Intelligent Agents

BERT, GPT and Further Developments for Natural Language Inference

Ralf Möller
Universität zu Lübeck
Institut für Informationssysteme

Recap: From ELMo via Transformers to BERT

- Language modeling is the "ultimate" NLP task
  - I.e., a perfect language model is also a perfect question answering/entailment/sentiment analysis model
  - Training a massive language model learns millions of latent features which are useful for these other NLP tasks
- E.g., for natural language inference
  - No internal "logical" representation
  - Use language directly to infer new propositions
    - ~~What kind of a thing is the meaning of a sentence?~~
    - What concrete phenomena do you have to deal with to understand a sentence?

BERT and Add-ons: 2019 Success Stories on SQuAD 2.0

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 May 21, 2019	XLNet (single model) *XLNet Team*	86.346	89.133
5 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886

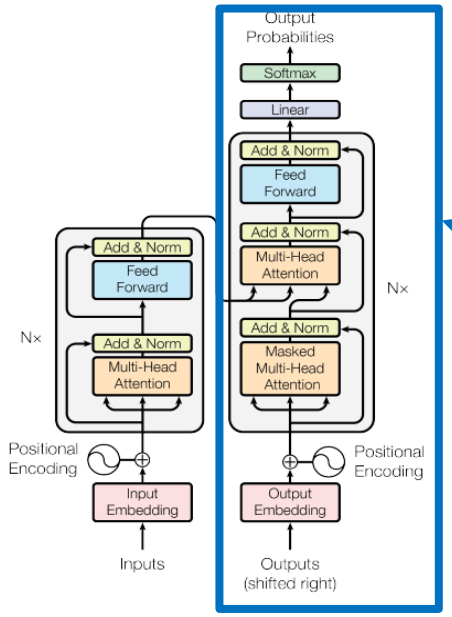
Exact Match and F1 scores

DAE stands for "**Data augmentation**" and "**Domain adaption**"

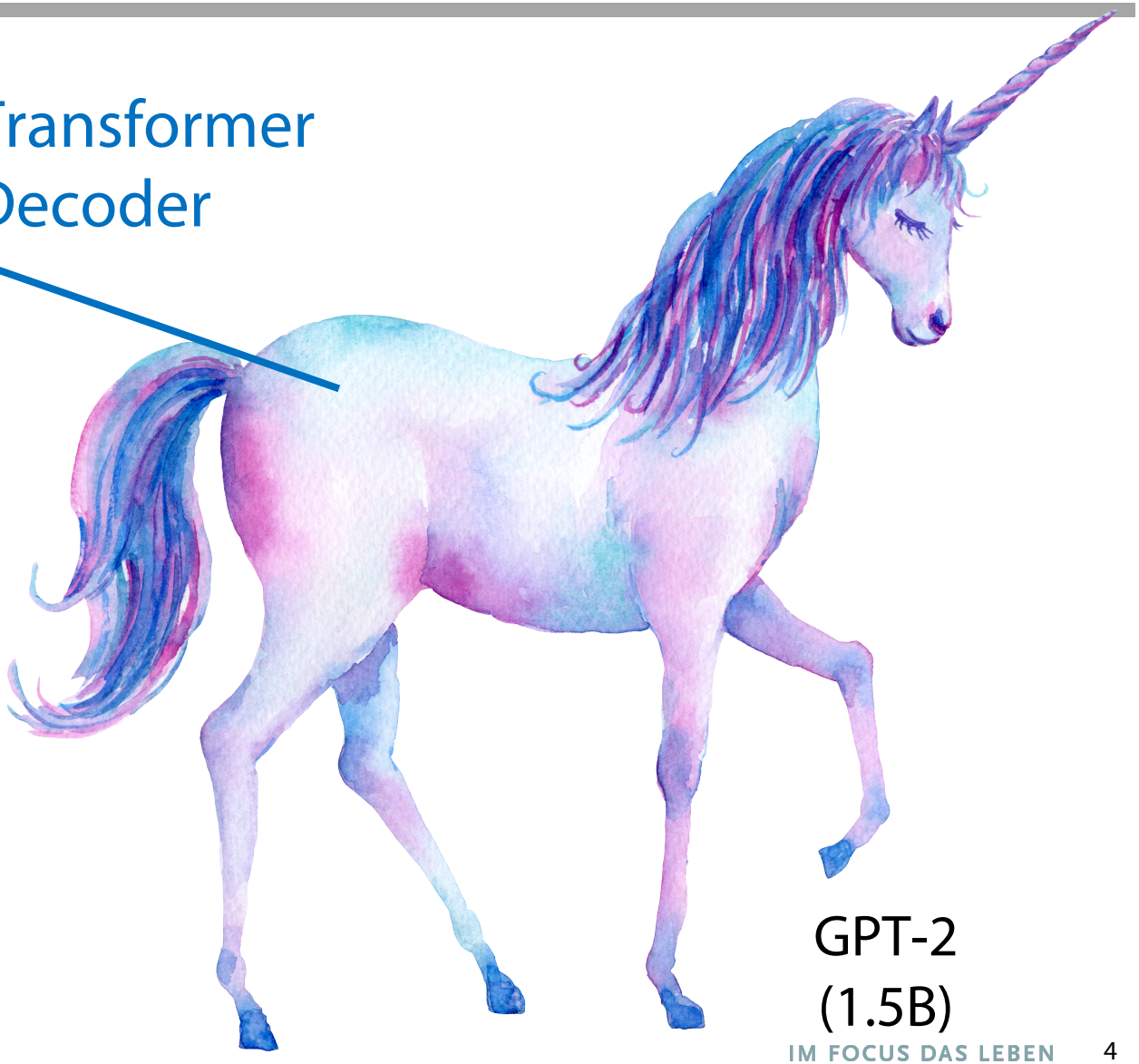
AoA stands for "**Attention-over-Attention**"

