8주차 NLP adv 논문 리뷰 과제

GPT-3 : Language Models are Few-Shot Learners

- GPT1 - Improving Language Understanding by Generative Pre-Training (2018)
- GPT2 - Language Models are Unsupervised Multitask Learners (2019)
- **GPT3 - Language Models are Few-Shot Learners (2020)**
- GPT4 - Technical Report (2023)

현재 GPT-4까지 나온 시점에서 가장 오랜 기간 동안 각광받았던, GPT-3에 대한 논문을 리뷰하고자 한다.

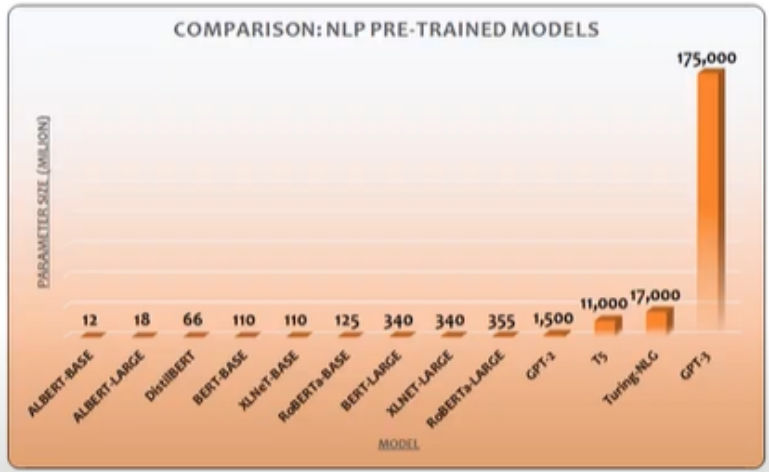
GPT : Genverative Pretrained Transformer (GPT uses Decoder part only.)

> 💡 **GPT3 : fine tuning을 배제하고 few-shot learning을 통해 고성능의 다용도 자연어 처리 모델을 만들고자 한다.**

0. Abstract

- 최근 NLP 엄청난 발전을 이뤄냈다. 예전 잘 훈련된 word embedding을 재사용했던 시대를 지나 이제는 잘 훈련된 NLP모델을 다운 받아서 개인의 task에 맞게 Fine-tuning을 통해 사용하는 시대이다.
- pre-training과 task-specific Fine-tuning을 이용한 모델들을 task-agnostic이라고도 불린다.
- 그러나 Fine-tuning과정에서 대용량의 dataset을 수집해야 하고, 레이블링, 추가 레이어를 넣고 등등 한계점이 존재한다.
- 이 논문에서는 언어모델의 사이즈를 대폭 키워, few-shot learning 모델을 통해 현재 task-agnostic SOTA와 맞먹을 정도이다.

> GPT는 transformer의 decoder 구조를 가지고 있고, BERT 모델은 transformer의 Encoder 구조를 가지고 있다.



gpt3의 파라미터 수는 동시대에 가장 높게 나왔던 MS사에서 출시한 T-NLC 모델에 비해 10배 이상의 수를 가진다.

- few shot learning 이전에 zero shot learning, one shot learning이 존재한다.

1. Introduction

문제

- 현재 사람들은 이미 잘 훈련된 자연어 처리 모델을 다운로드 받아서 개개인의 task에 맞게 Fine-tuning을 거쳐서 사용한다. 하지만 이러한 모델은 task-agnostic한 방식을 채택하고 있기 때문에, task에 대한 수많은 데이터셋과 Fine-tuning 과정을 필요로 한다.

1. 모든 새로운 task에 대해서 labeling된 data를 필요로 하기 때문에, 언어 모델의 확장성을 제한한다.
2. 인간은 새로운 언어에 대해서 학습할 때, 많은 데이터를 필요로 하지 않는다. (즉, task-agnostic은 새로운 언어를 배우는 데에 있어서 많은 자본과 데이터를 필요로 한다.)

모델 소개

- 이 논문에서는, 175 billion의 parameter를 사용하는 "GPT-3"이라는 language model을 실험하여, model의 parameter가 커질수록 성능이 향상됨을 확인한다.
- 또한, GPT-3을 각각 few-shot learning, one-shot learning, zero-shot learning을 통해 학습하고, 비교해 본다.
- *zero shot learning*

예시 : 얼룩말의 사진을 예측하고자 한다. 말과 젖소의 사진을 학습 시킨다. 그리고 얼룩말을 젖소의 색깔을 가진 말이라고 학습시킨다. 그리고 얼룩말 사진 보여줬을 때, 얼룩말이라고 예측하는 것이다. 이 과정에서 얼룩말 사진은 학습되지 않았다. (zero shot)

