Fixed Encoder Self-Attention Patterns in Transformer-Based Machine Translation

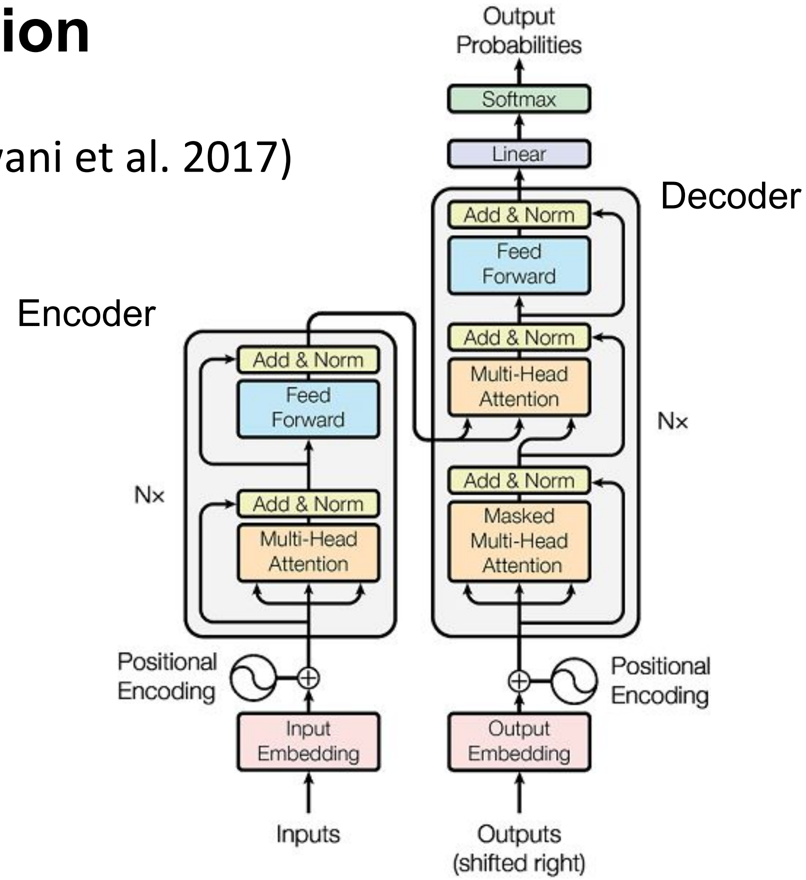
Alessandro Raganato Yves Scherrer Jörg Tiedemann

UNIVERSITY OF HELSINKI
FACULTY OF ARTS

EMNLP 2020

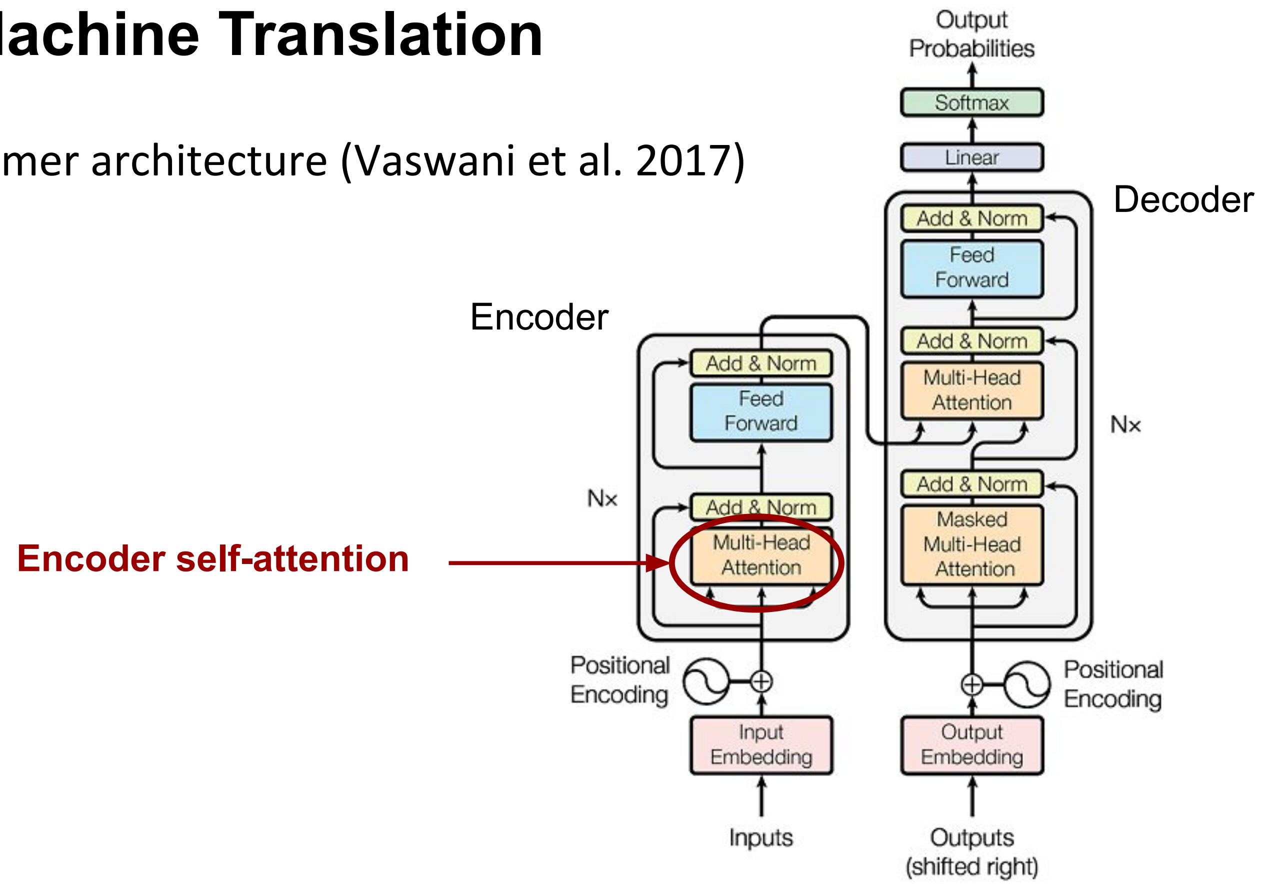
Neural Machine Translation

- Transformer architecture (Vaswani et al. 2017)