https://rajpurkar.github.io/SQuAD-explorer/

Generative Pre-Training (GPT)



Transformer Decoder



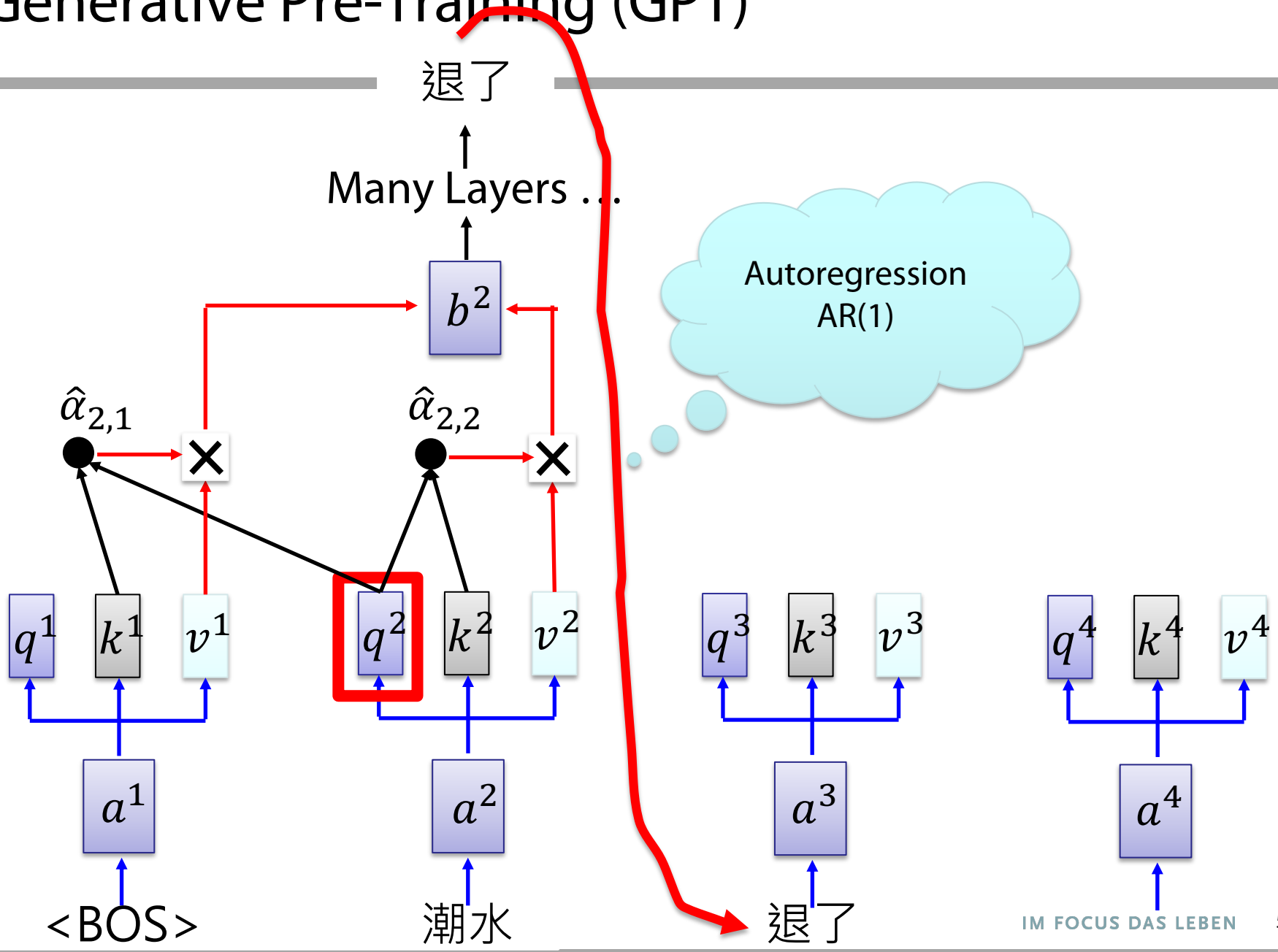
BERT (340M)

ELMO (94M)

