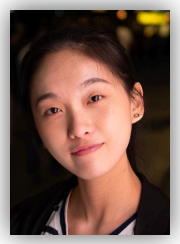
Hint-Based Training for Non-Autoregressive Translation

Zhuohan Li Zi Lin Di He Fei Tian Tao Qin Liwei Wang Tie-Yan Liu

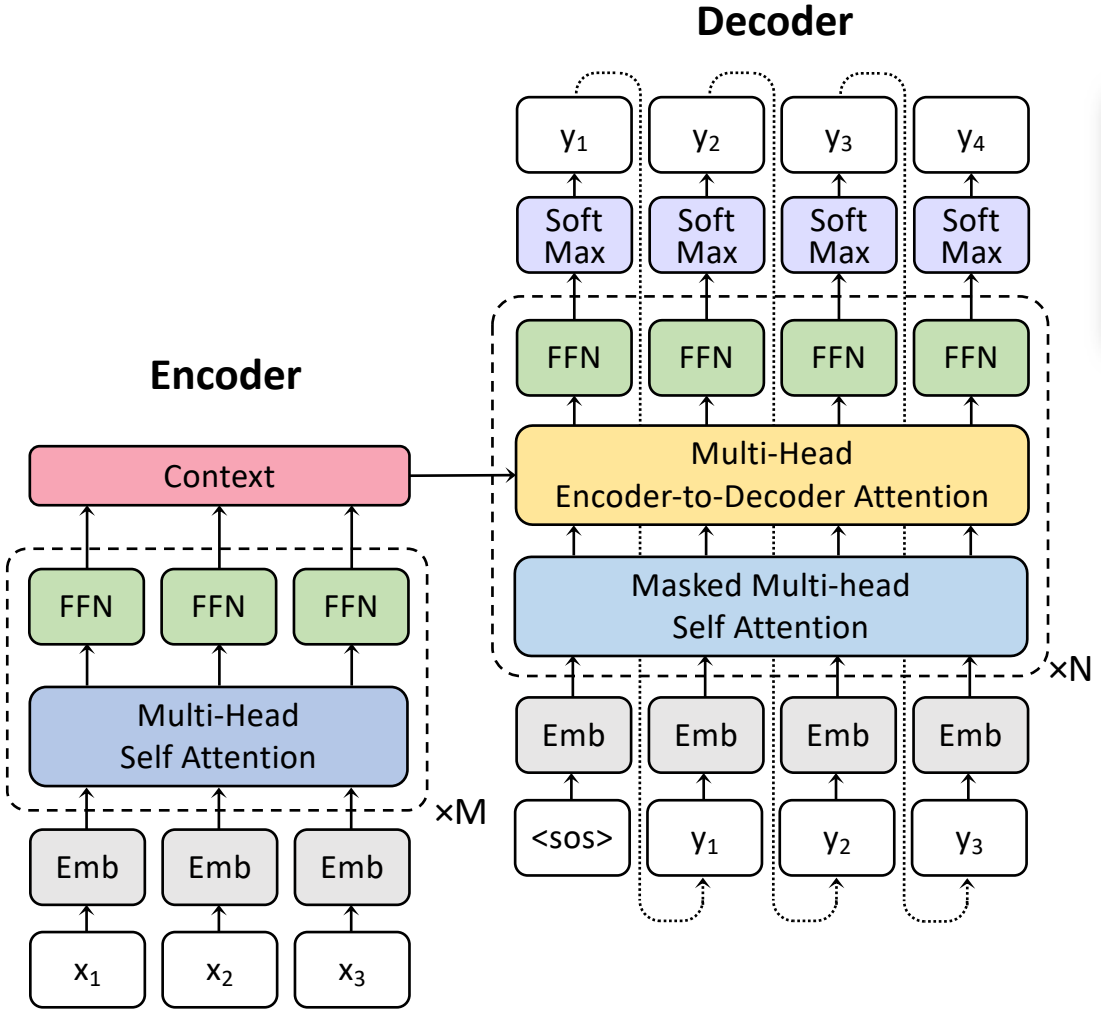


