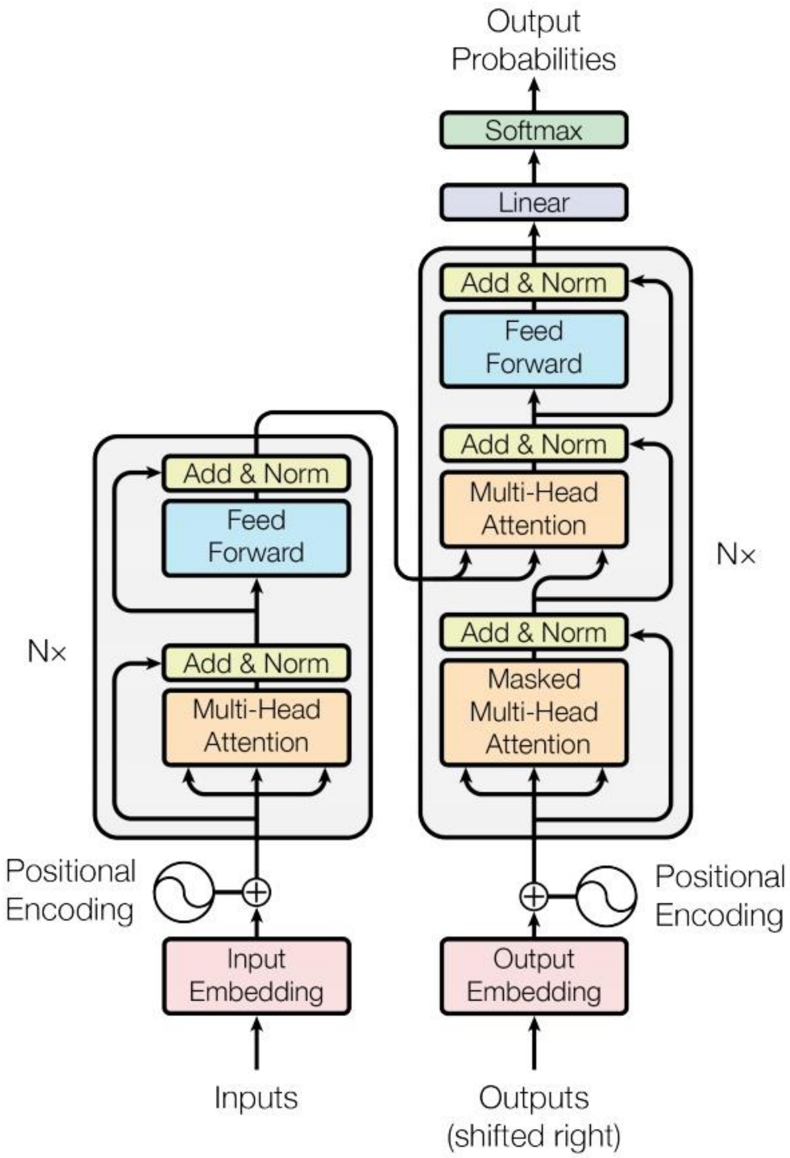
Survey on Analysis of Transformer

Wenxuan Wang

07/03/2019

Motivation



Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Paper Included

- How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures (ACL 2018)
- What Does BERT Look At? An Analysis of BERT's Attention(ACL 2019 Workshop)
- Is Attention Interpretable?(ACL 2019)
- A Multiscale Visualization of Attention in the Transformer Model (ACL 2019 System Demonstrations)
- Analyzing the Structure of Attention in a Transformer Language Model (ACL 2019 Workshop)
- Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned(ACL 2019)

How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures

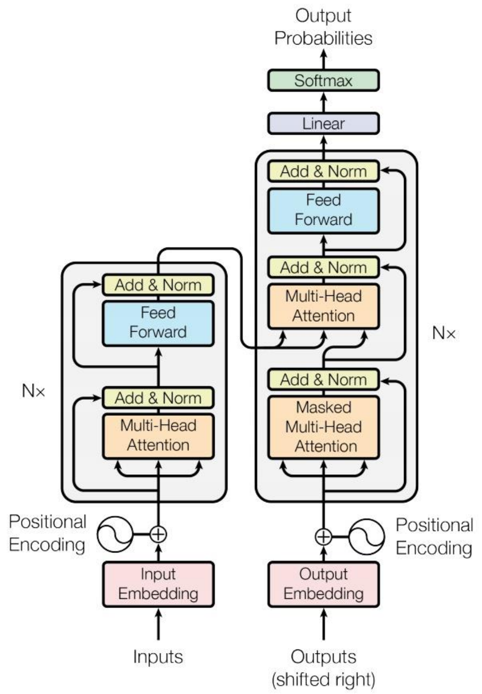
Tobias Domhan

Amazon

ACL 2018

How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures

- Motivation:
  - Why Transformer works?
  - The Transformer has self-attention, layer normalization, multiple source attention mechanisms, a multi-head dot attention mechanism, and residual feedforward layers.
  - How much each of these components matters?



How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures

- Architecture Definition Language (ADL) allowing for a flexible combination of common building blocks.

$$\mathit{linear}(\mathbf{h}_t, d_o) = \mathbf{W}\mathbf{h}_t + \mathbf{b},$$

$$\mathit{ff}(\mathbf{h}_t, d_o) = \mathit{dropout}(\max(0, \mathit{linear}(\mathbf{h}_t, d_o)))$$

$$\mathit{ffl}(\mathbf{h}_t) = \mathit{ff}(4d_{in}) \rightarrow \mathit{linear}(d_{in})$$

$$\mathbf{U}^{L_s} = \mathit{dropout} \rightarrow \mathit{birnn} \rightarrow \mathit{repeat}(n-1, \mathit{res\_d}(\mathit{rnn})).$$

$$t_{\text{enc}} = \mathit{res\_nd}(\mathit{mh\_dot\_self\_att}) \rightarrow \mathit{res\_nd}(\mathit{ffl})$$

$$t_{\text{dec}} = \mathit{res\_nd}(\mathit{mh\_dot\_self\_att}) \rightarrow \mathit{res\_nd}(\mathit{mh\_dot\_src\_att}) \rightarrow \mathit{res\_nd}(\mathit{ffl}).$$

How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures

- Recurrent and convolutional models can be very close to the Transformer performance.
- Multiple source attention layers and residual feed-forward layers are key.

