AI프렌즈 세미나 #51

GPT-3 세미나 (부제: 이게 된다고?)

2020. 08. 12.

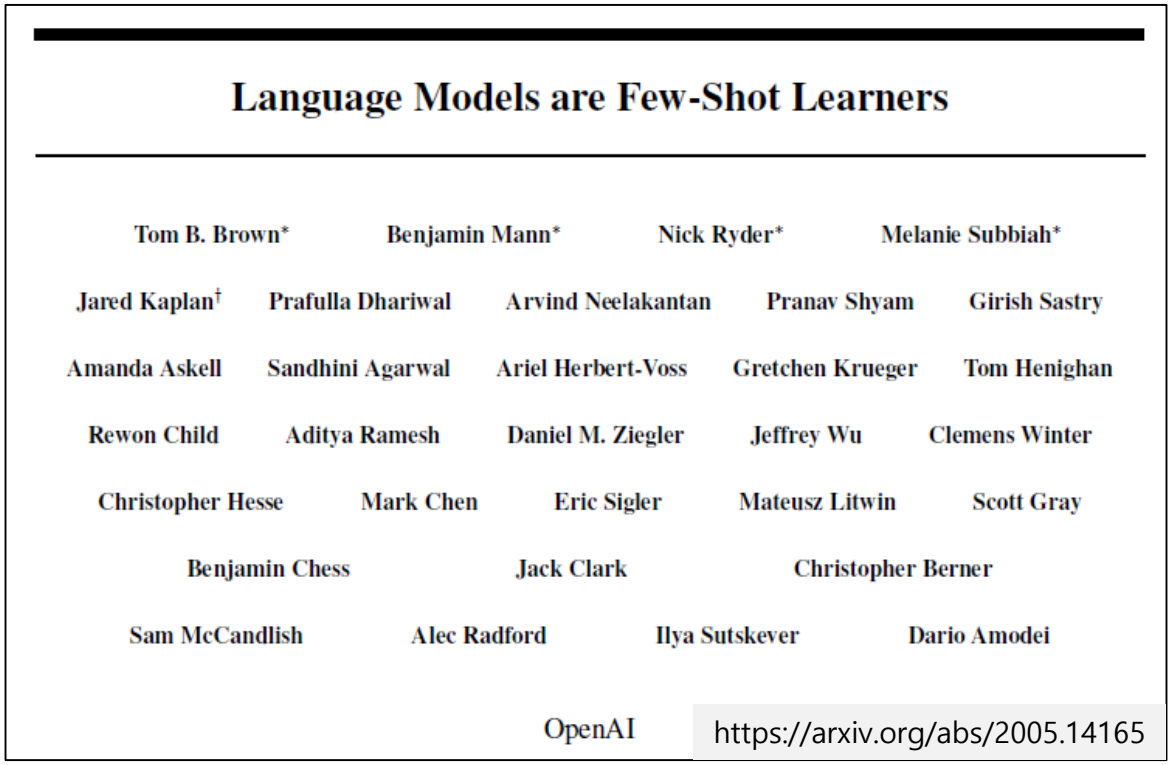
임 준 호

언어지능연구실 / 한국전자통신연구원

목차

- GPT-3 개요
- 배경지식
  - Transformer, BERT, GPT
- 모델, 데이터, 학습 과정
- 실험결과
- 한계점
- 생각해 볼 내용

GPT-3 개요



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

OpenAI

https://arxiv.org/abs/2005.14165

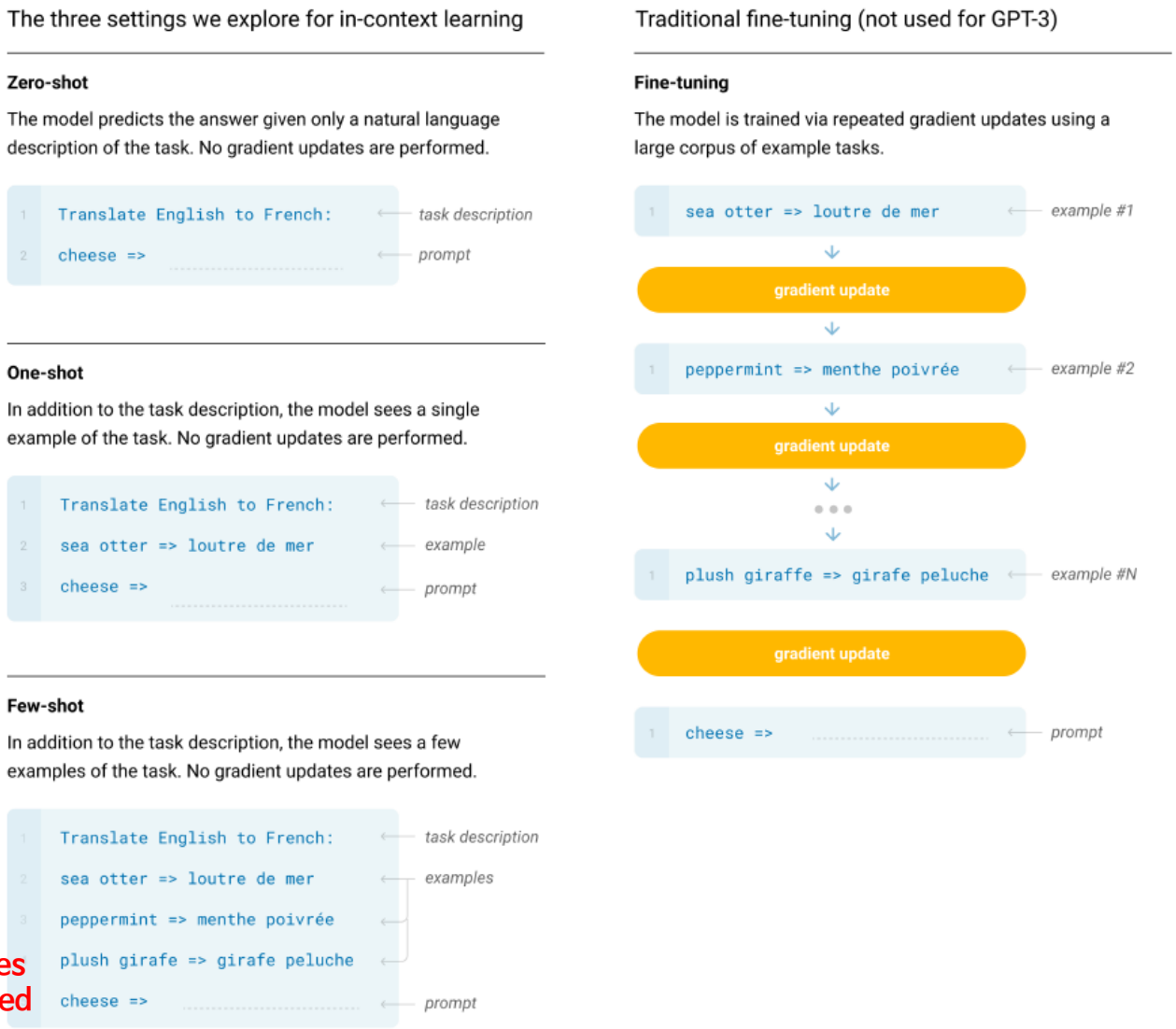
- 논문 성격
  - 새로운 모델/알고리즘 제안? => NO
  - 새로운 접근 방법? => NO
  - 논문의 기여? => 언어모델 기능에 대한 새로운 발견!
  - 논문의 주요 내용? => 다양한 실험, 공정한 평가!