- *one shot learning*

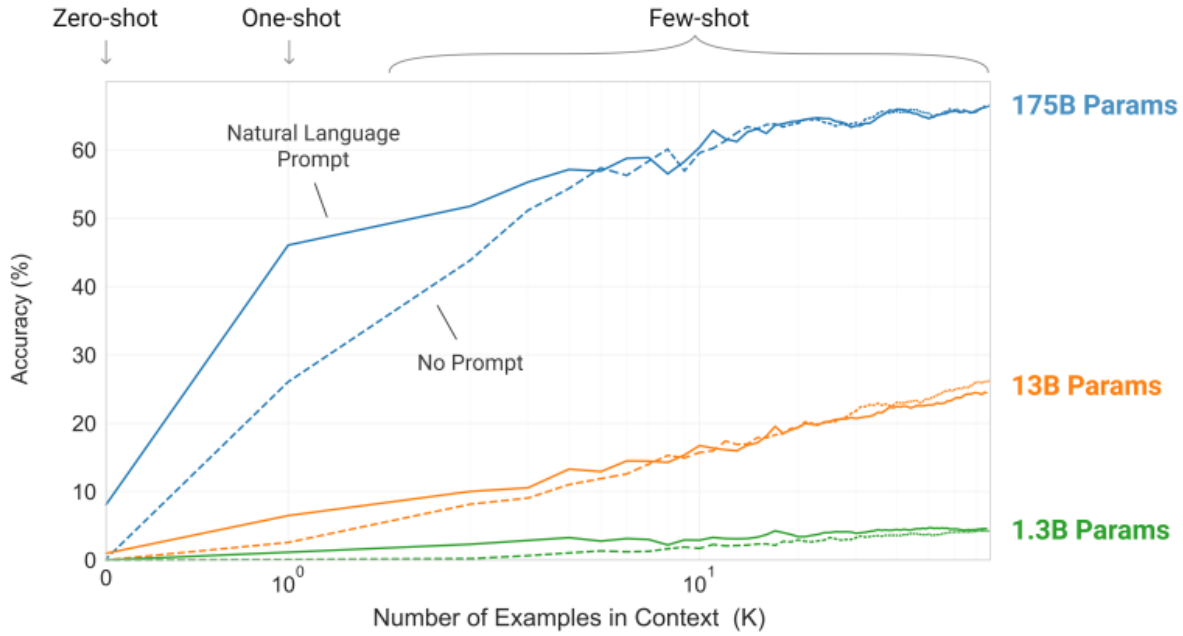
one shot learning은 zero shot learning 보다 쉬운 학습 방법이다. 한 장의 이미지를 학습시키고, 이와 관련된 다른 이미지를 보여줬을 때 예측하는 것이다.

- *few shot learning*

few shot learning은 one shot learning에 더 발전된 학습 방법으로 한 장 이상의 이미지를 학습시키는 것이다.

