Current State of Research in Neural Machine Translation

Christopher Brix

christopher.brix@rwth-aachen.de

January 14th, 2020

www.christopher-brix.de

Outline

About Me

Neural Machine Translation (NMT)

Old and New Architectures

2D-LSTMs

Other Research

Data Cleaning/Fairness

Document Level Translations

Sparsity

About Me

Education:

- 2014 - 2018: B.Sc. Computer Science, RWTH Aachen University
- 2018 - 2020: M.Sc. Computer Science, RWTH Aachen University

Research:

- Student Research Assistant since 2016
  - i6: Human Language Technology and Pattern Recognition (Prof. Dr.-Ing. Ney)
  - Supervisor: Parnia Bahar
- Coauthored paper "Empirical Investigation of Optimization Algorithms in Neural Machine Translation", published in the PBML
- Coauthored paper "Towards Two-Dimensional Sequence to Sequence Model in Neural Machine Translation", published in EMNLP

Highlights:

- LxMLS, Participant & Monitor
- Google NLP Summit 2019
- Google Research Intern 2020

Neural Machine Translation (NMT)

Machine translation:

- **translate source sentence f_1^J to target hypothesis $\hat{e}_1^{\hat{I}}$**
- $\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \left\{ \Pr(e_1^I | f_1^J) \right\}$

SMT:

- **decompose using Bayes theorem**
- $\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \left\{ \Pr(f_1^J | e_1^I) \cdot \Pr(e_1^I) \right\}$

NMT:

- **directly model $\Pr(e_1^I | f_1^J)$**
- **generate words using neural network (NN)**

Encoder-Decoder

Idea: Encode, then decode [Sutskever+ 14]

- **Summarize source sentence to fixed-sized vector**
- **Decode summary to target sentence**

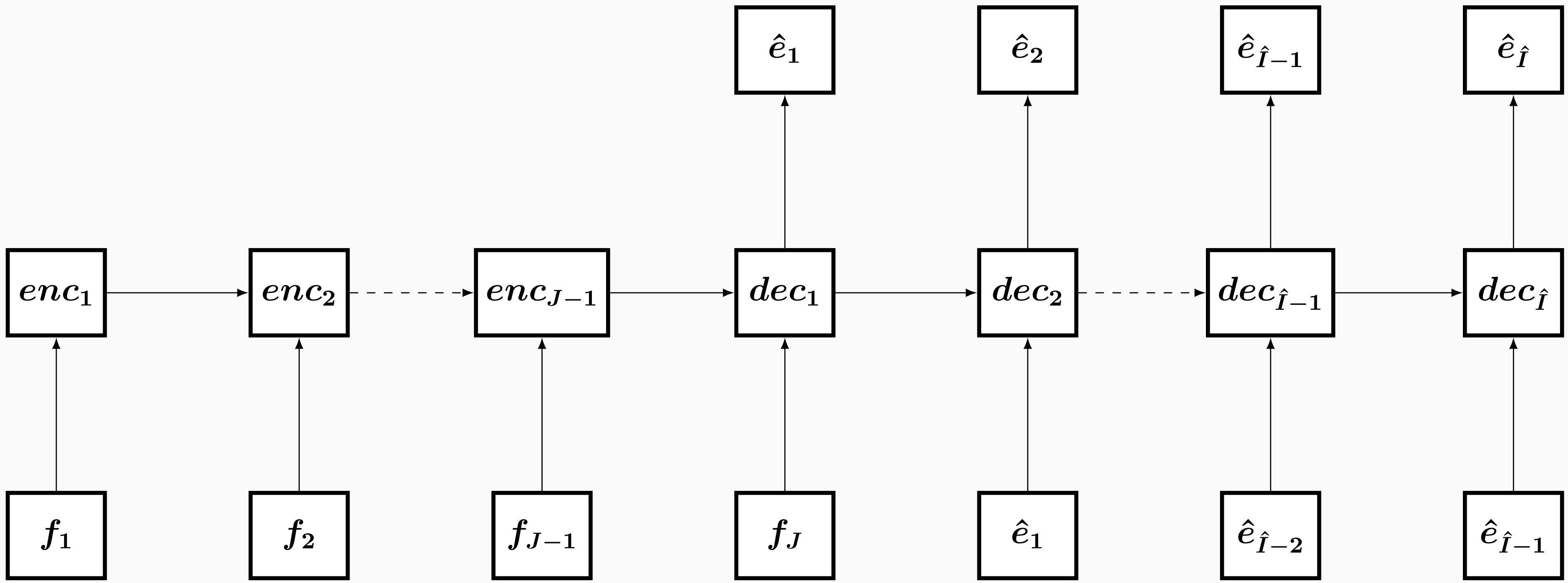


Figure: Architecture of an encoder-decoder NMT system

Attention

Idea: Focus on specific source words [Bahdanau+ 15]:

