Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned

Voita et al.

Elbakian Movses

National Research University Higher School of Economics

2019

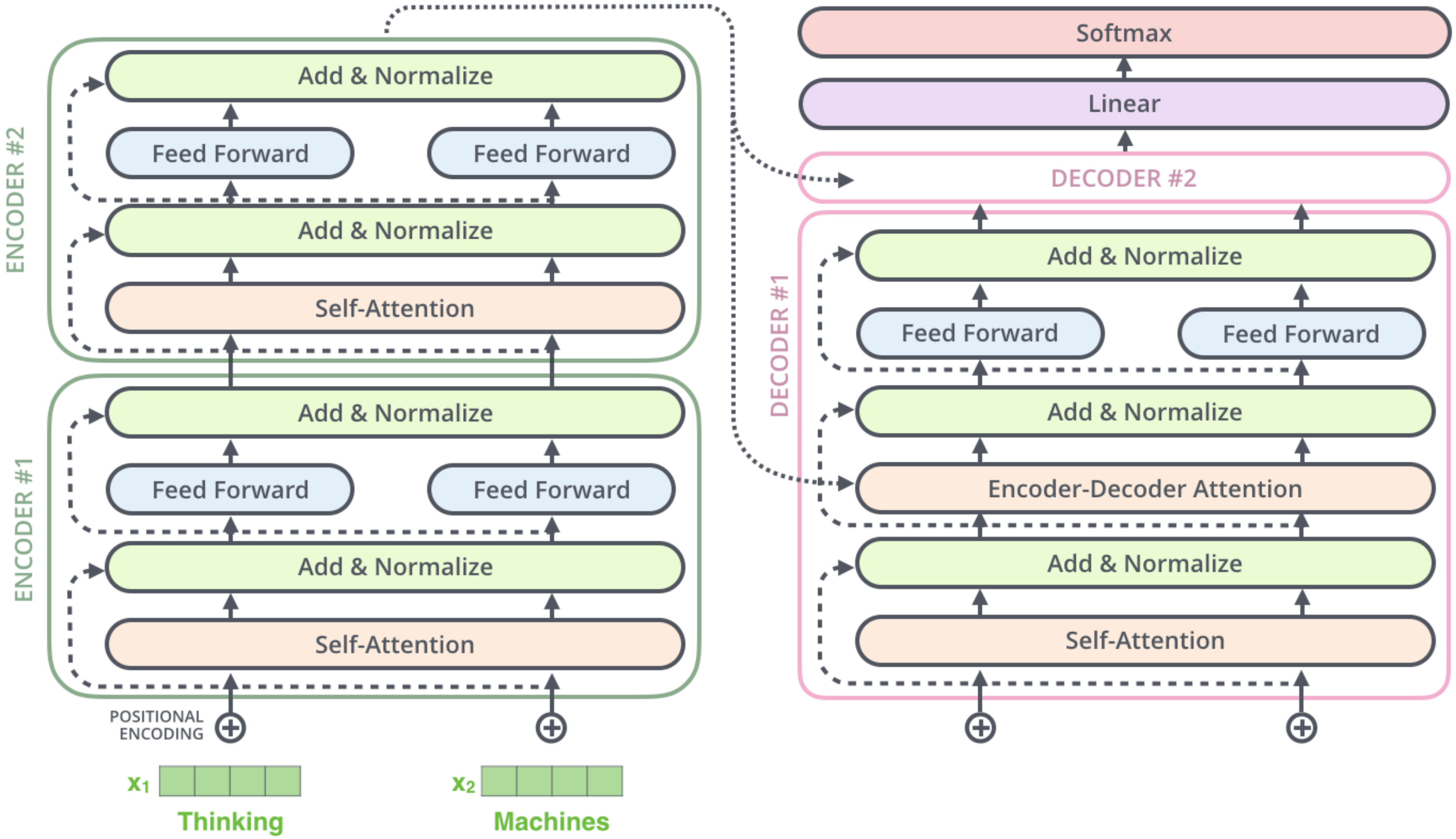
Plan

- Research objectives
- Brief reminder about Transformer
- Datasets
- Metrics for attention heads classification
- Attention heads classification
- Pruning attention heads
- Results
- References

