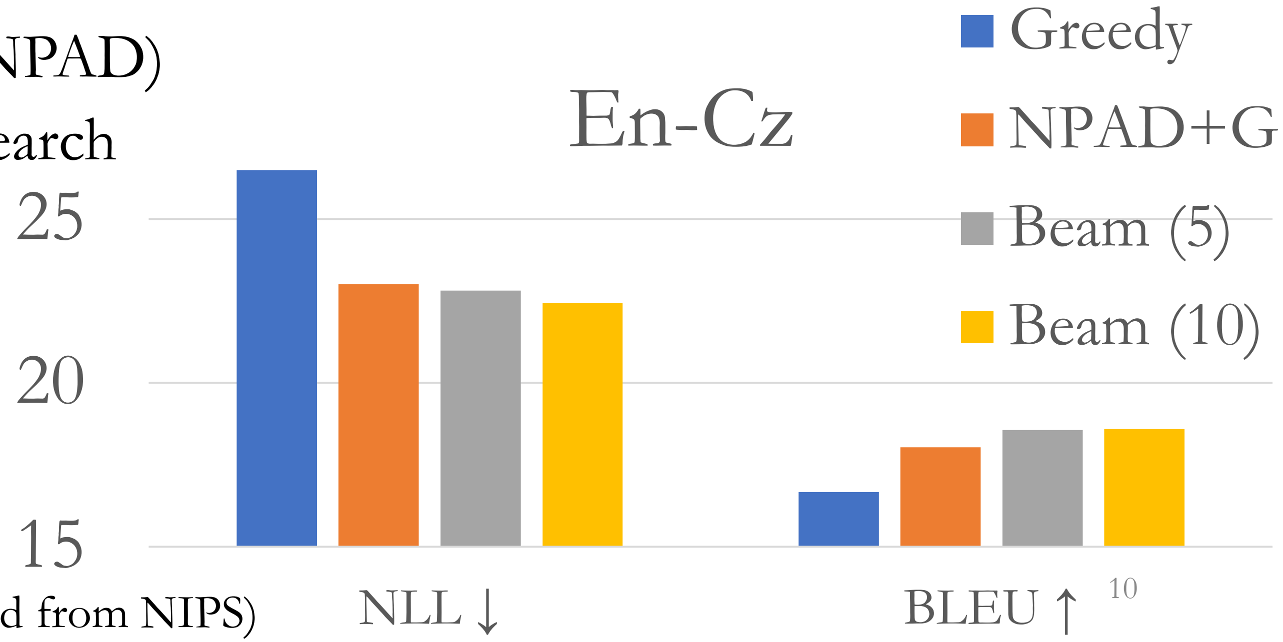
- English to Czech Translation: 12m training sentence pairs

Strategy	# Chains	Valid NLL↓	Valid BLEU↑	Test-1 NLL↓	Test-1 BLEU↑
Sto. Sampling	50	22.9818	15.64	26.2536	16.76
Greedy	-	27.879	15.5	26.4928	16.66
Beam	5	20.1842	17.03	22.8106	18.56
Beam	10	19.9173	17.13	22.4392	18.59

Cho (2016)

Inference is difficult and expensive (1)

- State space grows exponentially w.r.t. the (max) length of a sentence
  - $|V| + |V|^2 + \cdots + |V|^T$ possible sentences with $|V| \approx 10^3 \sim 10^6$ $T \approx 10 \sim 300$
  - No obvious way to reduce the search space: non-Markovian model
- Cheap, approximate search is often too approximate
  - Greedy decoding: cheap, approximate search
  - Noisy, parallel approximate decoding (NPAD)
  - Beam search: expensive, approximate search



(Cho, 2016 horribly rejected from NIPS)

Inference is difficult and expensive (2)

