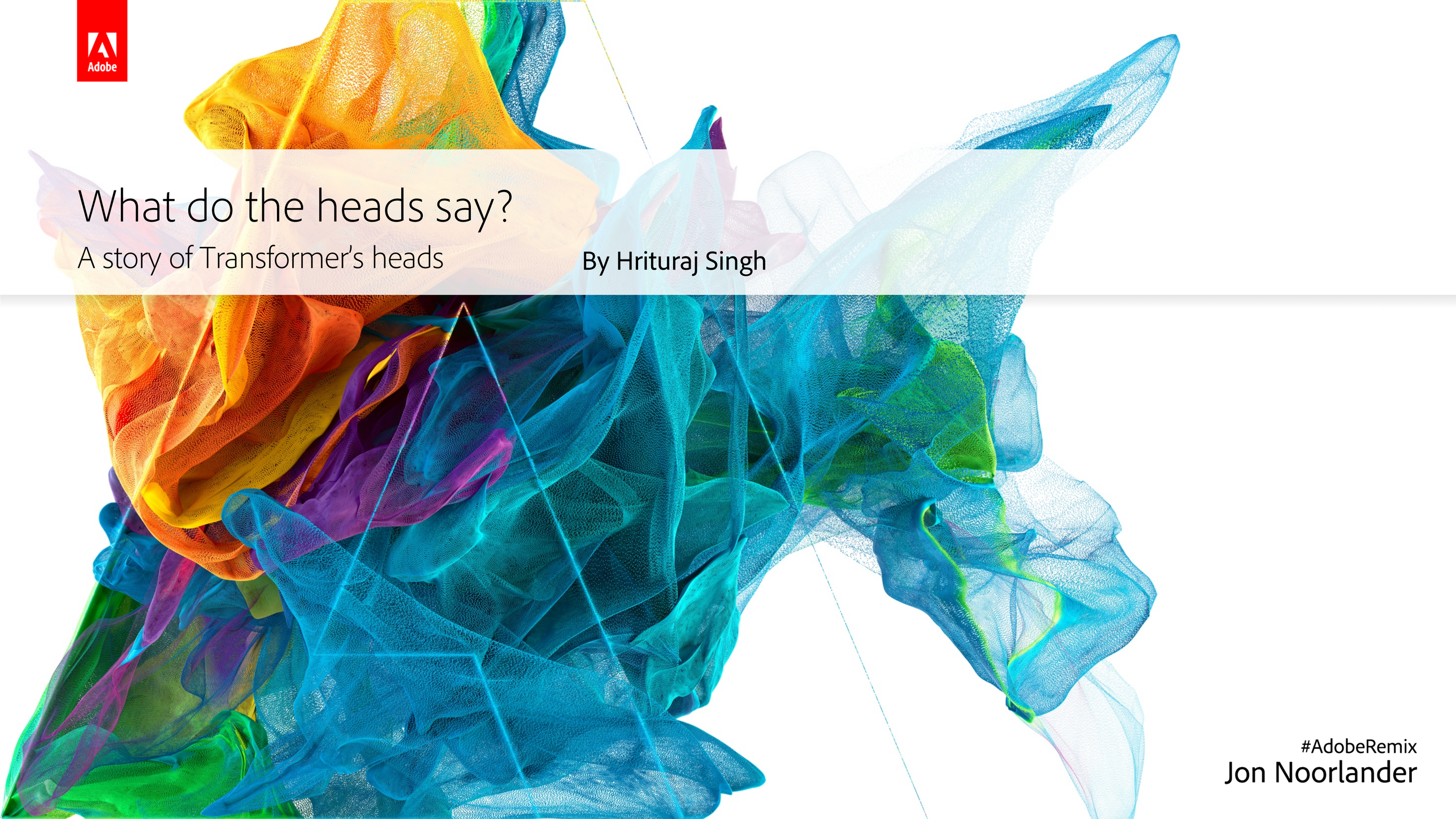


What do the heads say?

A story of Transformer's heads

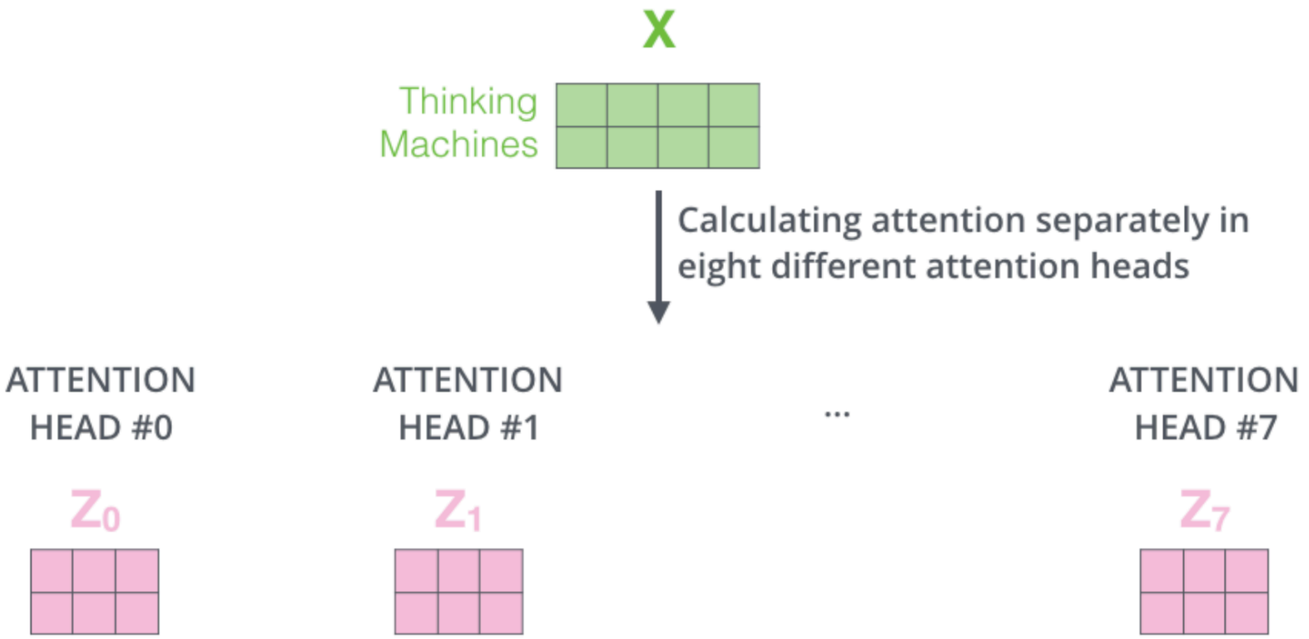
By Hrituraj Singh

#AdobeRemix

Jon Noorlander

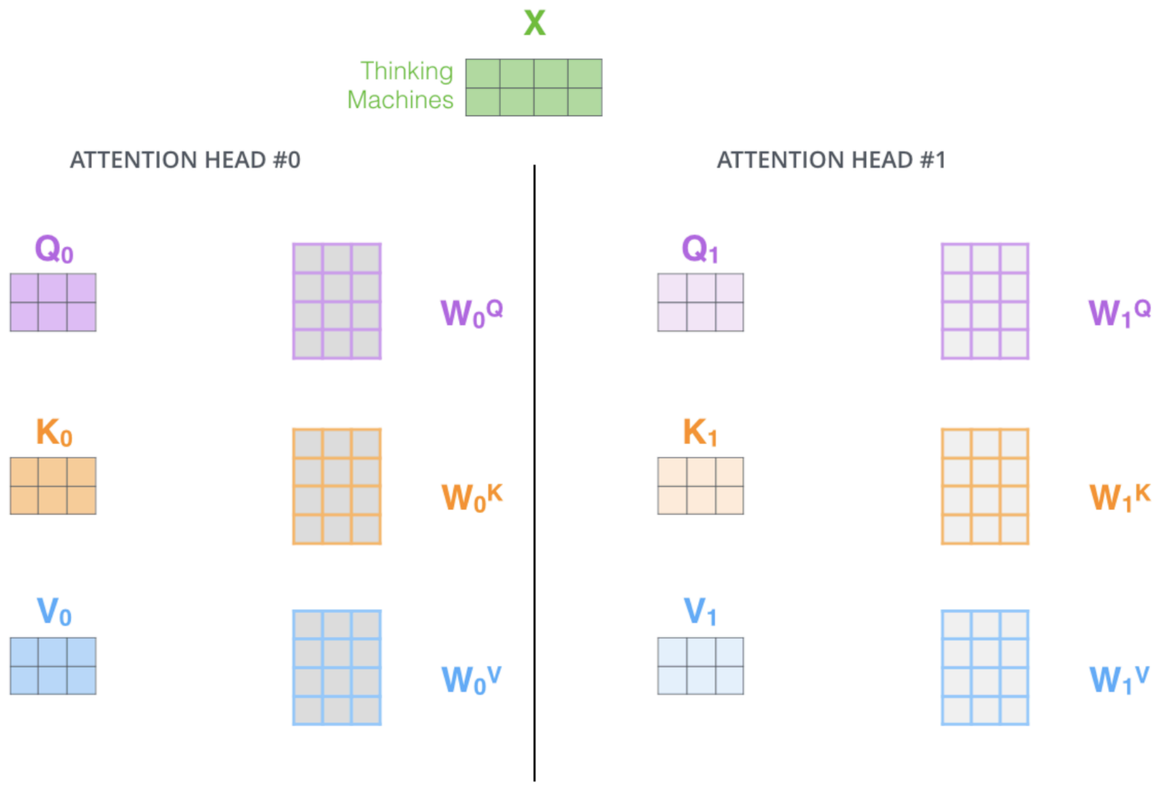
Introduction

When *Vaswani et al.* came out with *Attention is all you need,* the **Transformer** model that came out of it while calculating attention (self as well as enc-dec) used multiple 'heads' (the protagonists of our story)



Introduction

Basically, instead of calculating attention once, we do it multiple times, ***hoping*** (what else did you expect) each *head* captures a different type of attention.



