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

ACL 2019

Background

- Transformer (Vaswani et al,. 2017) is a leading modelling paradigm in neural machine translation.
- Transformer follows an encoder-decoder framework comprising stacked **multi-head self-attention layers** and **fully-connected layers**;
  - Transformer-base: 6 layers each side and 8 heads per layer (144 heads in total);
  - Transformer-big: 6 layers each side and 16 heads per layer (288 heads in total);
- Multi-head mechanism is demonstrated to be able to improve model capacity in comparison to single-head attention:
  - Single-head attention is 0.9 BLEU score worse than the 8-head attention model (25.8 BLEU).

Questions

- Which heads are the most important to translation quality?

- Do individual attention heads play consistent and interpretable roles?

- Can we significantly reduce the number of attention heads while preserving translation quality?

Transformer Architecture

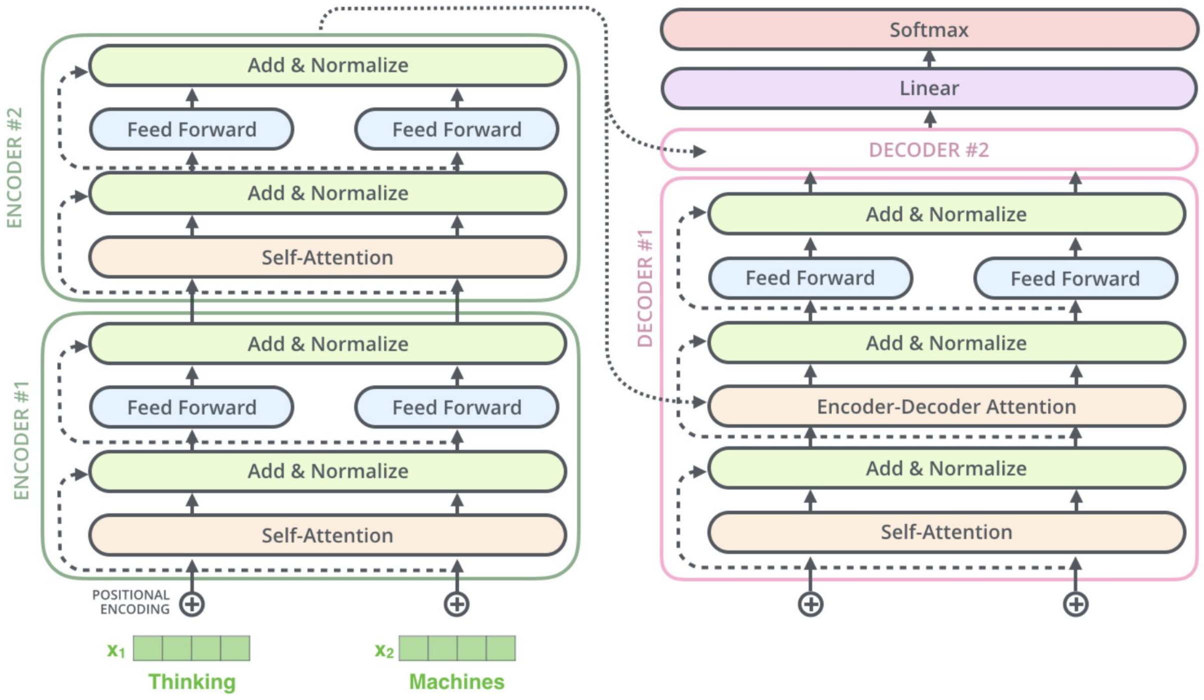


Figure from http://jalammar.github.io/illustrated-transformer/

Transformer Architecture

- Attention Computation

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, V are parameter matrices, d_k is the dimensionality of K.