GPT-3 개요

- 어떤 새로운 발견?
  - 충분히 큰 대용량의 언어모델은 퓨샷 학습이 가능함 (=논문 제목)
    - 퓨샷학습: 학습 예제를 0개(zero-shot), 1개(one-shot), 소수 개(few-shot) 만을 사용한 학습 방법
  - 비고: BERT 등 기존 언어모델은 응용 태스크 별로 대용량 학습데이터를 추가 학습(fine-tuning)하여야 적용 가능
    - GPT-3 수준의 대용량 언어모델은 소량(한 개~수십 개)의 학습데이터만으로도 응용 태스크에 바로 적용 가능

scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art finetuning approaches

- GPT-3 퓨샷 학습 접근방법



* No gradient updates are performed

- GPT-3 연구 배경 = GPT-2

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

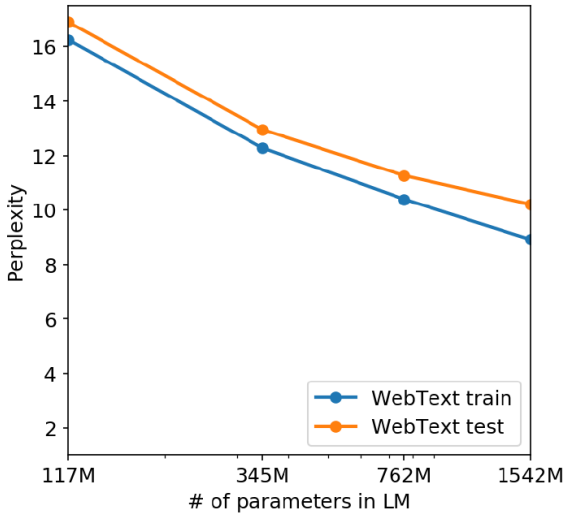
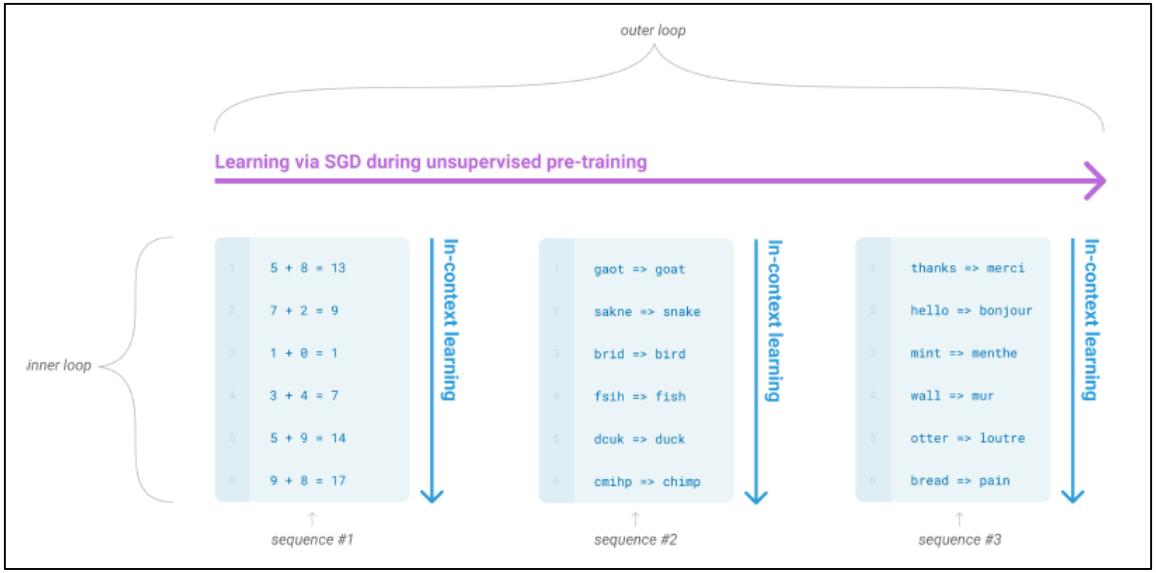


Figure 4. The performance of LMs trained on WebText as a function of model size.

- 잠시, 용어 정리
  - In GPT-2, zero-shot transfer learning 용어 사용
    - 기존 pre-training & fine-tuning 시, transfer learning 용어 사용
      - No weight update = zero-shot
      - But, weight update가 없더라도 퓨샷 학습 예제 제공 시, 실제 zero-example 이 아님 (혼동)
  - In GPT-3, meta learning 용어 사용
    - outer-loop, inner-loop (in-context learning) 용어 사용



- GPT-3 Hypothesis
  - 언어모델 크기와 학습 loss는 power-law 관계를 가짐
    - Since in-context learning involves absorbing many skills and tasks within the parameters of the model, it is plausible that in-context learning abilities might show similarly strong gains with scale.

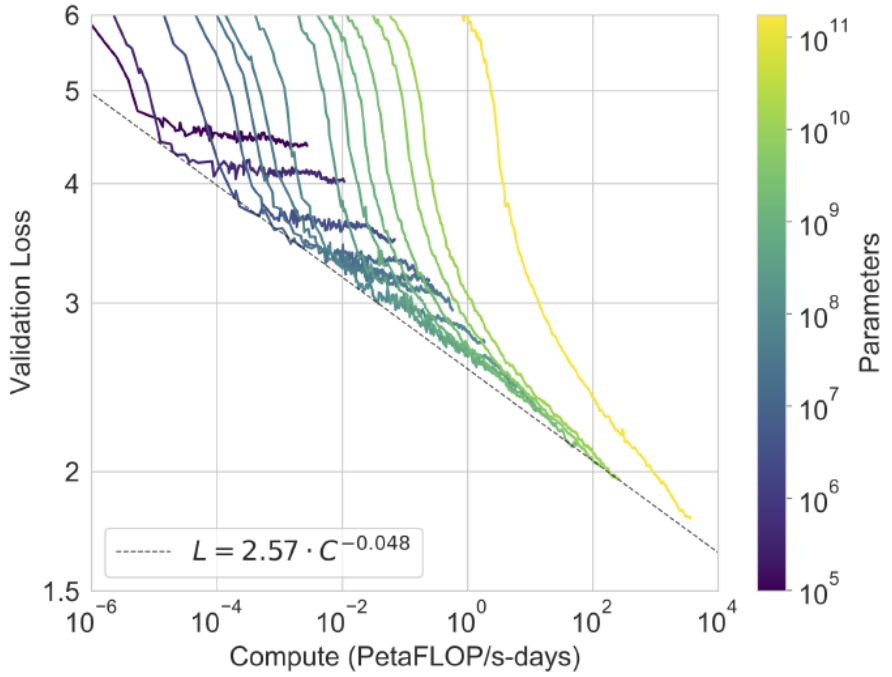
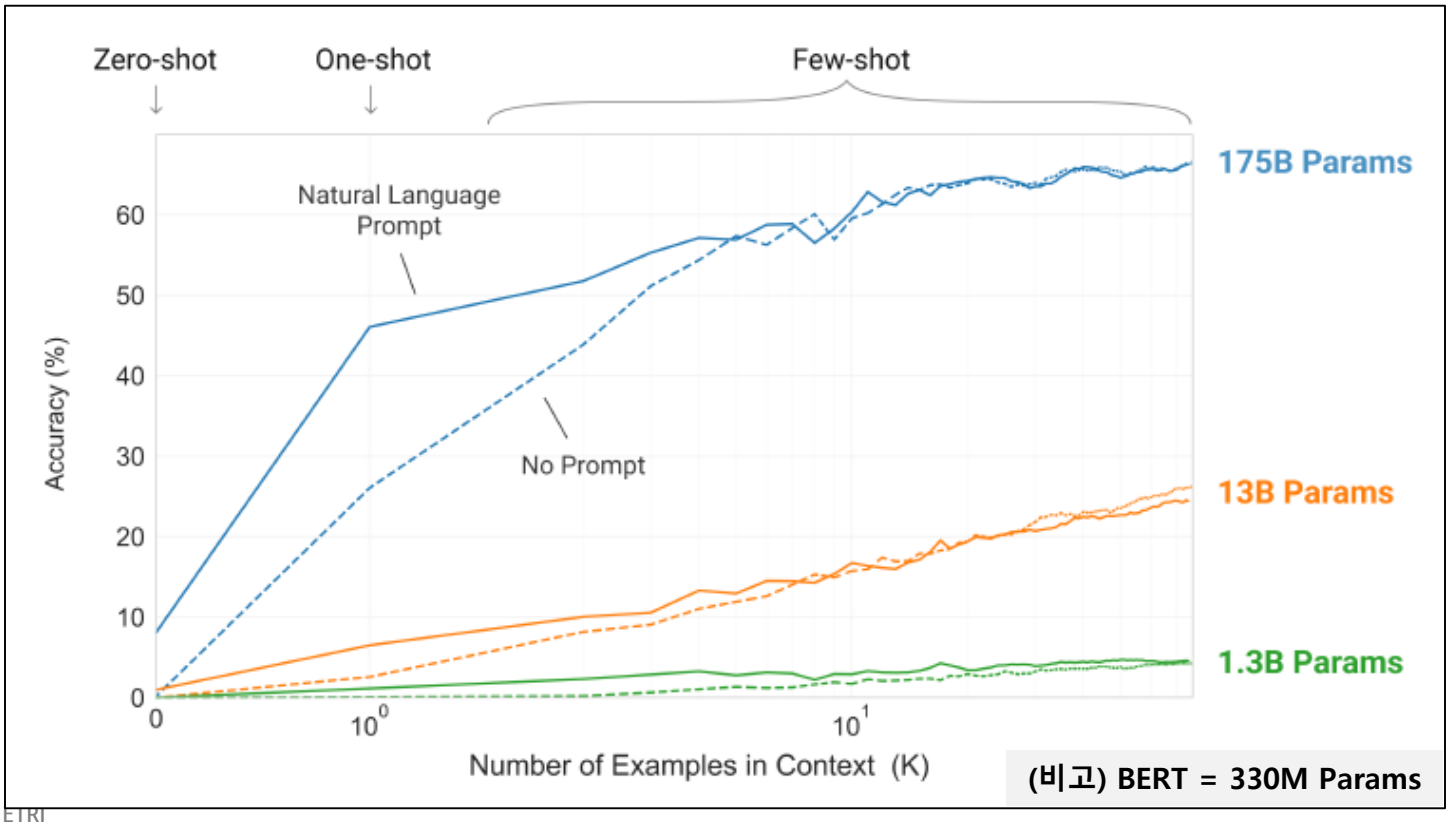