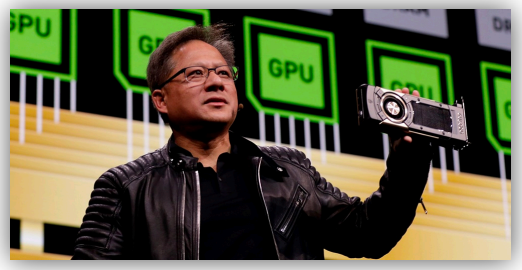


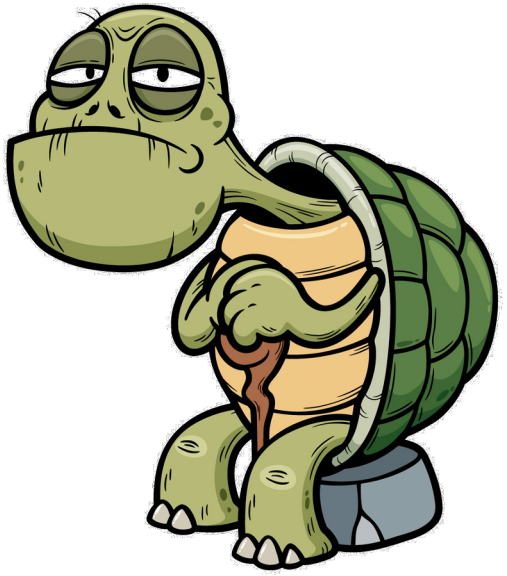

EMNLP-IJCNLP 2019

Berkeley UNIVERSITY OF CALIFORNIA
北京大学 PEKING UNIVERSITY
Microsoft

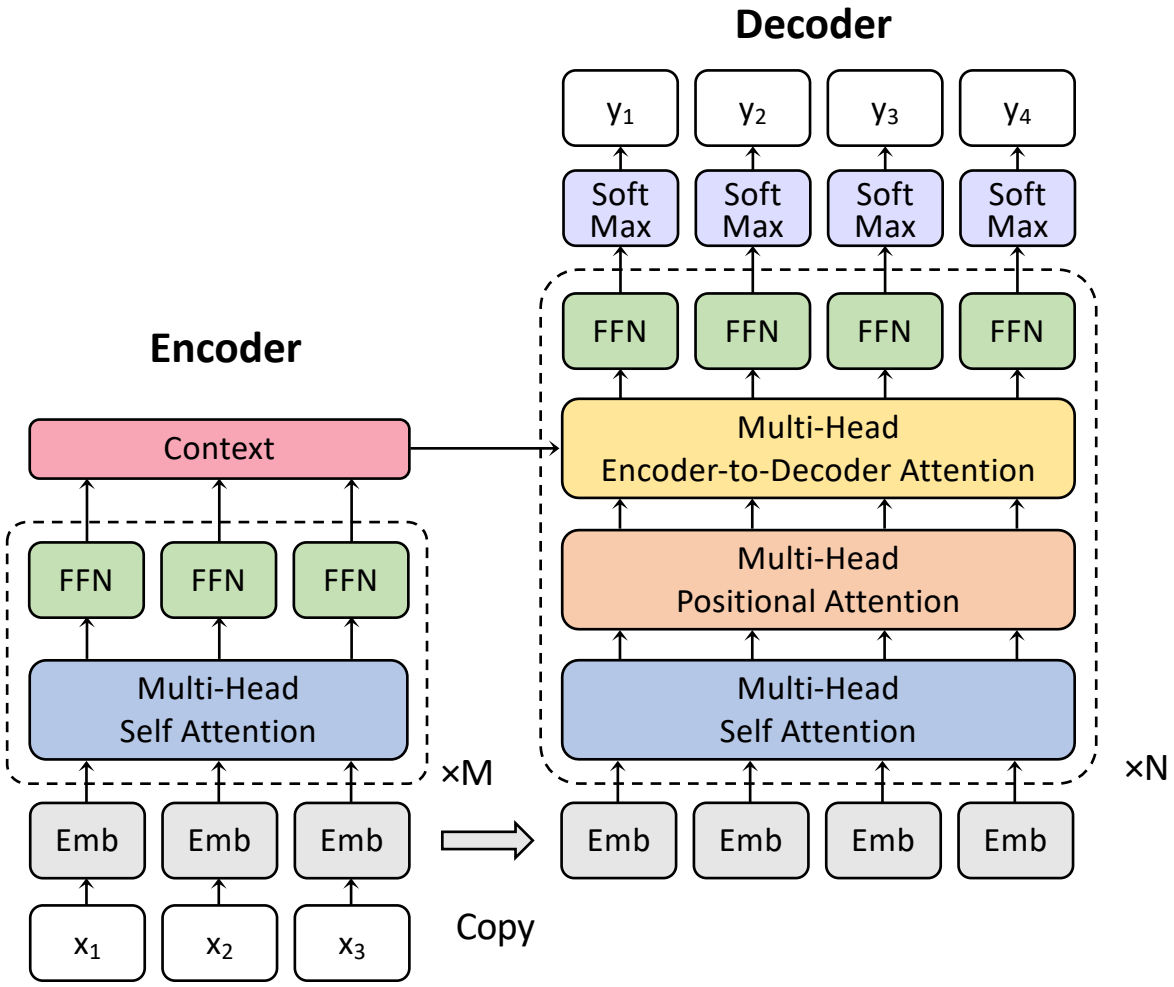
