자연어 처리 DAY 5
Self-supervised Pre-training Models

Jaegul Choo

Associate Professor, Graduate School of AI, KAIST

boostcamp AI Tech

DAVIAN

Data and Visual Analytics Lab

1.

Self-Supervised Pre-Training Models

- GPT-1
- BERT

Recent Trends

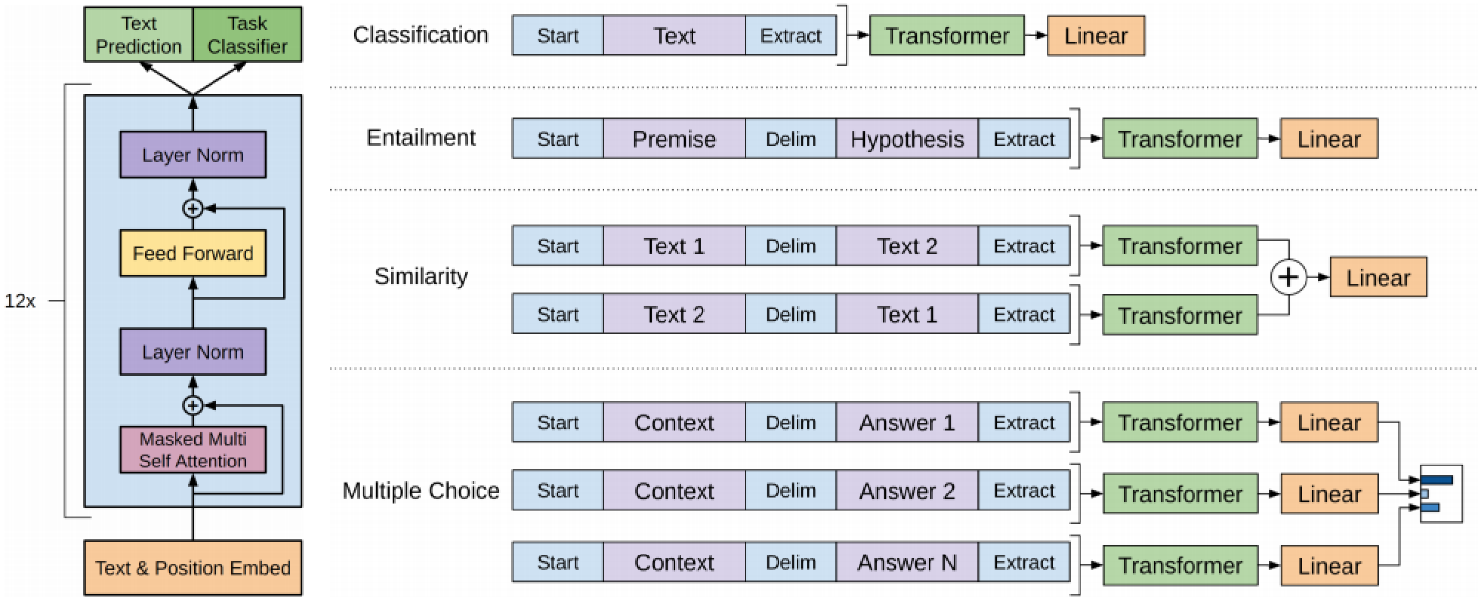
- Transformer model and its self-attention block has become a general-purpose sequence (or set) encoder and decoder in recent NLP applications as well as in other areas.
- Training deeply stacked Transformer models via a self-supervised learning framework has significantly advanced various NLP tasks through transfer learning, e.g., BERT, GPT-3, XLNet, ALBERT, RoBERTa, Reformer, T5, ELECTRA…
- Other applications are fast adopting the self-attention and Transformer architecture as well as self-supervised learning approach, e.g., recommender systems, drug discovery, computer vision, …
- As for natural language generation, self-attention models still requires a greedy decoding of words one at a time.

Attention Is All You Need, NeurIPS'17

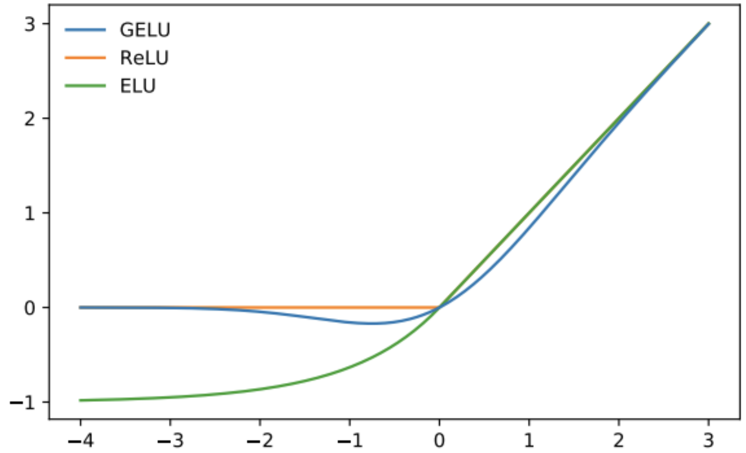
GPT-1

Improving Language Understanding by Generative Pre-training

- GPT-1
  - It introduces special tokens, such as <S> /<E>/ $, to achieve effective transfer learning during fine-tuning
  - It does not need to use additional task-specific architectures on top of transferred



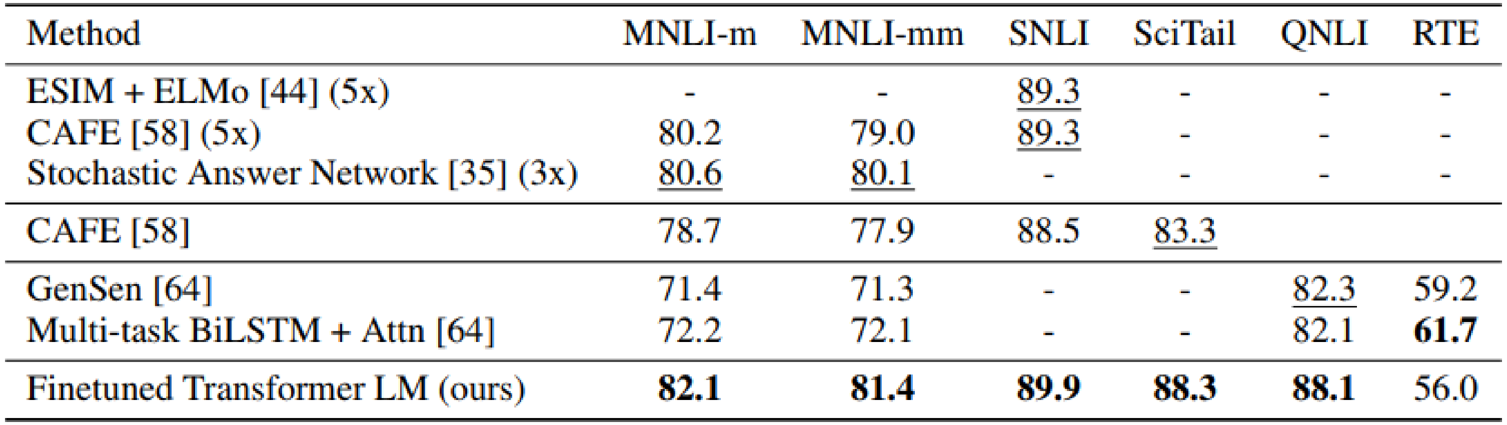
- 12-layer decoder-only transformer
- 12 head / 768 dimensional states
- GELU activation unit



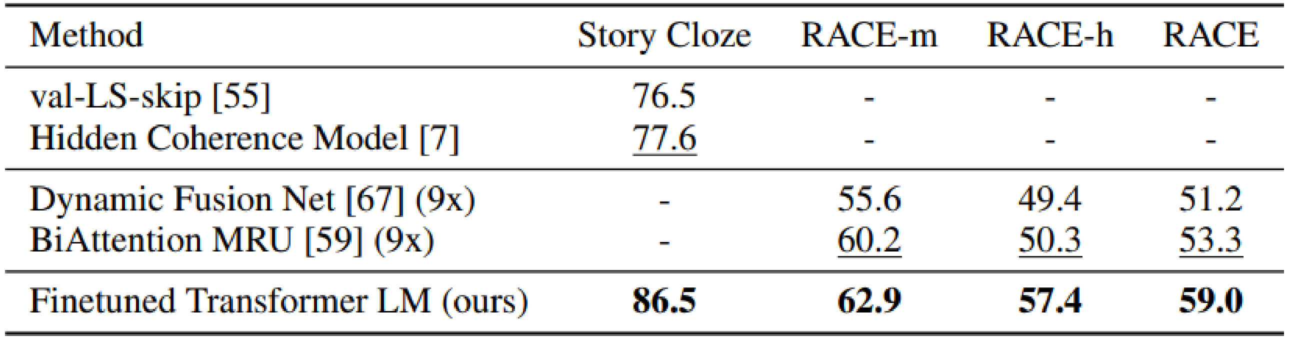
https://blog.openai.com/language-unsupervised/

Improving Language Understanding by Generative Pre-training

- Experimental Results



Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0



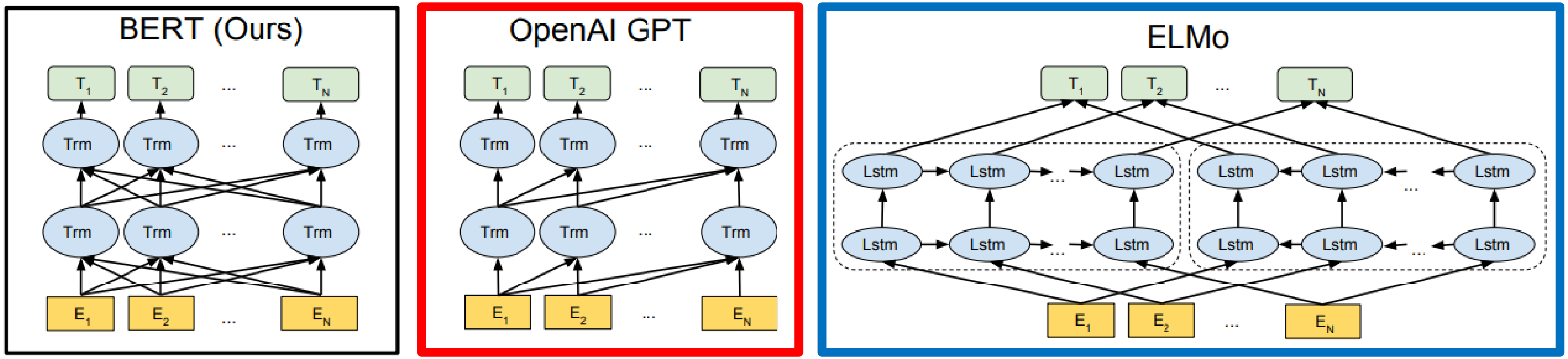
Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

https://blog.openai.com/language-unsupervised/

BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- Learn through masked language modeling task
- Use large-scale data and large-scale model



Unidirectional

Bi-LSTM

BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL'19

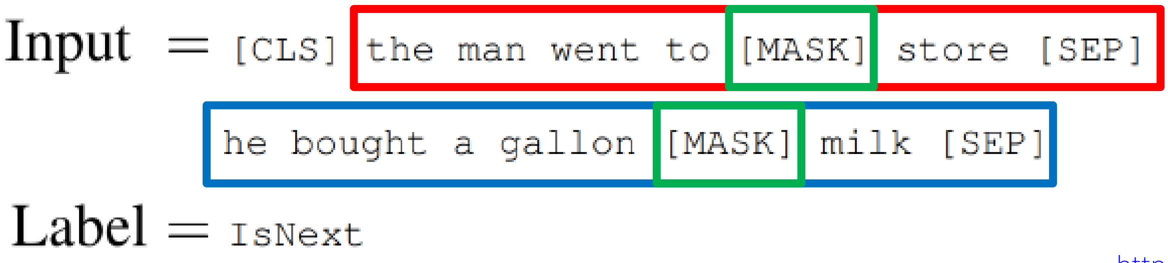
Masked Language Model

- Motivation
  - Language models only use left context or right context, but language understanding is bi-directional
- If we use bi-directional language model?
  - Problem: Words can "see themselves" (cheating) in a bi-directional encoder

Pre-training Tasks in BERT

- Masked Language Model (MLM)
  - Mask some percentage of the input tokens at random, and then predict those masked tokens.
  - 15% of the words to predict
    - 80% of the time, replace with [MASK]
    - 10% of the time, replace with a random word
    - 10% of the time, keep the sentence as same