Figure 3.1: Smooth scaling of performance with compute. Performance (measured in terms of cross-entropy validation loss) follows a power-law trend with the amount of compute used for training. The power-law behavior observed in [KMH+20] continues for an additional two orders of magnitude with only small deviations from the predicted curve. For this figure, we exclude embedding parameters from compute and parameter counts.

- (논문 핵심) 언어모델 크기 별 퓨샷 학습 성능
  - 응용 태스크: Random insertion in word (RI)
    - A random punctuation or space character is inserted between each letter of a word, and the model must output the original word.
      - Example: s.u!c/c!e.s s i/o/n = succession
      - (비고) 단위 = BPE 토큰 (임의 길이의 subword) 단위



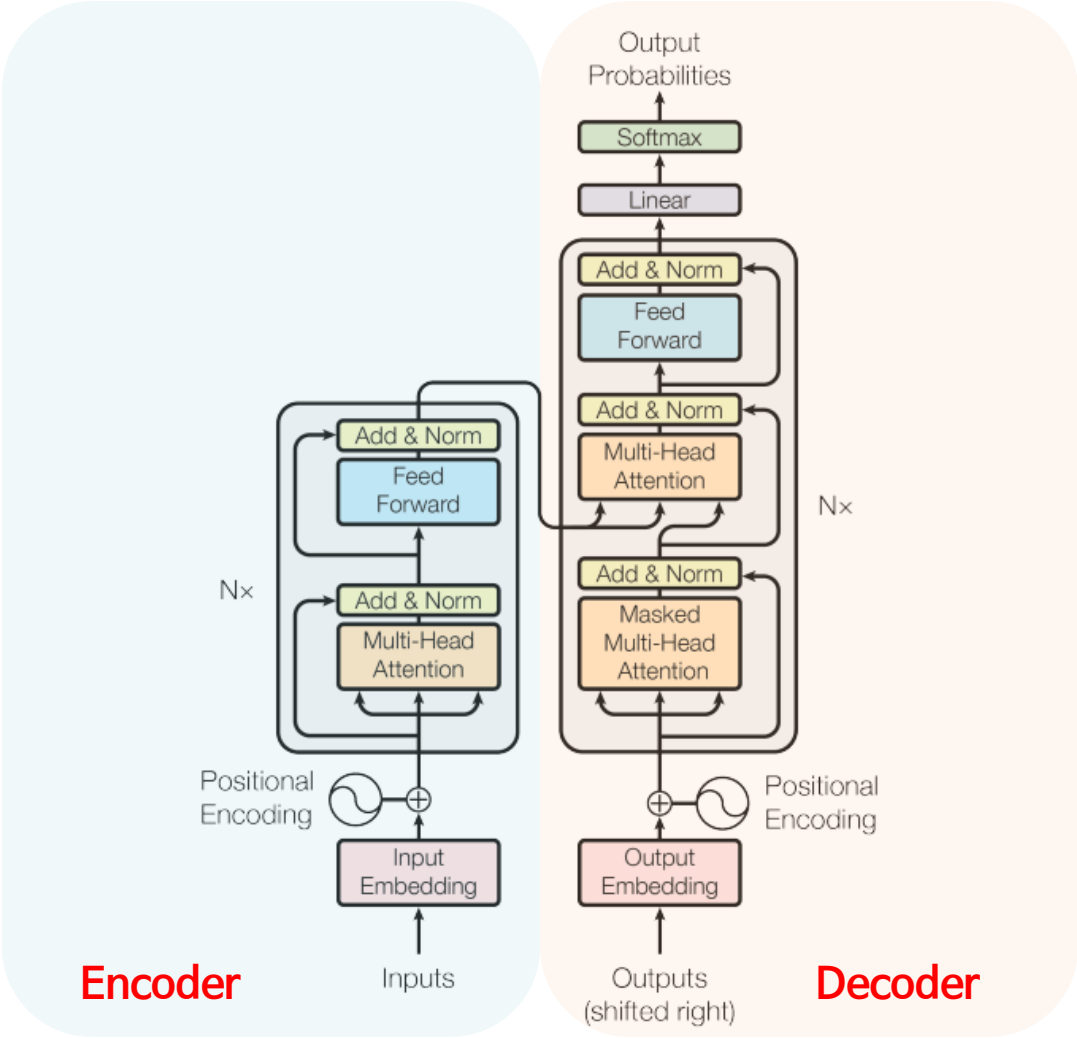
GPT-3 개요

- OpenAI에서 퓨샷-학습을 연구하는 동기
  - 1) the need for a large dataset of labeled examples for every new task limits the applicability of language models
    - There exists a very wide range of possible useful language tasks
    - it is difficult to collect a large supervised training dataset
  - 2) the generalization achieved under the pre-training & fine-tuning paradigm can be poor because the model is overly specific to the training distribution
  - 3) humans do not require large supervised datasets to learn most language tasks
    - It is sufficient to enable a human to perform a new task
      - a brief directive in natural language (e.g. “please tell me if this sentence describes something happy or something sad”)
      - at most a tiny number of demonstrations (e.g. “here are two examples of people acting brave; please give a third example of bravery”)

