BERTology Nuggets:

what we've learned about how BERT works

University of Helsinki, Nov. 25 2021

Anna Rogers
https://annargrs.github.io
@annargrs

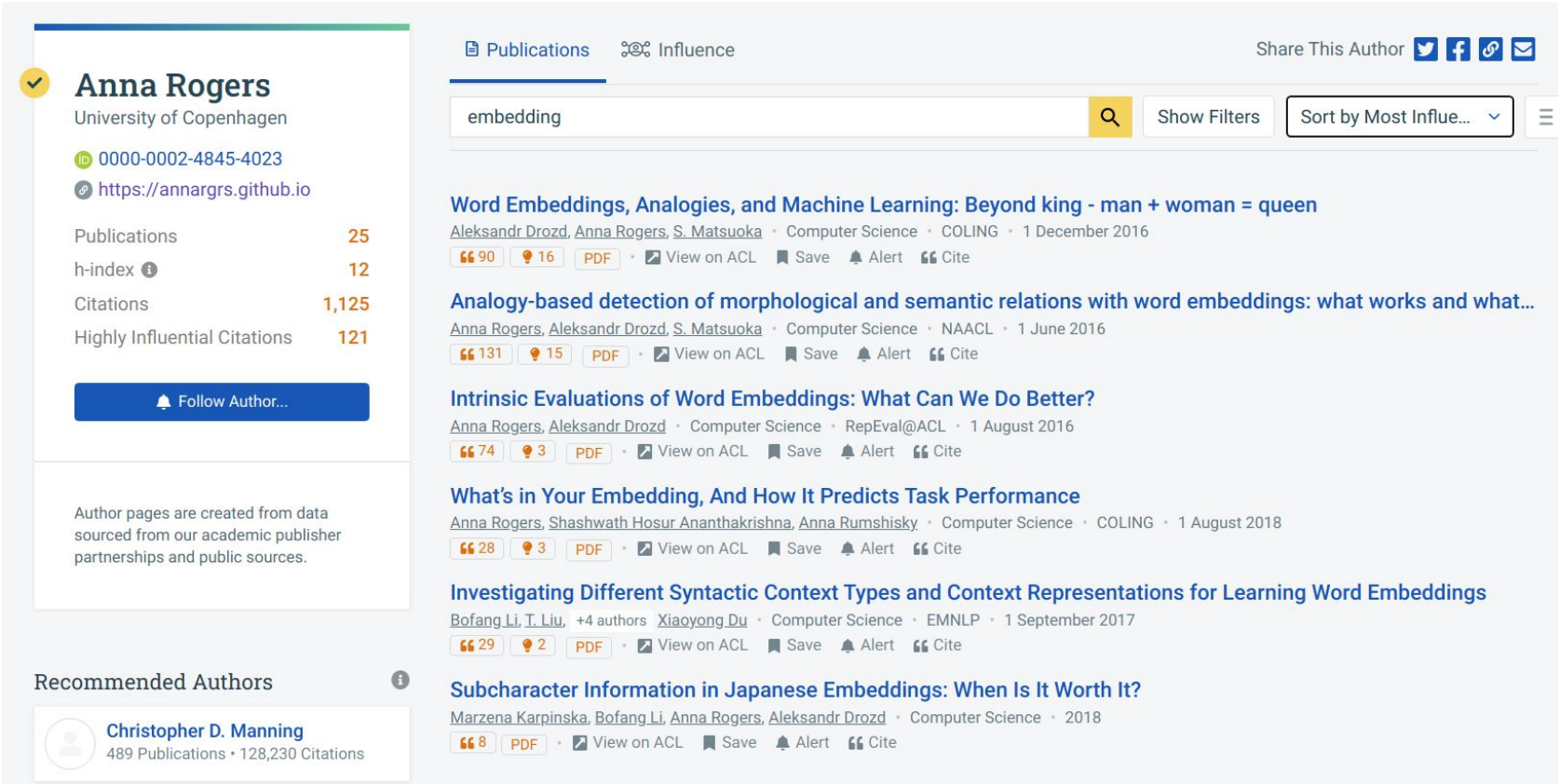
RIKEN
Center for
Computational Science

slides (with links to papers):
https://annargrs.github.io/talks

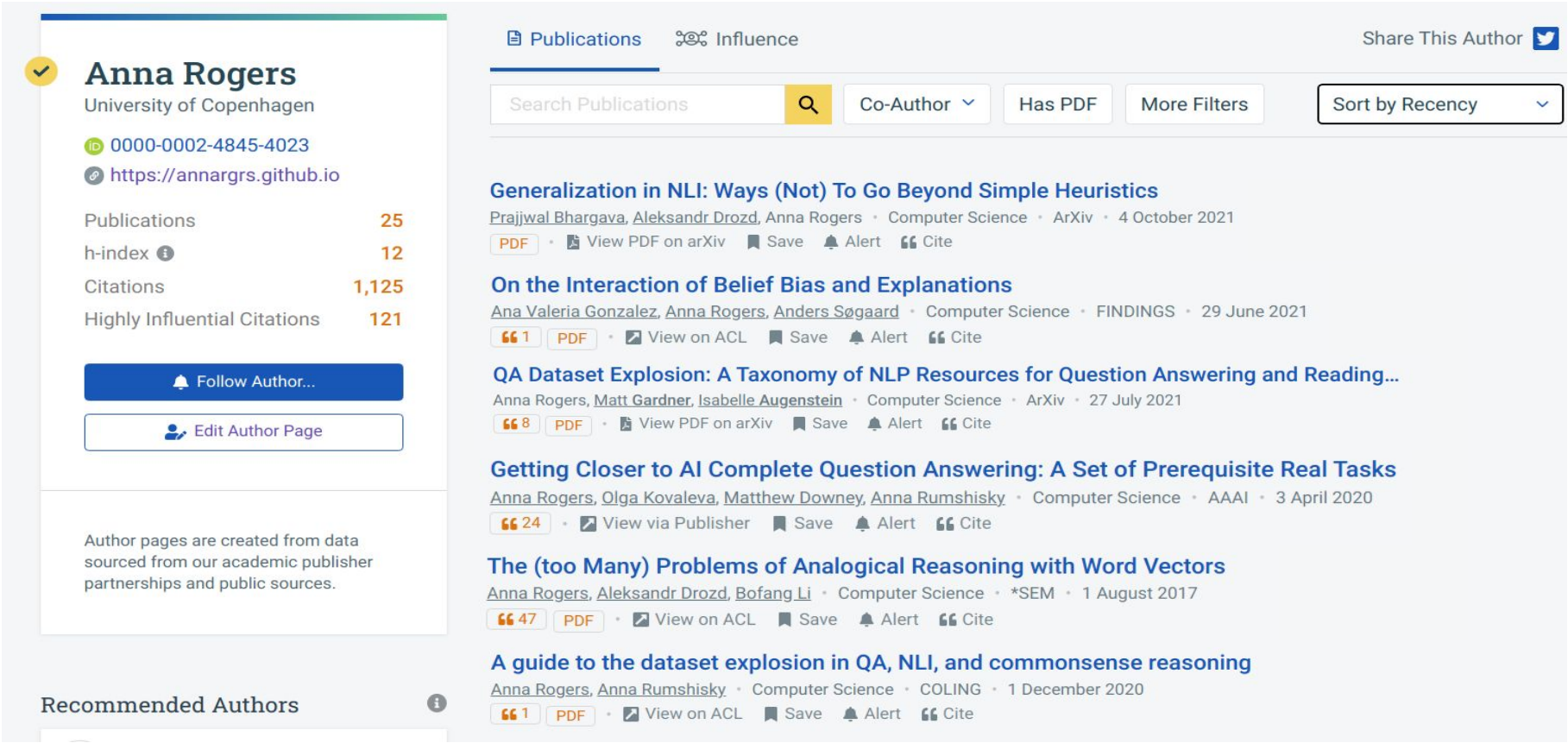
Anna Rogers

2017: Ph.D. in computational linguistics,
University of Tokyo
2017-2020: Postdoc in ML for NLP,
University of Massachusetts
2020 ~ : postdoc in social data science,
Copenhagen University
2021 ~ : visiting researcher,
RIKEN Center for Computational Science (Japan)

I study embeddings...



and whether they enable "reasoning"...



Anna Rogers
University of Copenhagen
0000-0002-4845-4023
https://annargrs.github.io

Publications	25
h-index	12
Citations	1,125
Highly Influential Citations	121

Follow Author...
Edit Author Page

Author pages are created from data sourced from our academic publisher partnerships and public sources.

Recommended Authors

Publications | Influence | Share This Author

Search Publications | Co-Author | Has PDF | More Filters | Sort by Recency

Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics
Prajjwal Bhargava, Aleksandr Drozd, Anna Rogers · Computer Science · ArXiv · 4 October 2021
PDF · View PDF on arXiv · Save · Alert · Cite

On the Interaction of Belief Bias and Explanations
Ana Valeria Gonzalez, Anna Rogers, Anders Søgaard · Computer Science · FINDINGS · 29 June 2021
1 · PDF · View on ACL · Save · Alert · Cite

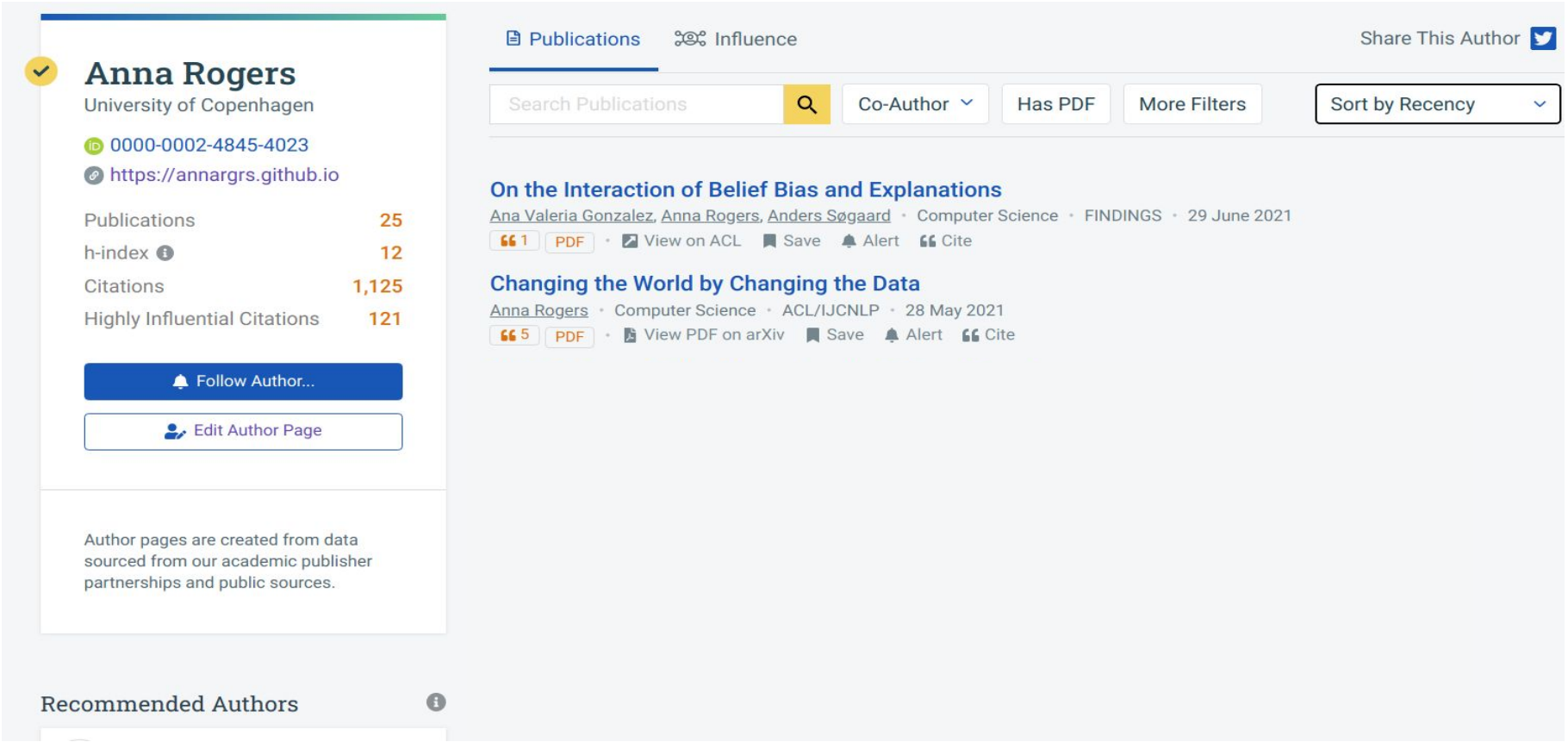
QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading...
Anna Rogers, Matt Gardner, Isabelle Augenstein · Computer Science · ArXiv · 27 July 2021
8 · PDF · View PDF on arXiv · Save · Alert · Cite

Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks
Anna Rogers, Olga Kovaleva, Matthew Downey, Anna Rumshisky · Computer Science · AAAI · 3 April 2020
24 · View via Publisher · Save · Alert · Cite

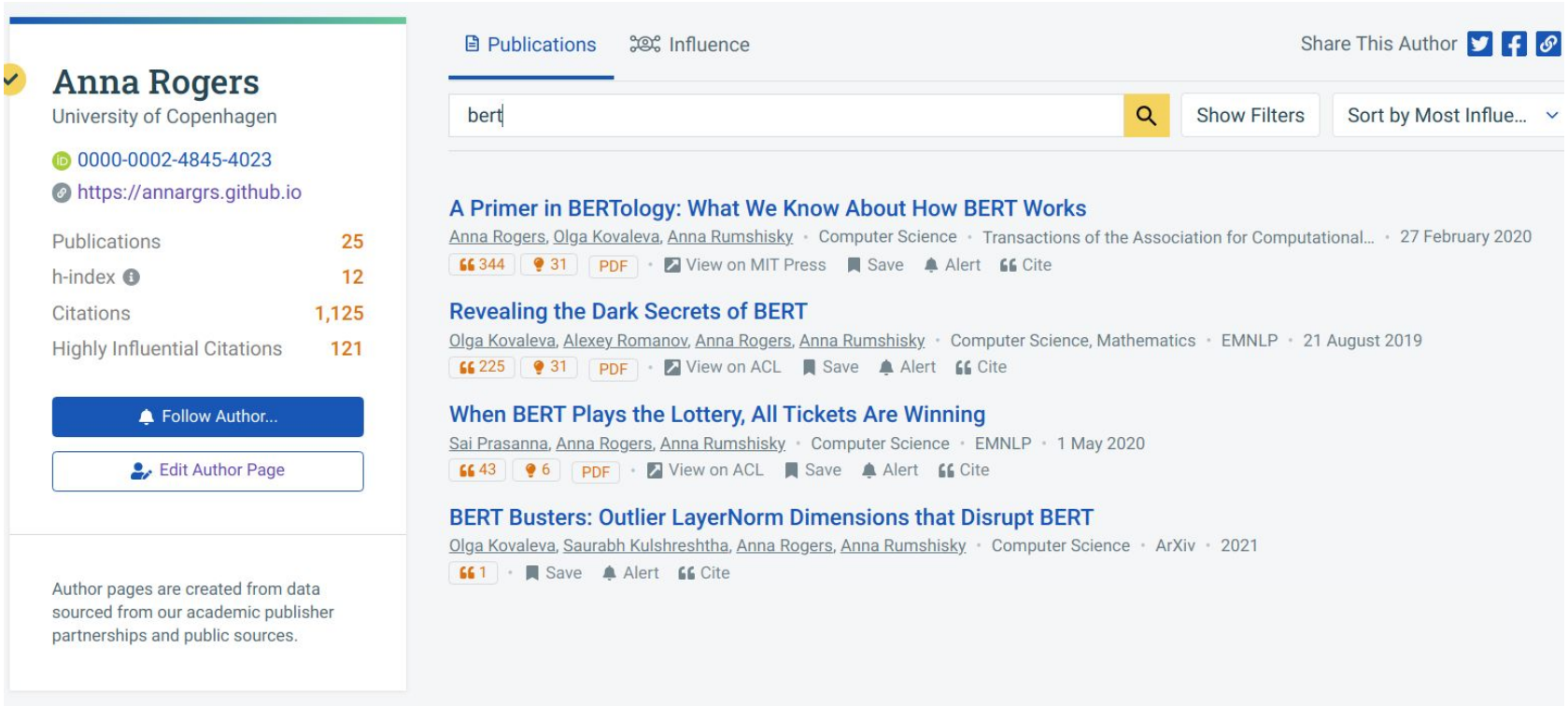
The (too Many) Problems of Analogical Reasoning with Word Vectors
Anna Rogers, Aleksandr Drozd, Bofang Li · Computer Science · *SEM · 1 August 2017
47 · PDF · View on ACL · Save · Alert · Cite

A guide to the dataset explosion in QA, NLI, and commonsense reasoning
Anna Rogers, Anna Rumshisky · Computer Science · COLING · 1 December 2020
1 · PDF · View on ACL · Save · Alert · Cite

… and how they can impact the world…



… and BERT.



Anna Rogers
University of Copenhagen
0000-0002-4845-4023
https://annargrs.github.io

Publications 25
h-index 12
Citations 1,125
Highly Influential Citations 121

Follow Author...
Edit Author Page

Author pages are created from data sourced from our academic publisher partnerships and public sources.

Publications | Influence
Share This Author

bert
Show Filters
Sort by Most Influe...

A Primer in BERTology: What We Know About How BERT Works
Anna Rogers, Olga Kovaleva, Anna Rumshisky · Computer Science · Transactions of the Association for Computational… · 27 February 2020
344 31 PDF · View on MIT Press Save Alert Cite

Revealing the Dark Secrets of BERT
Olga Kovaleva, Alexey Romanov, Anna Rogers, Anna Rumshisky · Computer Science, Mathematics · EMNLP · 21 August 2019
225 31 PDF · View on ACL Save Alert Cite

When BERT Plays the Lottery, All Tickets Are Winning
Sai Prasanna, Anna Rogers, Anna Rumshisky · Computer Science · EMNLP · 1 May 2020
43 6 PDF · View on ACL Save Alert Cite

BERT Busters: Outlier LayerNorm Dimensions that Disrupt BERT
Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, Anna Rumshisky · Computer Science · ArXiv · 2021
1 · Save Alert Cite

In this talk:

- BERT model recap
- What does attention do for BERT?
- How does BERT learn?
- Does lottery ticket hypothesis hold for BERT?
- What does it take to disrupt BERT?
- What does BERT learn?

My amazing collaborators

UMASS LOWELL

Olga Kovaleva
Anna Rumshisky

Sai Prasanna

UTD

Prajjwal Bhargava

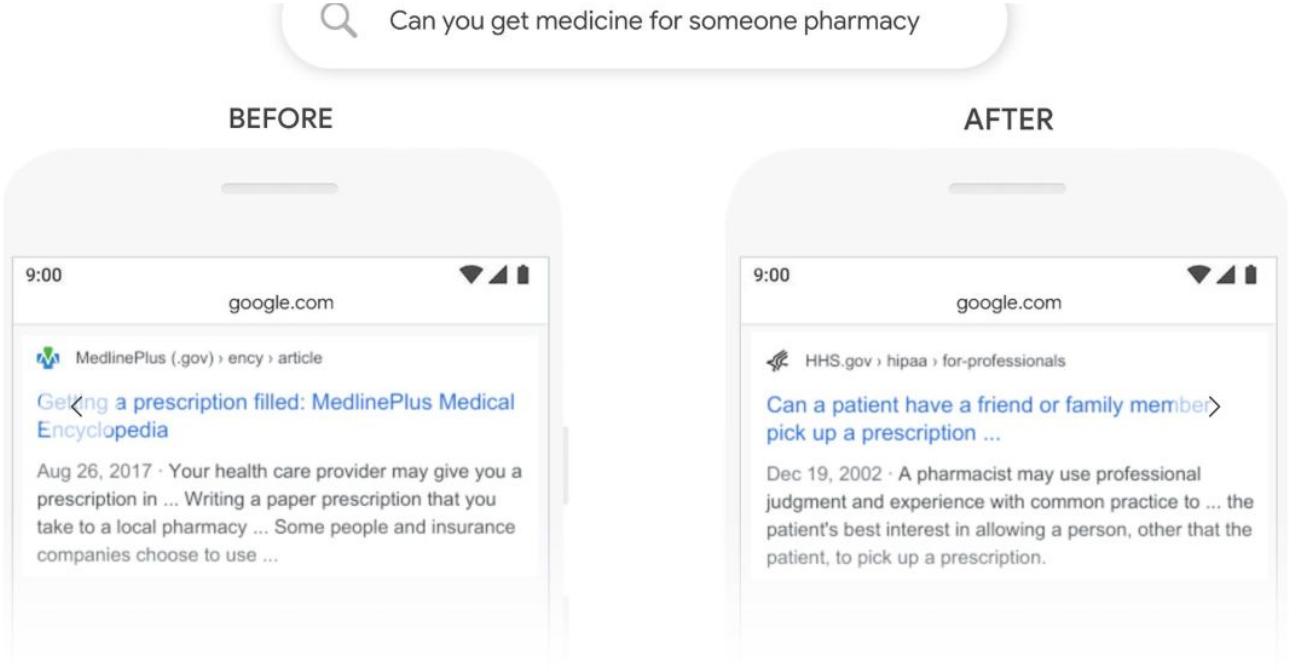
RIKEN

Aleksandr Drozd

Who is BERT?