… a beast with many heads

The story

We hoped that the *heads* will capture different aspects of the language and some people have recently explored that.

Head

Papers which form the storyline

1. Analysing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. *ACL 2019*

2. Are Sixteen Heads Really Better than One? *NIPS (Yeah, yeah, NeurIPS) 2019 (Poster)*

3. What Does BERT Look At? An Analysis of BERT's Attention *BlackBoxNLP (Best Paper)*

Two parts of story

1. Interpreting what a head is looking at – *Interpret*

2. How important is that head - *Importance*

Paper-Part Matrix

Paper\Part	Interpret	Importance
ACL 2019	Yes	Yes
NIPS	No	Yes
BlackBoxNLP	Yes	No

ACL 2019 – Analyzing Multi Head Attention

Dataset

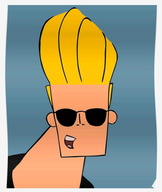
WMT data, Pairs of languages – English to German, French and Russian.

OpenSubtitles2018 – For English to Russian to see the impact when domain changes

ACL 2019 – Analyzing Multi Head Attention

- Heads Importance

Head
Confidence

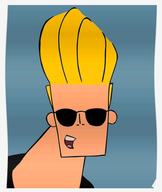


Layer wise
Relevance Propagation

ACL 2019 – Analyzing Multi Head Attention

Heads Importance

Head Confidence

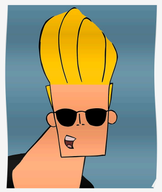


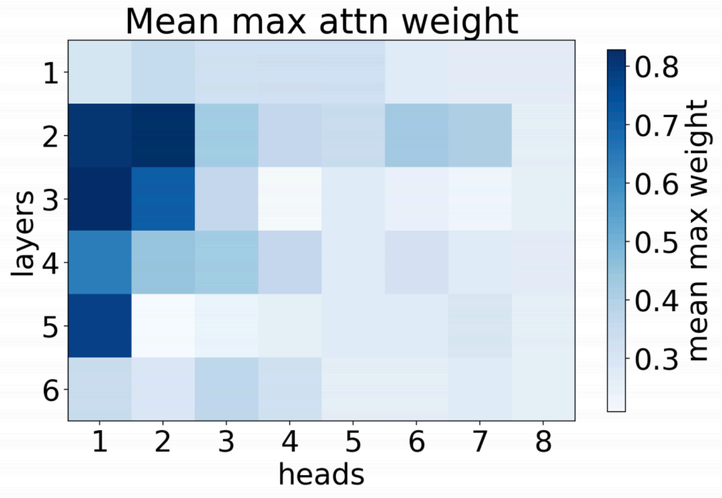
……. the "confidence" of a head as the average of its maximum attention weight excluding the end of sentence symbol,2 where average is taken over tokens in a set of sentences used for evaluation (development set).

ACL 2019 – Analyzing Multi Head Attention

Heads Importance

Head Confidence





ACL 2019 – Analyzing Multi Head Attention

Heads Importance

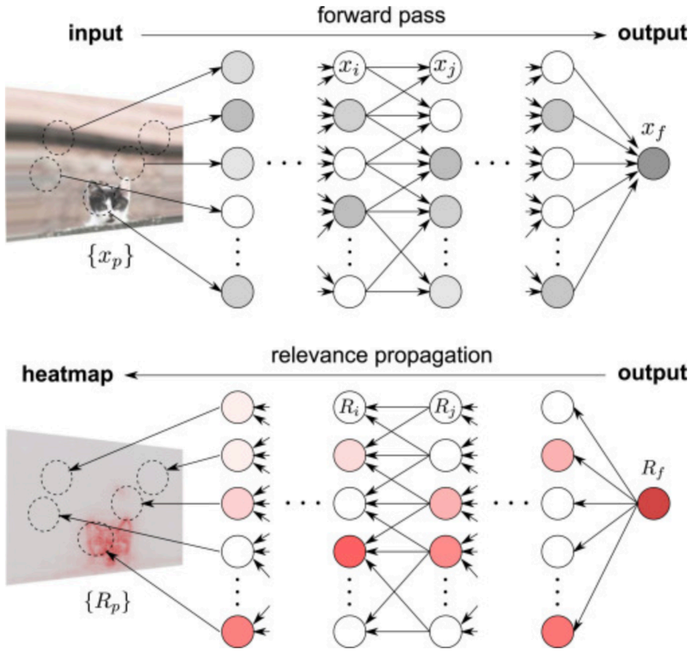
……. Layer-wise relevance propagation (LRP) is a method for computing the relative contribution of neurons at one point in a network to neurons at another.

Layer wise
Relevance Propagation

Adobe

ACL 2019 – Analyzing Multi Head Attention

- Heads Importance



Layer wise
Relevance Propagation

ACL 2019 – Analyzing Multi Head Attention

- Heads Importance

$$f = \ldots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \cdots = \sum_{d} R_d^{(1)}$$

Layer wise
Relevance Propagation

ACL 2019 – Analyzing Multi Head Attention

Heads Importance

Given, *I(v)* denotes all incoming neurons to a neuron *v* and *O(u)* denotes all outgoing neurons from neuron u:

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

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

Layer wise
Relevance Propagation

Adobe

ACL 2019 – Analyzing Multi Head Attention

Heads Importance

Redistribution rule:

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

Layer wise
Relevance Propagation

Adobe

ACL 2019 – Analyzing Multi Head Attention

Heads Importance

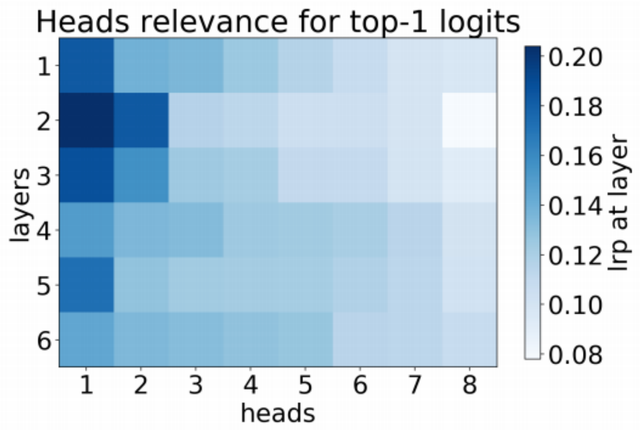
Here, relevance of a head to top-1 logit is calculated by calculating the sum of relevance of the following neurons (loosely called *context vector*) averaged over instances:

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

Layer wise
Relevance Propagation

ACL 2019 – Analyzing Multi Head Attention

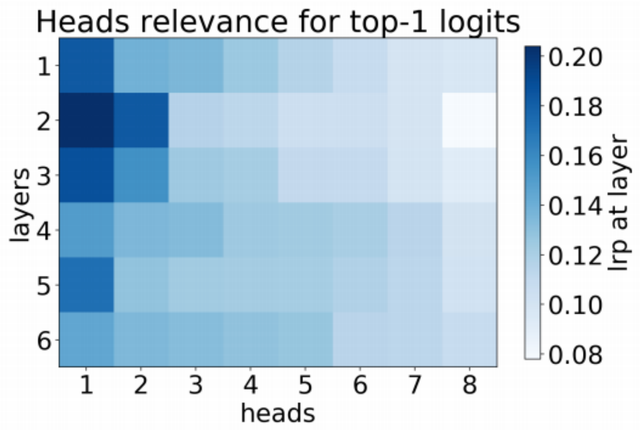
- Heads Importance

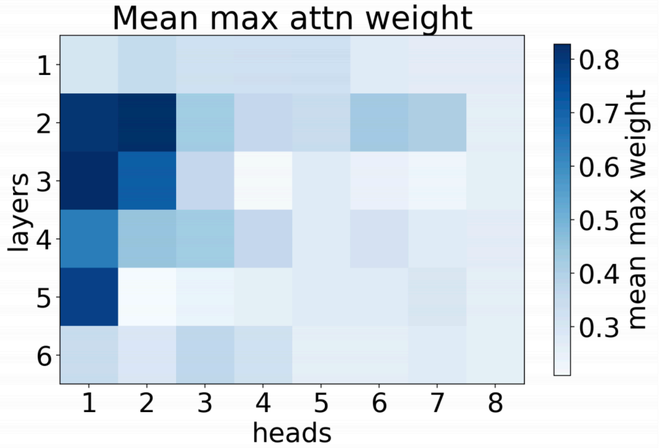


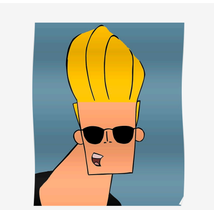
Layer wise
Relevance Propagation

ACL 2019 – Analyzing Multi Head Attention

- Heads Importance







ACL 2019 – Analyzing Multi Head Attention

Heads Interpretability

Positional

Syntactic

Rare

Adobe

ACL 2019 – Analyzing Multi Head Attention

Heads Interpretability

Positional

……..We refer to a head as "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)