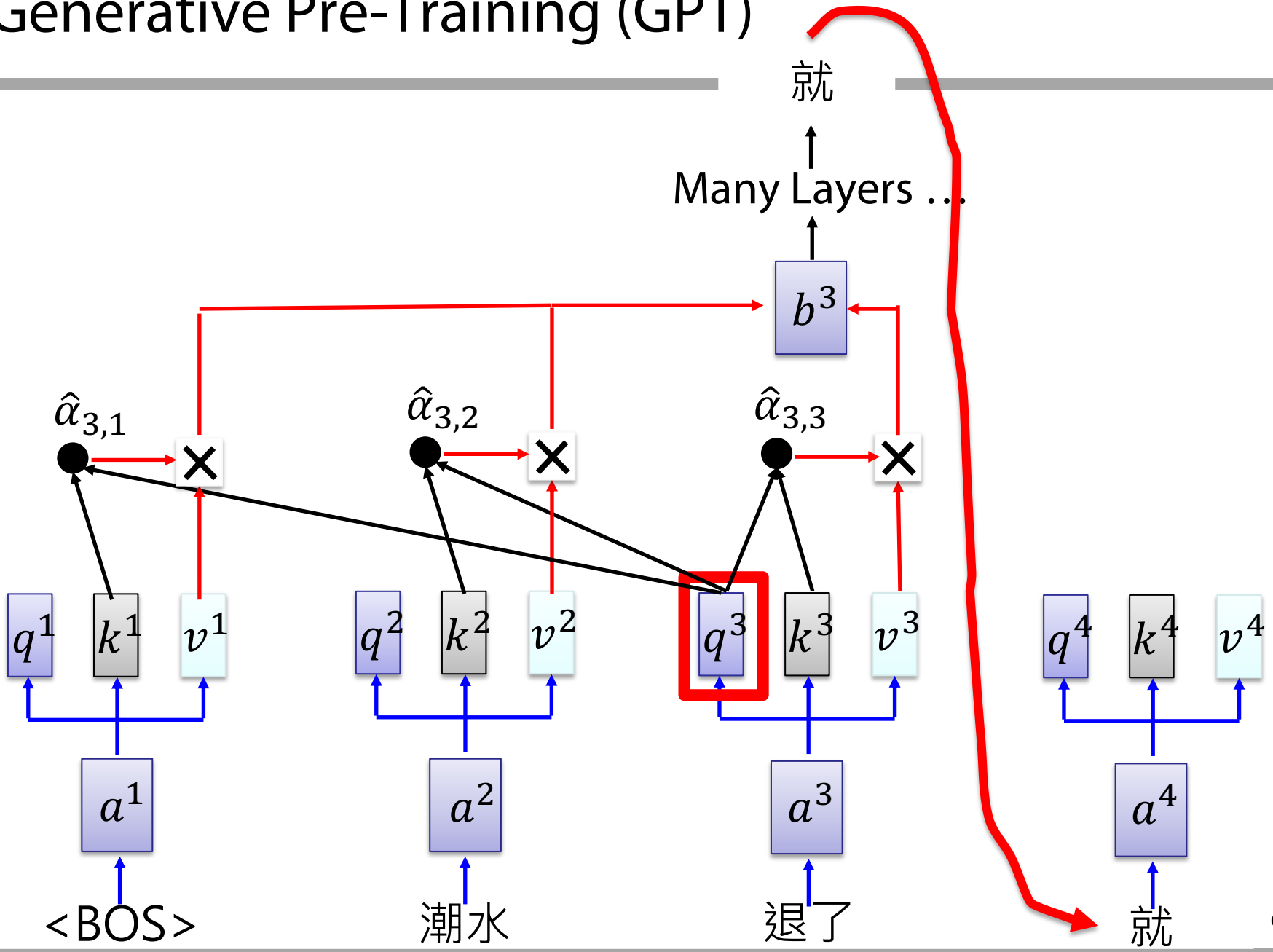
GPT-2 (1.5B)

https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Generative Pre-Training (GPT)



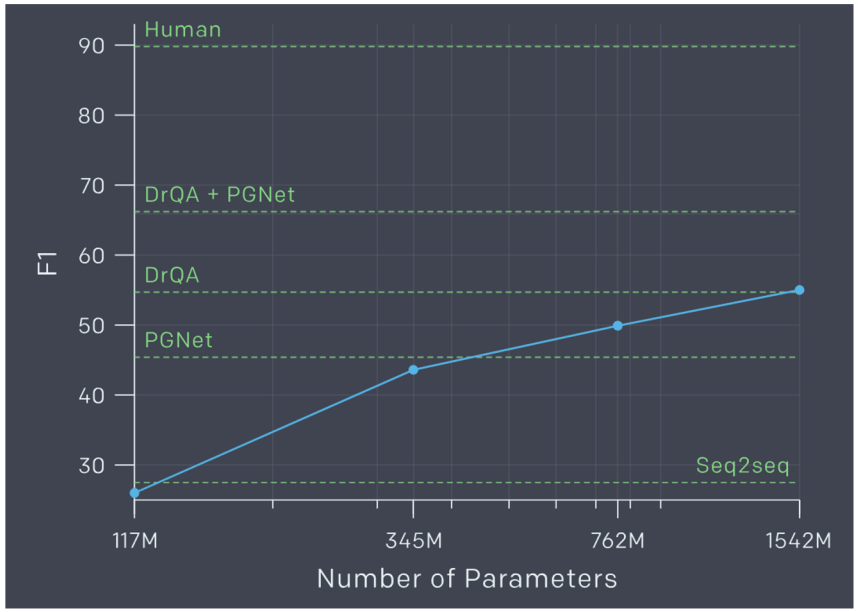
Generative Pre-Training (GPT)



One-shot or Few-shot Learning? GPT-3 175B

- Reading Comprehension

$d_1, d_2, \cdots, d_N,$
"Q:", $q_1, q_2, \cdots, q_N,$
"A:"



- Summarization $d_1, d_2, \cdots, d_N,$"TL;DR:"

TL;DR=Too long; didn't read.