- Next Sentence Prediction (NSP)
  - Predict whether Sentence B is an actual sentence that proceeds Sentence A, or a random sentence



Input = [CLS] the man went to [MASK] store [SEP]

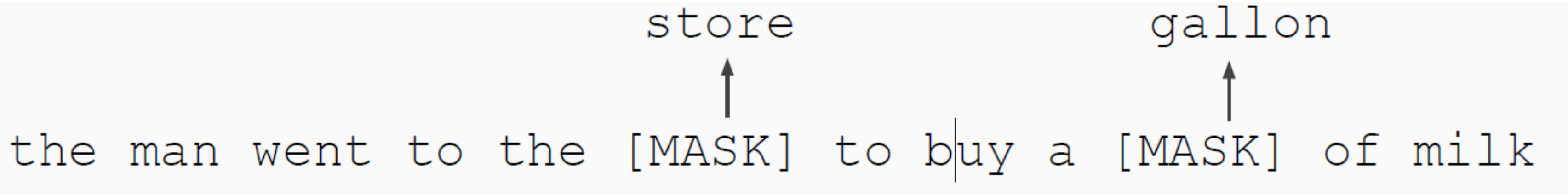
he bought a gallon [MASK] milk [SEP]

Label = IsNext

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

Pre-training Tasks in BERT: Masked Language Model

- **How to**
  - Mask out *k*% of the input words, and then predict the masked words
    - e.g., use *k*= 15%



store

gallon

the man went to the [MASK] to buy a [MASK] of milk

- Too little masking : Too expensive to train

- Too much masking : Not enough to capture context

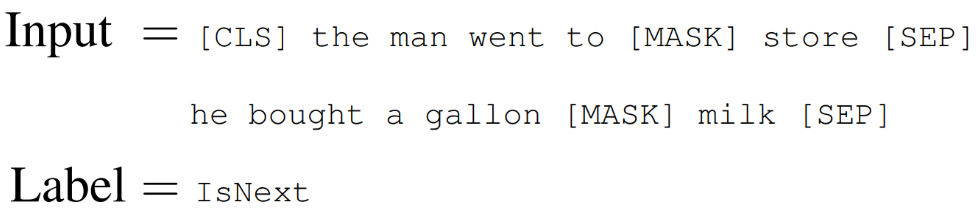
https://nlp.stanford.edu/seminar/details/jdevlin.pdf

Pre-training Tasks in BERT: Masked Language Model

- **Problem**
  - Mask token never seen during fine-tuning
- **Solution**
  - 15% of the words to predict, but don't replace with [MASK] 100% of the time. Instead:
    - 80% of the time, replace with [MASK]
      - went to the store → went to the [MASK]
    - 10% of the time, replace with a random word
      - went to the store → went to the running
    - 10% of the time, keep the same sentence
      - went to the store → went to the store

Pre-training Tasks in BERT: Next Sentence Prediction

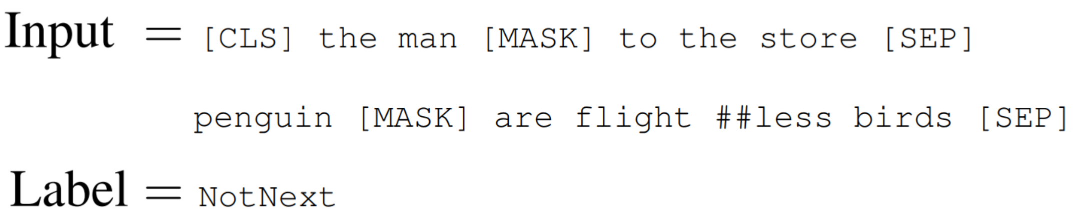
- To learn the relationships among sentences, predict whether Sentence B is an actual sentence that proceeds Sentence A, or a random sentence



Input = [CLS] the man went to [MASK] store [SEP]

he bought a gallon [MASK] milk [SEP]

Label = IsNext



Input = [CLS] the man [MASK] to the store [SEP]

penguin [MASK] are flight ##less birds [SEP]

Label = NotNext

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

BERT Summary

1. **Model Architecture**
   - BERT BASE: L = 12, H = 768, A = 12
   - BERT LARGE: L = 24, H = 1024, A = 16
2. **Input Representation**
   - WordPiece embeddings (30,000 WordPiece)
   - Learned positional embedding
   - [CLS] – Classification embedding
   - Packed sentence embedding [SEP]
   - Segment Embedding
3. **Pre-training Tasks**
   - Masked LM
   - Next Sentence Prediction

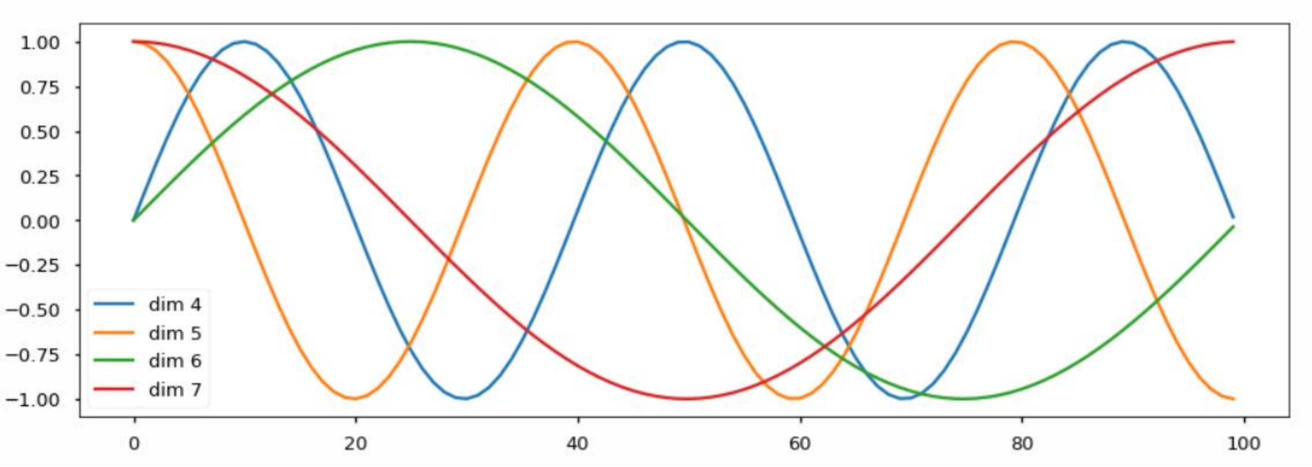
Transformer: Positional Encoding

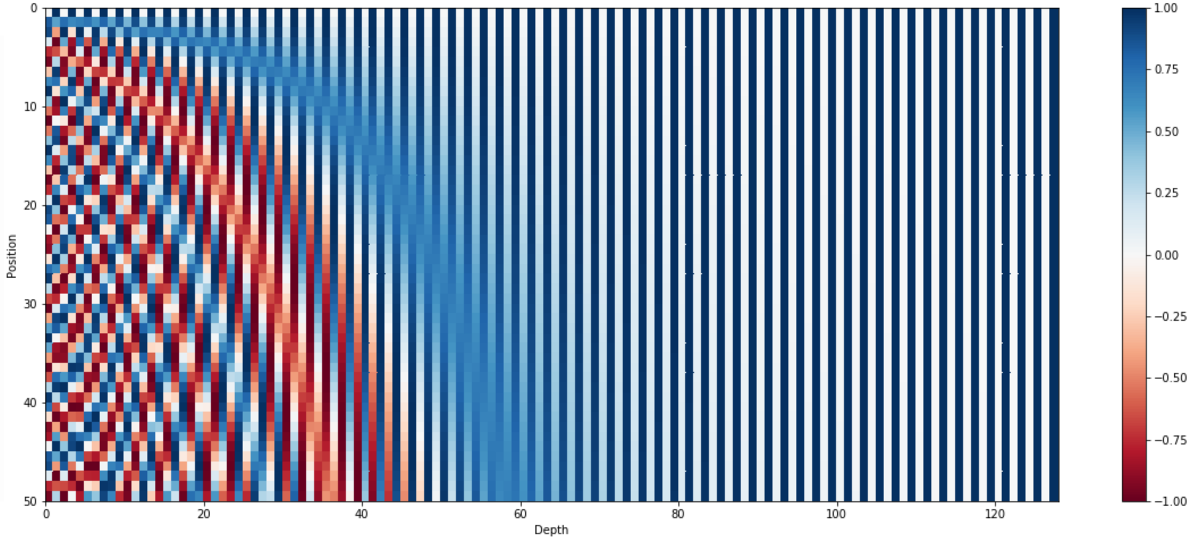
- Use sinusoidal functions of different frequencies

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

- Easily learn to attend by relative position, since for any fixed offset k, $PE_{(pos+k)}$ can be represented as linear function of $PE_{(pos)}$

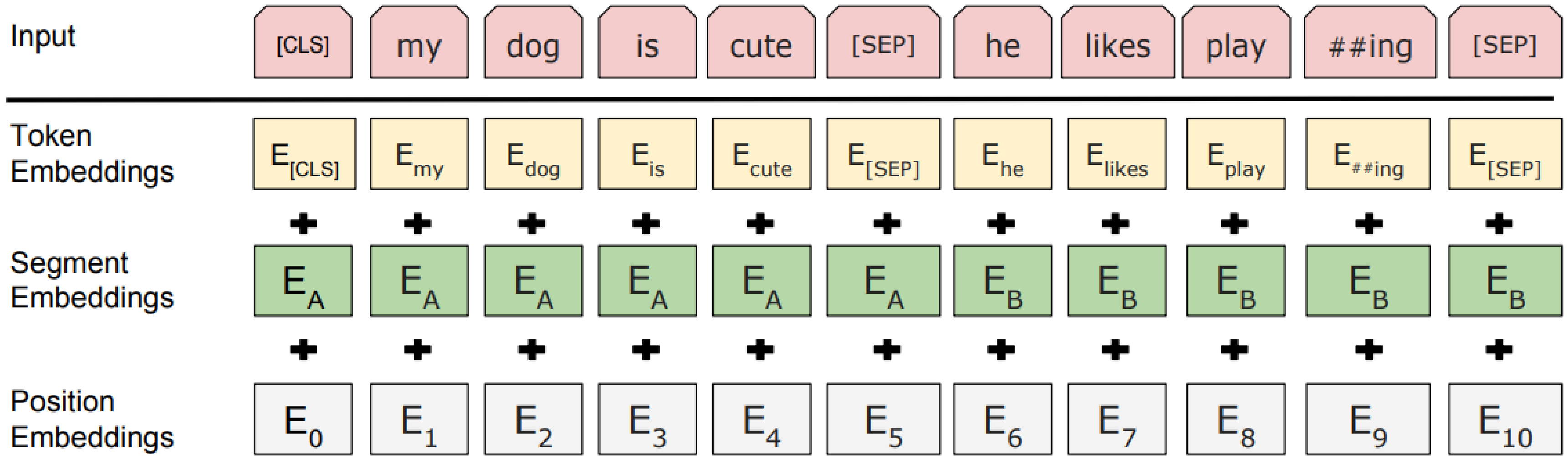




http://nlp.seas.harvard.edu/2018/04/03/attention

BERT: Input Representation

- The input embedding is the sum of the token embeddings, the segmentation embeddings and the position embeddings

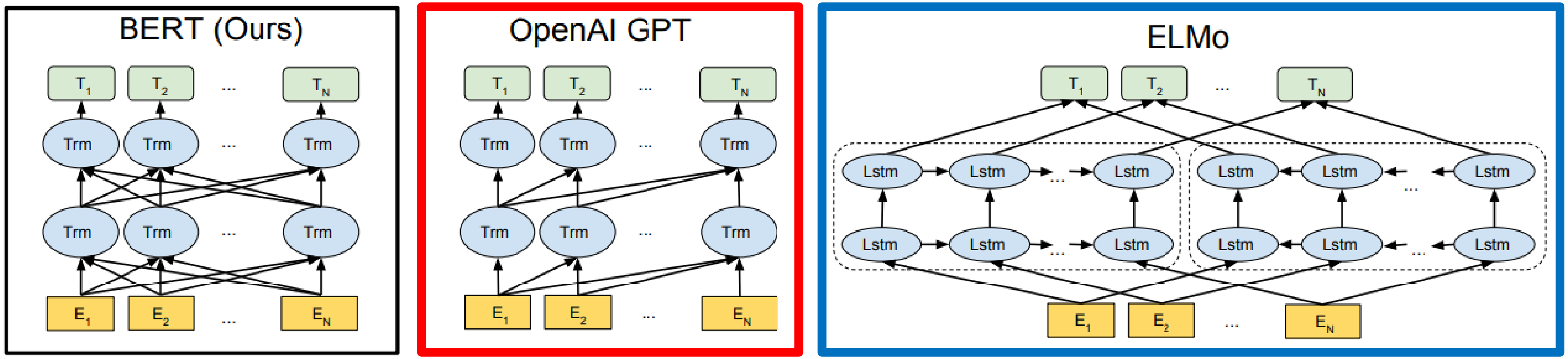


Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL'19

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- Learn through masked language modeling task
- Use large-scale data and large-scale model



