

# Attention Semantics

What **attention heads** actually know and why should we care

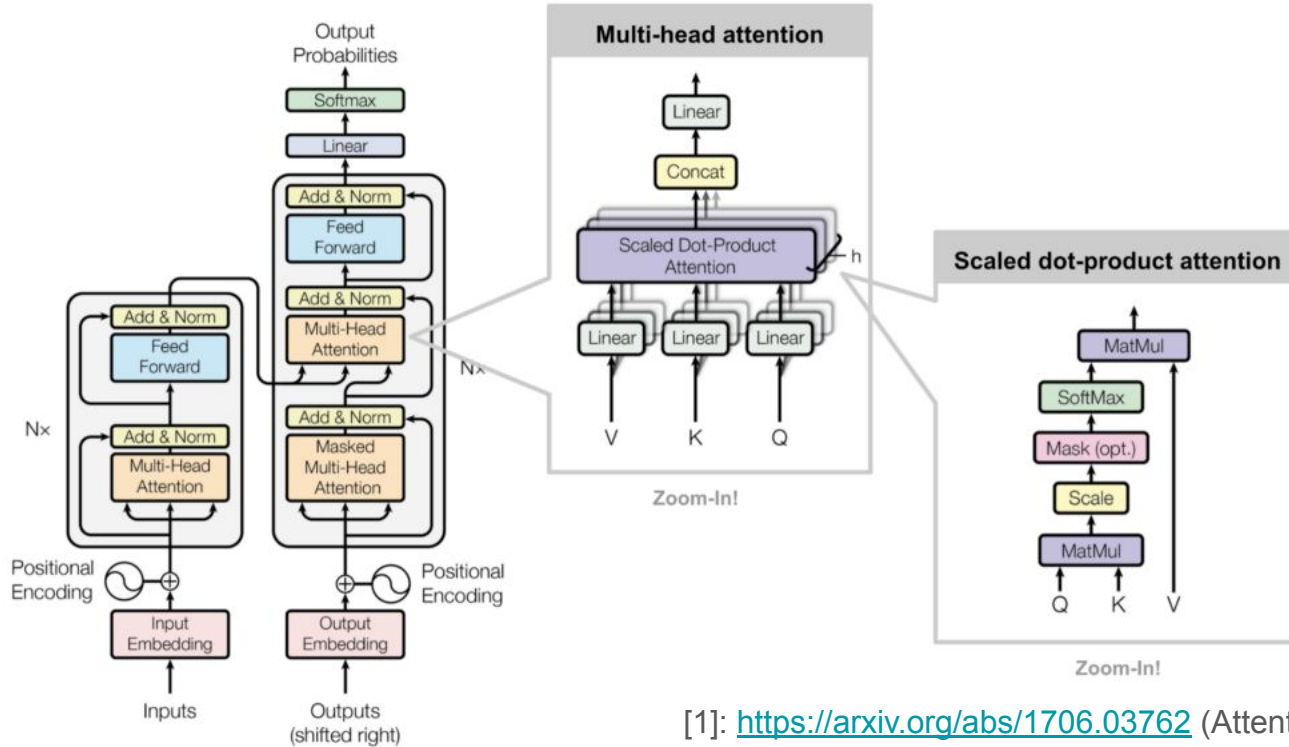


MUNI  
FI



FI:PV212: Readings in Digital ...  
Michal Štefánik  
[stefanik.m@mail.muni.cz](mailto:stefanik.m@mail.muni.cz)

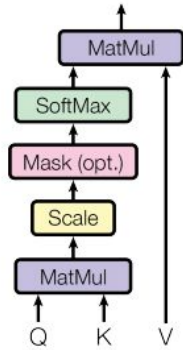
# Transformer [1]



[1]: <https://arxiv.org/abs/1706.03762> (Attention is All You Need)

# Attention [1]

Scaled Dot-Product Attention



Multi-Head Attention

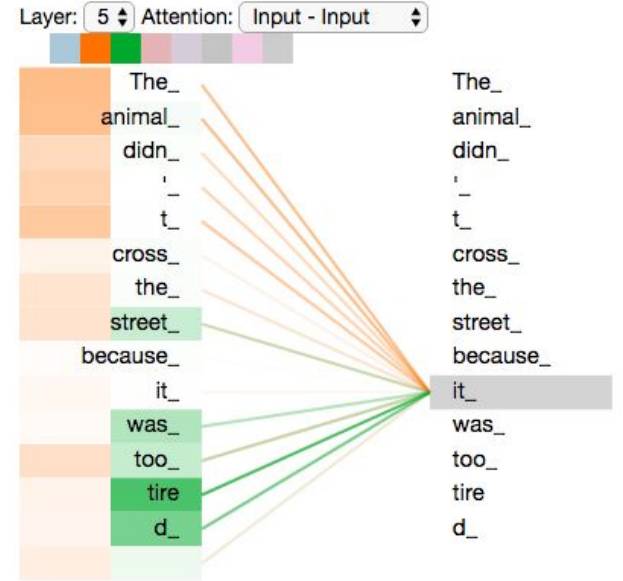
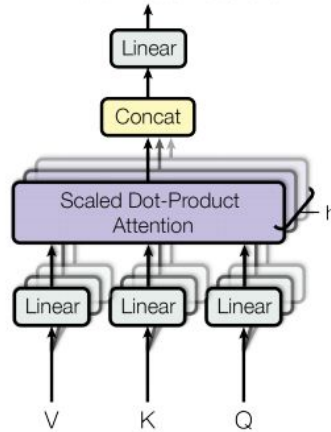
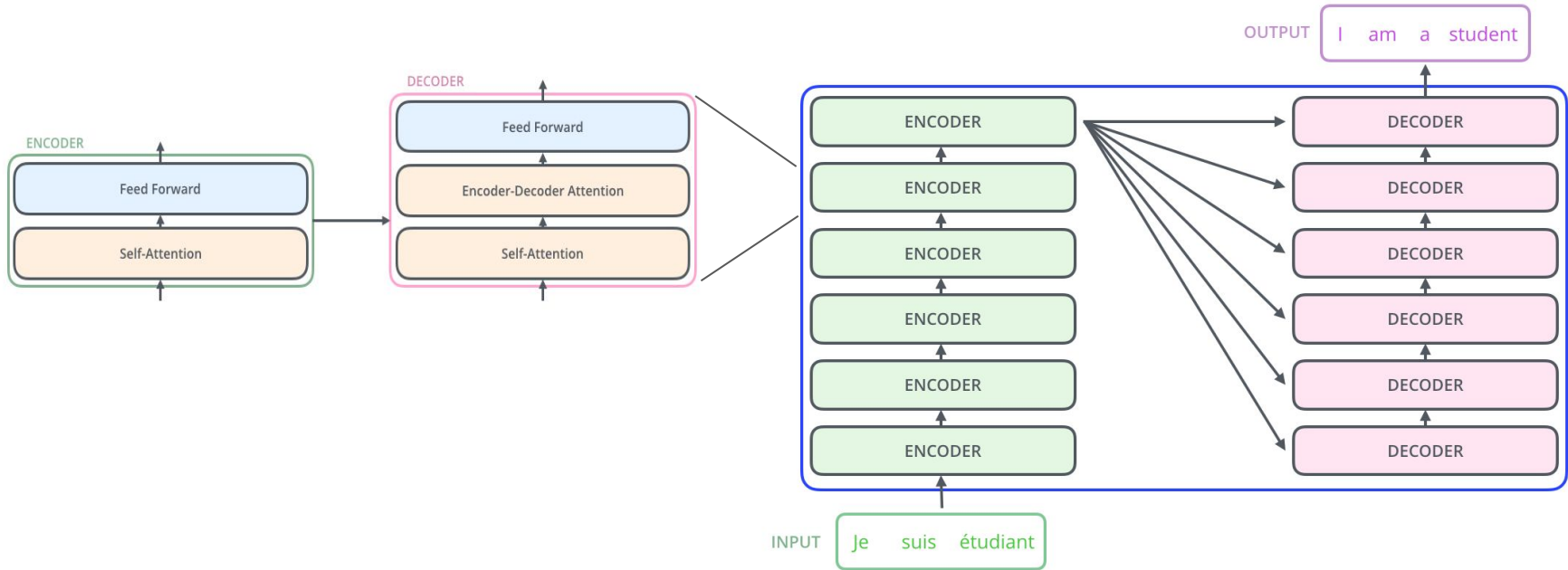