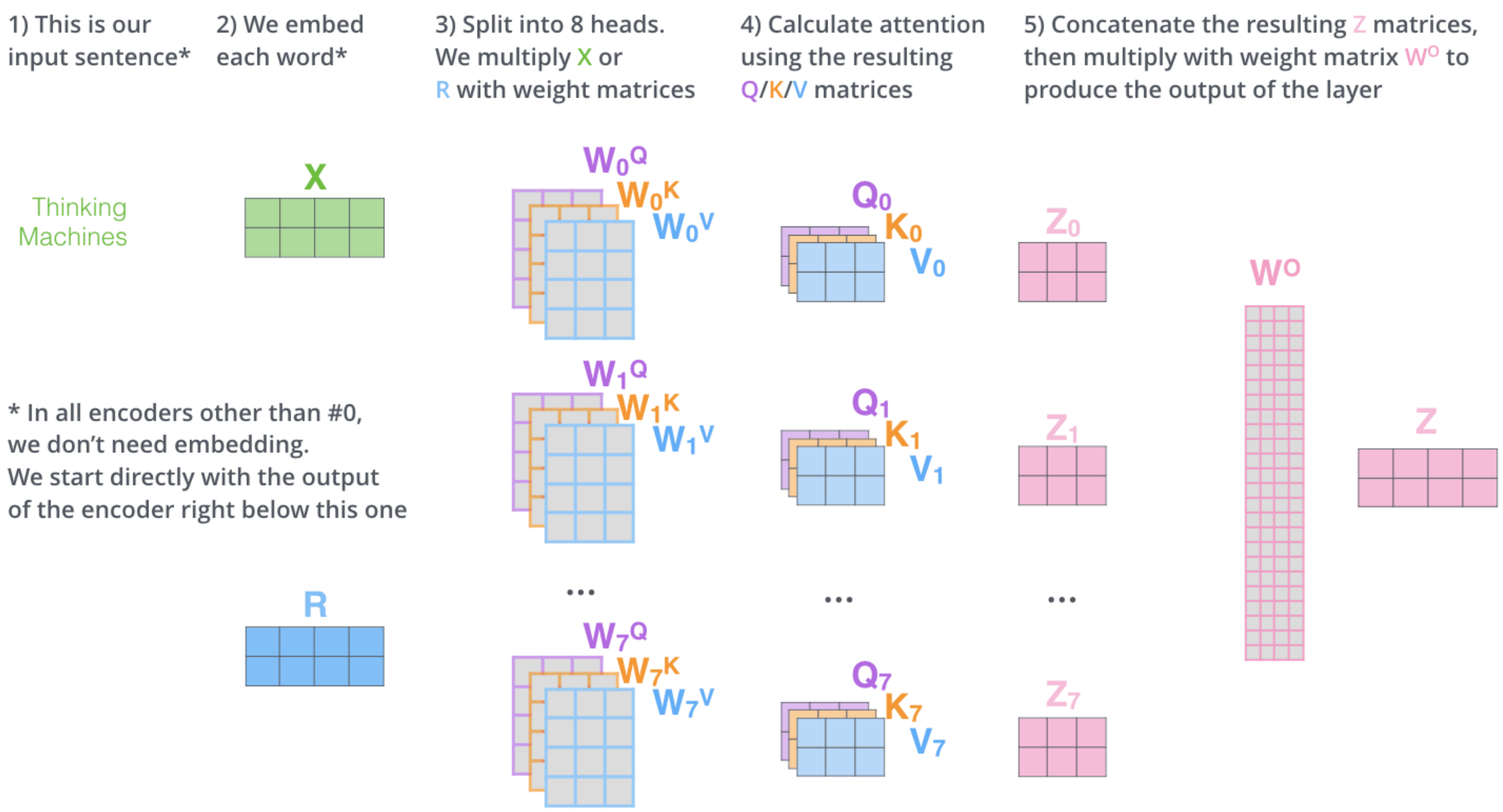
Research objectives

- To what extent does translation quality depend on individual encoder heads?
- Do individual encoder heads play consistent and interpretable roles? If so, which are the most important ones for translation quality?
- Which types of model attention (encoder self-attention, decoder self-attention or decoder-encoder attention) are most sensitive to the number of attention heads and on which layers?
- Can we significantly reduce the number of attention heads while preserving translation quality?

