AI504: Programming for Artificial Intelligence

Week 13: BERT & GPT

Edward Choi
Grad School of AI
edwardchoi@kaist.ac.kr

Index

- Word Embedding
  - Review
- Contextualized Word Embedding
  - ELMo
  - GPT
- BERT
- GPT-2 & GPT-3

Word Embedding

What is Word Embedding?

- Express a word as a vector
- 'cat' and 'kitty' are similar words, so they have similar vector representations → short distance
- 'hamburger' is not similar with 'cat' or 'kitty’, so they have different vector representations → far distance

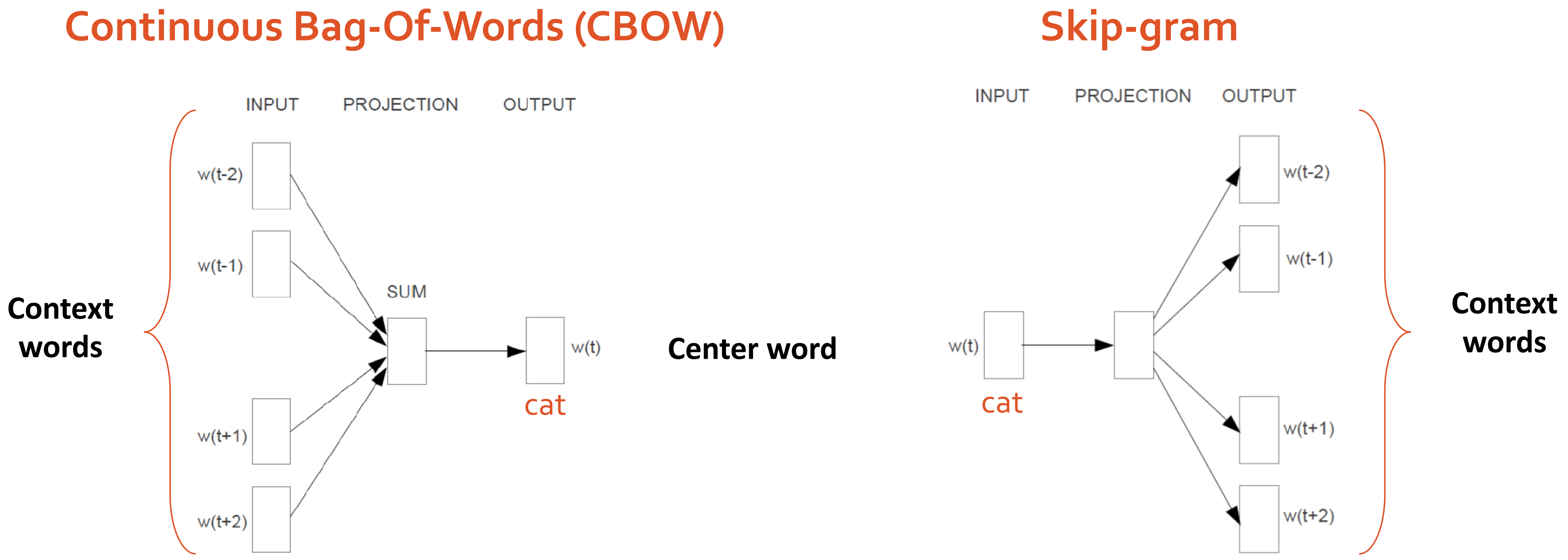
Pre-existing word representation method

- Each word can be represented by a one-hot encoding which each word takes up its respective dimension.
- horse = [0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0]^T
- zebra = [0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0]^T

- Inner-product similarity between two different words (e.g., horse and zebra) is a lways 0.
- Euclidean distance between them is always $\sqrt{2}$

- However, 'horse' and 'zebra' should be semantically similar than 'horse' and 'desk', since they are living creatures and mammals.

Two Models of Word2Vec



Property of Word2Vec – Analogy Reasoning

- More examples: http://wonjaekim.com/archives/50

	데모	http://w.elnn.kr/
버락_오바마-미국+러시아	블라디미르/Noun_푸틴/Noun	-
버락_오바마-미국+스타워즈	아나킨/Noun_스카이워커/Noun	-
아카라카-연세대학교+고려대학교	입실렌티/Noun	입실렌티/Noun
아이폰-휴대폰+노트북	아이패드/Noun	아이패드/Noun
컴퓨터공학-자연과학+인문학	법학/Noun	게임학/Noun
플레이스테이션-소니+마이크로소프트	엑스박스/Noun_360/Number	MSX/Alpha
한국-서울+파리	프랑스/Noun	프랑스/Noun
컴퓨터-기계+인간	운영체제/Noun	일반인/Noun
게임+공부	프로그래밍/Noun	덕질/Noun
박보영-배우+가수	애프터스쿨/Noun	허각/Noun
밥+했는지	끓였/Verb	저녁밥/Noun
사랑+이별	그리움/Noun	추억/Noun
삼성-한화	노트북/Noun	후지필름/Noun
소녀시대-소녀+아줌마	아이유/Noun	에이핑크/Noun
수학-증명	경영학/Noun	이산수학/Noun
스파게티-소시지+김치	칼국수/Noun	비빔국수/Noun
아버지-남자+여자	어머니/Noun	어머니/Noun
아이유-노래+연기	송중기/Noun	송중기/Noun
안드로이드-자유	iOS/Alpha	아이폰/Noun
우주-빛	태양계/Noun_밖/Noun	NASA/Alpha
인간-직업	짐승/Noun	볼뉴르크/Noun
최현석_셰프-허세+셰프	이연/Noun_복/Noun	-
패스트푸드-체인점	영국/Noun_요리/Noun	철물/Noun

Contextualized Word Embedding

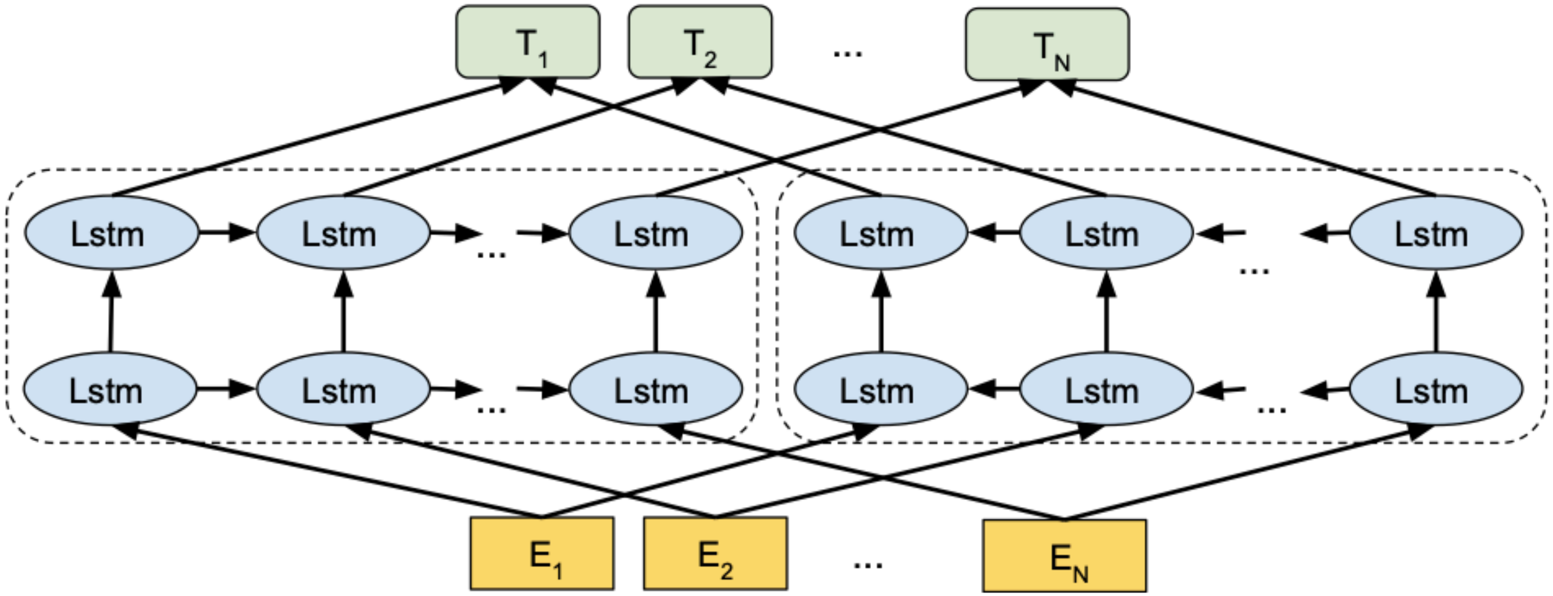
We should consider Context!

Contextualized Word Embedding

- Words have many meanings depending on the **context**!
  - Previous word embeddings represent fixed semantic.
- ELMo
  - "Deep contextualized word representations", Peters et al. 2018 NAACL
  - Word embeddings should be obtained **on the fly**!