Autoregressive MT models

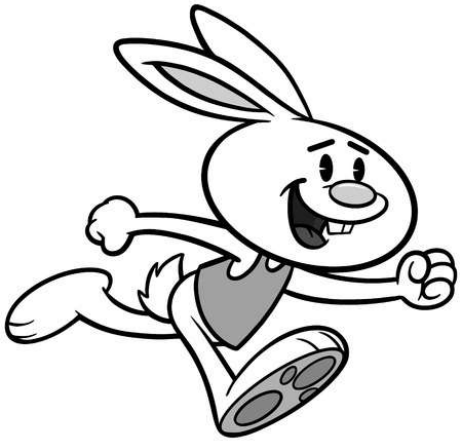


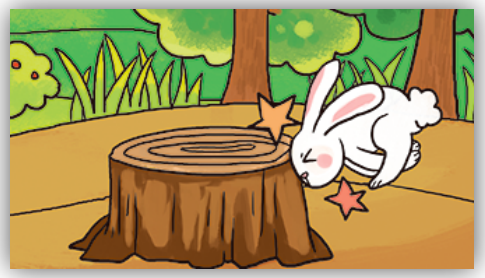




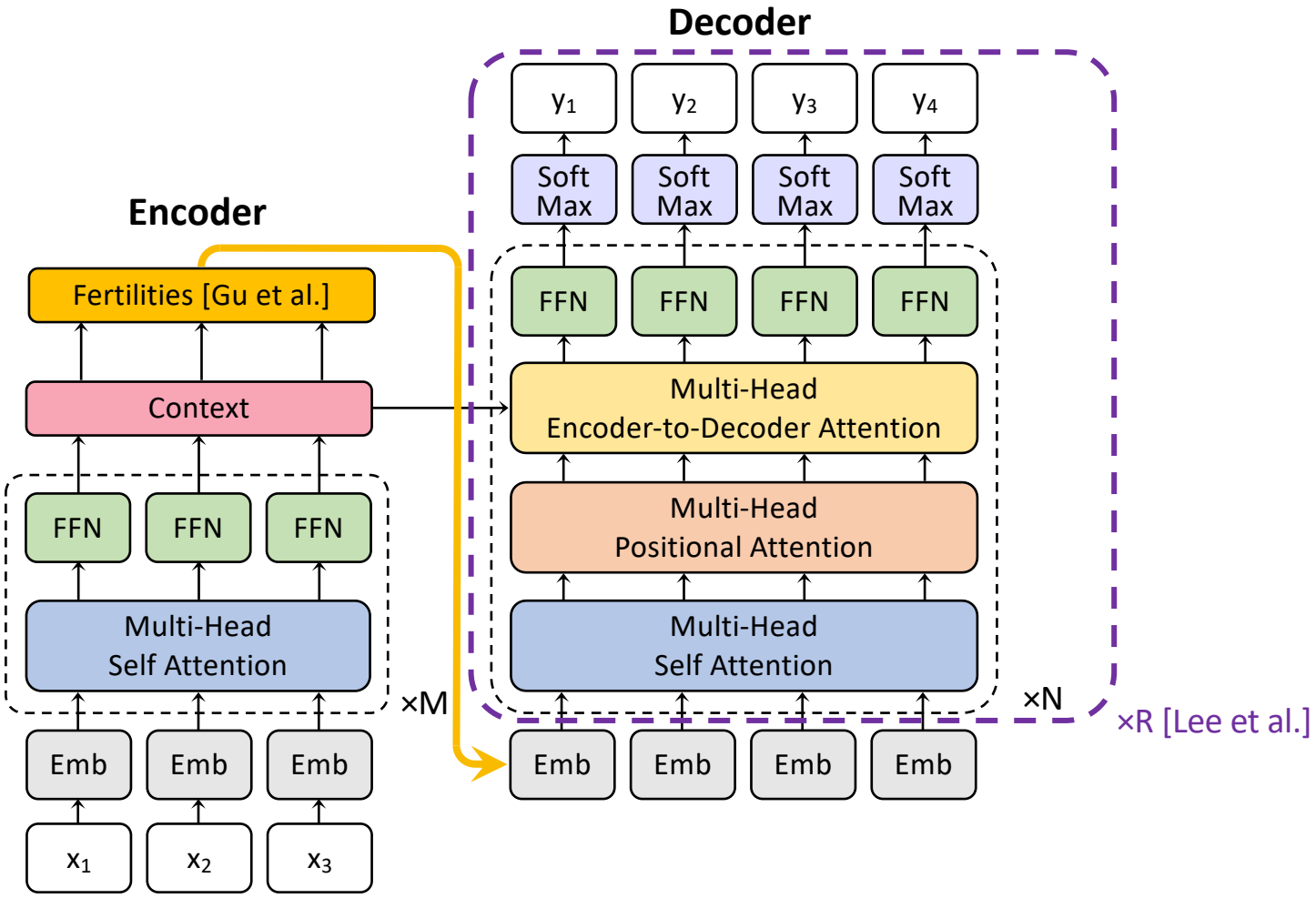
Non-autoregressive MT models