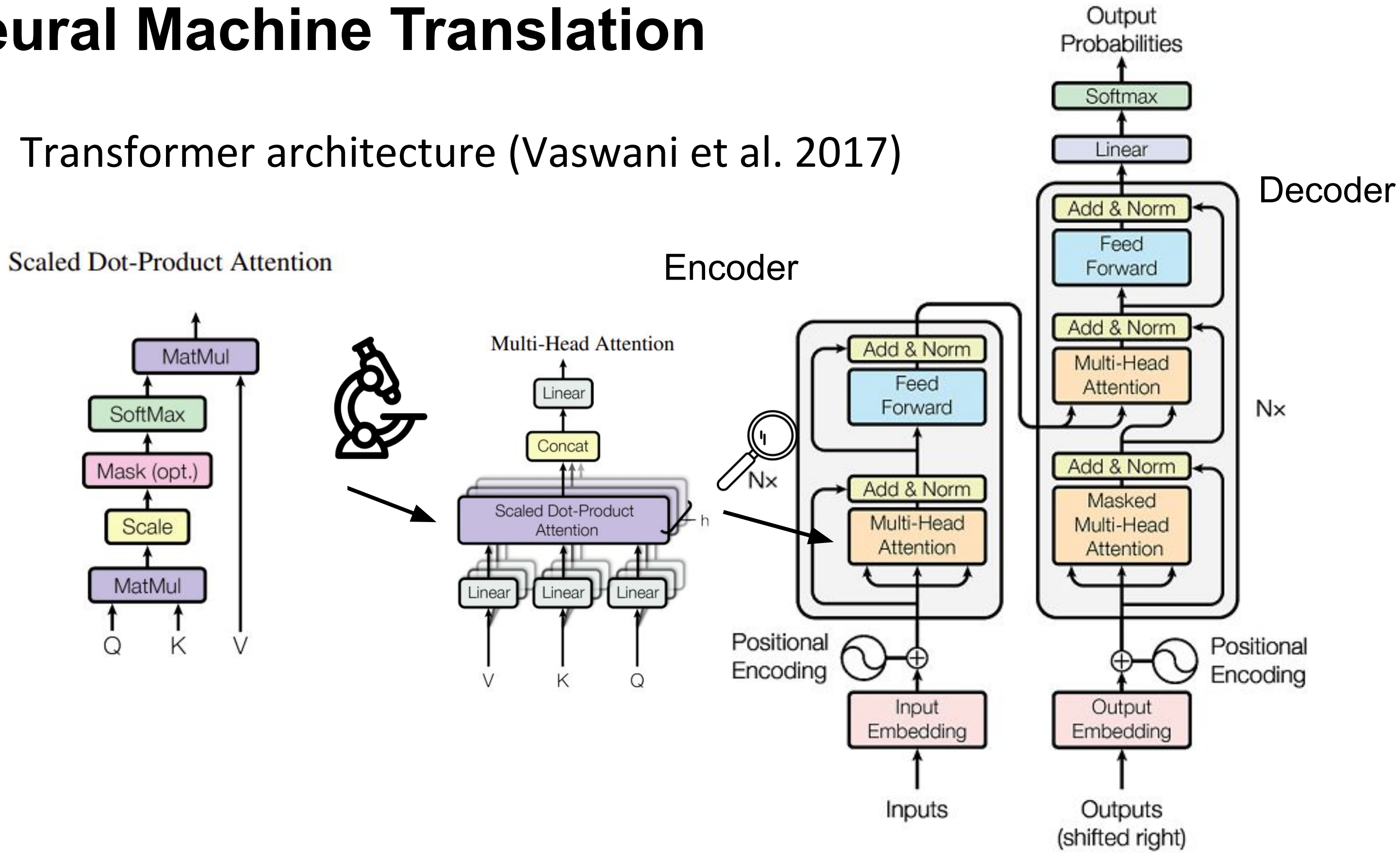
Neural Machine Translation

- Transformer architecture (Vaswani et al. 2017)



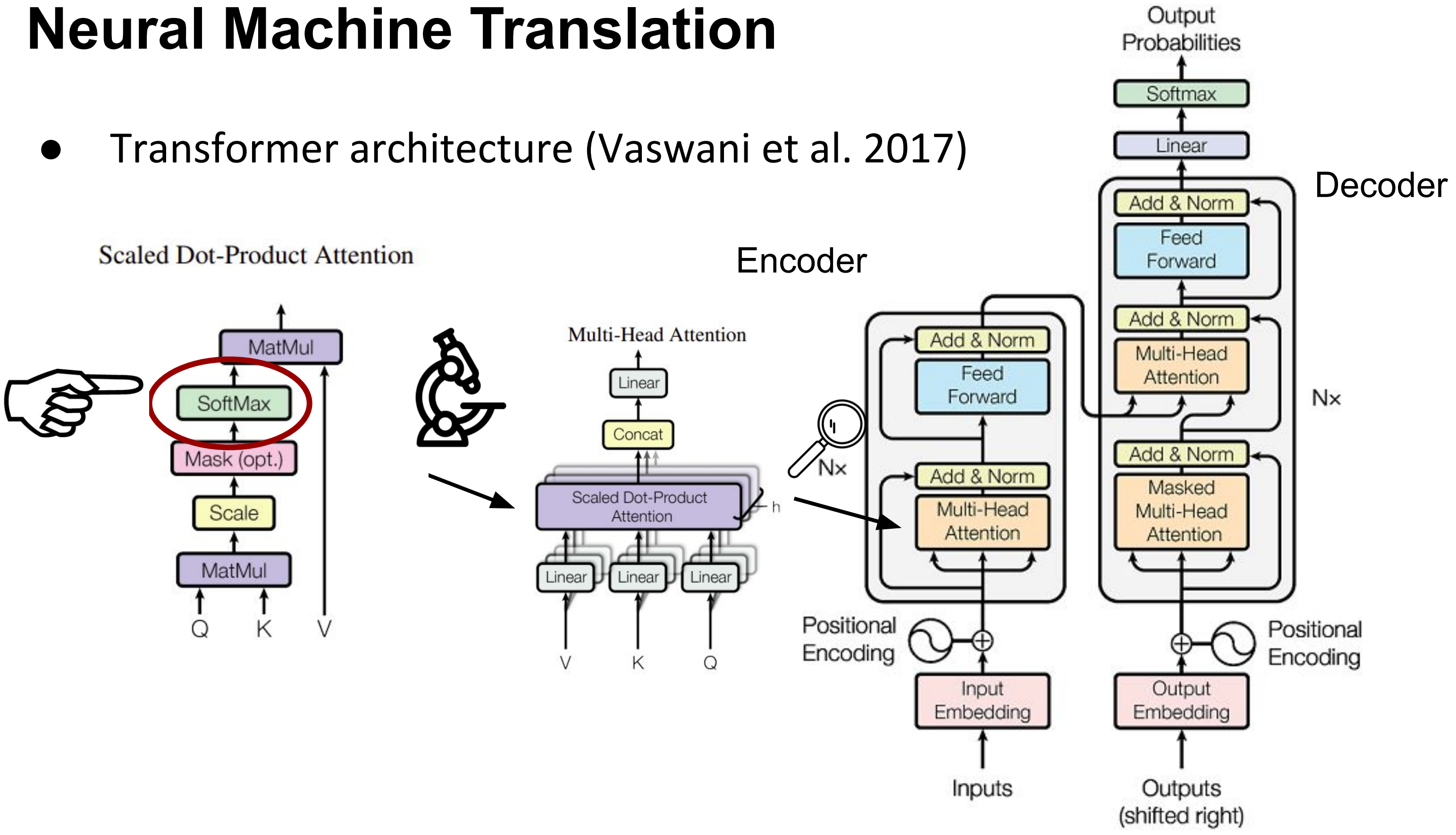
Neural Machine Translation

- Transformer architecture (Vaswani et al. 2017)