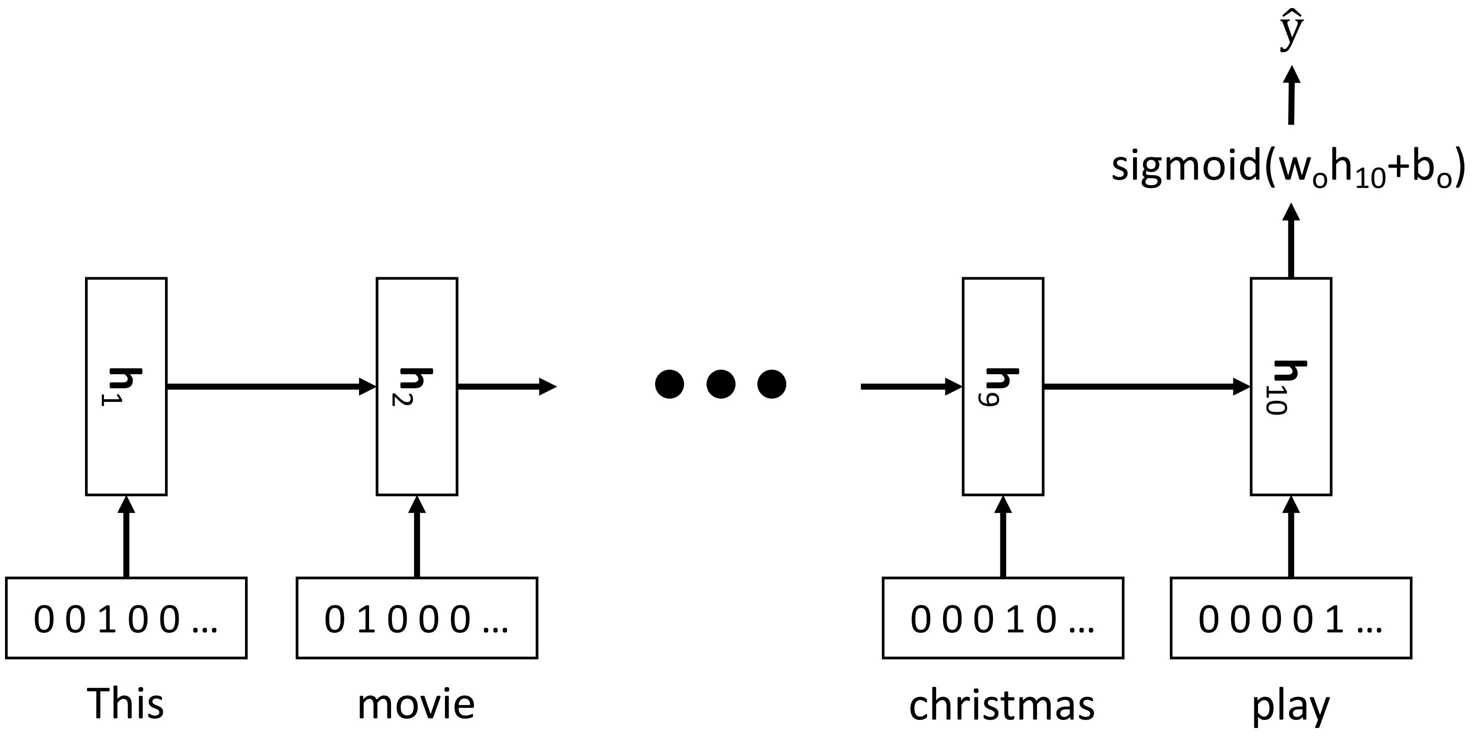
ELMo

- A simple bi-directional multi-layer LSTM
  - Residual connections between layers
- Trained via bi-directional language modeling



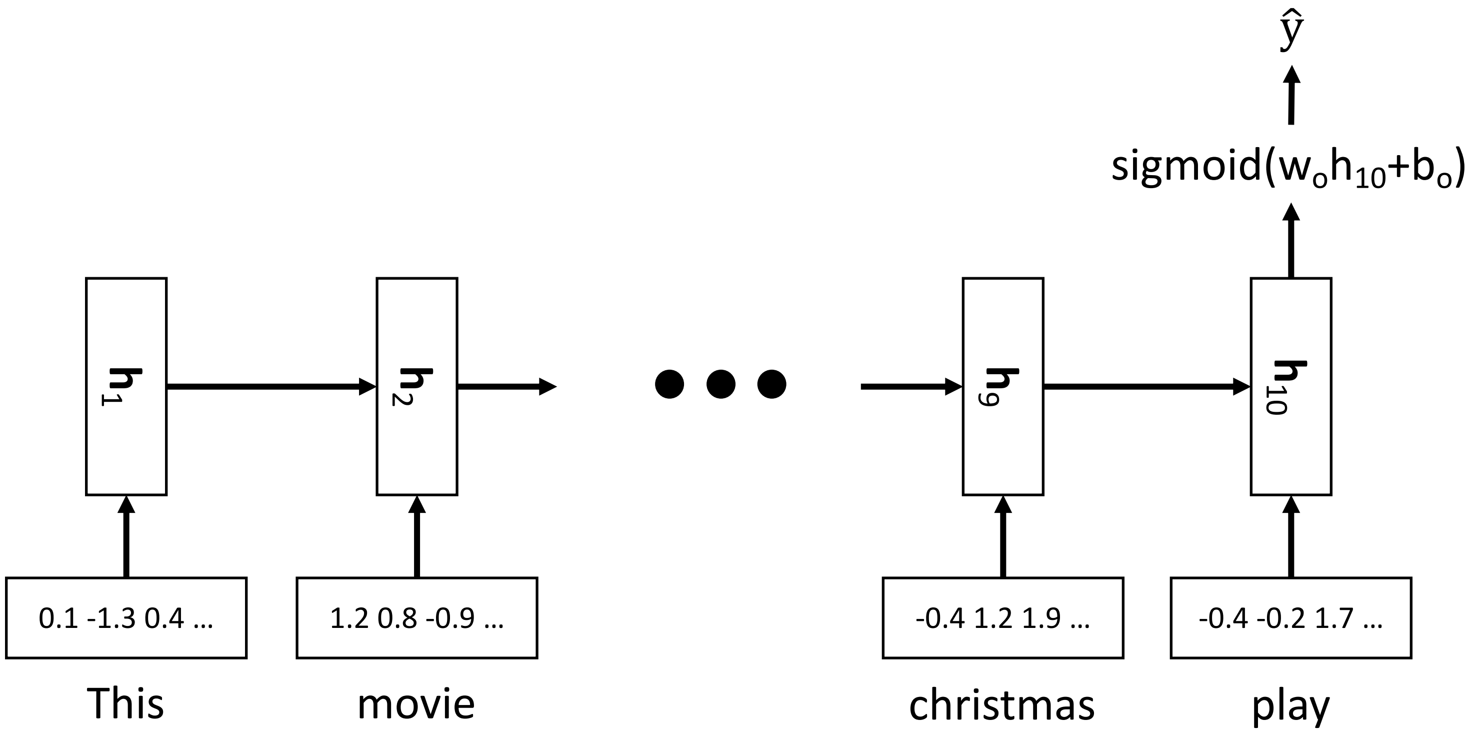
Sentiment Classification

- Sentiment classification with one-hot embedding



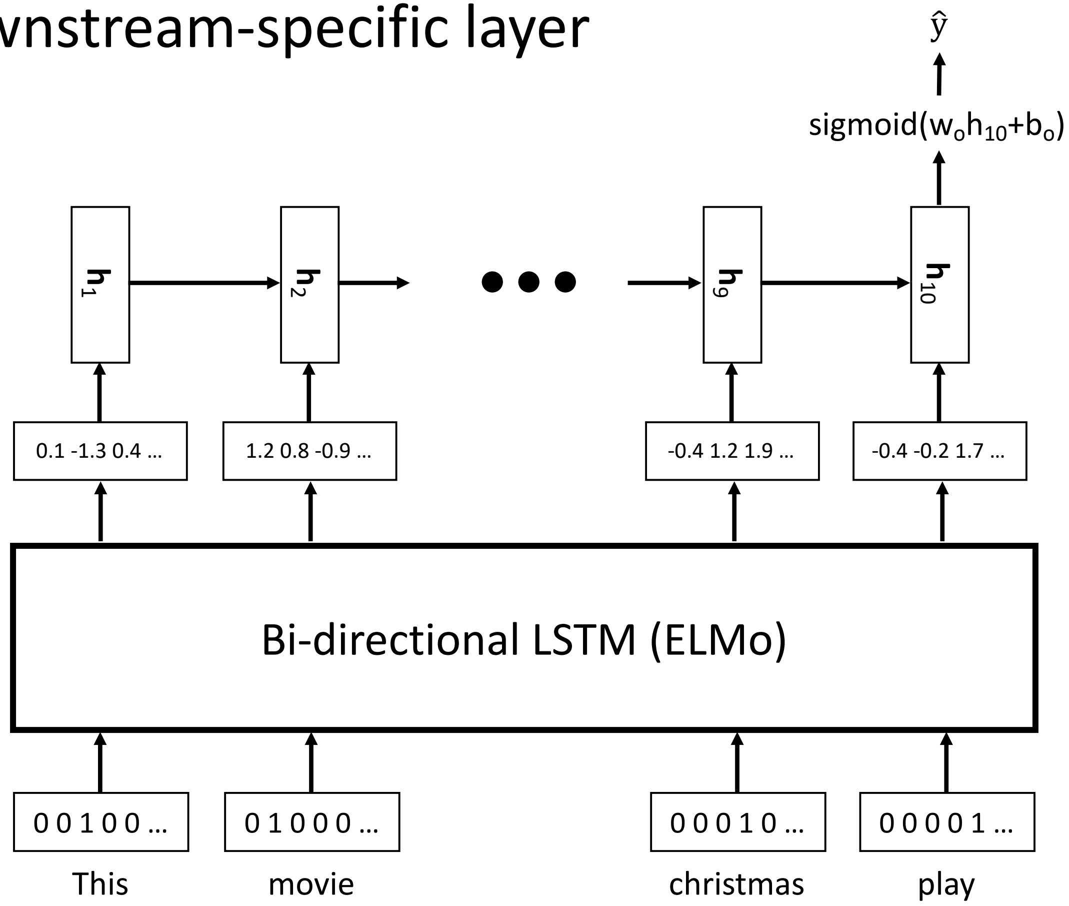
Sentiment Classification

- Sentiment classification with word2vec embedding



Sentiment Classification

- Sentiment classification with ELMo
  - ELMo is a replacement for the embedding layer
  - Still need a downstream-specific layer



ELMo

- Pre-train on 1B Word Dataset (forward-LM perplexity 39.7)
- Simply adding ELMo to the baseline gave SOTA performance
- Fine-tuning for LM using the downstream task dataset.
  - NOT fine-tune via the downstream supervised labels.

