Intro to BERT-ology

Dr. Rachael Tatman

- You can get equally good results with smaller models
- BERT is not a cognitive model
- We only know some of the security risks posed by BERT based models

But first! What is BERT?

Photo by See-ming Lee, shared under CC BY-SA 2.0

A specific, large transformer masked language model.

A specific, large transformer masked **language model**.

A language model is a statistical model of the probability of a sentence or phrase.

A specific, large transformer masked **language model**.

P(Rasa is open source) > P(Source is Rasa open)

RASA

@rctatman

A specific, large transformer

masked language model.

A language model trained by removing words and having the model fill in the _____.

One of the big contribution of BERT was proposing this way of training language models. ⬇️

A specific, large transformer

masked language model.

A language model trained by removing words and having the model fill in the _____.

A specific, large transformer

masked language model.

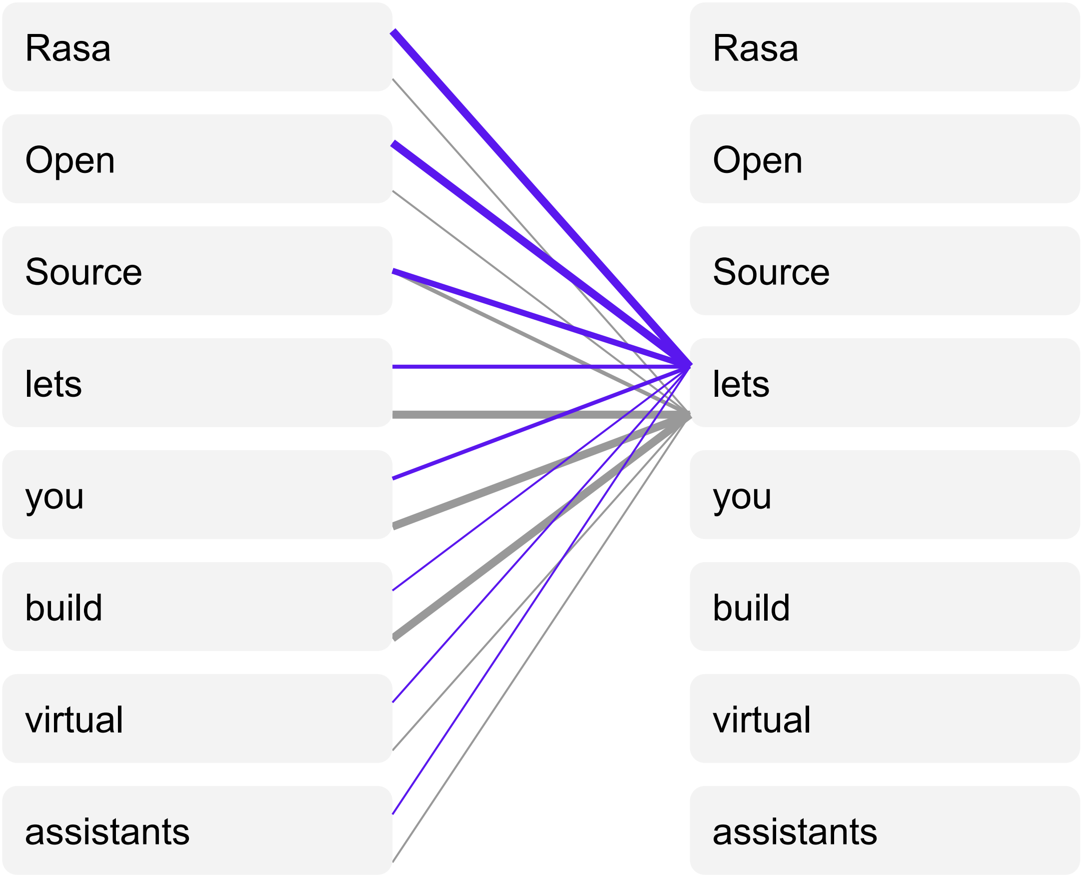
Masked language models are one kind of *contextual word embedding* and can be used as input embeddings.

A specific, large **transformer** masked language model.

Transformers are a fairly new family of neural network architectures.