사칙연산, 오타 검색 번역 등의 pattern few shot learning으로 학습

실험 결과

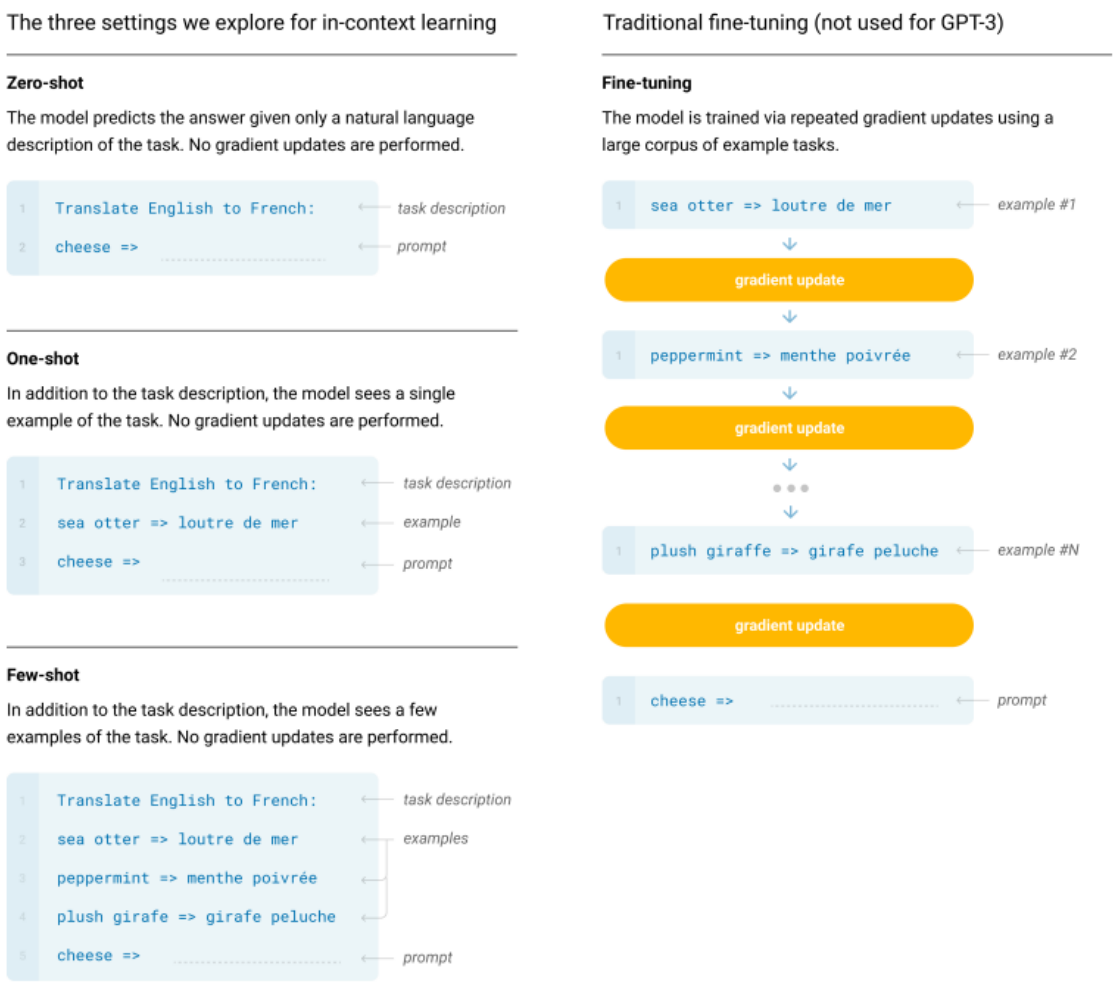


few shot에서 단어에서 관계없는 symbol을 지우는 간단한 task를 수행해 보았을 때, task에 대한 설명이 많을수록(zero to few) 성능이 향상된 것을 볼 수 있다. 또한 모델의 파라미터가 많을수록 성능이 급격하게 향상되는 것을 확인할 수 있다.

---

2. Approach

- model, data, training을 포함한 pre-training 과정은 model의 크기, dataset의 다양성, 길이, 크기가 커졌다는 것을 제외하곤 GPT-2의 방법과 유사하다.



- zero shot : task만 부여 // one shot : task + 1개의 example // few shot : task + 여러 개의 example // Fine-tuning : 이 과정에서 비싼 연산이 들어가게 된다.(각 task 학습시, 수 많은 data 요함)

2-1 Model and Architectures

- GPT-3 모델은 GPT-2와 같은 모델과 아키텍처를 사용했다.
- model size와 성능 간의 상관관계 확인을 위해, 8가지 다른 size의 모델을 사용했다.

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}

2-2 Training Dataset

- 데이터가 풍부한 Common Crawl Dataset을 사용했다. 하지만 필터링 않은 데이터가 많이 섞여있기 때문에, 학습 데이터의 품질 향상을 위해 3가지 방법을 추가했다.

1. Common Crawl Dataset에서 high-quality reference corpora와 비슷한 데이터들을 다운로드했다.
2. fuzzy duplication at the document level (중복 제거)
3. added known high-quality reference corpora (좋은 데이터들을 추가로 데이터셋에 포함)

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

총 300 billion tokens를 수집했지만, 대두사의 데이터가 중복이 없었다.

2-3 Training Process

- large model일수록 큰 batch size를 사용했다. 하지만 적은 learning rate를 필요로 한다.
- 학습 과정에서 gradient noise scale을 측정하여, batch size 선택에 사용했다.
- Out of memory를 막기 위해, model parallelism(분산 학습)을 사용했다.

B Details of Model Training

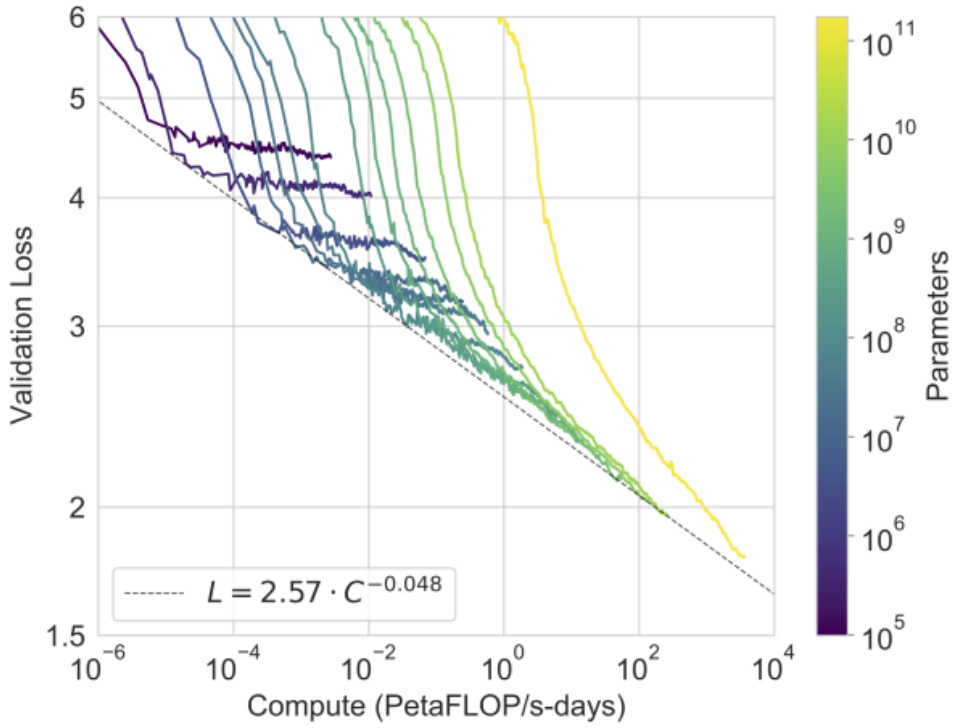
To train all versions of GPT-3, we use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-8}$, we clip the global norm of the gradient at 1.0, and we use cosine decay for learning rate down to 10% of its value, over 260 billion tokens (after 260 billion tokens, training continues at 10% of the original learning rate). There is a linear LR warmup over the first 375 million tokens. We also gradually increase the batch size linearly from a small value (32k tokens) to the full value over the first 4-12 billion tokens of training, depending on the model size. Data are sampled without replacement during training (until an epoch boundary is reached) to minimize overfitting. All models use weight decay of 0.1 to provide a small amount of regularization [LH17].