Model	IWSLT EN→DE	WMT'17 EN→DE		WMT'17 LV→EN	
	BLEU	BLEU	METEOR	BLEU	METEOR
Transformer	25.4 ± 0.1	27.6 ± 0.0	47.2 ± 0.1	18.5 ± 0.0	51.3 ± 0.1
RNMT	23.2 ± 0.2	25.5 ± 0.2	45.1 ± 0.1	-	-
- input feeding	23.1 ± 0.2	24.6 ± 0.1	43.8 ± 0.2	-	-
RNN	22.8 ± 0.2	23.8 ± 0.1	43.3 ± 0.1	15.2 ± 0.1	45.9 ± 0.1
+ mh	23.7 ± 0.4	24.4 ± 0.1	43.9 ± 0.1	16.0 ± 0.1	47.1 ± 0.1
+ pos	23.9 ± 0.2	24.1 ± 0.1	43.5 ± 0.2	-	-
+ norm	23.7 ± 0.1	24.0 ± 0.2	43.2 ± 0.1	15.2 ± 0.1	46.3 ± 0.2
+ multi-att-1h	24.5 ± 0.0	25.2 ± 0.1	44.9 ± 0.1	16.6 ± 0.2	49.1 ± 0.2
/ multi-att	24.4 ± 0.3	25.5 ± 0.0	45.3 ± 0.0	17.0 ± 0.2	49.4 ± 0.1
+ ff	25.1 ± 0.1	26.7 ± 0.1	46.4 ± 0.2	17.8 ± 0.1	50.5 ± 0.1

Table 3: Transforming an RNN into a Transformer style architecture. + shows the incrementally added variation. / denotes an alternative variation to which the subsequent + is relative to.

Model	IWSLT EN-DE	WMT'17 EN→DE		WMT'17 LV→EN	
	BLEU	BLEU	METEOR	BLEU	METEOR
Transformer	25.4 ± 0.1	27.6 ± 0.0	47.2 ± 0.1	18.5 ± 0.0	51.3 ± 0.1
CNN GLU	24.3 ± 0.4	25.0 ± 0.3	44.4 ± 0.2	16.0 ± 0.5	47.4 ± 0.4
+ norm	24.1 ± 0.1	-	-	-	-
+ mh	24.2 ± 0.2	25.4 ± 0.1	44.8 ± 0.1	16.1 ± 0.1	47.6 ± 0.2
+ ff	25.3 ± 0.1	26.8 ± 0.1	46.0 ± 0.1	16.4 ± 0.2	47.9 ± 0.2
CNN ReLU	23.6 ± 0.3	23.9 ± 0.1	43.4 ± 0.1	15.4 ± 0.1	46.4 ± 0.3
+ norm	24.3 ± 0.1	24.3 ± 0.2	43.6 ± 0.1	16.0 ± 0.2	47.1 ± 0.5
+ mh	24.2 ± 0.2	24.9 ± 0.1	44.4 ± 0.1	16.1 ± 0.1	47.5 ± 0.2
+ ff	25.3 ± 0.3	26.9 ± 0.1	46.1 ± 0.0	16.4 ± 0.2	47.9 ± 0.1

Table 4: Transforming a CNN based model into a Transformer style architecture.

How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures

- Self-attention is much more important for encoder side than decoder side.
- It is OK even if decoder is either an RNN or CNN when encoder has self-attention.

		IWSLT EN→DE	WMT'17 EN→DE		WMT'17 LV→EN	
Encoder	Decoder	BLEU	BLEU	METEOR	BLEU	METEOR
self-att	self-att	25.4 ± 0.2	27.6 ± 0.0	47.2 ± 0.1	18.3 ± 0.0	51.1 ± 0.1
self-att	RNN	25.1 ± 0.1	27.4 ± 0.1	47.0 ± 0.1	18.4 ± 0.2	51.1 ± 0.1
self-att	CNN	25.4 ± 0.4	27.6 ± 0.2	46.7 ± 0.1	18.0 ± 0.3	50.3 ± 0.3
RNN	self-att	25.8 ± 0.1	27.2 ± 0.1	46.7 ± 0.1	17.8 ± 0.1	50.6 ± 0.1
CNN	self-att	25.7 ± 0.1	26.6 ± 0.3	46.3 ± 0.1	16.8 ± 0.4	49.4 ± 0.4
RNN	RNN	25.1 ± 0.1	26.7 ± 0.1	46.4 ± 0.2	17.8 ± 0.1	50.5 ± 0.1
CNN	CNN	25.3 ± 0.3	26.9 ± 0.1	46.1 ± 0.0	16.4 ± 0.2	47.9 ± 0.2
self-att	*combined*	25.1 ± 0.2	27.6 ± 0.2	47.2 ± 0.2	18.3 ± 0.1	51.1 ± 0.1
self-att	*none*	23.7 ± 0.2	25.3 ± 0.2	43.1 ± 0.1	15.9 ± 0.1	45.1 ± 0.2

Table 5: Different variations of the encoder and decoder self-attention layer.