GPT-3 개요

- GPT-3 API 활용 예
  - 1) https://blog.pingpong.us/gpt3-review/
  - 2) https://github.com/elyase/awesome-gpt3
    - https://twitter.com/i/status/1282676454690451457
    - https://twitter.com/FaraazNishtar/status/12859346228916674557
  - 3) https://twitter.com/gdb
  - 4) https://gptcrush.com/

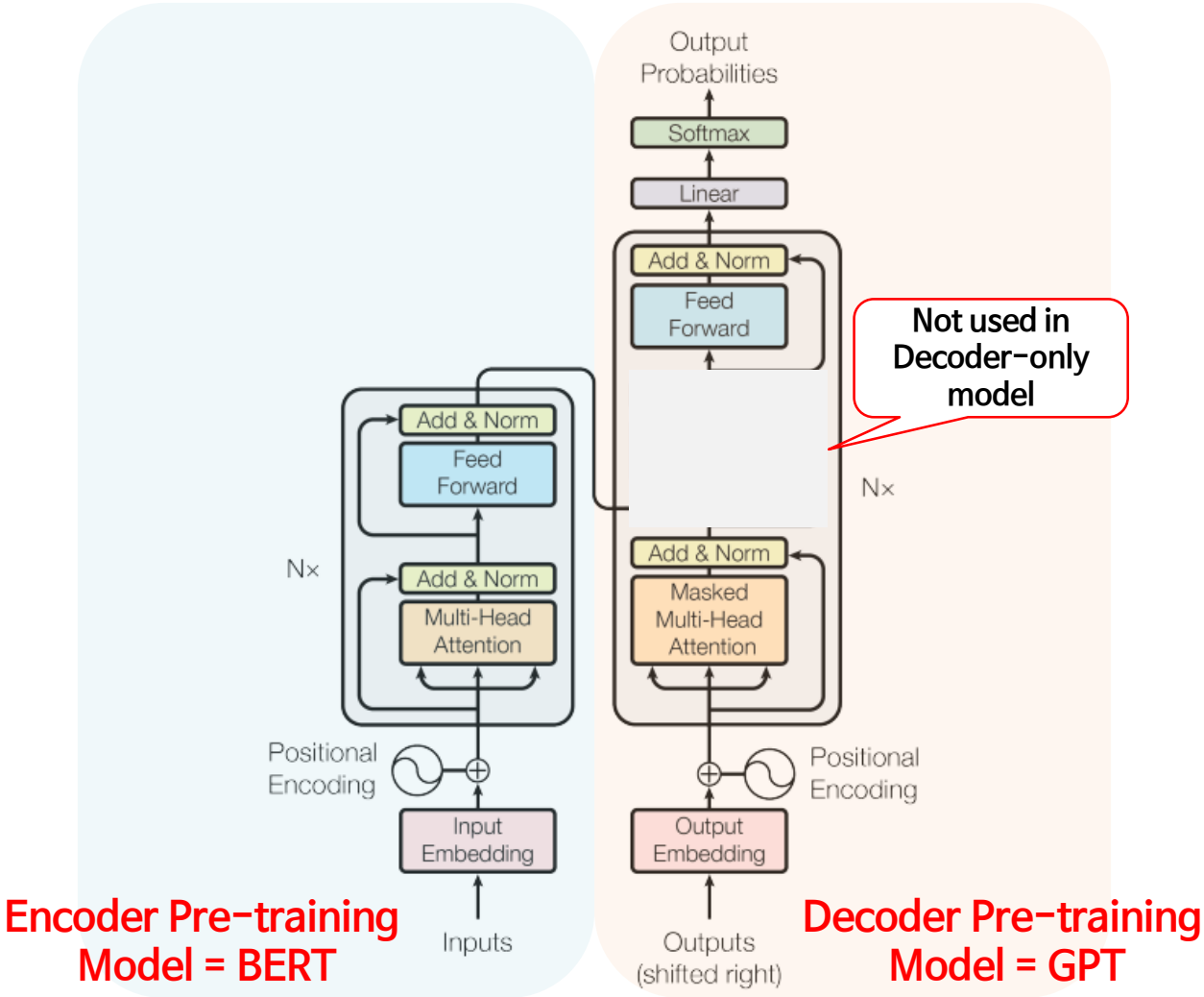
- Original Transformer architecture



나는 학교에 간다.

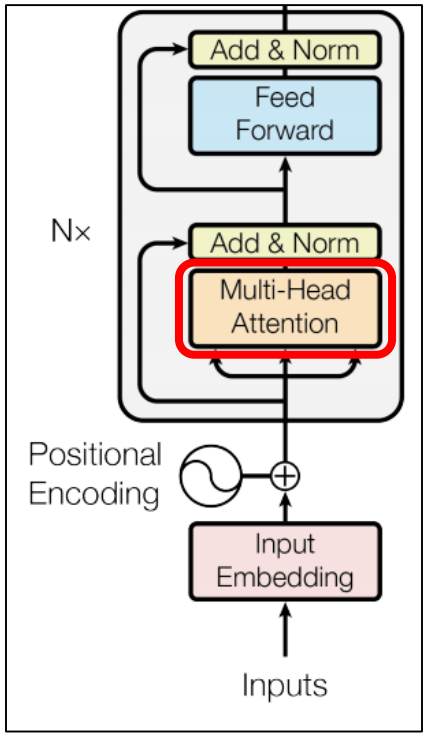
I -> go -> to -> school -> . (auto-regressive manner)

- BERT & GPT

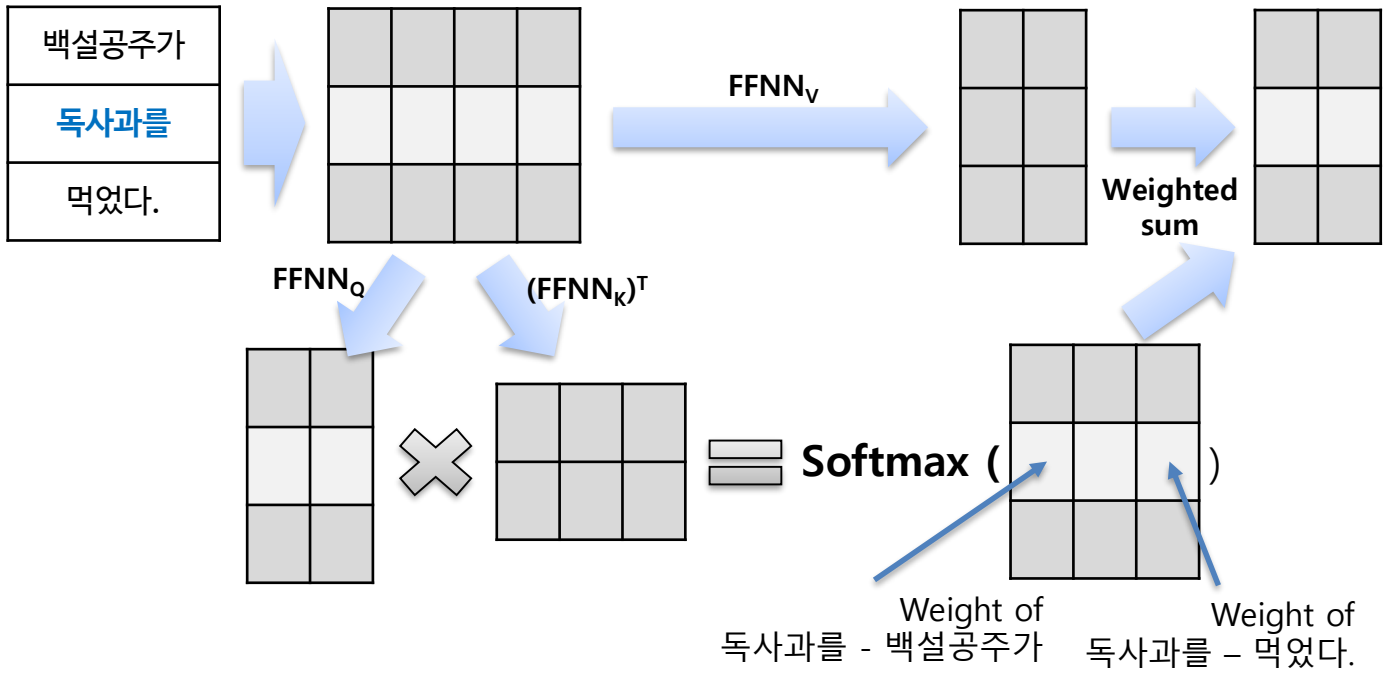


- (Building block #1) self-attention mechanism

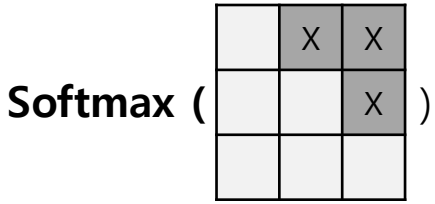
$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