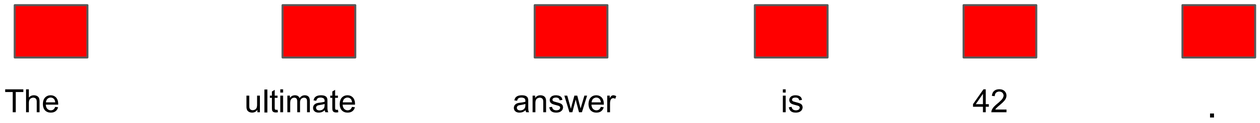
Neural Machine Translation

- Transformer architecture (Vaswani et al. 2017)



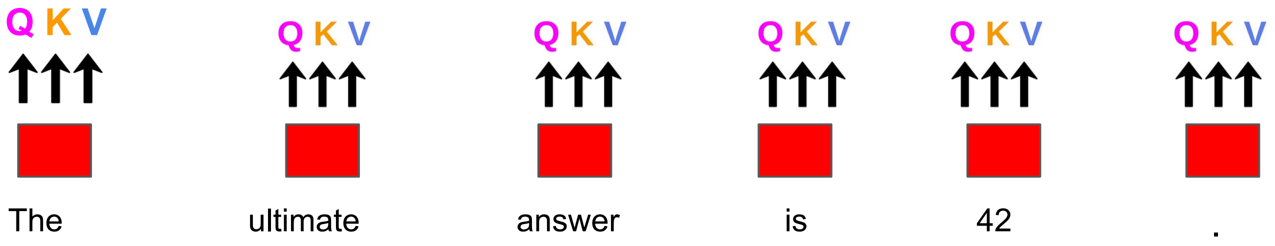
Transformer architecture

- Encoder self-attention



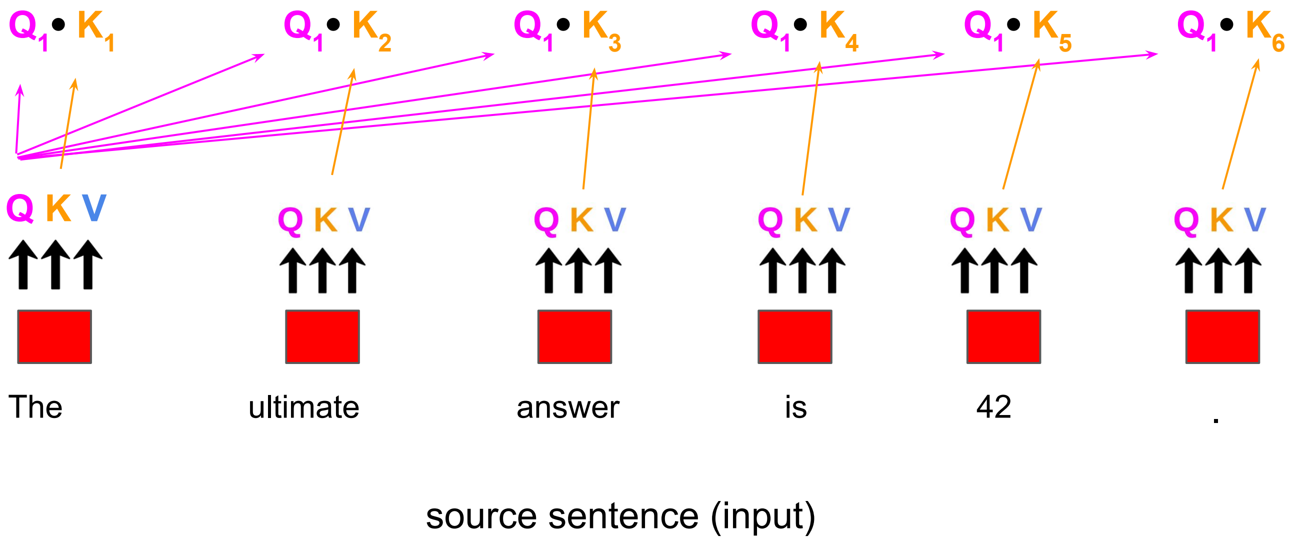
source sentence (input)

Transformer architecture



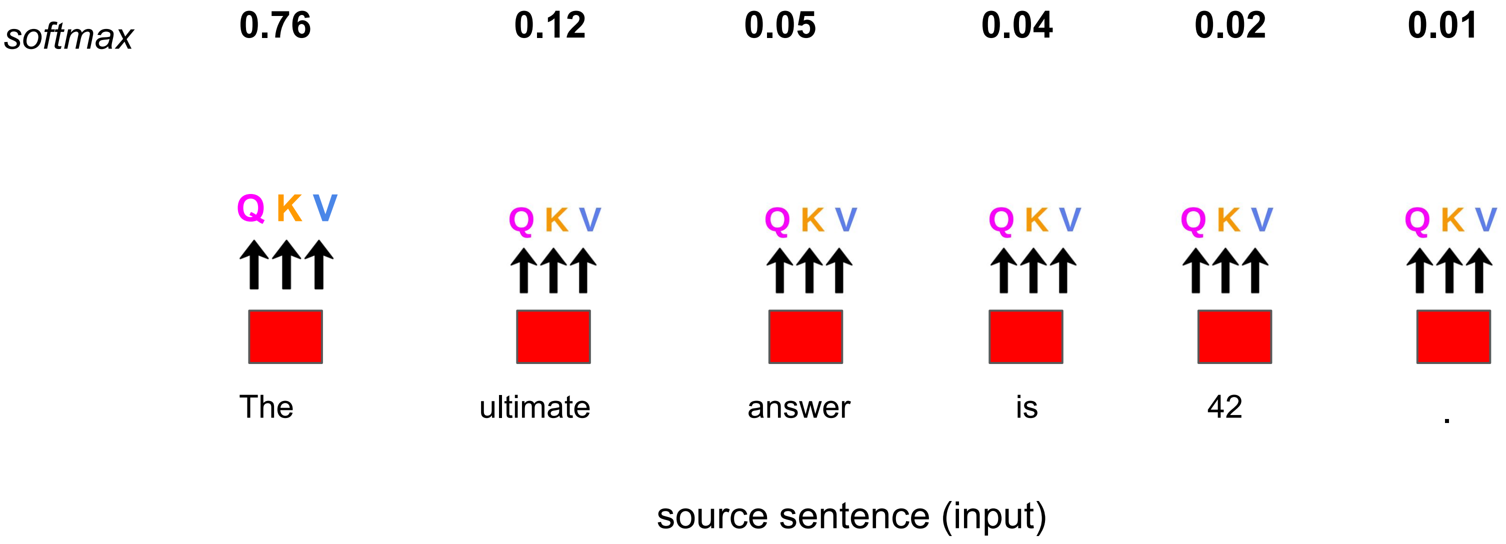
source sentence (input)