During training we always train on sequences of the full $n_{\text{ctx}} = 2048$ token context window, packing multiple documents into a single sequence when documents are shorter than 2048, in order to increase computational efficiency. Sequences with multiple documents are not masked in any special way but instead documents within a sequence are delimited with a special end of text token, giving the language model the information necessary to infer that context separated by the end of text token is unrelated. This allows for efficient training without need for any special sequence-specific masking.

---

3. Result

- size가 각기 다른 8개의 GPT03 모델의 learning curve를 비교했을 때, size가 큰 모델일수록 언어 모델의 성능이 향상됨을 확인할 수 있다.
- 이 과정에서, traigning compute와 performance는 power-law를 따른다고 알려졌는데, 모델 size가 일정 수준 이상에서는 power-law의 기댓값보다 더 좋은 성능을 보였다.
- 이것이 training dataset을 cross-entropy를 통해 학습해서(외워버려서) 그런 것 아닐까 하는 의심이 들 수도 있지만, cross-entropy가 다양한 NLP 분야의 task에서 일관적으로 성능 향상을 보임을 보인다.

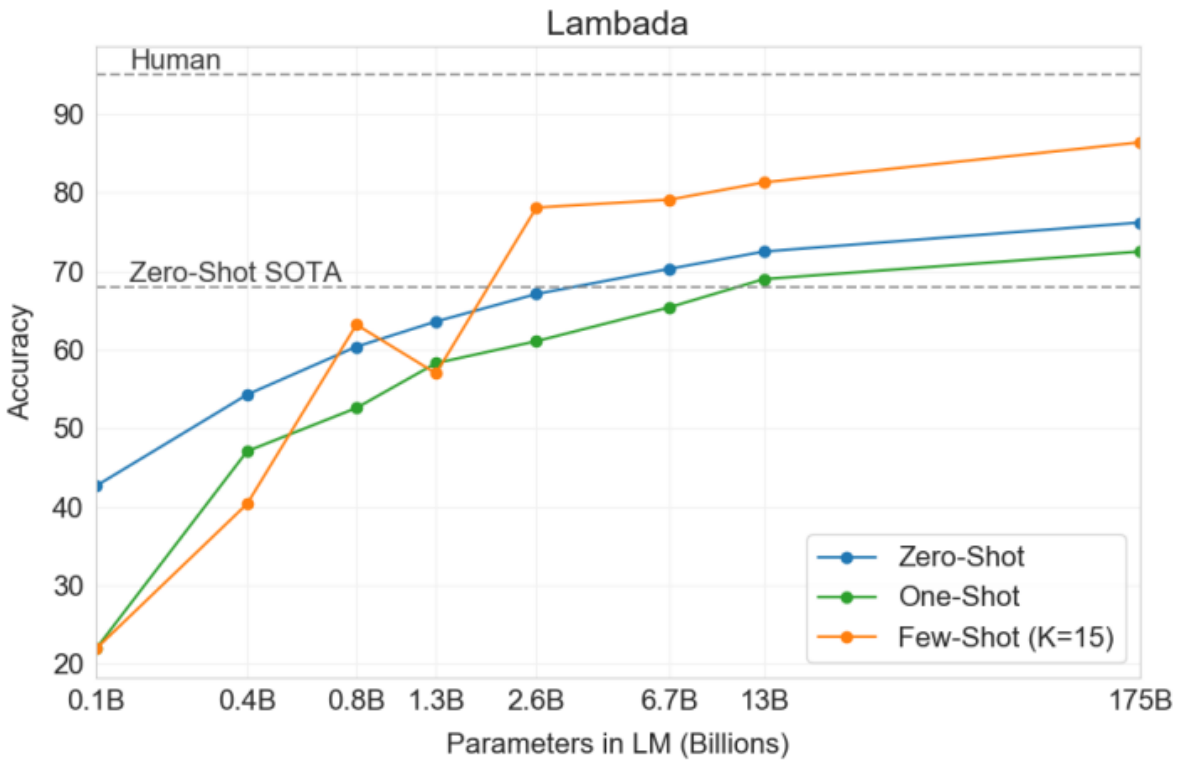


3-1 Language Modeling

- GPT-3 모델이 SOTA(기존 최고 성능평가) 기록을 압도했다.

