Session 3: NLP and Transformers

Youssef Mohamed

What is NLP?

"How computers can be used to understand and manipulate natural language text or speech to do useful things"[1]

HOW?

- Tokenization
- Part Of Speech (POS) Tagging
- Chunking

Applications

- Sentiment Analysis
- Speech Recognition
- Translation

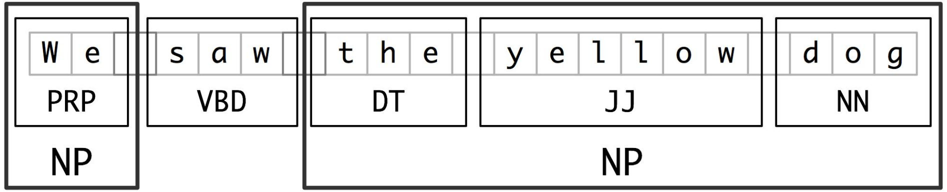
[2]

Natural Language Processing

['Natural', 'Language', 'Processing']

Lexical Term	Tag	Example
Noun	NN	Paris, France, Someone, Kurtis
Verb	VB	work, train, learn, run, skip
Determiner	DT	the, a

[3]



[4]

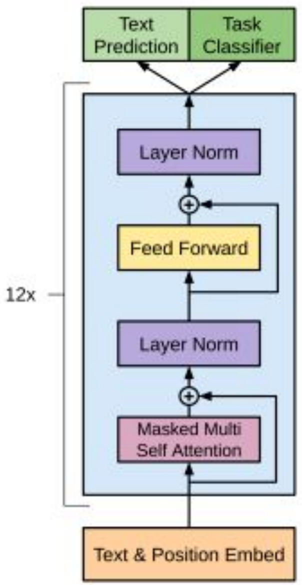
[1] Chowdhury, G.G., 2003. Natural language processing. *Annual review of information science and technology*, *37*(1), pp.51-89.
[2] https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6-unique-ways-perform-tokenization/
[3] https://towardsdatascience.com/part-of-speech-tagging-for-beginners-3a0754b2ebba
[4]https://towardsdatascience.com/chunking-in-nlp-decoded-b4a71b2b4e24

Improving Language Understanding by Generative Pre-Training

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever
OpenAI

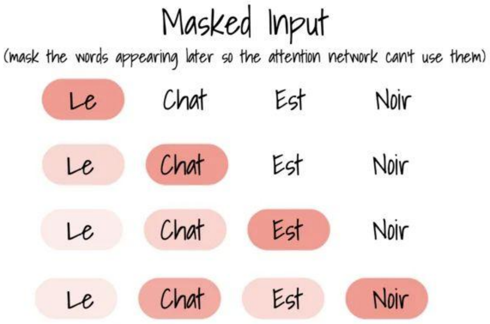
Pretraining

- BooksCorpus (800M words)
  - List of unpublished books

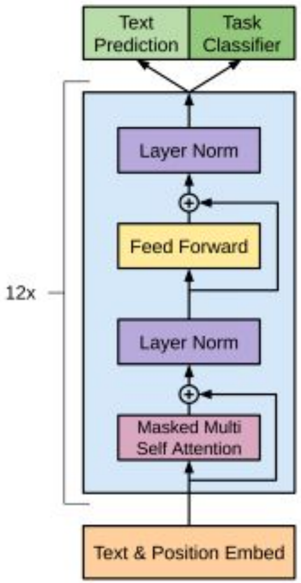


$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta) \quad (1)$$

$$\begin{aligned} h_0 &= U W_e + W_p \\ h_l &= \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n] \\ P(u) &= \texttt{softmax}(h_n W_e^T) \end{aligned} \quad (2)$$