ACL 2019 – Analyzing Multi Head Attention

Heads Interpretability

…….A head is said to be syntactic if it attends to tokens corresponding to any of the major syntactic relations in a sentence

Syntactic

ACL 2019 – Analyzing Multi Head Attention

- Heads Interpretability

1. Hold some sentences
2. Run Stanford CoreNLP Parser on them
3. Check how 'accurately' a head assigns its maximum attention weight to a token according to an assigned relationship

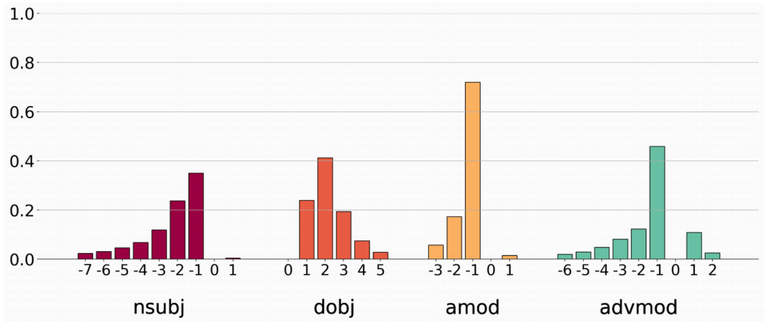
Syntactic

ACL 2019 – Analyzing Multi Head Attention

Heads Interpretability

Baseline:

Some syntactic relationships occur in particular relative positions:



Baseline is looking at the most frequent position for a particular relation

Syntactic

ACL 2019 – Analyzing Multi Head Attention

- Heads Interpretability

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

Syntactic

ACL 2019 – Analyzing Multi Head Attention

Heads Interpretability

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

Syntactic

ACL 2019 – Analyzing Multi Head Attention

Heads Interpretability

What is a rare token?
……least frequent token in a sentence which is not in the top500 most frequent tokens

Rare

ACL 2019 – Analyzing Multi Head Attention

- Heads Interpretability

The first head which we had discussed earlier:

1. OpenSubtitles: *points to the rarest token in 66% of cases, and to one of the two least frequent tokens in 83% of cases*
2. WMT: *points to one of the two least frequent tokens in more than 50% of such cases*