Transformer architecture



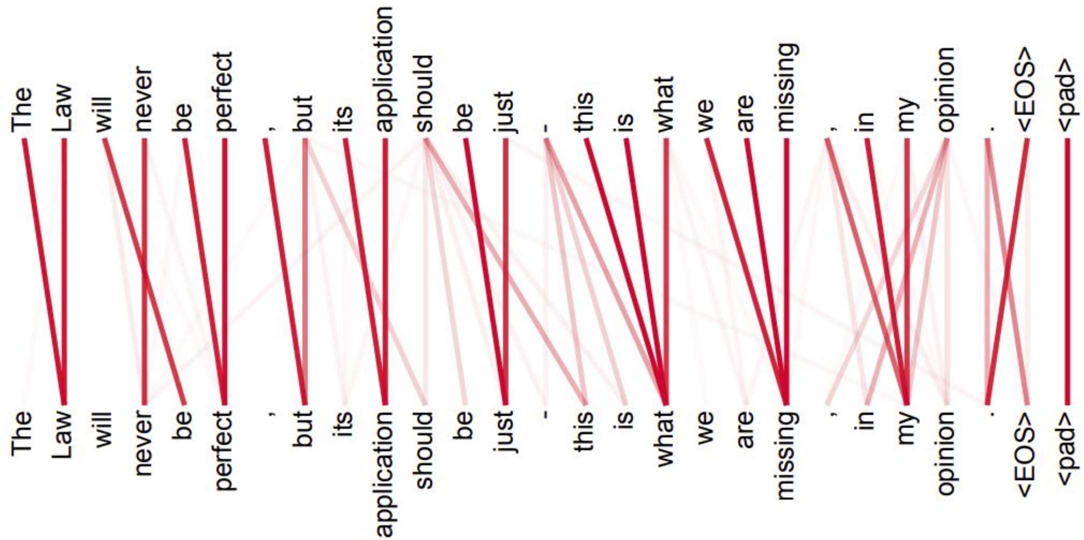
Transformer architecture

- Eight attention heads in the *base* version of the Transformer



Motivation

- Many of the attention heads exhibit behaviour that seems related to the structure of the sentence. The heads learned to perform different tasks. (Vaswani et al. 2017)
- Tremendous amount of works on how to interpret them



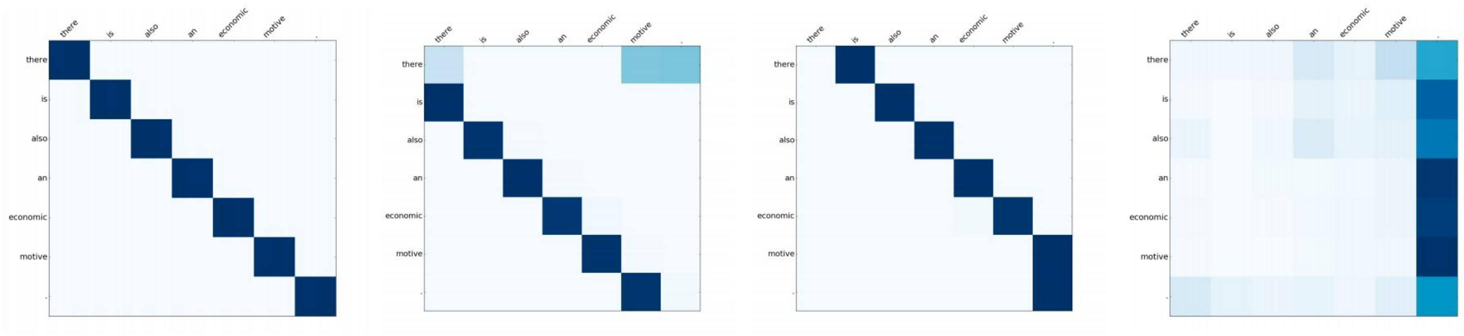
Motivation

❖ A portion of encoder self-attention patterns learned by the Transformer architecture reflect positional encoding of contextual information (Raganato and Tiedemann, 2018; Kovaleva et al., 2019; Voita et al., 2019ab; Correia et al., 2019, etc.)

Motivation

❖ A portion of encoder self-attention patterns learned by the Transformer architecture reflect positional encoding of contextual information (Raganato and Tiedemann, 2018; Kovaleva et al., 2019; Voita et al., 2019ab; Correia et al., 2019, etc.)

➢ four different positional patterns (Raganato and Tiedemann, 2018)



Motivation

- A portion of encoder self-attention patterns learned by the Transformer architecture reflect positional encoding of contextual information (Raganato and Tiedemann, 2018; Kovaleva et al., 2019; Voita et al., 2019ab; Correia et al., 2019, etc.)

➔ Instead of learning positional patterns, we can replace them by fixed *non-learnable* predefined patterns, reflecting the importance of locality, without the need of learning them!

- Without requiring any learnable parameters nor external knowledge!

Fixed encoder self-attention patterns

- We design seven intuitive and simple fixed attention patterns
  - example sentence: "*a master of science fic## tion .*"

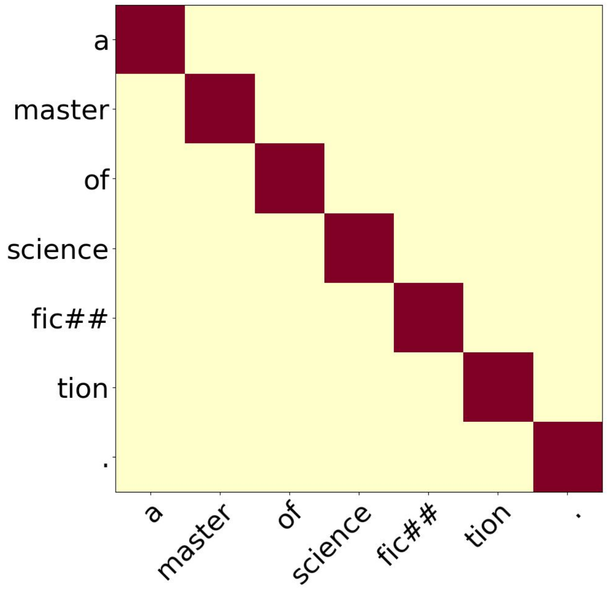
Given the *i-th* word within a sentence of length *n*, we define the following patterns:

Fixed encoder self-attention patterns

- Example sentence: "*a master of science fic## tion .*"

Given the *i-th* word within a sentence of length *n*, we define the following patterns:

1. **the current token**

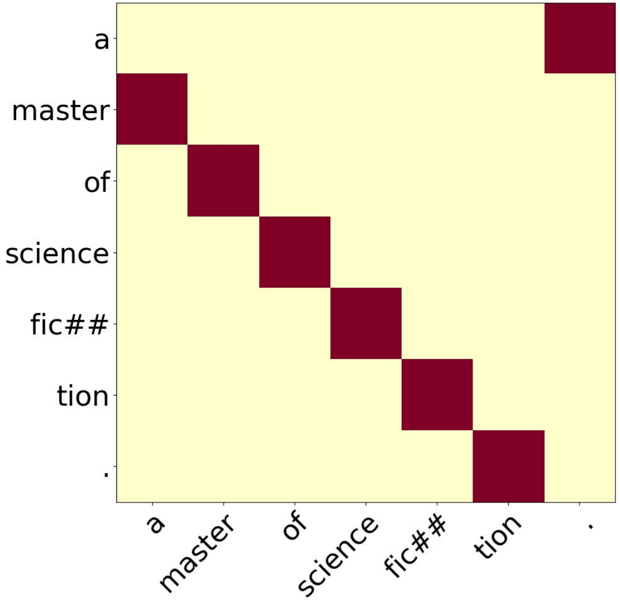


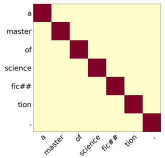
Fixed encoder self-attention patterns

- Example sentence: “*a master of science fic## tion .*”

Given the *i-th* word within a sentence of length *n*, we define the following patterns:

1. the current token
2. **the previous token**



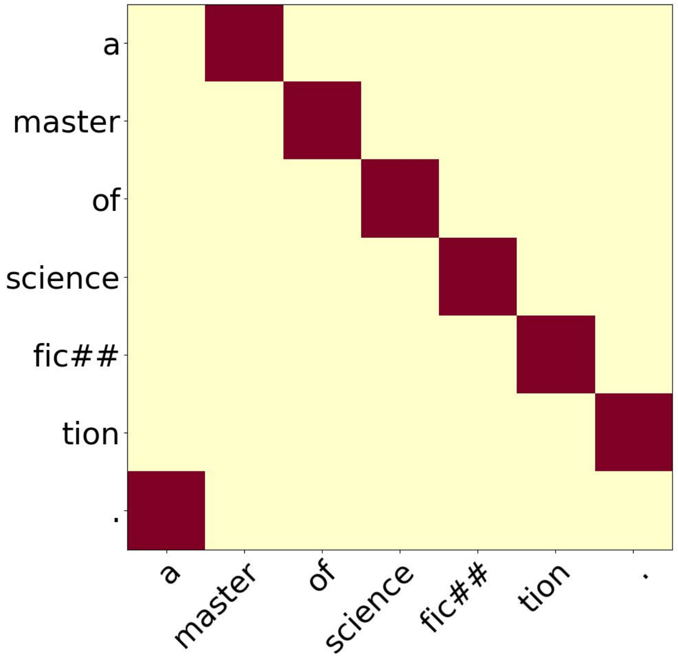


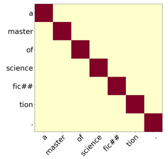
Fixed encoder self-attention patterns

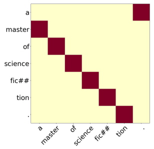
- Example sentence: “*a master of science fic## tion .*”

Given the *i-th* word within a sentence of length *n*, we define the following patterns:

1. the current token
2. the previous token
3. **the next token**





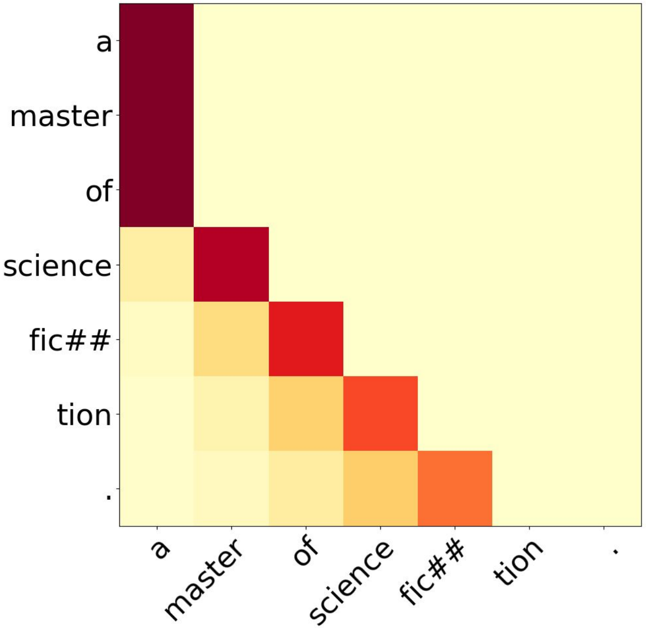


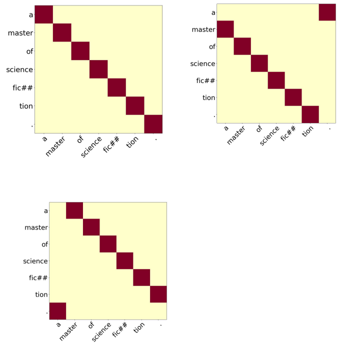
Fixed encoder self-attention patterns

- Example sentence: “*a master of science fic## tion .*”

Given the *i-th* word within a sentence of length *n*, we define the following patterns:

1. the current token
2. the previous token
3. the next token
4. **the larger left-hand context**



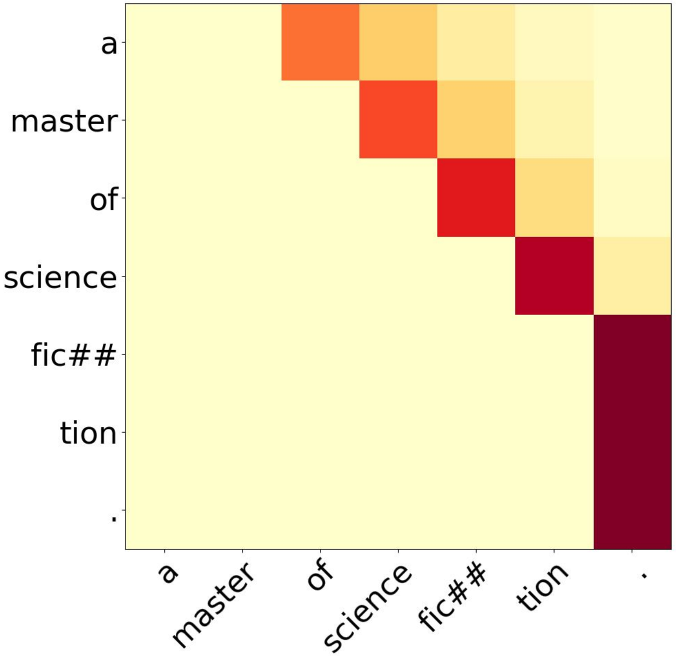


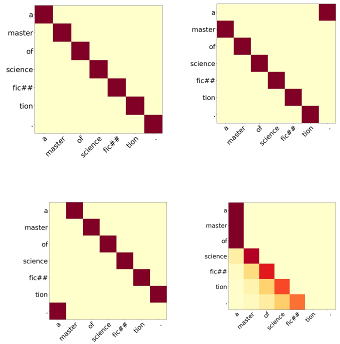
Fixed encoder self-attention patterns

- Example sentence: "*a master of science fic## tion .*"

Given the *i-th* word within a sentence of length *n*, we define the following patterns:

1. the current token
2. the previous token
3. the next token
4. the larger left-hand context
5. **the larger right-hand context**