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

GPT-1

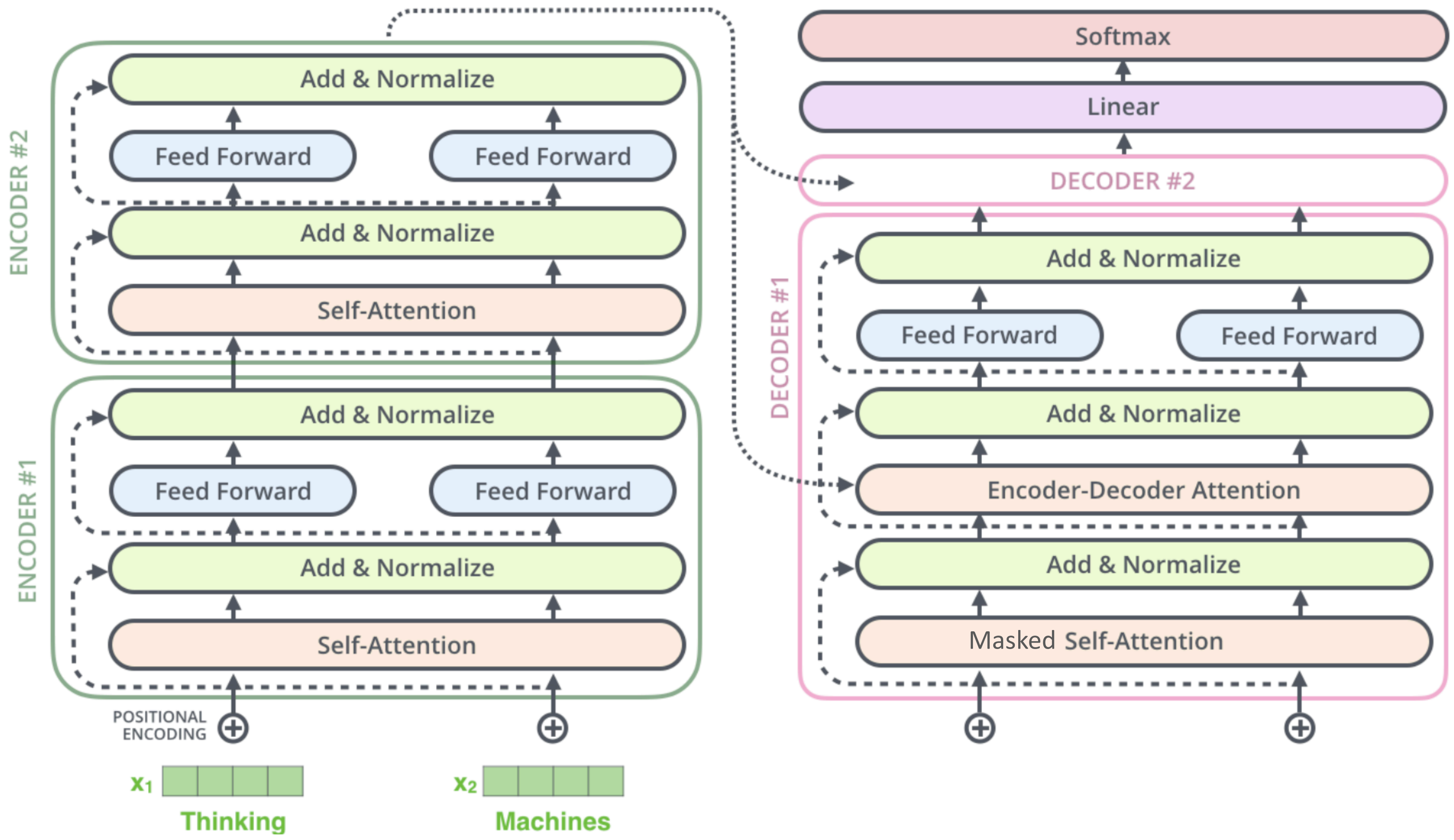
- "Improving Language Understanding by Generative Pre-Training"
  - Technical report from OpenAI, 2018

- Transformer decoder without *Enc-Dec attention*
  - Trained via only forward language modeling

- GPT-1 came out before BERT!