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

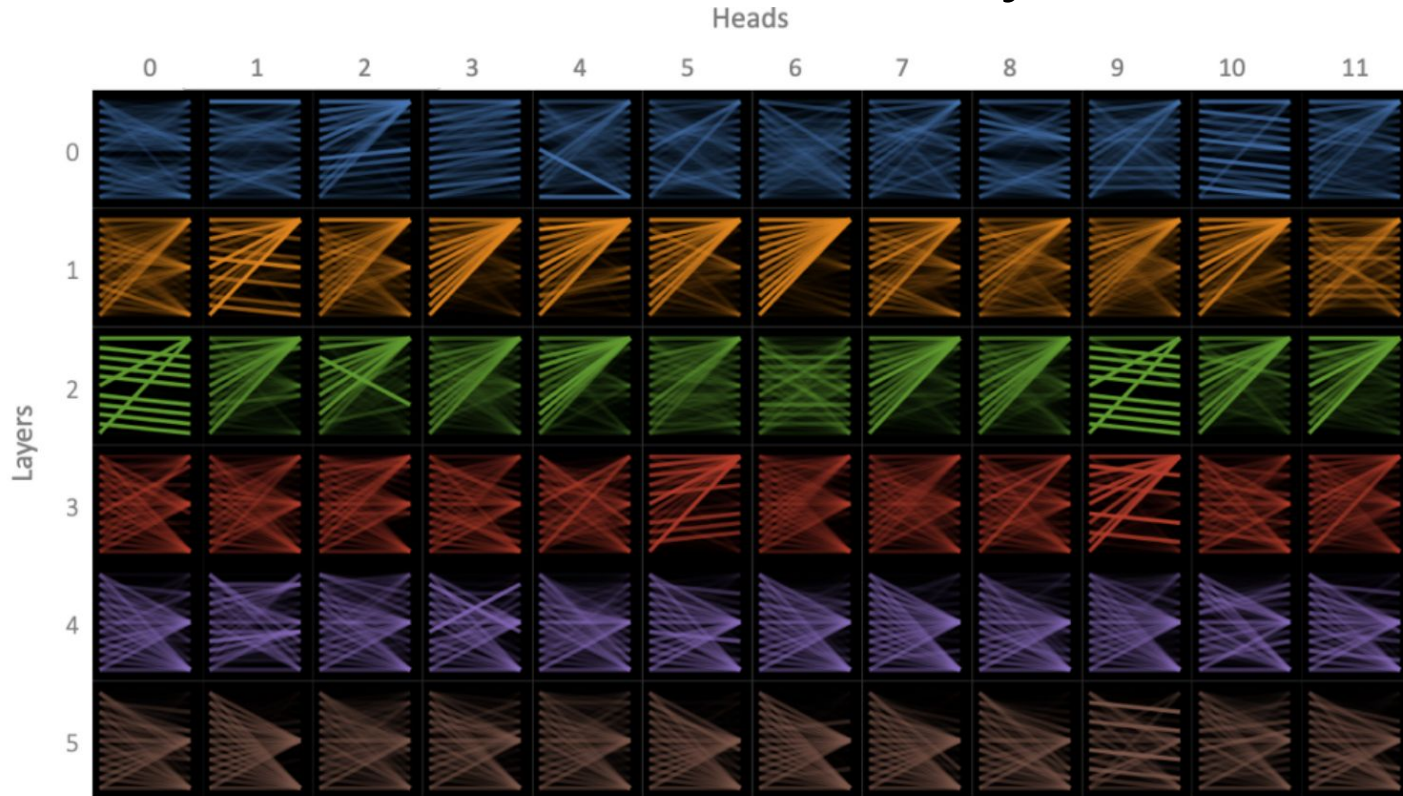
[1]: <https://arxiv.org/abs/1706.03762> (Attention is All You Need)

[3]: [https://colab.research.google.com/github/tensorflow/tensor2tensor/blob/master/tensor2tensor/notebooks/hello\\_t2t.ipynb](https://colab.research.google.com/github/tensorflow/tensor2tensor/blob/master/tensor2tensor/notebooks/hello_t2t.ipynb)

# Transformer as autoencoder



# Attention heads layout



Model view (first 6 layers) for input sentences “the rabbit quickly hopped” and “the turtle slowly crawled”.

# Specific heads semantics [2]

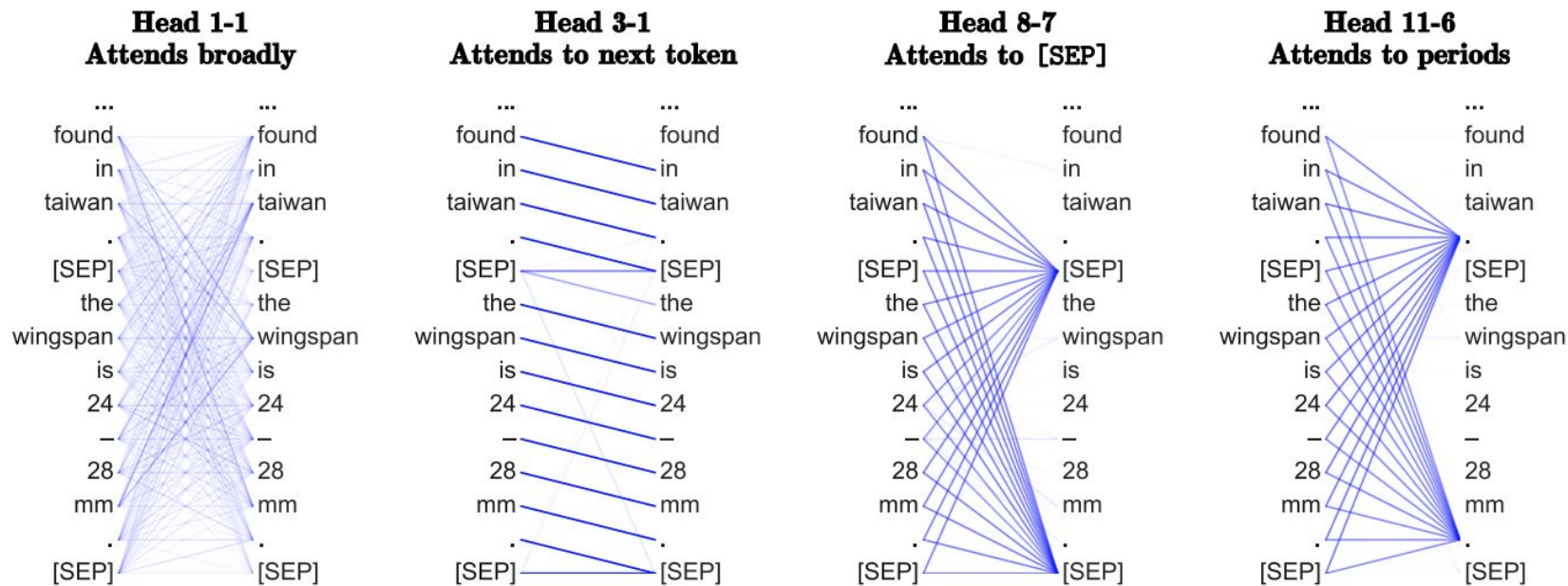


Figure 1: Examples of heads exhibiting the patterns discussed in Section 3. The darkness of a line indicates the strength of the attention weight (some attention weights are so low they are invisible).

[2]: <https://arxiv.org/pdf/1906.04341.pdf> (What Does BERT Look At?)