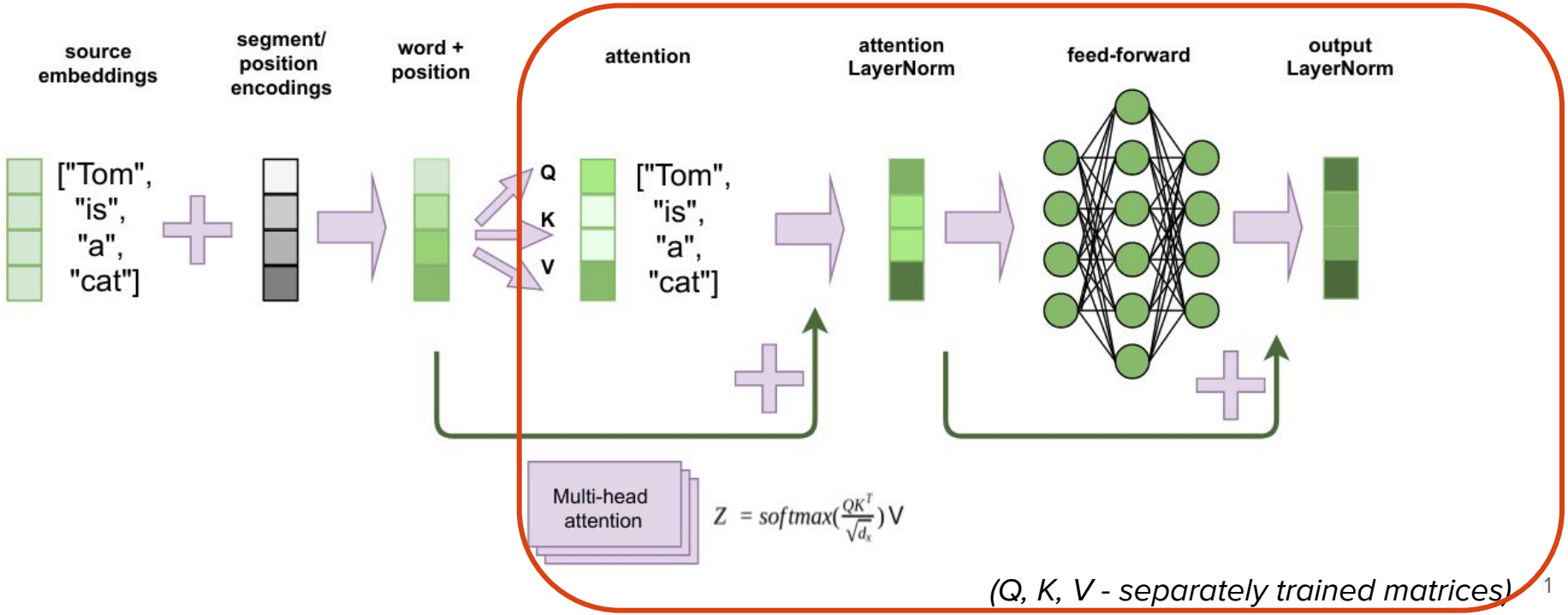
- Devlin et al.: BERT: Pre-training of **Deep Bidirectional Transformers** for Language Understanding
- Best paper at NAACL 2019;
- Developed and open-sourced by Google;
- (still) a must-have baseline
- Nearly **30K** citations on Google Scholar!

BERT improves ~10% of Google search queries!



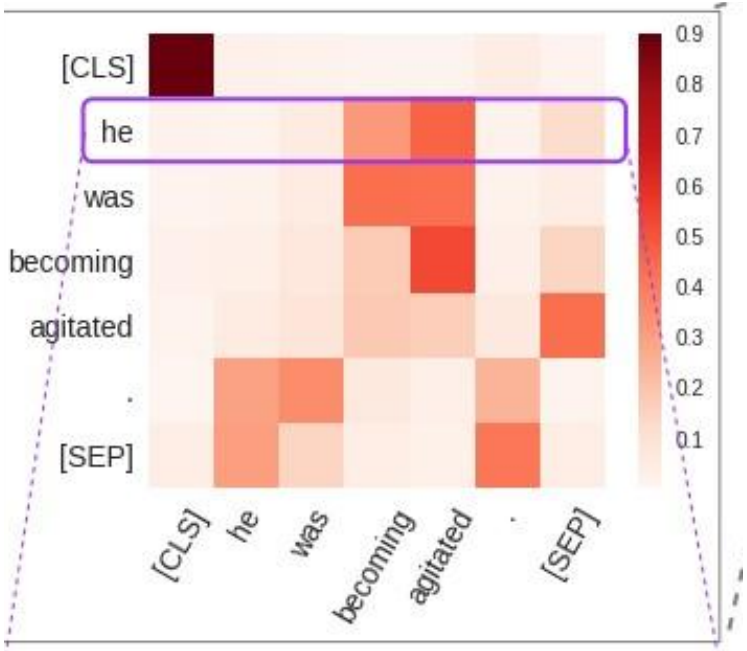
https://techcrunch.com/2019/10/25/google-brings-in-bert-to-improve-its-search-results/

Transformer Encoder



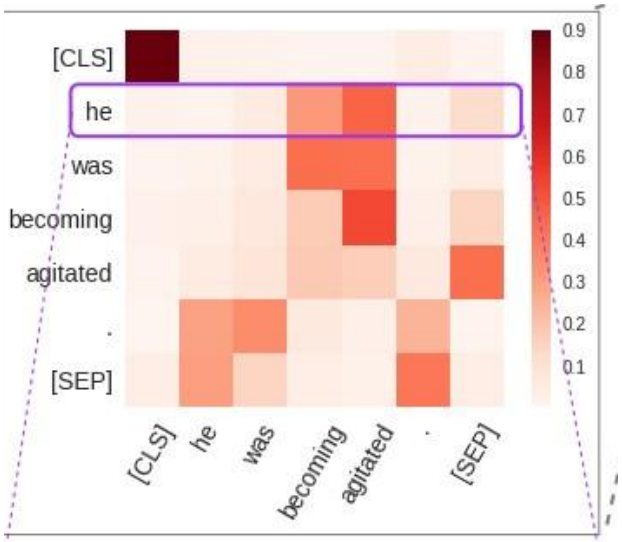
(Q, K, V - separately trained matrices)

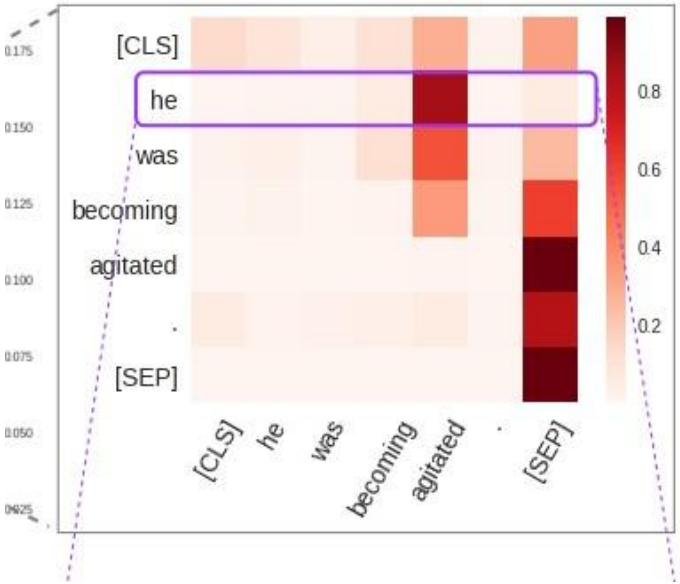
The secret sauce: attention weights encoding relations in sequences



http://mlexplained.com/2017/12/29/attention-is-all-you-need-explained/

Multiple heads could learn different patterns and create a rich representation





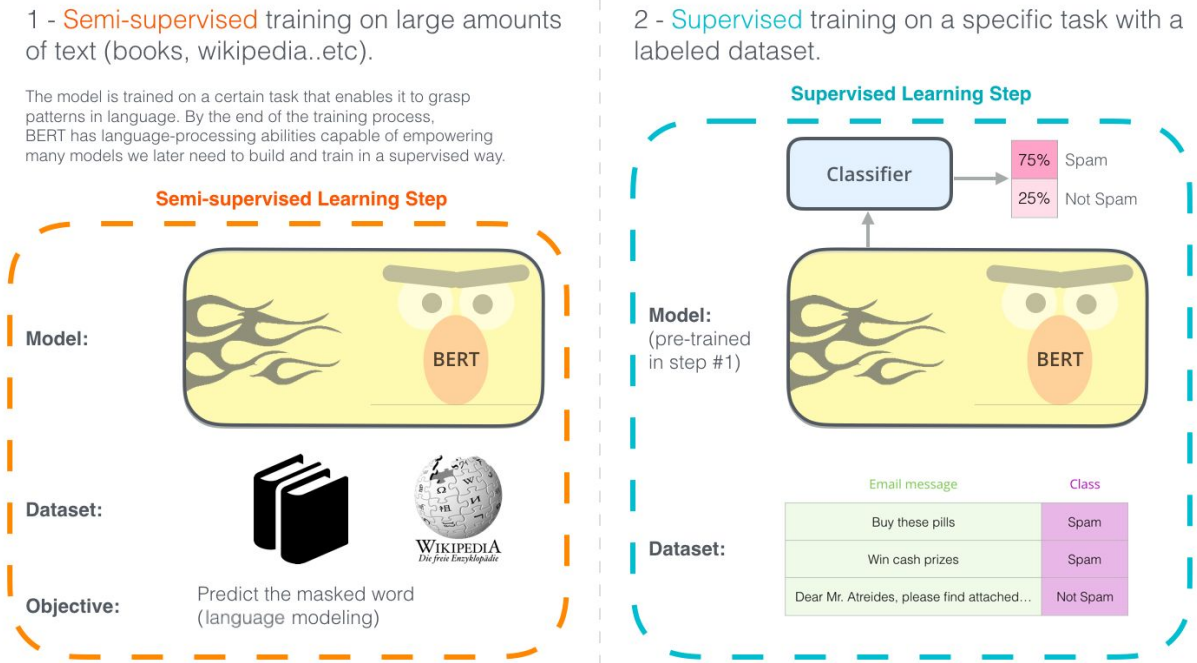
http://mlexplained.com/2017/12/29/attention-is-all-you-need-explained/

BERT embeddings: stacks of Transformer encoders

😎😎😎😎😎😎😎😎😎😎😎😎

- Multiple self-attention heads (12 and 16 in BERT-base and BERT-large, respectively)
- Multiple layers: 12 in BERT-base model, 24 in BERT-large, each with multiple heads
- Total parameters: 110 and 340 M for BERT-base and large;
- Train time: 4 and 16 Cloud TPUs for BERT-base and large (16 and 64 TPU chips total).

Pre-train + fine-tune architecture



https://medium.com/huggingface/introducing-fastbert-a-simple-deep-learning-library-for-bert-models-89ff763ad384

BERT pretraining tasks

Masked language model:

my dog is hairy → my dog is [MASK]

Next sentence prediction:

[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]

Label = IsNext

https://medium.com/huggingface/introducing-fastbert-a-simple-deep-learning-library-for-bert-models-89ff763ad384

Transformer-based architectures achieve superhuman performance on many NLP tasks!

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
+	1	Alibaba DAMO NLP	StructBERT		90.3	75.3	97.1	93.9/91.9	93.0/92.5	74.8/91.0	90.9	90.7	96.4	90.2	94.5	49.1
	2	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
	3	ERNIE Team - Baidu	ERNIE		90.1	72.8	97.5	93.2/91.0	92.9/92.5	75.2/90.8	91.2	90.8	96.1	90.9	94.5	49.4
	4	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
+	5	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
+	6	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
	7	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
	8	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
+	9	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	10	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

https://gluebenchmark.com/leaderboard

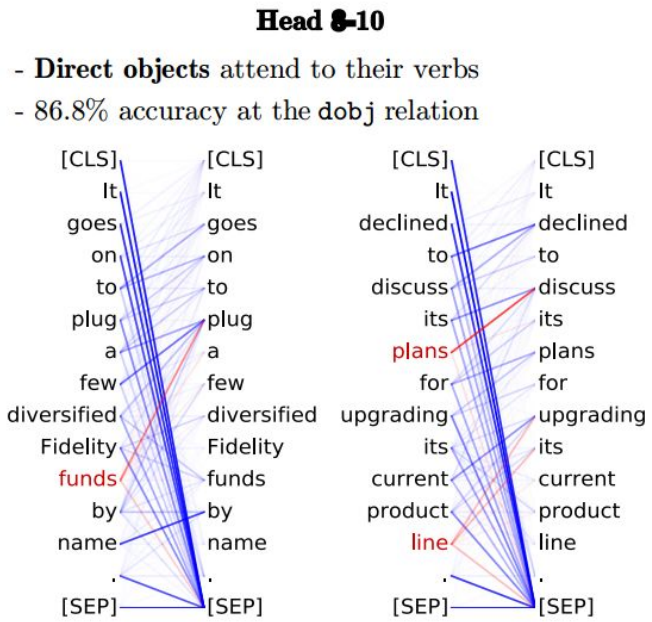
GLUE tasks: different aspects of natur language understanding

- **Text similarity and paraphrase:** MRPC, STS-B*, QQP
  how similar are 2 sentences?
- **Sentiment analysis:** SST-2
  positive of neg sentiment?
- **Entailment and inference**: RTE, QNLI, MNLI, WNLI
  does A entail B?
- **Linguistic acceptability judgements**: CoLA*
  how well-formed is this sentence?