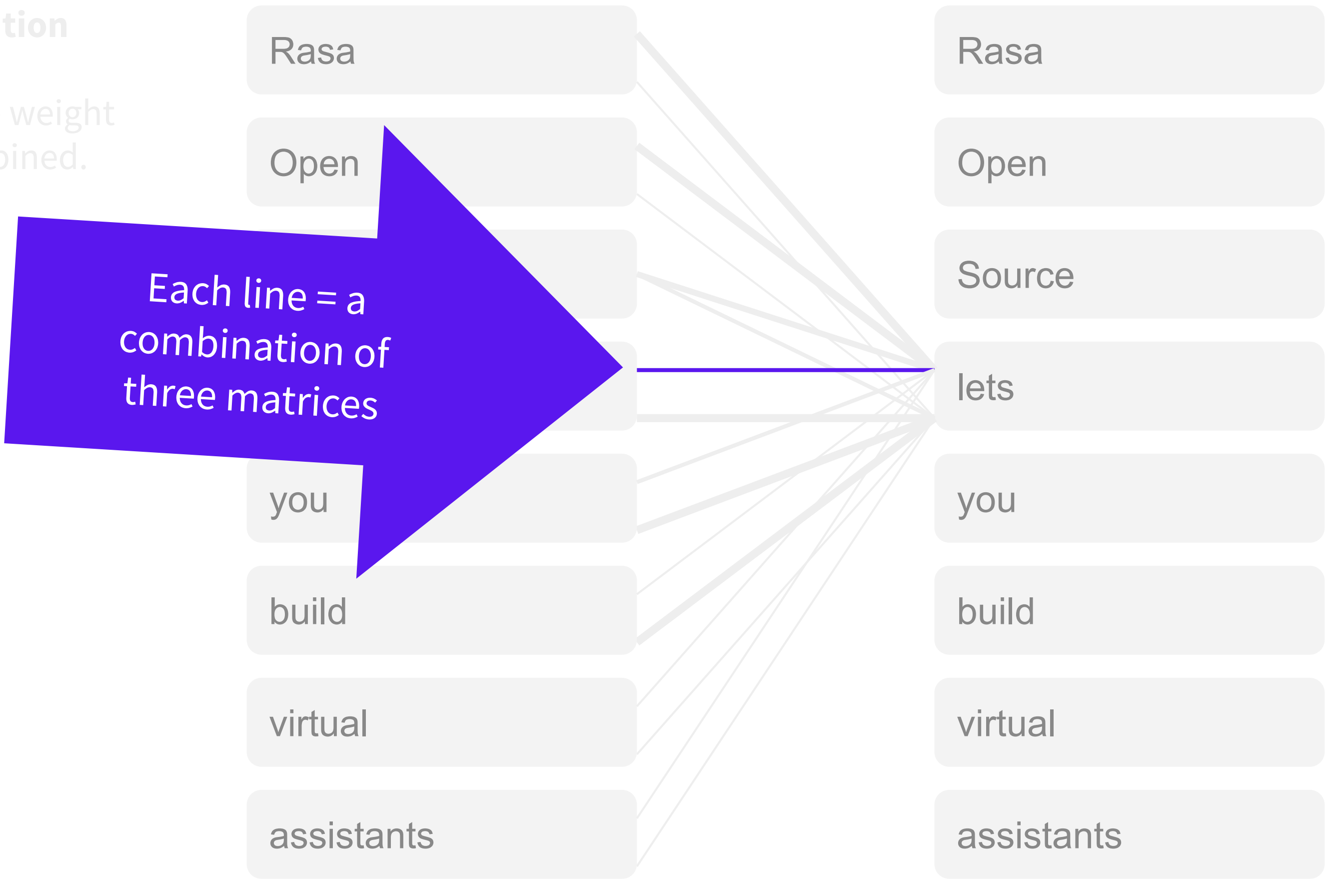
Multi-headed self attention

You learn multiple ways to weight the relationship of each item in the input sequence to all other items in the input



Multi-headed self attention

Each head is made three weight matrices which are combined.



A specific, **large** transformer masked language model.

BERT is extremely large: the large version has 340 million trainable parameters. (An earlier related model, ELMO, had only 93 million.)