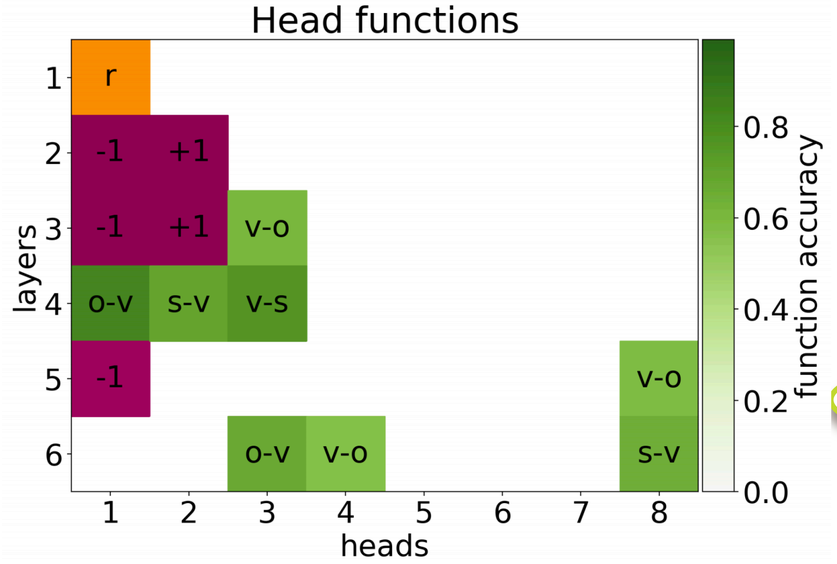
Rare

Adobe

ACL 2019 – Analyzing Multi Head Attention

Heads Interpretability

Positional

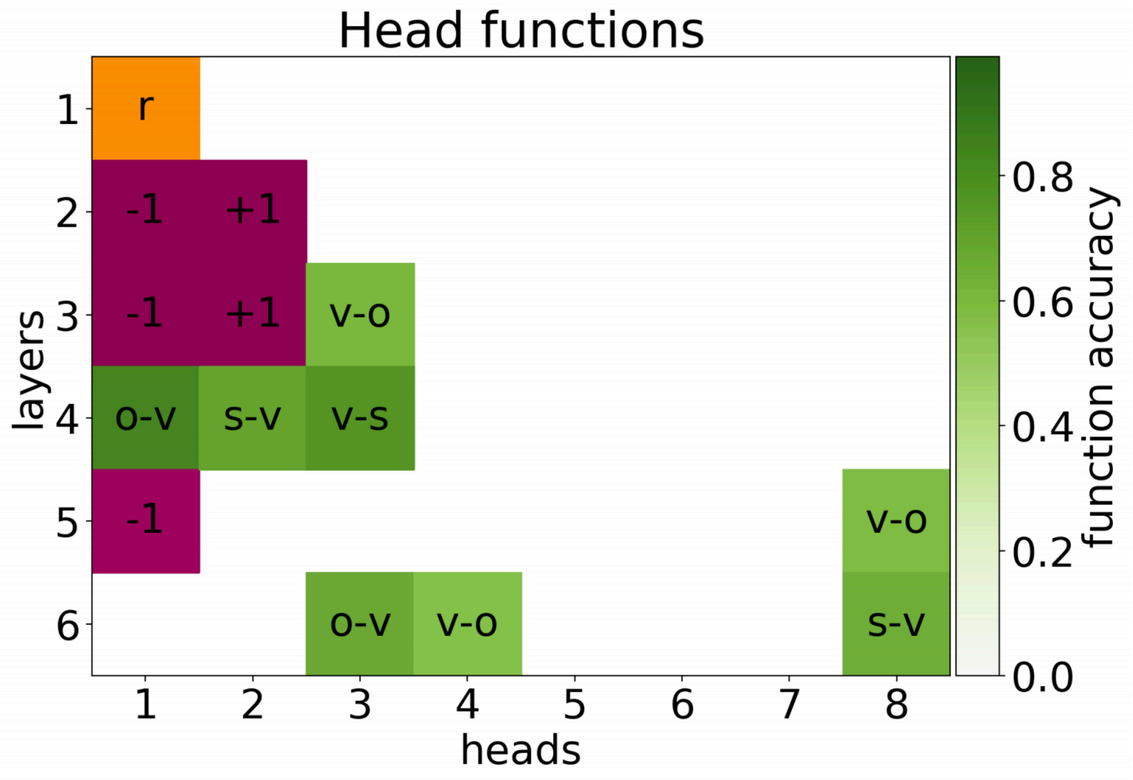


Syntactic

Rare

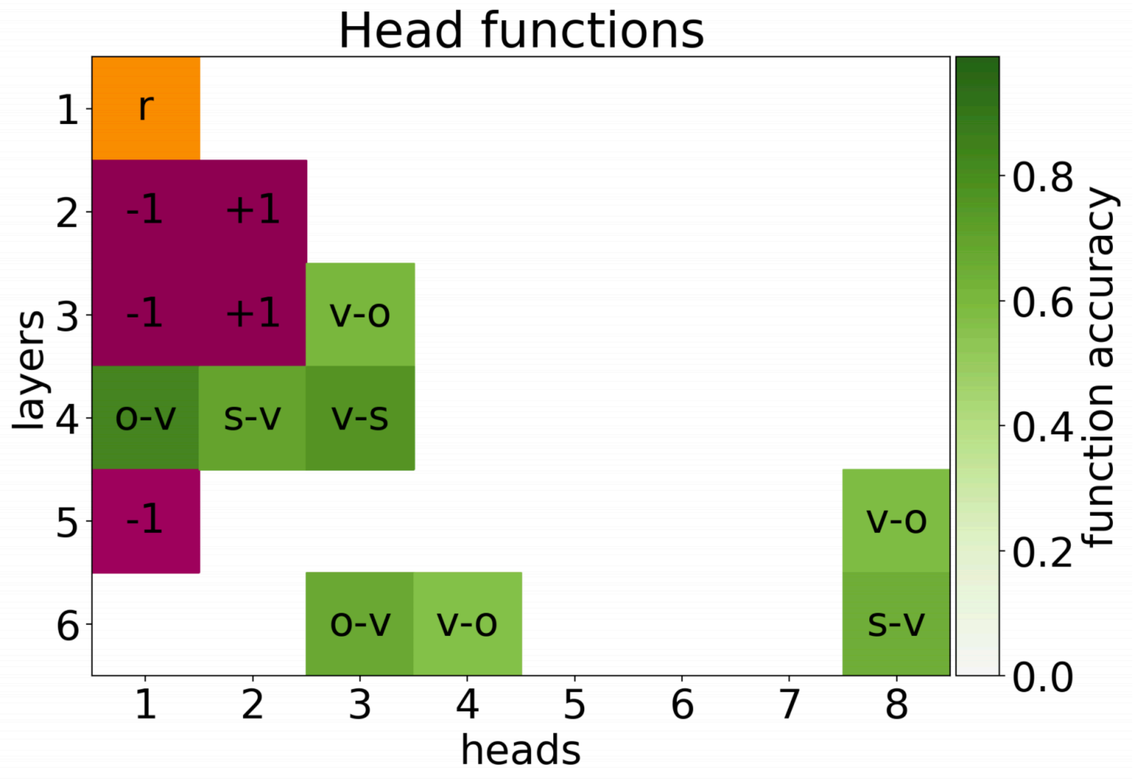
ACL 2019 – Analyzing Multi Head Attention

Heads Interpretability



ACL 2019 – Analyzing Multi Head Attention

Heads Interpretability



ACL 2019 – Analyzing Multi Head Attention

- Heads Importance again : Pruning

ACL 2019 – Analyzing Multi Head Attention

Heads Importance again : Pruning

Method:

1. Modify the original attention formulation as below:

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

2. Ideally, they would be using L0 norm to regularize and prune the heads:

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

ACL 2019 – Analyzing Multi Head Attention

Heads Importance again : Pruning

Method:

3. Since the formulation is non-differentiable, use stochastic relaxation:

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

4. They, that's how, define the final objective as:

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

ACL 2019 – Analyzing Multi Head Attention

- Heads Importance again : Pruning

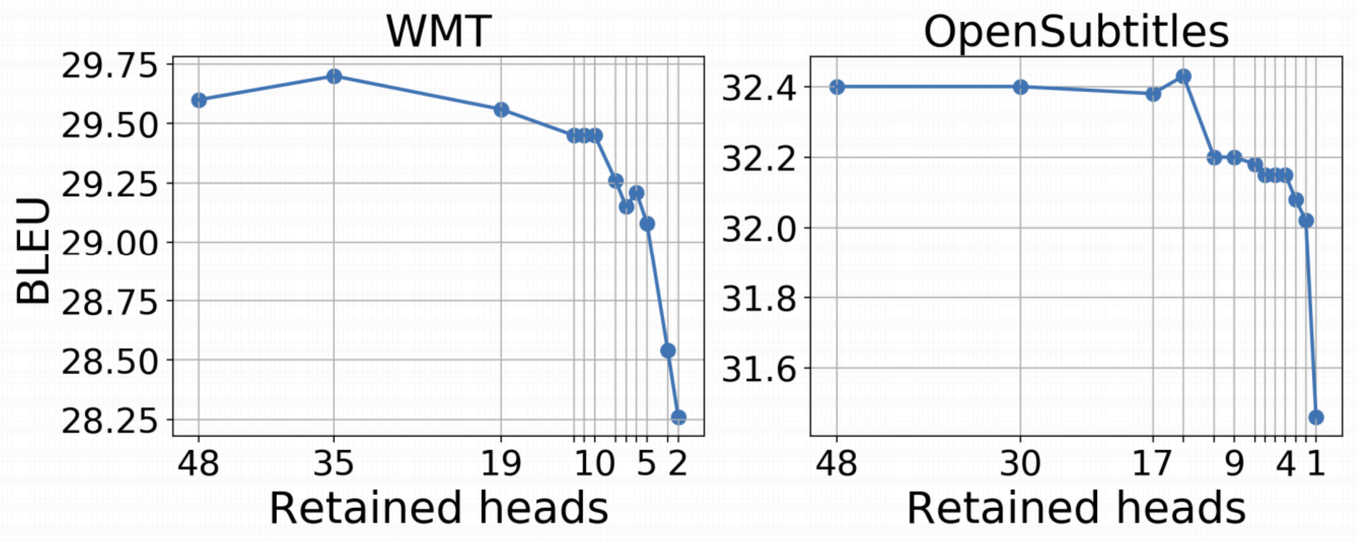
Method:

4. Start from converged model training without L0 penalty and then add these gates and train

5. Increase ƛ to prune more heads

ACL 2019 – Analyzing Multi Head Attention

- Heads Importance again : Pruning Results : Only Encoder Heads



ACL 2019 – Analyzing Multi Head Attention

Heads Importance again : Pruning Results : Only Encoder Heads

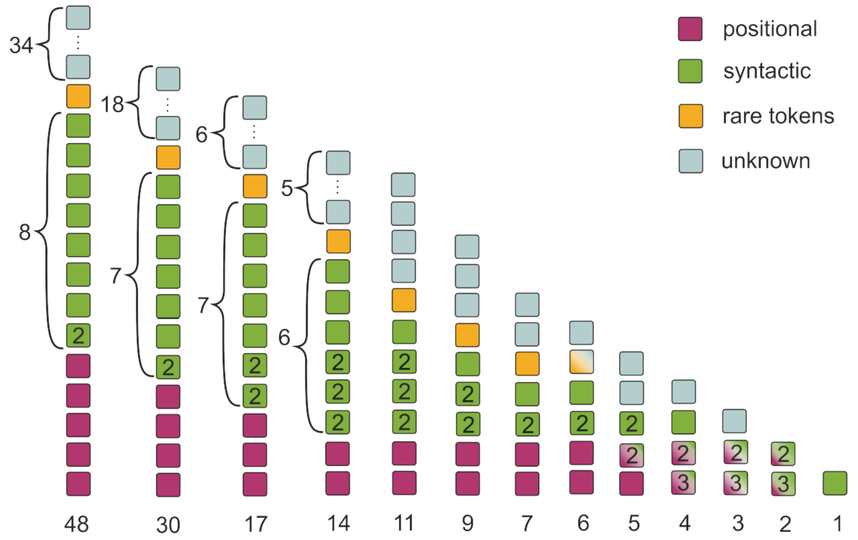


Figure 8: Functions of encoder heads retained after pruning. Each column represents all remaining heads after varying amount of pruning (EN-RU; Subtitles).

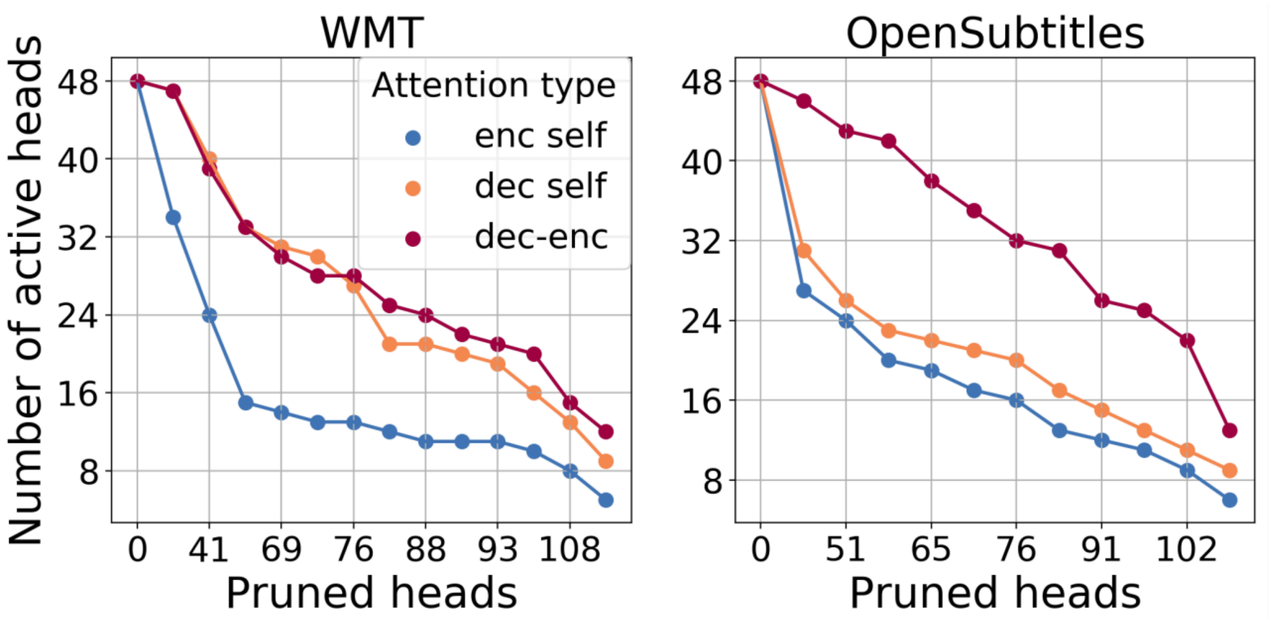
ACL 2019 – Analyzing Multi Head Attention

Heads Importance again : Pruning Results : Prune them all : Train from Scratch

	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

ACL 2019 – Analyzing Multi Head Attention

Heads Importance again : Pruning Results : Prune them all : Who goes first?