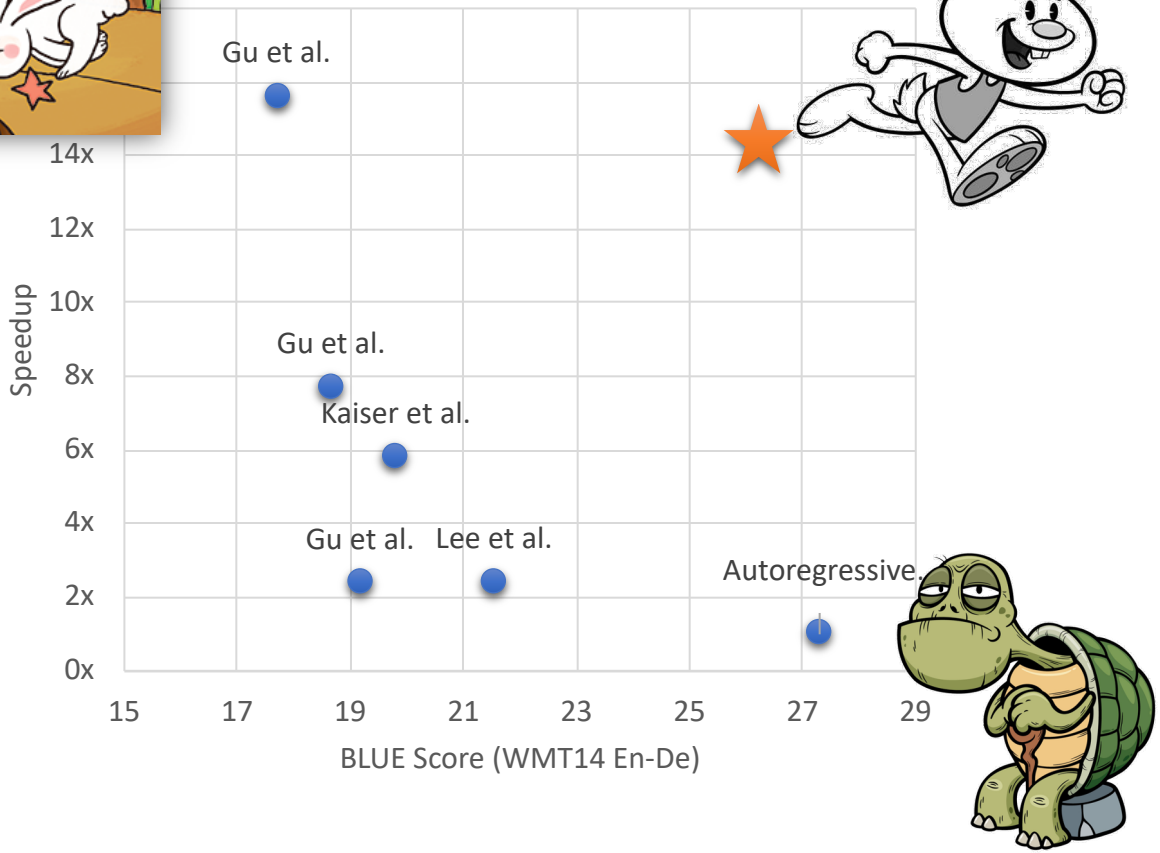


Previous works on non-autoregressive MT

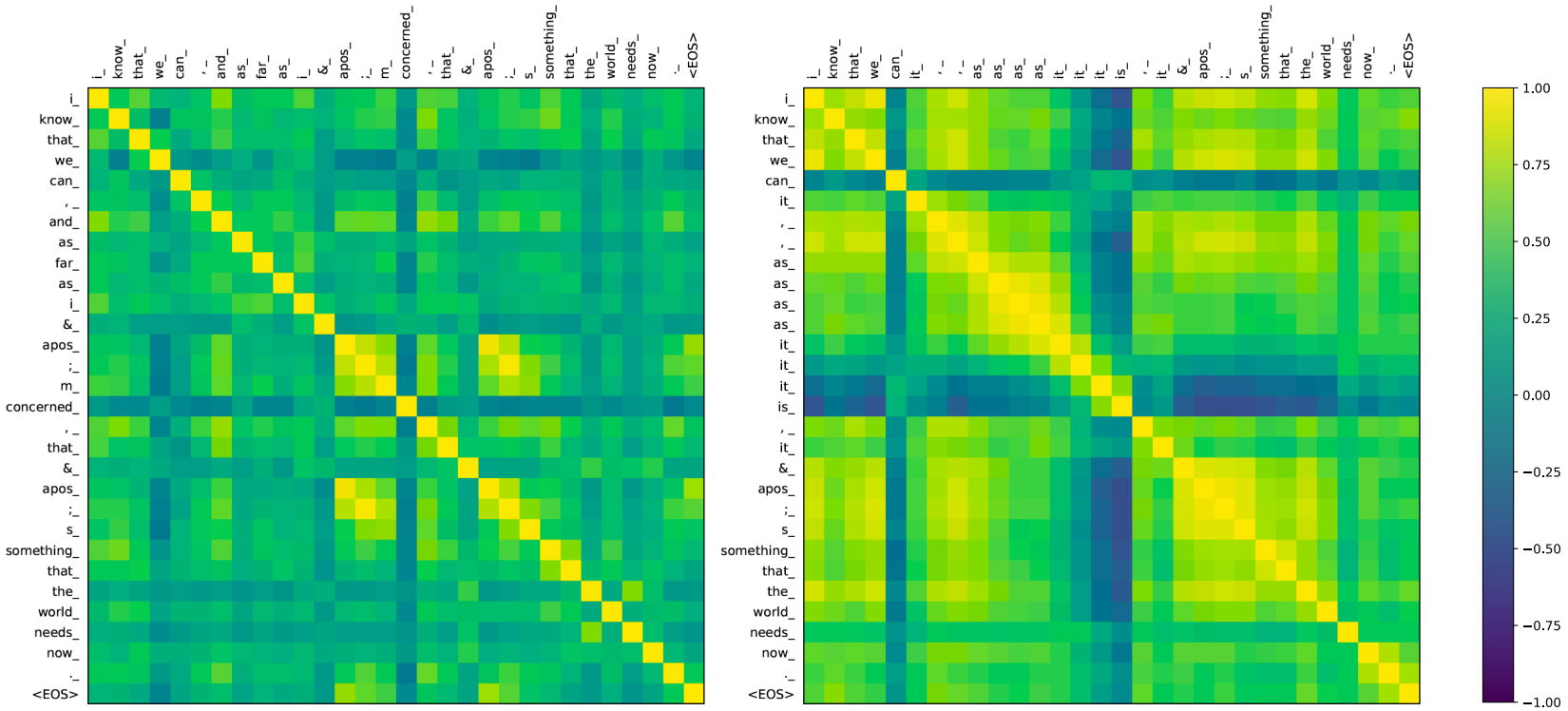


Quality-speedup trade-off





Hidden states similarity