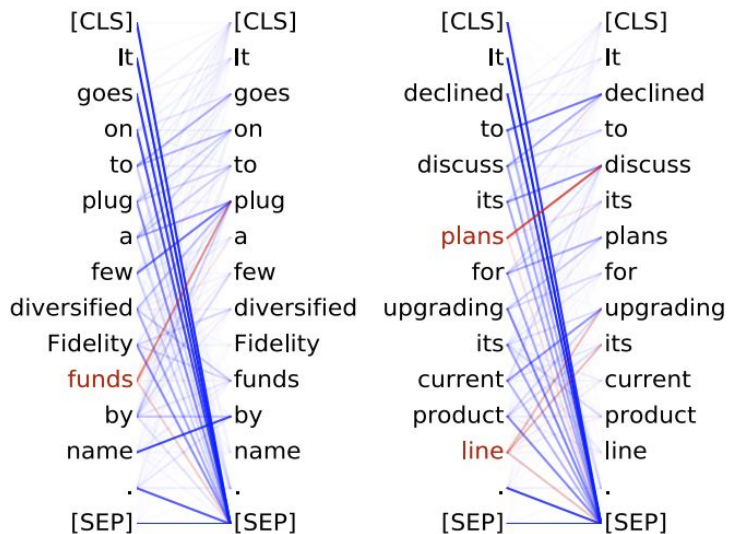
# Specific heads semantics [2]

- Many elementary patterns, “No-Op” attention to [SEP] (?)
  - *“Four attention heads (in layers 2, 4, 7, and 8) on average put >50% of their attention on the previous token and five attention heads (in layers 1, 2, 2, 3, and 6) put >50% of their attention on the next token.”*
- Transitive information propagation is beneficial [3], but we can not see any other heads later attending to [SEP], dots, of commas
  - *“Attention heads processing [SEP] almost entirely (more than 90%) attend to themselves and the other [SEP] token.”*
  - *“(…) the gradients for attention to [SEP] become very small. (…) attending more or less to [SEP] does not substantially change BERT’s outputs.”*

# Specific heads semantics [2]

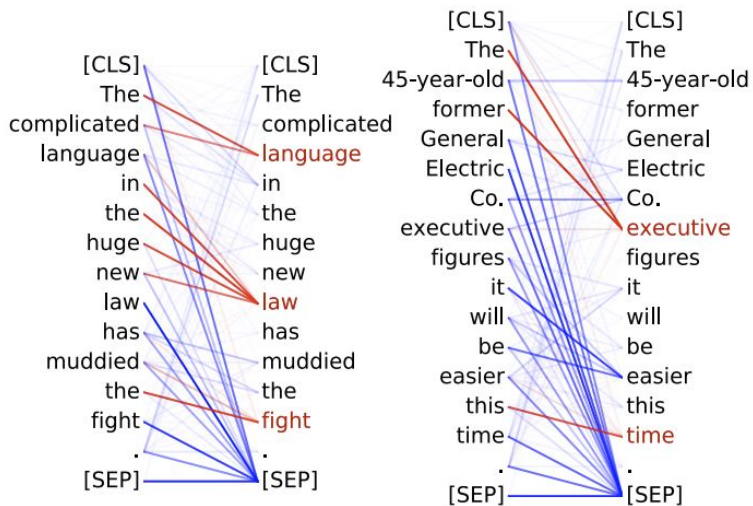
## Head 8-10

- **Direct objects** attend to their verbs
- 86.8% accuracy at the dobj relation



## Head 8-11

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the det relation

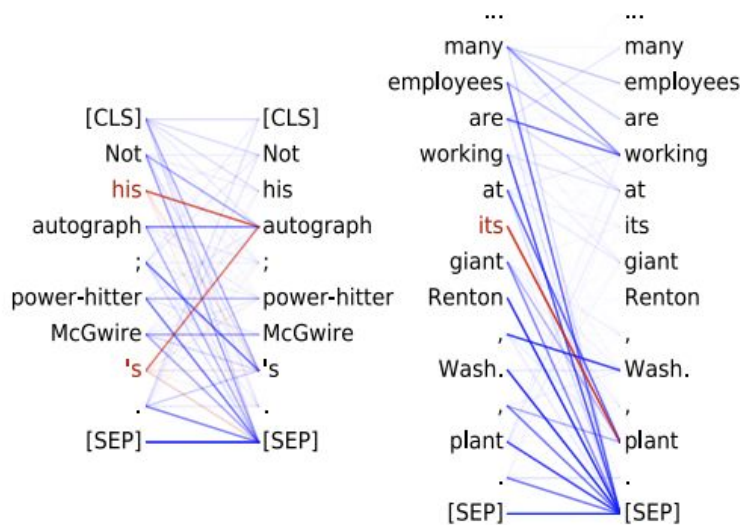




# Specific heads semantics [2]

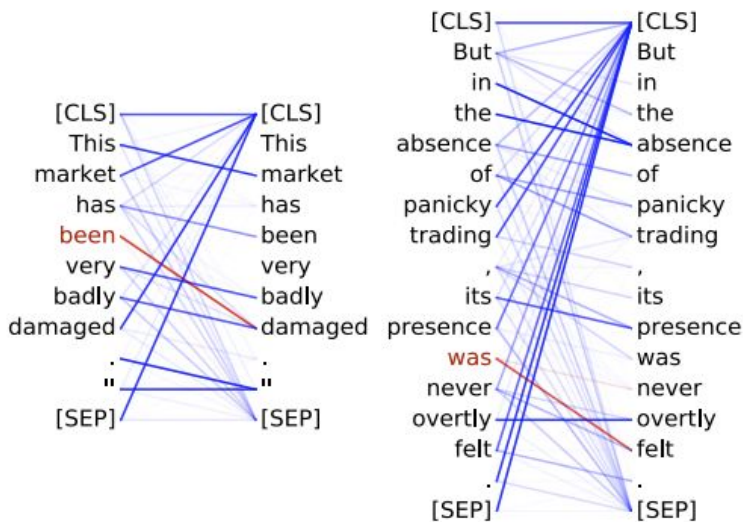
## Head 7-6

- **Possessive pronouns** and apostrophes attend to the head of the corresponding NP
- 80.5% accuracy at the poss relation



## Head 4-10

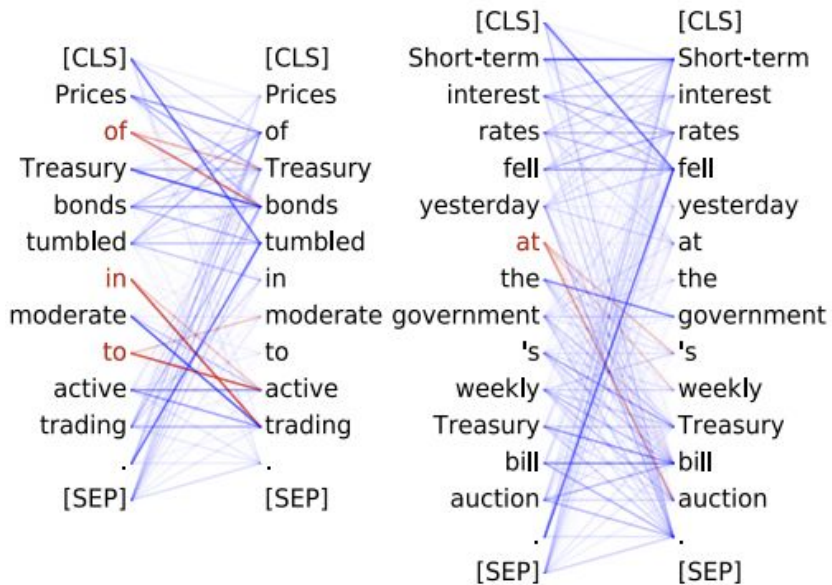
- **Passive auxiliary verbs** attend to the verb they modify
- 82.5% accuracy at the auxpass relation



# Specific heads semantics [2]

## Head 9-6

- **Prepositions** attend to their objects
- 76.3% accuracy at the pobj relation



## Head 5-4

- **Coreferent** mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent



# Syntactic heads [2]

- No “syntactic” heads
- But syntactic properties are decomposed to simpler tasks!

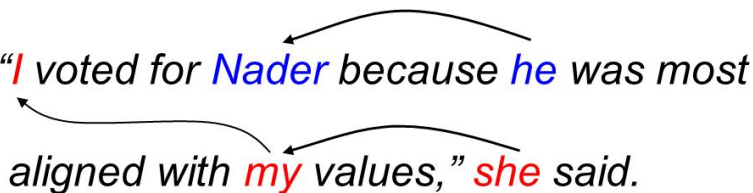
| Relation | Head | Accuracy    | Baseline  |
|----------|------|-------------|-----------|
| All      | 7-6  | 34.5        | 26.3 (1)  |
| prep     | 7-4  | 66.7        | 61.8 (-1) |
| pobj     | 9-6  | <b>76.3</b> | 34.6 (-2) |
| det      | 8-11 | <b>94.3</b> | 51.7 (1)  |
| nn       | 4-10 | 70.4        | 70.2 (1)  |
| nsubj    | 8-2  | 58.5        | 45.5 (1)  |
| amod     | 4-10 | 75.6        | 68.3 (1)  |
| dobj     | 8-10 | <b>86.8</b> | 40.0 (-2) |
| advmod   | 7-6  | 48.8        | 40.2 (1)  |
| aux      | 4-10 | 81.1        | 71.5 (1)  |
| poss     | 7-6  | <b>80.5</b> | 47.7 (1)  |
| auxpass  | 4-10 | <b>82.5</b> | 40.5 (1)  |
| ccomp    | 8-1  | <b>48.8</b> | 12.4 (-2) |
| mark     | 8-2  | <b>50.7</b> | 14.5 (2)  |
| prt      | 6-7  | <b>99.1</b> | 91.4 (-1) |

Table 1: The best performing attentions heads of BERT on WSJ dependency parsing by dependency type. Numbers after baseline accuracies show the best offset found (e.g., (1) means the word to the right is predicted as the head). We show the 10 most common relations as well as 5 other ones attention heads do well on. Bold highlights particularly effective heads.