- **Summarize partial source sentence**
- **Decode summary to target word**
- **Repeat**

Online Visualization: https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention

Attention

Encoder:

$$f_1^J \rightarrow \overrightarrow{h}_j = LSTM(f_j, \overrightarrow{h}_{j-1})$$

$$f_1^J \rightarrow \overleftarrow{h}_j = LSTM(f_j, \overleftarrow{h}_{j+1})$$

$$h_j = \begin{bmatrix} \overrightarrow{h}_j \\ \overleftarrow{h}_j \end{bmatrix}$$

Attention:

$$\alpha(j|i) = A_j(s_{i-1}, h_1^J)$$

$$c_i = \sum_{j=1}^{J} \alpha(j|i) \cdot h_j$$

Decoder:

$$e_i \leftarrow t_i = Y(e_{i-1}, s_{i-1}, c_i)$$

$$s_i = LSTM([e_i, c_i], s_{i-1})$$

$$p_i(e_i = w|e_1^{i-1}, f_1^J)$$
$$= softmax(t_i)_w$$

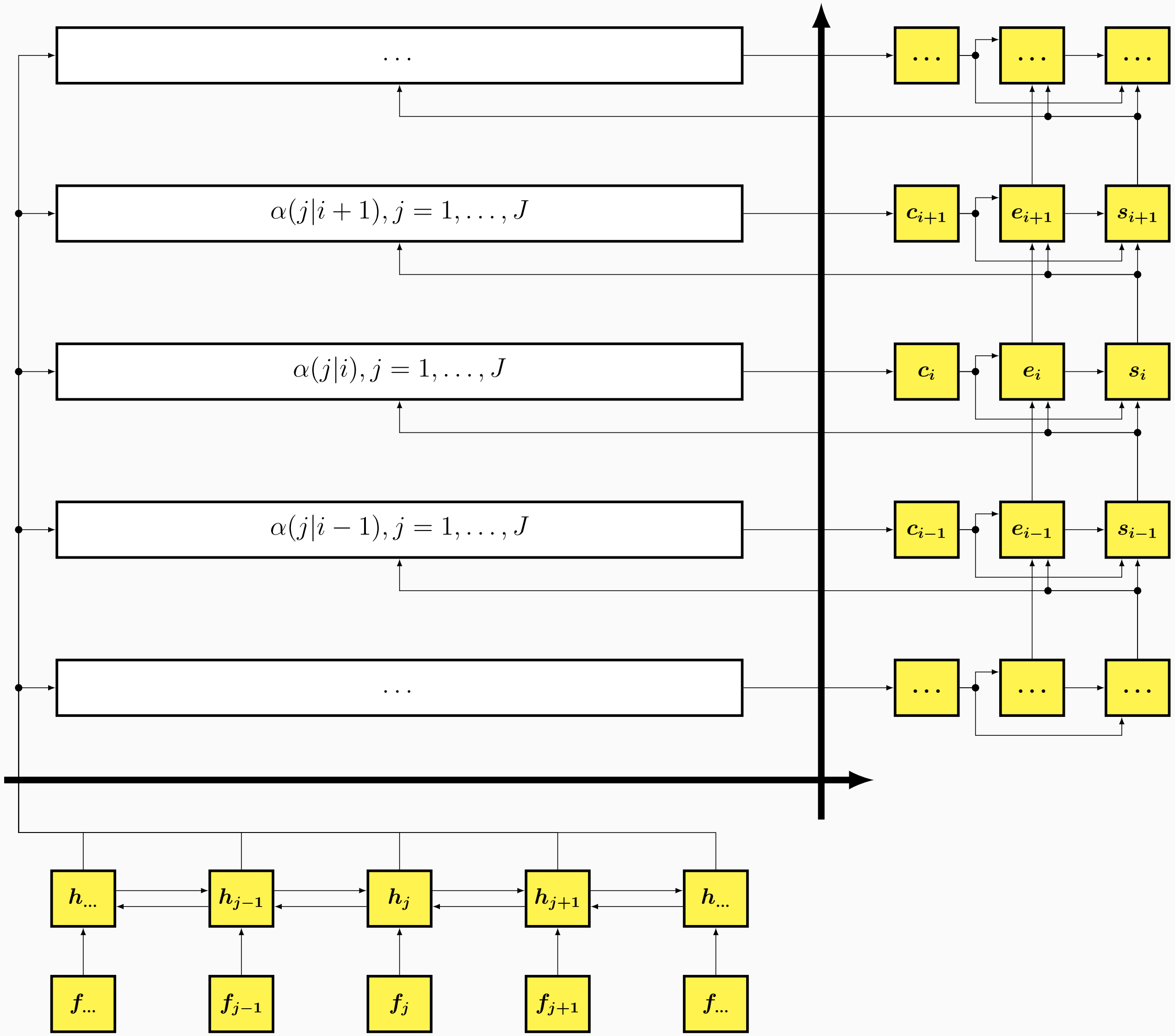


