State-of-the-art Neural Translation

Carlos Ventura s202450, Jakub Reha s184478, Silas Brack s174433
DTU Compute, Technical University of Denmark
Deep Learning 02456

Introduction

Machine translation German to English using three different Sequence2Sequence models with encoder-decoder architectures with two different datasets. We analyze the influence of attention. For every architecture word-level embeddings were used.

Theory

$$p_{\theta_{\text{enc}},\theta_{\text{dec}}}\left(\mathbf{Y}_{1:m} \mid \mathbf{X}_{1:n}\right)$$

$$f_{\theta_{\text{enc}}} : \mathbf{X}_{1:n} \to \overline{\mathbf{X}}_{1:n}$$

$$p_{\theta_{dec}}\left(\mathbf{Y}_{1:m} \mid \overline{\mathbf{X}}_{1:n}\right) = \prod_{i=1}^{m} p_{\theta_{dec}}\left(\mathbf{y}_i \mid \mathbf{Y}_{0:i-1}, \overline{\mathbf{X}}_{1:n}\right)$$

Above it can be seen the probabilistic definition of these models, where X is the input sequence, $\bar{X}$ the context vector and Y the target sequence.

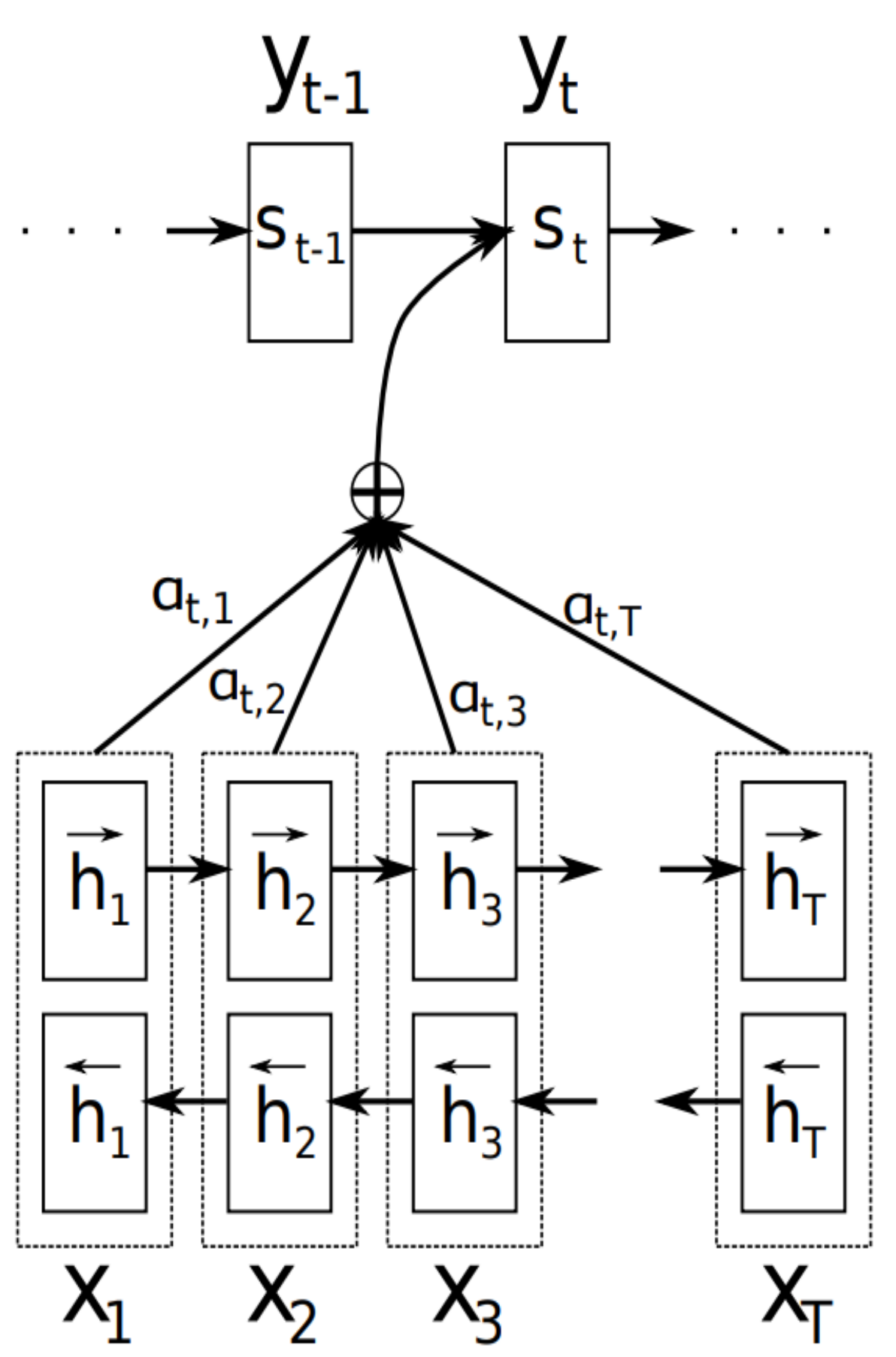
Future work

Knowledge distillation

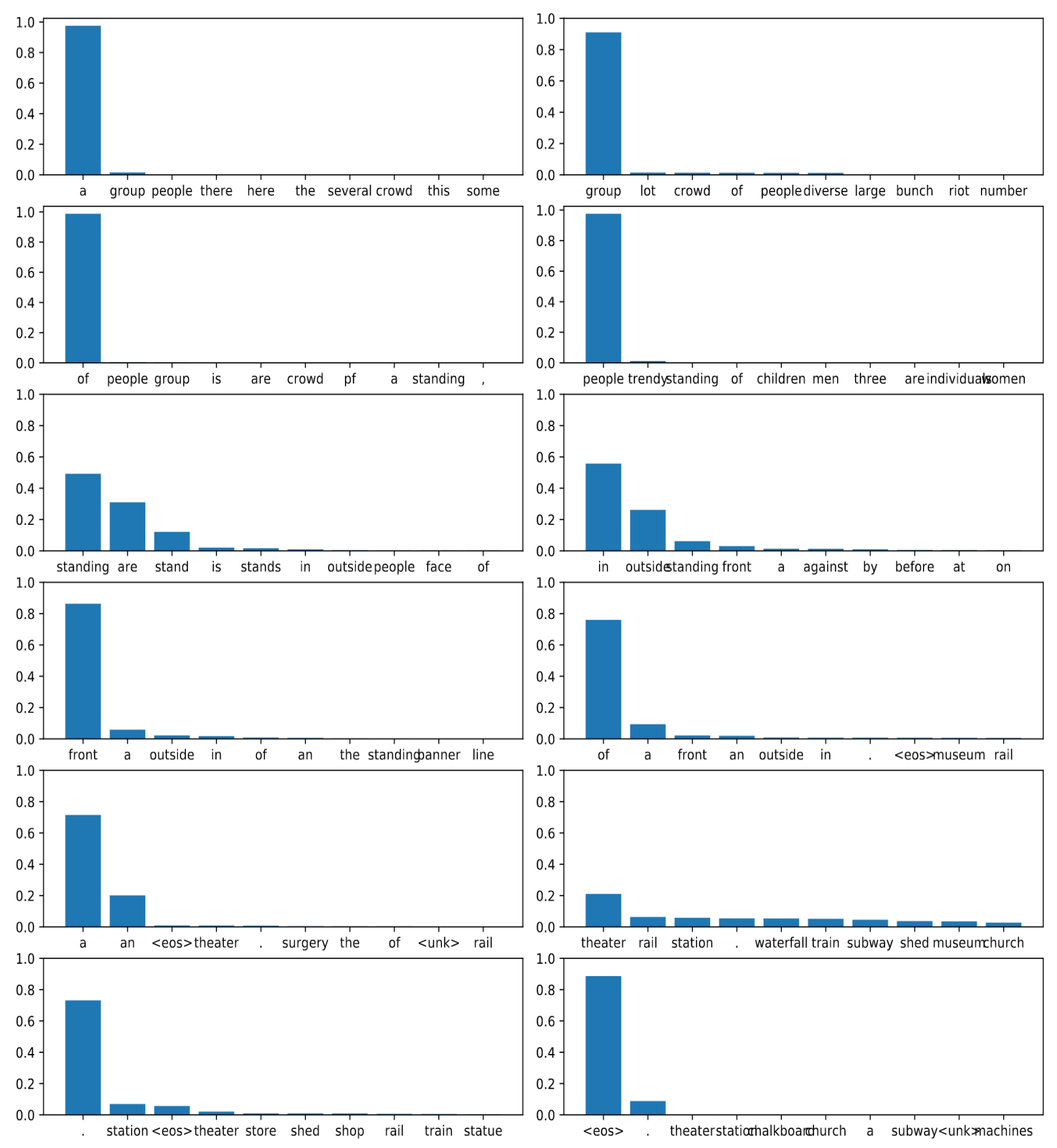
$$E(\mathbf{x}|t) = -t^2 \sum_i \hat{y}_i(\mathbf{x}|t) \log y_i(\mathbf{x}|t) - \sum_i \bar{y}_i \log y_i(\mathbf{x}|1)$$

$$y_i(\mathbf{x}|t) = \frac{e^{\frac{z_i(\mathbf{x})}{t}}}{\sum_j e^{\frac{z_j(\mathbf{x})}{t}}}$$