What Does BERT Look At? An Analysis of BERT's Attention

Kevin Clark, Urvashi Khandelwal, Omer Levy, Christopher D. Manning

Stanford University & Facebook AI Research

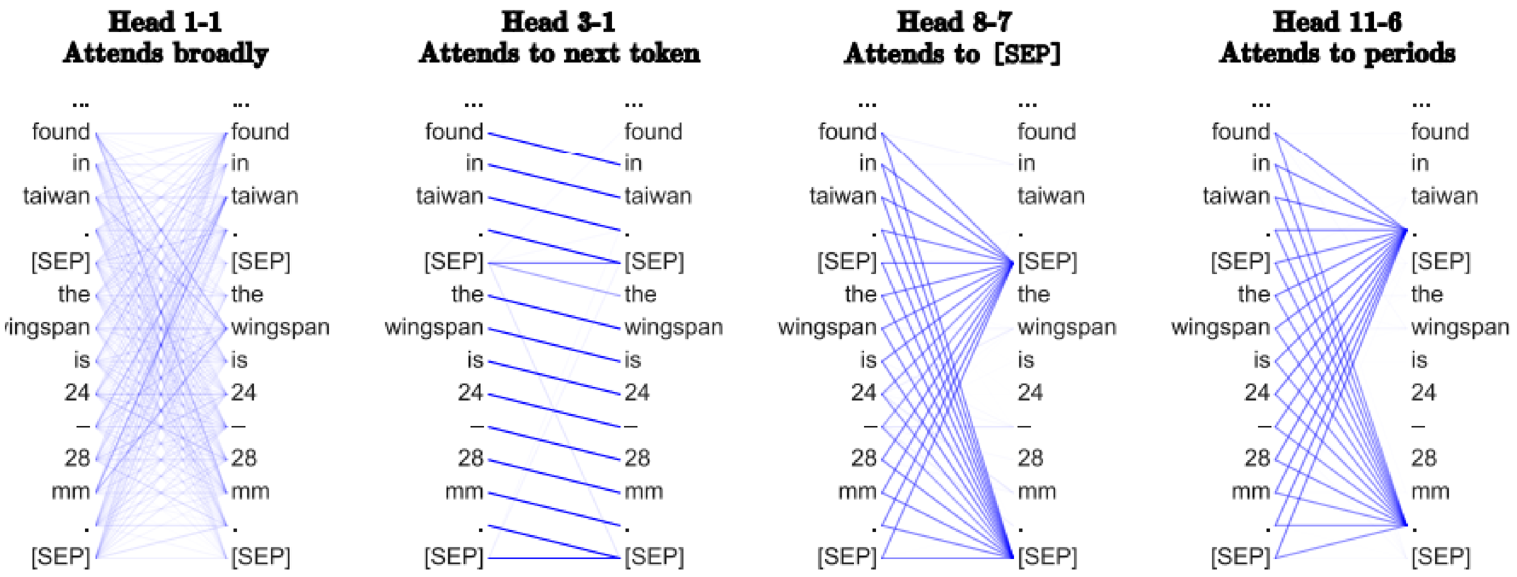
ACL 2019 Workshop: Blackbox NLP

What Does BERT Look At? An Analysis of BERT's Attention

- Motivation:
  - What Does BERT Look At?
  - Common Behaver?
  - Syntactic?
  - Semantic?

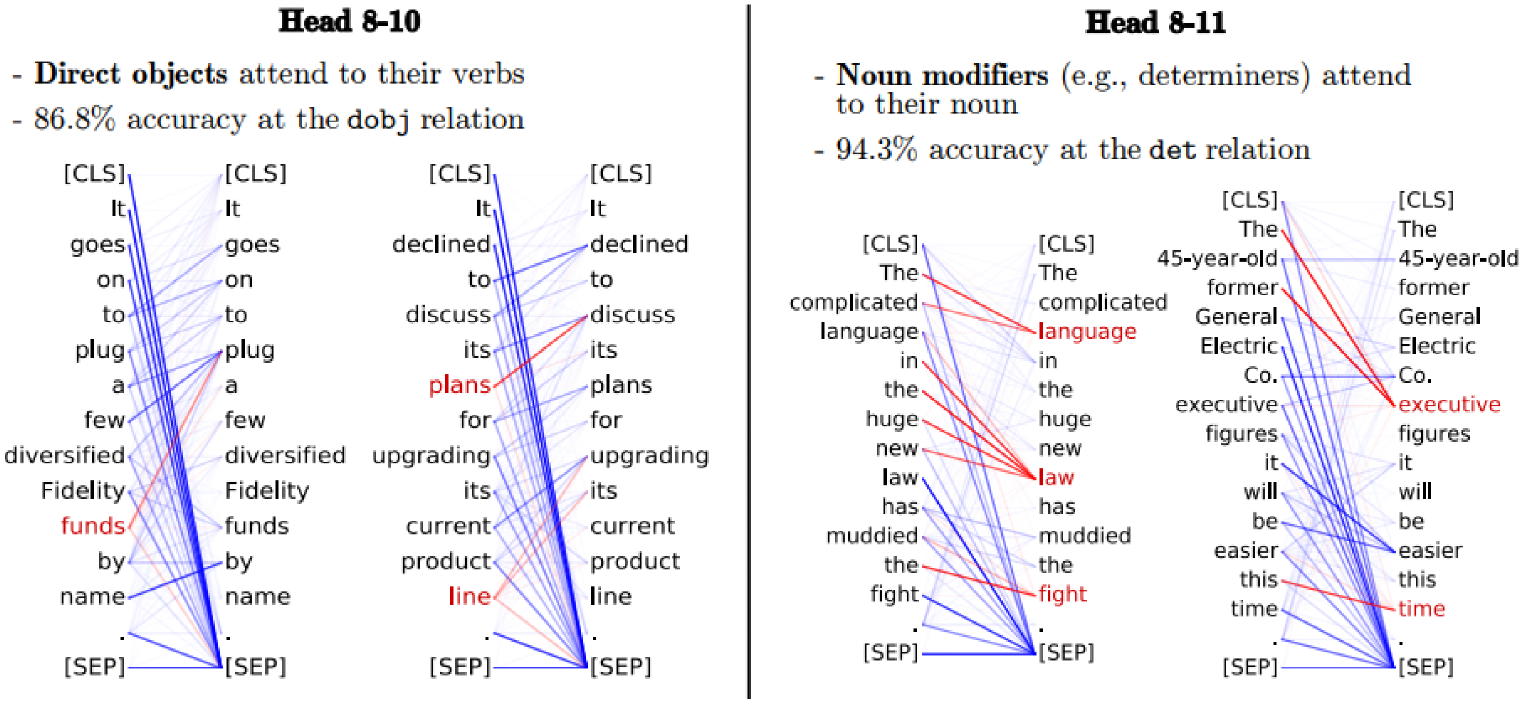
What Does BERT Look At? An Analysis of BERT's Attention