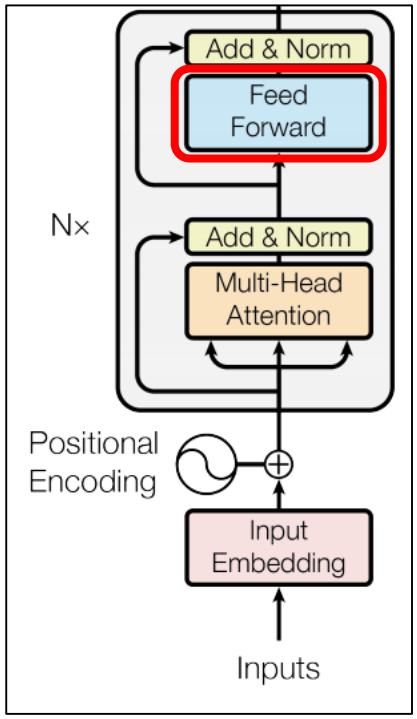
1) Transformer encoder model,



2) Transformer decoder model,



- (Building block #2) FFNN



백설공주가

독사과를

먹었다.

Multi-head Attention
Add & Norm

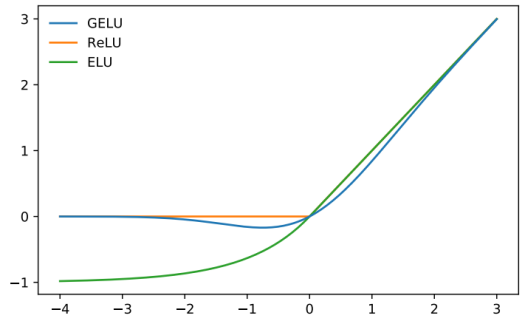
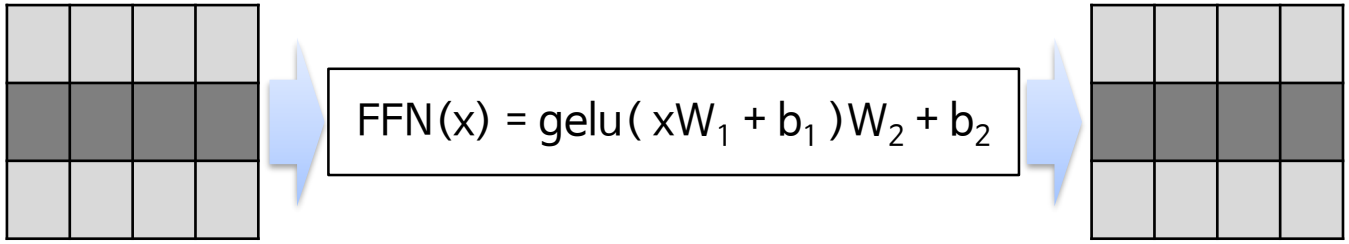


Figure 1: The GELU ($\mu = 0, \sigma = 1$), ReLU, and ELU ($\alpha = 1$).



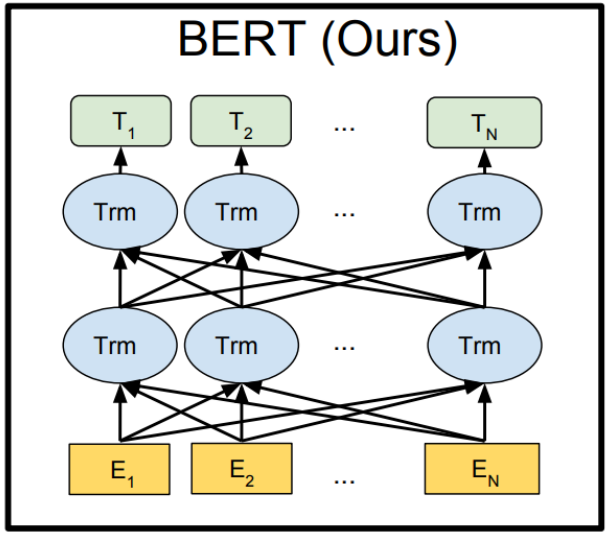
$$FFN(x) = gelu(xW_1 + b_1)W_2 + b_2$$

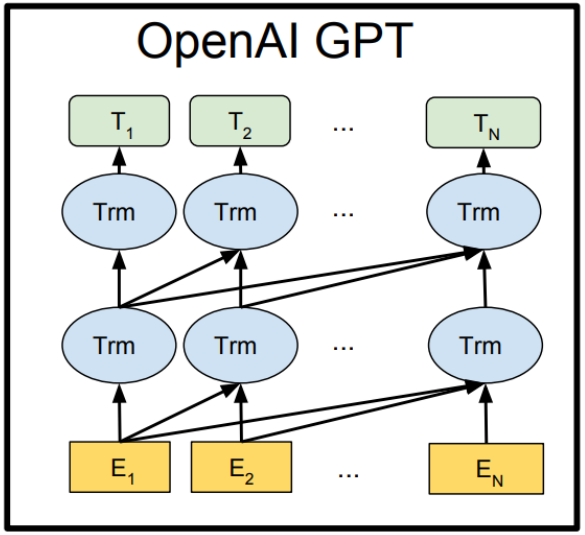
Note that,

* 중간계층 dim = hidden dim * 4
  - BERT-Large: 1024 x 4096
  - transformer: 512 x 2048

- BERT & GPT

	BERT	GPT
입력/출력	입력: N개의 단어 열 출력: N개 단어의 인코딩 벡터	입력: N개의 단어 열 출력: N+1번째 단어
모델 특징	양방향(bi-directional) 문맥 정보 활용	단방향(uni-directional) 문맥 정보 활용
사전학습 태스크	입력: 임의의 단어 masking 출력: 주위 문맥으로 해당 단어 맞추기	입력: 이전 단어 열 출력: 다음 단어 예측
비고	동일 문장이라도 random masking 위치에 따라 서로 다른 정보 학습 (→ random masking을 수 차례 반복하여 학습)	동일 문장이면 동일 정보 학습





모델, 데이터, 학습

- 모델
  - 1) GPT-2와 동일한 모델 사용
    - the modified initialization, pre-normalization, and reversible tokenization
    - alternating dense and locally banded sparse attention patterns in the layers of the transformer, similar to the Sparse Transformer
    - 최대 길이: n_ctx = 2048
  - 2) 모델 크기 별 성능 검증을 위해 8가지 크기의 모델 학습

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens. * 300B / 175B

모델, 데이터, 학습

- 데이터

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

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

1) The CommonCrawl data was downloaded from 41 shards of monthly CommonCrawl covering 2016 to 2019, constituting 45TB of compressed plaintext before filtering and 570GB after filtering, roughly equivalent to 400 billion byte-pair-encoded tokens.
2) an expanded version of the WebText dataset, collected by scraping links over a longer period of time
3) two internet-based books corpora (Books1 and Books2)
4) English-language Wikipedia.