# Coreference heads [2]

- “(...) *what percent of the time does the head word of a coreferent mention most attend to the head of one of that mention’s antecedents.*”
- Coreference (semantic task) is also resolved by particular heads

*“I voted for Nader because he was most aligned with my values,” she said.*



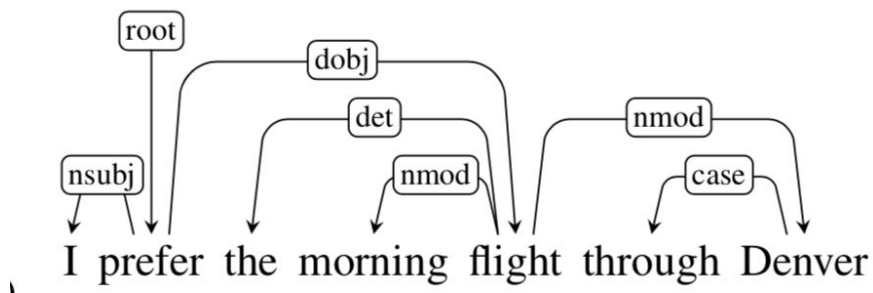
| Model        | All | Pronoun | Proper | Nominal |
|--------------|-----|---------|--------|---------|
| Nearest      | 27  | 29      | 29     | 19      |
| Head match   | 52  | 47      | 67     | 40      |
| Rule-based   | 69  | 70      | 77     | 60      |
| Neural coref | 83* | –       | –      | –       |
| Head 5-4     | 65  | 64      | 73     | 58      |

\*Only roughly comparable because on non-truncated documents and with different mention detection.

Table 2: Accuracies (%) for systems at selecting a correct antecedent given a coreferent mention in the CoNLL-2012 data. One of BERT’s attention heads performs fairly well at coreference.

# Dependency parsing groups of heads [2]

- Prediction of antecedents (heads) for each token
- “(...) linear combination of (all) attention weights.”
- “there is not much more syntactic information in BERT’s vector representations compared to its attention maps.”



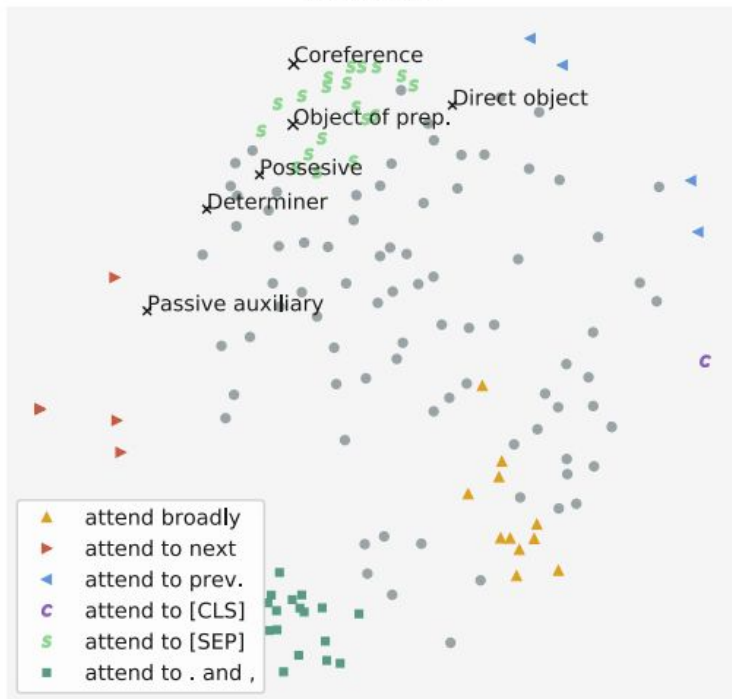
| Model                    | UAS     |
|--------------------------|---------|
| Structural probe         | 80 UAS* |
| Right-branching          | 26      |
| Distances + GloVe        | 58      |
| Random Init Attn + GloVe | 30      |
| Attn                     | 61      |
| Attn + GloVe             | 77      |

Table 3: Results of attention-based probing classifiers on dependency parsing. A simple model taking BERT attention maps and GloVe embeddings as input performs quite well. \*Not directly comparable to our numbers; see text.

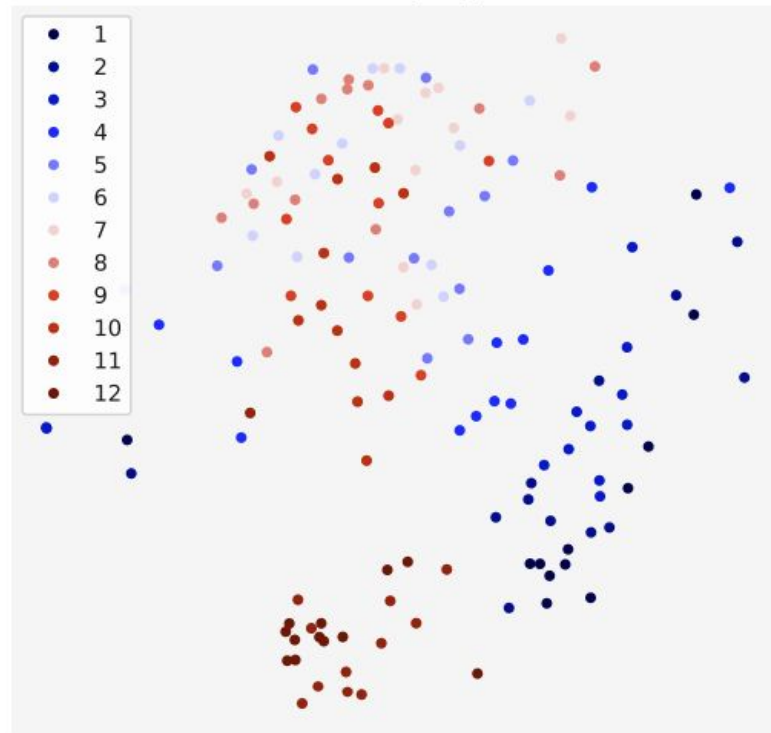
# Heads overview [2]

Embedded BERT attention heads

Behaviors



Colored by Layer



# Layers semantics: Probing [3]

## Scalar Mixing Weights

(...) for each task we introduce scalar parameters  $\gamma_\tau$  and  $a_\tau(0), a_\tau(1), \dots, a_\tau(L)$ , and let:

$$\mathbf{h}_{i,\tau} = \gamma_\tau \sum_{\ell=0}^L s_\tau^{(\ell)} \mathbf{h}_i^{(\ell)} \quad (1)$$

where  $s_\tau = \text{softmax}(a_\tau)$ . We learn these weights jointly with the probing classifier  $P_\tau$ , in order to allow it to extract information from the many layers of an encoder (...) we extract the learned coefficients in order to estimate the contribution of different layers to that particular task

## Cumulative Scoring

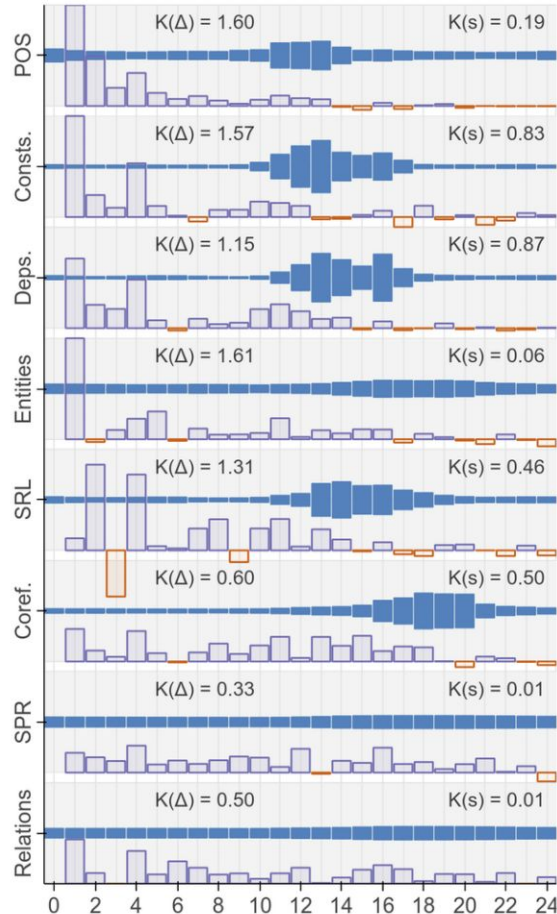
(...) we train a series of classifiers  $\{P_\tau(L)\}_L$  which use scalar mixing to attend to layer  $L$  as well as *all previous layers*.