Transformer



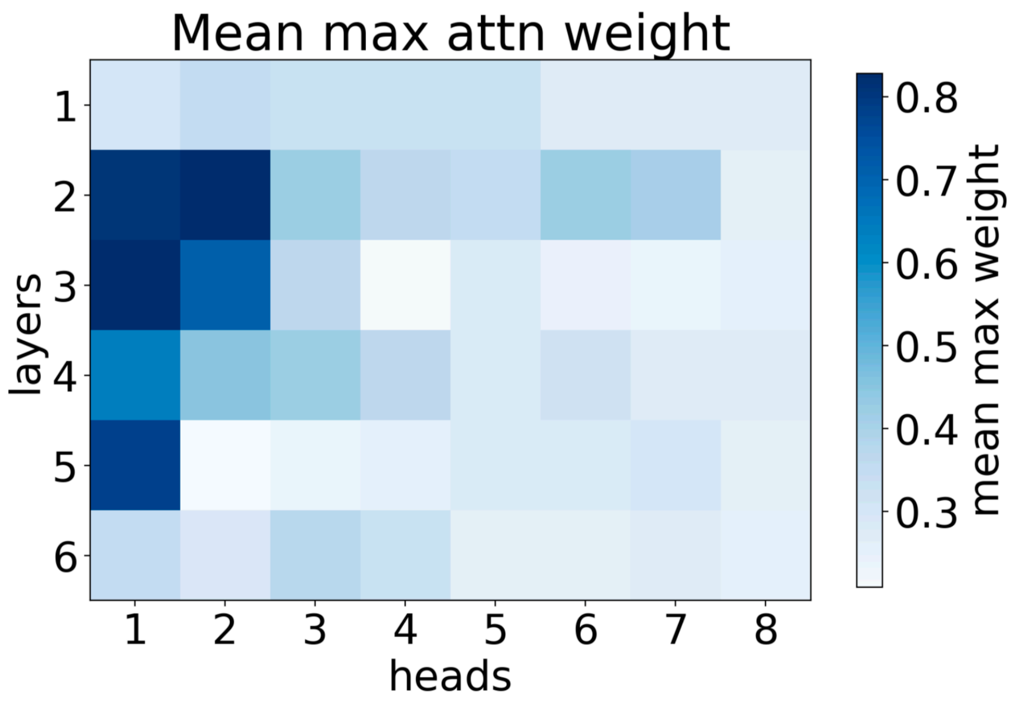
Transformer. Self-Attention



Datasets

- English is a source language. Russian, German and French are target languages.
- 2.5m pairs of sentences for training from WMT
- Also English-Russian OpenSubtitles2018 was used for pruning experiments

Metrics for attention heads classification. Confidence



- Confidence is an average of head’s maximum attention weight excluding the end of sentence symbol, where average is taken over tokens
- Confident head is one that usually assigns a high proportion of its attention to a single token.

Metrics for attention heads classification. Layer-wise relevance propagation. General Idea

- f – real-valued output of model
- $z = (z_d^{(l)})_{d=1}^{V(l)}$ – l-th layer
- $R_d^{(l)}$ – relevance score
- $f = \ldots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \cdots = \sum_{d} R_d^{(1)}$ – conservation principle

Total contribution of neurons at each layer is constant!