3-2 LAMBADA

- LAMBADA dataset은 text 내에서 long-range dependency를 테스트 하는 것이다. (context를 읽고, sentence의 마지막 word를 예측하는 문제)
- SOTA 보다 18%나 앞선 좋은 성능을 보였다.



3-3 HellaSwag

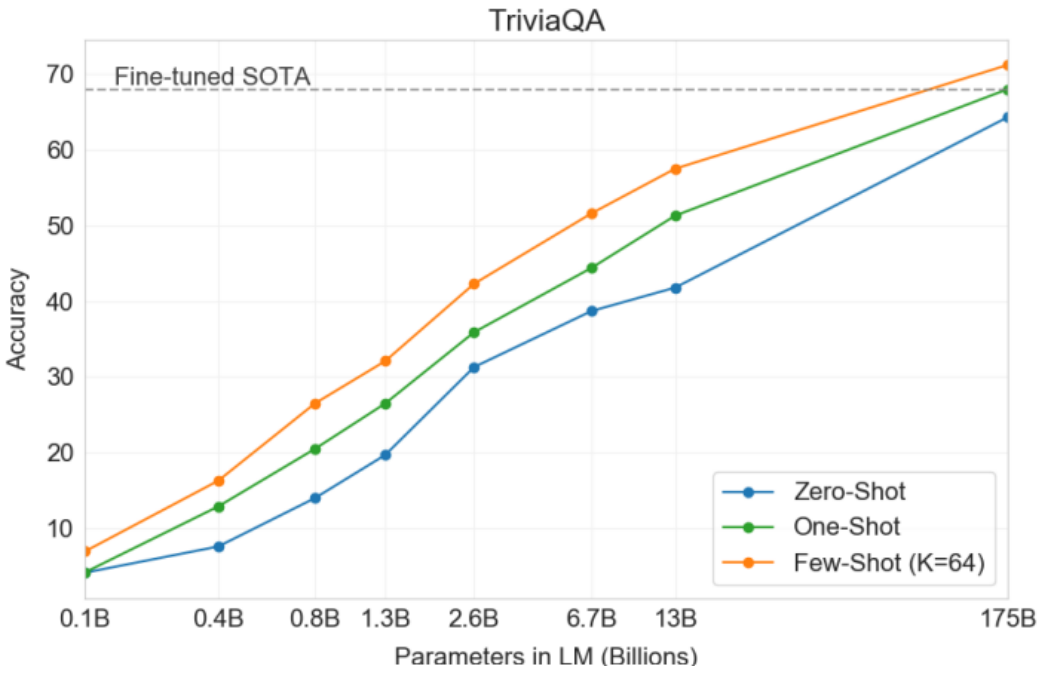
- HellaSwag dataset은 story와 몇 가지 instruction이 주어지면 가장 알맞는 ending을 뽑는 문제
- StoryCloze dataset은 story에 따른 가장 그럴듯한 ending sentence 뽑는 문제
- 두 문제 둘 다 SOTA보다는 성능이 떨어지지만, Fine-tuning 없이도좋은 성능을 뽑아냈다.

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3

3-4 Closed Book Question Answering

- Closed Book Question Answering은 다양한 지식에 대한 context가 없는 답변을 생성하는 문제이다.

- Model size가 커졌을 때, SOTA를 뛰어넘는 성능을 보여준다.