Figure: **Architecture of an attention NMT system**

Transformer

Idea: Self-Attention for high parallelizability [Vaswani+ 17]:

- **Every word computes importance of all other positions for itself**
- **Different indices are independent**
- $\alpha(j|j') = A_j(h_{j'}, h_1^J)$

$$\hat{h}_{j'} = \sum_{j=1}^{J} \alpha(j|j') \cdot h_j$$

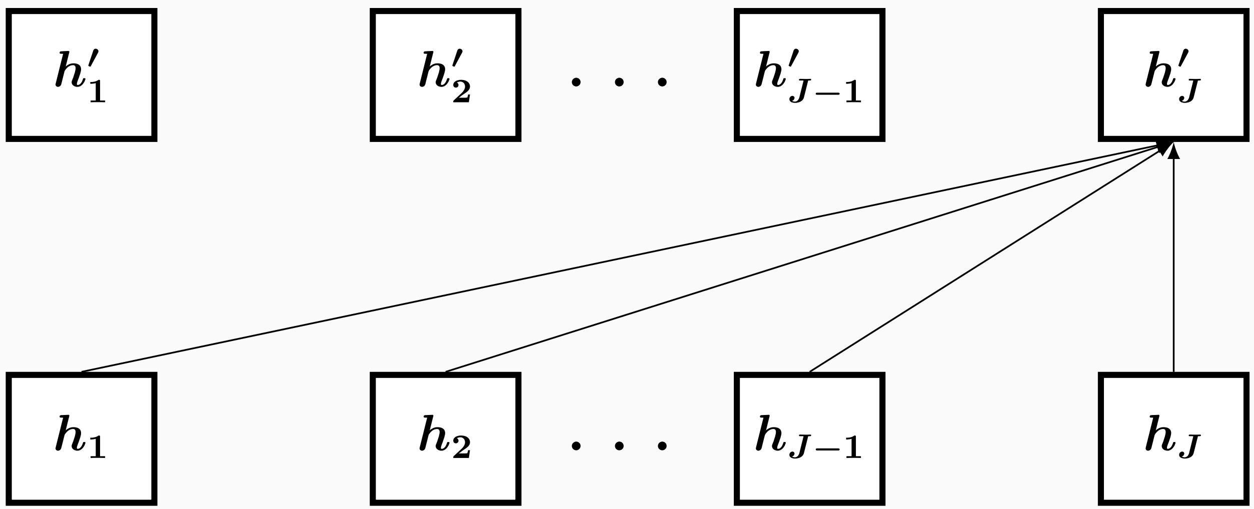


Figure: **Self-Attention**

Online Visualization: http://jalammar.github.io/illustrated-transformer

Transformer

Positional Encoding:

- **Sine/Cosine encoding of sentence index**

6 Encoder Layers:

- **Multi-Head Attention**
- **Feed Forward Layer**

6 Decoding Layers:

- **Masked Multi-Head Attention (on decoding sequence)**
- **Multi-Head Attention (on last encoder)**
- **Feed Forward Layer**