Fine-tuning



$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m). \tag{4}$$

$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y). \tag{3}$$

$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C}) \tag{5}$$

Performance

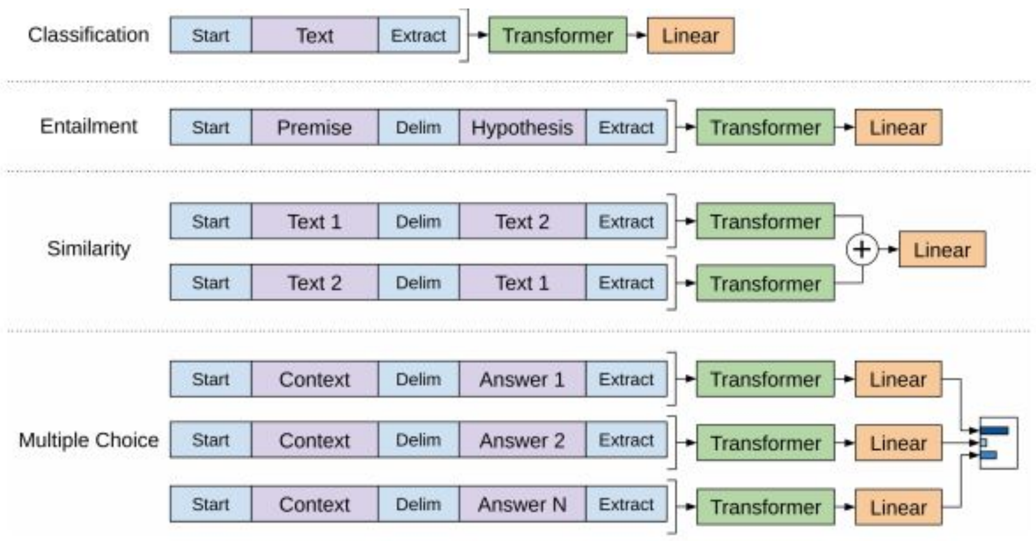
- NLI: showing the relationship between sentences (entailment, contradiction or neutral)
- QA: a passage then MCQ based on it
- SS: if two sentences are semantically similar or not
- C: grammatically correct or not,

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

“GPT-1 performed better than specifically trained supervised state-of-the-art models in 9 out of 12 tasks the models were compared on”

Task-Specific Input Transformations

- Start and end tokens
- Delimiter added so input could be sent as ordered sequence
- Minimal changes to the model



Language Models are Few-Shot Learners

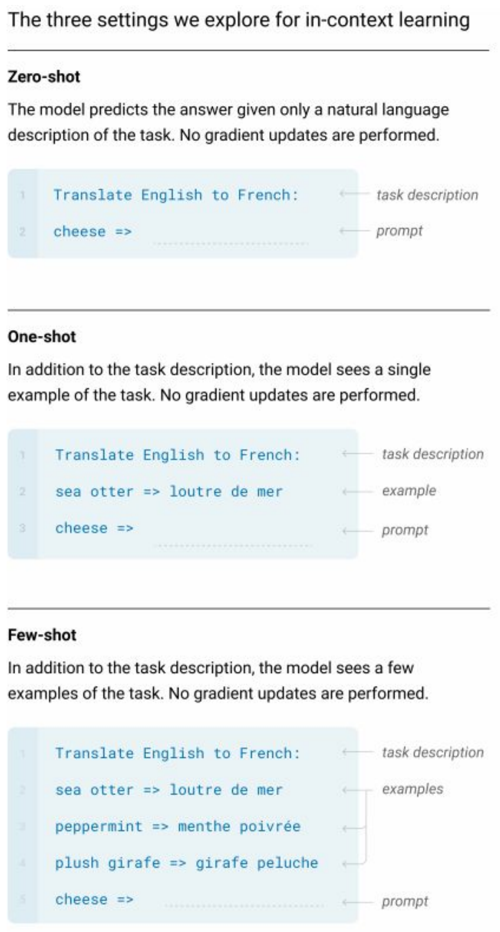
Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam ,Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei

Overview

- **GPT-1 vs GPT-3**
  - Context window size was increased from 512 for GPT-1 to 2048 tokens for GPT-3
  - 96 layers with each layer having 96 attention heads compared to 12 in GPT-1
  - 117M parameter in GPT-1 and 175B in GPT-3
- The usual pretraining and fine tuning (or is it ?)

Few shot learning

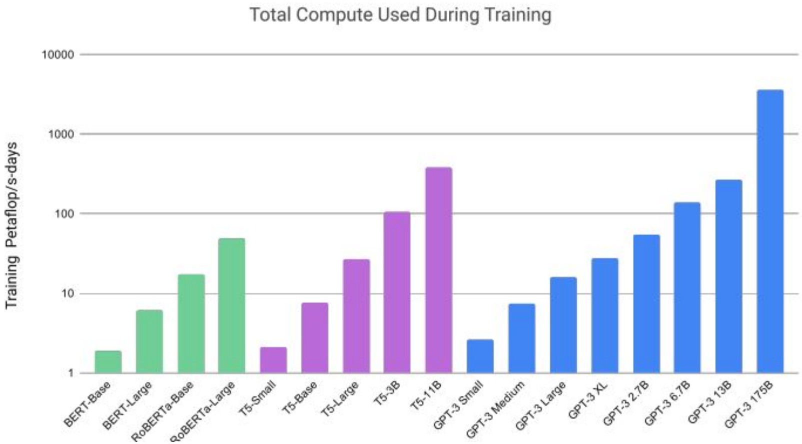
- Model does not need fine tuning
- only examples of the task