- You can get equally good results with smaller models
- BERT is not a cognitive model
- We only know some of the security risks posed by BERT based models

- **You can get equally good results with smaller models**
- BERT is not a cognitive model
- We only know some of the security risks posed by BERT based models

		Compression	Performance	Speedup	Model	Evaluation
Distillation	DistilBERT (Sanh et al., 2019)	×2.5	90%	×1.6	$BERT_6$	All GLUE tasks
	$BERT_6$-PKD (Sun et al., 2019a)	×1.6	97%	×1.9	$BERT_6$	No WNLI, CoLA and STS-B
	$BERT_3$-PKD (Sun et al., 2019a)	×2.4	92%	×3.7	$BERT_3$	No WNLI, CoLA and STS-B
	(Aguilar et al., 2019)	×2	94%	-	$BERT_6$	CoLA, MRPC, QQP, RTE
	BERT-48 (Zhao et al., 2019)	×62	87%	×77	$BERT_{12}{}^{*\dagger}$	MNLI, MRPC, SST-2
	BERT-192 (Zhao et al., 2019)	×5.7	94%	×22	$BERT_{12}{}^{*\dagger}$	MNLI, MRPC, SST-2
	TinyBERT (Jiao et al., 2019)	×7.5	96%	×9.4	$BERT_4{}^{*\dagger}$	All GLUE tasks
	MobileBERT (Sun et al.)	×4.3	100%	×4	$BERT_{24}{}^{\dagger}$	No WNLI
	PD (Turc et al., 2019)	×1.6	98%	$\times 2.5^3$	$BERT_6{}^{\dagger}$	No WNLI, CoLA and STS-B
	MiniBERT(Tsai et al., 2019)	$\times 6^{\S}$	98%	$\times 27^{\S}$	$mBERT_3{}^{\dagger}$	CoNLL-2018 POS and morphology
	BiLSTM soft (Tang et al., 2019)	×110	91%	$\times 434^{\ddagger}$	$BiLSTM_1$	MNLI, QQP, SST-2
Quant.	Q-BERT (Shen et al., 2019)	×13	99%	-	$BERT_{12}$	MNLI, SST-2
	Q8BERT (Zafrir et al., 2019)	×4	99%	-	$BERT_{12}$	All GLUE tasks
Other	ALBERT-base (Lan et al., 2019)	×9	97%	×5.6	$BERT_{12}{}^{**}$	MNLI, SST-2
	ALBERT-xxlarge (Lan et al., 2019)	×0.47	107%	×0.3	$BERT_{12}{}^{**}$	MNLI, SST-2
	BERT-of-Theseus (Xu et al., 2020)	×1.6	98%	-	$BERT_6$	No WNLI

"A Primer in BERTology: What we know about how BERT works"
Anna Rogers, Olga Kovaleva, Anna Rumshisky (Preprint 2019)

Train a small model to mimic the behavior or weights of a larger one.

		Compression	Performance	Speedup	Model	Evaluation
Distillation	DistilBERT (Sanh et al., 2019)	$\times$2.5	90%	$\times$1.6	BERT_6	All GLUE tasks
	BERT_6-PKD (Sun et al., 2019a)	$\times$1.6	97%	$\times$1.9	BERT_6	No WNLI, CoLA and STS-B
	BERT_3-PKD (Sun et al., 2019a)	$\times$2.4	92%	$\times$3.7	BERT_3	No WNLI, CoLA and STS-B
	(Aguilar et al., 2019)	$\times$2	94%	-	BERT_6	CoLA, MRPC, QQP, RTE
	BERT-48 (Zhao et al., 2019)	$\times$62	87%	$\times$77	$\text{BERT}_{12}^{*\dagger}$	MNLI, MRPC, SST-2
	BERT-192 (Zhao et al., 2019)	$\times$5.7	94%	$\times$22	$\text{BERT}_{12}^{*\dagger}$	MNLI, MRPC, SST-2
	TinyBERT (Jiao et al., 2019)	$\times$7.5	96%	$\times$9.4	$\text{BERT}_4^{*\dagger}$	All GLUE tasks
	MobileBERT (Sun et al.)	$\times$4.3	100%	$\times$4	$\text{BERT}_{24}^{\dagger}$	No WNLI
	PD (Turc et al., 2019)	$\times$1.6	98%	$\times 2.5^{3}$	$\text{BERT}_6^{\dagger}$	No WNLI, CoLA and STS-B
	MiniBERT(Tsai et al., 2019)	$\times 6^{\S}$	98%	$\times 27^{\S}$	$\text{mBERT}_3^{\dagger}$	CoNLL-2018 POS and morphology
	BiLSTM soft (Tang et al., 2019)	$\times$110	91%	$\times 434^{\ddagger}$	BiLSTM_1	MNLI, QQP, SST-2
Quant	Q-BERT (Shen et al., 2019)	$\times$13	99%	-	BERT_{12}	MNLI, SST-2
	Q8BERT (Zafrir et al., 2019)	$\times$4	99%	-	BERT_{12}	All GLUE tasks
Other	ALBERT-base (Lan et al., 2019)	$\times$9	97%	$\times$5.6	BERT_{12}^{**}	MNLI, SST-2
	ALBERT-xxlarge (Lan et al., 2019)	$\times$0.47	107%	$\times$0.3	BERT_{12}^{**}	MNLI, SST-2
	BERT-of-Theseus (Xu et al., 2020)	$\times$1.6	98%	-	BERT_6	No WNLI