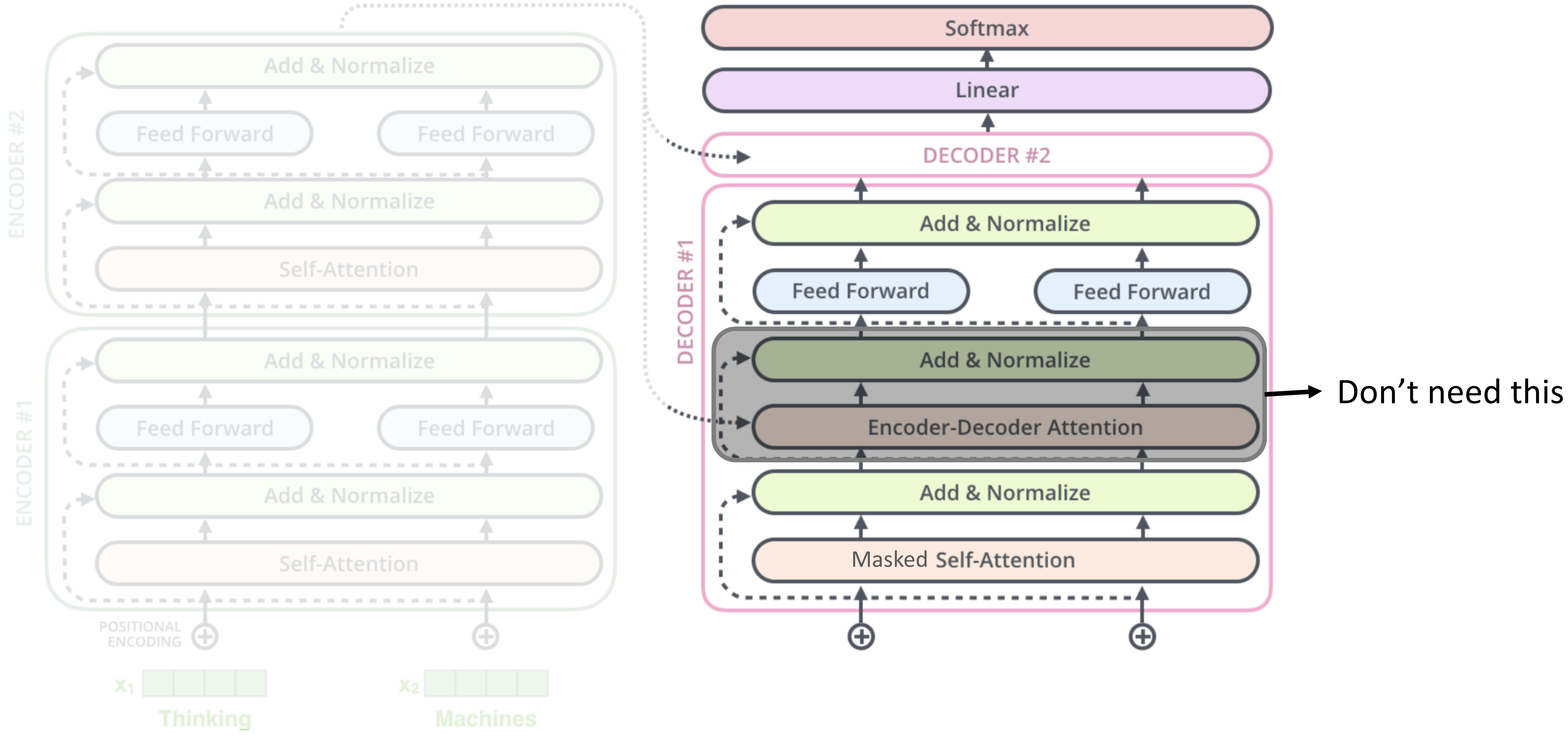
GPT-1

- Transformer architecture



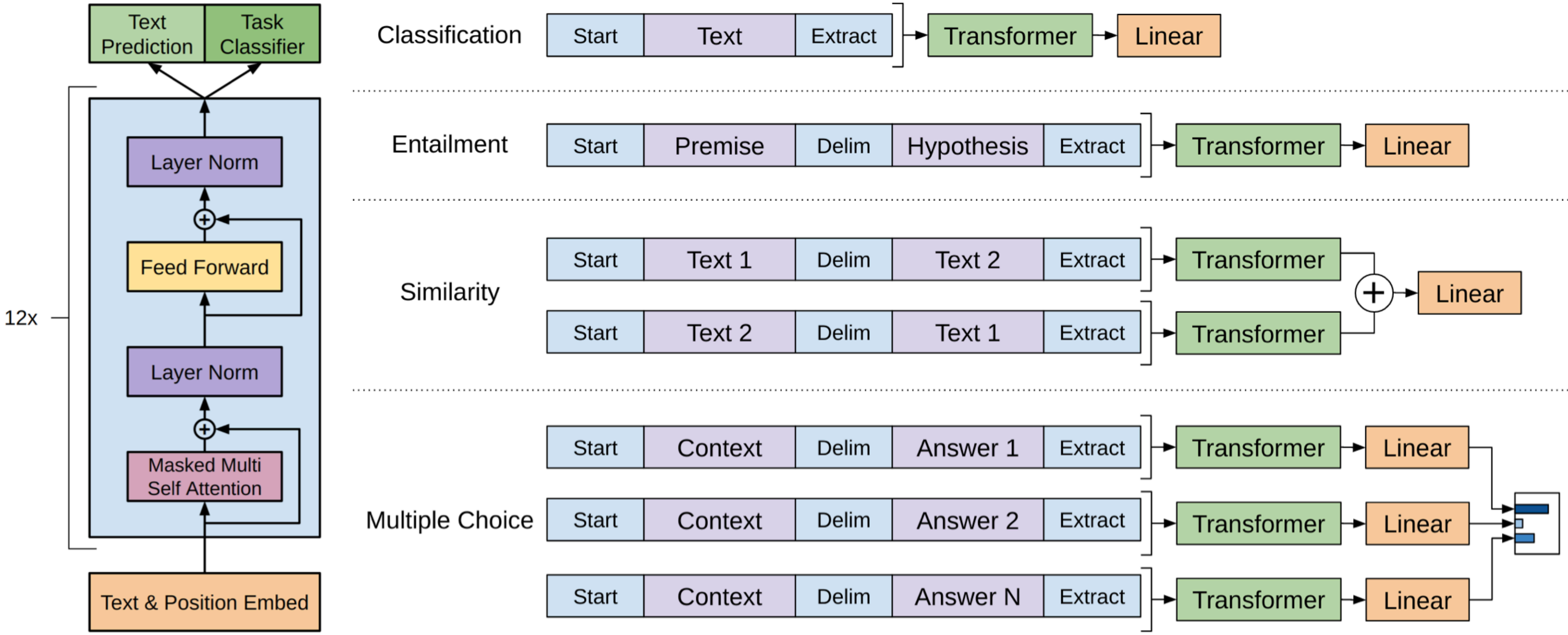
GPT-1

- Transformer architecture



GPT-1 on Downstream Tasks

- Pre-train on BooksCorpus (perplexity 18.4)
  - 12 layers, 768 hidden size, 12 attention heads (110M parameters)
- No downstream specific architecture, just one linear layer.
  - Fine-tuning the entire GPT with the supervised labels.
  - Downstream task inputs are all converted to sequence of tokens.



GPT-1 on Downstream Tasks

- Pre-train on BooksCorpus (perplexity 18.4)
- No downstream specific architecture, just one linear layer.
  - Fine-tuning the entire GPT with the supervised labels.
  - Downstream task inputs are all converted to sequence of tokens.

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

GPT-1 on Downstream Tasks

- Pre-train on BooksCorpus (perplexity 18.4)
- No downstream specific architecture, just one linear layer.
  - Fine-tuning the entire GPT with the supervised labels.
  - Downstream task inputs are all converted to sequence of tokens.

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**

GPT-1 on Downstream Tasks

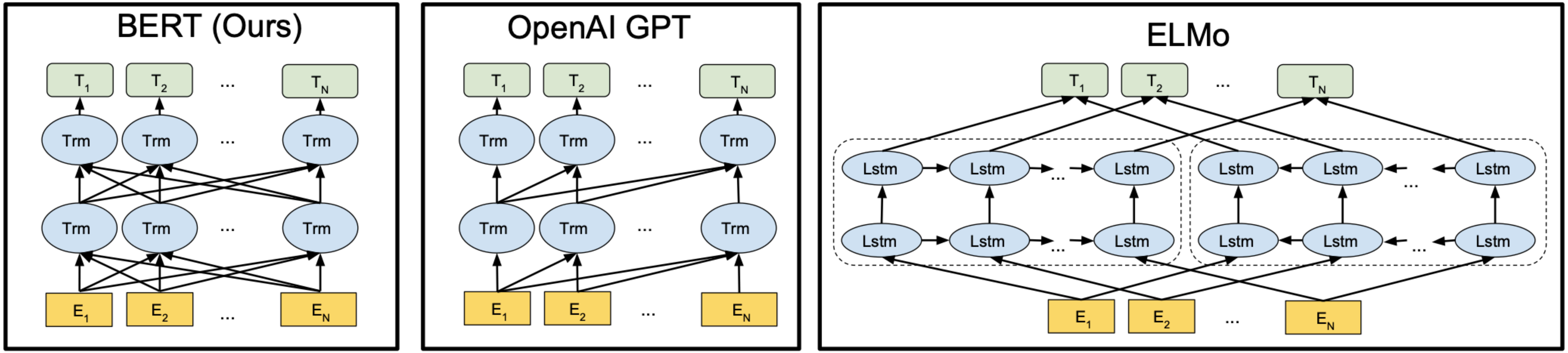
- Pre-train on BooksCorpus (perplexity 18.4)
- No downstream specific architecture, just one linear layer.
  - Fine-tuning the entire GPT with the supervised labels.
  - Downstream task inputs are all converted to sequence of tokens.

Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

BERT

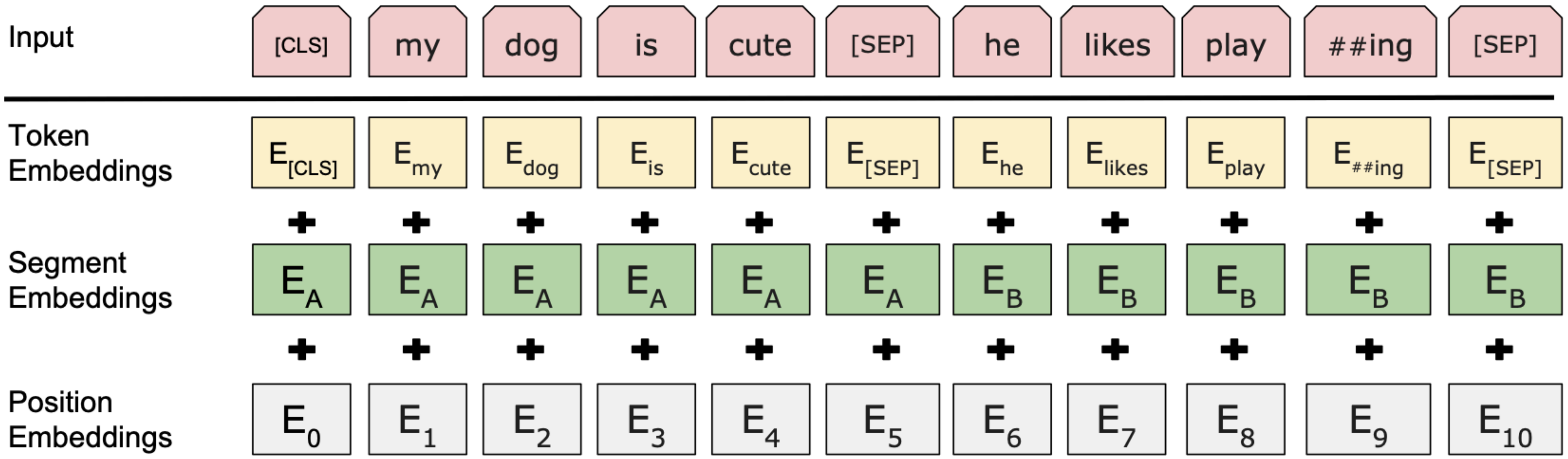
BERT

- Bidirectional Encoder Representations from Transformers
  - Possibly inspired by ELMo & GPT
  - Takes the best of the two: Bi-directionality & Powerful Transformer