- Multi-head Attention

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

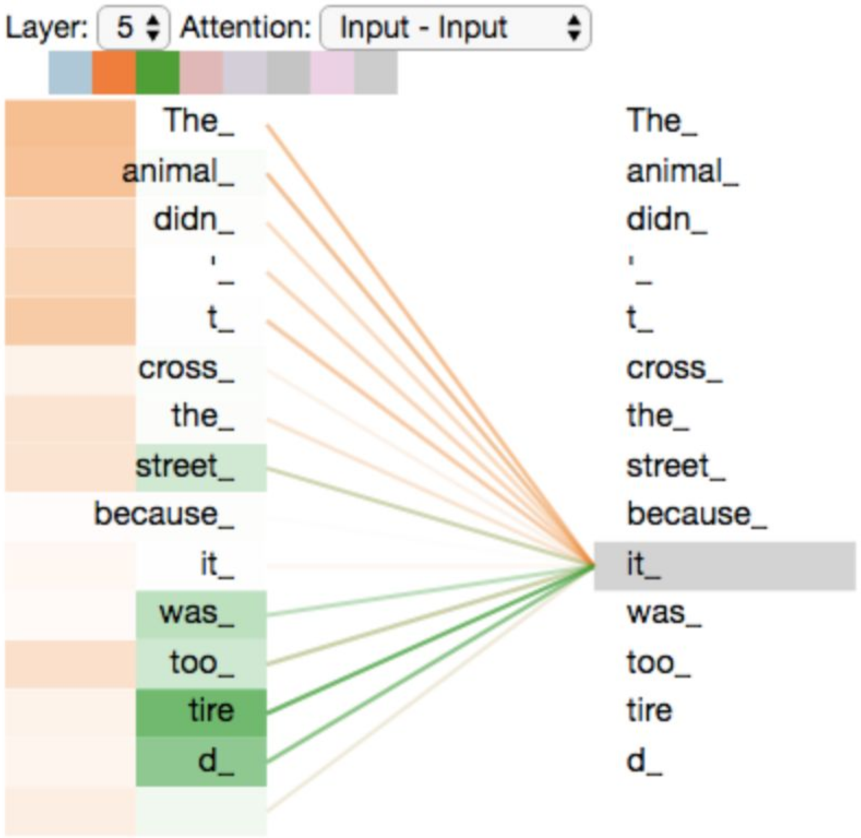
$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}_i(\mathrm{head}_i)W^O$$

where W_i and W^O are parameter matrices.

Transformer Architecture

Input sentence:
“The animal didn't cross the street because it was too tired”.

➔ What does “it” refer to?



Dataset Setting

- Task: machine translation
- Source language: English
- Target language: Russian, German, French
- Dataset: WMT (2.5M sentence pairs), OpenSubtitles2018corpus (6M sentence pairs)

Q1: Identify Important Heads

Metrics for Importance Measure:

- **confidence**: "confidence" of a head as the average of its maximum attention weight excluding the end of sentence symbol ("EOS"), where average is taken over tokens in a set of sentences used for evaluation (development set).

➔ A confident head is one that usually assigns a high proportion of its attention to a single token.

➔ Intuitively, we might expect confident heads to be important to the translation task.

Q1: Identify Important Heads

Metrics for Importance Measure:

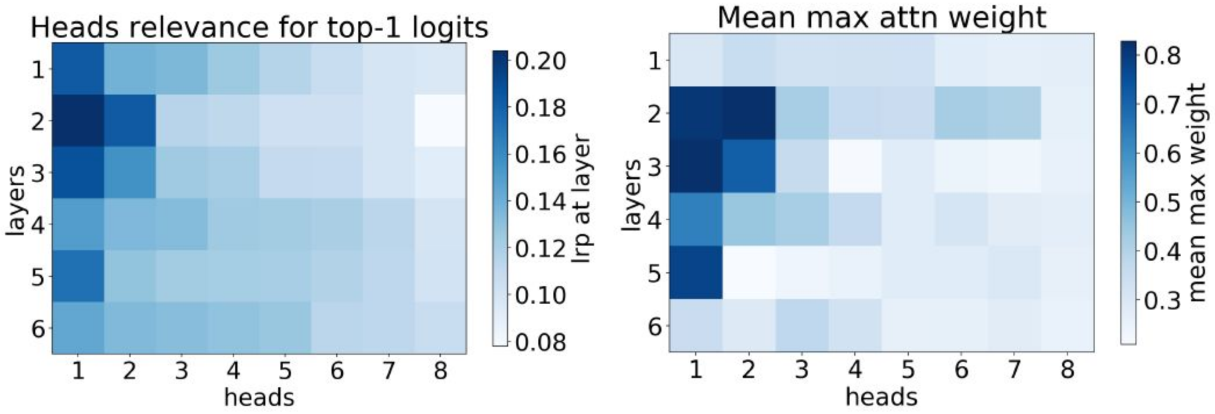
- **Layer-wise relevance propagation (LRP)**: a method for computing the relative contribution of neurons at one point in a network to neurons at another. (Ding et al., 2017)

➔ General idea: neurons in (L+1)-th layer are fully determined by neurons in L-th layer and the connected weights; thus, we can compute the "contribution" of a specific neuron to the outputs by back-propagating predictions.