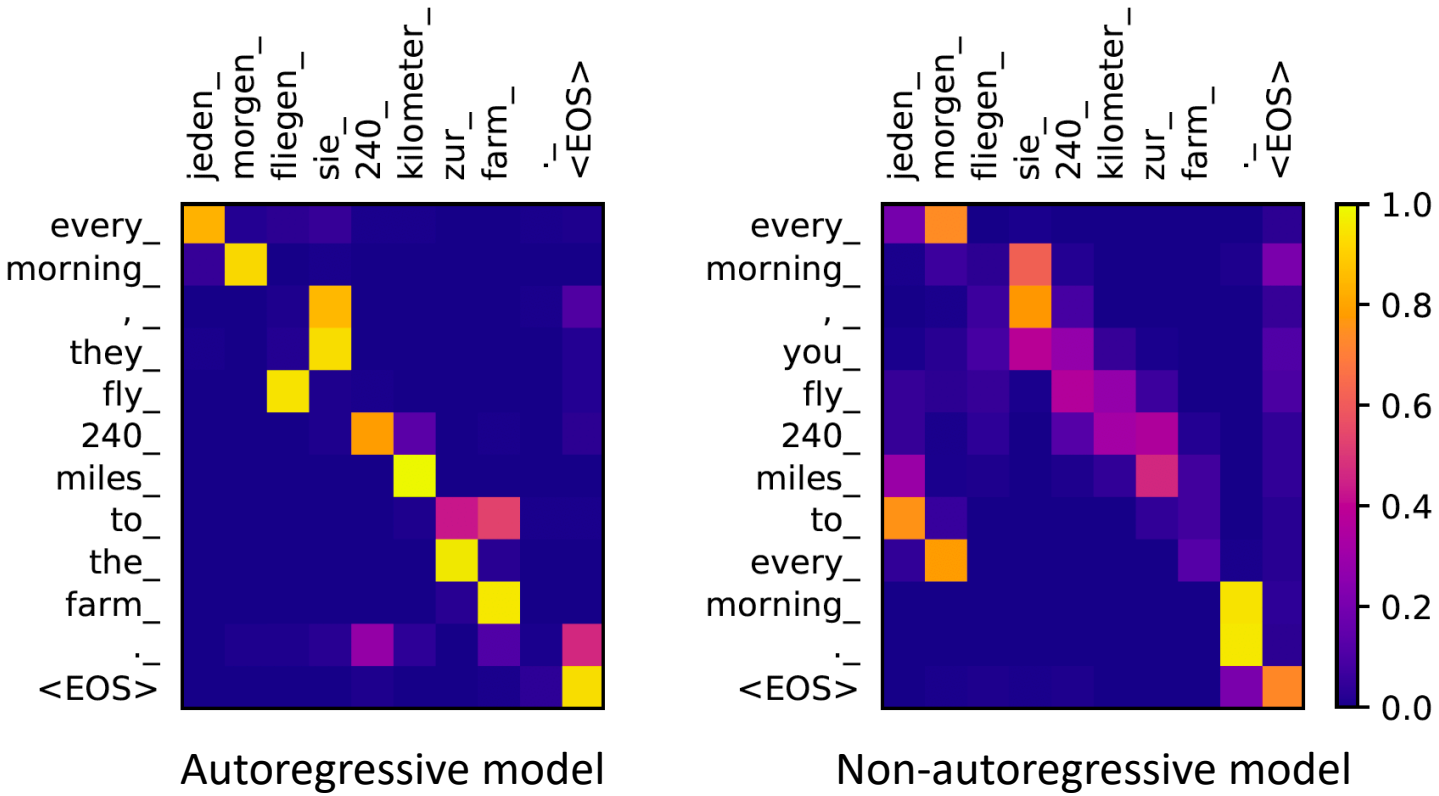


Autoregressive model

Non-autoregressive model

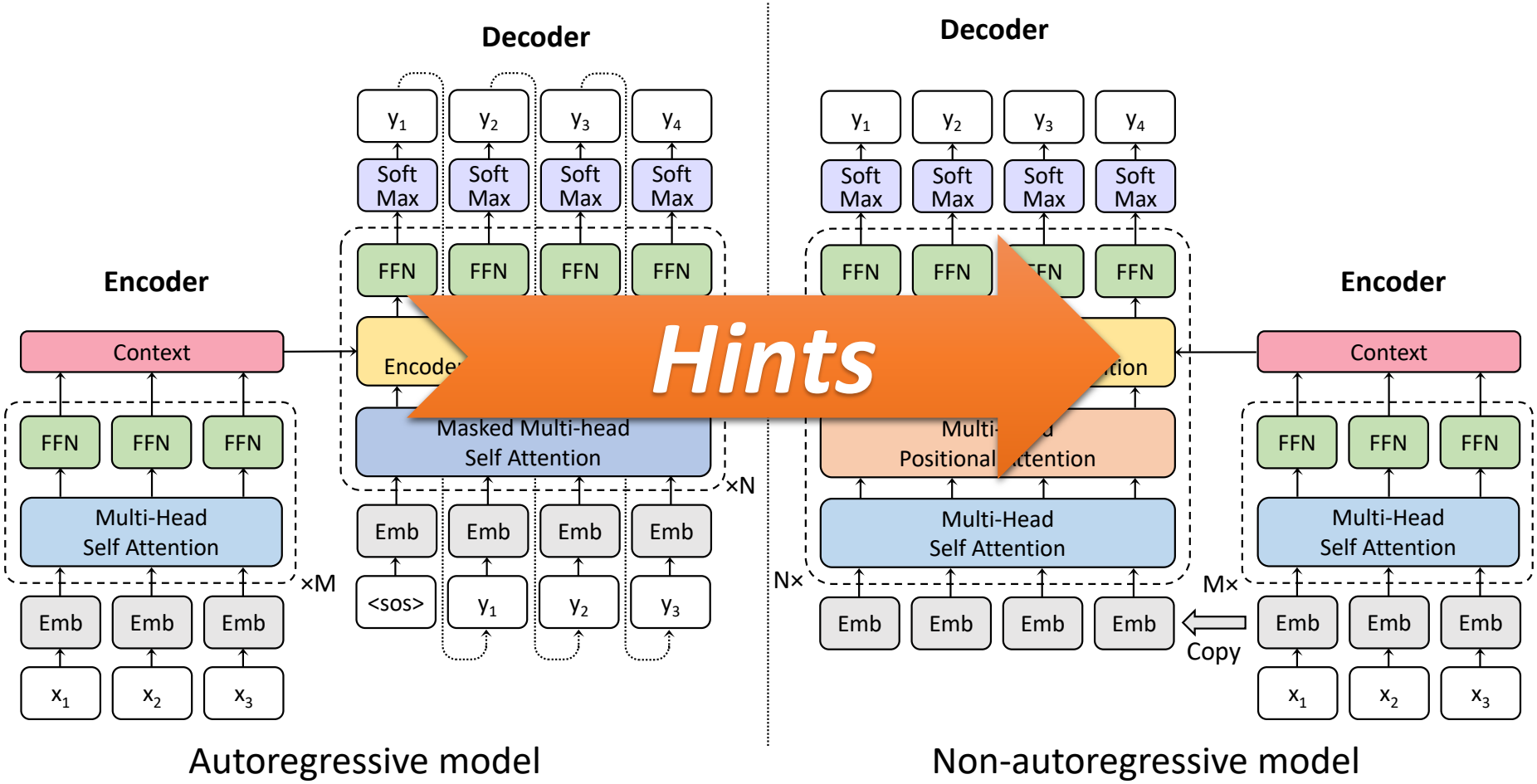
Hidden states cosine-similarity of a sampled sentence in IWSLT14 De-En.