- Translation

English sentence 1	=	French sentence 1
English sentence 2	=	French sentence 2
English sentence 3	=	

Tom B. Brown et al.
Language Models are Few-Shot Learners.
https://arxiv.org/abs/2005.14165.
2020.

Examples

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

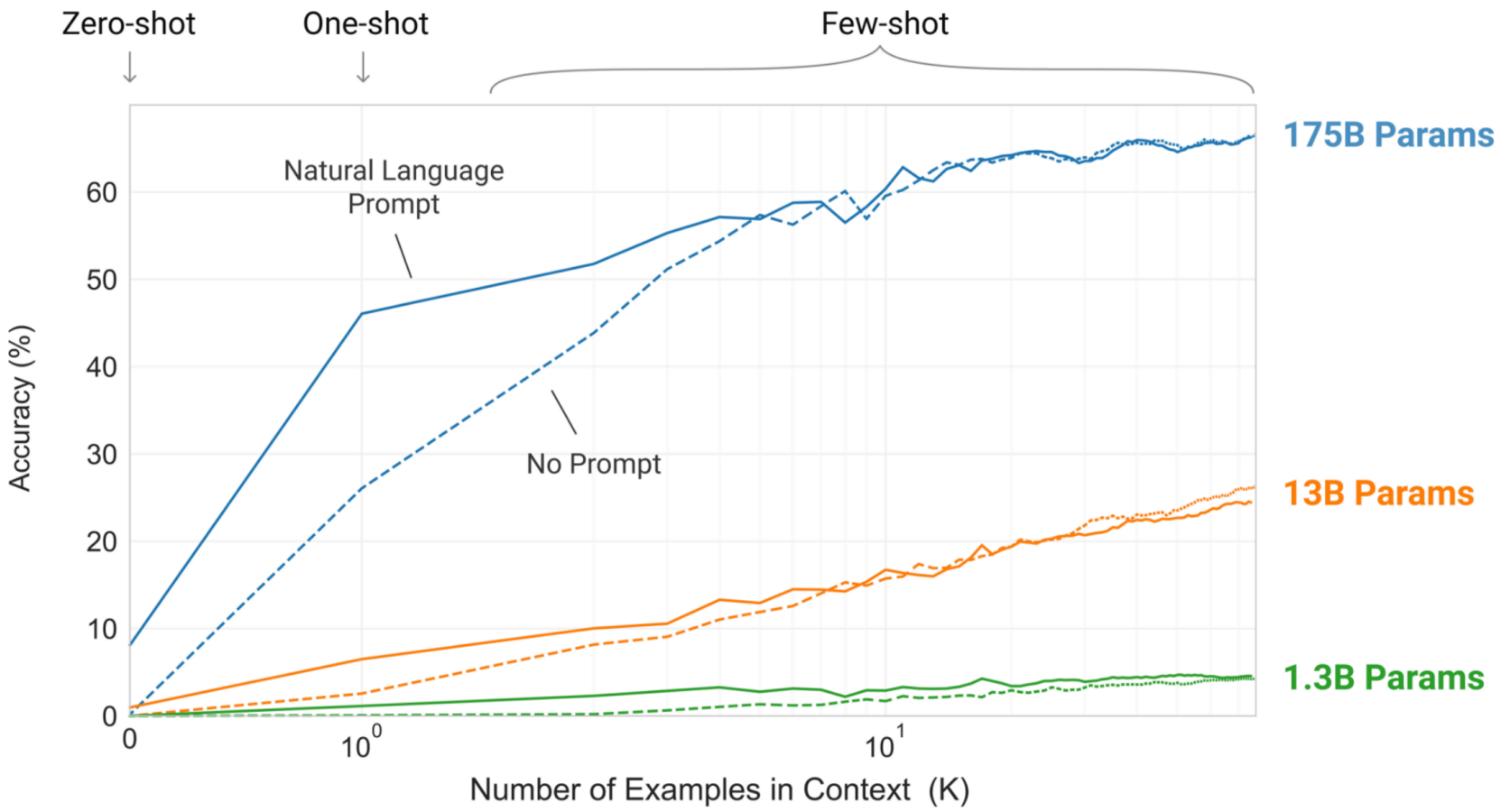
Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```


https://blog.inten.to/gpt-3-language-models-are-few-shot-learners-a13d1ae8b1f9

Evaluation


Zero-shot
One-shot
Few-shot
Natural Language Prompt
No Prompt
175B Params
13B Params
1.3B Params
Accuracy (%)
0
10
20
30
40
50
60
0
10^0
10^1
Number of Examples in Context (K)