We can then compute a differential score  $\Delta_\tau^{(\ell)}$ , which measures how much better we do on the probing task if we observe one additional encoder layer  $\ell$ :

$$\Delta_\tau^{(\ell)} = \text{Score}(P_\tau^{(\ell)}) - \text{Score}(P_\tau^{(\ell-1)}) \quad (3)$$

# Layers semantics [3]

- Ordering of the tasks in layers: syntactic < semantic
- Localizable resolution of syntactical tasks, distributed resolution of semantic tasks
- “Availability of heuristics” suspicion



[3]: <https://arxiv.org/pdf/1905.05950.pdf> (BERT Rediscovered the Classical NLP Pipeline)

Figure 2: Layer-wise metrics on BERT-large. Solid (blue) are mixing weights  $s_{\tau}^{(\ell)}$  (§3.1); outlined (purple) are differential scores  $\Delta_{\tau}^{(\ell)}$  (§3.2), normalized for each task. Horizontal axis is encoder layer.



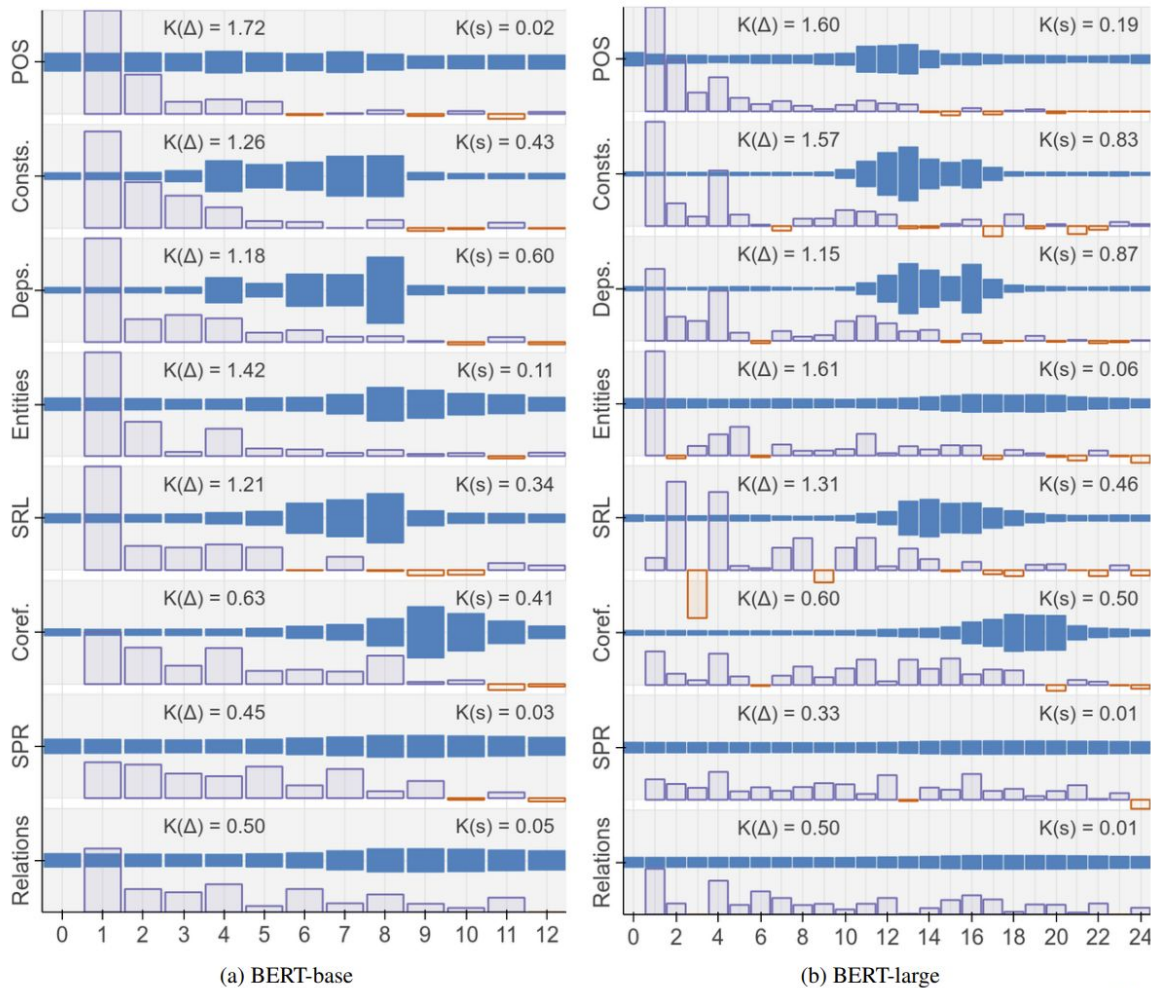
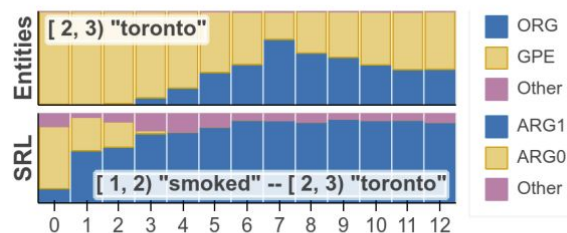


Figure A.3: Layer-wise metrics on BERT-base (left) and BERT-large (right). Solid (blue) are mixing weights  $s_\tau^{(\ell)}$ ; outlined (purple) are differential scores  $\Delta_\tau^{(\ell)}$ , normalized for each task. Horizontal axis is encoder layer.

# Layers semantics: per-case analysis [3]

- “Availability of heuristics” suspicion demonstrated on a few cases

(a) he smoked **toronto** in the playoffs with six hits, seven walks and eight stolen bases ...



(b) china **today** blacked out a cnn interview that was ...

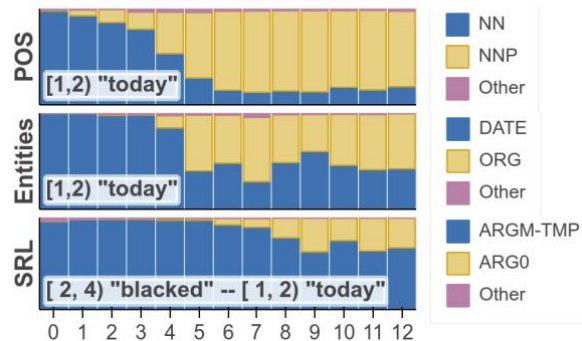


Figure 3: Probing classifier predictions across layers of BERT-base. Blue is the correct label; orange is the incorrect label with highest average score over layers. Bar heights are (normalized) probabilities  $P_{\tau}^{(\ell)}(\text{label}|s_1, s_2)$ . In the interest of space, only selected annotations are shown.