- *Is this what we want?*

$$\arg\max_{\theta} \sum_{t=1}^{T} \log p_{\theta}(x_t | x_{<t}, s)$$

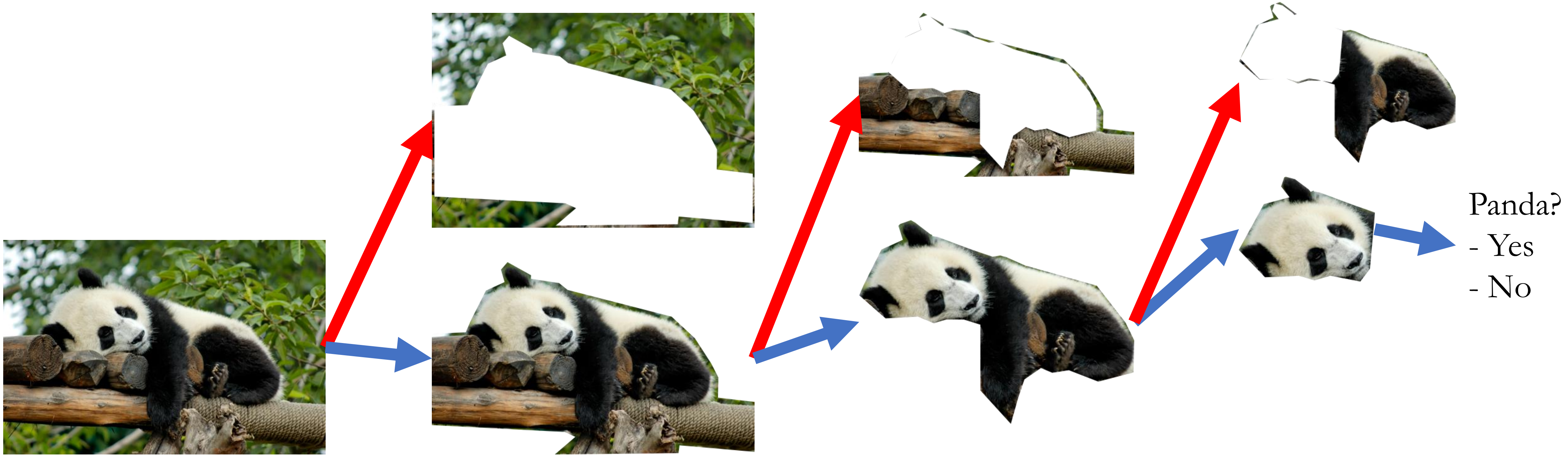
- Decoding objectives are *not known* in advance
  - MT for real-time conversation: quality ↑ vs. delay ↓
  - MT for K-12 students: quality ↑ vs. text difficulty ↓
  - On-device translation: quality ↑ vs. computational complexity ↓

- Even if so, little or no data available
  - Simultaneous interpretation: almost none with time stamps [He et al., 2016 NAACL]
  - Parallel corpora with controlled levels of difficulty: none

What can we do about it?