Acknowledgements

Contextual Word Representations with BERT and Other Pre-trained Language Models

Jacob Devlin
Google AI Language

SQuAD 2.0 (2020)

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Dec 01, 2020	EntitySpanFocusV2 (ensemble) *RICOH_SRCB_DML*	90.521	92.824
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.420	92.799
4 Sep 11, 2020	EntitySpanFocus+AT (ensemble) *RICOH_SRCB_DML*	90.454	92.748
4 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777
5 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.115	92.580
6 Nov 01, 2020	electra+nlayers+kdav (ensemble) *oppo.tensorlab*	90.002	92.497

https://rajpurkar.github.io/SQuAD-explorer/

SQuAD 1.1 (2020)

SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Apr 10, 2020	LUKE (single model) *Studio Ousia & NAIST & RIKEN AIP*	**90.202**	**95.379**
2 May 21, 2019	XLNet (single model) *Google Brain & CMU*	89.898	95.080
3 Dec 11, 2019	XLNET-123++ (single model) *MST/EOI* http://tia.today	89.856	94.903
3 Aug 11, 2019	XLNET-123 (single model) *MST/EOI*	89.646	94.930
4 Sep 25, 2019	BERTSP (single model) *NEUKG* http://www.techkg.cn/	88.912	94.584

RoBERTA

- *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (Liu et al, University of Washington and Facebook, 2019)
  - Trained BERT for more epochs and/or on more data
    - Showed that more epochs alone helps, even on same data
    - More data also helps
  - Masking and pre-training data slightly improved

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv, abs/1907.11692*. **2019**.

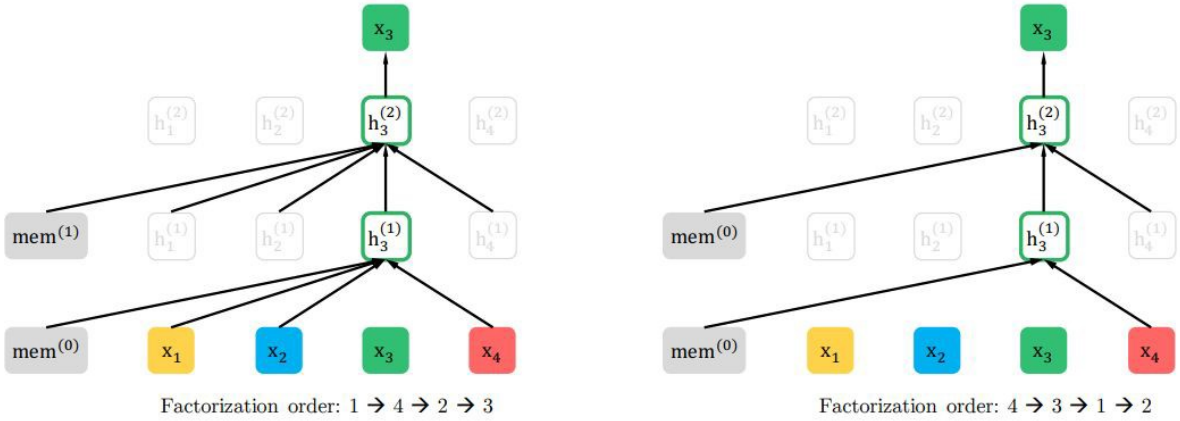
XLNet

- XLNet: Generalized Autoregressive Pretraining for Language Understanding (Yang et al, CMU and Google, 2019)
- Innovation #1: Relative position embeddings
  - Sentence: John ate a hot dog
  - Absolute attention: "How much should dog attend to hot (in any position), and how much should dog in position 4 attend to the word in position 3? (Or 508 attend to 507, …)"
  - Relative attention: "How much should dog attend to hot (in any position) and how much should dog attend to the previous word?"

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Proc. NeurIPS-19. **2019**.

XLNet

- Innovation #2: Permutation Language Modeling
  - In a left-to-right language model, every word is predicted based on all of the words to its left
  - Instead: Randomly permute the order for *every training sentence*
  - Equivalent to masking, but many more predictions per sentence
  - Can be done efficiently with Transformers



XLNet

- Also used more data and bigger models, but showed that innovations improved on BERT even with same data and model size
- XLNet results:

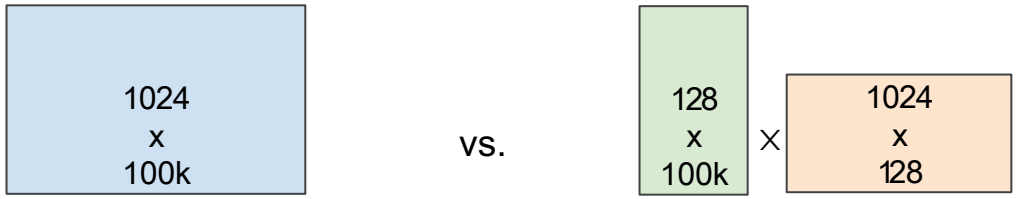
Multi-Genre Natural-Language-Inference

Model	MNLI	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B
Single-task single models on dev								
BERT [2]	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0
RoBERTa [21]	90.2/90.2	94.7	92.2	**86.6**	96.4	**90.9**	68.0	92.4
XLNet	**90.8/90.8**	**94.9**	**92.3**	85.9	**97.0**	90.8	**69.0**	**92.5**