[3]: <https://arxiv.org/pdf/1905.05950.pdf> (BERT Rediscovered the Classical NLP Pipeline)

# Are Sixteen Heads Really Better than One? [4]

- Previous experiments show a suspicion on redundancy of Transformers Attention heads
- Experiments show that some (many!) heads can be removed without harming the performance
  - But it depends on the task
  - The more complex, the less heads can go off

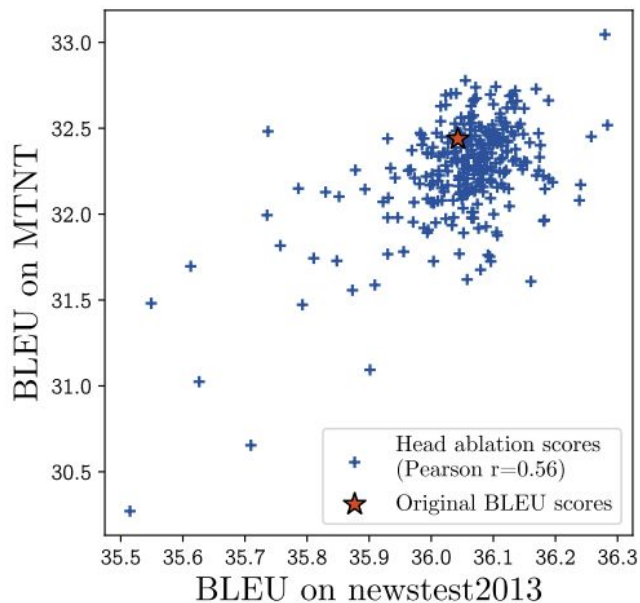
# Head ablations: ablating one head [4]

- Previous experiments show a suspicion on redundancy of Transformers Attention heads
- Experiments show that some (many!) heads can be removed without harming the performance
  - But it depends on the task
  - The more complex, the less heads can go off

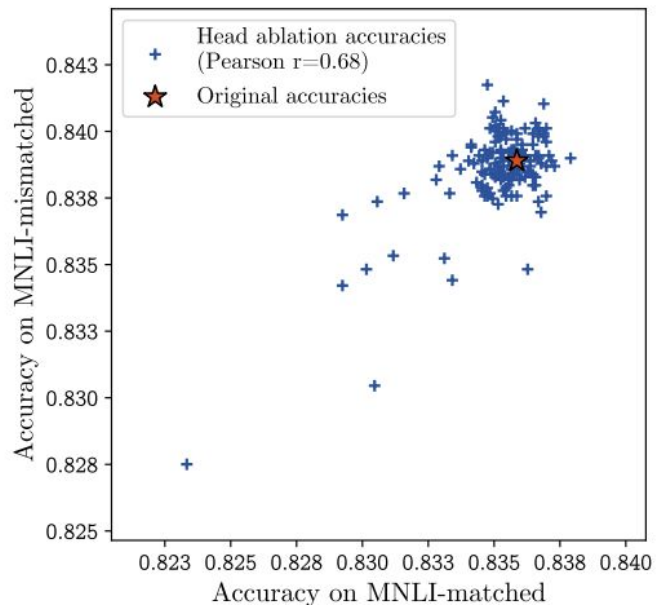
| Layer \ Head | 1            | 2     | 3            | 4           | 5     | 6            | 7    | 8            | 9     | 10          | 11          | 12   | 13    | 14          | 15    | 16    |
|--------------|--------------|-------|--------------|-------------|-------|--------------|------|--------------|-------|-------------|-------------|------|-------|-------------|-------|-------|
| 1            | 0.03         | 0.07  | 0.05         | -0.06       | 0.03  | <u>-0.53</u> | 0.09 | <u>-0.33</u> | 0.06  | 0.03        | 0.11        | 0.04 | 0.01  | -0.04       | 0.04  | 0.00  |
| 2            | 0.01         | 0.04  | 0.10         | <u>0.20</u> | 0.06  | 0.03         | 0.00 | 0.09         | 0.10  | 0.04        | <u>0.15</u> | 0.03 | 0.05  | 0.04        | 0.14  | 0.04  |
| 3            | 0.05         | -0.01 | 0.08         | 0.09        | 0.11  | 0.02         | 0.03 | 0.03         | -0.00 | 0.13        | 0.09        | 0.09 | -0.11 | <u>0.24</u> | 0.07  | -0.04 |
| 4            | -0.02        | 0.03  | 0.13         | 0.06        | -0.05 | 0.13         | 0.14 | 0.05         | 0.02  | 0.14        | 0.05        | 0.06 | 0.03  | -0.06       | -0.10 | -0.06 |
| 5            | <u>-0.31</u> | -0.11 | -0.04        | 0.12        | 0.10  | 0.02         | 0.09 | 0.08         | 0.04  | <u>0.21</u> | -0.02       | 0.02 | -0.03 | -0.04       | 0.07  | -0.02 |
| 6            | 0.06         | 0.07  | <u>-0.31</u> | 0.15        | -0.19 | 0.15         | 0.11 | 0.05         | 0.01  | -0.08       | 0.06        | 0.01 | 0.01  | 0.02        | 0.07  | 0.05  |

Table 1: Difference in BLEU score for each head of the encoder's self attention mechanism. Underlined numbers indicate that the change is statistically significant with  $p < 0.01$ . The base BLEU score is 36.05.

# Head ablations: ablating one head: per-task [4]



(a) BLEU on newstest2013 and MTNT when individual heads are removed from WMT. Note that the ranges are not the same on the X and Y axis as there seems to be much more variation on MTNT.



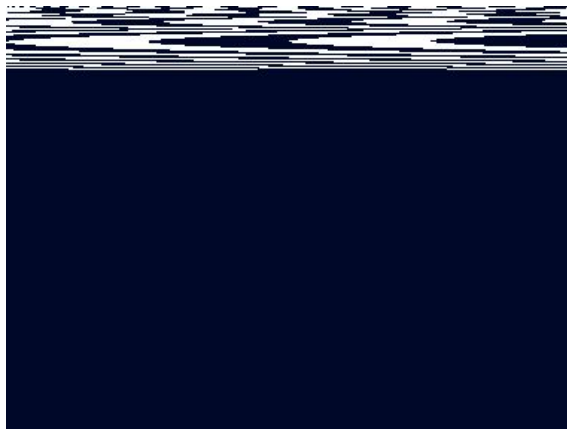
(b) Accuracies on MNL-matched and -mismatched when individual heads are removed from BERT. Here the scores remain in the same approximate range of values.

# Head ablations: incremental ablations

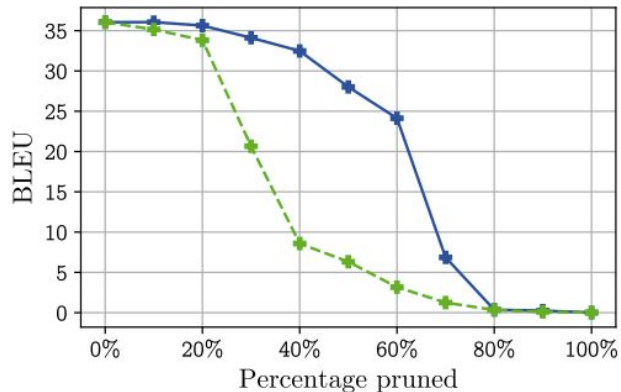
As a proxy score for head importance, we look at the expected sensitivity of the model to the mask variables  $\xi_h$  defined in §2.3:

$$I_h = \mathbb{E}_{x \sim X} \left| \frac{\partial \mathcal{L}(x)}{\partial \xi_h} \right| \quad (2)$$

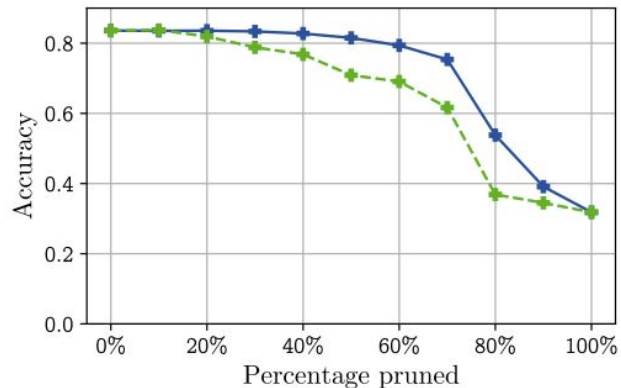
where  $X$  is the data distribution and  $\mathcal{L}(x)$  the loss on sample  $x$ . Intuitively, if  $I_h$  has a high value then changing  $\xi_h$  is liable to have a large effect on the model.



# Head ablations: incremental ablations



(a) Evolution of BLEU score on `newstest2013` when heads are pruned from WMT.



(b) Evolution of accuracy on the MultiNLI-matched validation set when heads are pruned from BERT.

Figure 3: Evolution of accuracy by number of heads pruned according to  $I_h$  (solid blue) and individual oracle performance difference (dashed green).