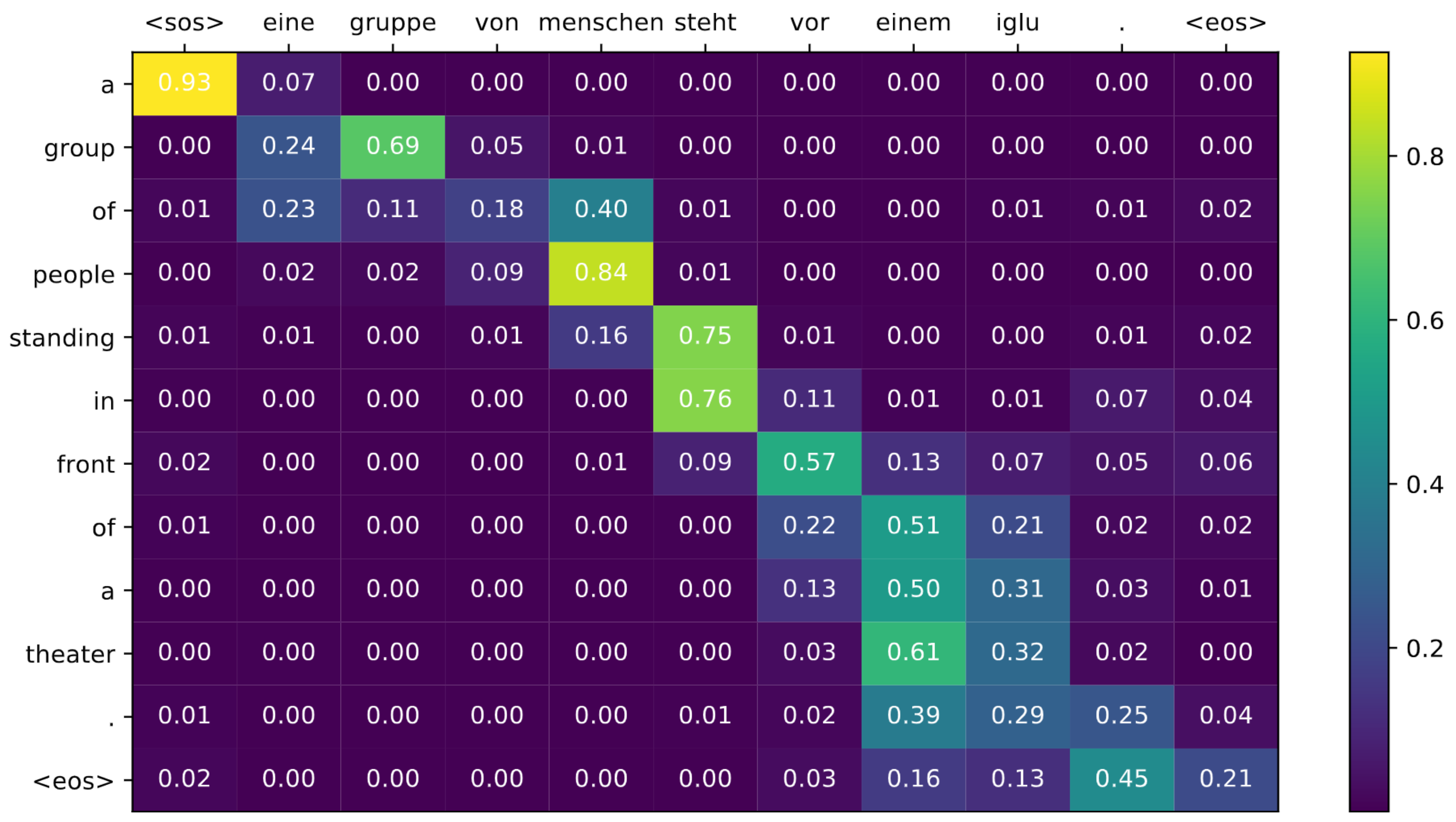
Seq2seq GRU w attn [3]