ACL 2019 – Analyzing Multi Head Attention

- Takeaways

1. Only a small subset of heads appear to be important for the translation task (*Hint : Three people who are very important)*

2. Important heads have one or more interpretable functions in the model, including attending to adjacent words and tracking specific syntactic relations

3. Specialized heads are the last to be pruned, confirming their importance directly. Moreover, the vast majority of heads, especially the encoder self-attention heads, can be removed without seriously affecting performance

NIPS– Are 16 heads really better than one?

Datasets

WMT – For original *Vaswani et al* Model

MultiNLI – For BERT Model

Only Pretrained Models are used

NIPS– Are 16 heads really better than one?

Heads Importance again : Pruning (Actually Just blatantly removing)

Directly replace the head results with zeros and notice the drop in accuracy on validation set

Before pruning:

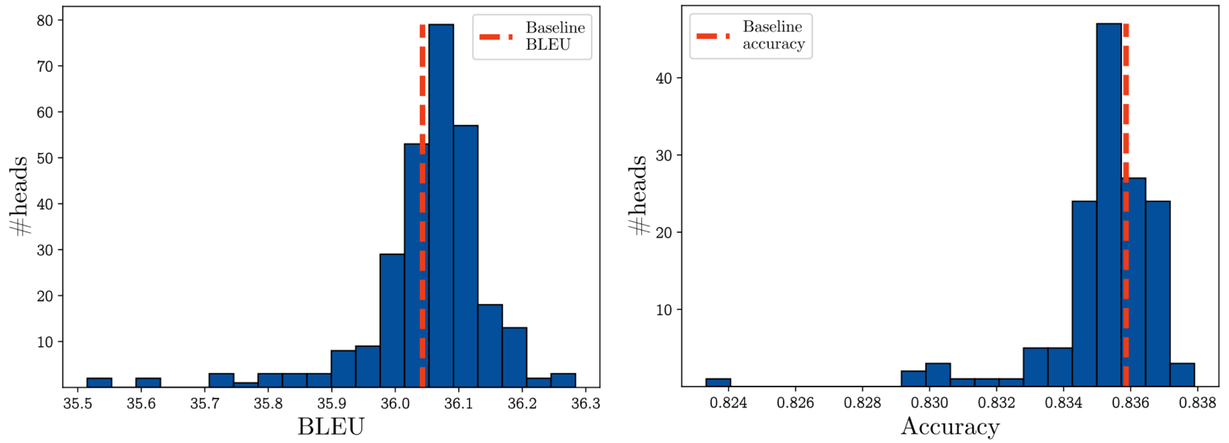
WMT Model has 96 heads (16 heads per layer, 6 layers)

BERT has 144 heads (12 heads per layer, 12 layers)

Adobe

NIPS– Are 16 heads really better than one?

Heads Importance again : Pruning (Actually Just blatantly removing) : Encoder heads



(a) WMT (b) BERT

Figure 1: Distribution of heads by model score after masking.

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	**-0.53**	0.09	**-0.33**	0.06	0.03	0.11	0.04	0.01	-0.04	0.04	0.00
2	0.01	0.04	0.10	**0.20**	0.06	0.03	0.00	0.09	0.10	0.04	**0.15**	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	**0.24**	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	**-0.31**	-0.11	-0.04	0.12	0.10	0.02	0.09	0.08	0.04	**0.21**	-0.02	0.02	-0.03	-0.04	0.07	-0.02
6	0.06	0.07	**-0.31**	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

NIPS– Are 16 heads really better than one?

- Heads Importance again : Ablating all heads in a single layer but one

Layer	Enc-Enc	Enc-Dec	Dec-Dec
1	**-1.31**	**0.24**	-0.03
2	-0.16	0.06	0.12
3	0.12	0.05	0.18
4	-0.15	-0.24	0.17
5	0.02	**-1.55**	-0.04
6	**-0.36**	**-13.56**	0.24

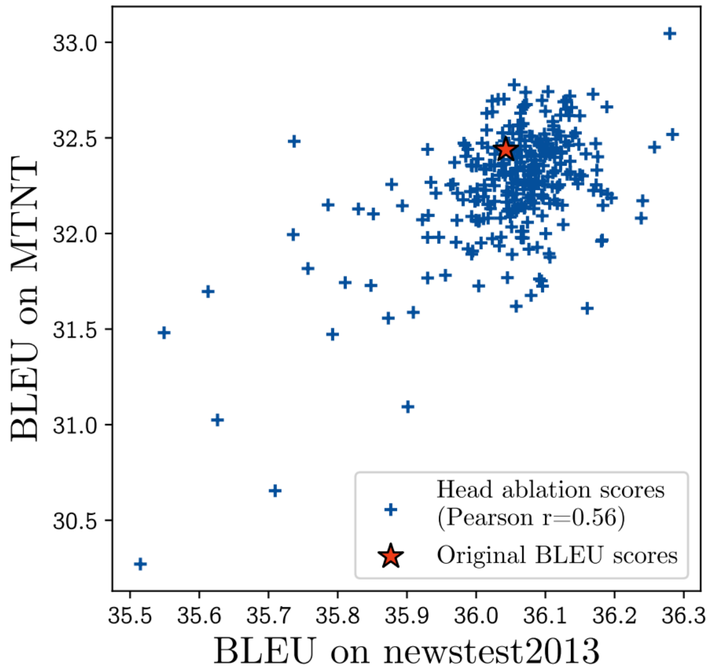
Table 2: Best delta BLEU by layer when only one head is kept in the WMT model. Underlined numbers indicate that the change is statistically significant with $p < 0.01$.

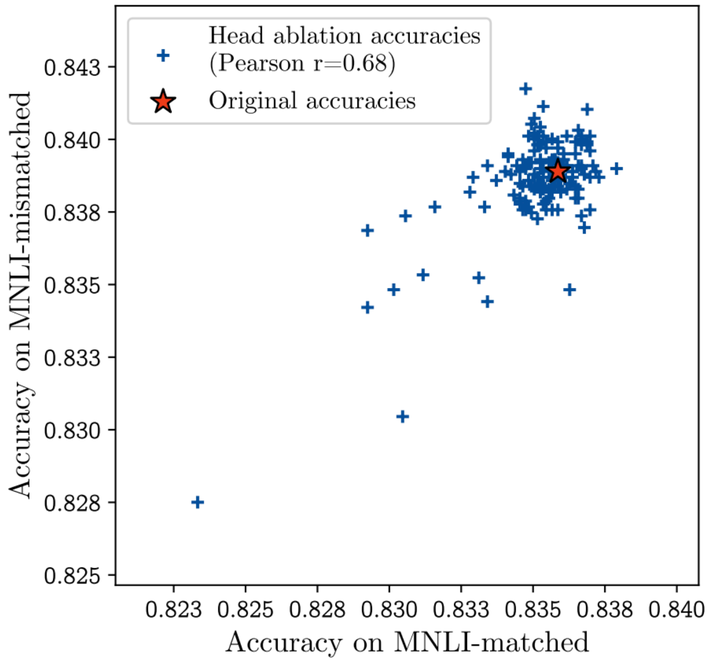
Layer		Layer	
1	-0.01%	7	0.05%
2	0.10%	8	-0.72%
3	-0.14%	9	-0.96%
4	-0.53%	10	0.07%
5	-0.29%	11	-0.19%
6	-0.52%	12	-0.12%

Table 3: Best delta accuracy by layer when only one head is kept in the BERT model. None of these results are statistically significant with $p < 0.01$.

NIPS– Are 16 heads really better than one?

Heads Importance again : Are Important Heads the Same Across Datasets?





NIPS– Are 16 heads really better than one?