3-5 Translation

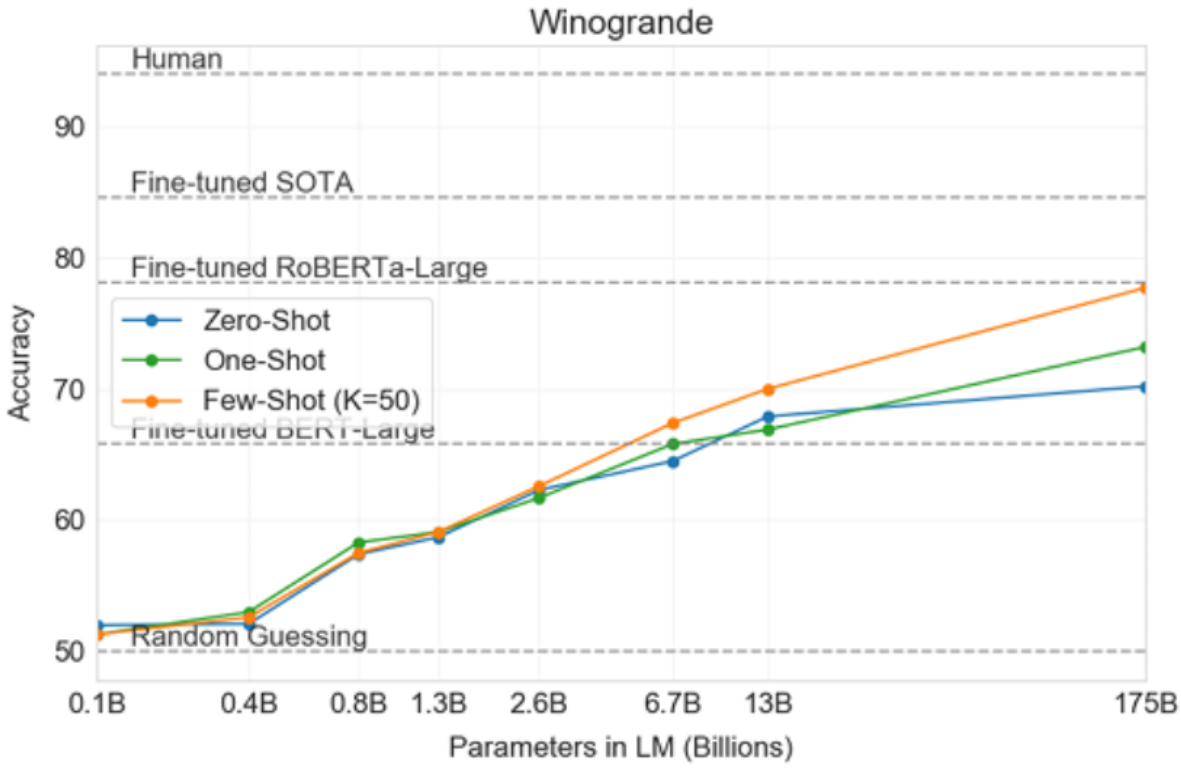
- 번역 모델을 위한 Fine-tuning 데이터는 엄청난 양이다. 하지만 GPT-3에서는 Fine-tuning 없이도 높은 성능을 보였다.
- Translation의 학습에서는 93% 텍스트가 영어고 7%만 다른 언어들을 포함했다. 별도의 목적함수를 사용하지는 않았다. (그냥 구분 없이 똑같이 학습했다.)
- 프랑스어 → 영어 // 독일어 → 영어에 대한 성능평가에서는 SOTA를 뛰어넘기도 했다.

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

3-6 Winograd-Style Task

- 해당 task는 대명사가 어떤 단어를 가리키는지 맞추는 문제 (추론 능력을 확인하는 문제이다)
- SOTA에 비해서는 낮은 성능평가 점수를 보였다.
- 그 이유를 생각해봤을 때, 양방향이 아닌 단방향으로만 정보를 습득하는 GPT의 기본 구조상 어려운 문제일 수 있다.
- 그러나 RoBEATa에 근접할 정도로 좋은 성능을 보였다.

Setting	Winograd	Winogrande (XL)
Fine-tuned SOTA	**90.1**[a]	**84.6**[b]
GPT-3 Zero-Shot	88.3*	70.2
GPT-3 One-Shot	89.7*	73.2
GPT-3 Few-Shot	88.6*	77.7



3-7 Common Sense Reasoning

- 일반 상식 task에서는 특정 dataset에서 SOTA를 뛰어넘었지만, 몇몇 dataset에서는 낮은 성능을 보였다.

Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	**92.0**[KKS+20]	**78.5**[KKS+20]	**87.2**[KKS+20]
GPT-3 Zero-Shot	**80.5***	68.8	51.4	57.6
GPT-3 One-Shot	**80.5***	71.2	53.2	58.8
GPT-3 Few-Shot	**82.8***	70.1	51.5	65.4