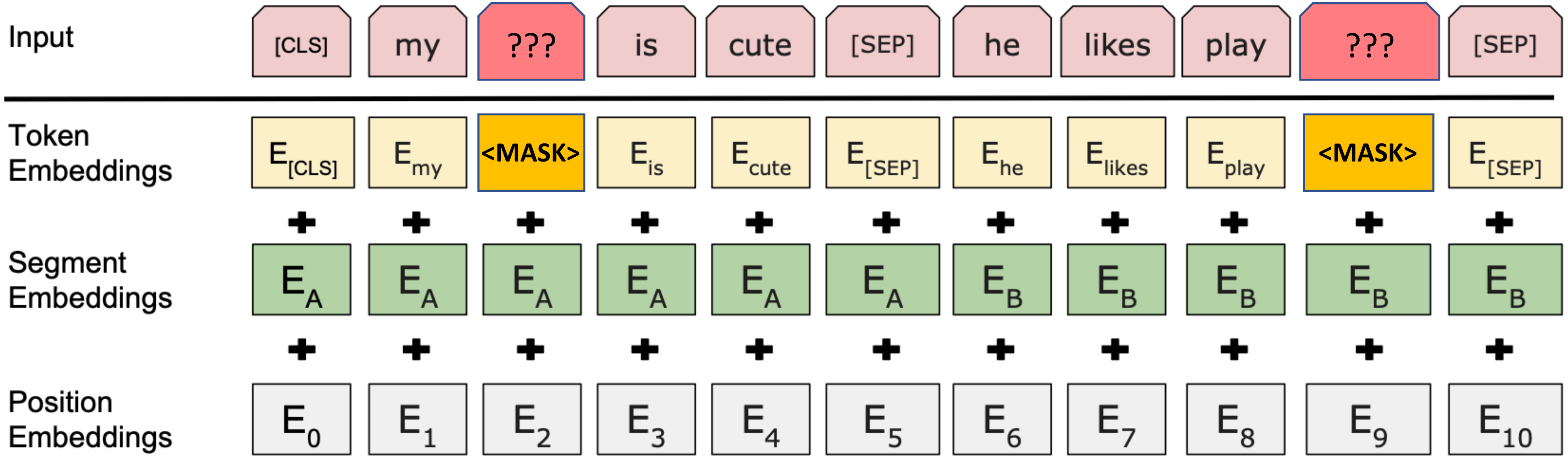
BERT Architecture

- Token Embedding (WordPiece)
  - + Positional Embedding
  - + Trainable Segment Embedding
- CLS: Special tokens representing all input



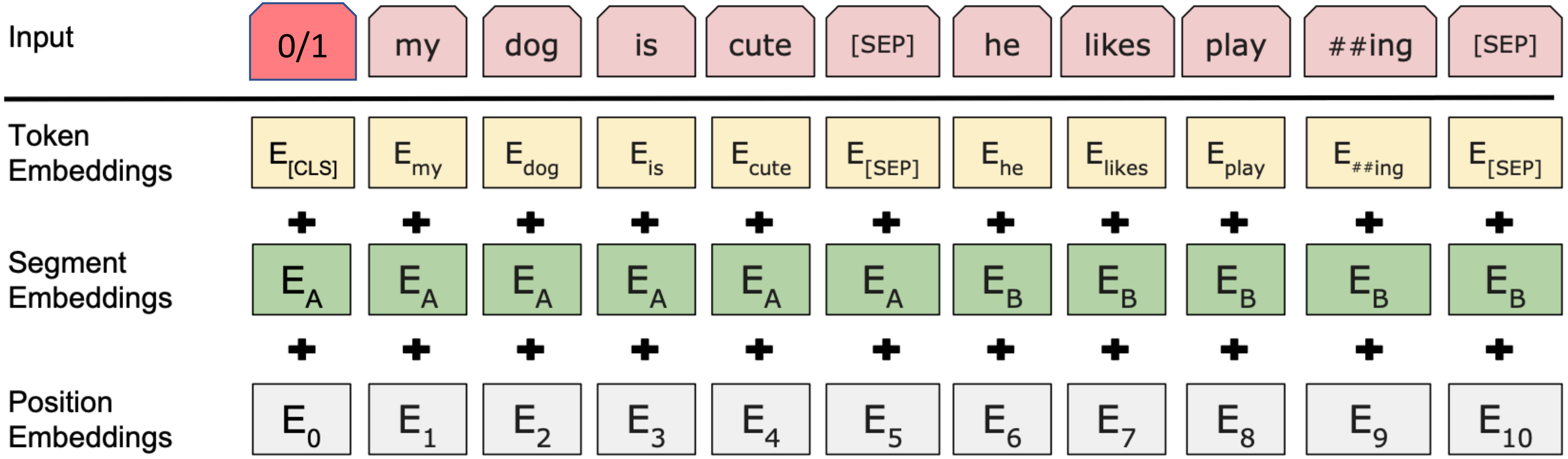
BERT Pre-training

- Masked Language Modeling
  - a.k.a Cloze Task
  - Identify 15% randomly masked tokens



BERT Architecture

- Next sentence prediction
  - Use CLS to perform binary task
  - If two sentences are actual neighbors ➔ 1, otherwise ➔ 0
- Helpful for some downstream tasks



BERT on Downstream Tasks

- Pre-train on BooksCorpus + English Wikipedia
  - Maksed LM perplexity of BERT-Base: ~4.
  - BERT-base: 110M parameters (12 layers, 768 hidden size, 12 attention heads)
- Different input/output for different downstream task (all fine-tuned).
  - Paraphrase detection
    - Input: Sentence A and sentence B
    - Output: CLS + Linear classifier
  - Sequence tagging
    - Input: Text and PAD
    - Output: Token embeddings + Linear classifier
  - Text classification
    - Input: Text and PAD
    - Output: CLS + Linear classifier

BERT on Downstream Tasks

- GLUE benchmark dataset
  - Collection of NLP tasks
  - Text similarity, paraphrase, inference, entailment, etc.

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	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

BERT on Downstream Tasks

- SQuAD 1.1 & 2.0

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

SQuAD 1.1

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

SQuAD 2.0

SQuAD Dataset

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

- Stanford Question Answering Dataset
- Total 98,178 questions
- Answers are segment of text from the passage
- Include non-entities and longer phrase

Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

SQuAD Versions

- 1.0/1.1
  - 100,000 question samples
  - Based on Wikipedia articles
- 2.0
  - Addition of 50,000 unanswerable questions

Article: Endangered Species Act
Paragraph: " . . . *Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition* was raised."

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

SQuAD Dataset

Evaluation method

- Exact match (EM): Percentage of correct answers
  - An answer is correct if it is one of any ground-truth answers

$$\mathrm{EM} = \frac{\#\text{ of correct answers}}{\#\text{ of questions}}$$

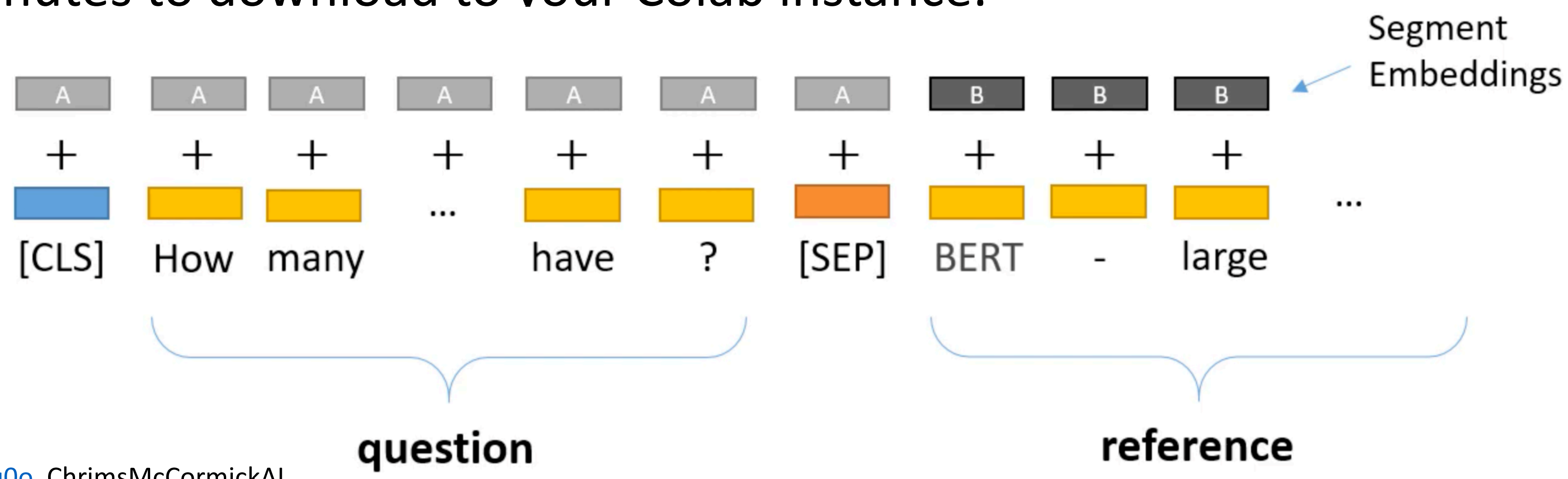
- F1-measure
  - Harmonic mean between precision and recall

$$\mathrm{F1-measure} = \frac{\text{precision} \times \text{recall}}{\frac{1}{2}(\text{precision} + \text{recall})}$$

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/2001.09694v2	90.578	92.978
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
5 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777
6 Jul 05, 2020	electra+nlayers (ensemble) *oppo.tensorlab*	90.126	92.622
7 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694v2	90.115	92.580
8 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
9 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215