Seq2seq GRU w attn Model

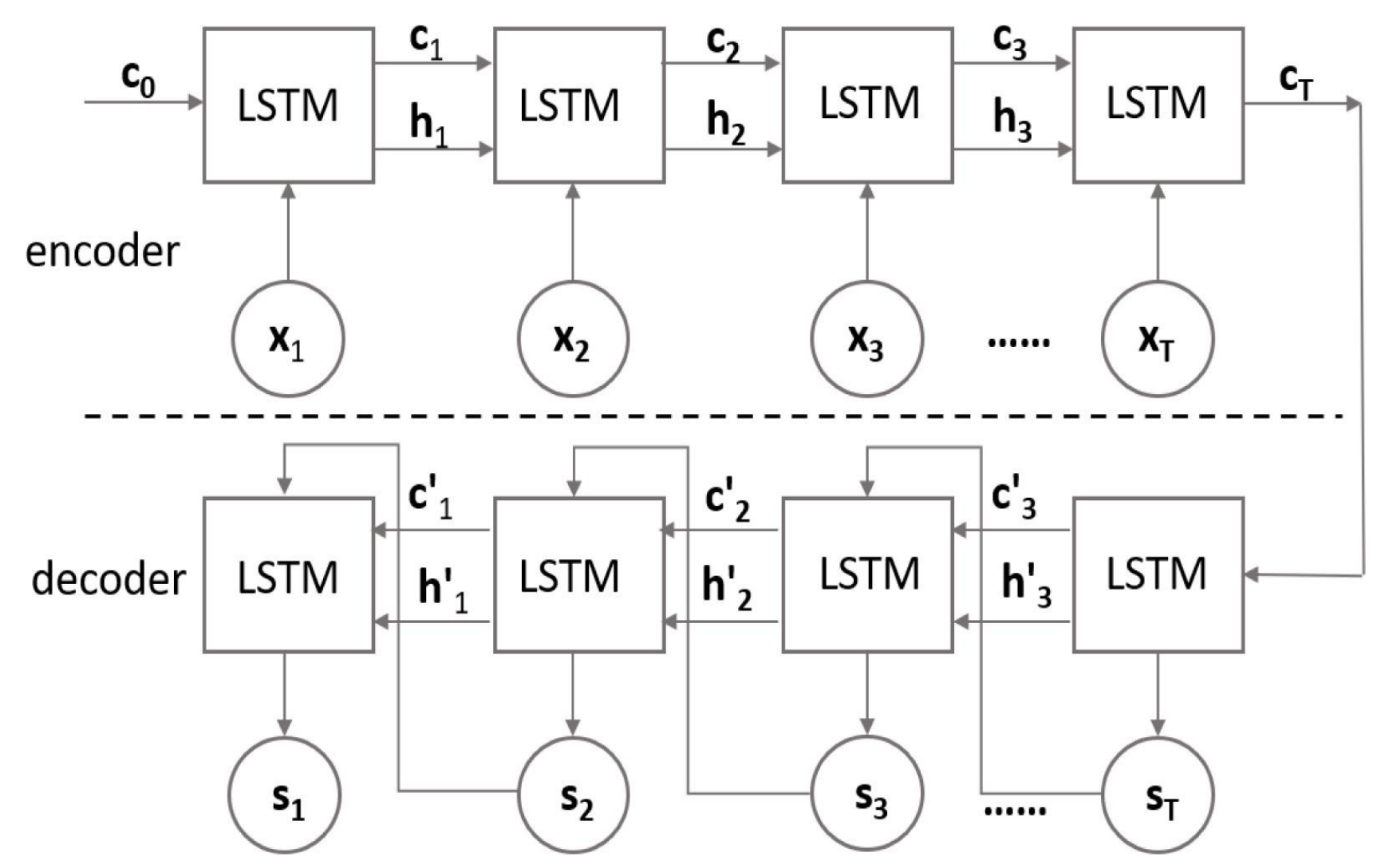


Conditional probability distribution



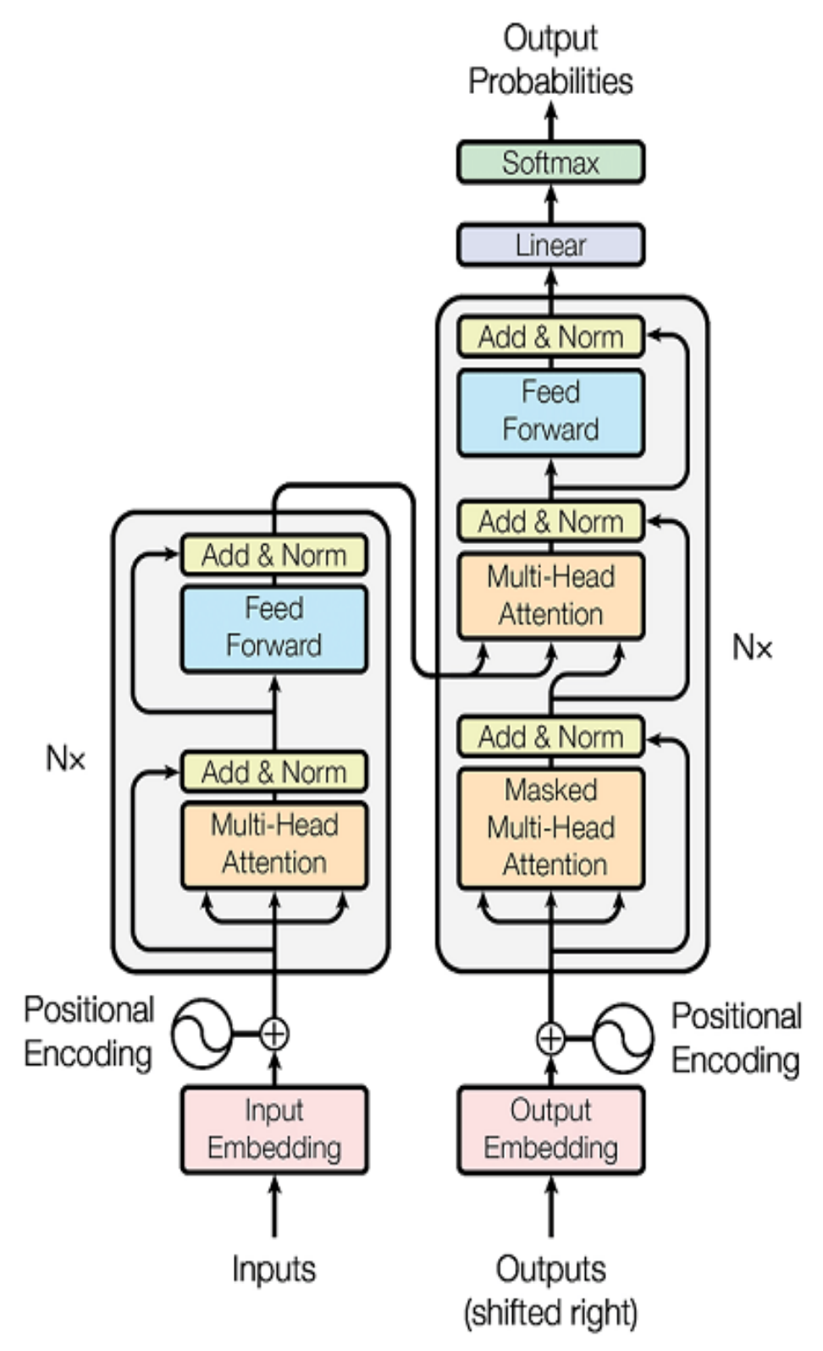