- How BERT’s attention heads behave?
  - Put little attention on the current token but attending heavily on next or previous token.
  - [SEP]: no-op
  - Attention heads in the same layer behave similarly
  - Lower layers has broad attention



What Does BERT Look At? An Analysis of BERT's Attention

- Probe each attention head for linguistic phenomena
  - Input a word, find the most-attended-to other word.
  - Evaluate the ability of the heads to classify various syntactic relations.
  - Find that particular heads correspond to particular relations, like direct objects of verbs, determiners of nouns, objects of prepositions.



Relation	Head	Accuracy	Baseline
All	7-6	34.5	26.3 (1)
prep	7-4	66.7	61.8 (-1)
pobj	9-6	**76.3**	34.6 (-2)
det	8-11	**94.3**	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	**86.8**	40.0 (-2)
advmod	7-6	48.8	40.2 (1)
aux	4-10	81.1	71.5 (1)
poss	7-6	**80.5**	47.7 (1)
auxpass	4-10	**82.5**	40.5 (1)
ccomp	8-1	**48.8**	12.4 (-2)
mark	8-2	**50.7**	14.5 (2)
prt	6-7	**99.1**	91.4 (-1)

What Does BERT Look At? An Analysis of BERT's Attention

- Attention heads for coreference resolution (semantic)
  - 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
  - One of BERT's attention heads achieves decent performance

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

What Does BERT Look At? An Analysis of BERT's Attention

- Probing Attention Head Combinations
  - dependency parsing
  - Given an input word, the classifier produces a probability distribution over other words in the sentence indicating how likely each other word is to be the syntactic head of the current one.
  - BERT's attention maps have a fairly thorough representation of English syntax

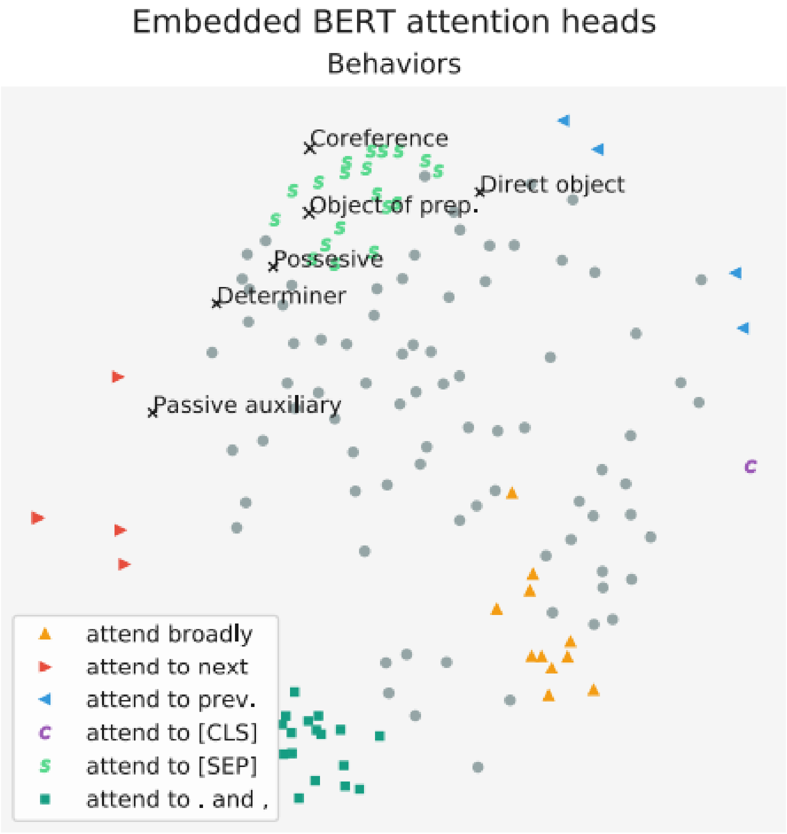
$$p(i|j) \propto \exp\left(\sum_{k=1}^{n} w_k \alpha_{ij}^k + u_k \alpha_{ji}^k\right)$$