모델, 데이터, 학습

- Note that, training tokens per params

	BERT	GPT
모델	BERT-Large 24 layers 1024 hidden dims, 16 multi heads	GPT-3 96 layers 12288 hidden dims, 96 multi heads
파라미터 수	330M	175,000M (175B)
Vocab.	30,000	50,257
학습 데이터	3.3B	300B
배치	256 * 512 (= 131,072)	3,200,000 (= 1,563 * 2048)
학습 횟수 (step)	1,000,000	93,750
Training Tokens	131B tokens	300B tokens
training tokens per params	**396.96**	**1.71**

모델, 데이터, 학습

- 학습 과정
  - 1) always train on sequences of the full n_ctx = 2048 token context window
    - packing multiple documents into a single sequence in order to increase computational efficiency
    - documents within a sequence are delimited with a special end of text token
  - 2) larger models can typically use a larger batch size, but require a smaller learning rate
    - the gradient noise scale during training and use it to guide our choice of batch size
  - 3) To train the larger models without running out of memory,
    - we use a mixture of model parallelism within each matrix multiply and model parallelism across the layers of the network
      - (c.f.) 175B parameter needs 700GB GPU memory

- 학습 과정
  - 4) H/W Environment
    - https://blogs.microsoft.com/ai/openai-azure-supercomputer/
      - The supercomputer developed for OpenAI is a single system with more than 285,000 CPU cores, 10,000 GPUs and 400 gigabits per second of network connectivity for each GPU server.
      - Compared with other machines listed on the TOP500 supercomputers in the world, it ranks in the top five, Microsoft says.

실험결과

- 공정한 평가 = Test set contamination study
  - 온라인 상의 대용량 텍스트를 사전학습 데이터로 사용하여, 평가셋의 데이터를 포함하고 있을 위험을 제거 (made a best effort)
    - remove text from training data by searching for 13 gram overlaps between all test/development sets used in this work
  - Unfortunately, a bug resulted in only partial removal of all detected overlaps from the training data. (^^)

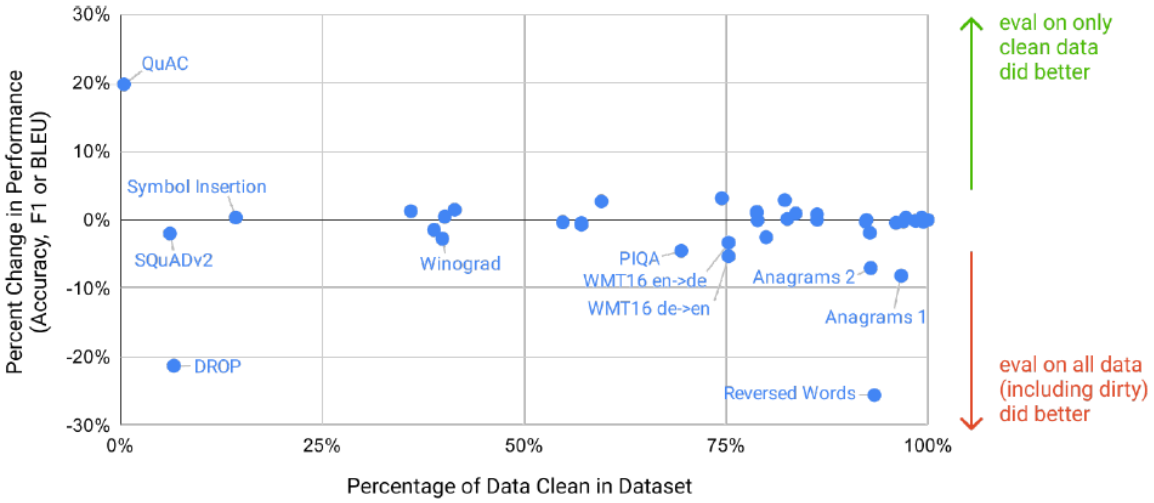


Figure 4.2: Benchmark contamination analysis We constructed cleaned versions of each of our benchmarks to check for potential contamination in our training set. The x-axis is a conservative lower bound for how much of the dataset is known with high confidence to be clean, and the y-axis shows the difference in performance when evaluating only on the verified clean subset. Performance on most benchmarks changed negligibly, but some were flagged for further review. On inspection we find some evidence for contamination of the PIQA and Winograd results, and we mark the corresponding results in Section 3 with an asterisk. We find no evidence that other benchmarks are affected.

실험결과

- 다양한 실험
  - 9가지 유형, 24개 평가셋 대상 실험
  - 1) 언어 생성
  - 2) 퓨샷 능력 평가를 위한 가상 태스크
  - 3) 기계독해, 질의응답
  - 4) 기계번역

실험결과

- 뉴스 기사 생성
  - 진짜 뉴스 기사와 동일한 제목, 부제목을 GPT-3에 입력하여 뉴스 기사를 생성하고, 평가자가 진짜 뉴스 기사와 GPT-3가 생성한 뉴스 기사를 구분
    - The articles we selected were not in the models' training data and the model outputs were formatted and selected programmatically to prevent human cherry-picking.

	Mean accuracy	95% Confidence Interval (low, hi)	*t* compared to control (*p*-value)	"I don't know" assignments
Control	88%	84%–91%	-	2.7%
GPT-3 175B	52%	48%–57%	12.7 (3.2*e*-23)	10.6%

Table 3.12: People's ability to identify whether ~ 500 word articles are model generated (as measured by the ratio of correct assignments to non-neutral assignments) was 88% on the control model and 52% on GPT-3 175B. This table shows the results of a two-sample T-Test for the difference in mean accuracy between GPT-3 175B and the control model (an unconditional GPT-3 Small model with increased output randomness).

- 뉴스 기사 생성 샘플

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.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

실험결과

- (가상 태스크) Learning and Using Novel Words
  - the ability to learn and utilize new words, for example using a word in a sentence after seeing it defined only once, or conversely inferring a word's meaning from only one usage.
    - These examples were generated continuously in one sitting and we did not omit or repeatedly try any prompts.

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.

Figure 3.16: Representative GPT-3 completions for the few-shot task of using a new word in a sentence. Boldface is GPT-3's completions, plain text is human prompts. In the first example both the prompt and the completion are provided by a human; this then serves as conditioning for subsequent examples where GPT-3 receives successive additional prompts and provides the completions. Nothing task-specific is provided to GPT-3 other than the conditioning shown here.

- (가상 태스크) Arithmetic
  - asking GPT-3 a simple arithmetic problem in natural language

- **2 digit addition (2D+)** – The model is asked to add two integers sampled uniformly from [0, 100), phrased in the form of a question, e.g. "Q: What is 48 plus 76? A: 124."
- **2 digit subtraction (2D-)** – The model is asked to subtract two integers sampled uniformly from [0, 100); the answer may be negative. Example: "Q: What is 34 minus 53? A: -19".
- **3 digit addition (3D+)** – Same as 2 digit addition, except numbers are uniformly sampled from [0, 1000).
- **3 digit subtraction (3D-)** – Same as 2 digit subtraction, except numbers are uniformly sampled from [0, 1000).
- **4 digit addition (4D+)** – Same as 3 digit addition, except uniformly sampled from [0, 10000).
- **4 digit subtraction (4D-)** – Same as 3 digit subtraction, except uniformly sampled from [0, 10000).
- **5 digit addition (5D+)** – Same as 3 digit addition, except uniformly sampled from [0, 100000).
- **5 digit subtraction (5D-)** – Same as 3 digit subtraction, except uniformly sampled from [0, 100000).
- **2 digit multiplication (2Dx)** – The model is asked to multiply two integers sampled uniformly from [0, 100), e.g. "Q: What is 24 times 42? A: 1008".
- **One-digit composite (1DC)** – The model is asked to perform a composite operation on three 1 digit numbers, with parentheses around the last two. For example, "Q: What is 6+(4*8)? A: 38". The three 1 digit numbers are selected uniformly on [0, 10) and the operations are selected uniformly from {+,-,*}.