* regression tasks

Q1: how important are attention heads?

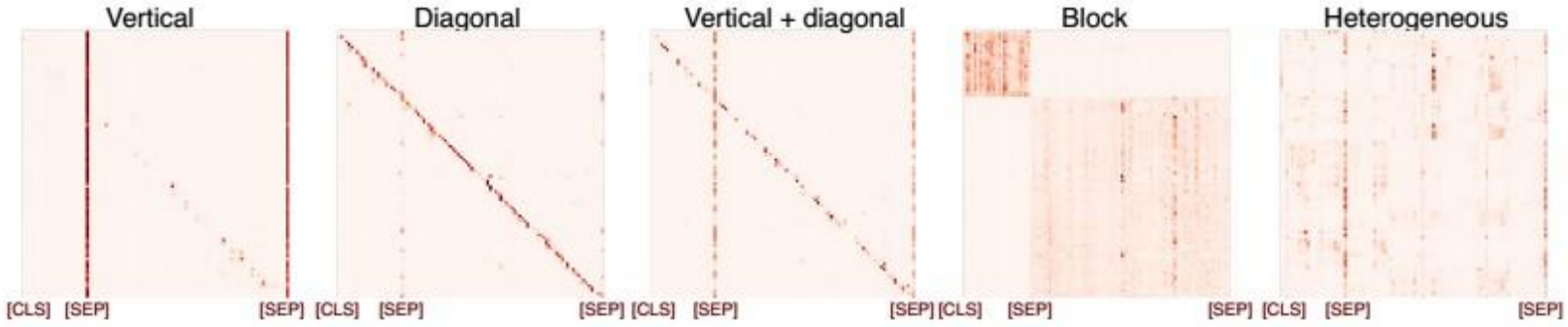
Self-attention sounds like a good mechanism to encode syntactic relations



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)

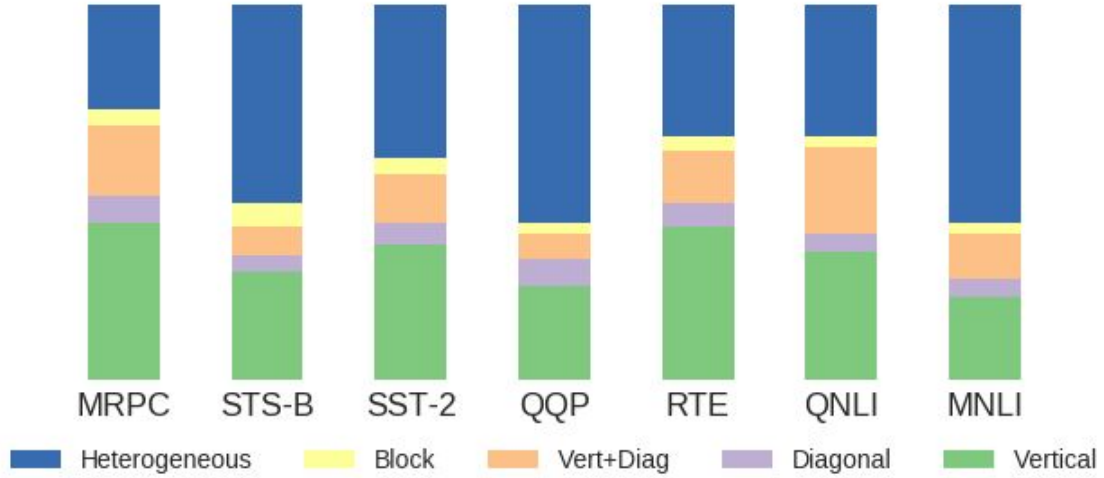
Clark et al. (2019) What Does BERT Look at? An Analysis of BERT's Attention

Self-attention pattern types



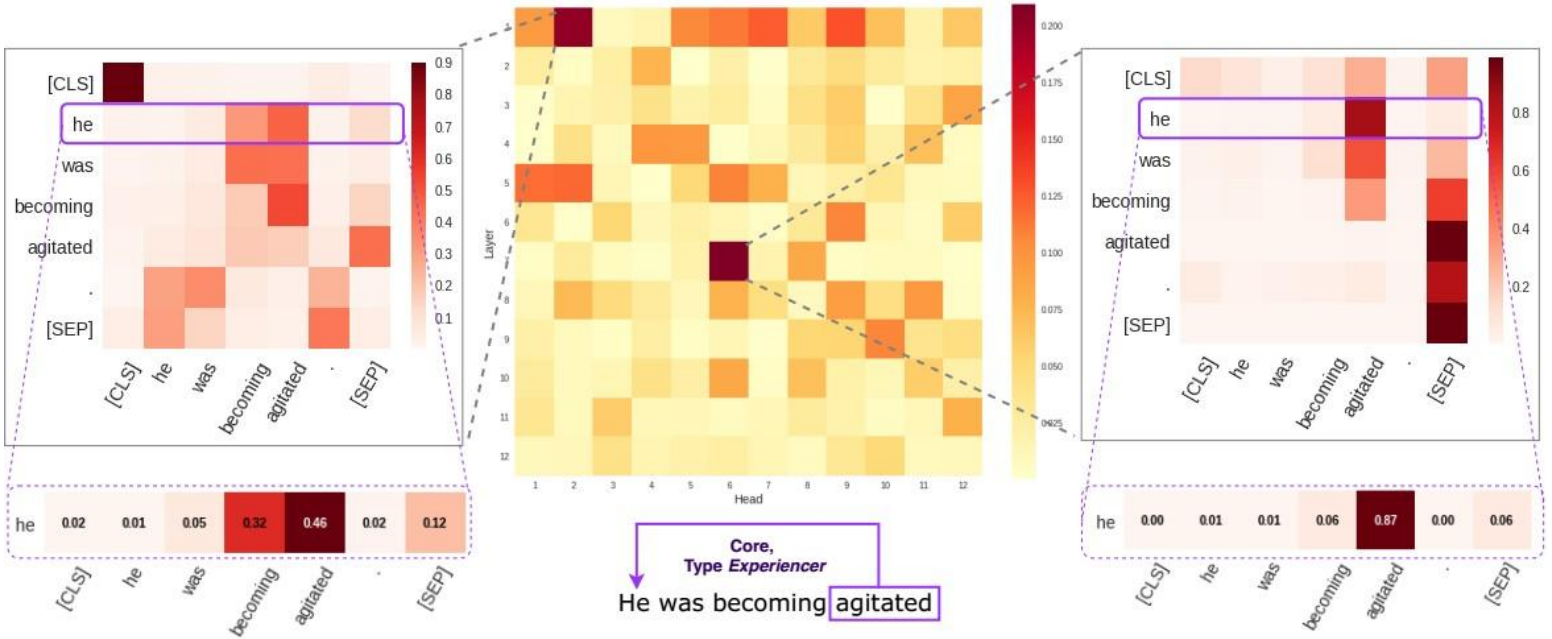
Kovaleva et al. (2019) Revealing the Dark Secrets of BERT

Most self-attention heads are NOT linguistically informative!



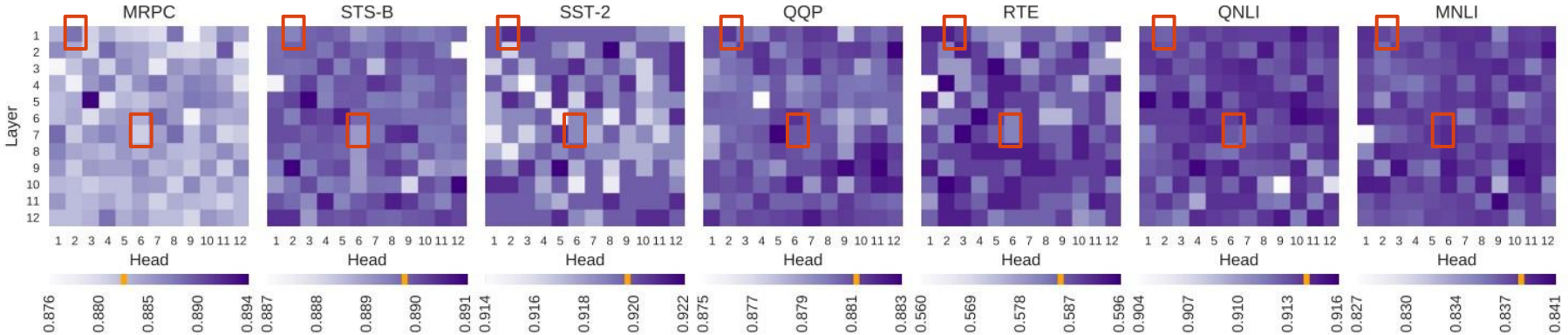
Kovaleva et al. (2019) Revealing the Dark Secrets of BERT

Some heads seem to encode specific relations…



Kovaleva et al. (2019) Revealing the Dark Secrets of BERT

...but even 'important' heads can be zeroed out!



Kovaleva et al. (2019) Revealing the Dark Secrets of BERT

See also:
Michel et al. (2019) Are sixteen heads really better than one?
Voita et al. (2019) Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned

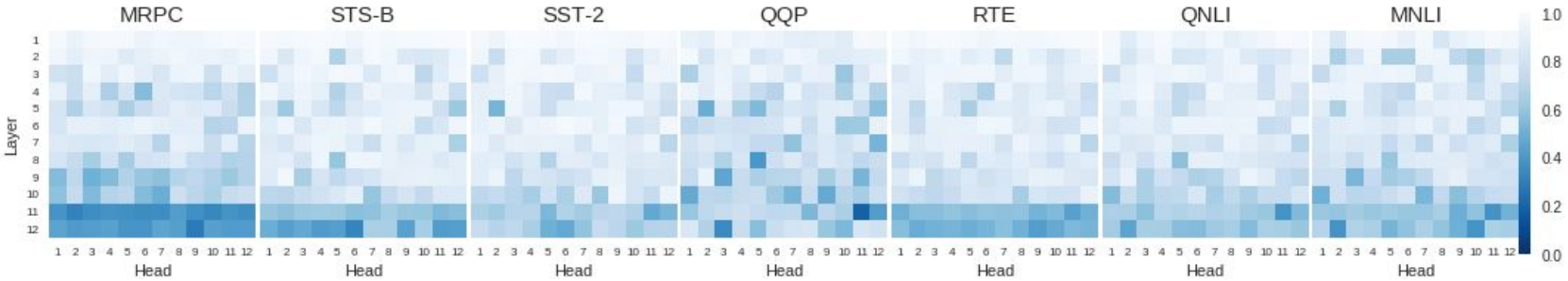
Q2: what do pre-training and fine-tuning do?

BERT works pretty well even without pre-training!