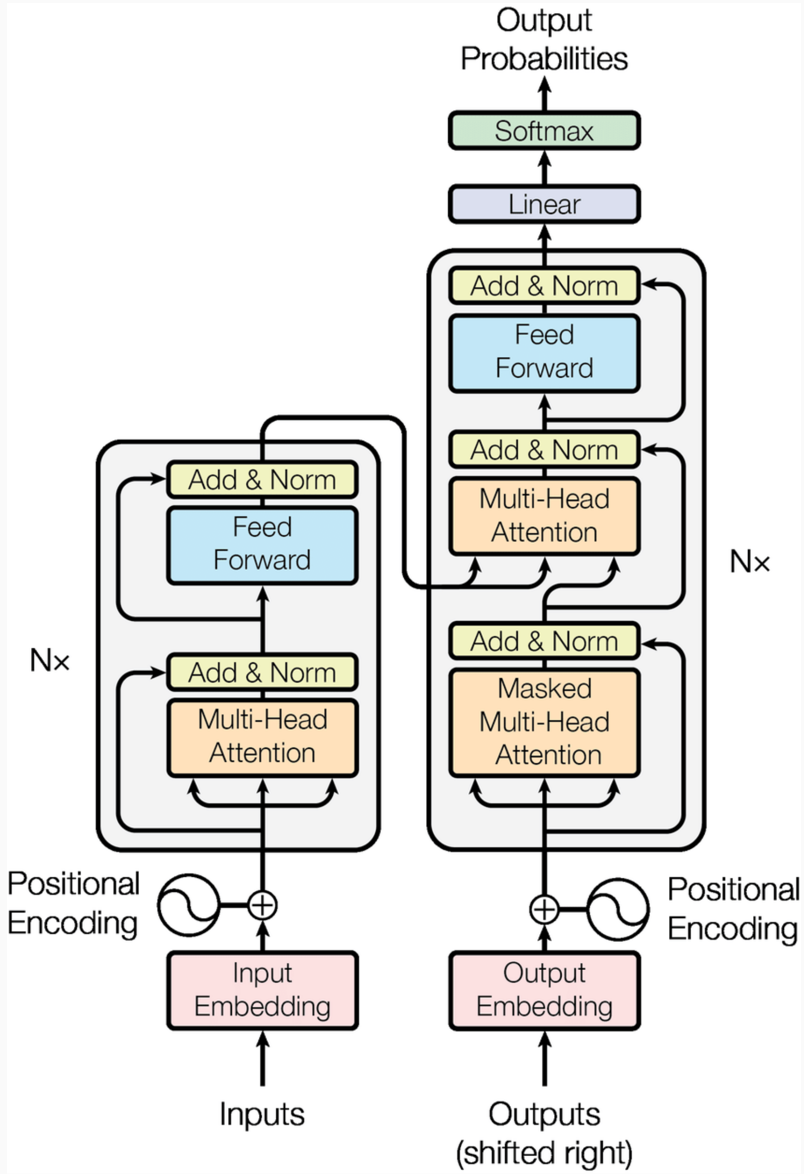


Figure: Architecture of a transformer NMT system

Two-Dimensional LSTM

- One-Dimensional LSTM processes one stream of data
- Often, data has more dimensions: eg. images
- LSTM can be extended to multiple dimensions [Graves+ 07]