➔ Here we propose to use LRP to evaluate the degree to which different heads at each layer contribute to the top-1 logit predicted by the model.

Q1: Identify Important Heads

Experiment Results:



(a) LRP (b) confidence

Model trained on 6m OpenSubtitles EN-RU data

Q2: Characterize Heads

Possible Head Roles

- **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.

Q2: Characterize Heads

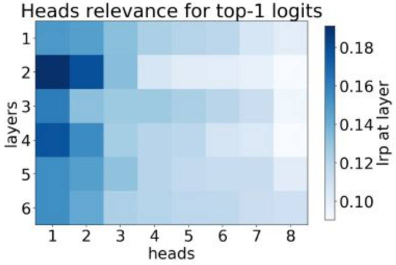
Positional Heads

- A head is “positional” if at least 90% of the time its maximum attention weight is assigned to a specific relative position (in practice either -1 or +1, i.e. attention to adjacent tokens).

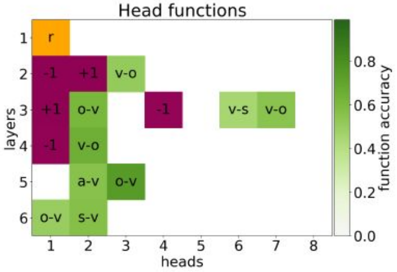
Q2: Characterize Heads

Positional Heads

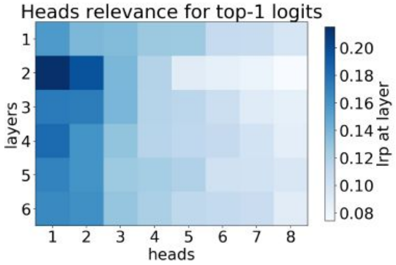
(Models trained on WMT)



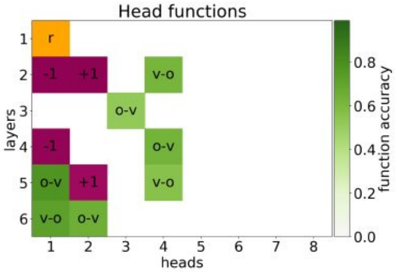
(a) LRP (EN-DE)



(b) head functions



(c) LRP (EN-FR)



(d) head functions

Q2: Characterize Heads

Syntactic Heads

- **Syntactic relations:**
  a. Nominal subject (**nsubj**), <noun, verb>, e.g., "**Clinton defeated** Dole."
  b. Direct object (**dobj**), <verb, object>, e.g., "She **gave** me a **raise**."
  c. Adjectival modifier (**amod**), <adj.m., noun>, e.g., "Sam eats **red meat**."
  d. Adverbial modifier (**advmod**), <verb, adv.m.>, e.g., "**Genetically modified** food."

➔ Ground truth is generated by CoreNLP (Manning et al, 2014).
➔ We calculate for each 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. -> defined as the "**head accuracy**".

Q2: Characterize Heads

Syntactic Heads

- Observation: Many dependency relations are frequently observed in specific relative positions (for example, often they hold between adjacent tokens)