* To spot-check whether the model is simply memorizing specific arithmetic problems,
we took the 3-digit arithmetic problems in our test set and searched for them in our training data
in both the forms "<NUM1> + <NUM2> =" and "<NUM1> plus <NUM2>".
Out of 2,000 addition problems we found only 17 matches (0.8%) and
out of 2,000 subtraction problems we found only 2 matches (0.1%),
suggesting that only a trivial fraction of the correct answers could have been memorized.

- (가상 태스크) Arithmetic

Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3

Table 3.9: Results on basic arithmetic tasks for GPT-3 175B. {2,3,4,5}D{+,-} is 2, 3, 4, and 5 digit addition or subtraction, 2Dx is 2 digit multiplication. 1DC is 1 digit composite operations. Results become progressively stronger moving from the zero-shot to one-shot to few-shot setting, but even the zero-shot shows significant arithmetic abilities.

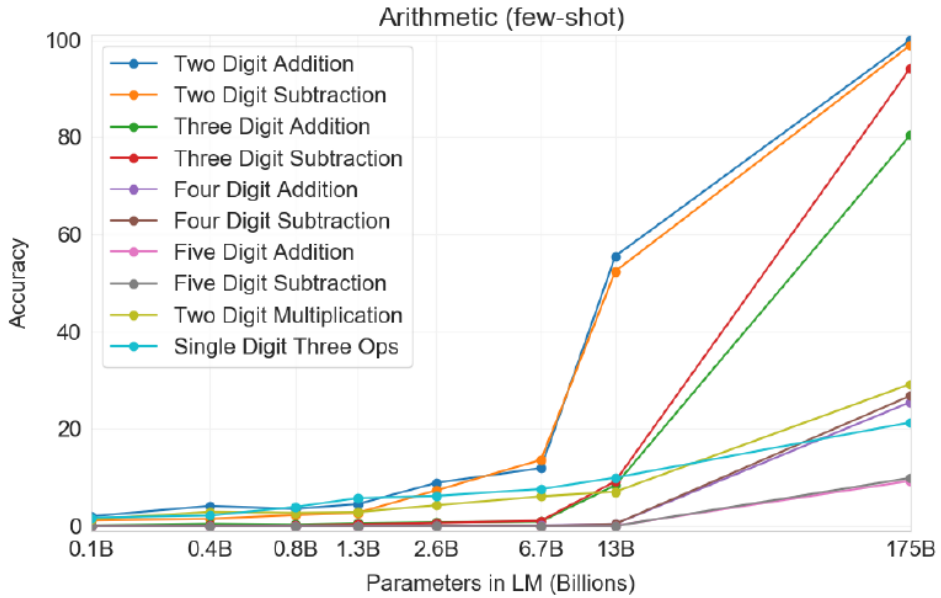


Figure 3.10: Results on all 10 arithmetic tasks in the few-shot settings for models of different sizes. There is a significant jump from the second largest model (GPT-3 13B) to the largest model (GPT-3 175), with the latter being able to reliably accurate 2 digit arithmetic, usually accurate 3 digit arithmetic, and correct answers a significant fraction of the time on 4-5 digit arithmetic, 2 digit multiplication, and compound operations. Results for one-shot and zero-shot are shown in the appendix.

실험결과

- 기계독해
  - 질문과 단락이 주어졌을 때, 정답을 인식하는 문제

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

Table 3.7: Results on reading comprehension tasks. All scores are F1 except results for RACE which report accuracy.
[a][JZC+19] [b][JN20] [c][AI19] [d][QIA20] [e][SPP+19]

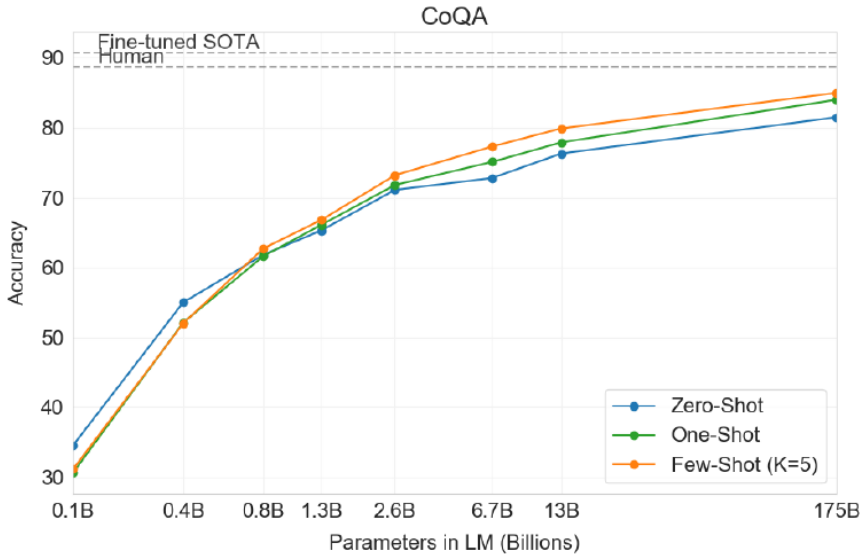
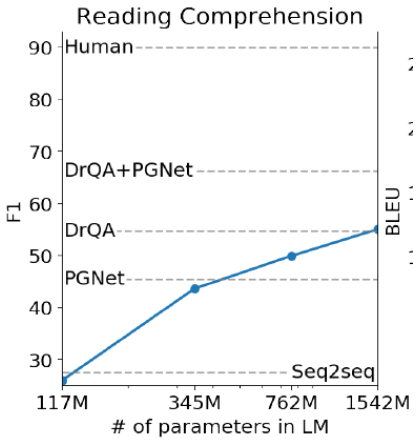


Figure 3.7: GPT-3 results on CoQA reading comprehension task. GPT-3 175B achieves 85 F1 in the few-shot setting, only a few points behind measured human performance and state-of-the-art fine-tuned models. Zero-shot and one-shot performance is a few points behind, with the gains to few-shot being largest for bigger models.

* 비고: GPT-2



실험결과

- Closed-book 환경 질의응답
  - A large language model can perform surprisingly well directly answering the questions without conditioning on auxilliary information.
    - They denote this more restrictive evaluation setting as “closed-book”.

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**

Table 3.3: Results on three Open-Domain QA tasks. GPT-3 is shown in the few-, one-, and zero-shot settings, as compared to prior SOTA results for closed book and open domain settings. TriviaQA few-shot result is evaluated on the wiki split test server.