ALBERT

- *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations* (Lan et al, Google and TTI Chicago, 2019)
- Innovation #1: Factorized embedding parameterization

Break down token **embeddings** into two small **embedding** matrixes. After applying this decomposition, **embeddings parameters** can be reduced from (number of tokens * hidden layer size) to (number of tokens * token **embedding** size + token **embedding** size * hidden layer size).



Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In: Proc. ICLR **2019**.

ALBERT

- Innovation #2: Cross-layer parameter sharing
  - Share all parameters between Transformer layers
- Results:

Models	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS
Single-task single models on dev								
BERT-large	86.6	92.3	91.3	70.4	93.2	88.0	60.6	90.0
XLNet-large	89.8	93.9	91.8	83.8	95.6	89.2	63.6	91.8
RoBERTa-large	90.2	94.7	**92.2**	86.6	96.4	**90.9**	68.0	92.4
ALBERT (1M)	90.4	95.2	92.0	88.1	96.8	90.2	68.7	92.7
ALBERT (1.5M)	**90.8**	**95.3**	**92.2**	**89.2**	**96.9**	**90.9**	**71.4**	**93.0**

- ALBERT is light in terms of *parameters*, not *speed*

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3	4.7x
	large	334M	92.2/85.5	85.0/82.2	86.6	93.0	73.9	85.2	1.0
ALBERT	base	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1	5.6x
	large	18M	90.6/83.9	82.3/79.4	83.5	91.7	68.5	82.4	1.7x
	xlarge	60M	92.5/86.1	86.1/83.1	86.4	92.4	74.8	85.5	0.6x
	xxlarge	235M	**94.1/88.3**	**88.1/85.1**	**88.0**	**95.2**	**82.3**	**88.7**	0.3x

T5

- *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* (Raffel et al, Google, 2019)
- Ablated many aspects of pre-training:
  - Model size
  - Amount of training data
  - Domain/cleanness of training data
  - Pre-training objective details (e.g., span length of masked text)
  - Ensembling
  - Finetuning recipe (e.g., only allowing certain layers to finetune)
  - Multi-task training

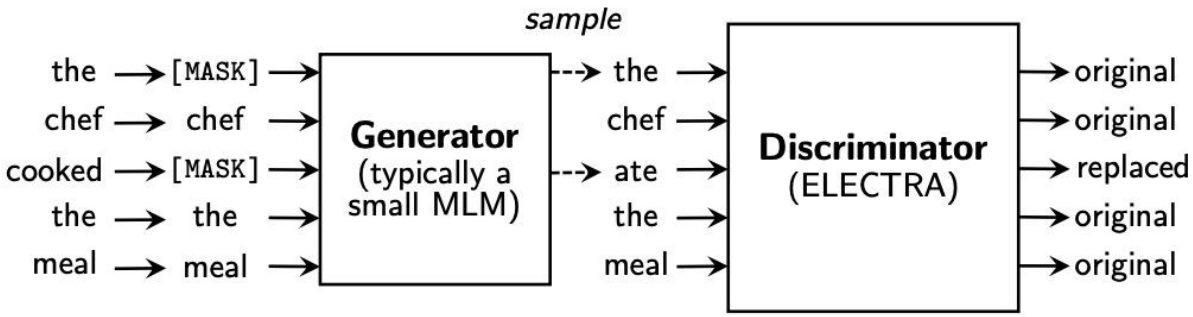
Raffel, Colin & Shazeer, Noam & Roberts, Adam & Lee, Katherine & Narang, Sharan & Matena, Michael & Zhou, Yanqi & Li, Wei & Liu, Peter. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. **2019**.

T5

- Conclusions:
  - Scaling up model size and amount of training data helps a lot
  - Best model is 11B parameters (BERT-Large is 330M), trained on 120B words of cleaned common crawl text
  - Exact masking/corruptions strategy doesn't matter that much
  - Mostly negative results for better finetuning and multi-task strategies