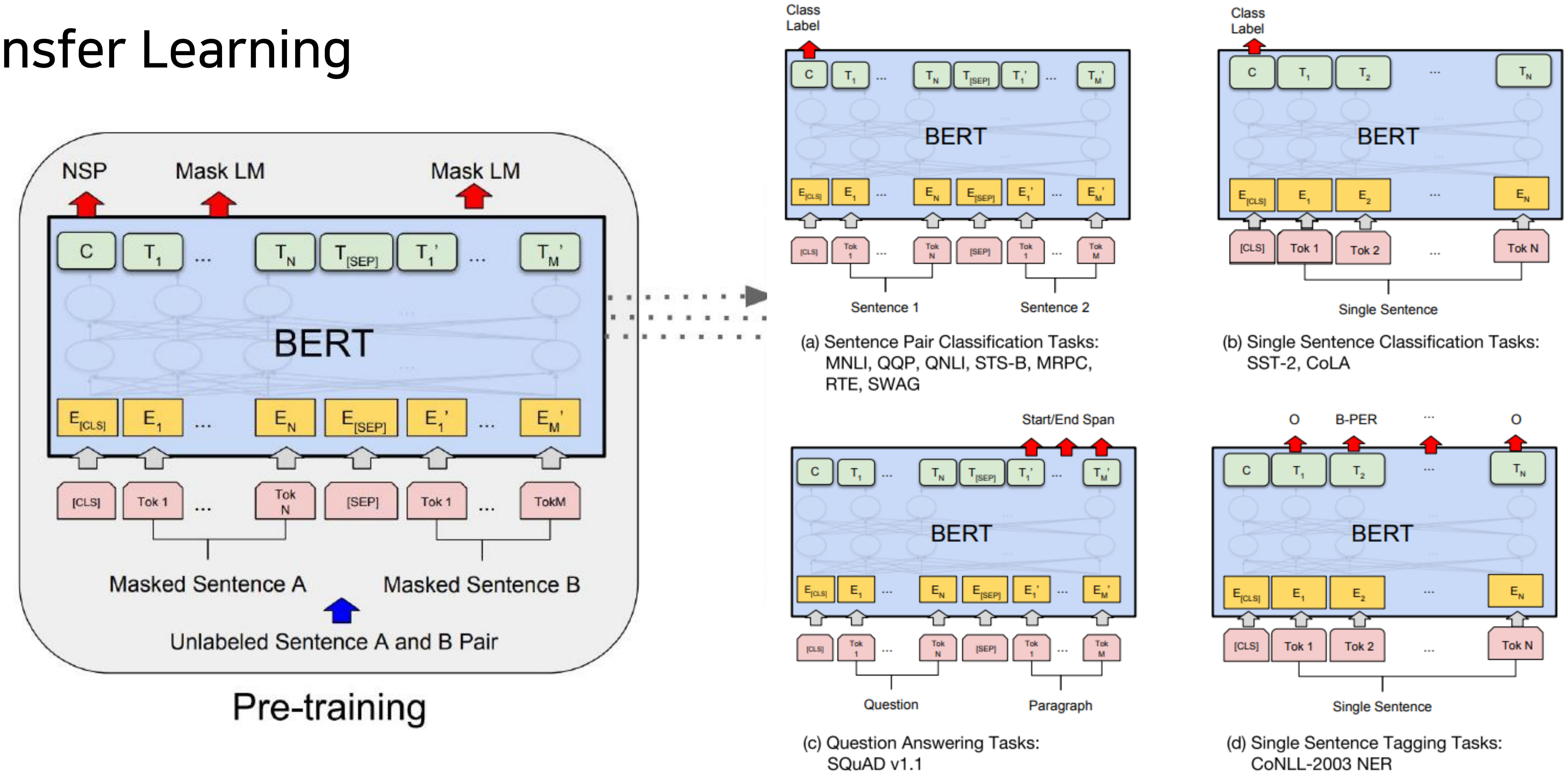
Unidirectional

Bi-LSTM

BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL'19

BERT: Fine-tuning Process

- Transfer Learning



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

https://blog.openai.com/language-unsupervised/

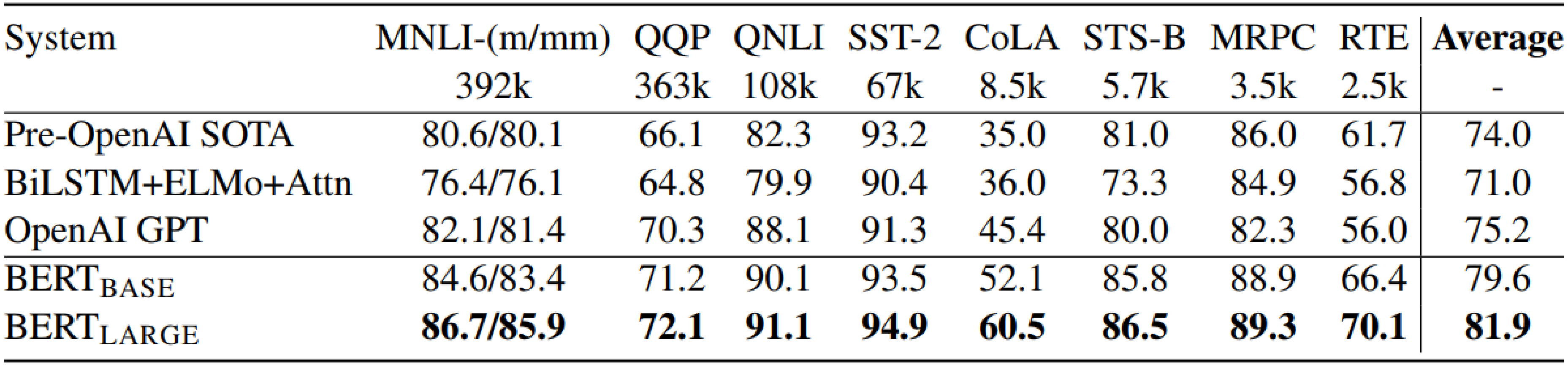
BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL'19

BERT vs GPT-1

- **Comparison of BERT and GPT-1**
  - Training-data size
    - GPT is trained on BookCorpus(800M words) ; BERT is trained on the BookCorpus and Wikipedia (2,500M words)
  - Training special tokens during training
    - BERT learns [SEP],[CLS], and sentence A/B embedding during pre-training
  - Batch size
    - BERT – 128,000 words ; GPT – 32,000 words
  - Task-specific fine-tuning
    - GPT uses the same learning rate of 5e-5 for all fine-tuning experiments; BERT chooses a task-specific fine-tuning learning rate.

BERT: GLUE Benchmark Results

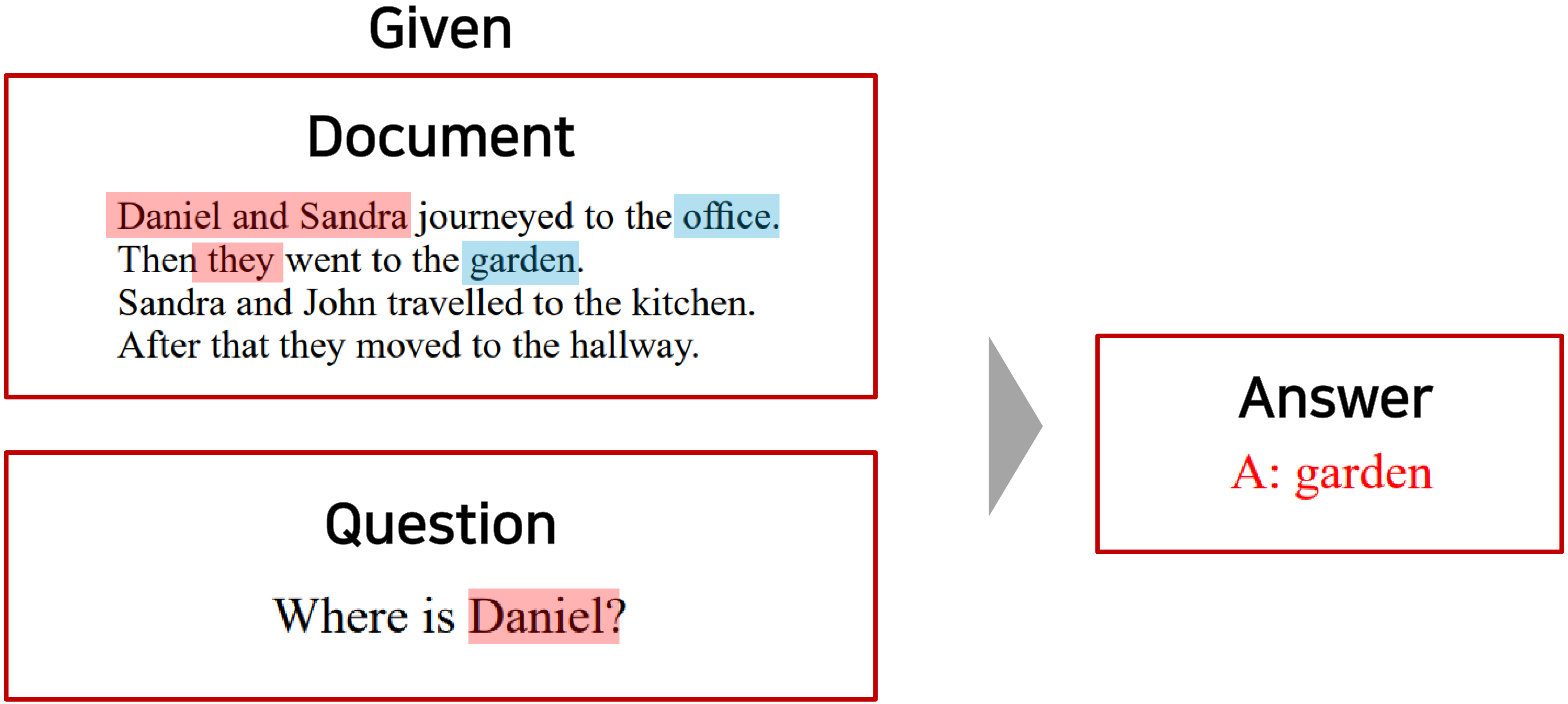
- GLUE Benchmark Results



System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT$_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT$_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL'19

Machine Reading Comprehension (MRC), Question Answering



Given

Document

Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.

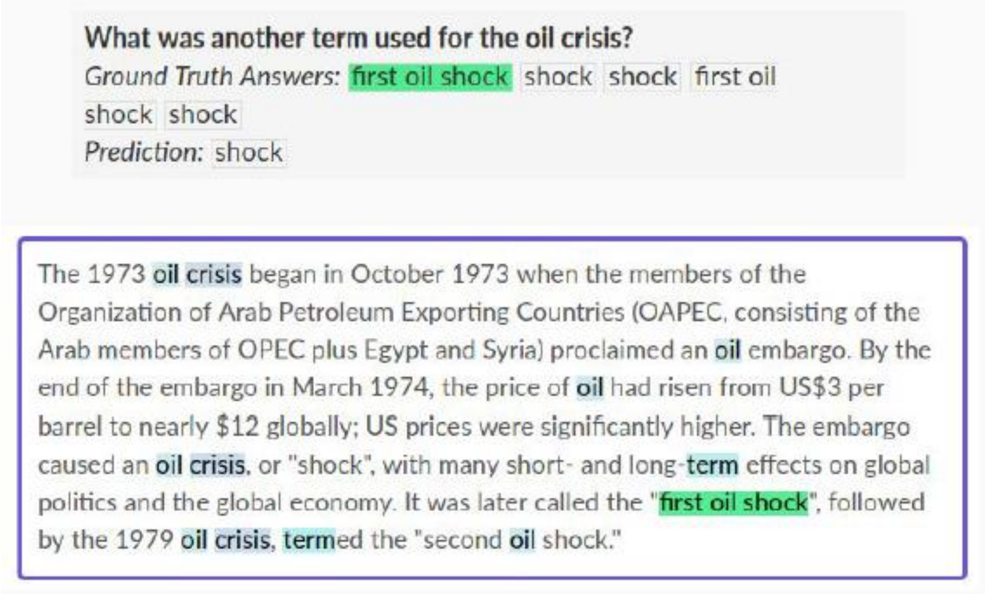
Question

Where is Daniel?