Heads Importance again : Pruning in a similar manner to ACL 2019

1. Formulate Multi Head Attention in following manner:

$$\mathbf{MHAtt}(\mathbf{x}, q) = \sum_{h=1}^{N_h} \xi_h \mathbf{Att}_{W_k^h, W_q^h, W_v^h, W_o^h}(\mathbf{x}, q)$$

2. Calculate the importance of head as how sensitive output is to this head's value:

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

NIPS– Are 16 heads really better than one?

Heads Importance again : Pruning in a similar manner to ACL 2019

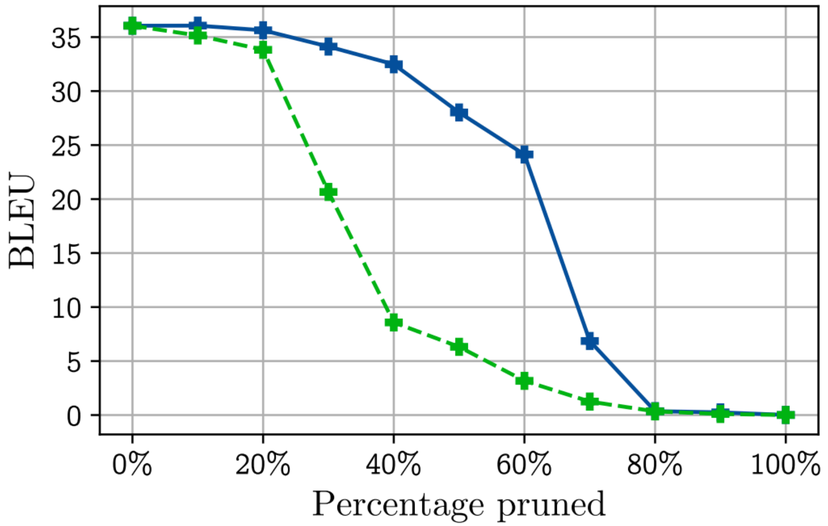
3. Now prune heads in order of importance score as obtained above.

4. Also prune heads in order of loss in BLEU Score

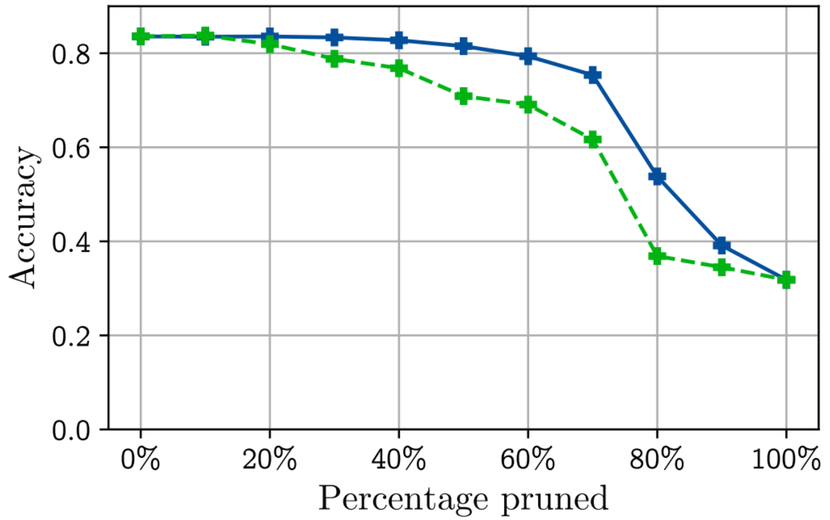
5. Compare

NIPS– Are 16 heads really better than one?

- Heads Importance again : Pruning in a similar manner to ACL 2019



(a) Evolution of BLEU score on `newstest2013` when heads are pruned from WMT according to I_h (solid blue) and BLEU difference (dashed green).



(b) Evolution of accuracy on the MultiNLI-matched validation set when heads are pruned from BERT according to I_h (solid blue) and accuracy difference (dashed green).

NIPS– Are 16 heads really better than one?

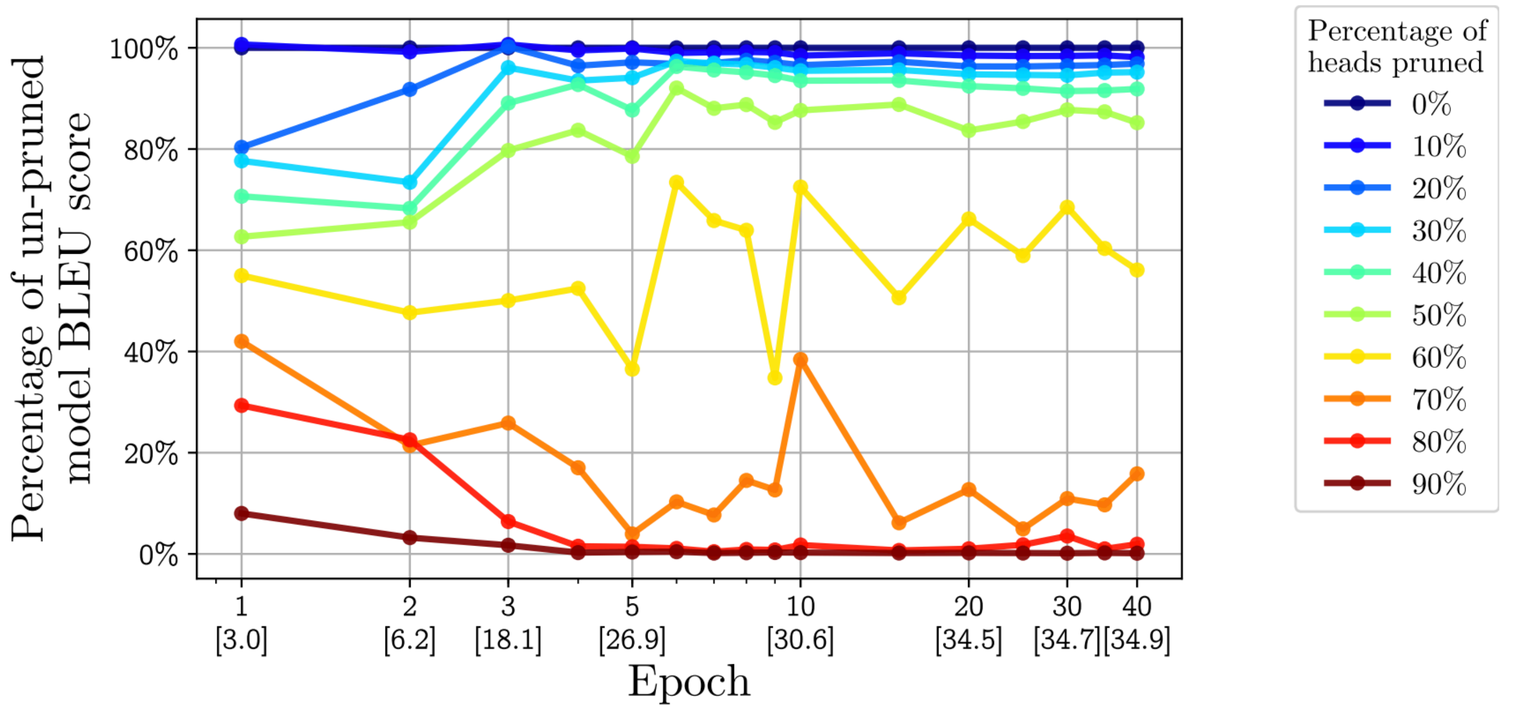
Heads Importance again : Pruning in a similar manner to ACL 2019 : Effect on Efficiency

Batch size	1	4	16	64
Original	17.0 ± 0.3	67.3 ± 1.3	114.0 ± 3.6	124.7 ± 2.9
Pruned (50%)	17.3 ± 0.6 (+1.9%)	69.1 ± 1.3 (+2.7%)	134.0 ± 3.6 (+17.5%)	146.6 ± 3.4 (+17.5%)

Table 4: Average inference speed of BERT on the MNLI-matched validation set in examples per second ($\pm$ standard deviation). The speedup relative to the original model is indicated in parentheses.

NIPS– Are 16 heads really better than one?

Heads Importance again : Pruning in a similar manner to ACL 2019 : When is Imp determined



NIPS– Are 16 heads really better than one?

- Takeaways

1. Same as earlier paper more or less
2. Not all heads are important (duh)
3. Enc-Dec heads are most important as compared to Enc only heads and dec only heads

BlackBoxNLO– What does BERT look at?