Metrics for attention heads classification. Layer-wise relevance propagation. Formulas

Weight ratio:

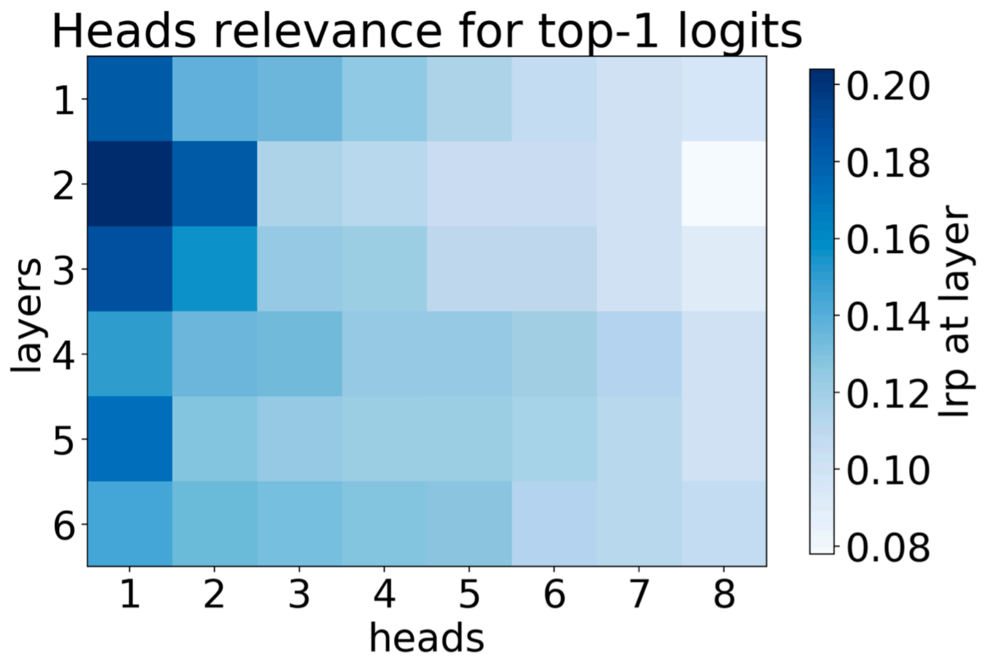
$$w_{u \to v} = \frac{W_{u,v} u}{\sum_{u' \in IN(v)} W_{u',v} u'} \quad \text{if } v = \sum_{u' \in IN(v)} W_{u',v} u',$$

$$w_{u \to v} = \frac{u}{\sum_{u' \in IN(v)} u'} \quad \text{if } v = \prod_{u' \in IN(v)} u'.$$

Relevance of neuron u to preceding neuron v:

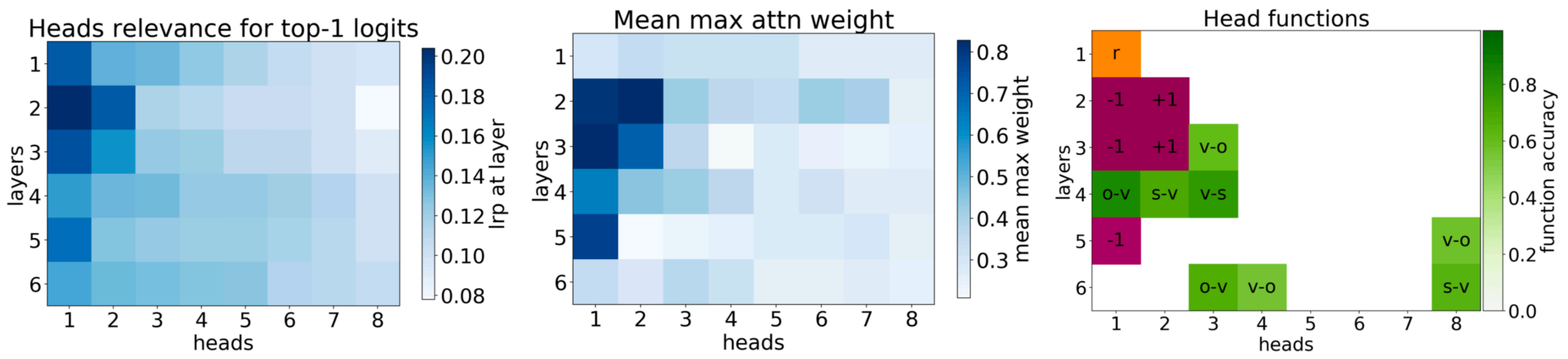
$$r_{u \leftarrow v} = \sum_{z \in OUT(u)} w_{u \to z} r_{z \leftarrow v}.$$