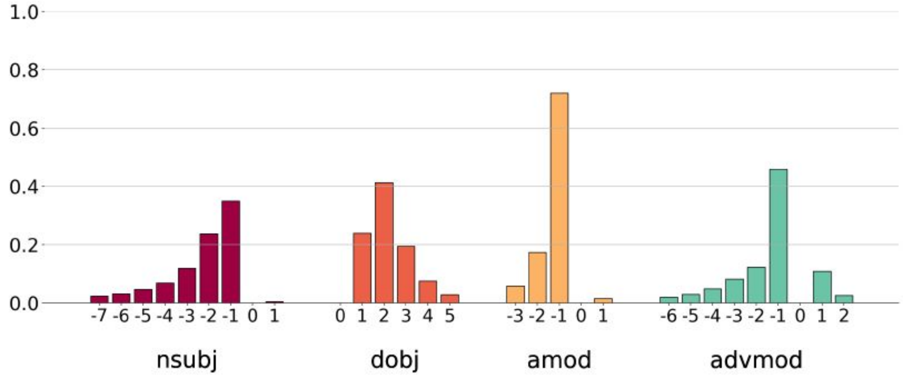


Figure 3: Distribution of the relative position of dependent for different dependency relations (WMT).

Q2: Characterize Heads

Syntactic Heads

- **Baseline**: looks at the most frequent relative position for a given dependency relation.
- **Syntactic Head**: a head is “syntactic” if its accuracy is at least **10%** higher than the baseline.

Q2: Characterize Heads

Syntactic Heads

dep.	direction	best head / baseline accuracy WMT	best head / baseline accuracy 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

Table 1: Dependency scores for EN-RU, comparing the best self-attention head to a positional baseline. Models trained on 2.5m WMT data and 6m OpenSubtitles data.

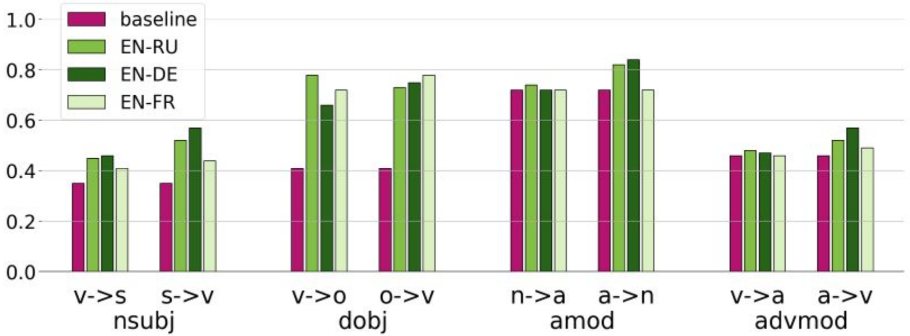
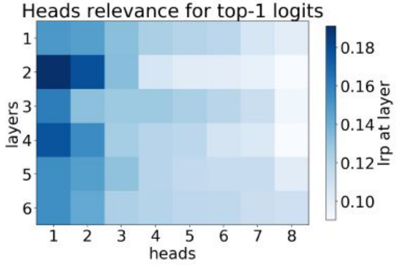


Figure 4: Dependency scores for EN-RU, EN-DE, EN-FR each trained on 2.5m WMT data.

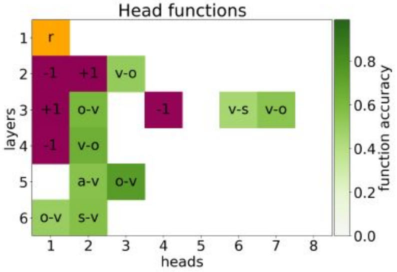
Q2: Characterize Heads

Syntactic Heads

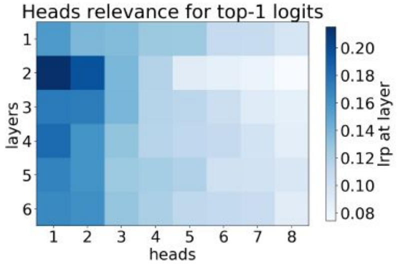
(Models trained on WMT)



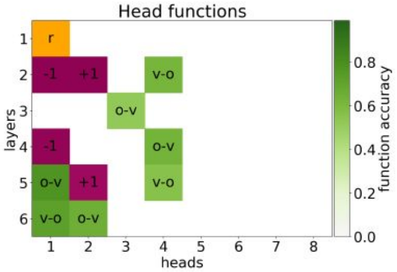
(a) LRP (EN-DE)



(b) head functions



(c) LRP (EN-FR)