Answer

A: garden

BERT: SQuAD 1.1



What was another term used for the oil crisis?

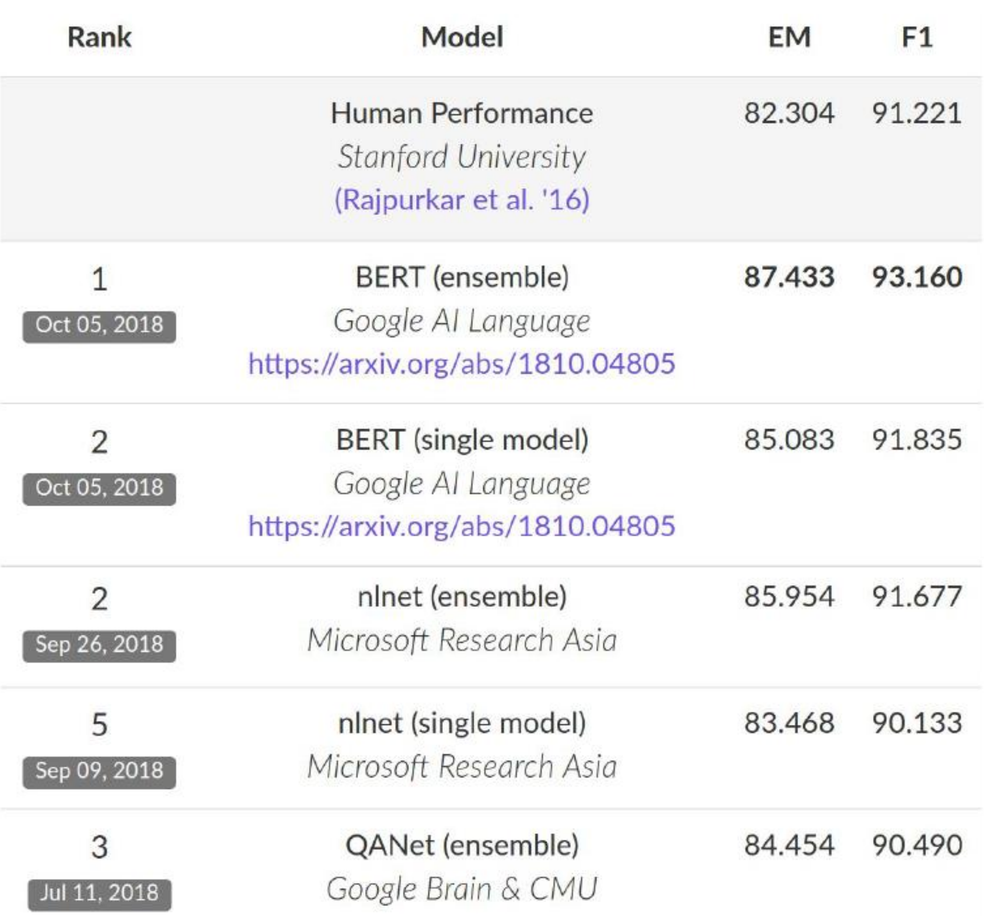
Ground Truth Answers: first oil shock shock shock first oil shock shock

Prediction: shock

The 1973 oil crisis began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC, consisting of the Arab members of OPEC plus Egypt and Syria) proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US$3 per barrel to nearly $12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "first oil shock", followed by the 1979 oil crisis, termed the "second oil shock."

Only new parameters: Start vector and end vector

$$P_i = \frac{e^{S\cdot T_i}}{\sum_j e^{S\cdot T_j}}$$



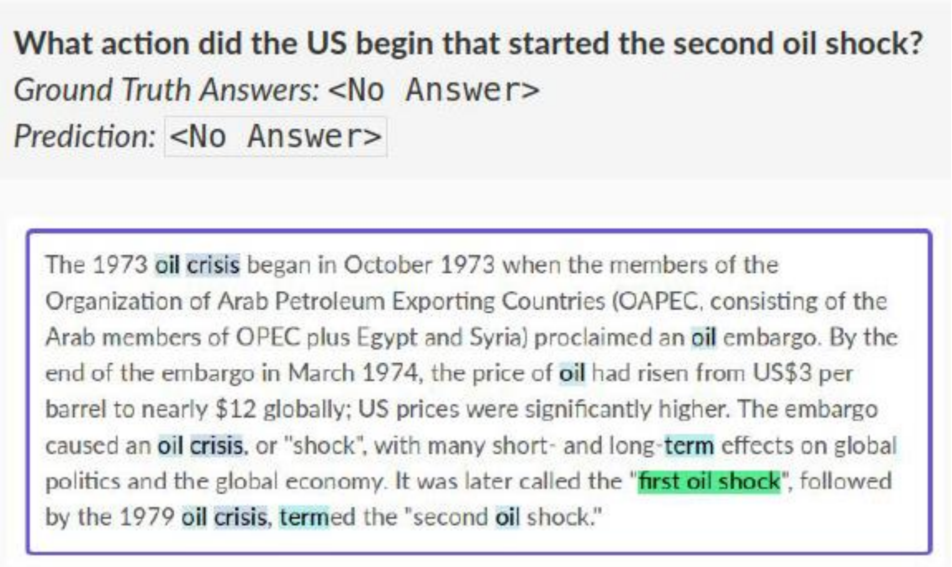
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
5 Sep 09, 2018	nlnet (single model) *Microsoft Research Asia*	83.468	90.133
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

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

BERT: SQuAD 2.0

- Use token 0 ([CLS]) to emit logit for “no answer”
- “No answer” directly competes with answer span
- Threshold is optimized on dev set

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

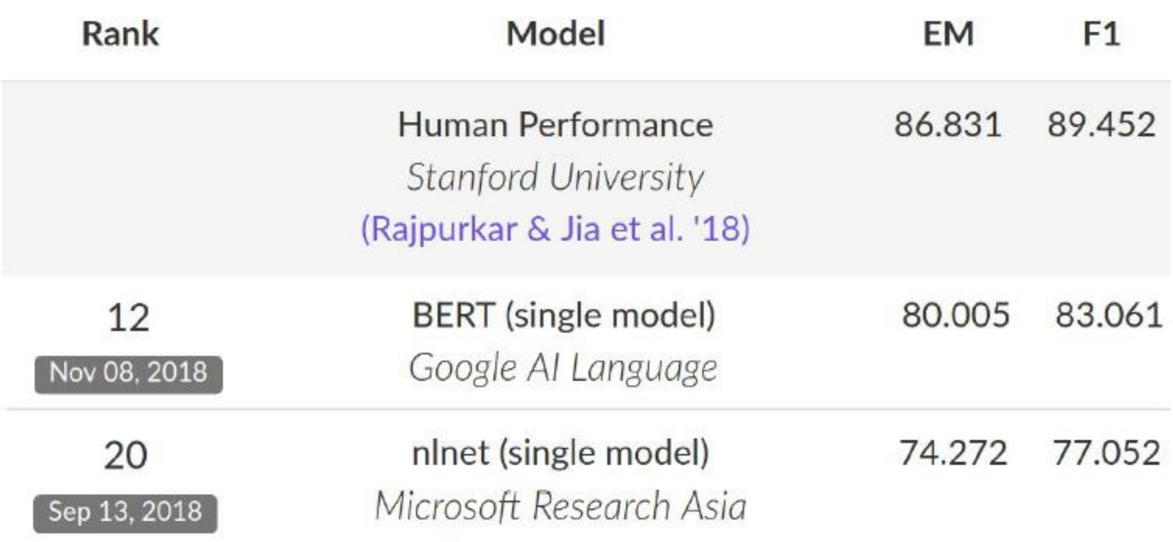


What action did the US begin that started the second oil shock?

Ground Truth Answers: `<No Answer>`

Prediction: `<No Answer>`

The 1973 oil crisis began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC, consisting of the Arab members of OPEC plus Egypt and Syria) proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US$3 per barrel to nearly $12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "first oil shock", followed by the 1979 oil crisis, termed the "second oil shock."



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
12 Nov 08, 2018	BERT (single model) *Google AI Language*	80.005	83.061
20 Sep 13, 2018	nlnet (single model) *Microsoft Research Asia*	74.272	77.052

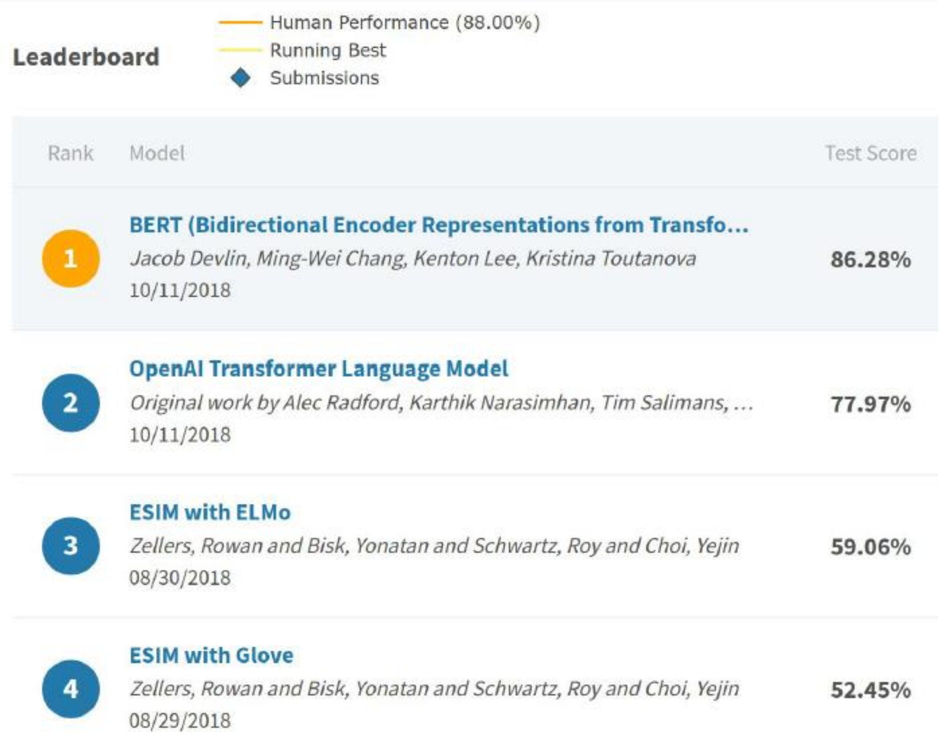
https://rajpurkar.github.io/SQuAD-explorer/

BERT: On SWAG

- Run each Premise + Ending through BERT
- Produce logit for each pair on token 0 ([CLS])

$$P_i = \frac{e^{V \cdot C_i}}{\sum_{j=1}^{4} e^{V \cdot C_j}}$$

A girl is going across a set of monkey bars. She
(i) jumps up across the monkey bars.
(ii) struggles onto the bars to grab her head.
(iii) gets to the end and stands on a wooden plank.
(iv) jumps up and does a back flip.