Attention distribution

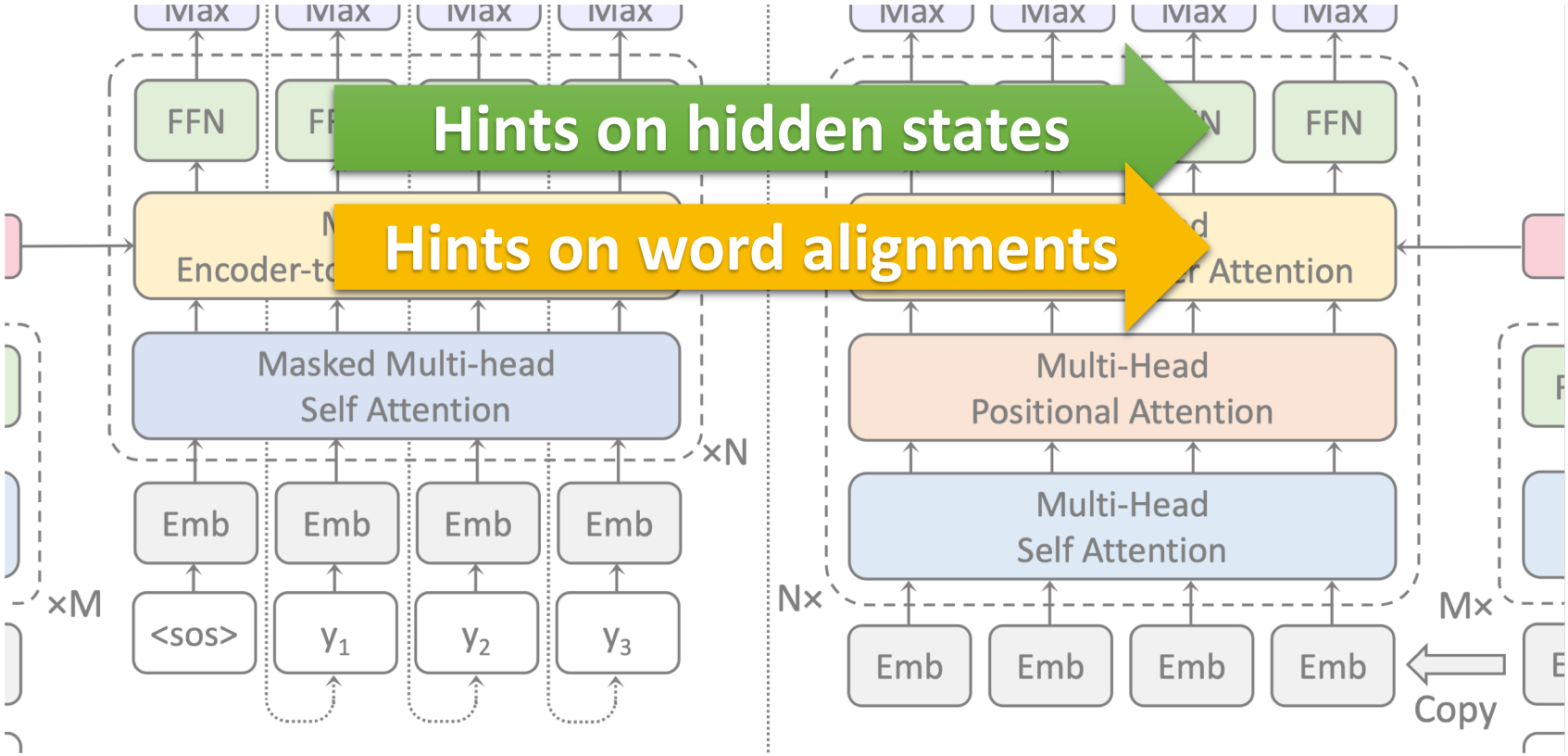


Encoder-to-decoder attention distribution of an informative head of a sampled sentence from IWSLT14 De-En.