"A Primer in BERTology: What we know about how BERT works"
Anna Rogers, Olga Kovaleva, Anna Rumshisky (Preprint 2019)

"A Primer in BERTology: What we know about how BERT works"
Anna Rogers, Olga Kovaleva, Anna Rumshisky

		Compression	Performance	Speedup	Model	Evaluation
Distillation	DistilBERT (Sanh et al., 2019)	×2.5	90%	×1.6	BERT$_6$	All GLUE tasks
	BERT$_6$-PKD (Sun et al., 2019a)	×1.6	97%	×1.9	BERT$_6$	No WNLI, CoLA and STS-B
	BERT$_3$-PKD (Sun et al., 2019a)	×2.4	92%	×3.7	BERT$_3$	No WNLI, CoLA and STS-B
	(Aguilar et al., 2019)	×2	94%	-	BERT$_6$	CoLA, MRPC, QQP, RTE
	BERT-48 (Zhao et al., 2019)	×62	87%	×77	BERT$_{12}^{*\dagger}$	MNLI, MRPC, SST-2
	BERT-192 (Zhao et al., 2019)	×5.7	94%	×22	BERT$_{12}^{*\dagger}$	MNLI, MRPC, SST-2
	TinyBERT (Jiao et al., 2019)	×7.5	96%	×9.4	BERT$_4^{*\dagger}$	All GLUE tasks
	MobileBERT (Sun et al.)	×4.3	100%	×4	BERT$_{24}^{\dagger}$	No WNLI
	PD (Turc et al., 2019)	×1.6	98%	×2.5^3	BERT$_6^{\dagger}$	No WNLI, CoLA and STS-B
	MiniBERT(Tsai et al., 2019)	×6§	98%	×27§	mBERT$_3^{\dagger}$	CoNLL-2018 POS and morphology
	BiLSTM soft (Tang et al., 2019)	×110	91%	×434‡	BiLSTM$_1$	MNLI, QQP, SST-2
Quant.	Q-BERT (Shen et al., 2019)	×13	99%	-	BERT$_{12}$	MNLI, SST-2
	Q8BERT (Zafrir et al., 2019)	×4	99%	-	BERT$_{12}$	All GLUE tasks
Other	ALBERT-base (Lan et al., 2019)	×9	97%	×5.6	BERT$_{12}$	MNLI, SST-2
	ALBERT-xxlarge (Lan et al., 2019)	×0.47	107%	×0.3	BERT$_{12}^{**}$	MNLI, SST-2
	BERT-of-Theseus (Xu et al., 2020)	×1.6	98%	-	BERT$_6$	No WNLI

Quantization = reducing precision, often also reducing memory footprint

Smaller

		Compression	Performance	Speedup	Model	Evaluation
Distillation	DistilBERT (Sanh et al., 2019)	$\times 2.5$	90%	$\times 1.6$	$BERT_6$	All GLUE tasks
	$BERT_6$-PKD (Sun et al., 2019a)	$\times 1.6$	97%	$\times 1.9$	$BERT_6$	No WNLI, CoLA and STS-B
	$BERT_3$-PKD (Sun et al., 2019a)	$\times 2.4$	92%	$\times 3.7$	$BERT_3$	No WNLI, CoLA and STS-B
	(Aguilar et al., 2019)	$\times 2$	94%	-	$BERT_6$	CoLA, MRPC, QQP, RTE
	BERT-48 (Zhao et al., 2019)	$\times 62$	87%	$\times 77$	$BERT_{12}{}^{*\dagger}$	MNLI, MRPC, SST-2
	BERT-192 (Zhao et al., 2019)	$\times 5.7$	94%	$\times 22$	$BERT_{12}{}^{*\dagger}$	MNLI, MRPC, SST-2
	TinyBERT (Jiao et al., 2019)	$\times 7.5$	96%	$\times 9.4$	$BERT_4{}^{*\dagger}$	All GLUE tasks
	MobileBERT (Sun et al.)	$\times 4.3$	100%	$\times 4$	$BERT_{24}{}^{\dagger}$	No WNLI
	PD (Turc et al., 2019)	$\times 1.6$	98%	$\times 2.5^3$	$BERT_6{}^{\dagger}$	No WNLI, CoLA and STS-B
	MiniBERT(Tsai et al., 2019)	$\times 6^{\S}$	98%	$\times 27^{\S}$	$mBERT_3{}^{\dagger}$	CoNLL-2018 POS and morphology
	BiLSTM soft (Tang et al., 2019)	$\times 110$	91%	$\times 434^{\ddagger}$	$BiLSTM_1$	MNLI, QQP, SST-2
Quant.	Q-BERT (Shen et al., 2019)	$\times 13$	99%	-	$BERT_{12}$	MNLI, SST-2
	Q8BERT (Zafrir et al., 2019)	$\times 4$	99%	-	$BERT_{12}$	All GLUE tasks
Other	ALBERT-base (Lan et al., 2019)	$\times 9$	97%	$\times 5.6$	$BERT_{12}{}^{**}$	MNLI, SST-2
	ALBERT-xxlarge (Lan et al., 2019)	$\times 0.47$	107%	$\times 0.3$	$BERT_{12}{}^{**}$	MNLI, SST-2
	BERT-of-Theseus (Xu et al., 2020)	$\times 1.6$	98%	-	$BERT_6$	No WNLI