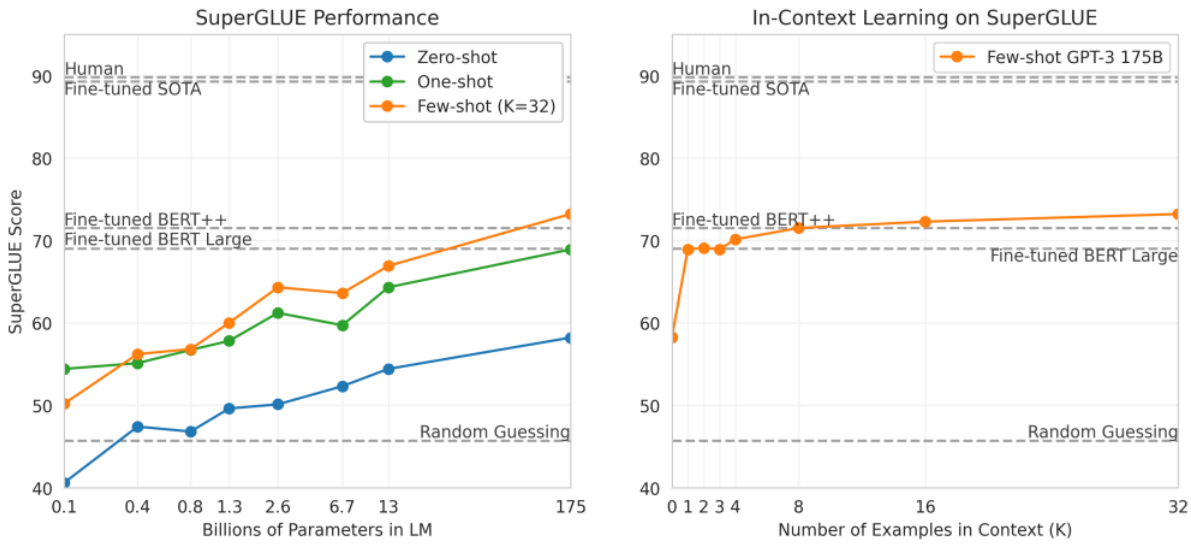
3-8 Reading Comprehesion

- 독해 능력에 대한 task는 아직 이 자연어 처리 모델이 사람만큼의 능력을 발휘하는 데에는 한계가 있다.
- GPT-3 또한 성능이 낮은 결과를 보인다.
- CoQA를 제외하고는 기존 SOTA와 매우 큰 차이를 보인다.

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

3-9 Super GLUE(벤치마크 dataset)

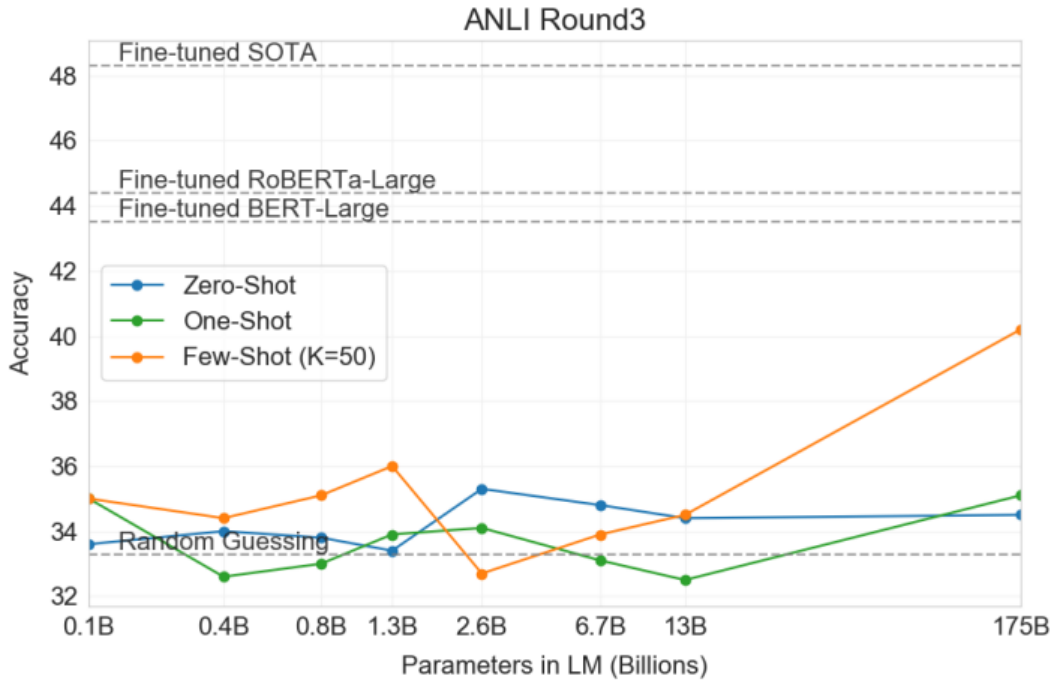
- BERT와 비교 할 때, SuperGLUE를 test한다.
- SOTA에는 미치지 못하지만, 대부분 BERT 성능 이상을 보인다.



3-10 NLI

- Natural Language Inference : 두 문장의 상관관계를 이해하는 task

- 이 task에서도 좋은 성능을 보이지 못했다.



3-11 Arithmetic

- GPT모델은 language 모델임에도 불구하고, 간단한 사칙연산을 few shot learning을 통해서 배워서 두 자리수 덧셈의 경우 한치의 오차도 없었다.

뉴스 기사 만들기

- GPT-2에서는 우수한 글짓기를 보여줬기에 GPT-3에서는 뉴스 기사를 만들어 보는 실험을 했다.

	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%

52% 정도의 성능을 보인 것으로 확인할 수 있다.

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%).

가장 사람처럼 쓴 기사 (12%에 대한 사람들만 이 기사가 사람이 쓰지 않았다라고 보였다. 88%가 사람처럼 썻다라고 판단했다.)