Leaderboard

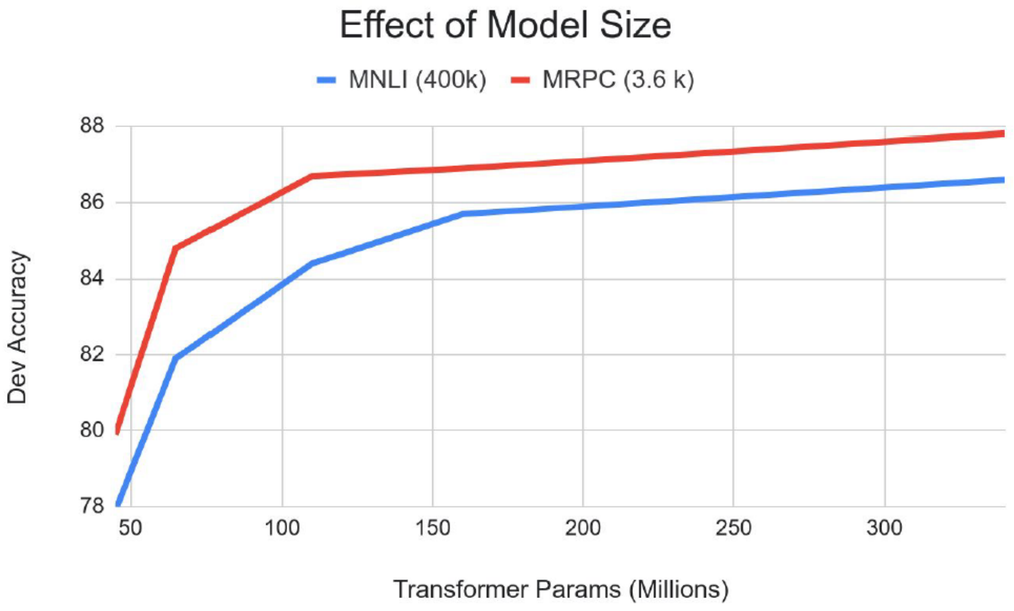
— Human Performance (88.00%)
— Running Best
◆ Submissions

Rank	Model	Test Score
1	**BERT (Bidirectional Encoder Representations from Transfo...** *Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova* 10/11/2018	**86.28%**
2	**OpenAI Transformer Language Model** *Original work by Alec Radford, Karthik Narasimhan, Tim Salimans, ...* 10/11/2018	**77.97%**
3	**ESIM with ELMo** *Zellers, Rowan and Bisk, Yonatan and Schwartz, Roy and Choi, Yejin* 08/30/2018	**59.06%**
4	**ESIM with Glove** *Zellers, Rowan and Bisk, Yonatan and Schwartz, Roy and Choi, Yejin* 08/29/2018	**52.45%**

https://leaderboard.allenai.org/swag/submissions/public

BERT: Ablation Study

- **Big models help a lot**
  - Going from 110M to 340M params helps even on datasets with 3,600 labeled examples
  - Improvements have not asymptoted



BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL’19

2.

Advanced Self-supervised Pre-training Models

GPT-2

GPT-3

ALBERT

ELECTRA

Light-weight Models

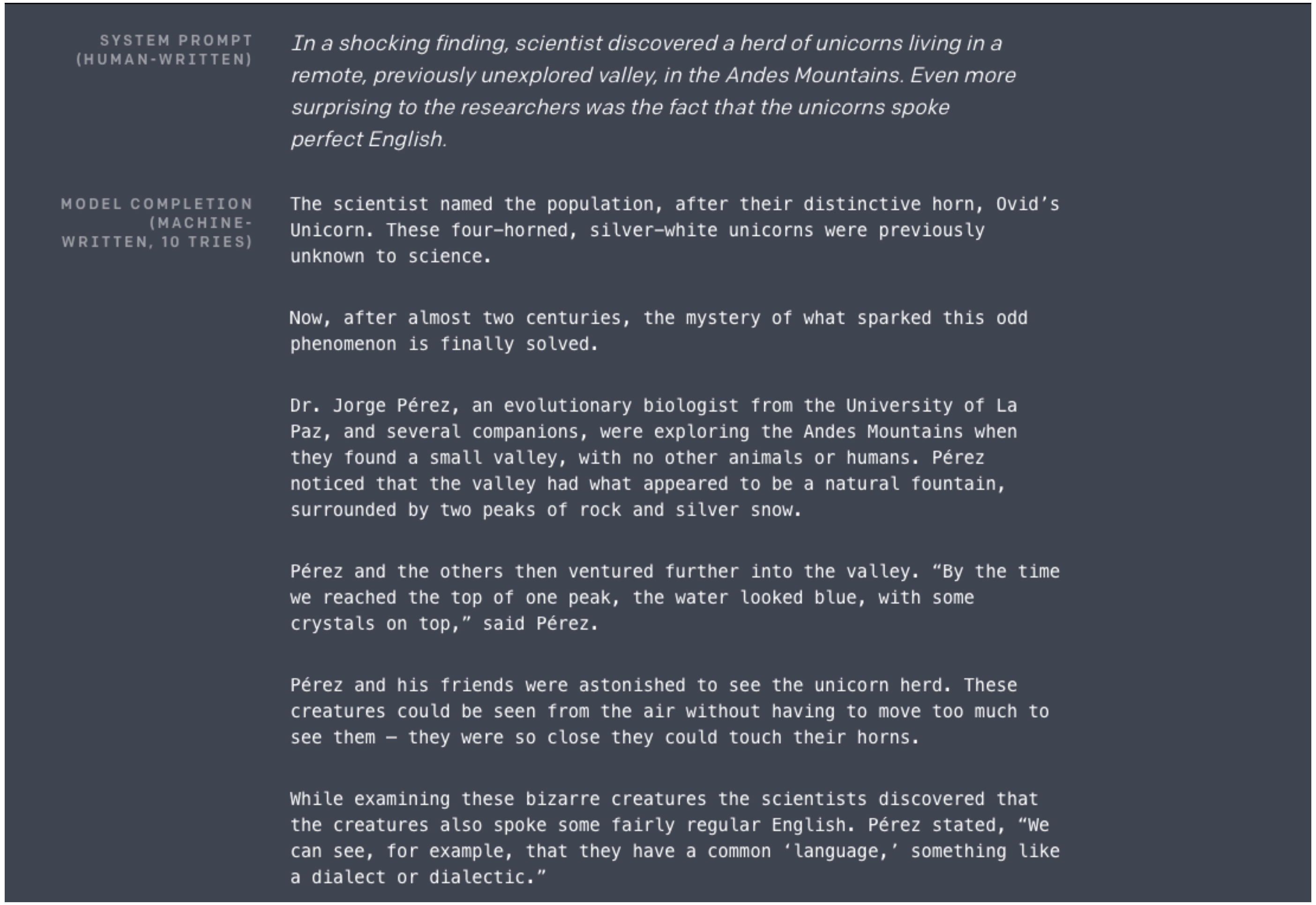
Fusing Knowledge Graph into Language Model

GPT-2

GPT-2: Language Models are Unsupervised Multi-task Learners

- Just a really big transformer LM
- Trained on 40GB of text
  - Quite a bit of effort going into making sure the dataset is good quality
  - Take webpages from reddit links with high karma
- Language model can perform down-stream tasks in a zero-shot setting – without any parameter or architecture modification

GPT-2: Language Models are Unsupervised Multi-task Learners



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

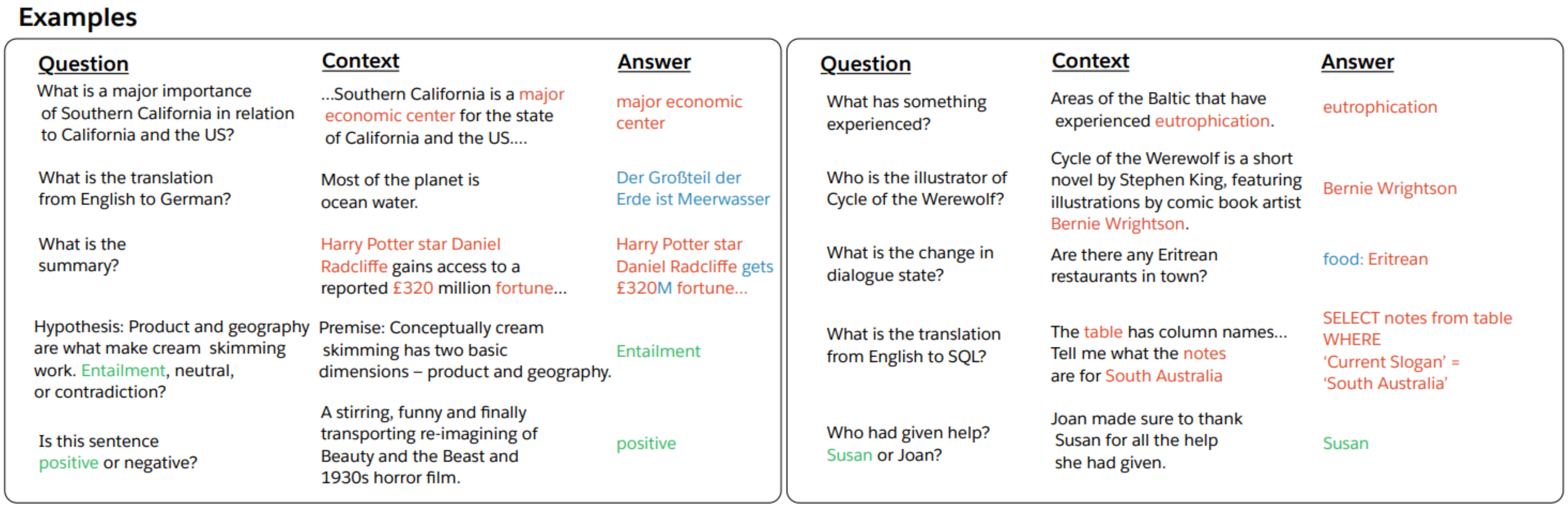
Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

https://blog.floydhub.com/gpt2/

GPT-2: Motivation (decaNLP)

- The Natural Language Decathlon: Multitask Learning as Question Answering
  - Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, Richard Socher



Examples

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

https://decanlp.com/

GPT-2: Datasets

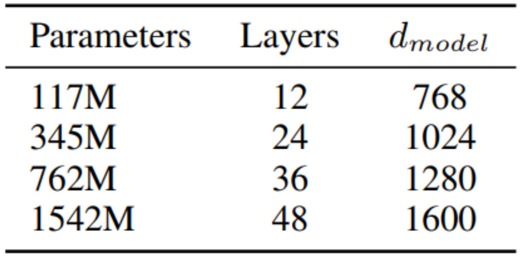
- **A promising source of diverse and nearly unlimited text is web scrape such as common crawl**
    - They scraped all outbound links from Reddit, a social media platform, WebText
        - 45M links
            - Scraped web pages which have been curated/filtered by humans
            - Received at least 3 karma (up-vote)
    - 8M removed Wikipedia documents
    - Use dragnet and newspaper to extract content from links

GPT-2: Datasets

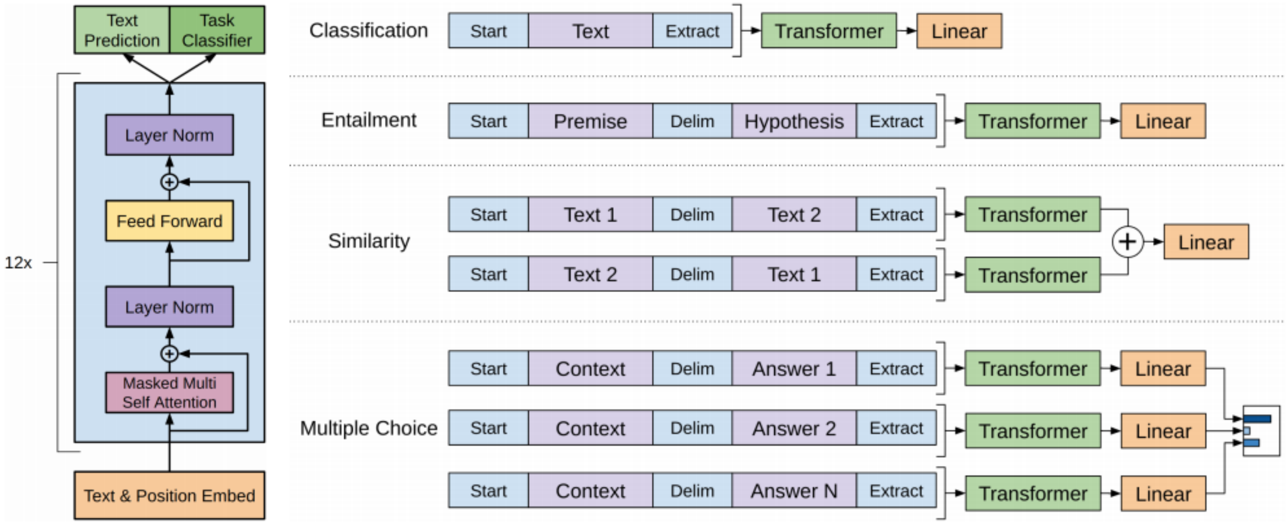
- **Preprocess**
  - Byte pair encoding (BPE)
  - Minimal fragmentation of words across multiple vocab tokens

GPT-2: Model

- **Modification**
  - Layer normalization was moved to the input of each sub-block, similar to a pre-activation residual network
  - Additional layer normalization was added after the final self-attention block.
  - Scaled the weights of residual layer at initialization by a factor of $1/\sqrt{n}$ where n is the number of residual layer



Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600



boostcamp AI Tech

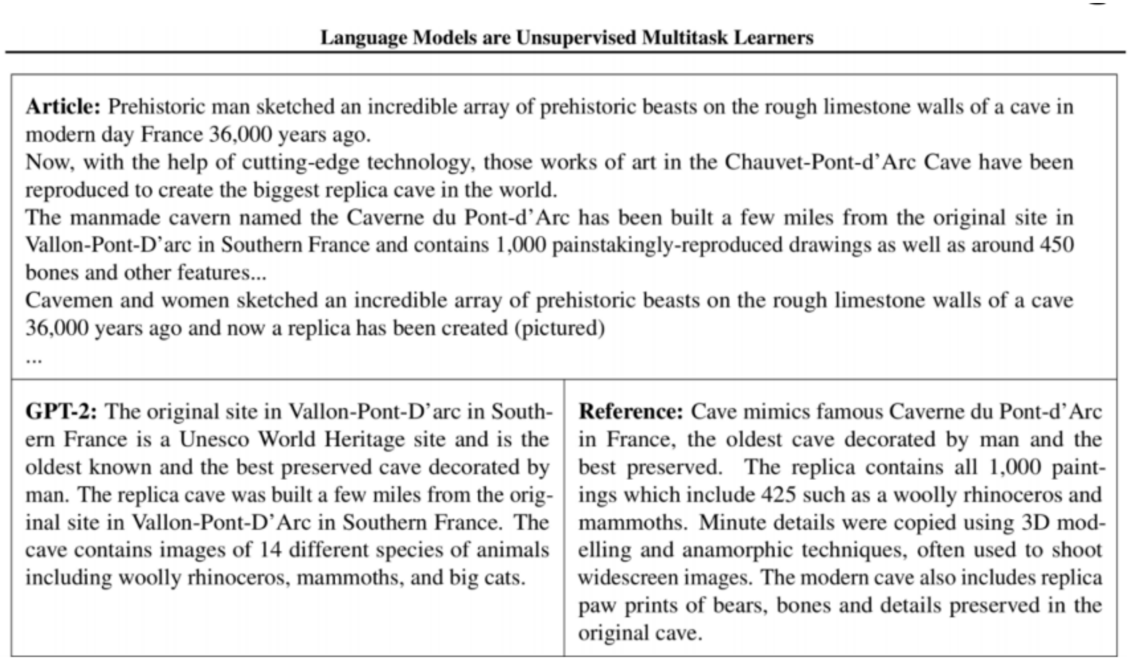
https://openai.com/blog/better-language-models/

GPT-2: Question Answering

- **Use conversation question answering dataset(CoQA)**
  - Achieved 55 F_1 score, exceeding the performance 3 out of 4 baselines without labeled dataset
  - Fine-tuned BERT achieved 89 F_1 performance

GPT-2: Summarization

- **CNN and Daily Mail Dataset**
  - Add text TL;DR: after the article and generate 100 tokens
  - (TL;DR: Too long, didn't read)

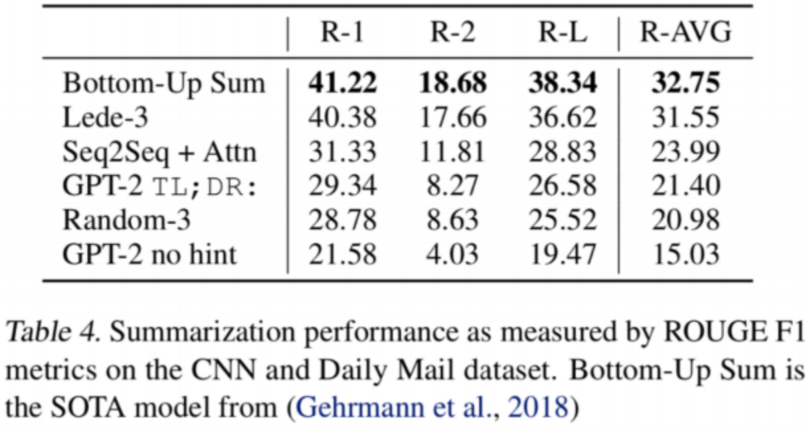


Language Models are Unsupervised Multitask Learners

Article: Prehistoric man sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave in modern day France 36,000 years ago.
Now, with the help of cutting-edge technology, those works of art in the Chauvet-Pont-d'Arc Cave have been reproduced to create the biggest replica cave in the world.
The manmade cavern named the Caverne du Pont-d'Arc has been built a few miles from the original site in Vallon-Pont-D'arc in Southern France and contains 1,000 painstakingly-reproduced drawings as well as around 450 bones and other features...
Cavemen and women sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave 36,000 years ago and now a replica has been created (pictured)
...

GPT-2: The original site in Vallon-Pont-D'arc in Southern France is a Unesco World Heritage site and is the oldest known and the best preserved cave decorated by man. The replica cave was built a few miles from the original site in Vallon-Pont-D'Arc in Southern France. The cave contains images of 14 different species of animals including woolly rhinoceros, mammoths, and big cats.

Reference: Cave mimics famous Caverne du Pont-d'Arc in France, the oldest cave decorated by man and the best preserved. The replica contains all 1,000 paintings which include 425 such as a woolly rhinoceros and mammoths. Minute details were copied using 3D modelling and anamorphic techniques, often used to shoot widescreen images. The modern cave also includes replica paw prints of bears, bones and details preserved in the original cave.



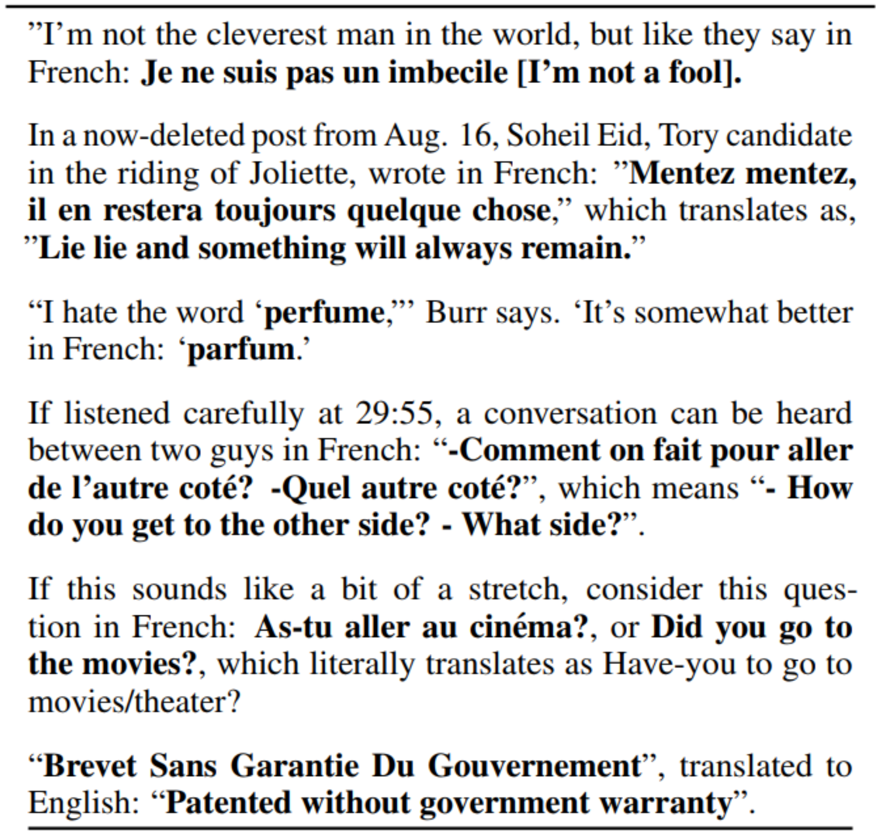
	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 `TL;DR:`	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

GPT-2: Translation

- **User WMT14 en-fr dataset for evaluation**
  - Use LMs on a context of example pairs of the format:
    - English sentence = French sentence
  - Achieve 5 BLEU score in word-by-word substitution
    - Slightly worse than MUSE (Conneau et al., 2017)



”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

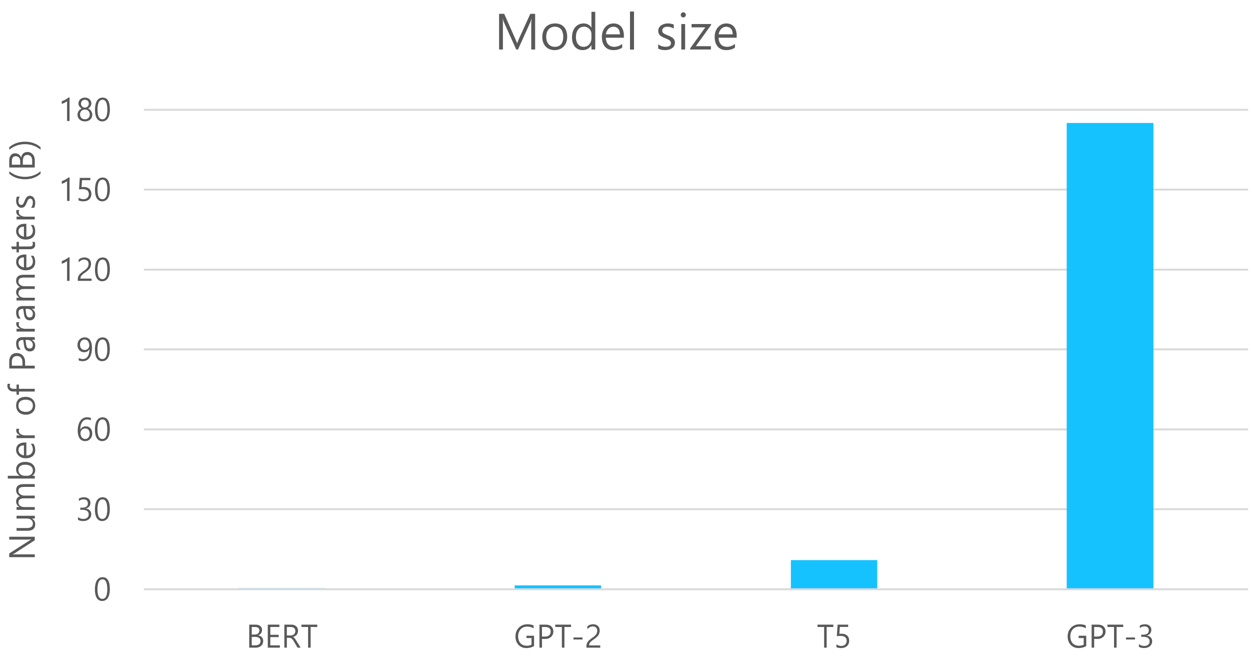
https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

GPT-3

GPT-3: Language Models are Few-Shot Learners

Language Models are Few-shot Learners

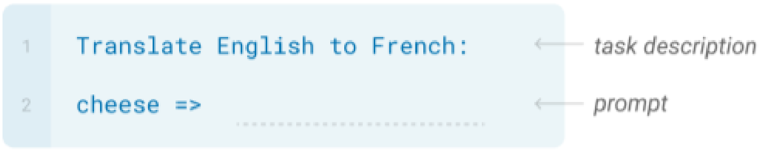
- Scaling up language models greatly improves task-agnostic, few-shot performance
- An autoregressive language model with 175 billion parameters in the few-shot setting
- 96 Attention layers, Batch size of 3.2M



GPT-3: Language Models are Few-Shot Learners

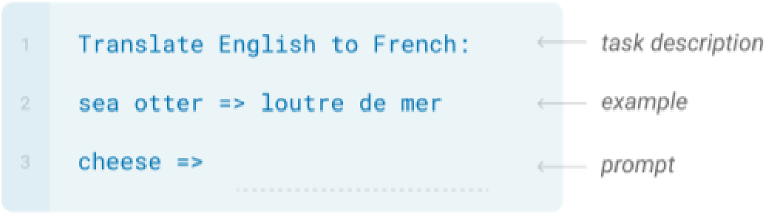
Language Models are Few-shot Learners

- Prompt: the prefix given to the model
- Zero-shot: Predict the answer given only a natural language description of the task
- One-shot: See a single example of the task in addition to the task description
- Few-shot: See a few examples of the task