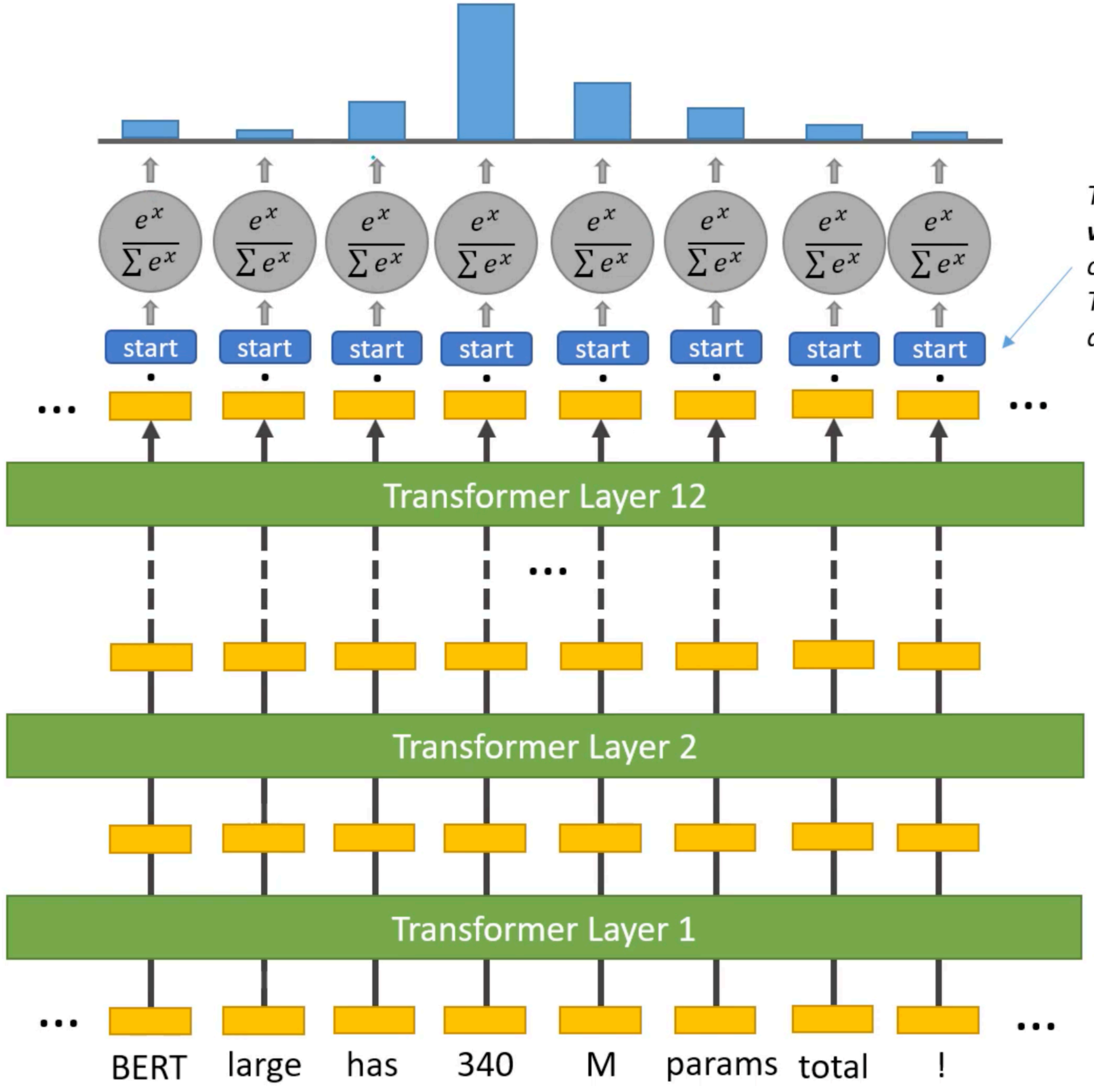
Fine-Tuning for SQuAD

- Question
  - How many parameters does BERT-large have?
- Reference
  - BERT-large is really big… it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.3GB, so expect it to take a couple minutes to download to your Colab instance.