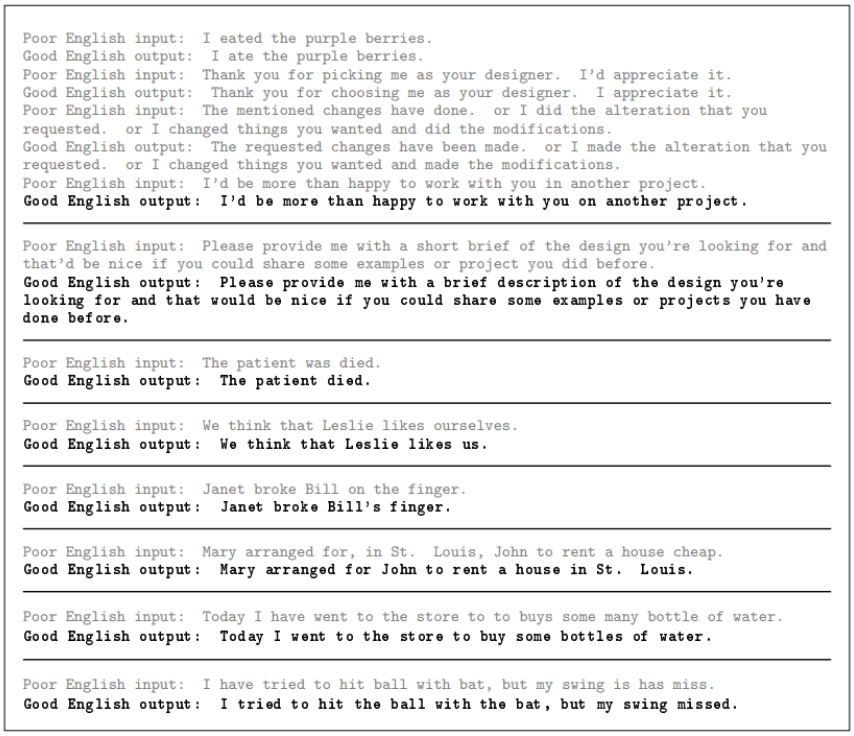
Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. 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."

Figure 3.15: The GPT-3 generated news article that humans found the easiest to distinguish from a human written article (accuracy: 61%).

이 기사가 가장 기계처럼 쓴 기사라고 한다(61%) 하지만 실제로 정말 구별하기 어려운 것 같다... (확실한 건 나보다는 글을 훨씬 잘 쓰는 것 같다..)

Correcting English Grammer

- 조금의 예제만 부여했음에도 불구하고, 문법이 틀린 문장에 대해서 높은 정확도로 문법을 교정해주는 것을 확인할 수 있다.



5. Limitations

- 성능적 한계 : 몇 가지 NLP task에 대해서는 좋지 않은 성능을 보여준다.
- 구조 및 알고리즘의 한계 : GPT-3는 bidirectional 구조나 denoising 같은 NLP 분야의 성능을 향상하는 방법들을 고려하지 않았다.
- 비싼 연산값 : GPT-3의 parameter가 매우 많기 때문에, training & inference cost가 매우 크다.
- 가장 큰 문제는 GPT가 단방향으로 정보를 습득하다 보니까, 양방향 취득하는 모델에 비해서 문장 전체에서 문맥을 파악하는데 취약한 단점이 있다.

→ 이 문제점은 이번에 나온 GPT-4에서 극복했다.

++ GPT-3 모델의 편향성 분석

- GPT-3가 다음 단어를 예측할 때, 주어진 단어에 따라 성별, 종교, 인종들에 대해서 편향성을 가지고 있다.

Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

남성 또는 여성으로 편향되는 단어 목록

결론

💡 GPT-3는 아직 완벽한 모델은 아니다. Fine-Tuning 없이도 몇몇 자연어 처리 성능 평가에서는 기존 최고의 점수를 능가하고, 몇몇 성능 평가에서는 버금가는 성능을 보여준다. 아직 자연어 처리 머신러닝에 적극 활용되지 않던 few shot learning을 통해서 인간처럼 아주 적은 data로 좋은 성능을 보여줬다. 또한 zero shot learning 보다는 one shot learning이, one shot learning 보다는 few shot learning이 더 좋은 성능을 보여주고, 단연 파라미터가 더 많은 모델이 성능이 더 좋다는 점을 보여준다. 나는아직 GPT-4에 대해서는 논문을 확인하지 않았다. GPT-3도 굉장히 훌륭하지만 이에 대한 한계점을 극복한 GPT-4 모델이 더욱 기대된다.

▼ 참고자료

유튜브

1. https://www.youtube.com/watch?v=p24JUVgDkQk&t=393
2. https://www.youtube.com/watch?v=xNdp3_Zrr8Q&t=243s

티스토리

1. https://devhwi.tistory.com/35
2. https://littlefoxdiary.tistory.com/44