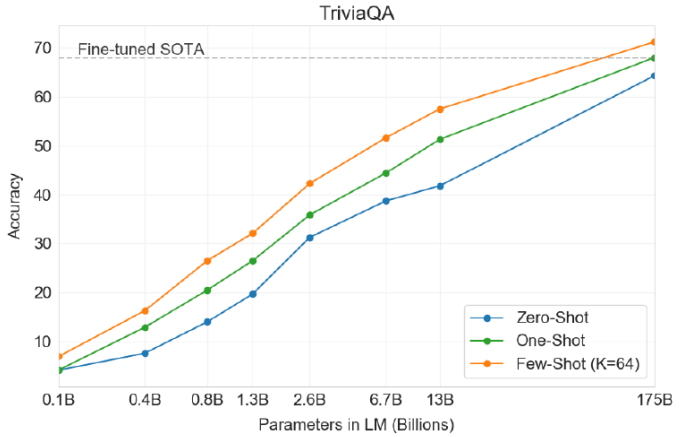


Figure 3.3: On TriviaQA GPT3’s performance grows smoothly with model size, suggesting that language models continue to absorb knowledge as their capacity increases. One-shot and few-shot performance make significant gains over zero-shot behavior, matching and exceeding the performance of the SOTA fine-tuned open-domain model, RAG [LPP+20]

- 기계번역

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

Table 3.4: Few-shot GPT-3 outperforms previous unsupervised NMT work by 5 BLEU when translating into English reflecting its strength as an English LM. We report BLEU scores on the WMT'14 Fr↔En, WMT'16 De↔En, and WMT'16 Ro↔En datasets as measured by `multi-bleu.perl` with XLM's tokenization in order to compare most closely with prior unsupervised NMT work. SacreBLEU[f] [Pos18] results reported in Appendix H. Underline indicates an unsupervised or few-shot SOTA, bold indicates supervised SOTA with relative confidence. [a][EOAG18] [b][DHKH14] [c][WXH+18] [d][oR16] [e][LGG+20] [f][SacreBLEU signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.intl+version.1.2.20]

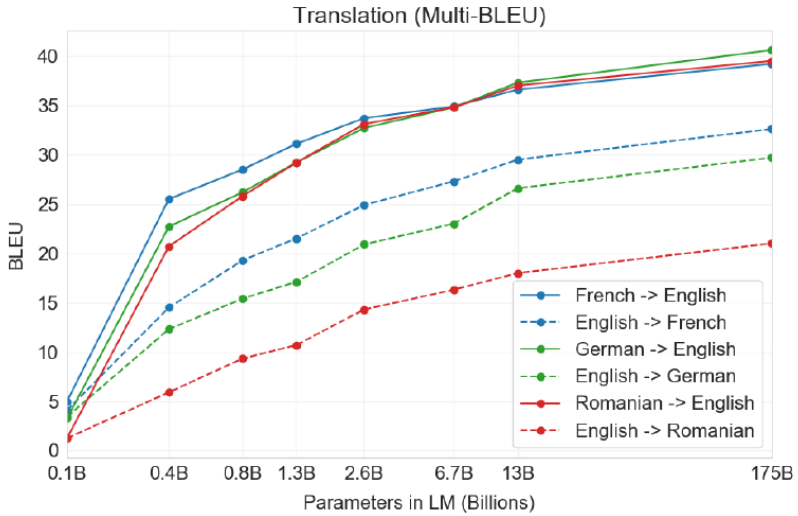
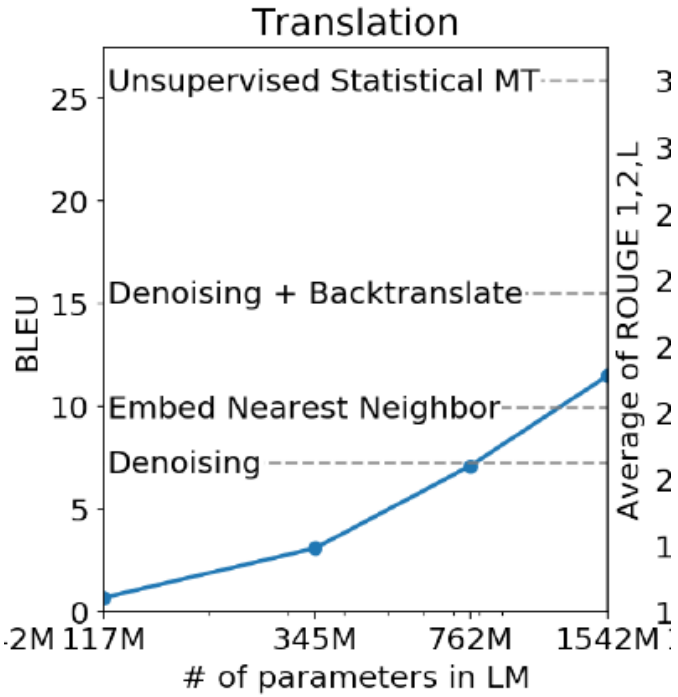


Figure 3.4: Few-shot translation performance on 6 language pairs as model capacity increases. There is a consistent trend of improvement across all datasets as the model scales, and as well as tendency for translation into English to be stronger than translation from English.

- 기계번역: GPT-2



* 비고: GPT-2 (Fr → En)

"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.

한계점

- (1) in-context learning으로 새로운 task를 추론 여부의 uncertainty

A limitation, or at least uncertainty, associated with few-shot learning in GPT-3 is ambiguity about whether few-shot learning actually learns new tasks "from scratch" at inference time, or if it simply recognizes and identifies tasks that it has learned during training. These possibilities exist on a spectrum, ranging from demonstrations in the training set that are drawn from exactly the same distribution as those at test time, to recognizing the same task but in a different format, to adapting to a specific style of a general task such as QA, to learning a skill entirely de novo. Where GPT-3 is on this spectrum may also vary from task to task. Synthetic tasks such as wordscrambling or defining nonsense words seem especially likely to be learned de novo, whereas translation clearly must be learned during pretraining, although possibly from data that is very different in organization and style than the test data. Ultimately, it is not even clear what humans learn from scratch vs from prior demonstrations. Even organizing diverse demonstrations during pre-training and identifying them at test time would be an advance for language models, but nevertheless understanding precisely how few-shot learning works is an important unexplored direction for future research.

한계점

- (2) still has notable weaknesses in text synthesis and several NLP tasks.
  - still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs.
  - it does little better than chance when evaluated one-shot or even few-shot on some “comparison” tasks, such as determining if two words are used the same way in a sentence, or if one sentence implies another (WIC and ANLI respectively), as well as on a subset of reading comprehension tasks.

한계점

- (3) several structural and algorithmic limitations
  - do not include any bidirectional architectures or other training objectives such as denoising
- (4) limits of the pretraining objective
  - current objective weights every token equally and lacks a notion of what is most important to predict and what is less important.
  - ultimately, useful language systems (for example virtual assistants) might be better thought of as taking goal-directed actions rather than just making predictions
  - not grounded in other domains of experience, such as video or real-world physical interaction
- (5) Others
  - poor sample efficiency during pre-training
  - both expensive and inconvenient to perform inference
  - its decisions are not easily interpretable

생각해 볼 내용

- 1) 언어모델 approach로 어디까지 할 수 있을까?
  - next word prediction / masked word prediction
  - 다음 단어 또는 masked 단어를 맞추기 위해 알아야 하는 상식, 추론까지 학습할 수 있을까?

- 2) Pre-training 이후, Fine-tuning이 아닌 (continual) Post-training 필요
  - Weight Update 없는 in-context learning의 한계
    - weight update가 없다는 것은 모델에 새로운 지식 학습이 없다는 것
  - 사람과 같이 Multi-task 기본 능력을 유지하면서 새로운 지식 학습 방법 필요

생각해 볼 내용

- 3) 언어모델의 확장 방향
  - 알고리즘 개선 (현재 단방향, 단순 supervised learning)
  - 외부 메모리와 연동 방법 (open 지식 활용)
    - 시기에 따라 답이 달라지는 문제에 대응 불가 (예: 우리나라의 대통령은?)
  - 모달리티 확장 방법

- 4) more bigger와 다른 연구 방향 필요
  - Restricted computation, data 환경에서의 효율적 LM 모델링

- 5) GPT-3의 다음은?
  - OpenAI가 생각하는 GPT-3 이후의 질문은?
  - 구글, 페이스북, AllenAI는 GPT-3를 보고 어떤 질문을 할까?
  - i.e.: 충분히 큰 BERT 모델의 퓨샷 학습 성능은?

생각해 볼 내용

- 실험 결과에 대한 개인적 질문
  - 1) 요약 태스크 평가 미수행
  - 2) 퓨샷 예제에 따른 성능 차이 (안정적?)
  - 3) 성능이 saturation 되는 예제 수량

감사합니다