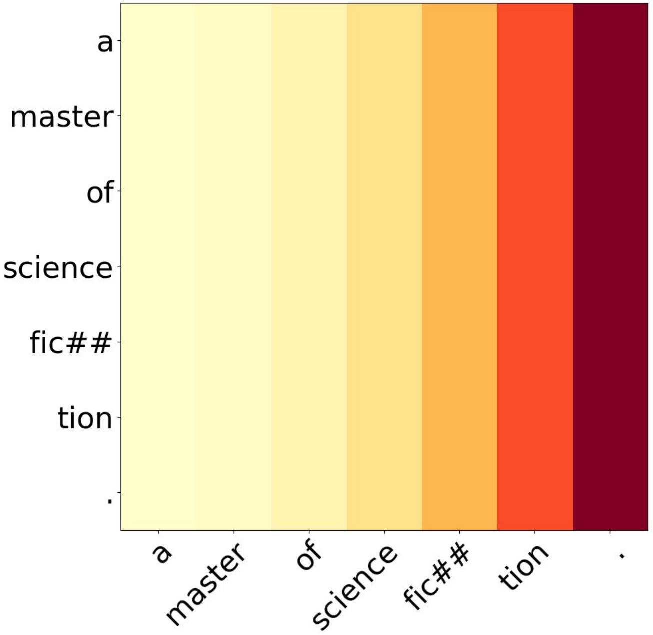


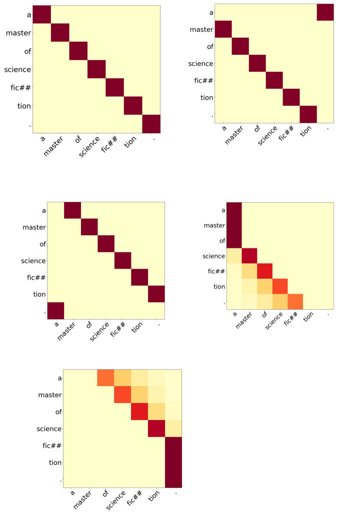
Fixed encoder self-attention patterns

- Example sentence: “*a master of science fic## tion .*”

Given the *i-th* word within a sentence of length *n*, we define the following patterns:

1. the current token
2. the previous token
3. the next token
4. the larger left-hand context
5. the larger right-hand context
6. **the end of the sentence**



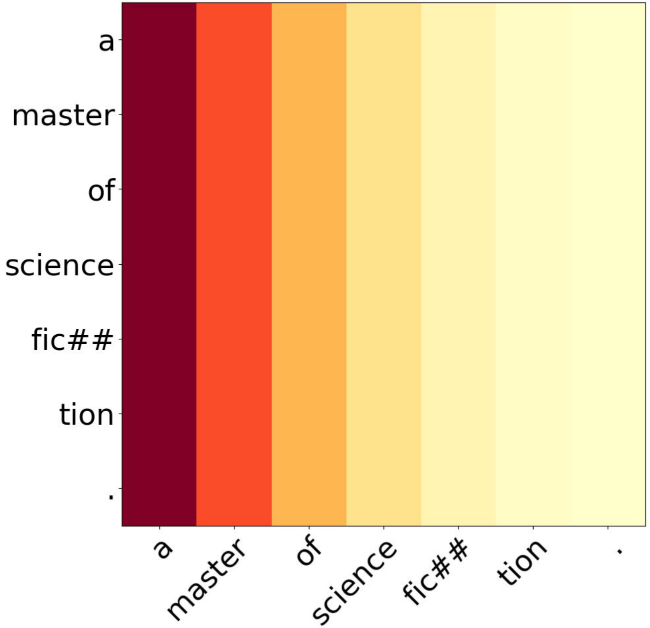


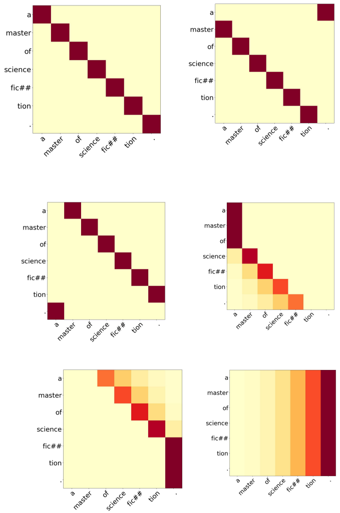
Fixed encoder self-attention patterns

- Example sentence: “*a master of science fic## tion .*”

Given the *i-th* word within a sentence of length *n*, we define the following patterns:

1. the current token
2. the previous token
3. the next token
4. the larger left-hand context
5. the larger right-hand context
6. the end of the sentence
7. **the start of the sentence**





Fixed encoder self-attention patterns

- Example sentence: “*a master of science fic## tion .*”

Given the *i-th* word within a sentence of length *n*, we define the following patterns:

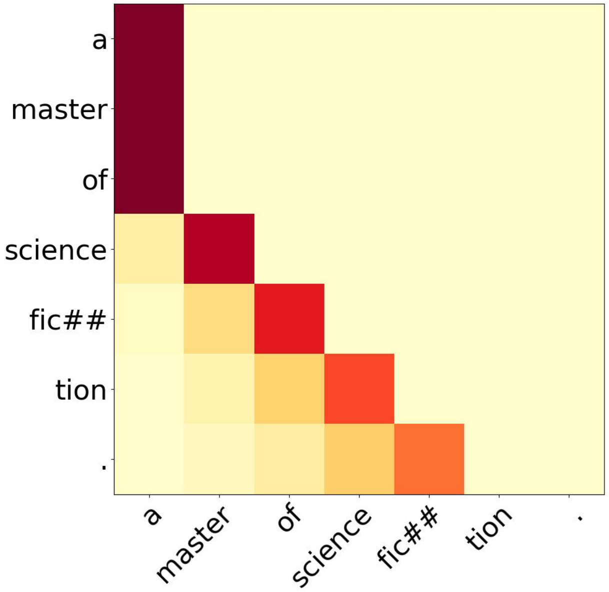
4. **the larger left-hand context:** a function *f* over the positions *0* to *i−2*

$$\xi_{i,j}^{(2)} = \begin{cases} 1 & \text{if } j = i - 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\xi_{i,j}^{(4)} = \begin{cases} f^{(4)}(j) & \text{if } j \leq i - 2 \\ 0 & \text{otherwise} \end{cases}$$