Dataset	Pre-trained	Fine-tuned, initialized with normal distr.	Fine-tuned, initialized with pre-trained	Metric	Size
MRPC	0/31.6	81.2/68.3	87.9/82.3	F1/Acc	5.8K
STS-B	33.1	2.9	82.7	Acc	8.6K
SST-2	49.1	80.5	92	Acc	70K
QQP	0/60.9	0/63.2	65.2/78.6	F1/Acc	400K
RTE	52.7	52.7	64.6	Acc	2.7K
QNLI	52.8	49.5	84.4	Acc	130K
MNLI-m	31.7	61.0	78.6	Acc	440K

Kovaleva et al. (2019) Revealing the Dark Secrets of BERT

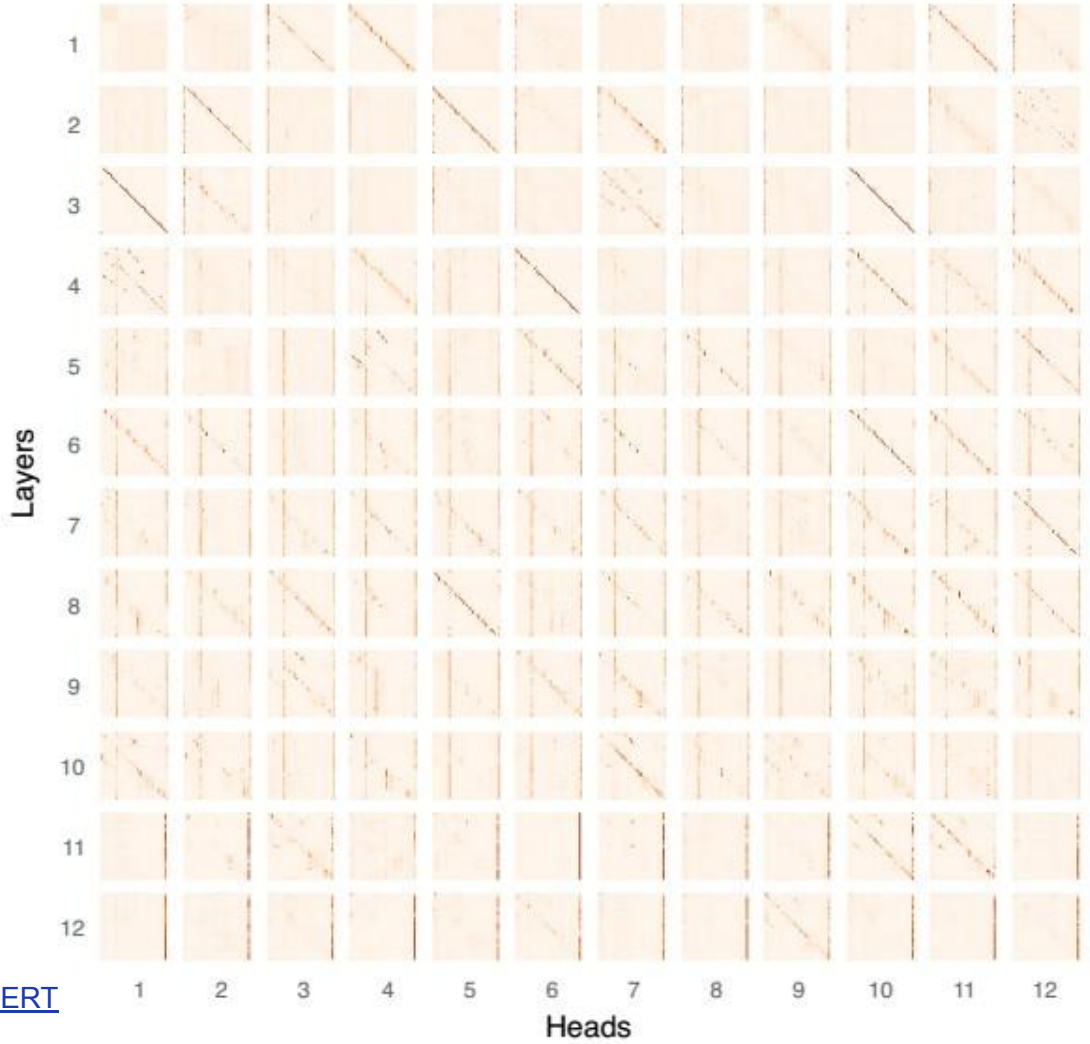
Final layers are the most task-specific and the most affected by fine-tuning



Kovaleva et al. (2019) Revealing the Dark Secrets of BERT

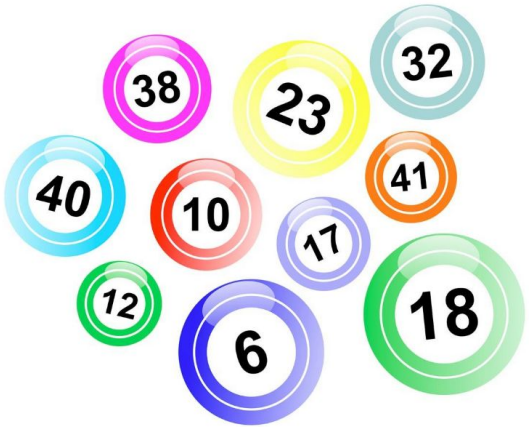
Fine-tuning does NOT necessarily 'teach' informative attention patterns

(BERT fine-tuned on QNLI)



Kovaleva et al. (2019) Revealing the Dark Secrets of BERT

Q3: Does lottery ticket hypothesis hold for pre-trained BERT?



What about pre-trained Transformers?

~~**Dense, randomly-initialized, feed-forward networks**~~ contain subnetworks (winning tickets) that -- when trained in isolation -- reach test accuracy comparable to the original network in a similar number of iterations

Prasanna et al. (2020) When BERT plays the lottery, all tickets are winning

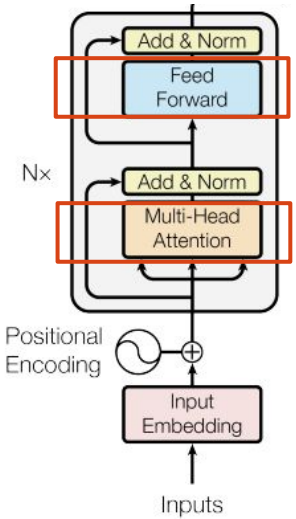
Method 1: unstructured magnitude pruning (m-pruning)

1. Iteratively prune 10% of the least magnitude weights across the entire fine-tuned model (except the embeddings) and evaluate on dev set.
2. Repeat (1) so long as performance of the pruned subnetwork is above 90% of the full model.

Method 2: structured pruning heads and MLPs based on importance scores (s-pruning)

$$I_h^{(h,l)} = E_{x \sim X} \left| \frac{\partial \mathcal{L}(x)}{\partial \xi^{(h,l)}} \right|$$

$$I_{mlp}^{(l)} = E_{x \sim X} \left| \frac{\partial \mathcal{L}(x)}{\partial \nu^{(l)}} \right|$$

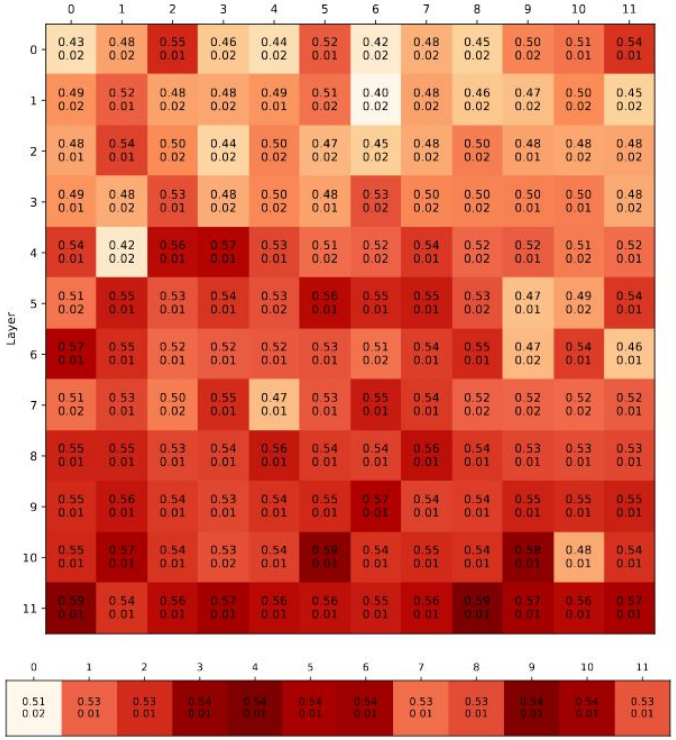


Iterative pruning till 90% performance of full model:

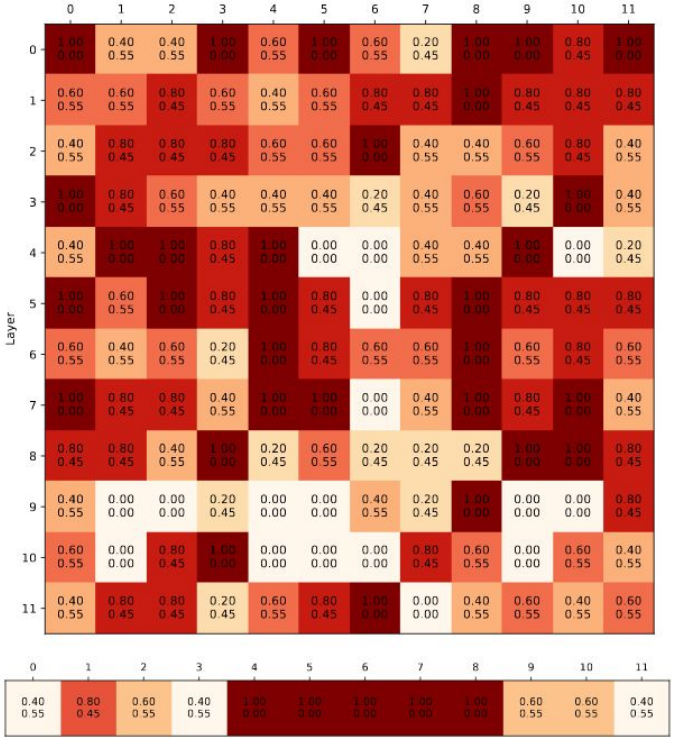
10% heads + 1 MLP per iteration

Michel et al. (2019) Are sixteen heads really better than one?