ELECTRA

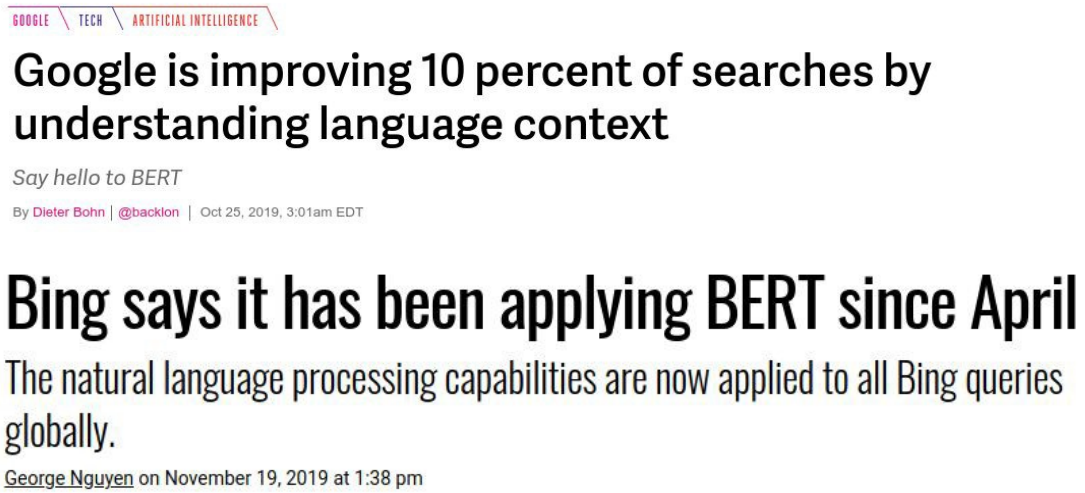
- *ELECTRA:Pre-training Text Encoders as Discriminators Rather Than Generators* (Clark et al, 2020)
- Train model to discriminate locally plausible text from real text
- Used, e.g., with ALBERT



Clark, K., Luong, M., Le, Q.V., & Manning, C.D.. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ArXiv, abs/2003.10555. **2020**.

Applying Models to Production Services?

- BERT and other pre-trained language models are extremely large and expensive
- How are companies applying them to low-latency production services?



GOOGLE | TECH | ARTIFICIAL INTELLIGENCE

Google is improving 10 percent of searches by understanding language context

Say hello to BERT

By Dieter Bohn | @backlon | Oct 25, 2019, 3:01am EDT

Bing says it has been applying BERT since April

The natural language processing capabilities are now applied to all Bing queries globally.

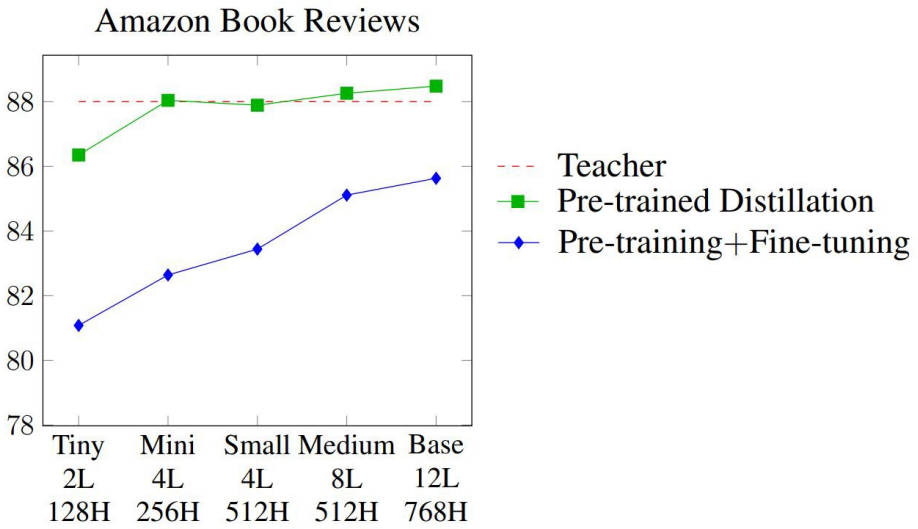
George Nguyen on November 19, 2019 at 1:38 pm

Distillation

- Answer: Distillation (a.k.a., model compression)
- Idea has been around for a long time:
  - *Model Compression* (Bucila et al, 2006)
  - *Distilling the Knowledge in a Neural Network* (Hinton et al, 2015)
- Simple technique:
  - Train "Teacher": Use SOTA pre-training + fine-tuning technique to train model with maximum accuracy
  - Label a large amount of unlabeled input examples with Teacher
  - Train "Student": Much smaller model (e.g., 50x smaller) which is trained to mimic Teacher output
  - Student objective is typically Mean Square Error or Cross Entropy

Distillation

- Example distillation results
  - 50k labeled examples, 8M unlabeled examples



Well-Read Students Learn Better: On the Importance of Pre-training Compact Models (Turc et al, 2020)

- Distillation works *much* better than pre-training + fine-tuning with smaller model