Metrics for attention heads classification.
Layer-wise relevance propagation.
Heads relevance



Attention heads classification

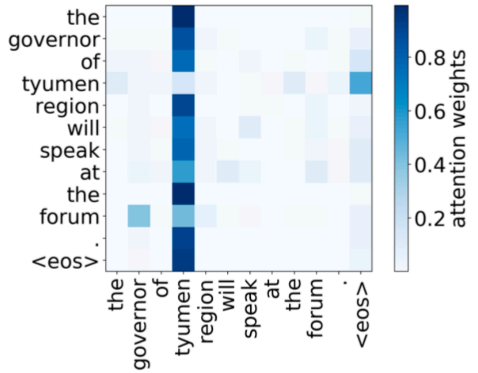
- positional: the head points to an adjacent token
- syntactic: the head points to tokens in a specific syntactic relation
- rare words: the head points to the least frequent tokens in a sentence.



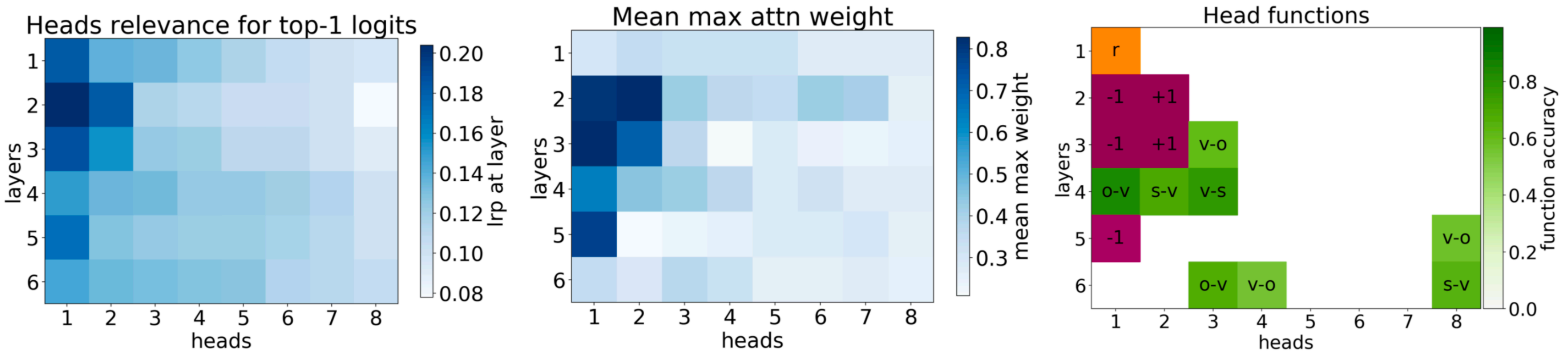
(a) LRP (b) confidence (c) head functions

Attention heads classification. Positional and rare words

- Head is “positional” if at least 90% of the time its maximum attention weight is assigned to a specific relative position(i.e. ±1)
- Head is “rare” if it points at the rarest word in a sentence more than in 50% of cases