Data-sets

- **3.14E+23!!** flops in total
- **355 years!!** to train GPT-3 on a V100 [1]
- **$4,600,000!!** to train GPT-3 using the lowest cost GPU cloud provider [1]

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



[1]https://lambdalabs.com/blog/demystifying-gpt-3/

Tasks and Results

Question Answering

- Performs better in factual question answering
- Better than Open-Domain model

Example 1
Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

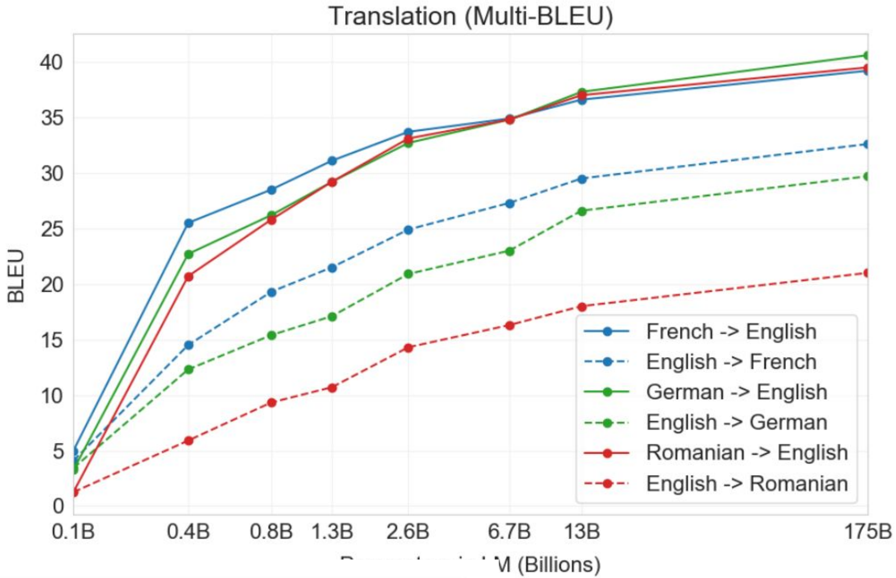
Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

Short answer: BOOLEAN:NO

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

- To English performs better



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]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	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	39.2	29.7	40.6	21.0	39.5

Reading Comprehension (Reasoning)

- Performs worse for reasoning tasks.

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

F	There's a lot of trash on the *bed* of the river — I keep a glass of water next to my *bed* when I sleep
F	*Justify* the margins — The end *justifies* the means
T	*Air* pollution — Open a window and let in some *air*
T	The expanded *window* will give us time to catch the thieves — You have a two-hour *window* of clear weather to finish working on the lawn

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

Premise: I tipped the bottle. What happened as a RESULT?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.

Premise: I knocked on my neighbor's door. What happened as a RESULT?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.

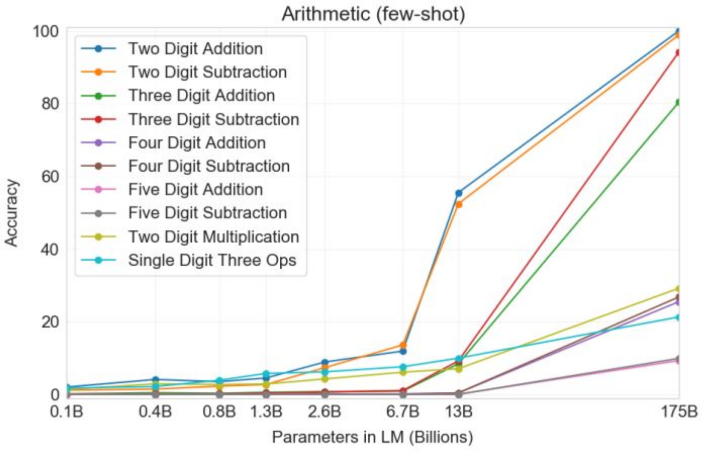
Made up tasks

Arithmetic

- Gap between three digit addition and four digit addition ?