"Good" subnetwork example: QNLI



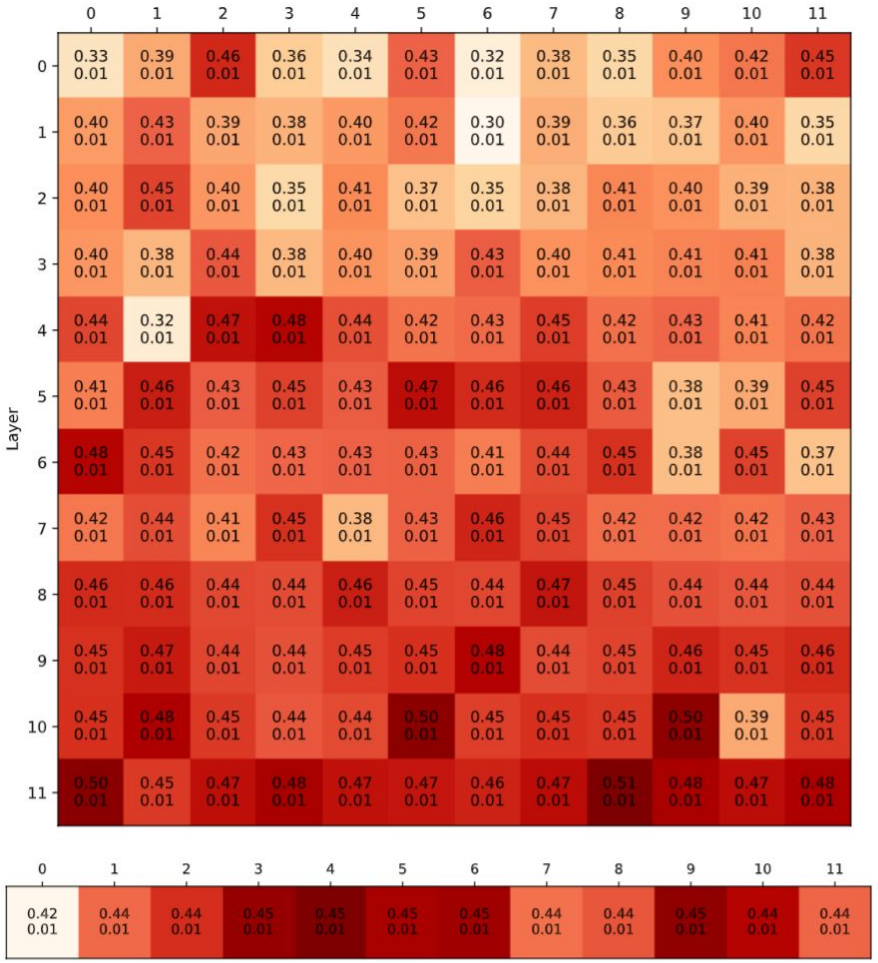
(a) M-pruning: each cell gives the percentage of surviving weights and it's std across 5 random seeds.



(b) S-pruning: each cell gives the mean and std of the binary outcome of survival of a given head/MLP across 5 random seeds.

M-pruning:

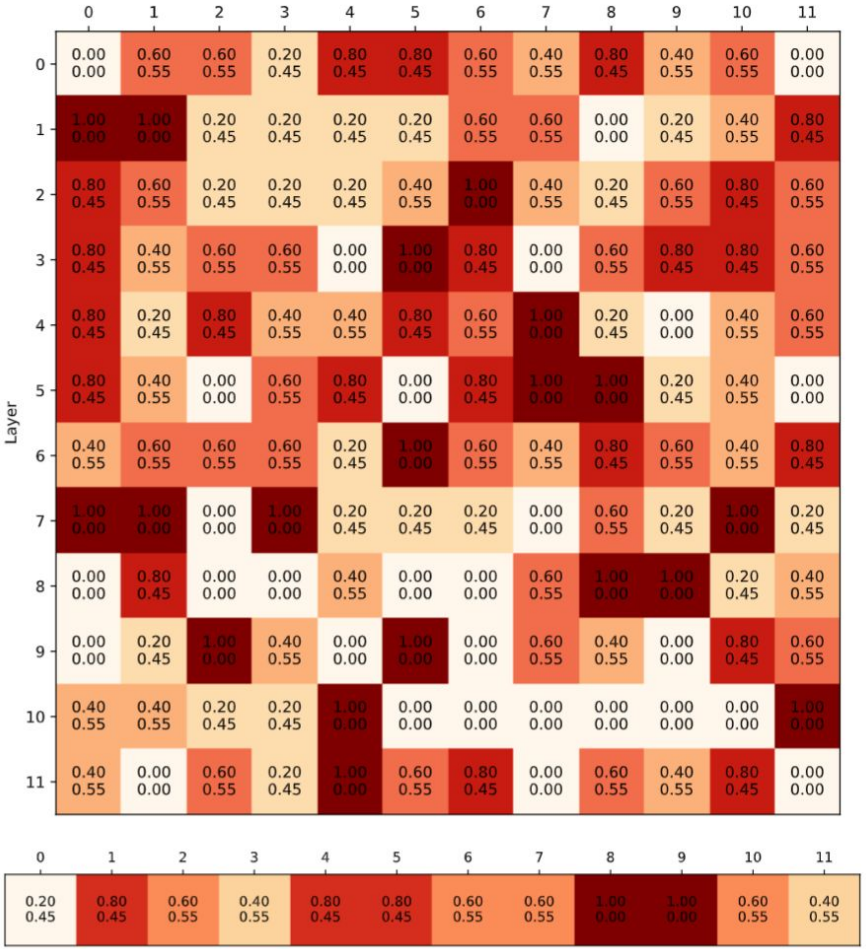
- **Quite stable!**
- **0-0.01 std across random seeds**



SST-2

S-pruning:

- very unstable
- 0-0.55 std across 5 random seeds
- Fleiss kappa for head/MLP survival masks across random seeds: 0.1~0.3



SST-2

Why so unstable?

Most heads are about equally (un)important!

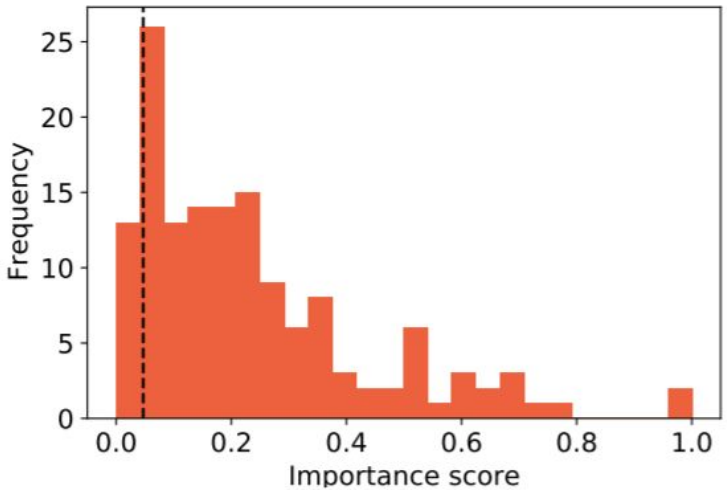


Figure 3: Head importance scores distribution (this example shows CoLA, pruning iteration 1)

S-pruned subnetworks are not stable across related tasks!

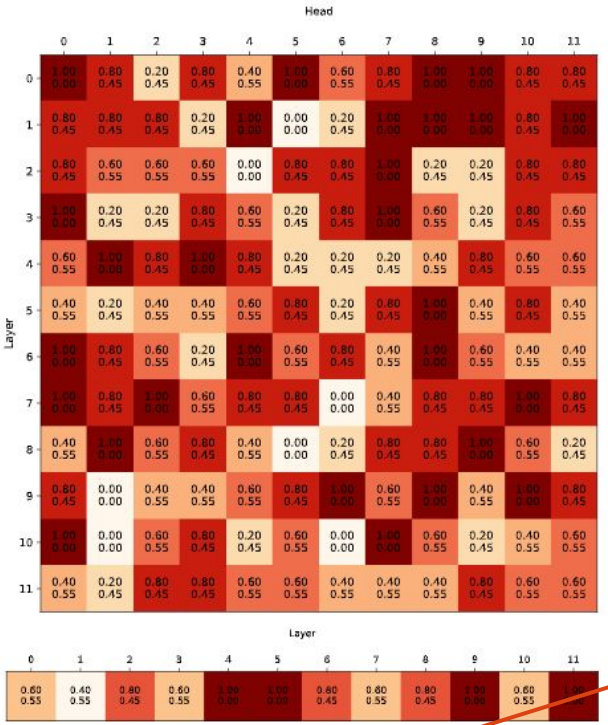


Figure 6: MRPC

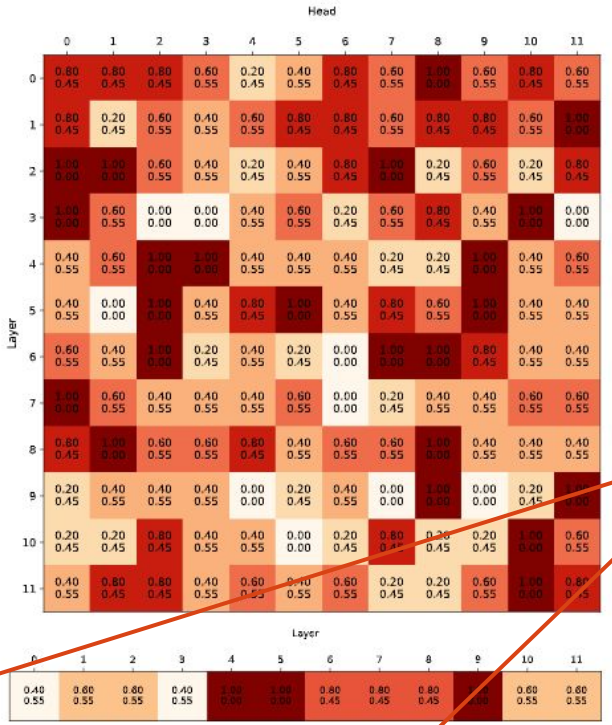


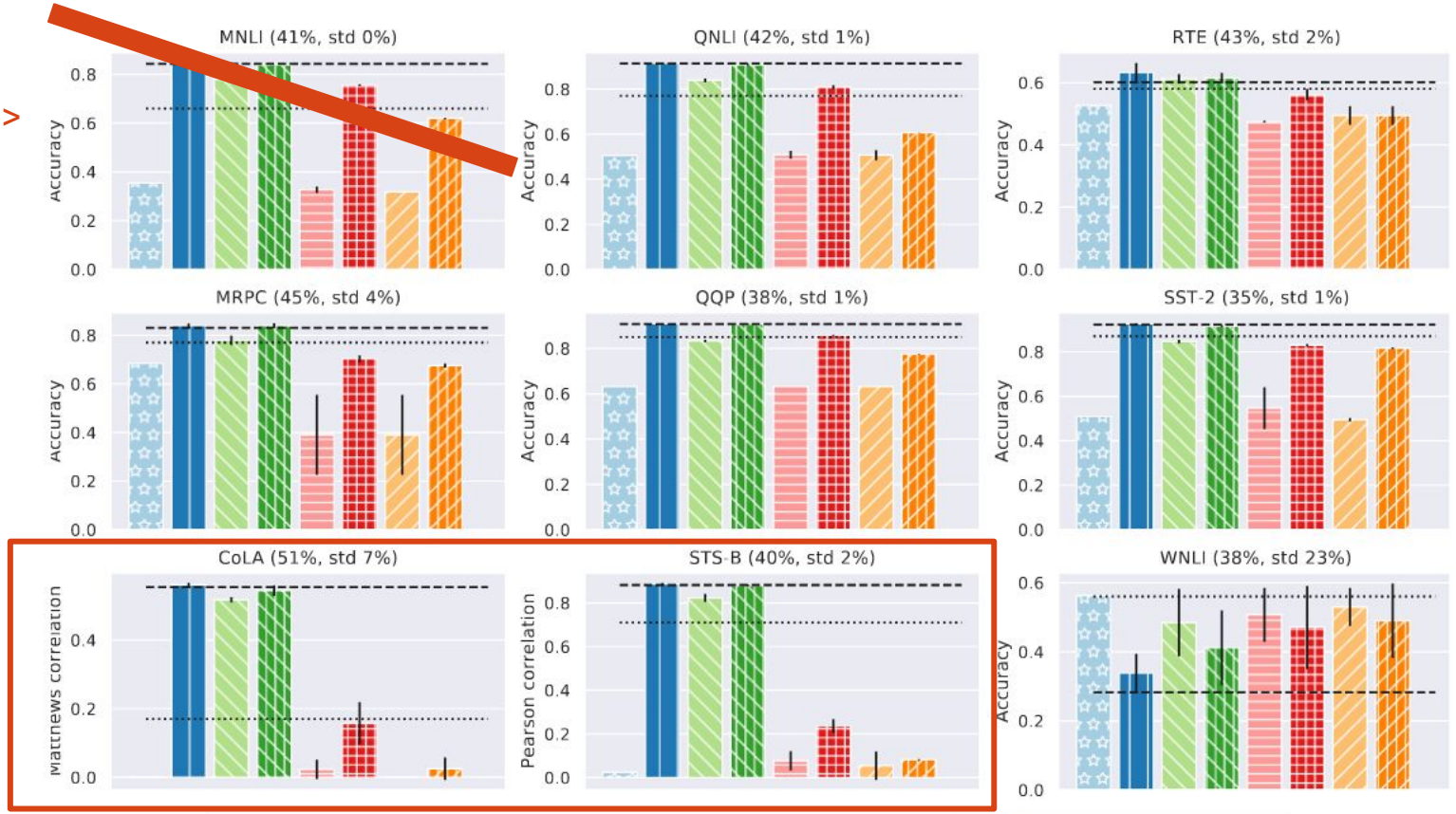
Figure 8: QQP

Both tasks target the ability to detect paraphrases!