Learning to decode

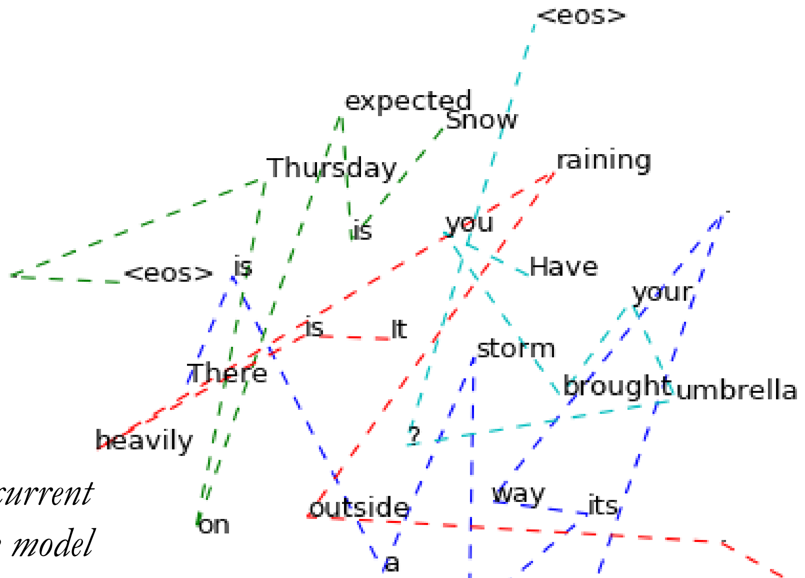
Neural network = Forgetting machine



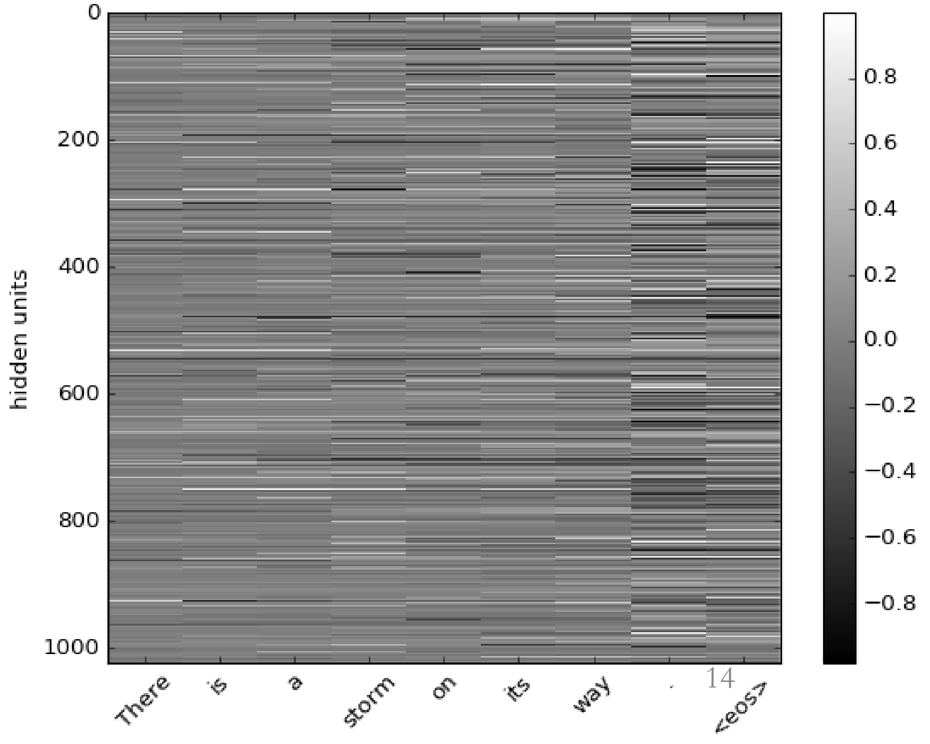
- A deep neural net iteratively disentangles relevant and irrelevant features
- Irrelevant features are discarded as information propagates
- In other words, *hidden layers contain rich info beyond the task*!

Exploiting the hidden activation

- What is captured by the hidden layers?
  - Deep Visualization: Edges/corners → textures → object parts → entire objects [Zeiler&Fergus, 2014 ECCV; Yosinski et al., 2016 DL; and many more]
  - Long-range dependency: closing brackets, agreement, … [Karpathy et al., 2015 arXiv; Tran et al., 2016 NAACL]
  - Sentiment! [Radford et al., 2017 OpenAI]
- Fairly limited understanding *especially* with text
- Why?