Datasets

They use a pretrained standard BERT model and gather the attention results on 1000 random Wikipedia sentences/segments

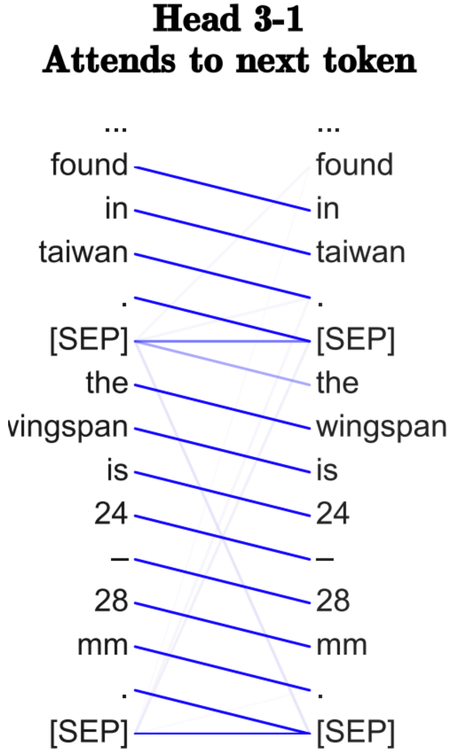
Also, while interpreting dependencies, they use Penn Treebank annotated with Stanford dependencies

CoNLL-2012 dataset for coreference resolution

BlackBoxNLP– What does BERT look at?

Heads Interpretability : Surface level Patterns : Positional Heads *(déjà vu?)*

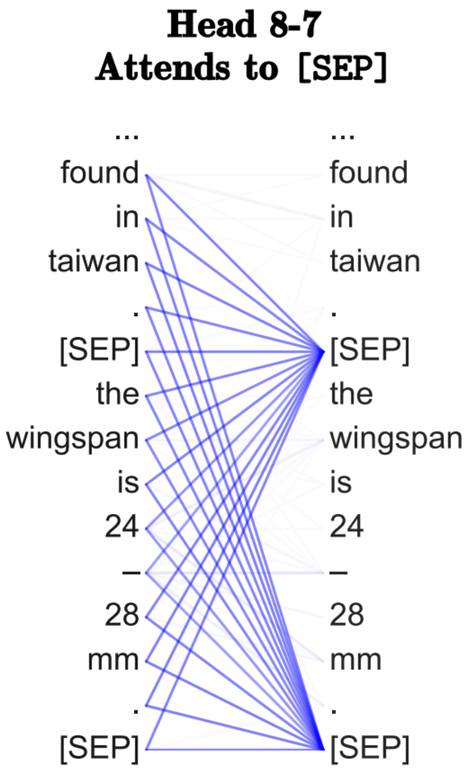
There are heads that specialize to attending heavily on the next or previous token, especially in earlier layers of the network.

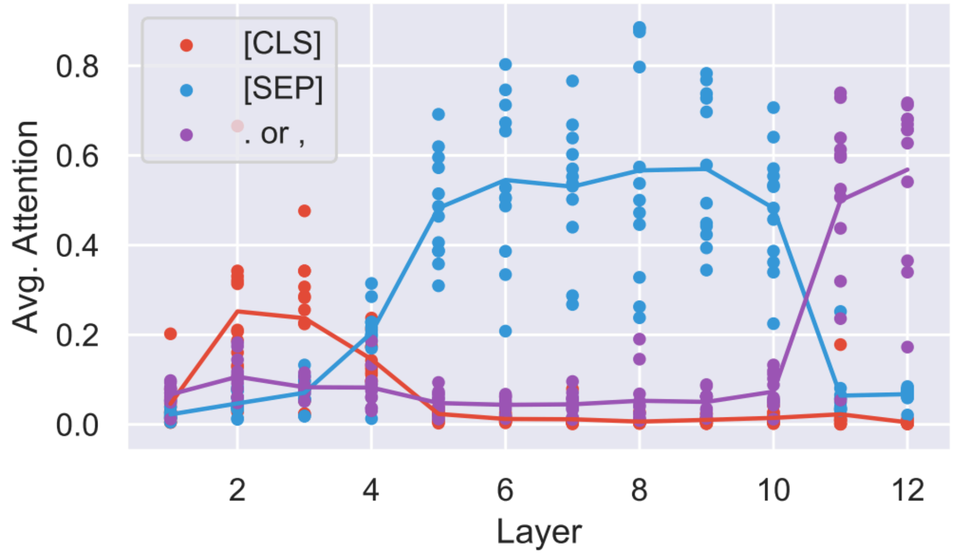


BlackBoxNLP– What does BERT look at?

Heads Interpretability : Surface level Patterns : Attending to SEP token

An abnormal/substantial amount of attention of BERT is on SEP token



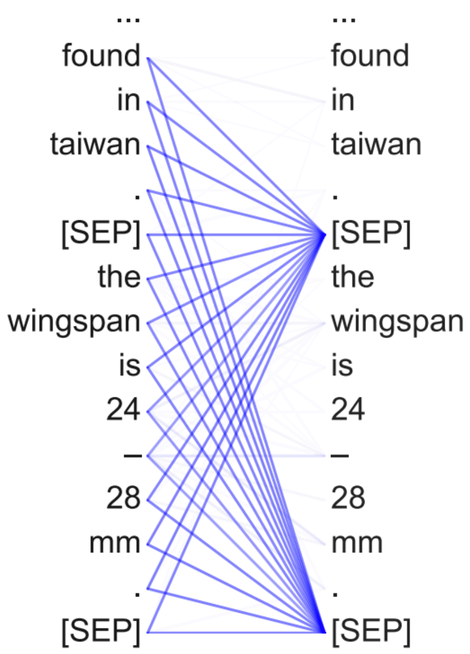


BlackBoxNLP– What does BERT look at?

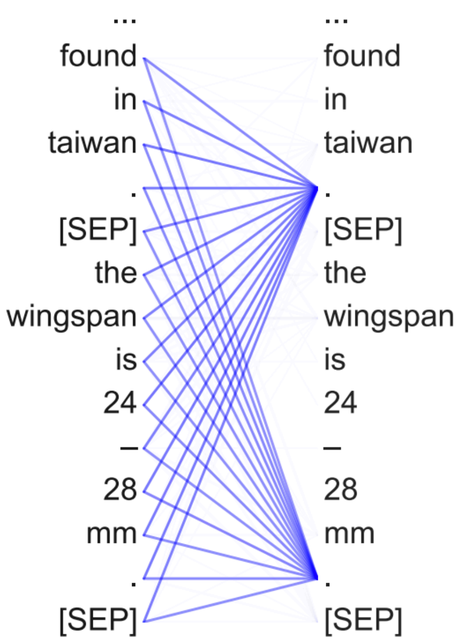
- Heads Interpretability : Surface level Patterns : Attending to SEP token

.... And on some other specific tokens

Head 8-7
Attends to [SEP]



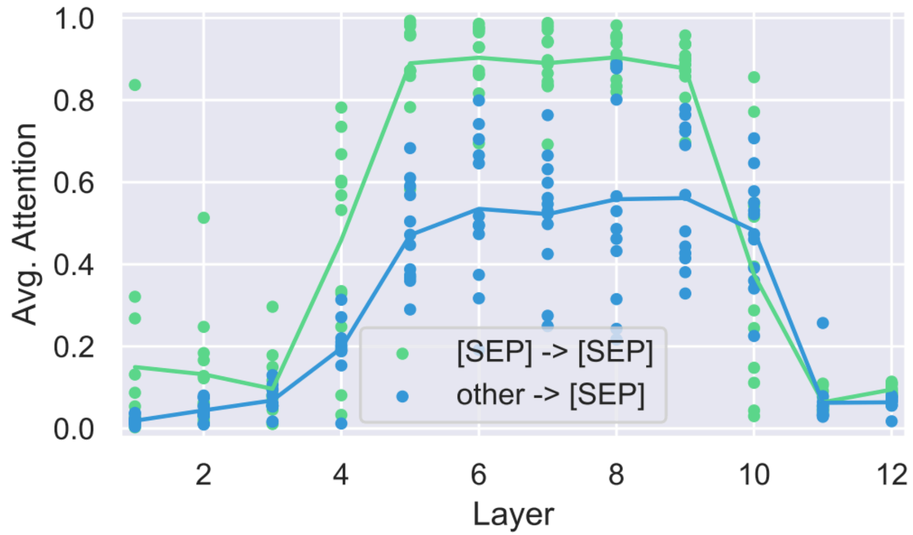
Head 11-6
Attends to periods



BlackBoxNLP– What does BERT look at?

Heads Interpretability : Surface level Patterns : Attending to SEP token : Explanation and Details

1. [SEP] is used to aggregate segment-level information which can then be read by other heads? If that was true, we would expect attention heads processing [SEP] to attend broadly over the whole segment to build up these representations. However, it almost always attend to itself



BlackBoxNLP– What does BERT look at?