(d) head functions

Q2: Characterize Heads

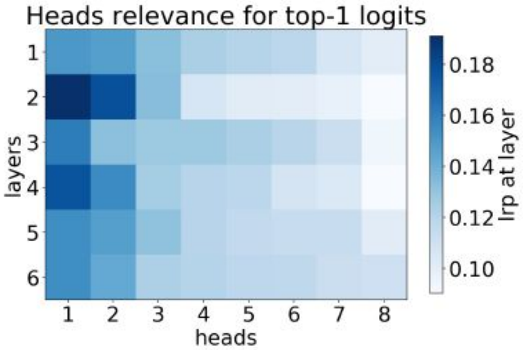
Rare Words Heads

- A head points to the least frequent tokens in a sentence.
- Typically, the most important head in the first layer is a Rare Words Head.
    - In models trained on openSubtiltes, this head points to one of the two least frequent tokens in **83%** of cases.
    - In models trained on WMT, this head points to one of the two least frequent tokens in **more than 50%** of cases.

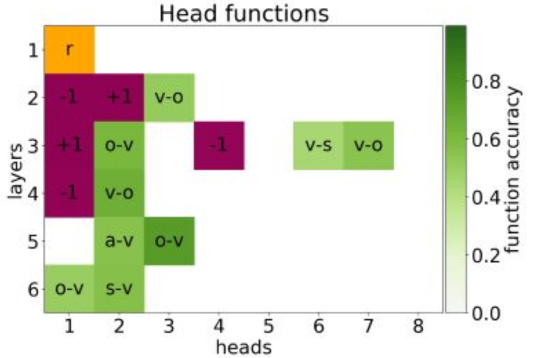
Q2: Characterize

Rare Words Heads

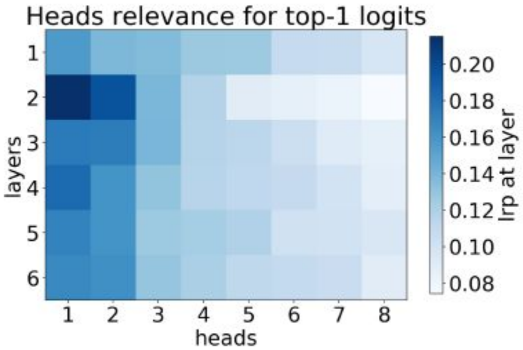
(Models trained on WM



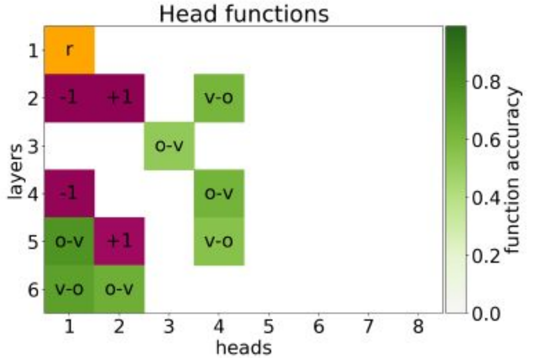
(a) LRP (EN-DE)



(b) head functions



(c) LRP (EN-FR)



(d) head functions

Q3: Prune Attention Heads

Methodology

- Regularization pruning: prune attention heads by adding regularization terms in the loss function.
- Gate variable g_i in [0, 1] for each attention head,

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

- L0 regularization:

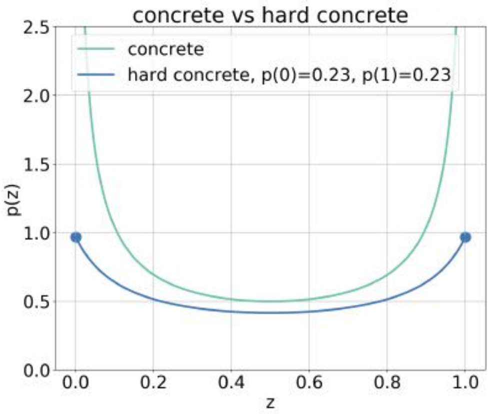
$$L_0(g_1, \ldots, g_h) = \sum_{i=1}^{h}(1 - [[g_i = 0]])$$