https://youtu.be/l8ZYCvgGu0o, ChrimsMcCormickAI

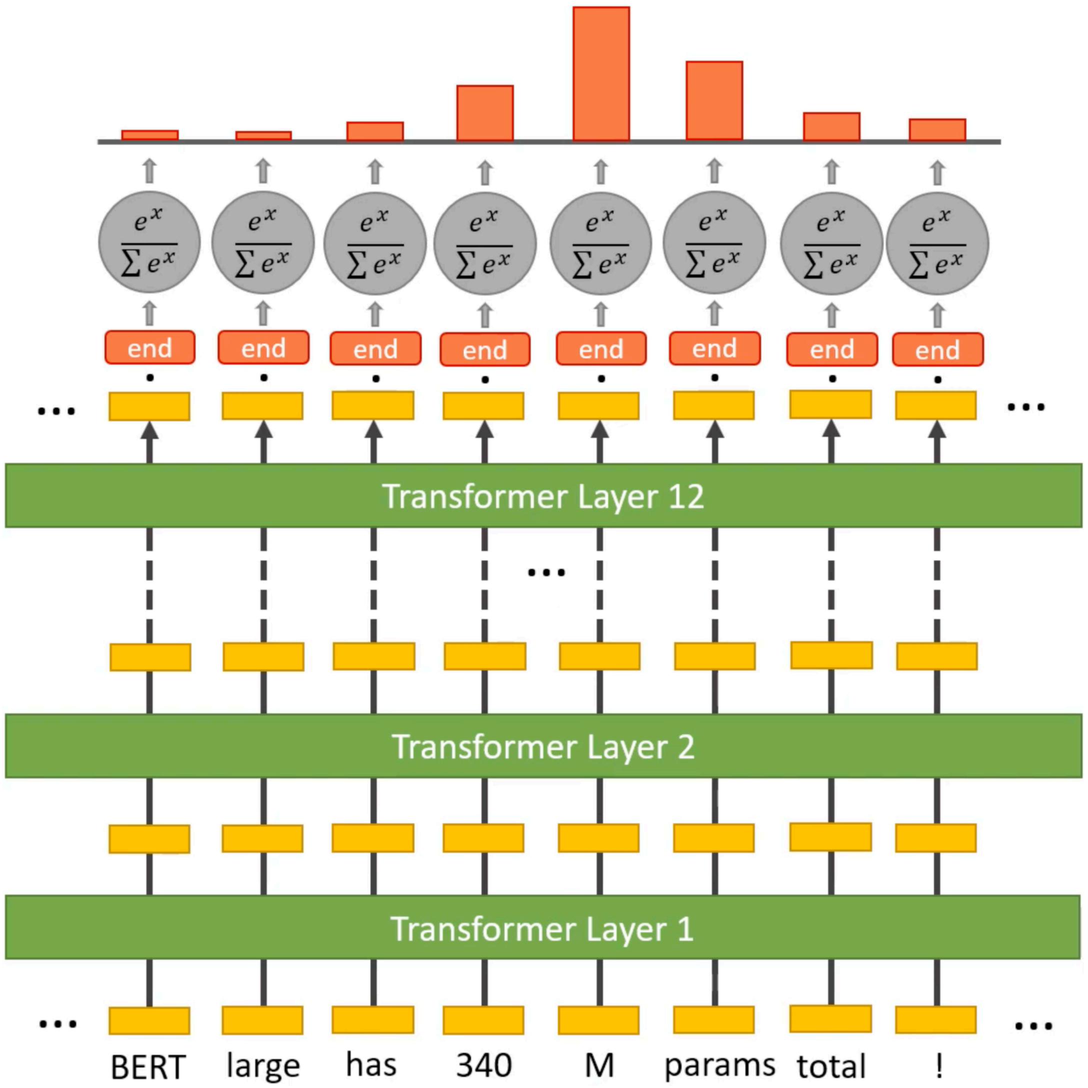
Predicting the Start Span



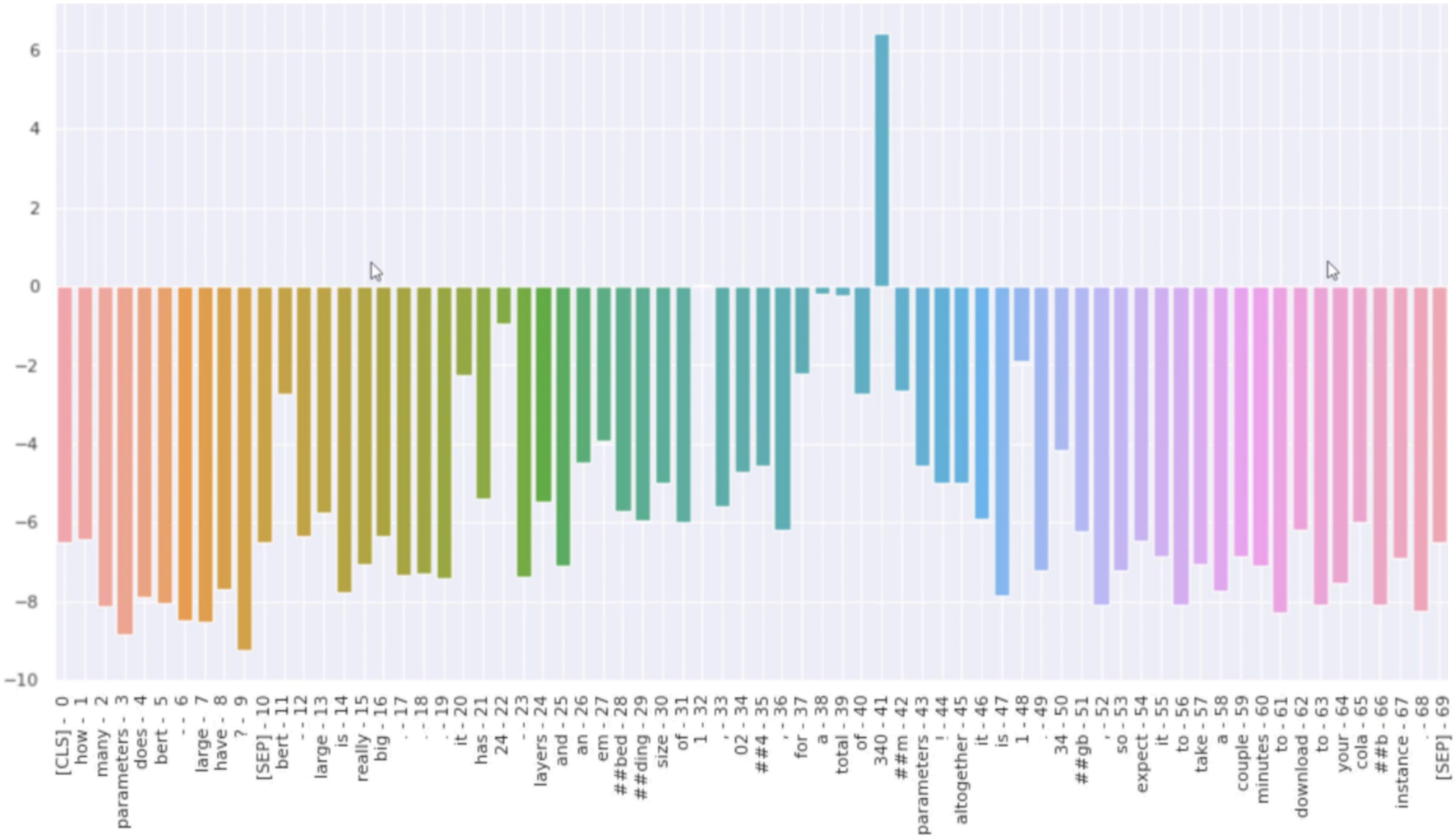
This length 768 vector is the ***weights*** *for the start token classifier.*
The ***same weights*** *are applied to* ***every position****.*

https://youtu.be/l8ZYCvgGu0o, ChrimsMcCormickAI

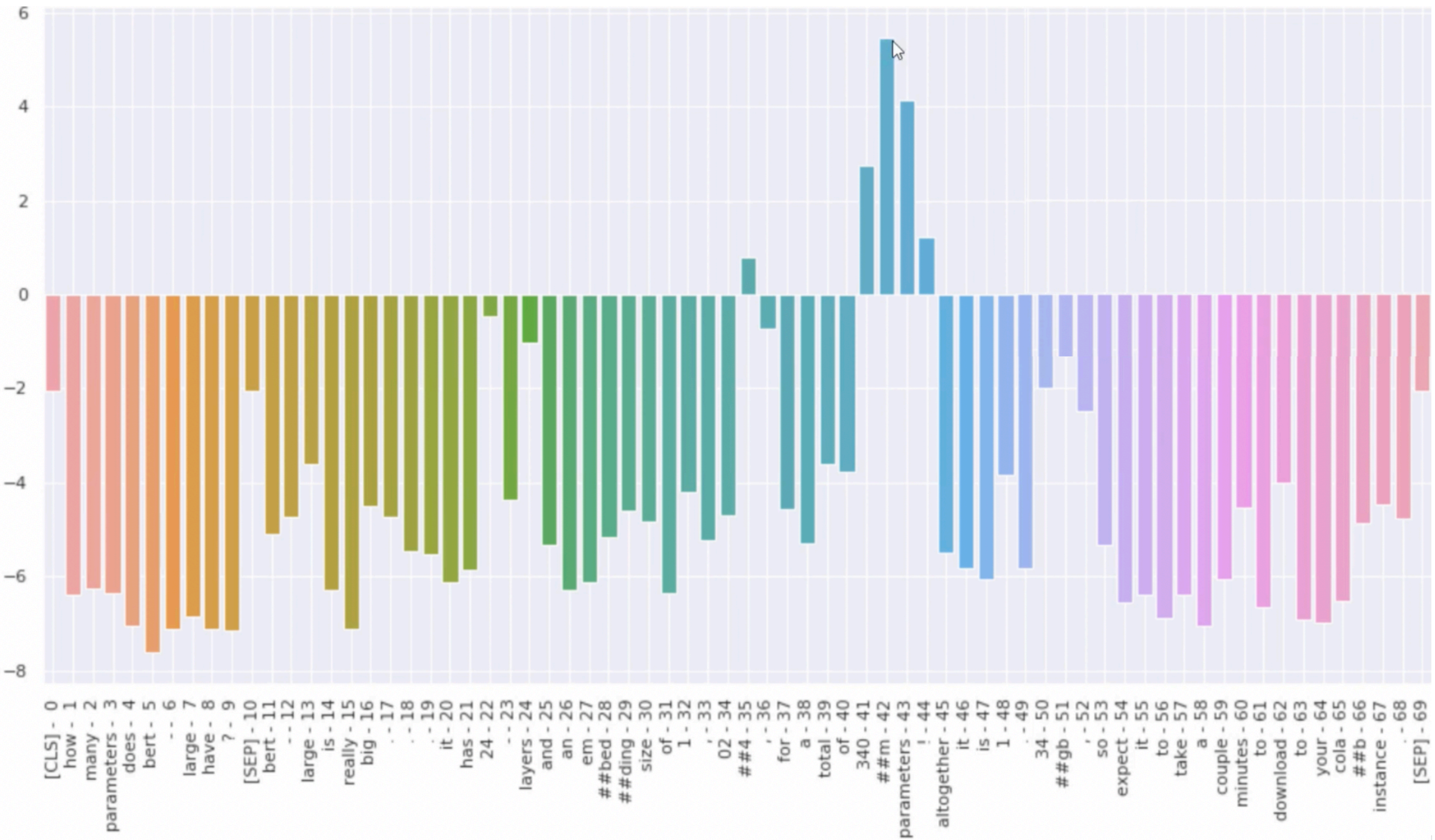
Predicting the End Span



BERT Demonstration Start Prediction



BERT Demonstration End Prediction



GPT-2 & GPT-3

GPT-2

- "Language Models are Unsupervised Multitask Learners"
  - OpenAI, 2019
  - 10 times bigger than GPT-1 (1.5B parameters)
    - 48 layers, 1600 hidden size (context size: 1024 token)
- Trained on WebText (perplexity 10~11)
  - Outbound links from Reddit posts with at least 3 Karmas
  - 8 million documents (40GB text)
  - Exclude Wikipedia, since many benchmark dataset relies on Wikipedia.
- Only perform zero-shot tasks!

GPT-2 on Zero-shot Tasks

- Need to tell GPT-2 which downstream task.
  - How to tell a language model to perform a specific task?
- Use textual input to condition GPT-2 for a specific task.
  - Summarization
    - Condition the model with a few examples, then start sampling tokens autoregressively.
    - (Long text, “TL;DR;”, summarization) x N, Long text, “TL;DR;”
  - Translation
    - (English sentence, “=”, French sentence) x N, English sentence, “=”

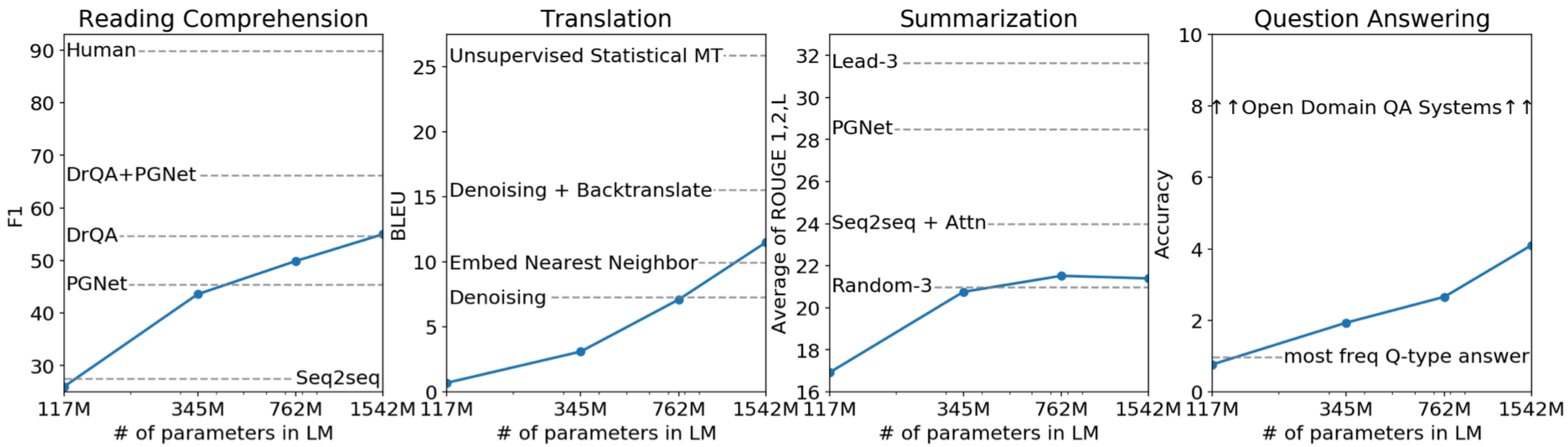
GPT-2 on Downstream Tasks

- Weirdly good zero-shot performance on multiple NLP tasks
  - No fine-tuning!