"A Primer in BERTology: What we know about how BERT works"
Anna Rogers, Olga Kovaleva, Anna Rumshisky (Preprint 2019)

Faster

		Compression	Performance	Speedup	Model	Evaluation
Distillation	DistilBERT (Sanh et al., 2019)	×2.5	90%	×1.6	$BERT_6$	All GLUE tasks
	$BERT_6$-PKD (Sun et al., 2019a)	×1.6	97%	×1.9	$BERT_6$	No WNLI, CoLA and STS-B
	$BERT_3$-PKD (Sun et al., 2019a)	×2.4	92%	×3.7	$BERT_3$	No WNLI, CoLA and STS-B
	(Aguilar et al., 2019)	×2	94%	-	$BERT_6$	CoLA, MRPC, QQP, RTE
	BERT-48 (Zhao et al., 2019)	×62	87%	×77	$BERT_{12}{}^{*\dagger}$	MNLI, MRPC, SST-2
	BERT-192 (Zhao et al., 2019)	×5.7	94%	×22	$BERT_{12}{}^{*\dagger}$	MNLI, MRPC, SST-2
	TinyBERT (Jiao et al., 2019)	×7.5	96%	×9.4	$BERT_4{}^{*\dagger}$	All GLUE tasks
	MobileBERT (Sun et al.)	×4.3	100%	×4	$BERT_{24}{}^{\dagger}$	No WNLI
	PD (Turc et al., 2019)	×1.6	98%	$\times 2.5^3$	$BERT_6{}^{\dagger}$	No WNLI, CoLA and STS-B
	MiniBERT(Tsai et al., 2019)	$\times 6^{\S}$	98%	$\times 27^{\S}$	$mBERT_3{}^{\dagger}$	CoNLL-2018 POS and morphology
	BiLSTM soft (Tang et al., 2019)	×110	91%	$\times 434^{\ddagger}$	$BiLSTM_1$	MNLI, QQP, SST-2
Quant.	Q-BERT (Shen et al., 2019)	×13	99%	-	$BERT_{12}$	MNLI, SST-2
	Q8BERT (Zafrir et al., 2019)	×4	99%	-	$BERT_{12}$	All GLUE tasks
Other	ALBERT-base (Lan et al., 2019)	×9	97%	×5.6	$BERT_{12}{}^{**}$	MNLI, SST-2
	ALBERT-xxlarge (Lan et al., 2019)	×0.47	107%	×0.3	$BERT_{12}{}^{**}$	MNLI, SST-2
	BERT-of-Theseus (Xu et al., 2020)	×1.6	98%	-	$BERT_6$	No WNLI

"A Primer in BERTology: What we know about how BERT works"
Anna Rogers, Olga Kovaleva, Anna Rumshisky (Preprint 2019)