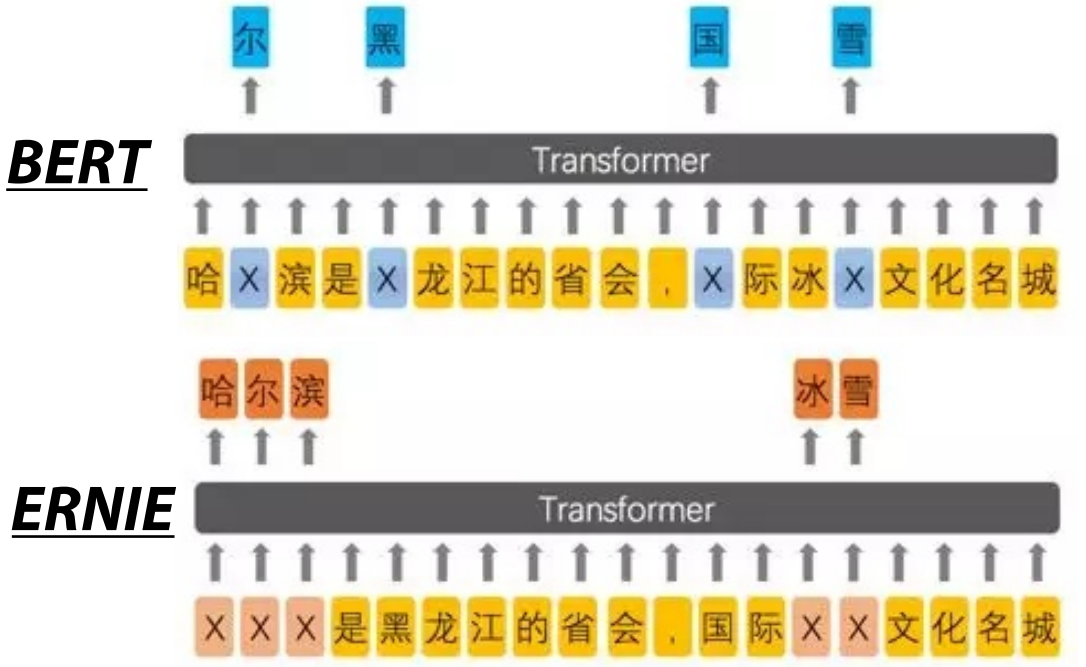
Why does distillation work so well?

A hypothesis:

- Finetuning mostly just picks up and tweaks existing latent features
- This requires an oversized model, because only a subset of the features are useful for any given task
- Distillation allows the model to only focus on those features
- Supporting evidence: Simple self-distillation of a small model (e.g., distilling a smaller BERT model) doesn't work very well

Enhanced Representation through Knowledge Integration (ERNIE)

- Incorporation of knowledge graphs
- Designed for Chinese (ERNIE-baidu)



Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu.
ERNIE: Enhanced Language Representation with Informative Entities.
In: Proc. ACL-19, 1441–1451. **2019**. https://arxiv.org/abs/1904.09223

Knowledge-aware **P**retrained **L**anguage **M**odels

Bert | Bert-wwm | ERNIE-baidu | SpanBert

ERNIE-thu | KnowBert | K-Bert | KEPLER | GLM

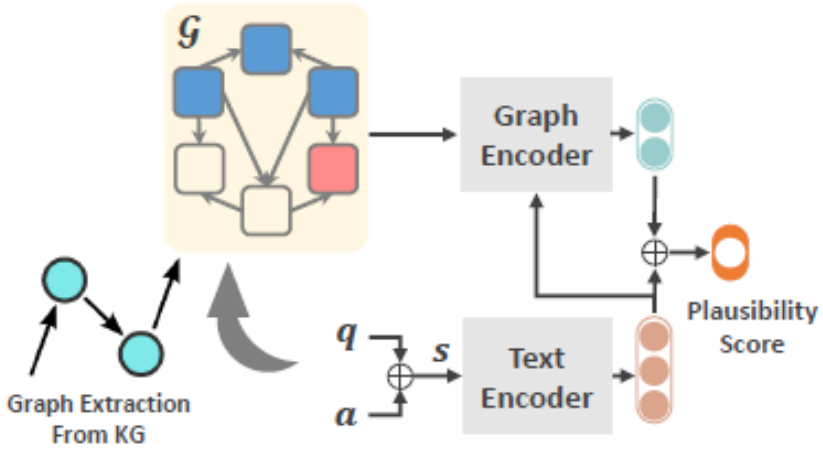
Knowledge-aware PLMs guided by KG

Knowledge-aware KG-enhanced QA

KagNet | CSQA | PG | MHGRN

BERT: Pre-training of deep bidirectional transformers for language understanding (NAACL 19)
Bert-wwm: Pre-Training with Whole Word Masking for Chinese BERT (Arxiv 19)
SpanBERT: Improving Pre-training by Representing and Predicting Spans (TACL 20)
ERNIE-baidu: Enhanced representation through knowledge integration (Arxiv 19)
ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding (AAAI 20)
ERNIE-thu: Enhanced Language Representation with Informative Entities (ACL 19)
K-BERT: Enabling Language Representation with Knowledge Graph (AAAI 20)
KnowBert: Knowledge enhanced contextual word representations (EMNLP 19)
KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation (Arxiv 19)
GLM: Exploiting Structured Knowledge in Text via Graph-Guided Representation Learning (Arxiv 20)
KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning (EMNLP 19)
CSQA: Graph-Based Reasoning over Heterogeneous External Knowledge for Common sense Question Answering (AAAI 20)
PG: Connecting the Dots: A Knowledgeable Path Generator for Commonsense Question Answering (EMNLP 2020 finding)
MHGRN: Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering (EMNLP 2020)

Unified Architecture in KG-enhanced QA



Graph Encoder: GNN, Relational Network…

Text Encoder: Bert, XLNet…

Next topic:
Graph encoding