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

GPT-2 on Zero-shot Tasks

- Weirdly good zero-shot performance on multiple NLP tasks
  - No fine-tuning!



GPT-2 Text Generation Example

Context (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.

GPT-2: 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.

GPT-3

- "Language Models are Few-shot Learners", OpenAI, 2020
  - 100 times bigger than GPT-2 (175B parameters)
    - 96 layers, 12,288 hidden size, 96 attention heads (context size 2048 tokens)
  - 31 authors (vs 7 authors of GPT-2)
  - Estimated 4~5 million dollars per training
- Trained on multiple sources (at least 570GB)

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

GPT-3 Downstream Task Mode

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:      ← task description
2  cheese => ..................      ← prompt
```

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
```

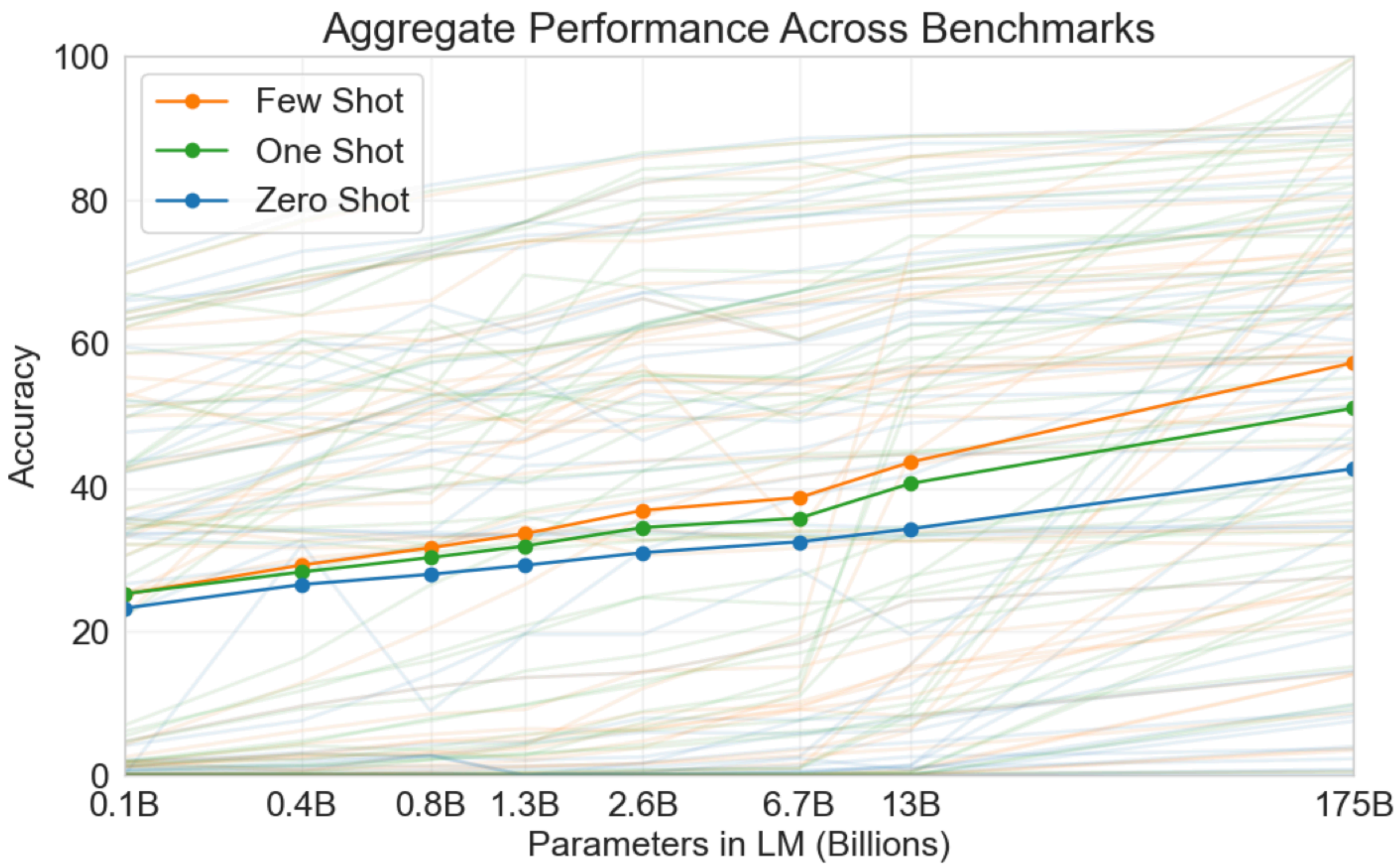
Few-shot

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
```

The Bigger, the Better

- Average performance across 42 benchmark datasets



Question Answering

- Better than fine-tuned SOTA in some cases

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

Translation

- Much better than GPT-2
  - GPT-2 En-Fr was 11.5 BLEU

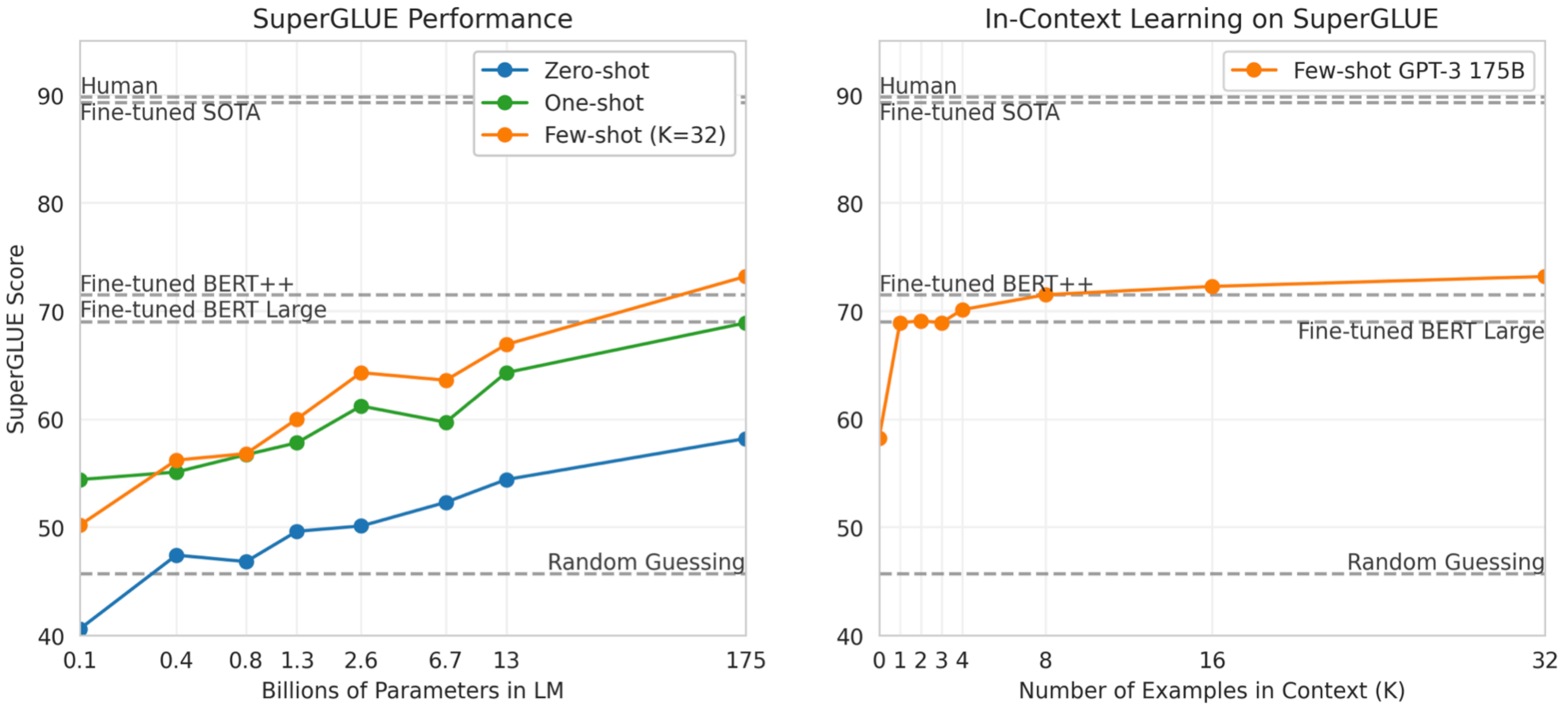
Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG+20]	-	-	<u>29.8</u>	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

Machine Reading Comprehension

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

SuperGLUE Benchmark

- More difficult language understanding tasks than GLUE



Can Humans Recognize Fake Text?

- Human accuracy in identifying whether short (∼200 word) news articles are model generated

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don't know" assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

Text most confused by humans (12% acc)

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Text most recognized by humans (61% acc)

Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: **A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what.**
Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

“Use it in a sentence”

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

AI504: Programming for Artificial Intelligence

Week 13: BERT & GPT

Edward Choi
Grad School of AI
edwardchoi@kaist.ac.kr