# So what is it for?

- Pruning of selected attention heads **increases speed** and **decrease model size**
  - and is already integrated into fine-tuning pipeline of some libraries, e.g. Transformers [5]
- Identifying head's functionality, or knowing how to identify it, can be useful, when you want to **utilize attention for specific task**



# Utilizing attention for specific task

A light peek into my research

- Attention is also an inherent way of denoting, which parts of text are **important** for given output
  - One head does only linear transformation of the input
  - Stacking them can create complex non-linearity
- But now that we can interpret heads' functionality, we can **pick** the ones that we know are **relevant for our problem**

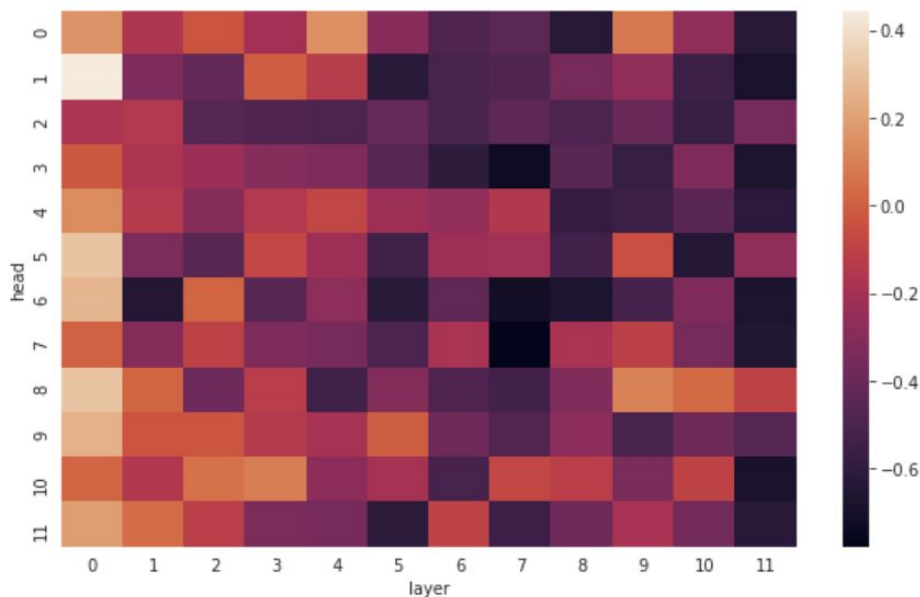
# Utilizing attention for weighting text

- **Can we weight text using Attention?**
- To find out, we propose a set of experiments, inspired by previous literature
- Itself, identification of **key phrases** in the text can be beneficial for many tasks: Summarization, Information Retrieval, Keyword extraction, indexing
- Arguably similar to Correferencing, that was associated with specific heads

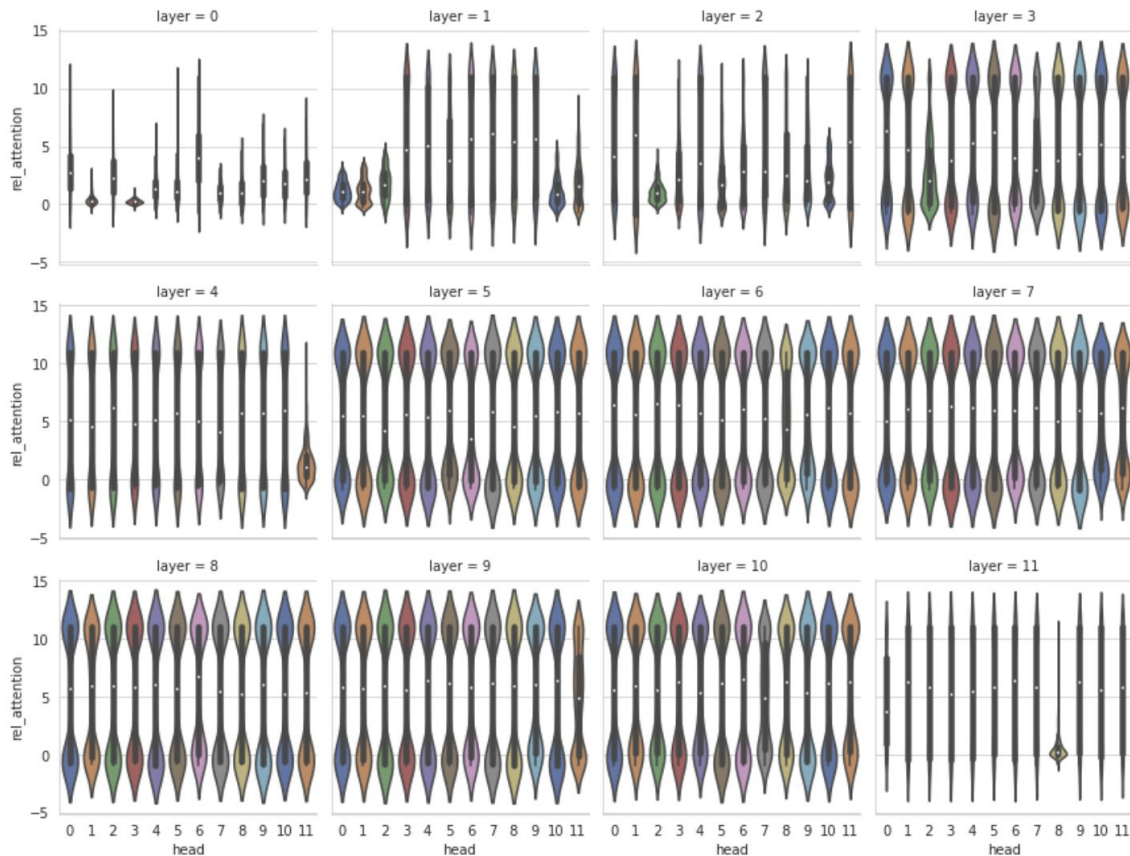
# Utilizing attention for weighting text: experiments

## Static analysis of the model

- Identification of heads, whose attention can best distinguish key parts from less important
- **Mean relative attention** =  $\text{mean\_key\_segments} - \text{mean\_other\_segments}$



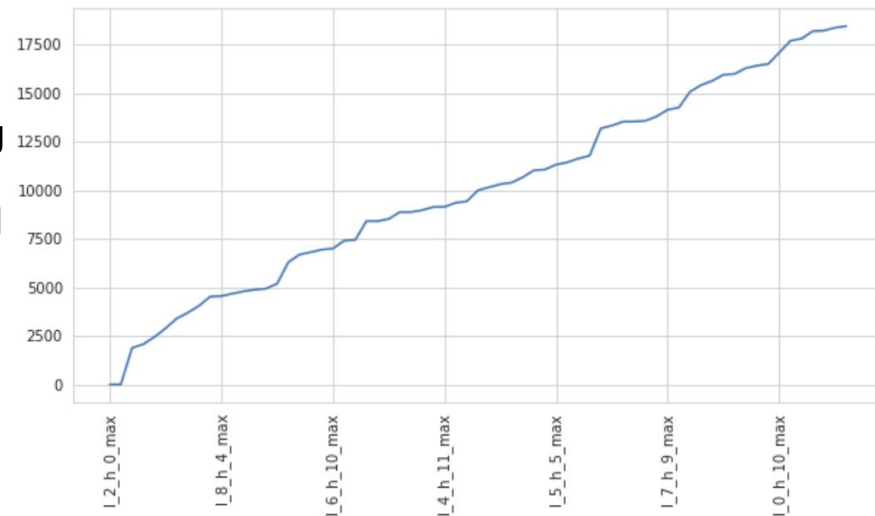
# Utilizing attention for weighting text: experiments



# Utilizing attention for weighting text: experiments

## Weighting of attention heads

- Identification of heads, that are best for **predicting keywords**
- Cross-entropy [6], and linear weights of each head
  - Reproduction of **Scalar Mixing Weights** from [2]