The Good, the Bad and the ~~Ugly~~ Random

- **The Good** - elements selected from the full model by either pruning strategy.
- **The Bad** - elements that did not survive the pruning and few sampled from the remaining to match the good subnetwork size.
- **The Random -** elements randomly sampled from the full model to match the good subnetwork size.

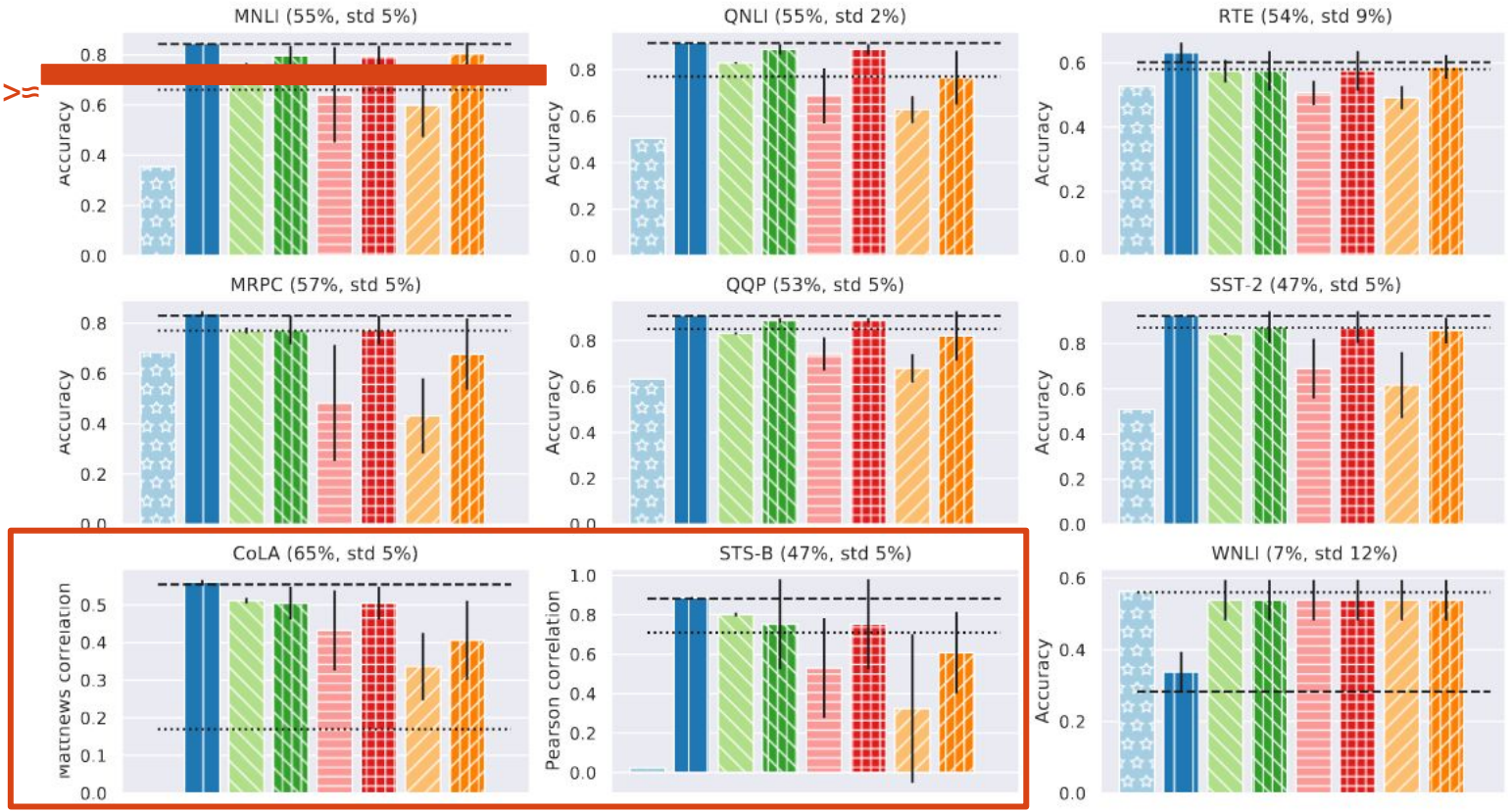
Good > random > bad



Regression tasks suffer dramatically

Magnitude pruning

Good ≈ random >≈ bad



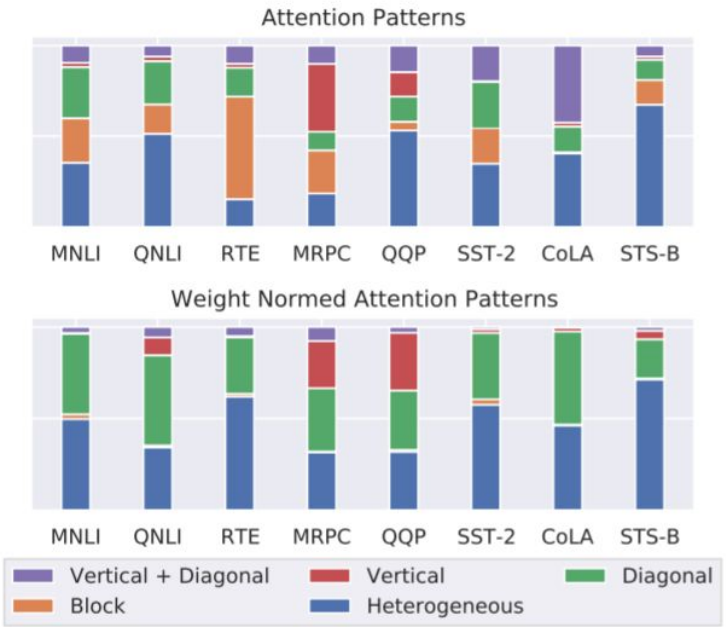
Structured pruning

Regression tasks doing much better than with m-pruning

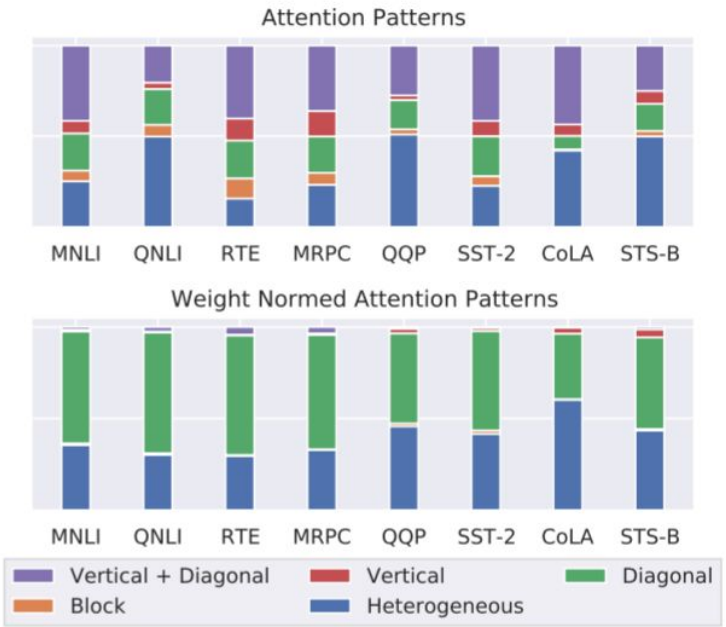
The "bad" s-pruned subnets are not too bad!

Model	CoLA	SST-2	MRPC	QQP	STS-B	MNLI	QNLI	RTE	WNLI	Average
Majority class baseline	0.00	0.51	0.68	0.63	0.02	0.35	0.51	0.53	**0.56**	0.42
CBOW	**0.46**	0.79	0.75	0.75	0.70	0.57	0.62	0.71	**0.56**	0.61
BILSTM + GloVe	0.17	0.87	**0.77**	0.85	**0.71**	0.66	**0.77**	**0.58**	**0.56**	0.66
BILSTM + ELMO	0.44	**0.91**	0.70	**0.88**	0.70	**0.68**	0.71	0.53	0.56	**0.68**
'Bad' subnetwork (s-pruning)	0.40	0.85	0.67	0.81	0.60	0.80	0.76	0.58	0.53	0.67
'Bad' subnetwork (m-pruning)	0.24	0.81	0.67	0.77	0.08	0.61	0.6	0.49	0.49	0.51
Random init + random s-pruning	0.00	0.78	0.67	0.78	0.14	0.63	0.59	0.53	0.50	0.52

Heads surviving importance-based pruning are NOT necessarily linguistically informative!



(b) Super-survivor heads, fine-tuned



(c) All heads, fine-tuned

[Prasanna et al. (2020) When BERT Plays the Lottery, All Tickets Are Winning]

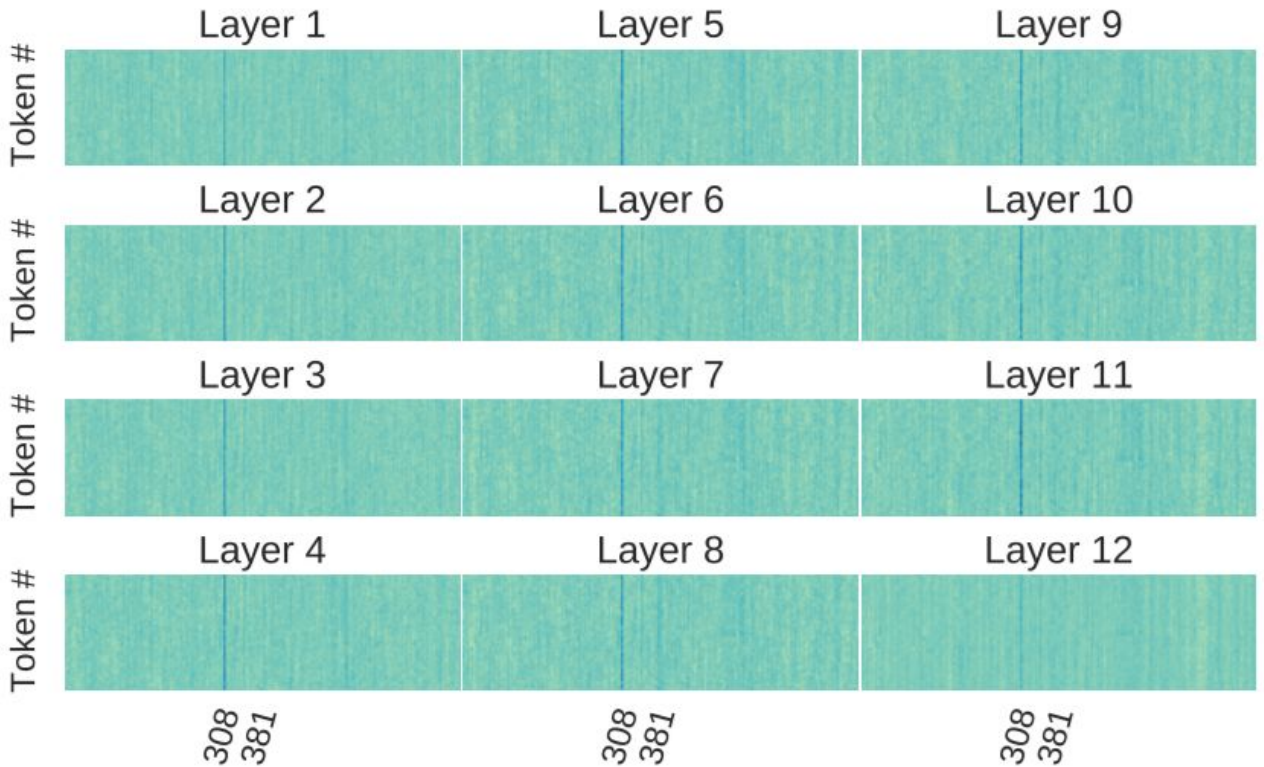
Q4:
so... is BERT invincible?