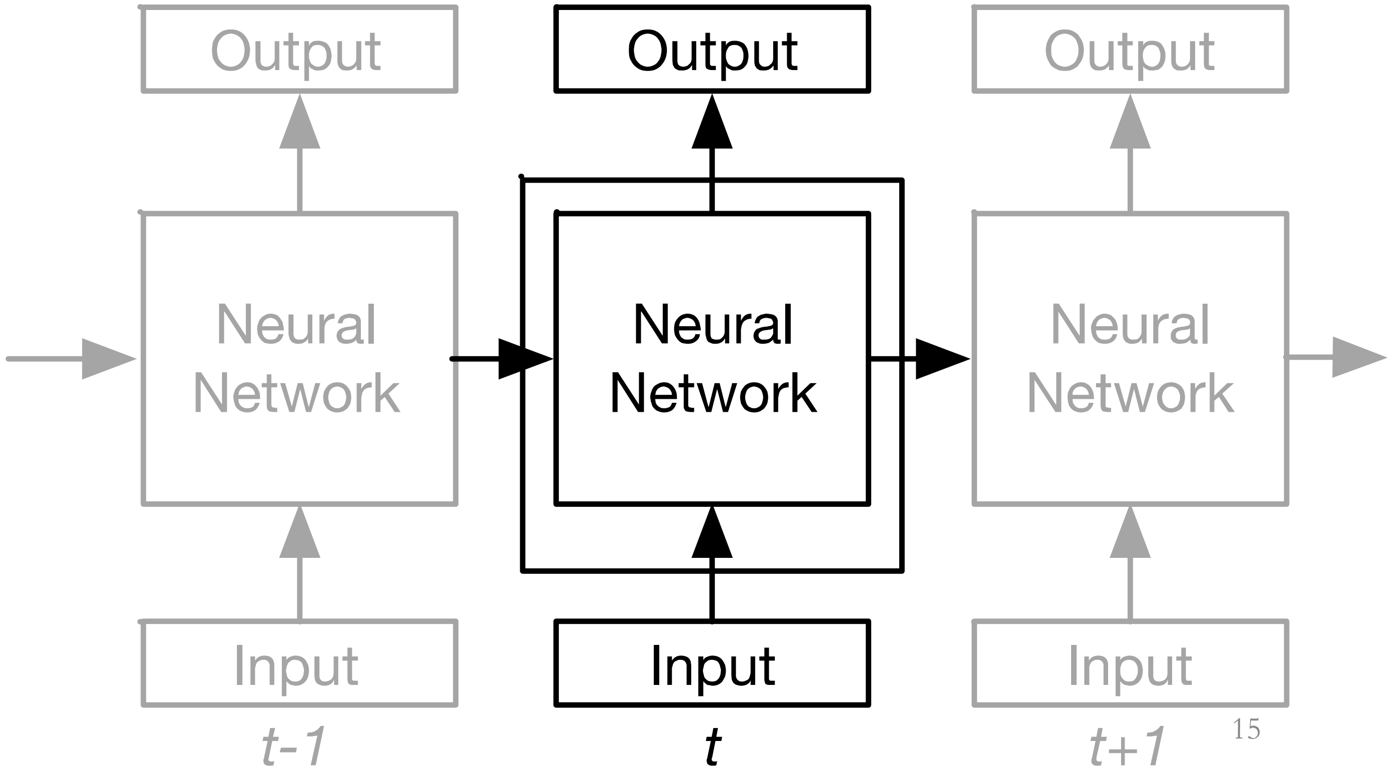


Hidden activations of a small recurrent language model



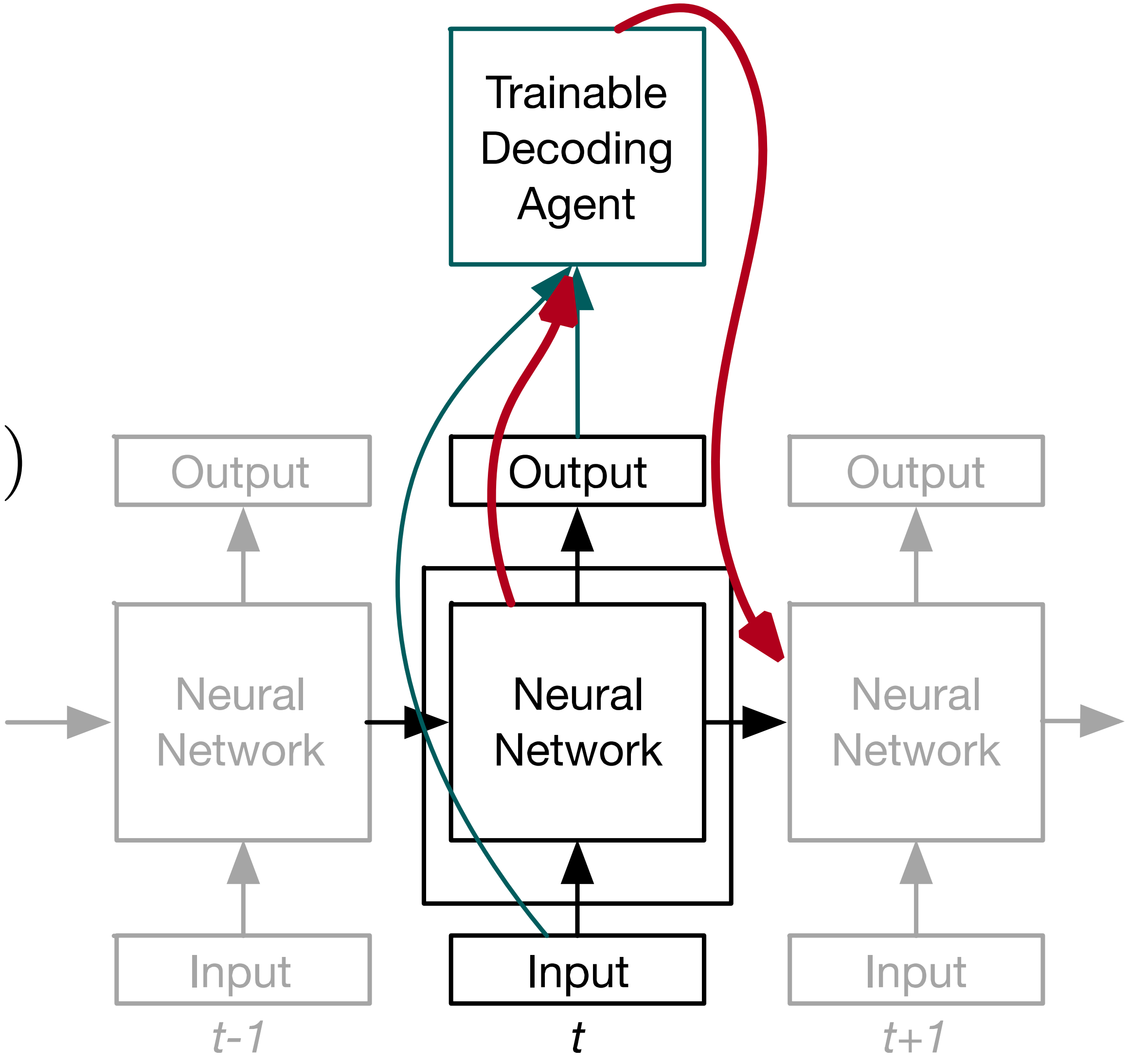
Trainable Decoding (1)