Attention maps of the rare words head

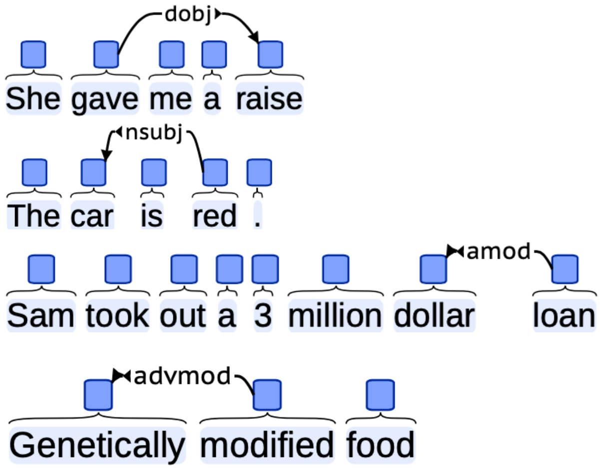


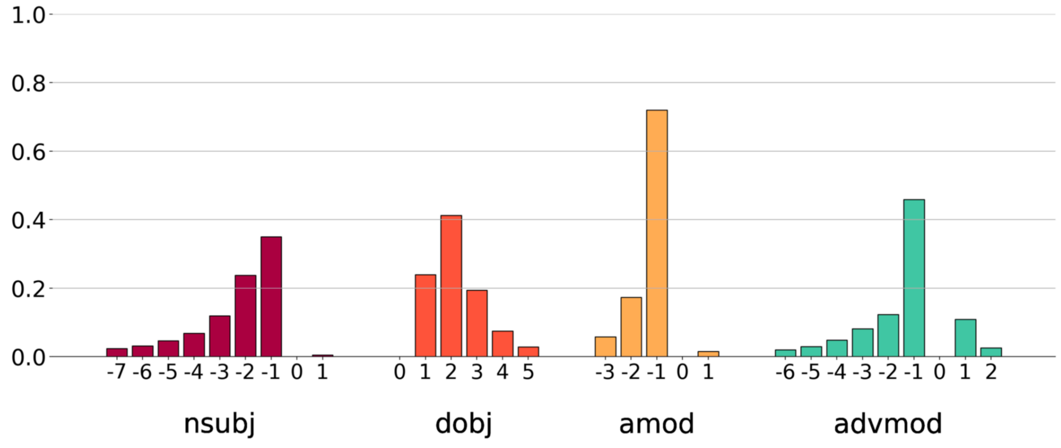
(a) LRP (b) confidence (c) head functions

Attention heads classification. Syntactic heads

Analyzed dependency relations:

- nominal subject (nsubj)
- direct object (dobj)
- adjectival modifier (amod)
- adverbial modifier (advmod)

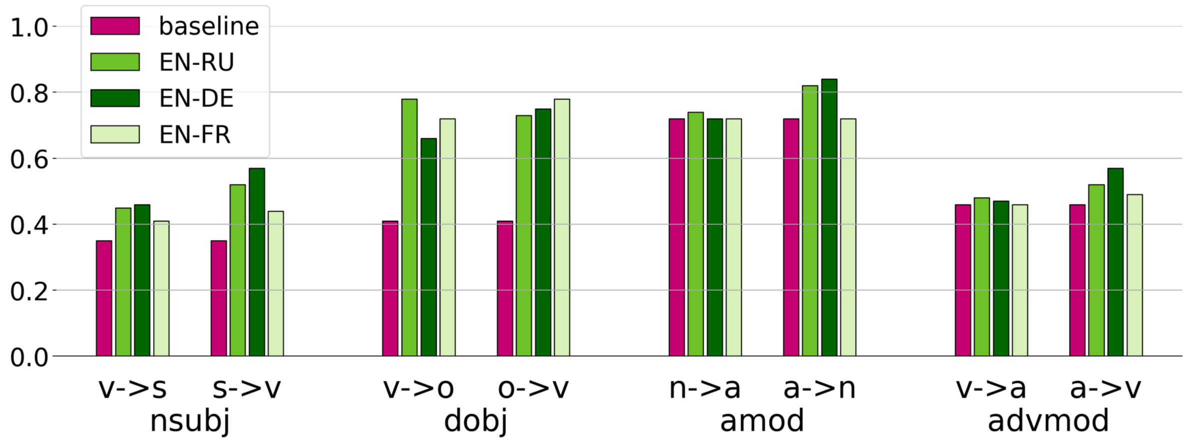




"Accuracy" of head = how often it assigns its maximum attention weight (excluding EOS) to a token with which it is in one of the aforementioned dependency relations

Attention heads classification.
Syntactic heads

Head is "syntactic" if its accuracy is at least 10% higher than the baseline that looks at the most frequent relative position for this dependency relation.