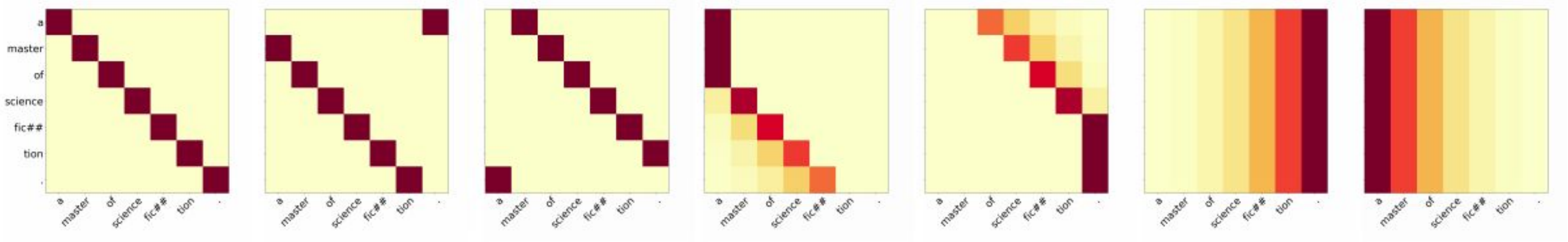
where

$$f^{(4)}(j) = \frac{(j+1)^3}{\sum_{j=0}^{i-2}(j+1)^3}$$



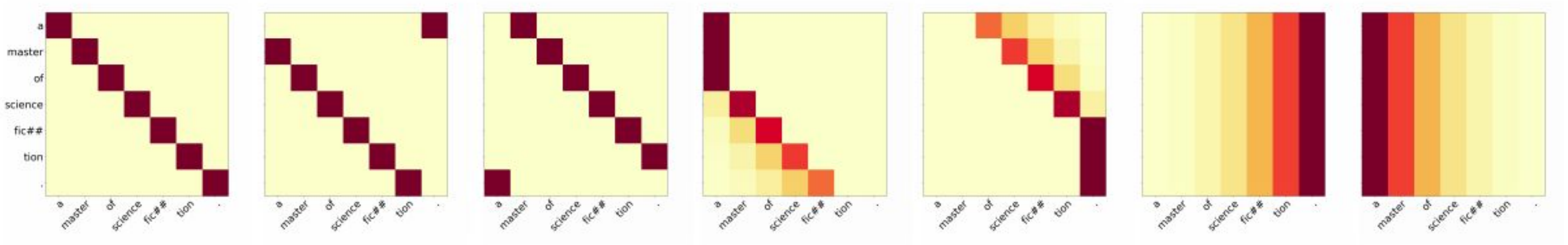
Fixed encoder self-attention patterns

- *Token-based* fixed attention patterns

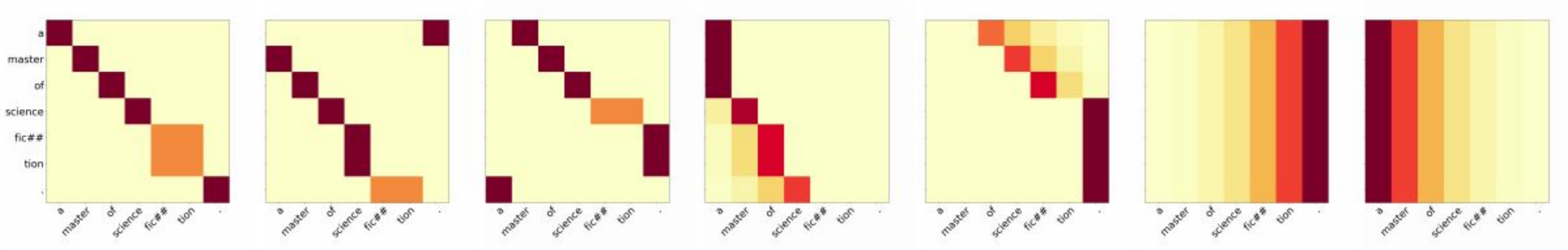


Fixed encoder self-attention patterns

- *Token-based* fixed attention patterns



- *Word-based* fixed attention patterns



Experimental setup

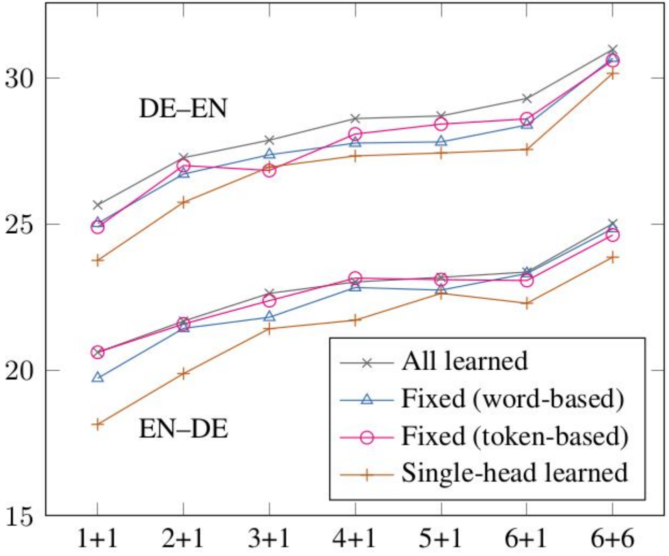
- Transformer models:
    - **8L**: all 8 attention heads in each layer are learnable,
    - **7Ftoken+1L**: 7 fixed token-based attention heads and 1 learnable head per encoder layer,
    - **7Fword+1L**: 7 fixed word-based attention patterns and 1 learnable head per encoder layer,
    - **1L**: a single learnable attention head per encoder layer.

- Evaluation settings:
    - High resource scenario:
        - German <-> English, 11.5M training sentences
    - Mid-size scenario:
        - German <-> English, 2.9M training sentences
    - Low-resource scenario:
        - German -> English, 159K training sentences
        - Korean -> English, 90K training sentences
        - Vietnamese <-> English, 133K training sentences

- Evaluation metric:
    - BLEU score

Experiments and results

- Mid-size scenario:
  - German <-> English, 2.9M training sentences



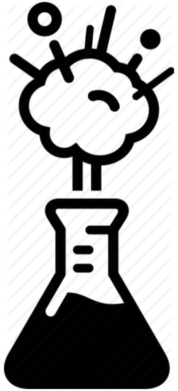
- The x-axis shows different configurations of encoder and decoder layers

Experiments and results

- Low-resource scenario:
  - German -> English, 159K training sentences
  - Korean -> English, 90K training sentences
  - Vietnamese <-> English, 133K training sentences

Enc. heads	DE–EN	KO–EN	EN–VI	VI–EN
8L	30.86	6.67	29.85	26.15
$7F_{token}+1L$	**32.95**	8.43	31.05	**29.16**
$7F_{word}+1L$	32.56	**8.70**	**31.15**	28.90
1L	30.22	6.14	28.67	25.03
Prior work	† 33.60	† 10.37	$^{\uplus}$ 27.71	$^{\uplus}$ 26.15

Experiments and results



- High resource scenario:
    - German <-> English, 11.5M training sentences

Encoder heads	EN–DE	DE–EN
8L	26.75	**34.10**
$7F_{token}$+1L	26.52	33.50
$7F_{word}$+1L	**26.92**	33.17
1L	26.26	32.91