- A conditional recurrent neural net *defines* an environment
- State:
  - Previous hidden state h_{t-1}
  - Current input $\hat{x}_{t-1}$
  - Source context $c_t(s)$
- Action: any modification
  - Next input $\hat{x}_t$
  - Source s
- Reward: arbitrary



Trainable Decoding (2)

- A conditional recurrent neural net *defines* an environment
- A decoder is an agent:
  - Observes the state via $p(x_t|\hat{x}_{<t}, s)$
  - Acts by selecting $\hat{x}_t$
- Limited, because it doesn't exploit rich info captured in h_t
- *Can we extend it by training a neural network decoder?*



Yes, we can!

- Simultaneous Translation
  - Jiatao Gu, Kyunghyun Cho, Victor OK Li. Trainable greedy decoding for neural machine translation. EMNLP 2017
- Trainable Greedy Decoding
  - Jiatao Gu, Graham Neubig, Kyunghyun Cho, Victor OK Li. Learning to translate in real-time with neural machine translation. EACL 2017.
  - Yun Chen, Victor Li, Sam Bowman, Kyunghyun Cho. A Stable and Effective Learning Strategy for Trainable Greedy Decoding. (under review)

Simultaneous Translation (1)

- Inspired by simultaneous interpretation
- Source words arrive one at a time
- Translation starts before the complete sentence arrives
- Objective: quality ↑ delay ↓

Interpreters at the Nuremberg Trial (1945-1946)
https://www.pri.org/stories/2014-09-29/how-do-all-those-leaders-un-communicate-all-those-languages