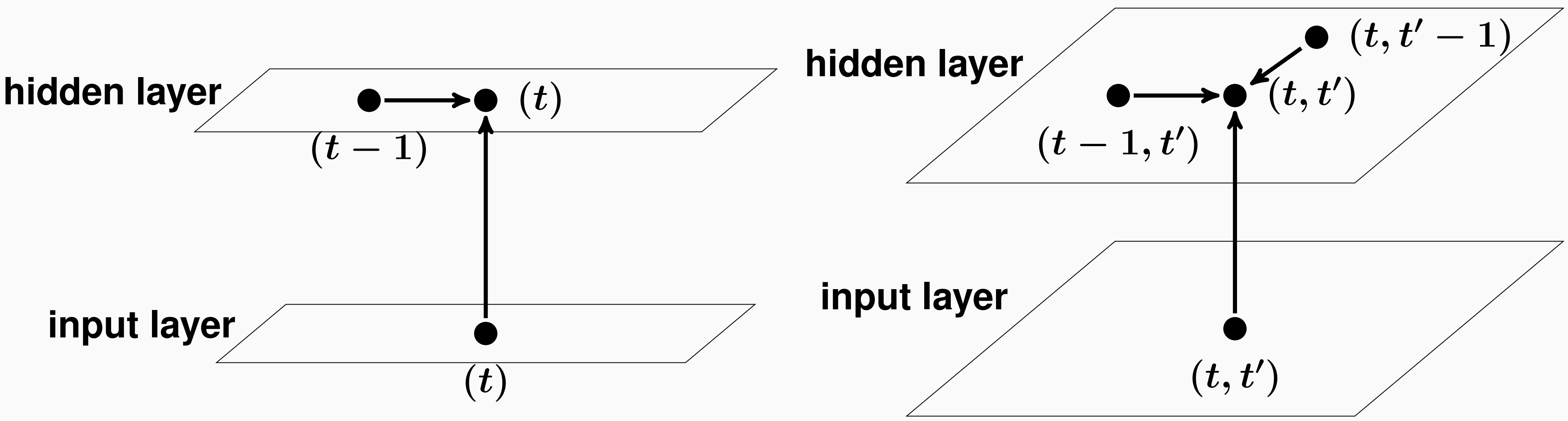


Figure: Extension of LSTM to two dimensions

Two-Dimensional LSTM

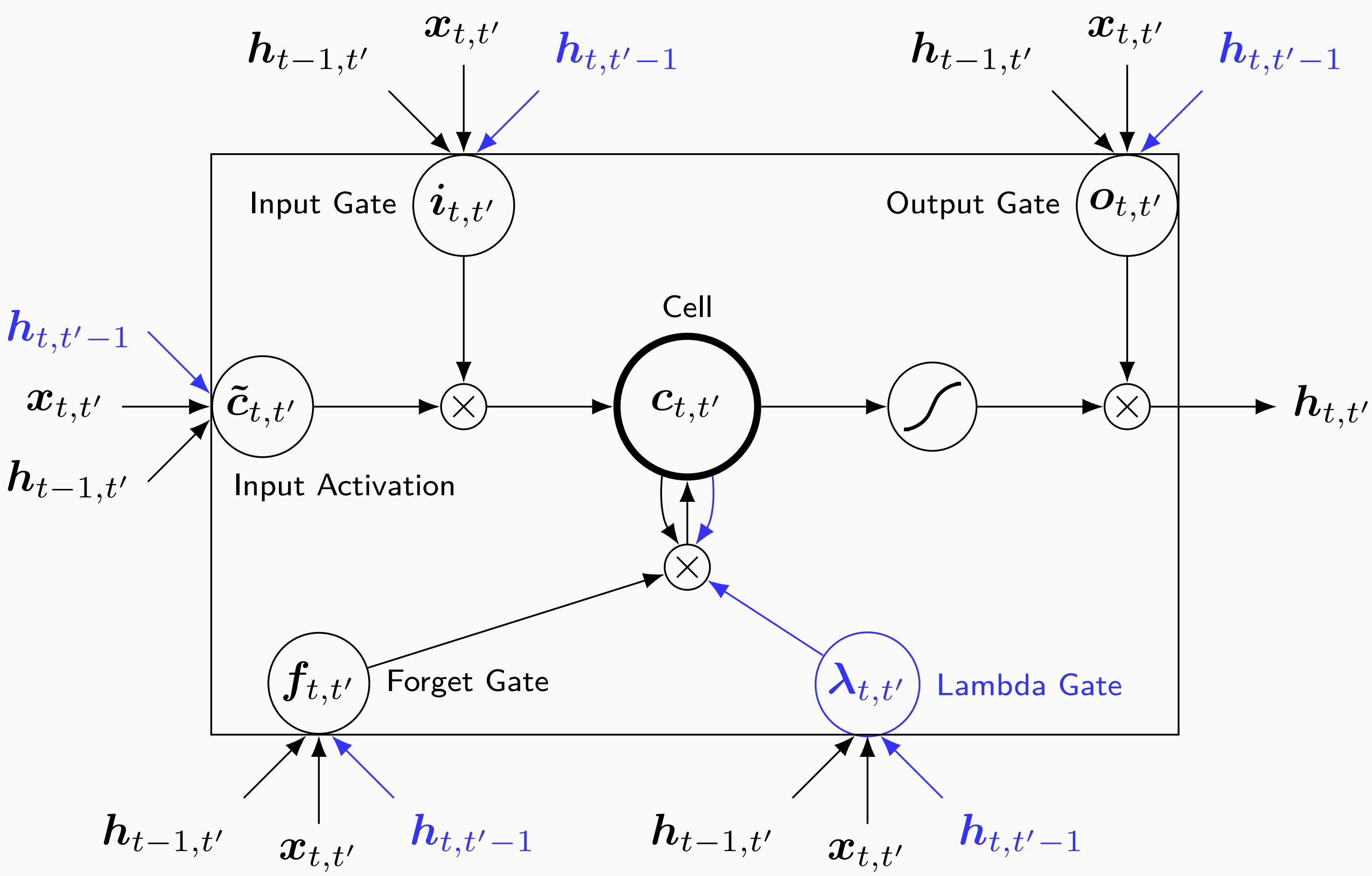
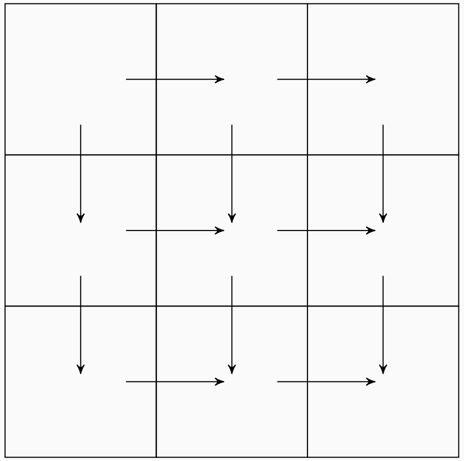