Ablation study

- We mask out one attention pattern across all encoder layers at test time

Ablation study

- We mask out one attention pattern across all encoder layers at test time

Disabled head	6+1 layers		6+6 layers	
	EN–DE	DE–EN	EN–DE	DE–EN
1 Current word	-0.15	0.11	0.12	-0.04
2 Previous word	**-5.72**	**-5.21**	**-3.05**	**-3.26**
3 Next word	**-1.80**	**-1.98**	**-2.08**	**-1.36**
4 Prev. context	**-4.73**	**-5.20**	**-1.42**	**-2.85**
5 Next context	-0.72	-0.34	-0.47	-0.66
6 Start context	-0.17	-0.12	0.14	0.13
7 End context	-0.02	0.12	-0.30	0.10
8 Learned head	**-2.22**	**-4.05**	-0.58	-0.78

Ablation study

- We mask out one attention pattern across all encoder layers at test time

Disabled head	6+1 layers		6+6 layers	
	EN–DE	DE–EN	EN–DE	DE–EN
1 Current word	-0.15	0.11	0.12	-0.04
2 Previous word	**-5.72**	**-5.21**	**-3.05**	**-3.26**
3 Next word	**-1.80**	**-1.98**	**-2.08**	**-1.36**
4 Prev. context	**-4.73**	**-5.20**	**-1.42**	**-2.85**
5 Next context	-0.72	-0.34	-0.47	-0.66
6 Start context	-0.17	-0.12	0.14	0.13
7 End context	-0.02	0.12	-0.30	0.10
8 Learned head	**-2.22**	**-4.05**	-0.58	-0.78

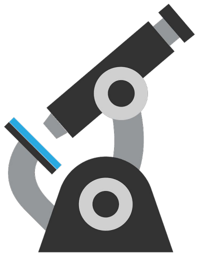
Disabled head	6+1 layers		6+6 layers	
	EN–VI	VI–EN	EN–VI	VI–EN
1 Current word	0.12	-0.14	0.16	-0.05
2 Previous word	**-2.32**	**-2.67**	**-2.71**	**-3.04**
3 Next word	**-1.12**	**-1.61**	**-1.35**	**-2.15**
4 Prev. context	**-4.11**	**-4.32**	**-2.82**	**-3.09**
5 Next context	-0.27	-0.50	-0.83	-0.77
6 Start context	-0.29	-0.08	-0.04	0
7 End context	0.28	-0.29	-0.23	-0.19
8 Learned head	-0.57	-0.88	-0.18	0.36

Eight fixed heads

- **8Ftoken**: extreme scenario where the eighth attention head is fixed as well:
  - eighth attentive pattern focuses on the last token, with a fixed weight of 1.0 at position *n*

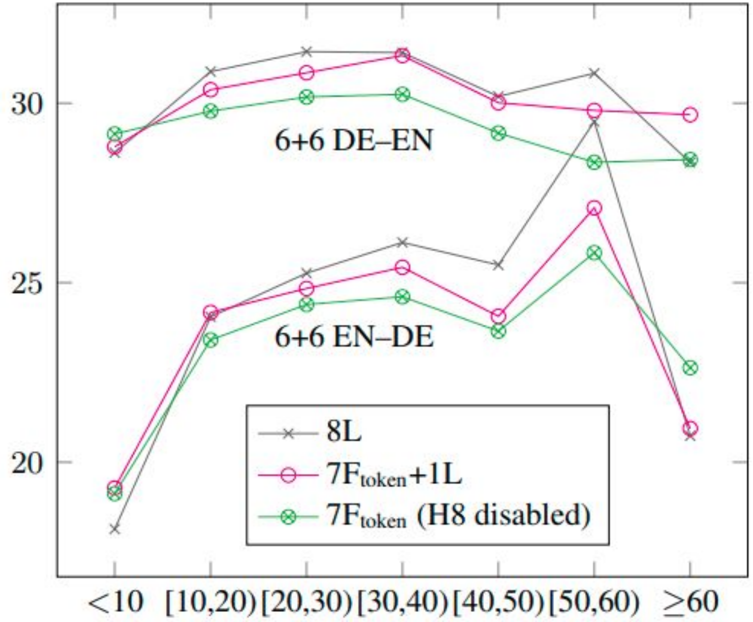
Enc. heads	#Param.	EN–DE	DE–EN	EN–VI	VI–EN
8L	91.7M	25.02	30.99	29.85	26.15
$7F_{token}$+1L	88.9M	24.63	30.61	31.05	29.16
$8F_{token}$	88.5M	24.64	30.56	31.45	28.97

Analysis:



Analysis: Sentence length

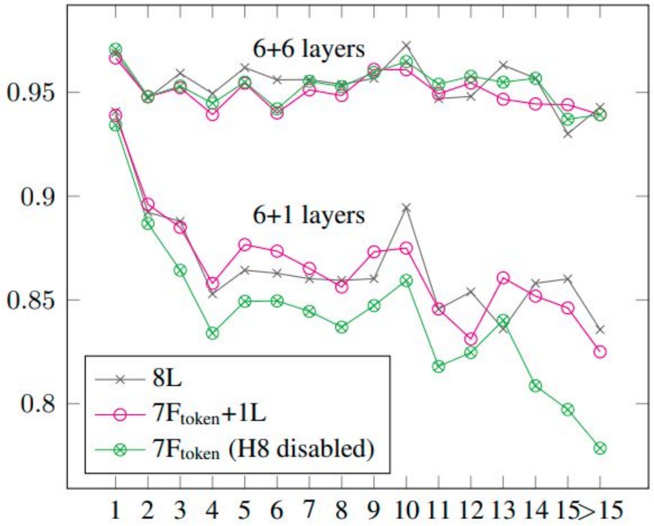
- BLEU scores for different ranges of sentence lengths



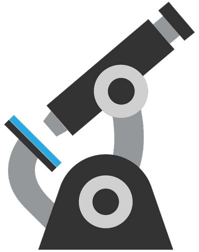
Analysis: Subject-verb agreement

- Contrastive test suite -- LingEval97 (Sennrich, 2017)
- Metric: accuracy score

- The *x*-axis shows distances between the subject and the verb.



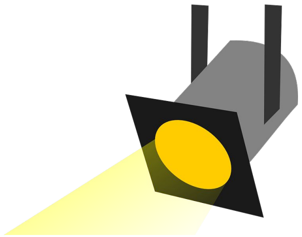
Analysis: Word Sense Disambiguation



- Contrastive test suites on word sense disambiguation:
  - ContraWSD (Rios Gonzales et al., 2017)
  - MuCoW (Raganato et al., 2019)
- Metric: accuracy score

	ContraWSD		MuCoW	
Encoder heads	6+1	6+6	6+1	6+6
8L	**0.804**	0.831	**0.741**	0.761
$7F_{token}$+1L	0.793	**0.834**	0.734	**0.772**
$7F_{token}$ (H8 disabled)	0.761	0.816	0.721	0.757

Conclusions



- Encoder self-attention can be simplified drastically, reducing parameter footprint at training time without degradation in translation quality

- Our extensive analyses show that:
    - only adjacent and previous token attentive patterns contribute significantly to the translation performance
    - the trainable encoder head can also be disabled without hampering translation quality if the number of decoder layers is deep enough
    - encoder attention heads based on locality patterns are beneficial in low-resource scenarios, but may affect the semantic feature extraction necessary for addressing lexical ambiguity phenomena

Thank you!