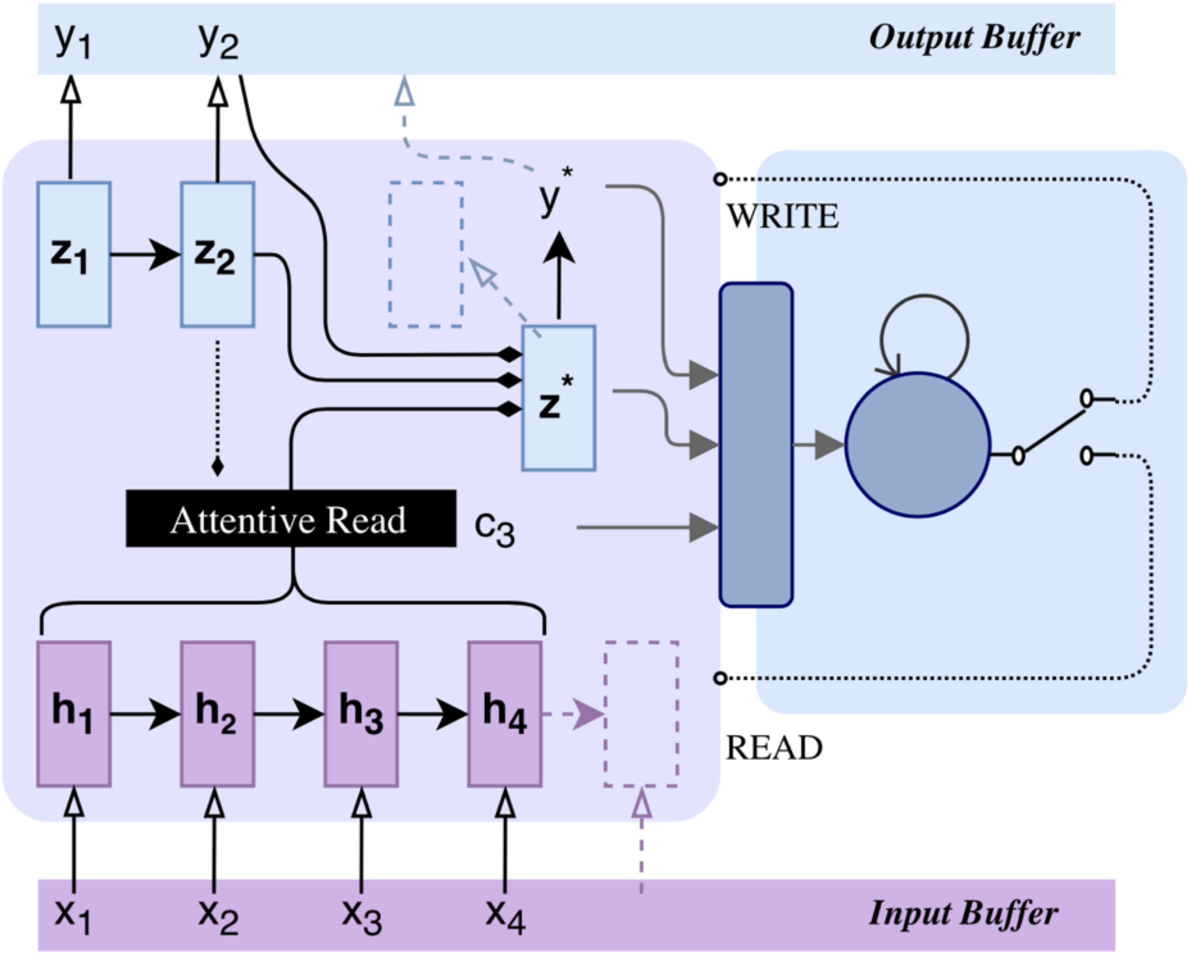
Simultaneous Translation (2)

Decoding

1. Start with a well-trained NMT
2. A simultaneous decoder intercepts and interprets the incoming signal
3. The simultaneous decoder forces the pretrained model to either
    1. output a target symbol, or
    2. wait for the next source symbol