Figure: 2DLSTM cell

Parallel Processing

- **1DLSTM iterating over n inputs: $\mathcal{O}(n)$ operations**



(a) Dependencies

1	2	3
4	5	6
7	8	9

(b) Ordered processing

1	2	3
2	3	4
3	4	5

(c) Parallel processing

- **2DLSTM can be optimized to only $\mathcal{O}(n+m)$ operations [Voigtlaender+ 16]**

2D Sequence to Sequence (2D seq2seq)

Novel architecture [Bahar+ 18]:

- **no explicit encoder**
- **no explicit decoder**
- **complexity $\mathcal{O}(I+J)$**

2DLSTM:

$$a_{0,0} = 0$$

$$a_{j,i} = 2DLSTM([f_j, e_{i-1}], a_{j-1,i}, a_{j,i-1})$$

Prediction:

$$\hat{e}_i \leftarrow \text{softmax}(a_{J,i})$$

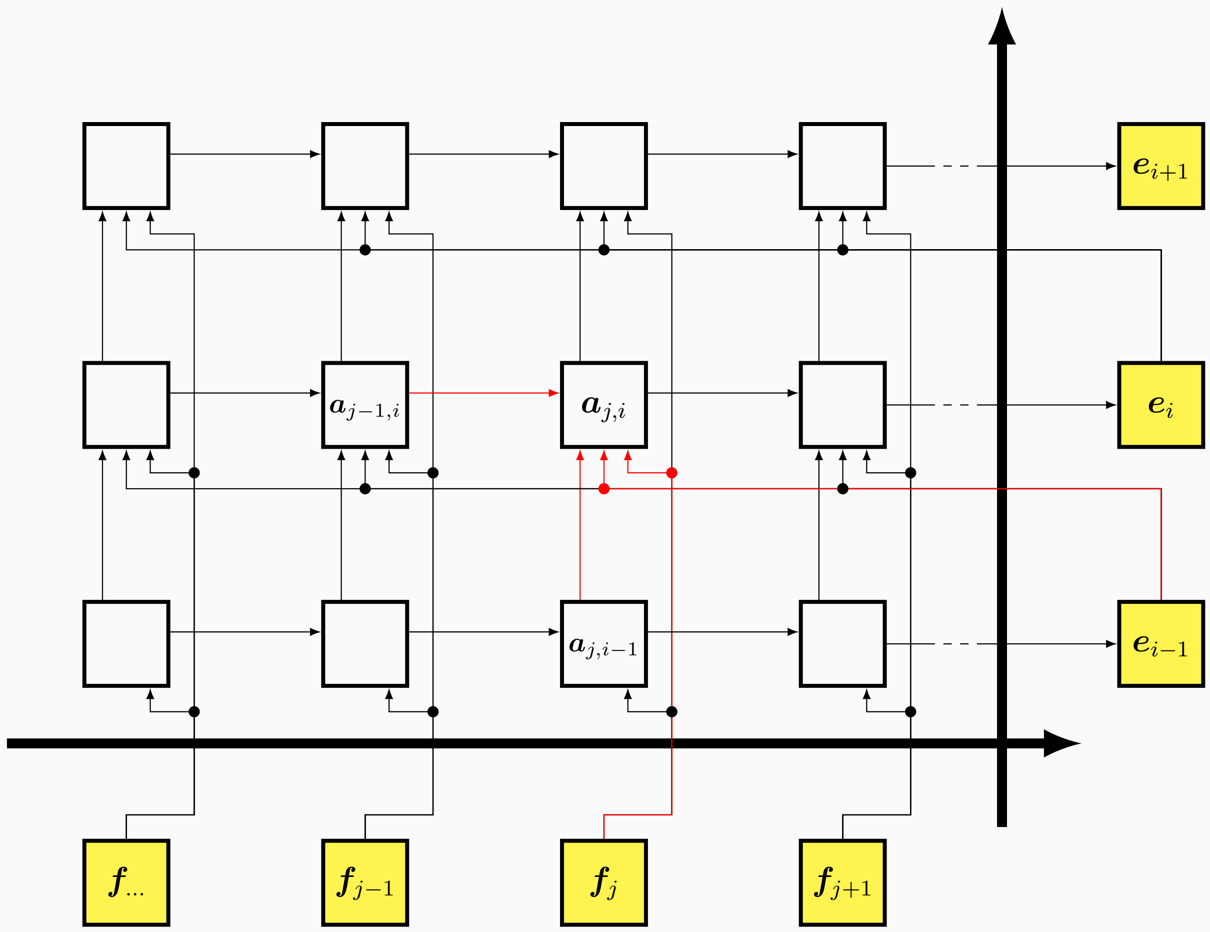


Figure: 2D seq2seq architecture

2D seq2seq - Results

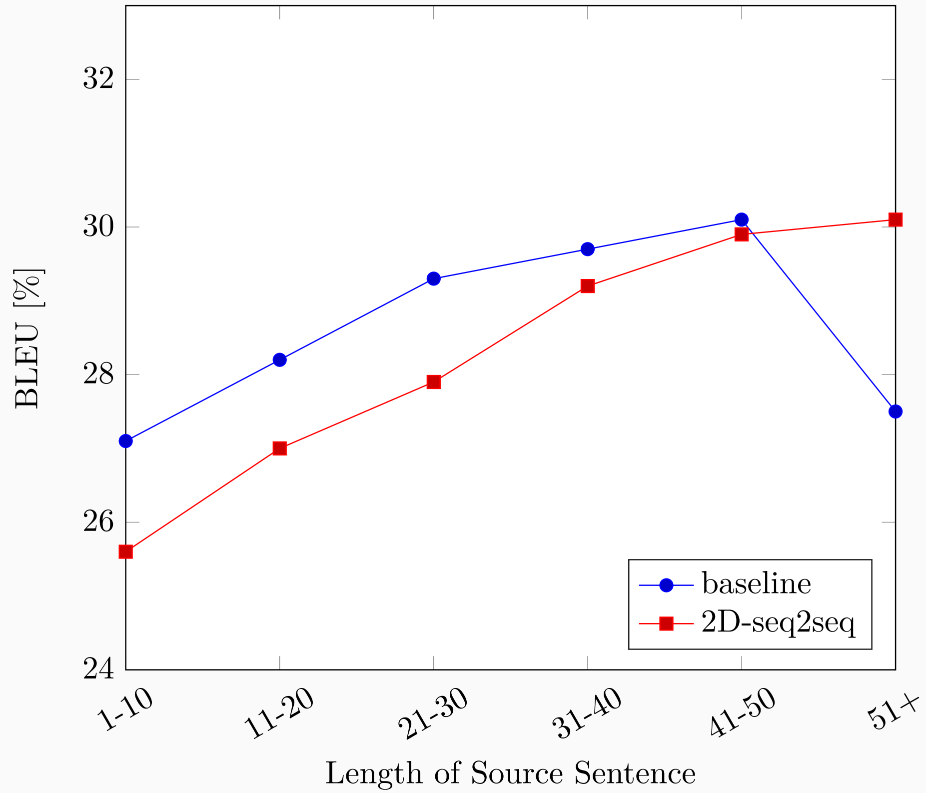
Table: WMT 2016/17, with an encoder/attention/decoder/2DLSTM size of 1000.