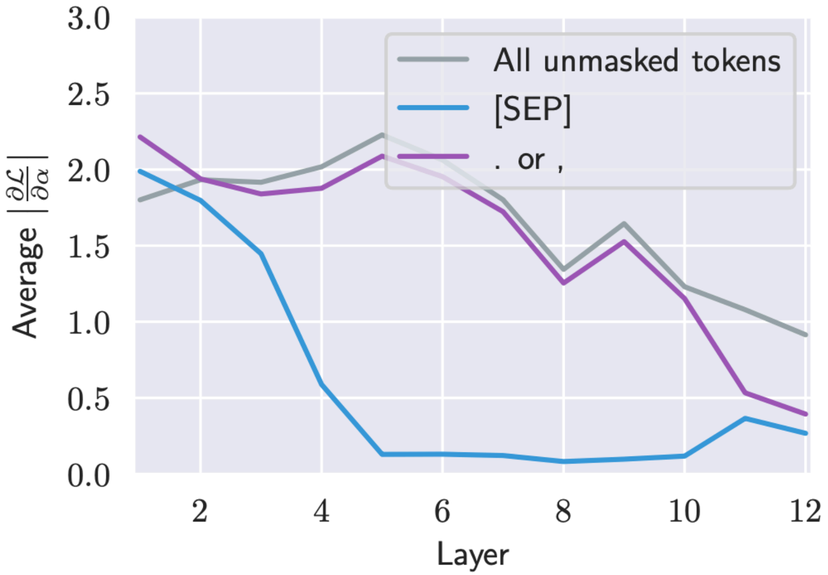
Heads Interpretability : Surface level Patterns : Attending to SEP token : Explanation and Details

2. Qualitative analysis shows that heads with specific functions attend to SEP when the function is not called for. May be it is being used as *no-op?*

BlackBoxNLP– What does BERT look at?

- Heads Interpretability : Surface level Patterns : Attending to SEP token : Explanation and Details

3. Apply gradient-based measures of feature importance. Starting in layer 5 – the same layer where attention to [SEP] becomes high – the gradients for attention to [SEP] become very small. Conclusion? It is indeed being used as no-op



BlackBoxNLP– What does BERT look at?

Heads Interpretability : Syntactic : He is back

Syntactic

BlackBoxNLP– What does BERT look at?

- Heads Interpretability : Syntactic

Just like ACL 2019, they also put simple fixed offset as baseline and compare

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)

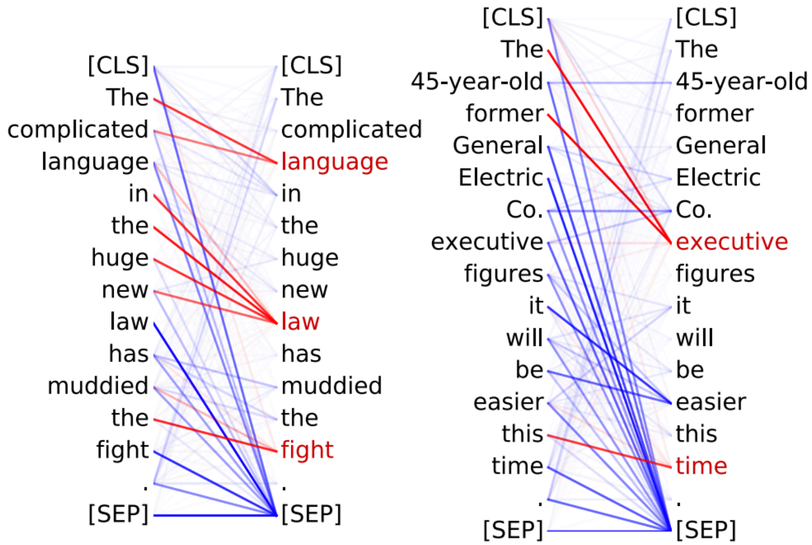
BlackBoxNLP– What does BERT look at?

Heads Interpretability : Syntactic

One head generally does good job at one relation but no head does good enough for all

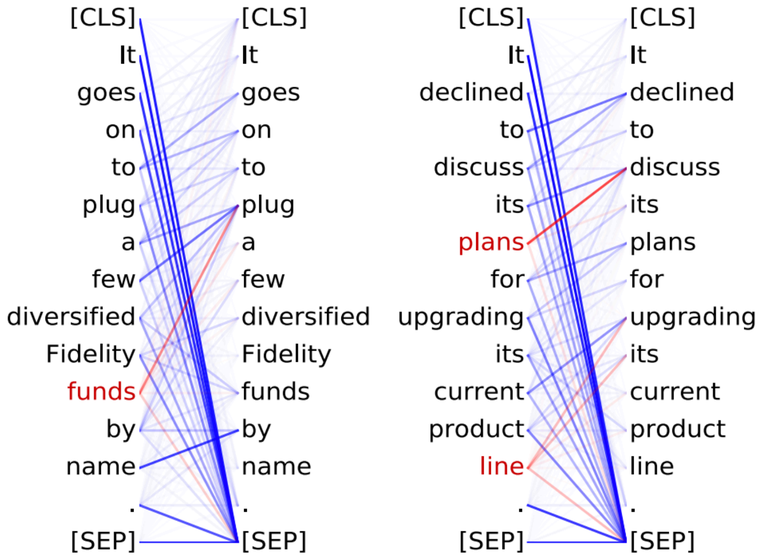
Head 8-11

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



Head 8-10

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



BlackBoxNLP– What does BERT look at?

Heads Interpretability : Coreference Resolution

Best head of BERT achieves reasonable preference on coreference resolution task

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.

BlackBoxNLP– What does BERT look at?

- Heads Interpretability : Going beyond single head : Probing the heads together : Attention Only

  Use the following combination of attention weights to calculate the probability of word *i* being word *j*'s syntactic head and train them using supervised learning :

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

BlackBoxNLP– What does BERT look at?

Heads Interpretability : Going beyond single head : Probing the heads together : Attention + Word

They added word's information too along with attention weights:

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

BlackBoxNLP– What does BERT look at?

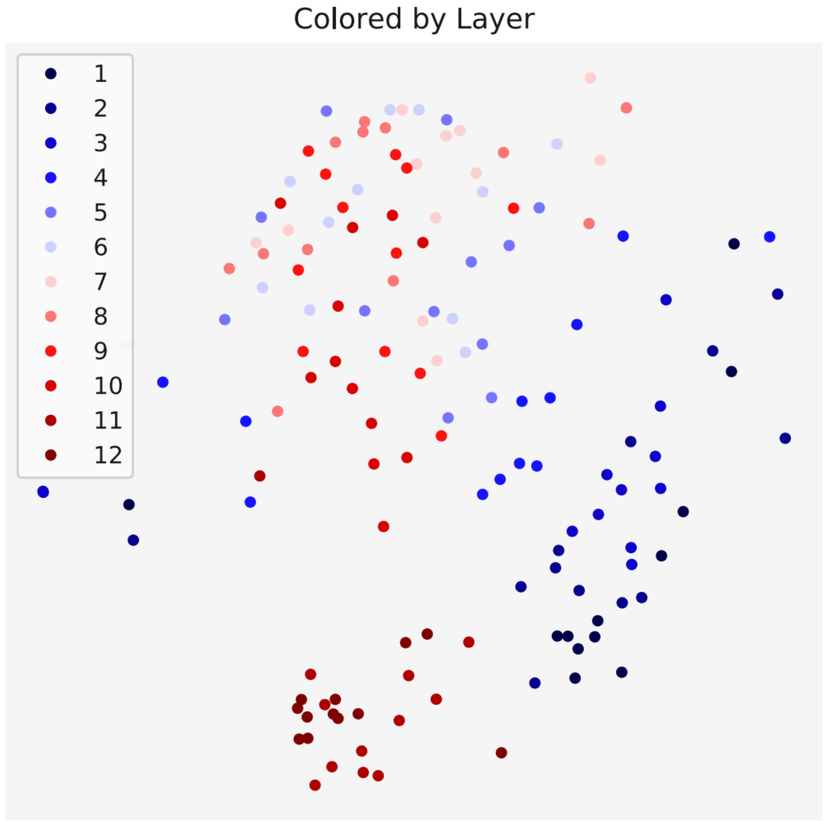
Heads Interpretability : Going beyond single head : Probing the heads together : Results

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

BlackBoxNLP– What does BERT look at?

Heads Interpretability : Similar heads in same layer : Clustering heads

Measure distance between two heads as : $\sum_{\text{token} \in \text{data}} JS(\mathbf{H}_i(\text{token}), \mathbf{H}_j(\text{token}))$



BlackBoxNLP– What does BERT look at?

- Takeaways

1. Many of the heads can be very well interpreted to be performing different specific tasks

2. BERT's attention carries a sufficient amount of syntactical structure of English language

3. Probing attention maps provides a sufficient empirical proof to point 2

4. BERT uses SEP token for no-op purpose

What do the heads say? Takeaway

1. A recent surge in dissecting attention maps has led to interesting findings
2. With transformers, came the MultiHead attention which performs *attention* multiple times hoping different heads will become specialized in different roles which they do
3. Despite that, many heads are mostly redundant and can be easily pruned with zero to negligible impact on the performance of models
4. Heads capture a lot of syntactic information and become specialized in different roles such as focusing on rare words, finding different syntactical relations or focusing on different positional offsets
5. Heads say a lot. You just need right ears (probes) to hear them.

Discuss


Adobe


Adobe