Hint-based training
from autoregressive teacher to non-autoregressive student



Hint-based training
from autoregressive teacher to non-autoregressive student



Hints on hidden states

- Directly regression fails because of the discrepancy between two models
- We penalize hidden states that are highly similar:

$$\mathcal{L}_{hid} = \frac{2}{(T_y - 1)T_y N} \sum_{s=1}^{T_y - 1} \sum_{t=s+1}^{T_y} \sum_{l=1}^{N} \phi(d_{st}, d_{tr})$$

  - d_{st} and d_{tr} are cosine similarities s-th and t-th of hidden states at layer l of the student and teacher models
  - ϕ is a fixed function only penalizes when the student's hidden states are similar while the teacher's not

Hints on word alignments

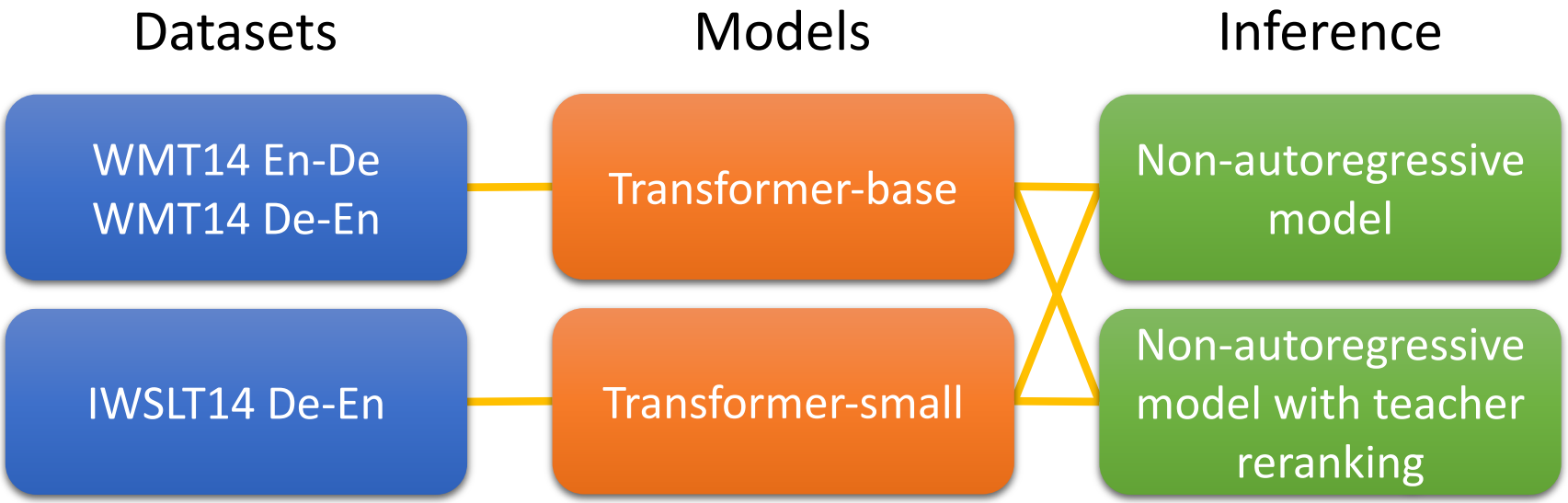
- We minimize the KL-divergence between the per-head encoder-to-decoder attention distribution of the student and teacher models

$$\mathcal{L}_{align} = \frac{1}{T_y NH} \sum_{t=1}^{T_y} \sum_{l=s+1}^{N} \sum_{h=1}^{H} D_{KL}(a_{t,l,h}^{tr} \parallel a_{t,l,h}^{st})$$

Total Loss

$$\mathcal{L} = \mathcal{L}_{nll} + \lambda \mathcal{L}_{hid} + \mu \mathcal{L}_{align}$$

Experimental settings

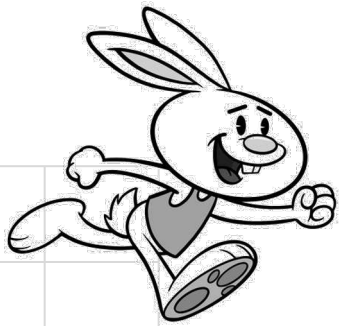


Experimental results

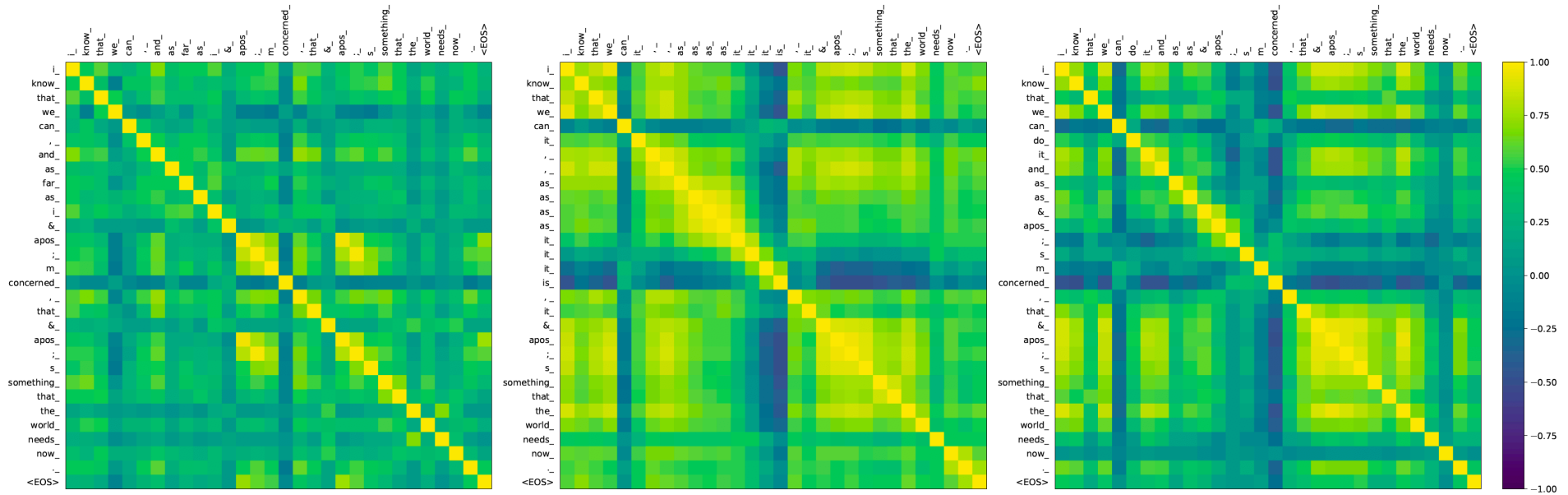
Models	WMT14 En-De	WMT14 De-En	IWSLT14 De-En	Latency	Speedup
Autoregressive models					
LSTM-based S2S (Wu et al., 2016; Bahdanau et al., 2016)	24.60	/	28.53	/	/
ConvS2S (Gehring et al., 2017; Edunov et al., 2017)	26.43	/	32.84	/	/
Transformer (Vaswani et al., 2017)	27.30	31.29	33.26	784 ms‡	1.00×
Non-autoregressive models					
FT (Gu et al., 2017)	17.69	20.62	/	39 ms†	15.6×
FT (rescoring 10 candidates)	18.66	22.41	/	79 ms†	7.68×
FT (rescoring 100 candidates)	19.17	23.20	/	257 ms†	2.36×
IR (Lee et al., 2018, adaptive refinement steps)	21.54	25.43	/	/	2.39×
LT (Kaiser et al., 2018)	19.8	/	/	105 ms†	5.78×
LT (rescoring 10 candidates)	21.0	/	/	/	/
LT (rescoring 100 candidates)	22.5	/	/	/	/
NART w/ hints	**21.11**	**25.24**	**25.55**	**26 ms**‡	**30.2×**
NART w/ hints ($B = 4$, 9 candidates)	**25.20**	**29.52**	**28.80**	**44 ms**‡	**17.8×**