Attention heat map, GRU

Seq2seq LSTM

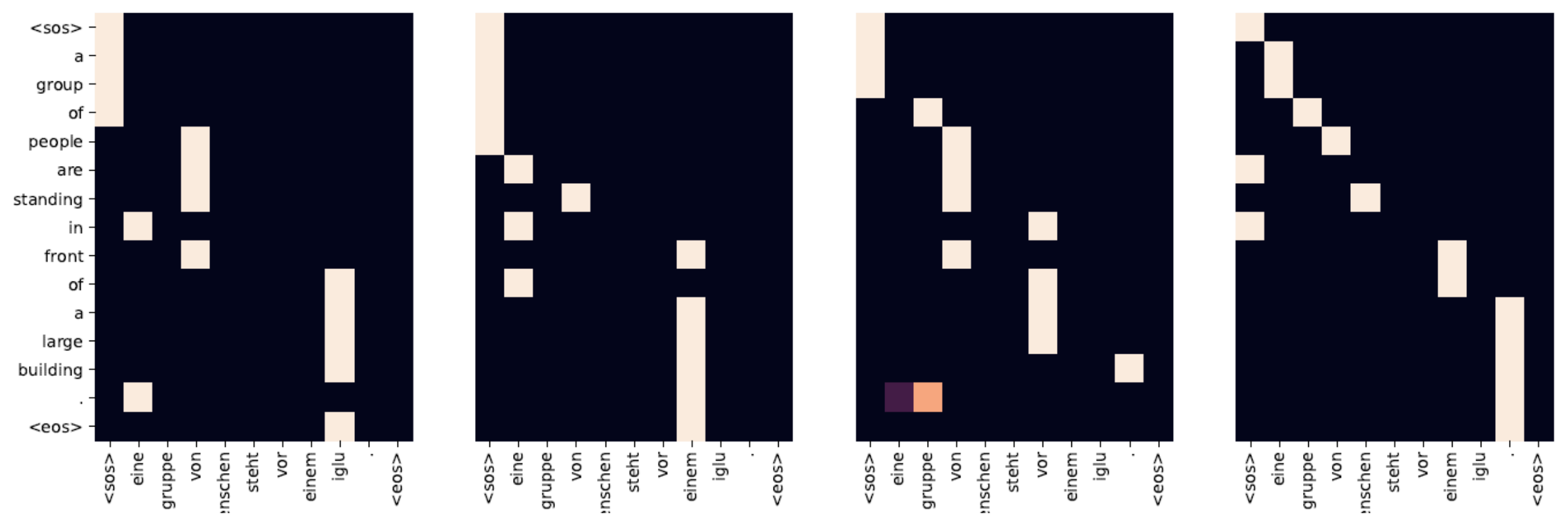


LSTM Model

Transformer [1][2]