	German→English				English→German			
	BLEU [%]		TER [%]		BLEU [%]		TER [%]	
	2016	2017	2016	2017	2016	2017	2016	2017
Baseline	33.1	29.0	47.5	51.9	27.4	22.9	53.9	60.2
2D seq2seq	33.7	29.3	46.9	51.9	28.9	23.2	52.6	59.5

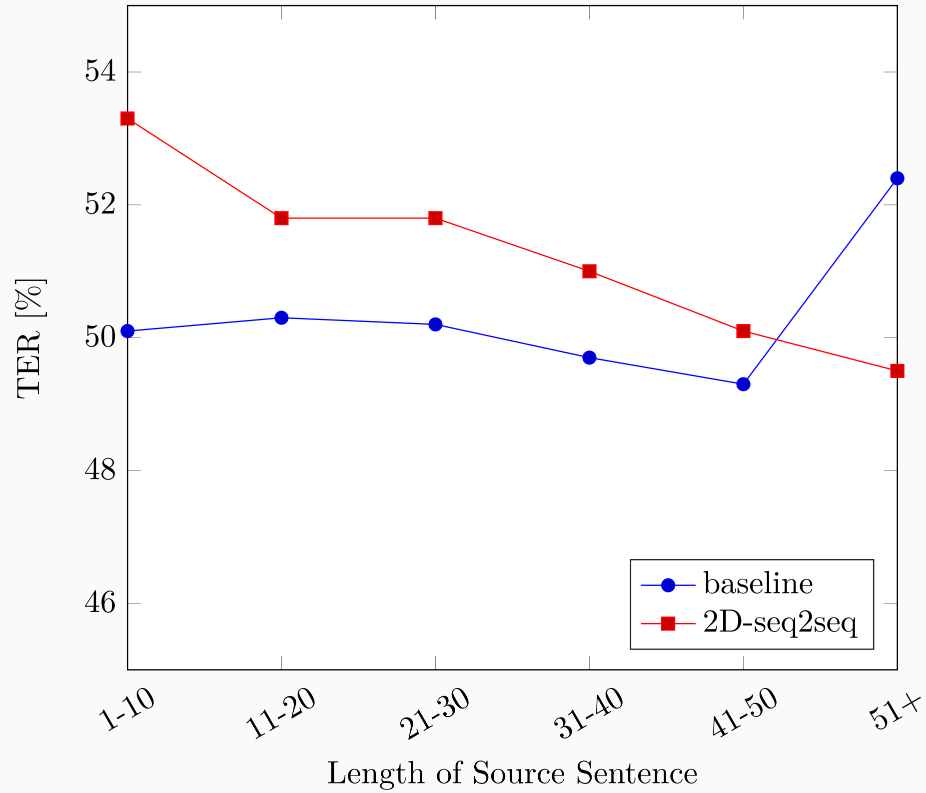
Table: Training and Decoding Speed.

	Training [tokens/s]	Decoding [tokens/s]
Baseline	2,944	48
2D seq2seq	791	0.7

2D seq2seq - Performance w.r.t. Sequence Length



(a) BLEU score w.r.t. the sequence length



(b) TER score w.r.t. the sequence length

Figure: WMT 2017 `newstest2015`, `newstest2016` and `newstest2017` German→English

- Groups contain 1455, 3081, 2133, 990, 344 and 169 sentences, respectively
- 2D seq2seq does not suffer from long sequences

Data Cleaning/Augmentation/Fairness

Data Cleaning:

- Paracrawl corpus: 5.000.000.000 German-English sentence pairs
- Very noisy

Data Augmentation:

- Translate monolingual data with model A to train model B on bilingual data
- Useful for small corpora

Data Fairness:

- Biased corpora create biased models
- Provide additional information (eg. gender) to model

Document Level Translations

Problem:

- **Sentence-wise translations may be inconsistent**
  - **Gender**
  - **Technical terms**
  - **Missing context**

Possible solutions:

- **Attention over previous sentence**
- **Additional document summaries**

Sparsity

Idea:

- **Remove part of the network to save space/computation time**

Different kinds of sparsity:

- **Structured sparsity**
  - **Delete whole layers**
  - **Delete individual neurons**
  - **Delete blocks of connections**
- **Unstructured sparsity**
  - **Delete individual connections**

Thank you for your attention

Christopher Brix

`Christopher.Brix@rwth-aachen.de`

`www.christopher-brix.de`

References

- **P. Bahar, C. Brix, H. Ney.**
  Towards two-dimensional sequence to sequence model in neural machine translation.
  In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3009–3015, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

- **D. Bahdanau, K. Cho, Y. Bengio.**
  Neural machine translation by jointly learning to align and translate.
  In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

- **A. Graves, S. Fernández, J. Schmidhuber.**
  Multi-dimensional recurrent neural networks.
  In [?], pp. 549–558.

I. Sutskever, O. Vinyals, Q. V. Le.
Sequence to sequence learning with neural networks.
In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. Curran Associates, Inc., 2014.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin.
Attention is all you need.
CoRR, Vol. abs/1706.03762, 2017.