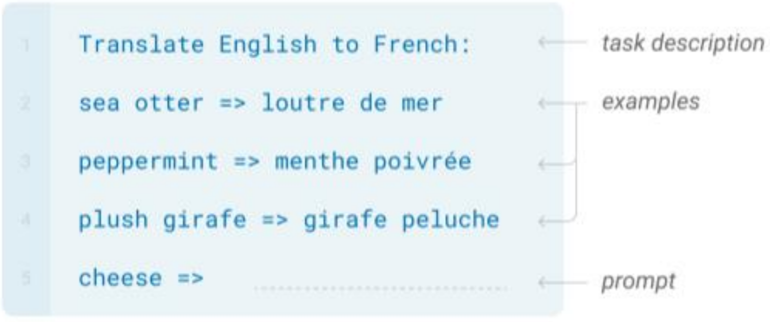
```
Translate English to French:
cheese =>
```

Zero-shot



```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

One-shot



```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

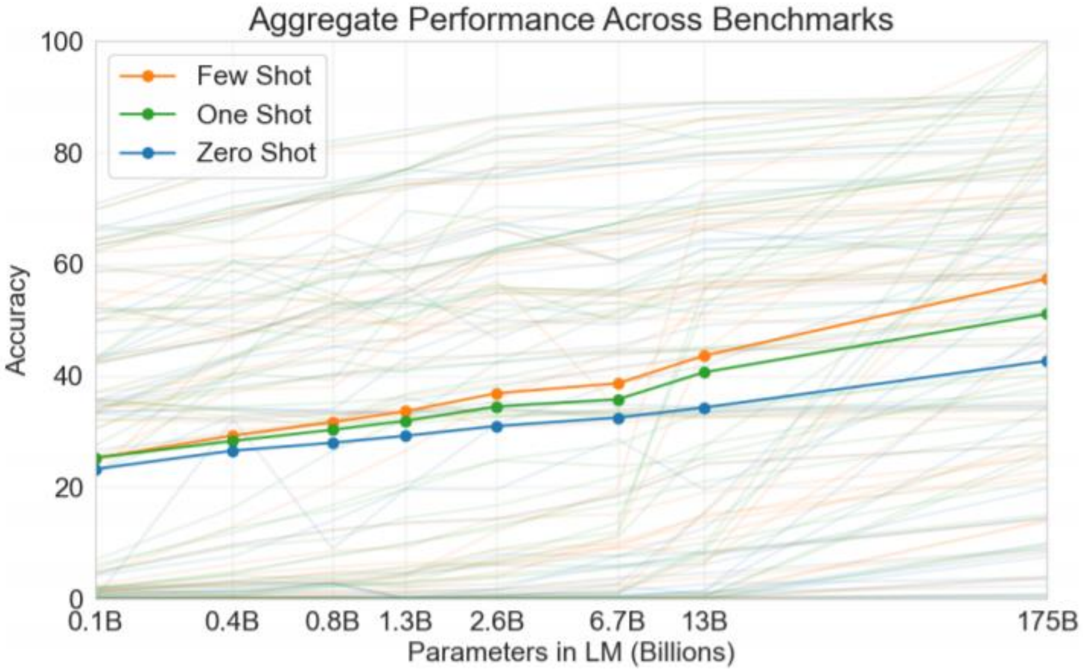
Few-shot

Language Models are Few-show Learners, NeurIPS'20

GPT-3: Language Models are Few-Shot Learners

Language Models are Few-shot Learners

- Zero-shot performance improves steadily with model size
- Few-shot performance increases more rapidly



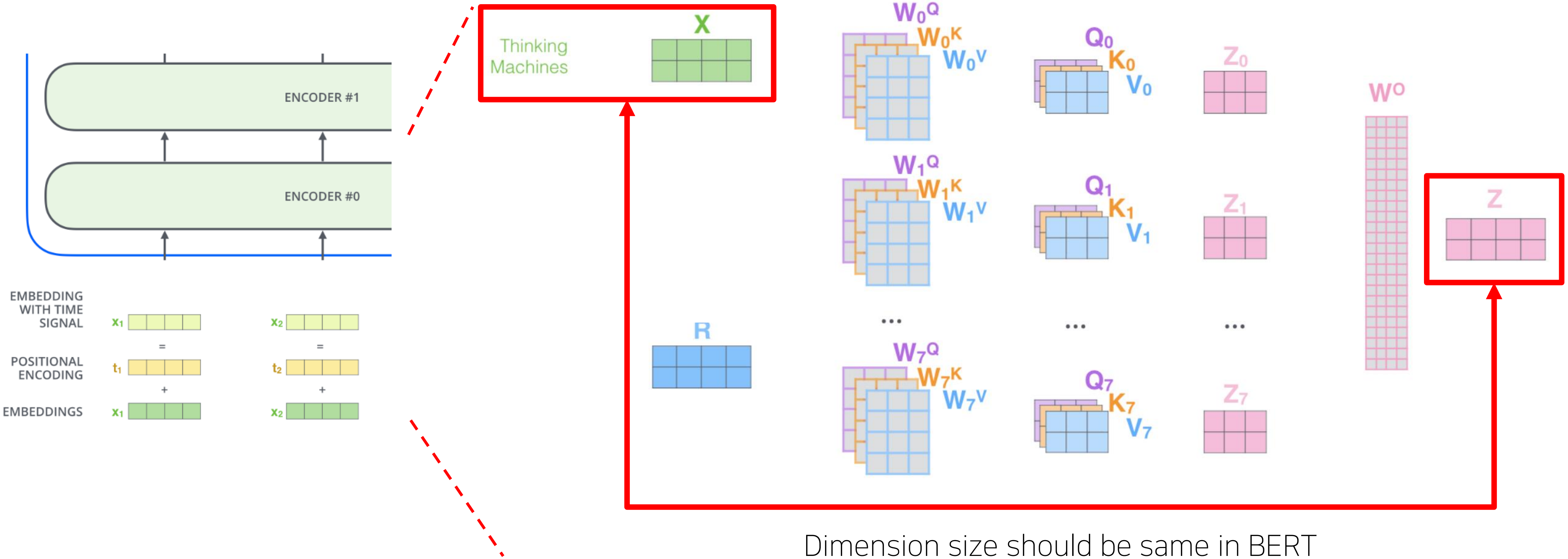
Language Models are Few-show Learners, NeurIPS'20

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

- Is having better NLP models as easy as having larger models?
  - Obstacles
    - Memory Limitation
    - Training Speed
  - Solutions
    - Factorized Embedding Parameterization
    - Cross-layer Parameter Sharing
    - (For Performance) Sentence Order Prediction

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

- **Factorized Embedding Parameterization**

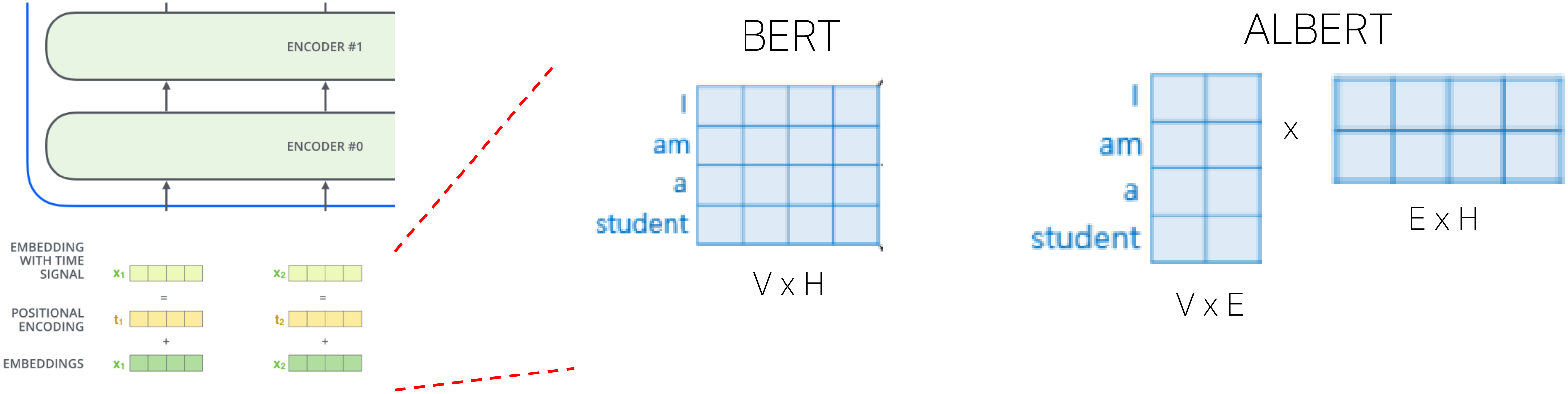


Dimension size should be same in BERT

http://jalammar.github.io/illustrated-transformer/

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

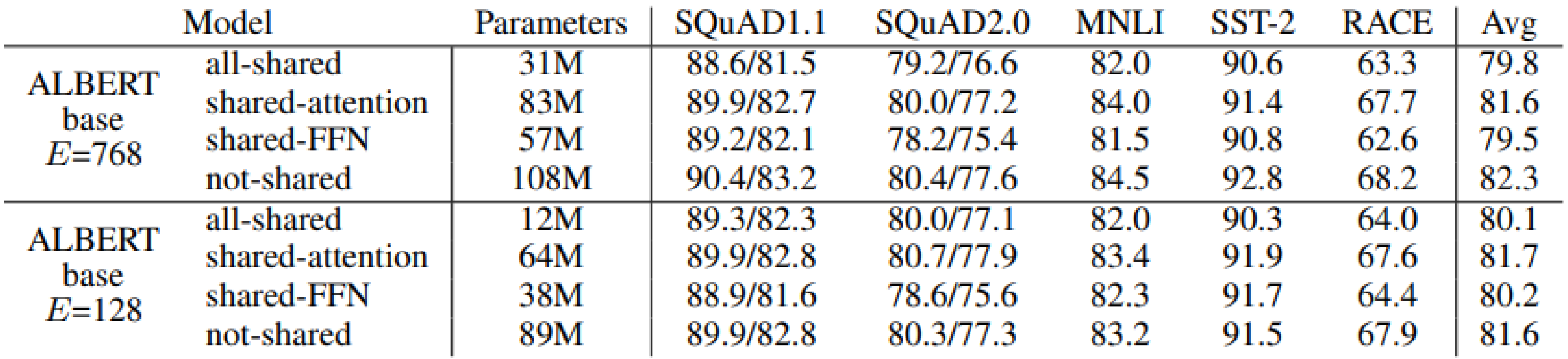
- **Factorized Embedding Parameterization**
  - V = Vocabulary size
  - H = Hidden-state dimension
  - E = Word embedding dimension



http://jalammar.github.io/illustrated-transformer/

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

- **Cross-layer Parameter Sharing**
  - **Shared-FFN**: Only sharing feed-forward network parameters across layers
  - **Shared-attention**: Only sharing attention parameters across layers
  - **All-shared**: Both of them

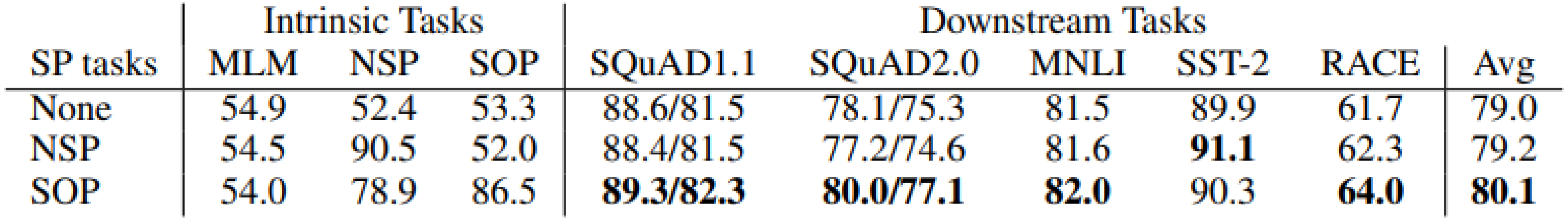


Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
ALBERT base *E*=768	all-shared	31M	88.6/81.5	79.2/76.6	82.0	90.6	63.3	79.8
	shared-attention	83M	89.9/82.7	80.0/77.2	84.0	91.4	67.7	81.6
	shared-FFN	57M	89.2/82.1	78.2/75.4	81.5	90.8	62.6	79.5
	not-shared	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3
ALBERT base *E*=128	all-shared	12M	89.3/82.3	80.0/77.1	82.0	90.3	64.0	80.1
	shared-attention	64M	89.9/82.8	80.7/77.9	83.4	91.9	67.6	81.7
	shared-FFN	38M	88.9/81.6	78.6/75.6	82.3	91.7	64.4	80.2
	not-shared	89M	89.9/82.8	80.3/77.3	83.2	91.5	67.9	81.6