Transformer Model



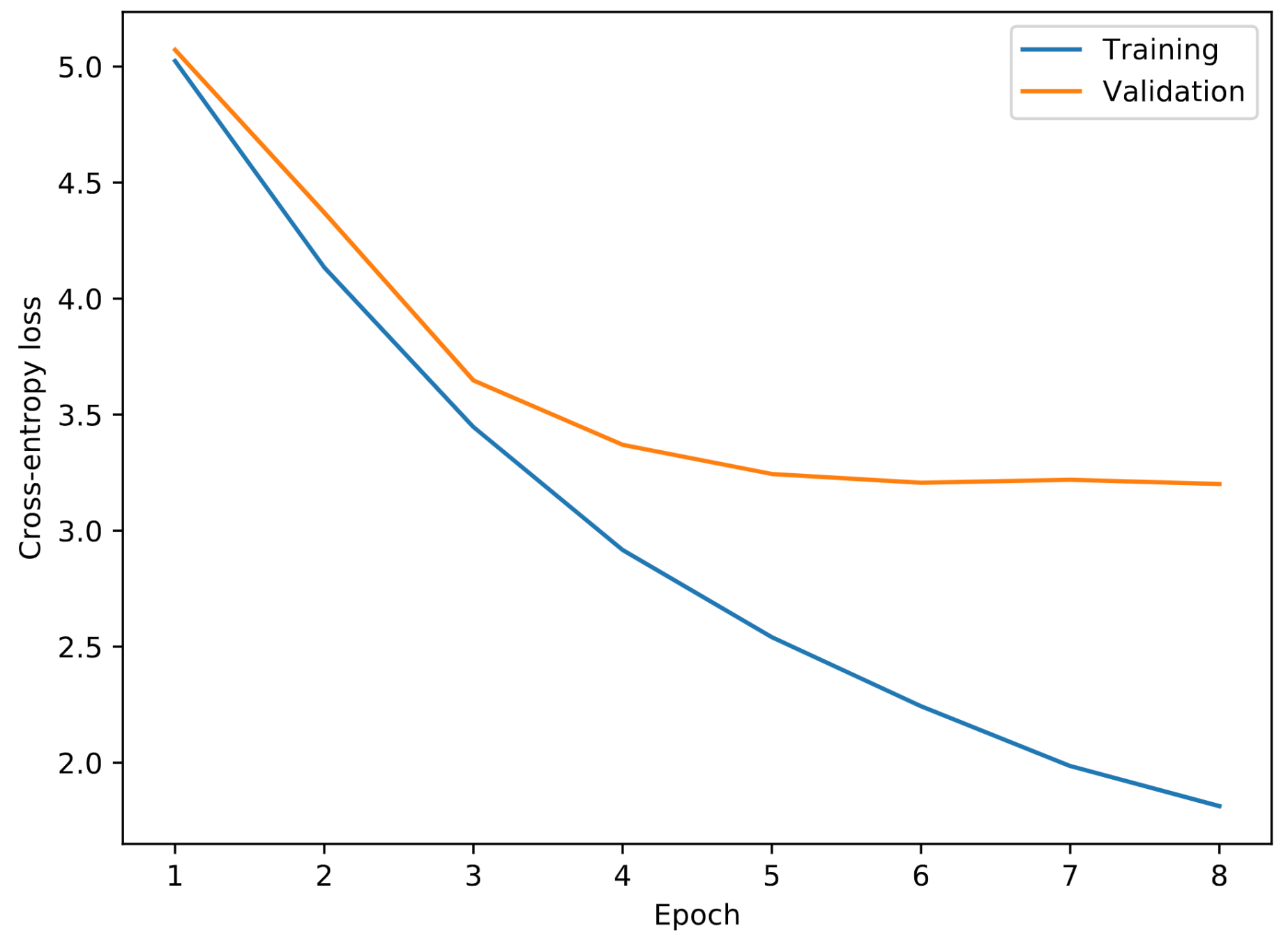
Attention heat map, Transformer decoder layer 6 (source)