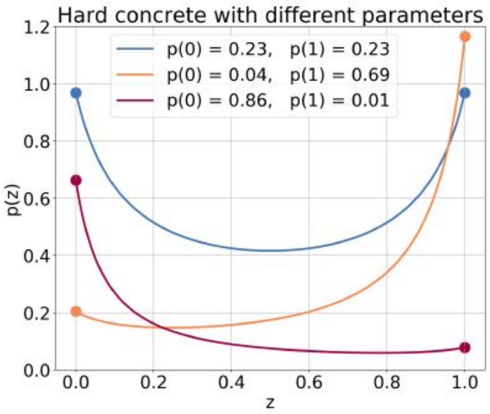
➔ **Non-differentiable!**

Q3: Prune Attention Heads

Methodology

- A stochastic relaxation: each g_i is independently drawn from a head-specific distribution, which is controlled by a learnable parameter phi_i.
- Hard Concrete Distribution (Louizos et al., 2018): a parameterized family of mixed discrete-continuous distributions over the closed interval [0, 1].





Q3: Prune Attention Heads

Methodology

- Relaxed L0 norm:

$$L_C(\phi) = \sum_{i=1}^{h} (1 - P(g_i = 0|\phi_i))$$

- The new objective function:

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

  where theta are network parameters, L_xent is cross-entropy loss.

- Fine-tuning from a converged model trained without L_C. By varying coefficient lambda, we obtain models of different retained attention heads.

Q3: Prune Attention Heads

- **Prune encoder heads only**.

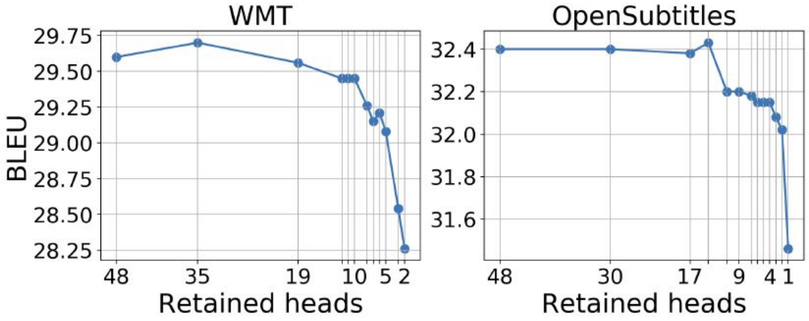


Figure 7: BLEU score as a function of number of retained encoder heads (EN-RU). Regularization applied by fine-tuning trained model.

➔ For OpenSubtitles, we lose only 0.25 BLEU when we prune all but 4 heads out of 48;
➔ For WMT, 10 heads in the encoder are sufficient to stay within 0.15 BLEU of the full model.

Q3: Prune Attention Heads

- **Prune encoder heads only**.

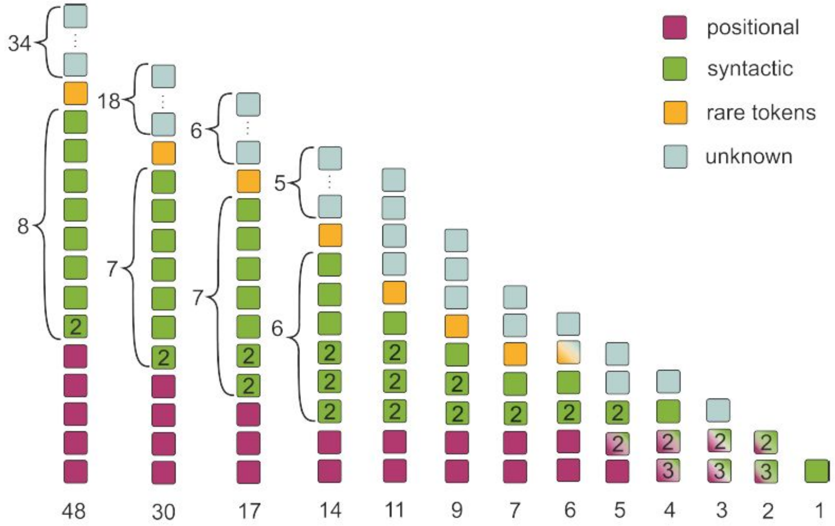


Figure 8: Functions of encoder heads retained after pruning. Each column represents all remaining heads after varying amount of pruning (EN-RU; Subtitles).