6 versions of BERT + XLNet, BART, ELECTRA, & GPT-2

You can 'kill' ~~BERT~~ by disabling <0.0001% of model weights!

high-magnitude weights **in the same position** in the **final operation** across Transformer layers!

- BERT & co: scaling factors and biases in the output LayerNorm
- GPT-2: output dense layer



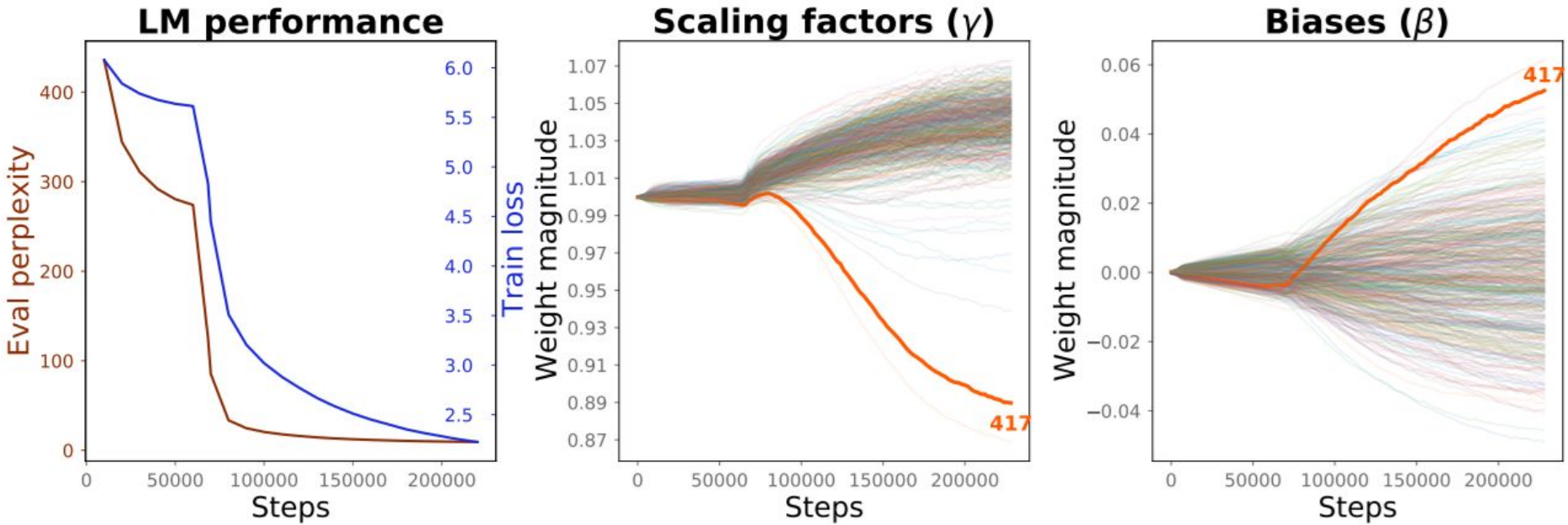
Kovaleva et al. (2021) BERT Busters: Outlier Dimensions that Disrupt Transformers

RoBERTa masked LM with disabled outlier features:

Original paragraph	Ghostbusters was [released] on June 8 , [1984] , to critical [acclaim] and became a cultural phenomenon . It was well [received] for its deft blend of comedy, [action] , and horror , and Murray ' s performance was [repeatedly] singled out for praise .

Kovaleva et al. (2021) BERT Busters: Outlier Dimensions that Disrupt Transformers

Outliers emerge early in pre-training: (BERT-medium)

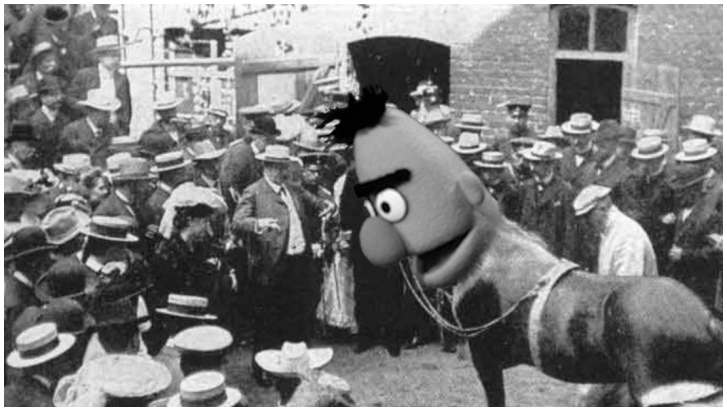


Kovaleva et al. (2021) BERT Busters: Outlier Dimensions that Disrupt Transformers

Q5:
what exactly do we teach BERT?

Our datasets are too easy!

- BERT learns shallow heuristics in NLI (McCoy et al. 2019, Zellers et al 2019, Jin et al. 2020)
- BERT learns shallow heuristics in reading comprehension (Si et al 2019, Rogers et al. 2020, Sugawara et al 2020)
- probably also everywhere else



QuAIL reading comprehension: paraphrase challenge set

T: John bought a used car on Tuesday morning.

Easy question:

What kind of car did John buy?

Paraphrased question:

What kind of vehicle did the man purchase?

Rogers et al. (2020) Getting Closer to AI-complete Question Answering: A Set of Prerequisite Real Tasks

QuAIL reading comprehension: paraphrase challenge set

Qtype	TriAN	BERT
Temporal order	0.51 (0.06)	0.24 (-0.25)
Coreference	0.42 (-0.11)	0.31 (-0.13)
Factual	0.32 (-0.21)	0.4 (-0.12)
Causality	0.33 (-0.2)	0.3 (-0.27)
Subsequent state	0.23 (-0.12)	0.3 (-0.18)
Event duration	0.62 (0.0)	0.48 (-0.13)
Entity properties	0.31 (-0.06)	0.42 (0.02)
Belief states	0.26 (-0.27)	0.31 (-0.39)
Unanswerable	0.55 (-0.1)	0.48 (-0.13)
All questions	0.4 (-0.11)	0.36 (-0.17)

Rogers et al. (2020) Getting Closer to AI-complete Question Answering: A Set of Prerequisite Real Tasks

Shallow heuristics in NLI task: lexical overlap

Biased training data (MNLI)

Premise: The banker near the judge saw the actor.

Hypothesis: The banker saw the actor.

Adversarial test data (HANS)

Premise: The judge by the actor stopped the banker.

Hypothesis: The banker stopped the actor.

Random baseline for 2-class classification: 50

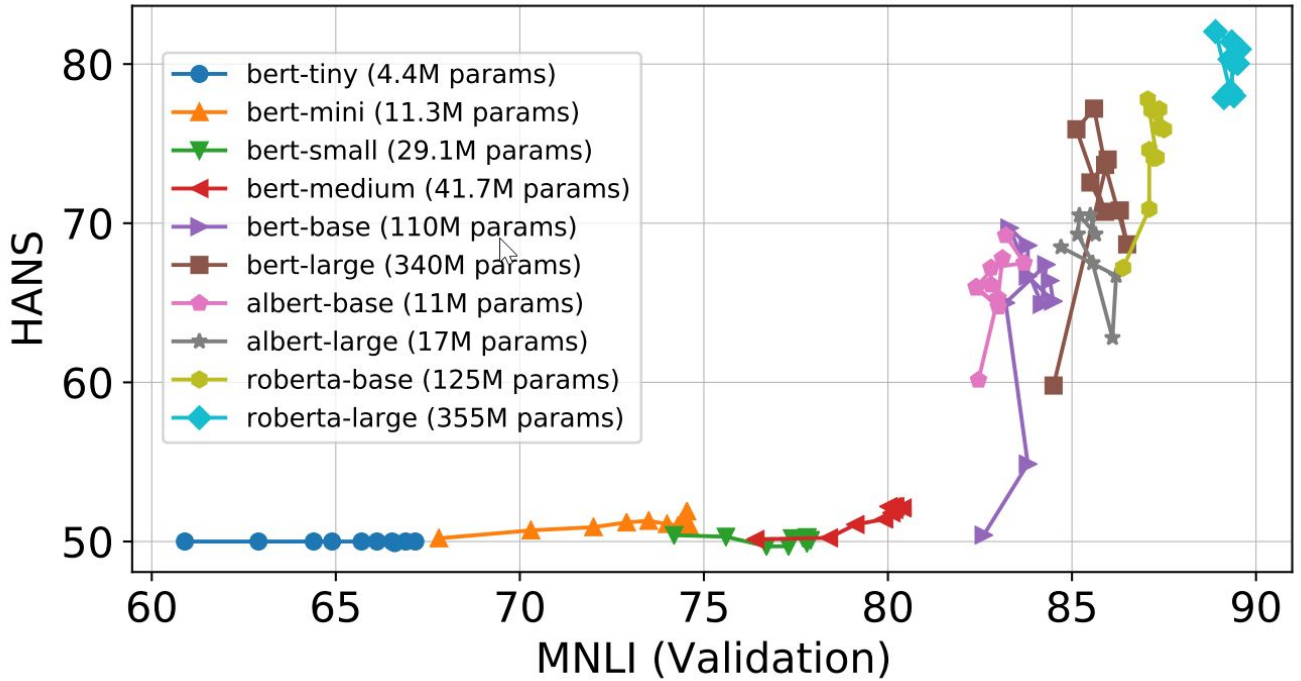
McCoy et al. (2019) Right for the wrong reasons.

We tried to help BERT-base to generalize better…
but none of the modeling approaches worked!

Architecture
Vanilla BERT
Siamese BERT (frozen encoder)
Siamese BERT (trainable encoder)
BERT + adapters
BERT + explicit debiasing with HEX projection

Bhargava et al. (2021) Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics

… but larger models generalize better!



Bhargava et al. (2021) Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics

(Some) follow-up questions:

- **Can we learn to generalize with smaller models?**
- **Can we pre-train Transformers more efficiently, given the outlier phenomenon and attention head redundancy?**
- **Can we get the Transformers to learn more high-level and/or interpretable linguistic/logical patterns for NLU tasks?**
- **Can we tell when BERT is right for the right reasons?**
- **… ?**

Let's discuss!

Slides (with links) available at
https://annargrs.github.io/talks

Anna Rogers
https://annargrs.github.io
@annargrs