Roughly as good

		Compression	Performance	Speedup	Model	Evaluation
Distillation	DistilBERT (Sanh et al., 2019)	$\times 2.5$	90%	$\times 1.6$	BERT$_6$	All GLUE tasks
	BERT$_6$-PKD (Sun et al., 2019a)	$\times 1.6$	97%	$\times 1.9$	BERT$_6$	No WNLI, CoLA and STS-B
	BERT$_3$-PKD (Sun et al., 2019a)	$\times 2.4$	92%	$\times 3.7$	BERT$_3$	No WNLI, CoLA and STS-B
	(Aguilar et al., 2019)	$\times 2$	94%	-	BERT$_6$	CoLA, MRPC, QQP, RTE
	BERT-48 (Zhao et al., 2019)	$\times 62$	87%	$\times 77$	BERT$_{12}$*†	MNLI, MRPC, SST-2
	BERT-192 (Zhao et al., 2019)	$\times 5.7$	94%	$\times 22$	BERT$_{12}$*†	MNLI, MRPC, SST-2
	TinyBERT (Jiao et al., 2019)	$\times 7.5$	96%	$\times 9.4$	BERT$_4$*†	All GLUE tasks
	MobileBERT (Sun et al.)	$\times 4.3$	100%	$\times 4$	BERT$_{24}$†	No WNLI
	PD (Turc et al., 2019)	$\times 1.6$	98%	$\times 2.5^3$	BERT$_6$†	No WNLI, CoLA and STS-B
	MiniBERT(Tsai et al., 2019)	$\times 6^{\S}$	98%	$\times 27^{\S}$	mBERT$_3$†	CoNLL-2018 POS and morphology
	BiLSTM soft (Tang et al., 2019)	$\times 110$	91%	$\times 434^{\ddagger}$	BiLSTM$_1$	MNLI, QQP, SST-2
Quant.	Q-BERT (Shen et al., 2019)	$\times 13$	99%	-	BERT$_{12}$	MNLI, SST-2
	Q8BERT (Zafrir et al., 2019)	$\times 4$	99%	-	BERT$_{12}$	All GLUE tasks
Other	ALBERT-base (Lan et al., 2019)	$\times 9$	97%	$\times 5.6$	BERT$_{12}$**	MNLI, SST-2
	ALBERT-xxlarge (Lan et al., 2019)	$\times 0.47$	107%	$\times 0.3$	BERT$_{12}$**	MNLI, SST-2
	BERT-of-Theseus (Xu et al., 2020)	$\times 1.6$	98%	-	BERT$_6$	No WNLI

"A Primer in BERTology: What we know about how BERT works"
Anna Rogers, Olga Kovaleva, Anna Rumshisky (Preprint 2019)

- You can get equally good results with smaller models
- **BERT is not a cognitive model**
- We only know some of the security risks posed by BERT based models

BERT (and other masked language models) aren't human like

- They don't do things in a human-like way
  - With sufficient post-processing, you can extract linguistic structures from the weights of the model, but that's also true of the plain text input
- They're not **grounded**
  - Don't have generalizable, structured knowledge about the world
  - Not capable of reasoning ("If I have two apples and I give one away, I will have _____ apples.")
- BERT can be easily fooled by spurious statistical correspondences in the fine-tuning data
  - "Probing Neural Network Comprehension of Natural Language Arguments" by Timothy Niven, Hung-Yu Kao (ACL 2019)

BERT (and other masked language models) aren't human like

- They don't do things in a human-like way
  - With sufficient post-processing, you can extract linguistic structures from the weights of the model, but that's also true of the plain text input
- They're not **grounded**
  - Don't have generalizable, structured knowledge about the world
  - Not capable of reasoning ("If I have two apples and I give one away, I will have _____ apples.")
- BERT can be easily fooled by spurious statistical correspondences in the fine-tuning data
  - "Probing Neural Network Comprehension of Natural Language Arguments" by Timothy Niven, Hung-Yu Kao (ACL 2019)

It can be helpful to think of these sorts of models as if they had Wernicke's aphasia: they produce language but without any understanding of what it means.

- You can get equally good results with smaller models
- BERT is not a cognitive model
- **We only know some of the security risks posed by BERT based models**

Universal triggers are learnable strings that can dramatically decrease task performance or cause generated text to be racist/homophobic.

Task	Input (red = trigger)	Model Prediction
Sentiment Analysis	**zoning tapping fiennes** Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride. . .	Positive $\rightarrow$ Negative
	zoning tapping fiennes As surreal as a dream and as detailed as a photograph, as visually dexterous as it is at times imaginatively overwhelming.	Positive $\rightarrow$ Negative

"Universal Adversarial Triggers for Attacking and Analyzing NLP"
Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, Sameer Singh
(Preprint 2019)

- You can get equally good results with smaller models
- BERT is not a cognitive model
- We only know some of the security risks posed by BERT based models

I'm not trying to tear down BERT! It's an important NLP paper and made a large impact on the field.

We just have a lot to learn about masked language models and transformers.

Thanks! Questions?

@rctatman