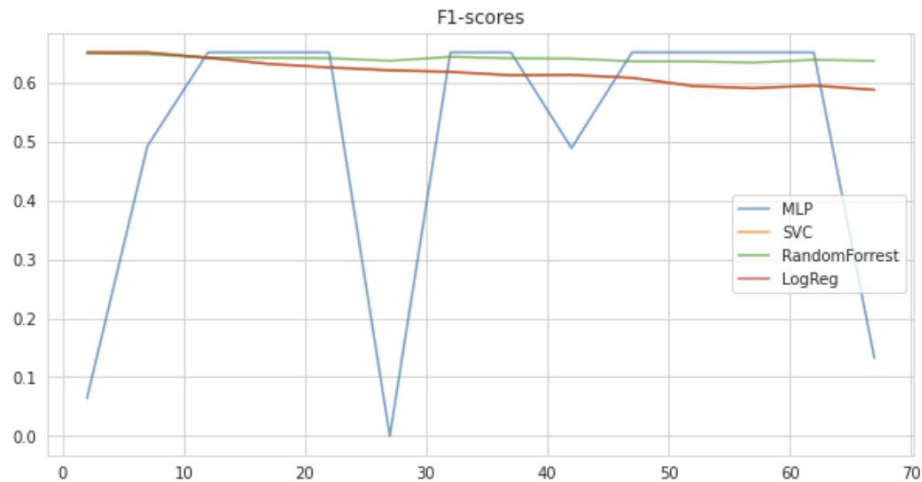
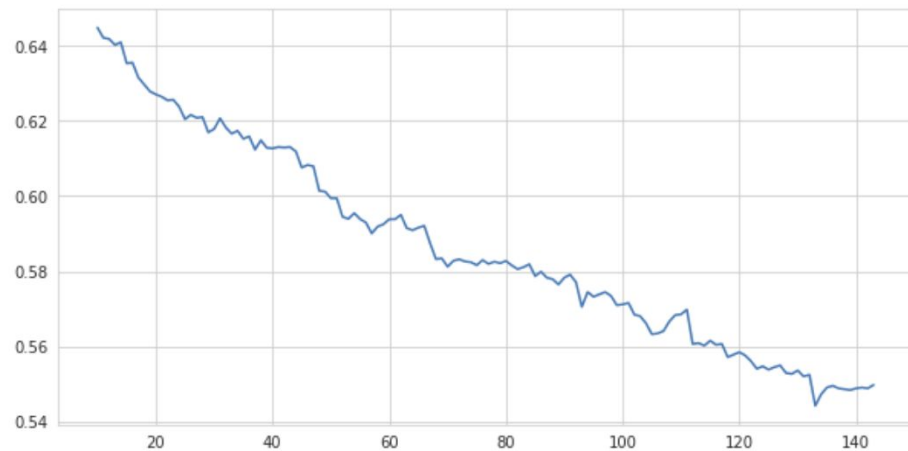
```
entropies.sort_values().head(10)
```

```
l_2_h_0_max    20.466552
l_2_h_9_max    23.297212
l_7_h_3_max    1894.318051
l_7_h_7_max    2083.535294
l_7_h_6_max    2455.226861
l_5_h_7_max    2899.598139
l_5_h_1_max    3407.217381
l_8_h_5_max    3707.706600
l_8_h_0_max    4063.854563
l_9_h_4_max    4535.810778
```

# Utilizing attention for weighting text: experiments

```
# Logistic Regression  
pd.Series(fscores).plot(figsize=(10, 5))
```

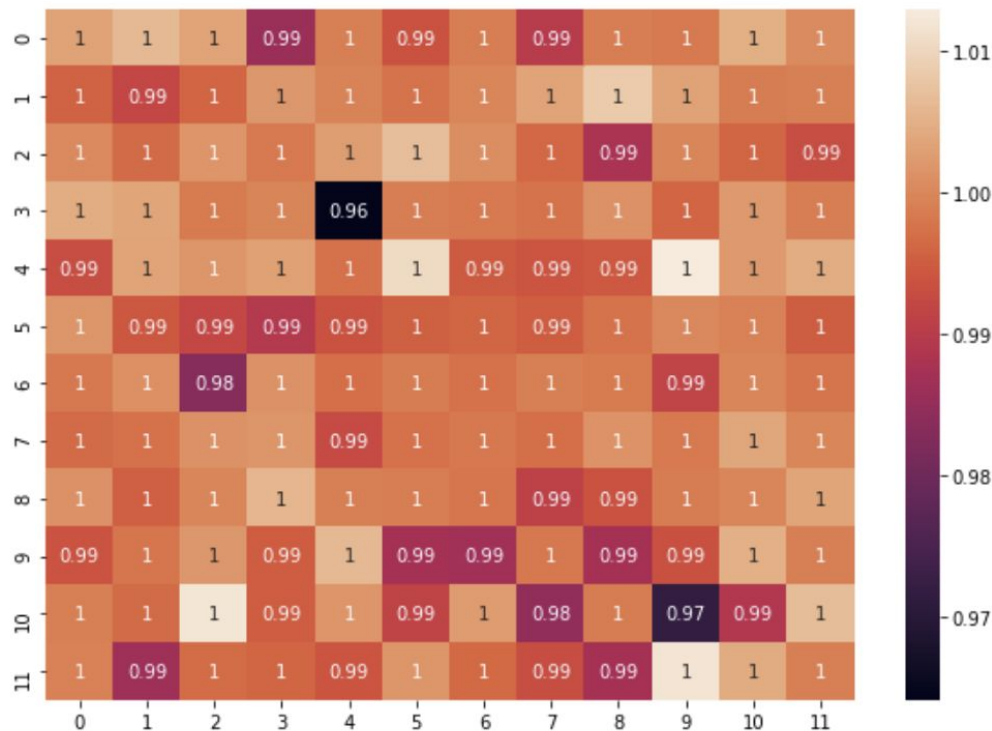
<AxesSubplot:>



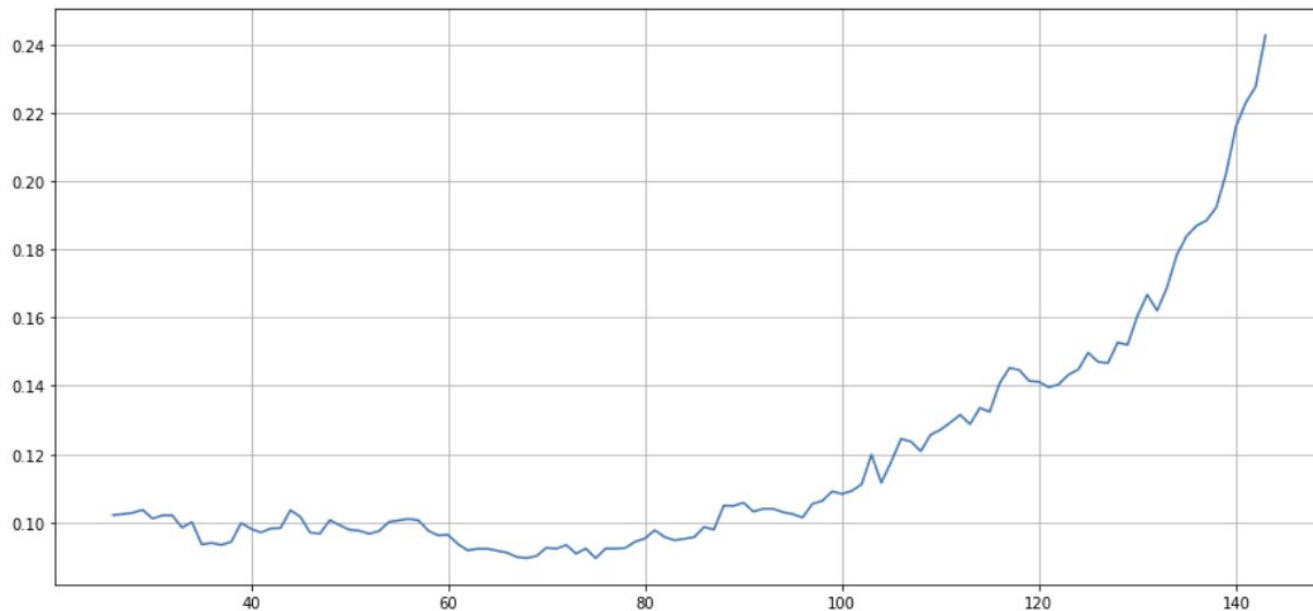
# Utilizing attention for weighting text: experiments

## Fine-tuning model and analysing heads

- We train the BERT-base model end-to-end for keyword identification
  - reaching F1 = 0.36/0.37 pruned
  - SOTA=0.42
- We reproduce “remove-one” ablation experiment
- We also rank head by their “ablation drop”



# Utilizing attention for weighting text: experiments



Ablation of worst-to-best-performing Attention heads





# Thanks!

Feel free to check out our theses:

<https://is.muni.cz/auth/rozpis/tema> tag **MIR**

or contact us later!

MUNI  
FI



Michal Štefánik

stefanik.m@mail.muni.cz