Translation samples

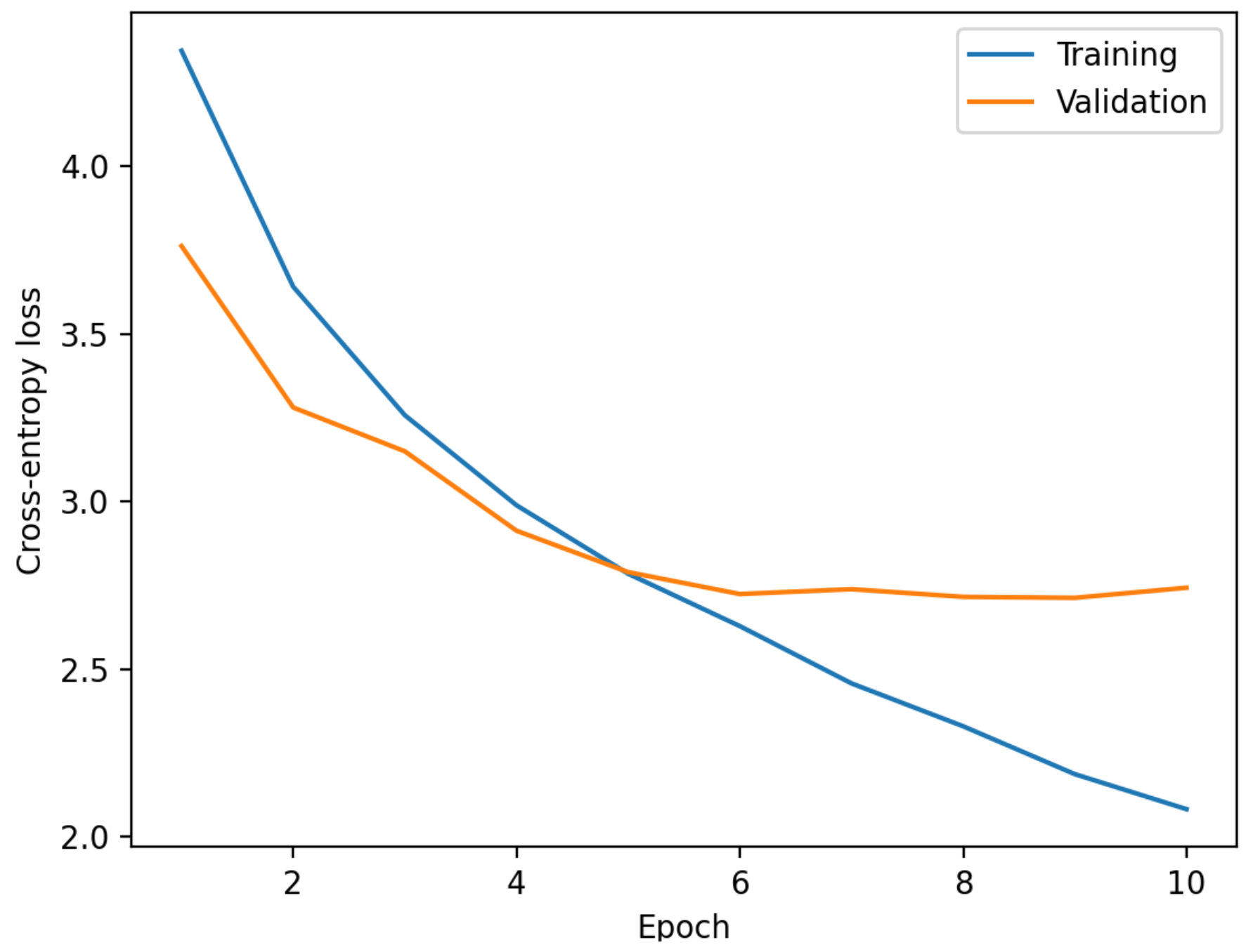
Original: eine gruppe von menschen steht vor einem iglu .
Translation: a group of people stands in front of an igloo .
Seq2Seq LSTM: a group of people standing in front of a <unk> booth .
Seq2Seq GRU w/ Attn: a group of people standing in front of a theater .
Transformer: a group of people are standing in front of a large building .
Bert2Bert: a group of people standing in front of an igloo.

Original: ein mann mit kariertem hut in einer schwarzen jacke und einer schwarz-weiß gestreiften hose spielt auf einer bühne mit einem sänger und einem weiteren gitarristen im hintergrund auf einer e-gitarre.
Translation: a man in a black jacket and checkered hat wearing black and white striped pants plays an electric guitar on a stage with a singer and another guitar player in the background .
Seq2Seq LSTM: a man in a black hat and black shirt plays a a with a a a a a a a a in a a in a background .
Seq2Seq GRU w/ Attn: a man in a plaid hat , jacket and black striped striped striped striped striped shirt , playing a guitar with a guitar with a guitar with a guitar in a treadmill.
Transformer: a man in a white shirt and jeans is playing a guitar on a stage .
Bert2Bert: a man with a checkered hat in a black jacket and a black-and-white striped pants plays on a balcony with a singer and another guitarist behind on an e-guitar