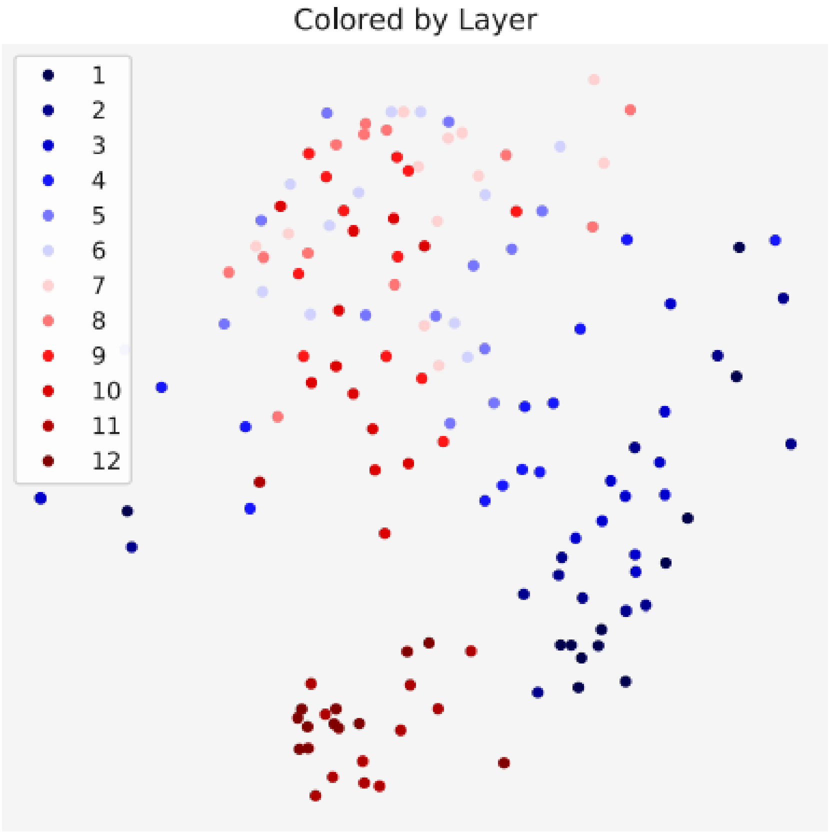
$$p(i|j) \propto \exp\left(\sum_{k=1}^{n} W_{k,:}(v_i \oplus v_j)\alpha_{ij}^k + U_{k,:}(v_i \oplus v_j)\alpha_{ji}^k\right)$$

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

What Does BERT Look At? An Analysis of BERT's Attention

- Clustering Attention Heads
  - Are attention heads in the same layer similar to each other or different?
  - Yes!





Is Attention Interpretable?

Sofia Serrano, Noah A. Smith
University of Washington
ACL 2019

Is Attention Interpretable?

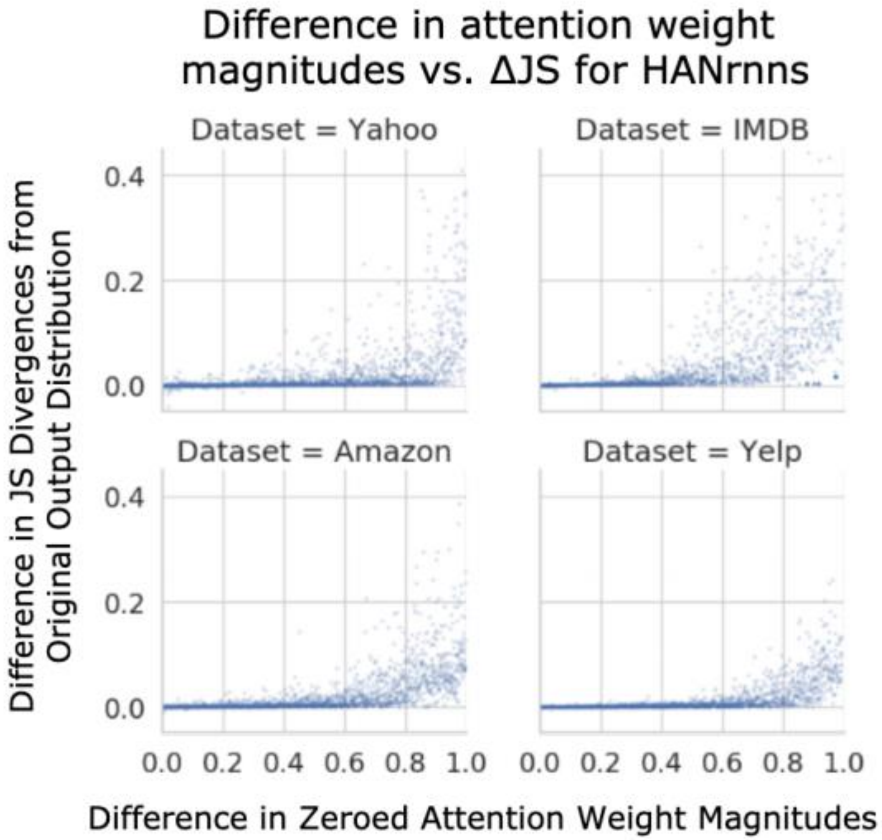
- Motivation:
  - Attention weights means importance ?
  - Bigger weights -> more important ?
  - Those explanations accurately represent the true reasons for the **model's decision**.

Is Attention Interpretable?

- Text classification task
- With and without attention weight w_i, the difference between output distribution p and q is huge -> important and interpretable
- Two ways for difference measurement
  - Jensen-Shannon divergence
  - decision flip

Is Attention Interpretable?