Quality-speedup trade-off





Hidden states similarity



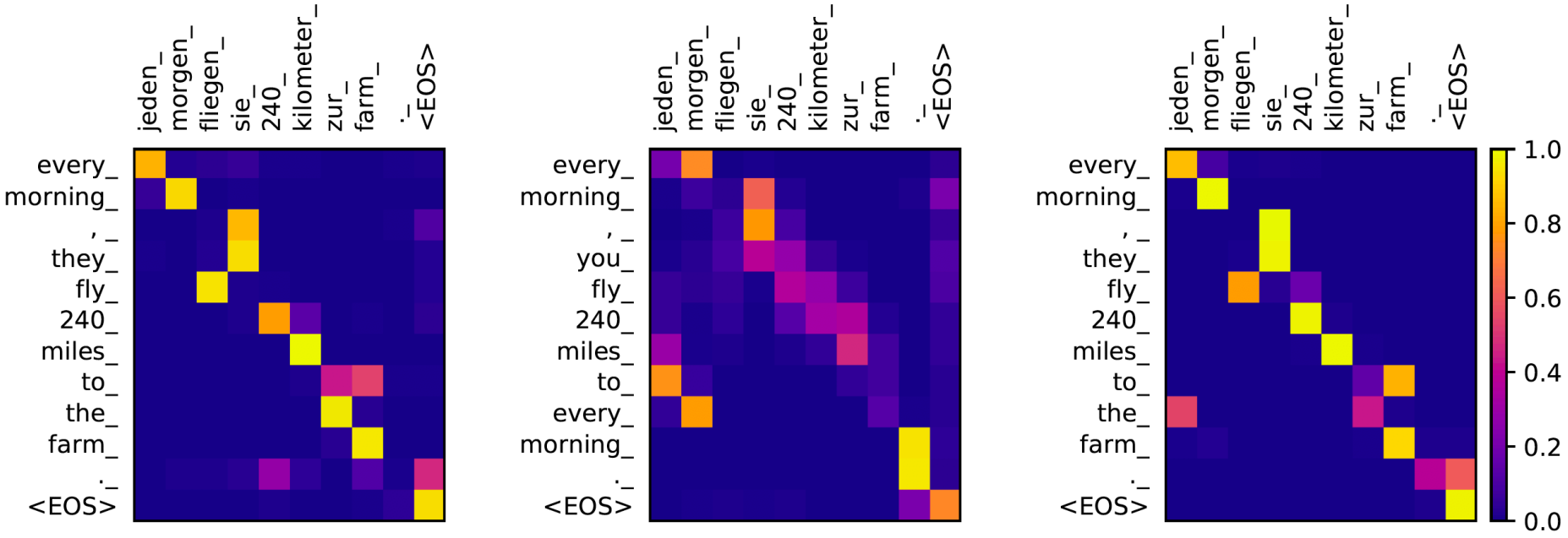
Autoregressive model

Non-autoregressive model without hints

Non-autoregressive model with hints

Hidden states cosine-similarity of a sampled sentence in IWSLT14 De-En.

Attention distribution



Autoregressive model

Non-autoregressive model without hints

Non-autoregressive model with hints

Encoder-to-decoder attention distribution of an informative head of a sampled sentence from IWSLT14 De-En.

Ablation studies

Ablation studies on IWSLT14 De-En. Results are BLEU scores without teacher rescoring.

Model	$\mathcal{L}_{nll}$	$\mathcal{L}_{nll} + \mathcal{L}_{align}$	$\mathcal{L}_{nll} + \mathcal{L}_{align} + \mathcal{L}_{hid}$
BLEU	23.08	24.76	**25.55**
Long-sentence BLEU	17.48	19.24	**20.63**

Summary

Instead of adding new modules that can slow down the model, we proposed a method to leverage the ***hints*** from the autoregressive model to help the training of the non-autoregressive model.

Thanks!

Q&A

Zhuohan Li (zhuohan@cs.berkeley.edu)