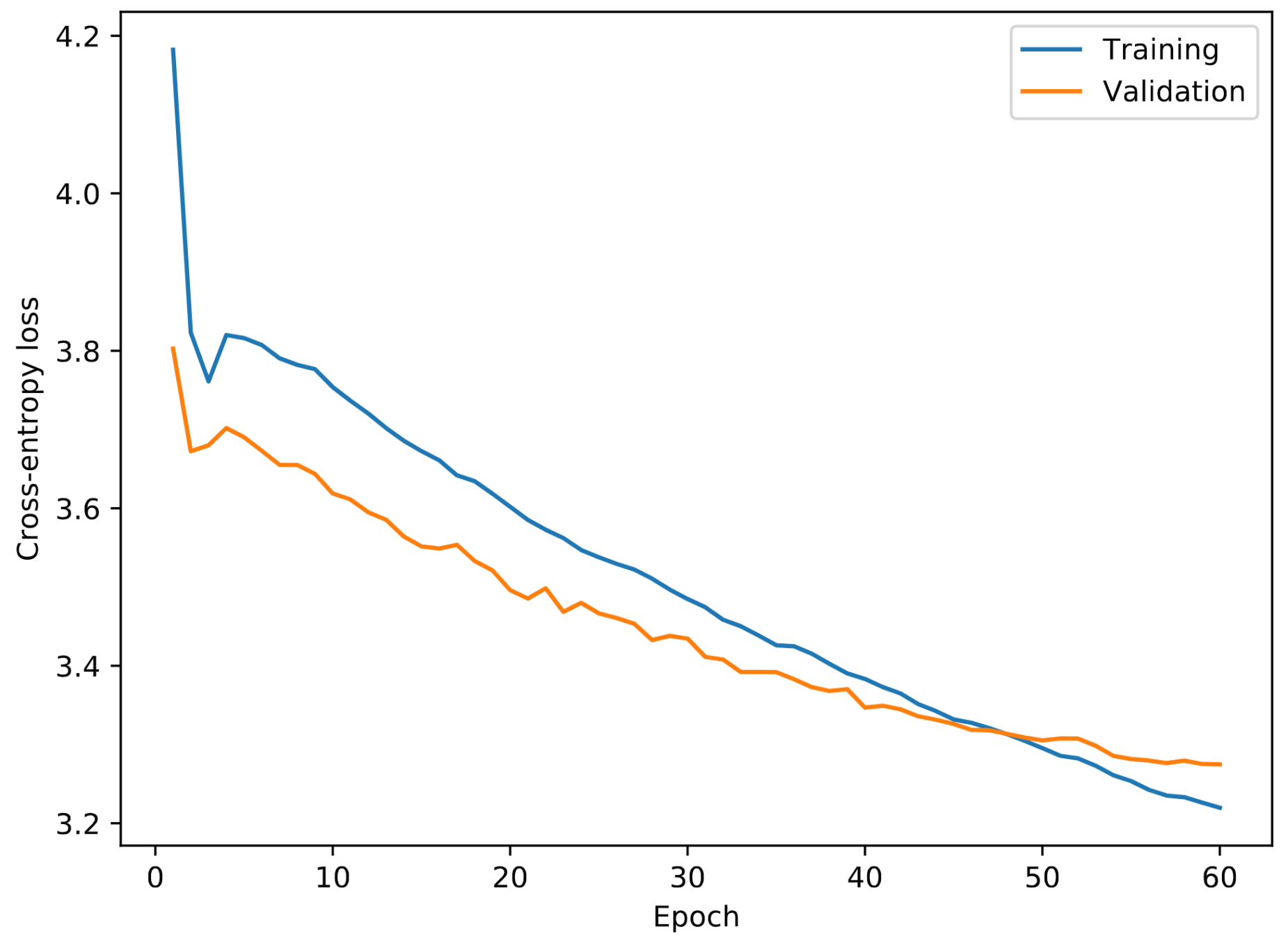
Seq2seq GRU w/ attn



Seq2seq LSTM



Transformer



Comparisons

	Seq2seq LSTM	Seq2seq GRU	Transformer	Bert2Bert
CE Train*	2.19	1.81	3.28	---
CE Val*	2.71	3.20	3.31	---
CE Test*	2.84	3.25	3.287	---
Perplexity*	17.116	25.774	26.751	---
BLEU*	21.7	29.3	10.5	27.7
Params	~35M	~20M	~50M	~770M
BLEU WMT14	0.13	0.21	2.8	17

*evaluated on Multi30k unless stated otherwise

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," CoRR, vol. abs/1706.03762, 2017.
[2] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush, "Opennmt: Open-source toolkit for neural machine translation," in Proc. ACL, 2017.
[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," 2016.