- Single attention weights' importance



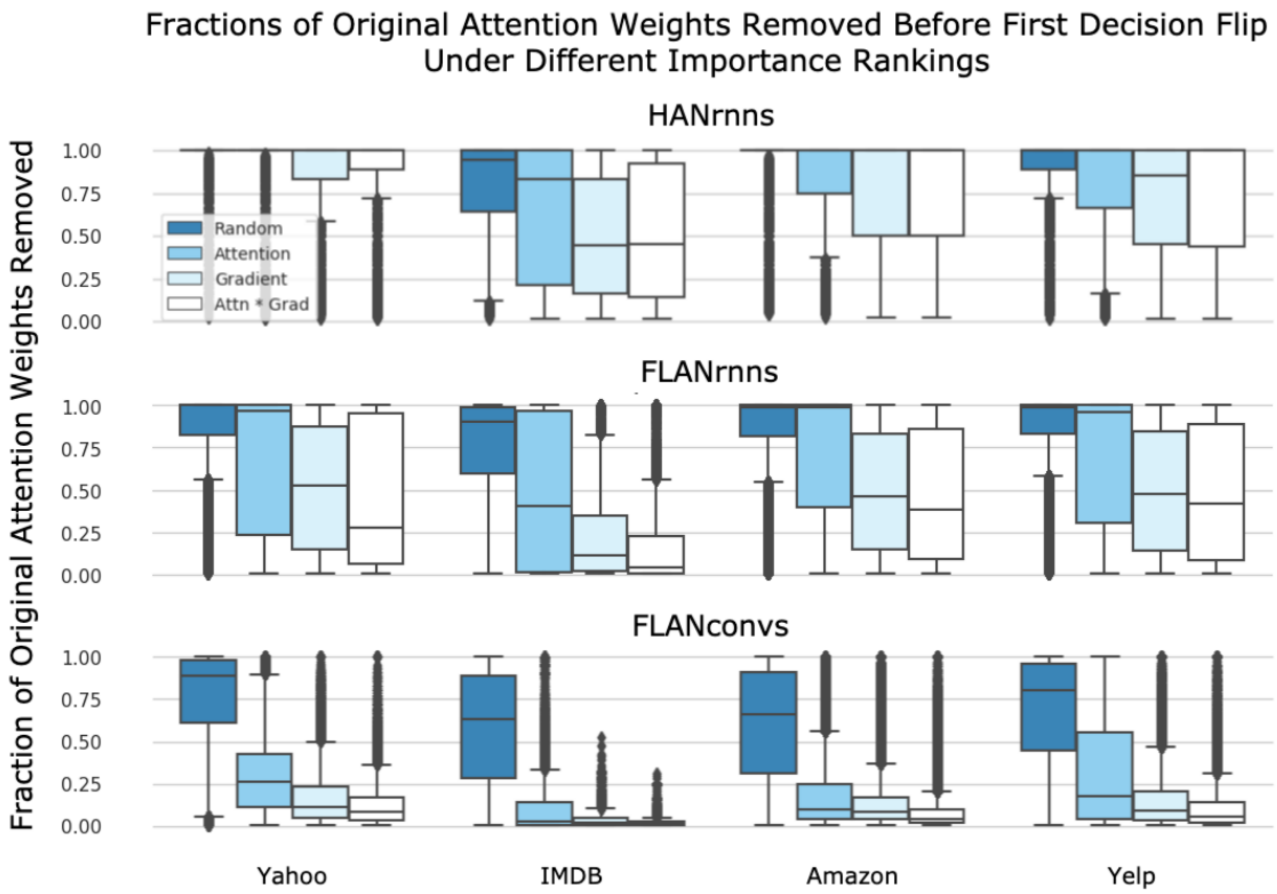
Remove random: Decision flip?

Remove i^*: Decision flip?	Yahoo: Yes	Yahoo: No	IMDB: Yes	IMDB: No
Yes	0.5	8.7	2.2	12.2
No	1.3	89.6	1.4	84.2

Remove i^*: Decision flip?	Amazon: Yes	Amazon: No	Yelp: Yes	Yelp: No
Yes	2.7	7.6	1.5	8.9
No	2.7	87.1	1.9	87.7

Is Attention Interpretable?

- Importance of Sets of Attention Weights
- Ranking and removing
- Attention Does Not Optimally Describe Model Decisions
- Decision Flips Often Occur Late