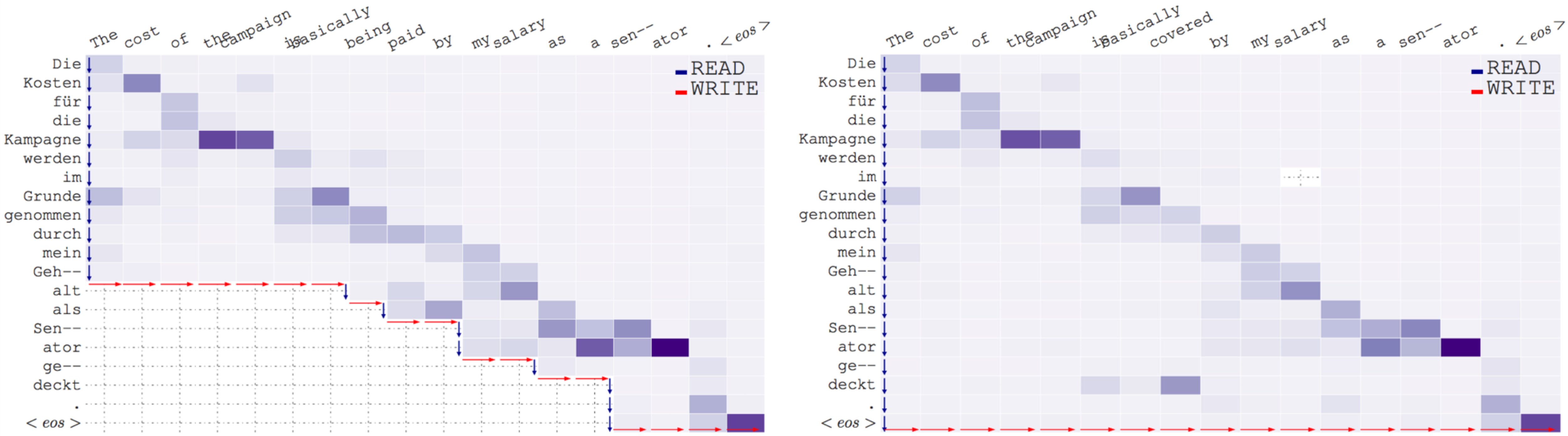
Learning

1. Trade-off between delay and quality
2. Policy gradient (REINFORCE)



[Gu, Neubig, Cho & Li, EACL 2017]

Simultaneous Translation (3)

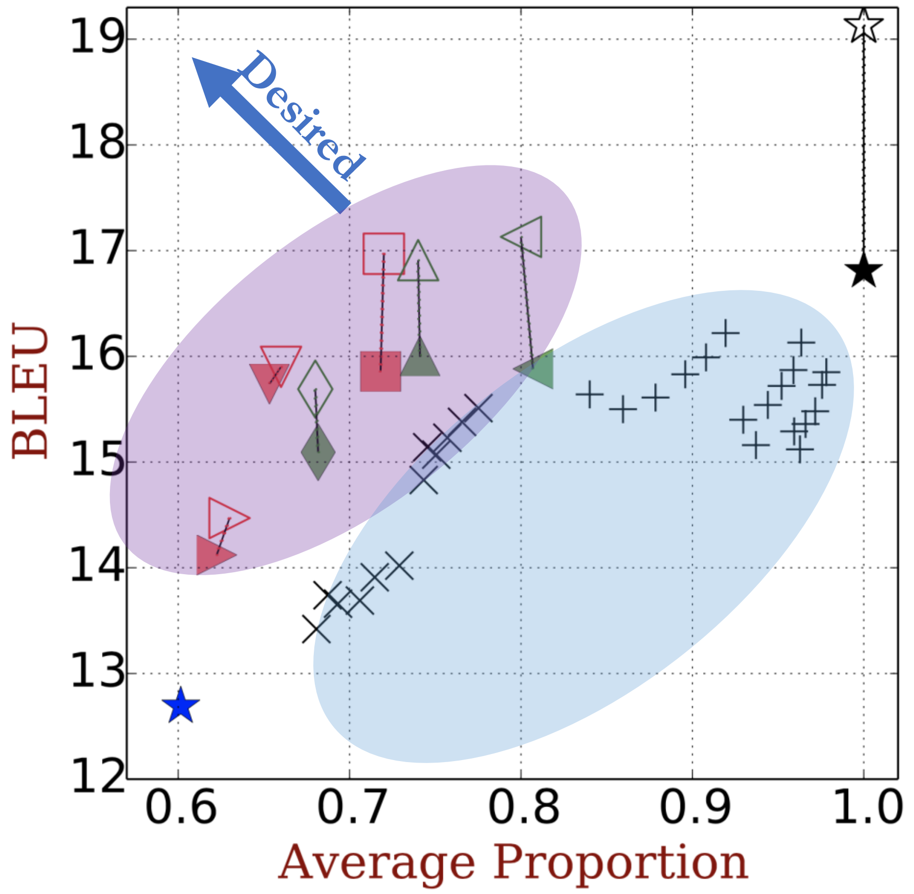


(a) Simultaneous Neural Machine Translation

(b) Neural Machine Translation

[Gu, Neubig, Cho & Li, EACL 2017]

Simultaneous Translation (4)



★ consecutive translation

★ word-by-word translation

+x simultaneous translation without using h_t

■ ▲ simultaneous translation (trainable decoding)

- *Better simultaneous translation by exploiting the rich info captured by the hidden state*

[Cho & Esipova, 2016 horribly rejected from EMNLP]
[Gu, Neubig, Cho & Li, EACL 2017]