Article Generation

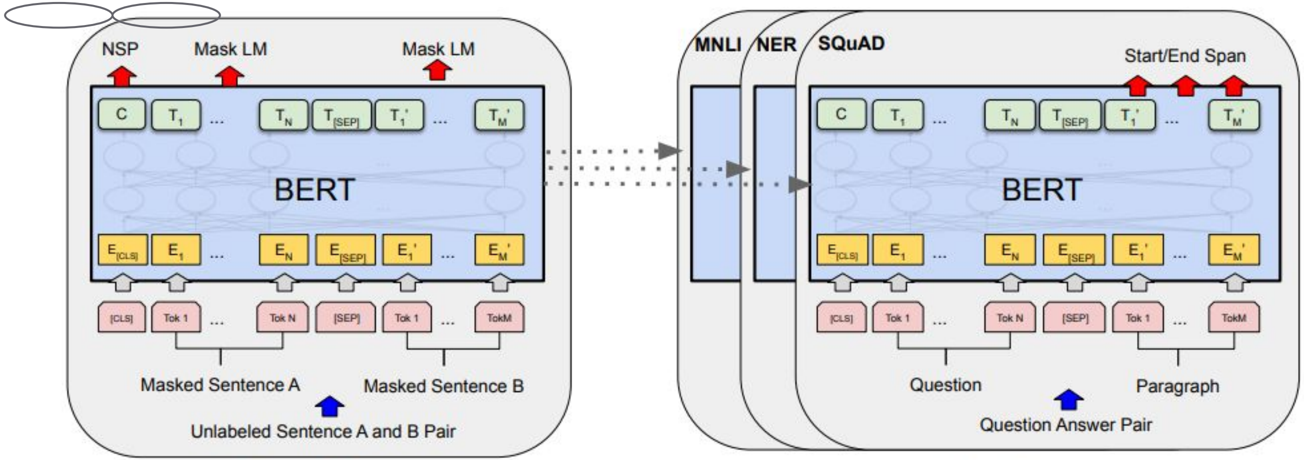
- performs worse the bigger the model is



	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%

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

BERT



Pre-training

Fine-Tuning

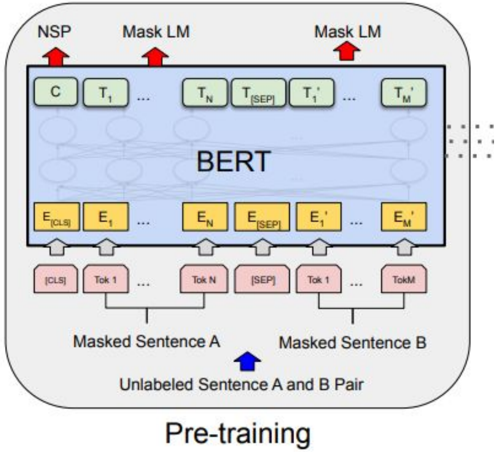
Pre-training

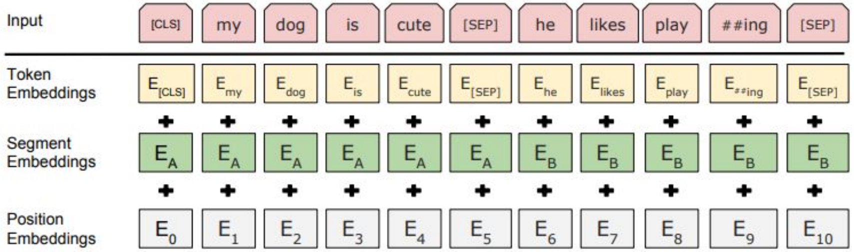
Masked LM

- Random masked words
- Each embedding is a word integrated with segment and position
- T is a word vector

Next Sentence Prediction (NSP)

- If two sentences follow each other
- binary classification (C)





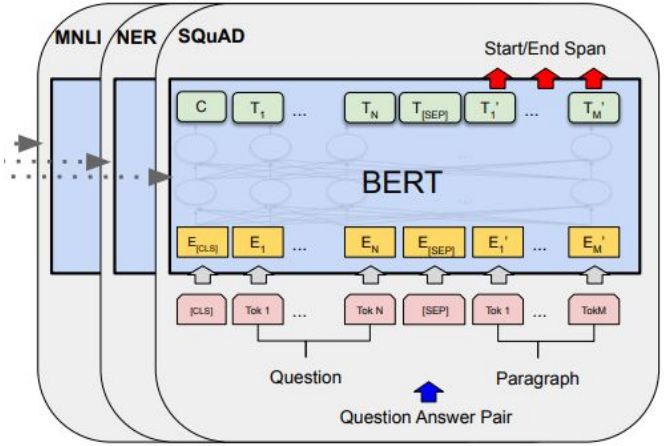
Data sets

- BooksCorpus (800M words)
  - List of unpublished books

- English Wikipedia (2,500M words)
  - Wikipedia articles

Fine-tuning

- Supervised training based on the task
- Replacing the output layer
- Modifying the input layer
- Start and end words



Fine-Tuning