A Multiscale Visualization of Attention in the Transformer Model

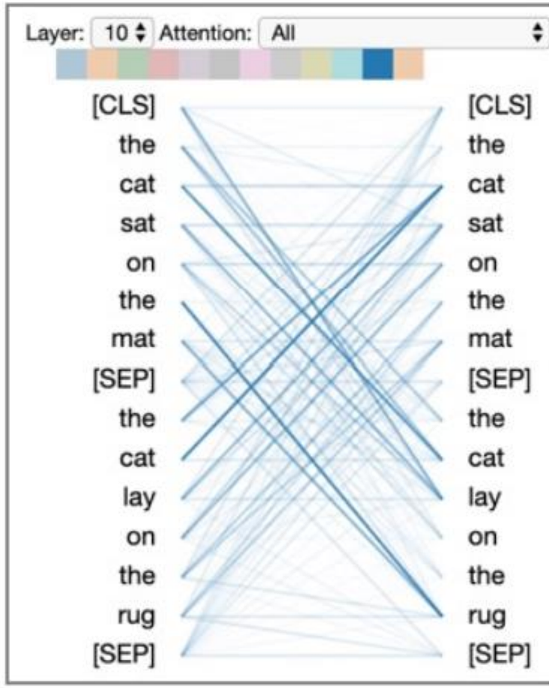
Jesse Vig

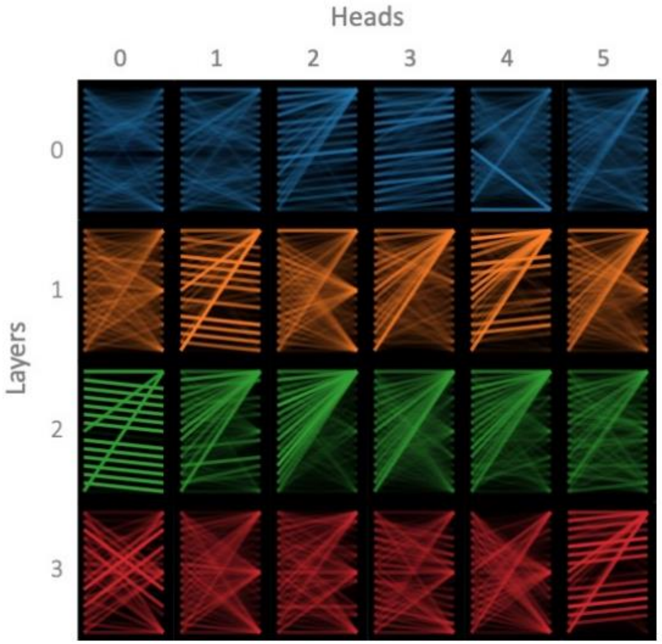
Palo Alto Research Center

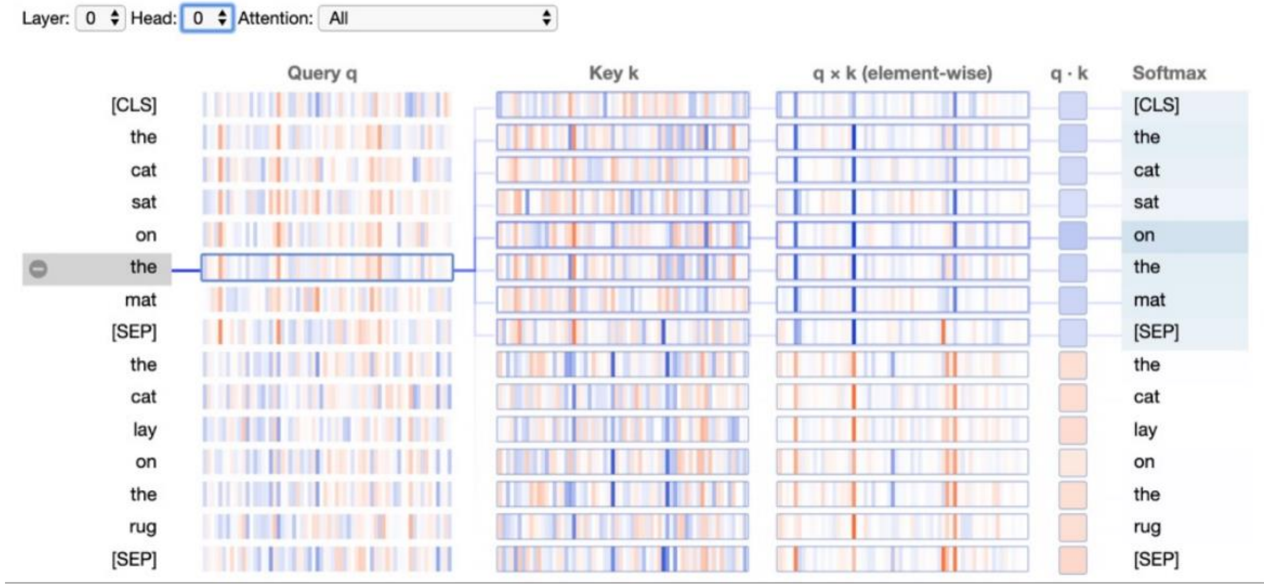
ACL 2019 System Demonstrations

A Multiscale Visualization of Attention in the Transformer Model

- Open source
- Attention Head, Model and Neuron level to visualize attention
- BERT and GPT-2 are included