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, ICLR'20

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

- **Sentence Order Prediction**
  - Next Sentence Prediction pretraining task in BERT is too easy
  - Predict the ordering of two consecutive segments of text
    - Negative samples the same two consecutive segments but with their order swapped

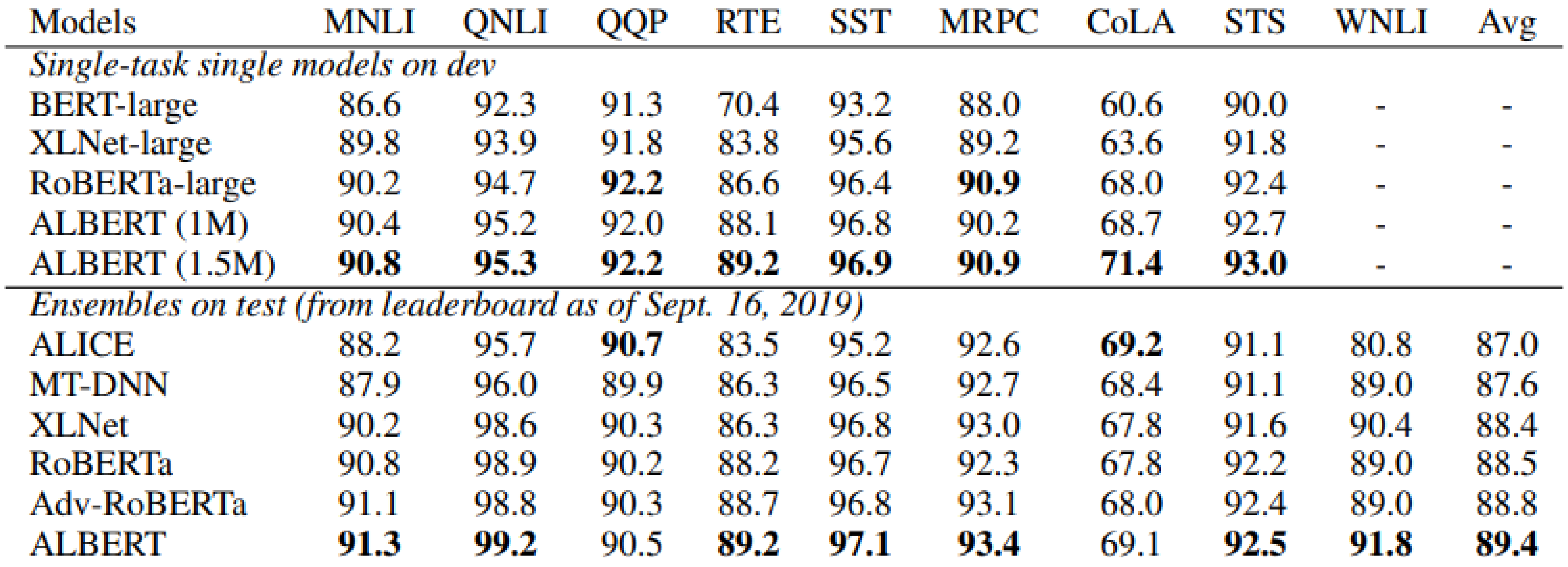


	Intrinsic Tasks			Downstream Tasks					
SP tasks	MLM	NSP	SOP	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
None	54.9	52.4	53.3	88.6/81.5	78.1/75.3	81.5	89.9	61.7	79.0
NSP	54.5	90.5	52.0	88.4/81.5	77.2/74.6	81.6	**91.1**	62.3	79.2
SOP	54.0	78.9	86.5	**89.3/82.3**	**80.0/77.1**	**82.0**	90.3	**64.0**	**80.1**

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, ICLR'20

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

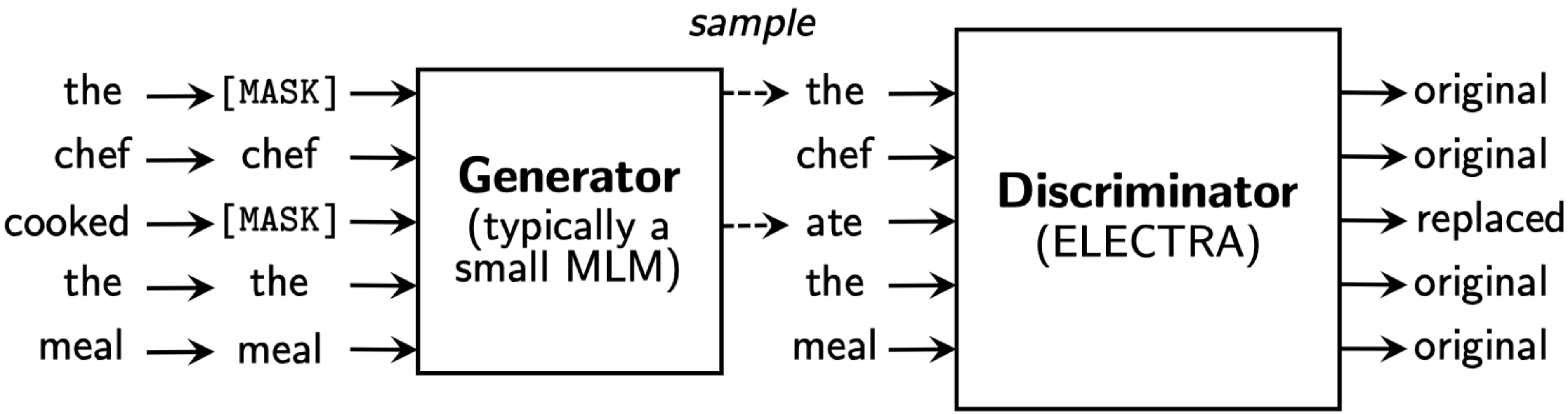
- GLUE Results



Models	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
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**	-	-
Ensembles on test (from leaderboard as of Sept. 16, 2019)										
ALICE	88.2	95.7	**90.7**	83.5	95.2	92.6	**69.2**	91.1	80.8	87.0
MT-DNN	87.9	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2	98.6	90.3	86.3	96.8	93.0	67.8	91.6	90.4	88.4
RoBERTa	90.8	98.9	90.2	88.2	96.7	92.3	67.8	92.2	89.0	88.5
Adv-RoBERTa	91.1	98.8	90.3	88.7	96.8	93.1	68.0	92.4	89.0	88.8
ALBERT	**91.3**	**99.2**	90.5	**89.2**	**97.1**	**93.4**	69.1	**92.5**	**91.8**	**89.4**

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, ICLR'20

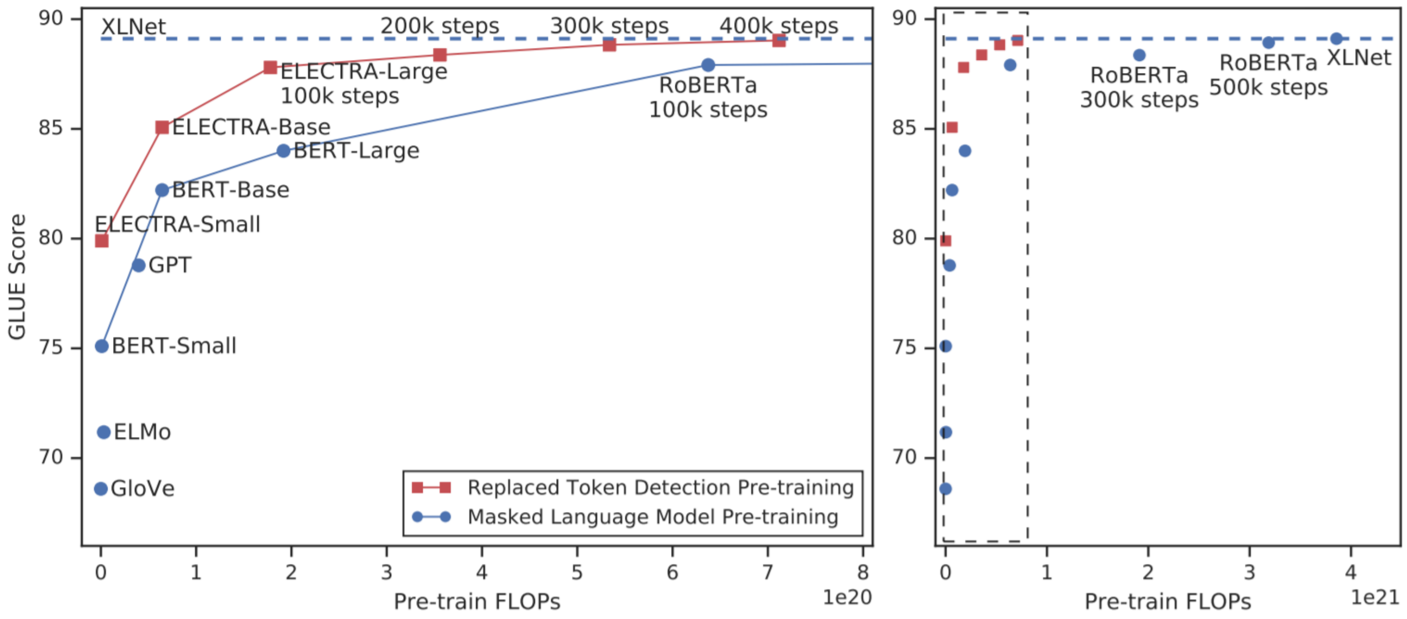
- Efficiently Learning an Encoder that Classifies Token Replacements Accurately
  - Learn to distinguish real input tokens from plausible but synthetically generated replacements
  - Pre-training text encoders as discriminators rather than generators
  - Discriminator is the main networks for pre-training.



ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, ICLR'20

ELECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately

- Replaced token detection pre-training vs masked language model pre-training
  - Outperforms MLM-based methods such as BERT given the same model size, data, and compute



ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, ICLR'20

Light-weight Models

- DistillBERT (NeurIPS 2019 Workshop)
  - A triple loss, which is a distillation loss over the soft target probabilities of the teacher model leveraging the full teacher distribution
- TinyBERT (Findings of EMNLP 2020)
  - Two-stage learning framework, which performs Transformer distillation at both the pre-training and task-specific learning stages

Fusing Knowledge Graph into Language Model

- **ERNIE: Enhanced Language Representation with Informative Entities (ACL 2019)**
  - Informative entities in a knowledge graph enhance language representation
  - Information fusion layer takes the concatenation of the token embedding and entity embedding
- **KagNET: Knowledge-Aware Graph Networks for Commonsense Reasoning (EMNLP 2019)**
  - A knowledge-aware reasoning framework for learning to answer commonsense questions
  - For each pair of question and answer candidate, it retrieves a sub-graph from an external knowledge graph to capture relevant knowledge

References

- GPT-1
  - https://blog.openai.com/language-unsupervised/
- BERT : Pre-training of deep bidirectional transformers for language understanding, NAACL'19
  - https://arxiv.org/abs/1810.04805
- SQuAD: Stanford Question Answering Dataset
  - https://rajpurkar.github.io/SQuAD-explorer/
- SWAG: A Large-scale Adversarial Dataset for Grounded Commonsense Inference
  - https://leaderboard.allenai.org/swag/submissions/public
- How to Build OpenAI's GPT-2: " The AI That Was Too Dangerous to Release"
  - https://blog.floydhub.com/gpt2/

References

- GPT-2
  - https://openai.com/blog/better-language-models/
  - https://cdn.openai.com/better-language models/language_models_are_unsupervised_multitask_learners.pdf
- Language Models are Few-shot Learners, NeurIPS'20
  - https://arxiv.org/abs/2005.14165
- Illustrated Transformer
  - http://jalammar.github.io/illustrated-transformer/
- ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, ICLR'20
  - https://arxiv.org/abs/1909.11942
- ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, ICLR'20
  - https://arxiv.org/abs/2003.10555

References

- DistillBERT, a distilled version of BERT: smaller, faster, cheaper and lighter
  - https://arxiv.org/abs/1910.01108
- TinyBERT: Distilling BERT for Natural Language Understanding, Findings of EMNLP'20
  - https://arxiv.org/abs/1909.10351
- ERNIE: Enhanced Language Representation with Informative Entities
  - https://arxiv.org/abs/1905.07129
- KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning
  - https://arxiv.org/abs/1909.02151

End of Document
Thank You.