Q3: Prune Attention Heads

- **Prune all types of attention heads**.

(EN-RU translation task)

	attention heads (e/d/d-e)	**BLEU** from trained	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

Q3: Prune Attention Heads

- **Heads Importance for Different Attention Types**

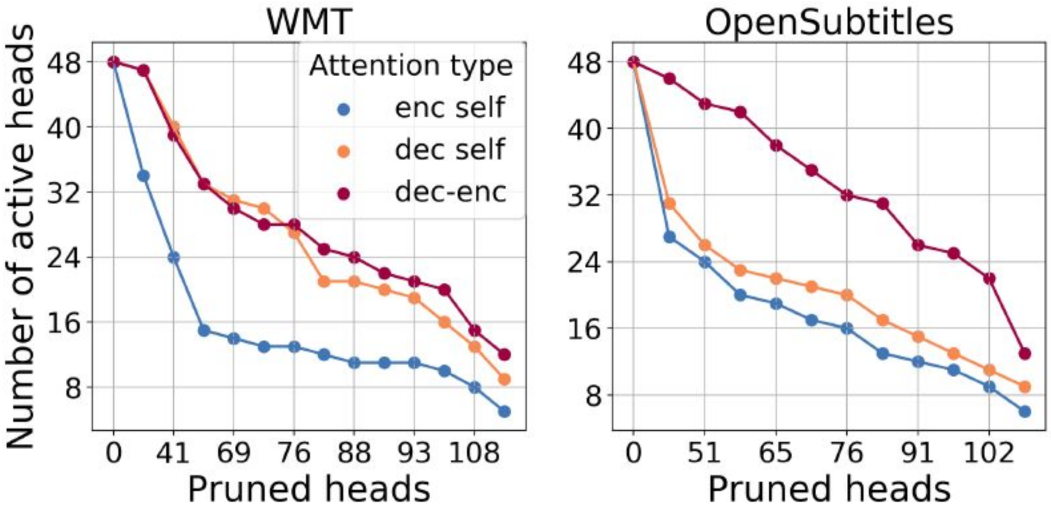


Figure 9: Number of active heads of different attention type for models with different sparsity rate

Q3: Prune Attention Heads

- **Heads Importance for Different Layers**

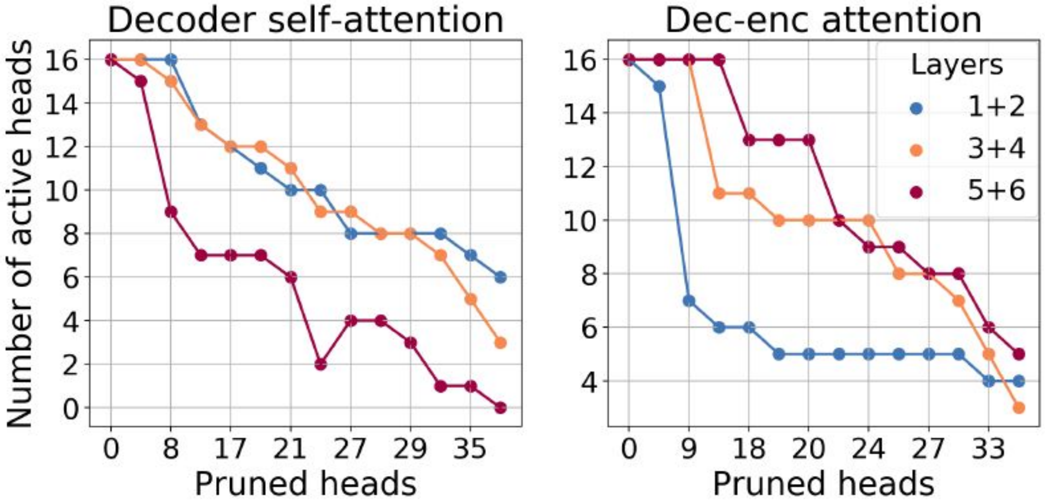


Figure 10: Number of active heads in different layers of the decoder for models with different sparsity rate (EN-RU, WMT)

Q&A