Analyzing the Structure of Attention in a Transformer Language Model

Jesse Vig & Yonatan Belinkov

Palo Alto Research Center & MIT

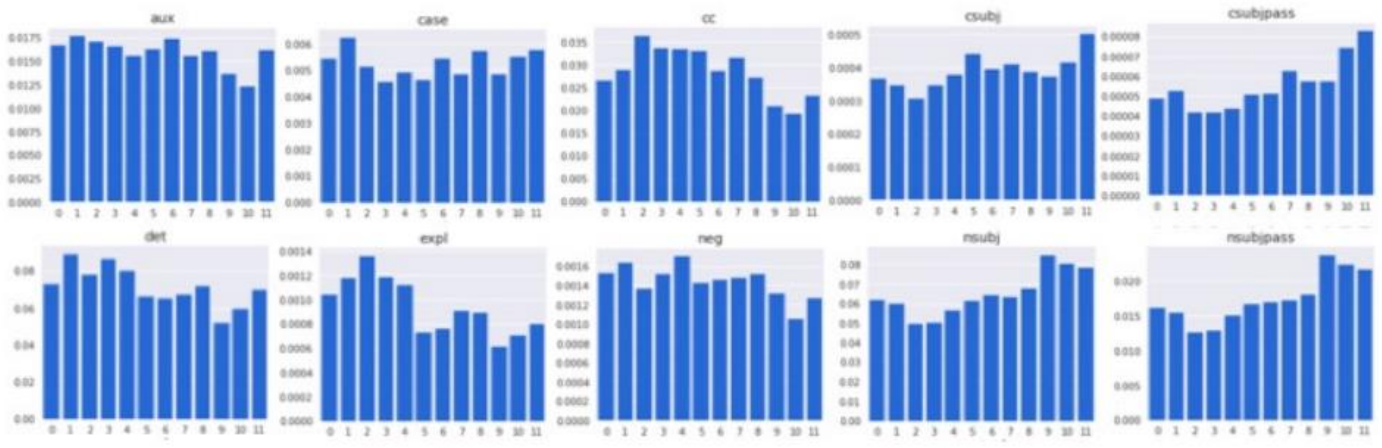
ACL 2019 Workshop

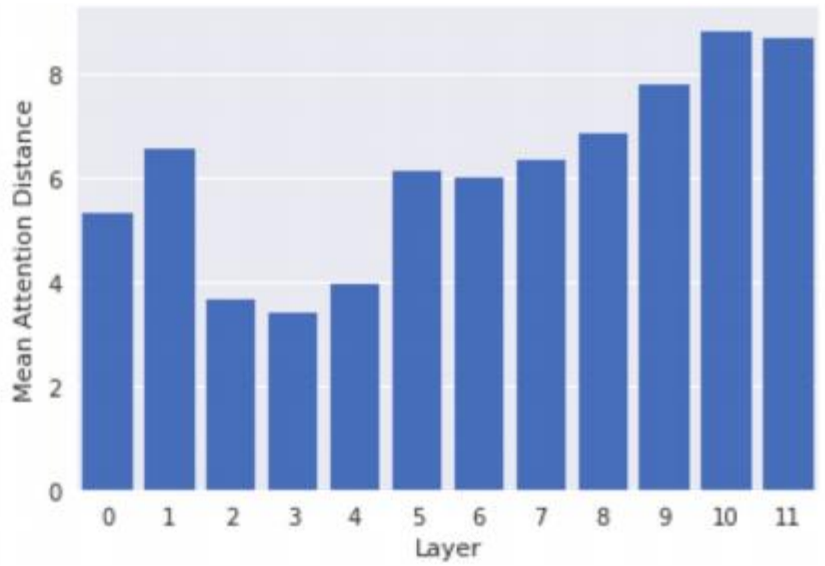
Analyzing the Structure of Attention in a Transformer Language Model

- Motivation:
  - Does attention align with syntactic dependency relations?
  - Which attention heads attend to which partof-speech tags?
  - How does attention capture long-distance relationships versus short-distance ones?

Analyzing the Structure of Attention in a Transformer Language Model

- Analysis on GPT-2
- Many attention heads specialize in particular part-of-speech tags and that different tags are targeted at different layer depths
- The deepest layers capture the most distant relationships





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

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, Ivan Titov

University of Edinburgh& University of Amsterdam

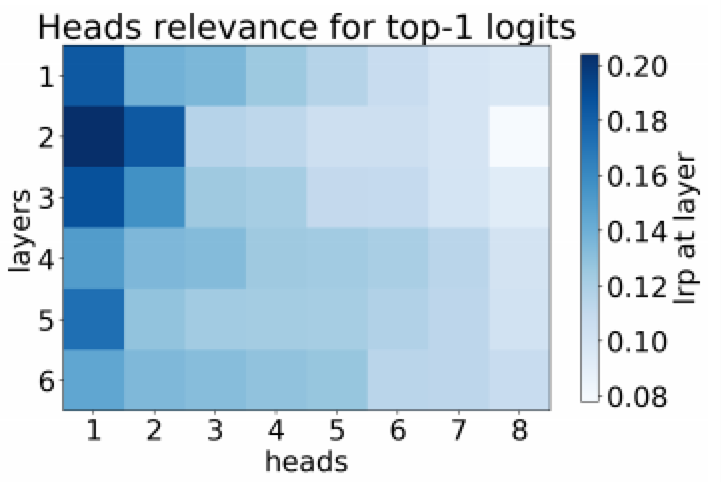
ACL 2019

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

- Motivation:
  - Do individual encoder heads play consistent and interpretable roles?
  - which are the most important ones for translation quality?
  - Can we significantly reduce the number of attention heads while preserving translation quality?

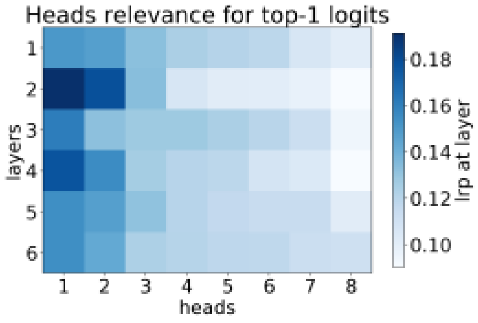
Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned

- Identifying Important Heads
- Layer-wise relevance propagation
- Only a small subset of heads are important for translation

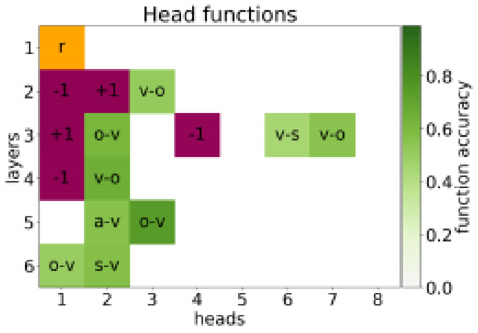


Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned

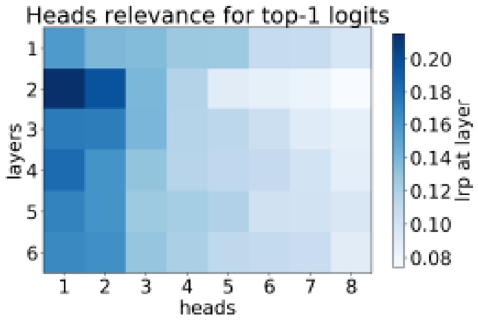
- Important heads have one or more specialized and interpretable functions in the model
  - Positional
  - Syntactic
  - Rare words



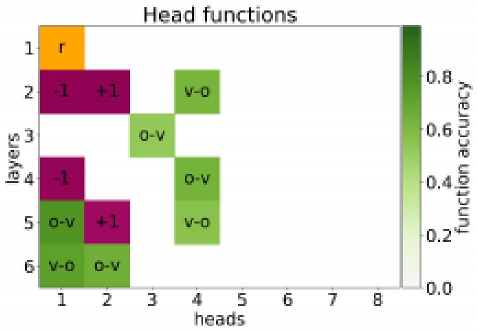
(a) LRP (EN-DE)



(b) head functions



(c) LRP (EN-FR)



(d) head functions

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

- Head pruning
- Add a scalar gate g with L0 norm

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

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

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

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

	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