dep.	direction	best head / baseline accuracy	
		WMT	**OpenSubtitles**
nsubj			
	v → s	45 / 35	77 / 45
	s → v	52 / 35	70 / 45
dobj			
	v → o	78 / 41	61 / 46
	o → v	73 / 41	84 / 46
amod			
	noun → adj.m.	74 / 72	81 / 80
	adj.m. → noun	82 / 72	81 / 80
advmod			
	v → adv.m.	48 / 46	38 / 33
	adv.m. → v	52 / 46	42 / 33

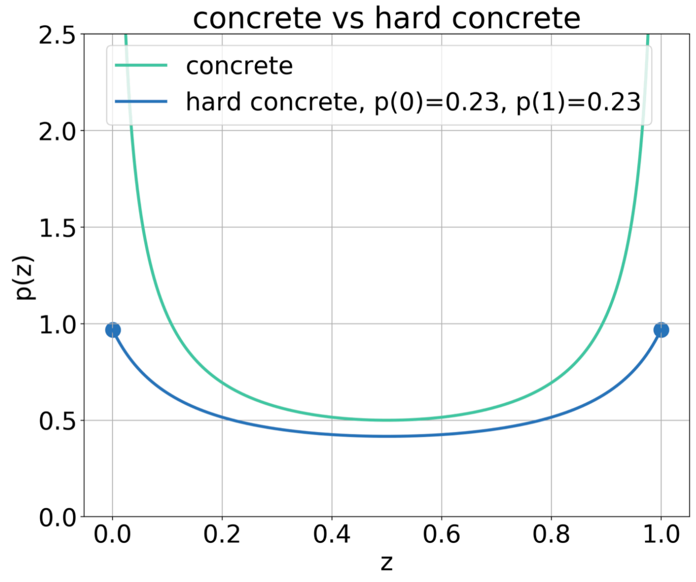
Pruning attention heads

$$\text{MultiHead}(Q,K,V) = \text{Concat}_i(g_i \cdot \text{head}_i)W^O$$

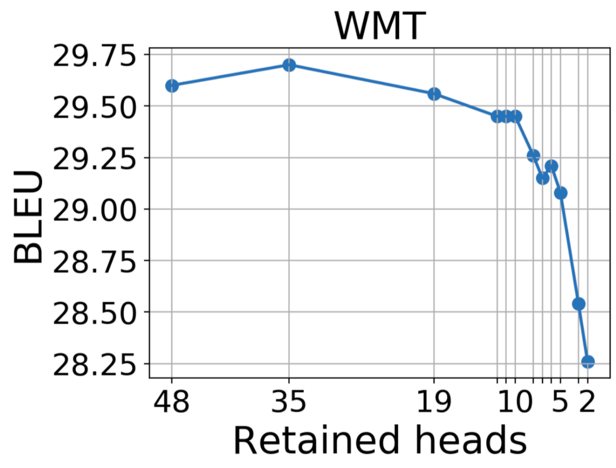
$$L_0(g_1,\ldots,g_h) = \sum_{i=1}^{h}(1 - [[g_i = 0]]) \longrightarrow L_C(\phi) = \sum_{i=1}^{h}(1 - P(g_i = 0|\phi_i)) \quad g_i \sim HardConcrete(\phi_i)$$

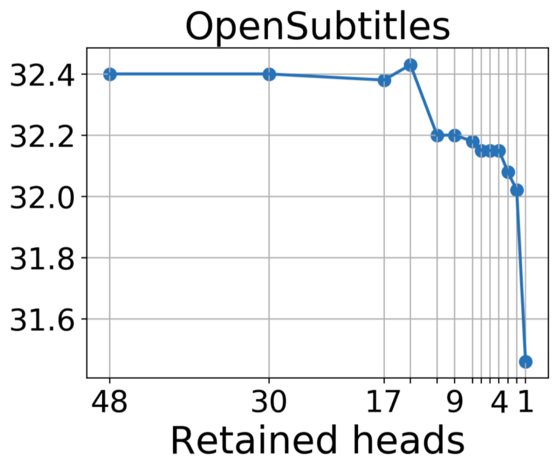
Loss:

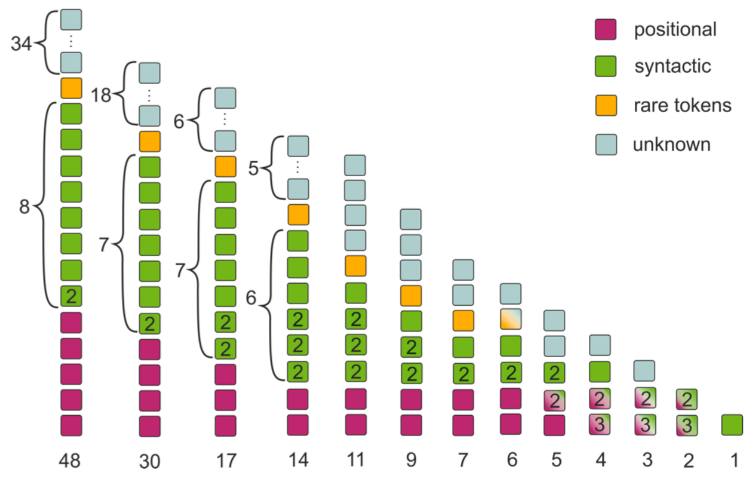
$$L(\theta,\phi) = L_{xent}(\theta,\phi) + \lambda L_C(\phi)$$



Pruning attention heads. Results

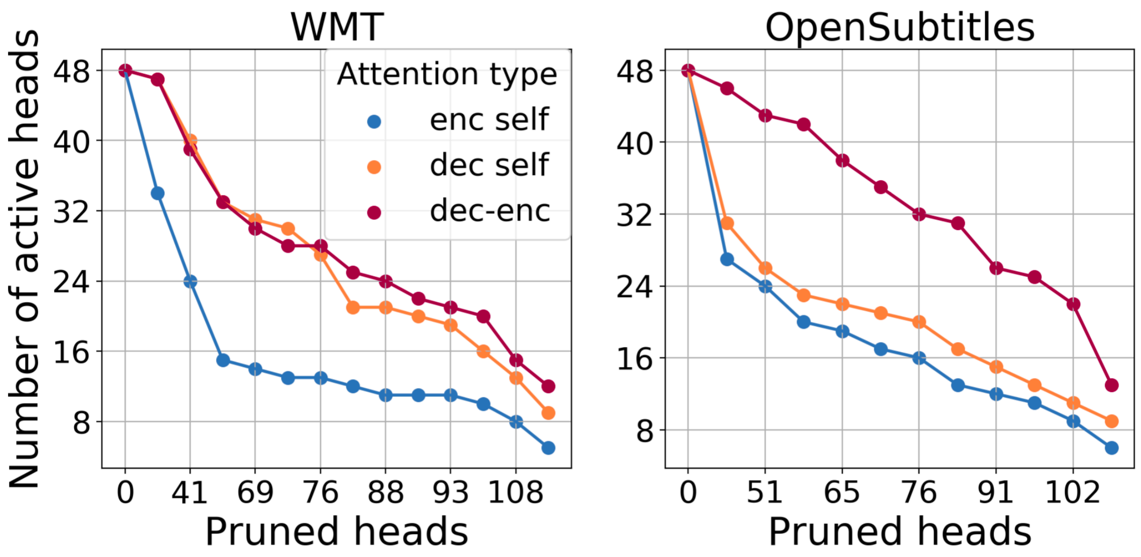






Pruning attention heads. Results

	attention heads (e/d/d-e)	BLEU from trained	BLEU from scratch
WMT, 2.5m			
baseline	48/48/48	**29.6**	
sparse heads	14/31/30	29.62	29.47
	12/21/25	29.36	28.95
	8/13/15	29.06	28.56
	5/9/12	28.90	28.41
OpenSubtitles, 6m			
baseline	48/48/48	**32.4**	
sparse heads	27/31/46	32.24	32.23
	13/17/31	32.23	31.98
	6/9/13	32.27	31.84



Results

- Found consistent roles played by attention heads
- Investigated effective pruning method without model's quality loss
- Only a small subset of heads ap- pear to be important for the translation task

References

- https://arxiv.org/abs/1905.09418
- https://arxiv.org/abs/1706.03762
- http://jalammar.github.io/illustrated-transformer/