CS 324
Large Language Models
(Winter 2022)

1. Introduction
2. Capabilities

발제자 : 박현빈

phyeonbin01@gmail.com

CS 324의 목적

- students will learn the fundamentals about the modeling, theory, ethics, and systems aspects of large language models, as well as gain hands-on experience working with them.

1. Introduction

1. What is a language model?
2. A brief history
3. Why does this course exist?
4. Structure of this course

What is a language model?

- Classic definition : 토큰 시퀀스에 대한 확률 분포
  - Input : 토큰들의 시퀀스
  - Output : 시퀀스의 확률. $p(x_{1:L})$

$$p(\text{the, mouse, ate, the, cheese}) = 0.02,$$

$$p(\text{the, cheese, ate, the, mouse}) = 0.01,$$

$$p(\text{mouse, the, the, cheese, ate}) = 0.0001.$$

Autoregressive Language Models

- 확률 $p(x_{1:L})$ 구하는 방법 : Chain Rule

$$p(x_{1:L}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_L \mid x_{1:L-1}) = \prod_{i=1}^{L} p(x_i \mid x_{1:i-1})$$

$$\begin{aligned} p(\text{the, mouse, ate, the, cheese}) = & \; p(\text{the}) \\ & p(\text{mouse} \mid \text{the}) \\ & p(\text{ate} \mid \text{the, mouse}) \\ & p(\text{the} \mid \text{the, mouse, ate}) \\ & p(\text{cheese} \mid \text{the, mouse, ate, the}) \end{aligned}$$

Autoregressive Language Models

- Generation

$$\text{for } i = 1, \ldots, L:$$
$$x_i \sim p(x_i \mid x_{1:i-1})^{1/T}$$

T : Temperature

T = 0 : 매번 확률 가장 높은 토큰 선택

T = 1 : 분포 p에서 sampling

T = ∞ : 모든 토큰의 확률이 같아 짐

T로 인해 확률의 합이 1이 안 되므로 다시 한 번 Normalization을 해야 한다.

A brief history

- Information Theory
  - 정보를 효율적으로 표현하고 전달하는 방법을 연구하는 학문
  - Entropy (정보량, 불확실성)

$$H(p) = \sum_x p(x) \log \frac{1}{p(x)}$$

the mouse ate the cheese $\Rightarrow 0001110101$

If $p(x) = \frac{1}{8}$, we should allocate $\log_2(8) = 3$ bits

A brief history

- Cross-Entropy
  - p는 “True” distribution
  - q는 x를 압축하는 모델
  - 샘플링은 모델 p에서, 압축 방식은 모델q
  - 모델 q가 p랑 비슷할수록 H(p, q)는 H(p)와 유사해짐
  - 모델 q는 N-gram 또는 사람이 될 수 있음

$$H(p,q) = \sum_x p(x) \log \frac{1}{q(x)}$$

A brief history

- N-gram
  - i번째 token은 이전 n-1개 토큰에 의해서만 결정된다.

$$p(x_i \mid x_{1:i-1}) = p(x_i \mid x_{i-(n-1):i-1})$$

$$p(\text{cheese} \mid \text{the, mouse, ate, the}) = p(\text{cheese} \mid \text{ate, the})$$

  - 장점
    - 계산 cost가 적기 때문에 방대한 데이터로 학습 가능하다.
    - 5-gram 모델 훈련 데이터 : 2 trillion tokens
    - GPT-3 훈련 데이터 : 300 billion tokens
  - 단점
    - 시퀀스 전체를 볼 수 없어 문맥을 온전히 이해할 수 없다

A brief history

- Neural Language Model
  - RNN
  - Transformer

- 초창기에는 비용이 너무 높아 N-gram을 더 많이 사용

Why does this course exist?

- 왜 **Large** Language Model을 배워야 하는가?
  - 4년만에 모델 크기가 5000배(94M □ 530B) 커졌다.
  - Scale 상승은 모델 성능 향상으로 이어졌다.
  - Prompt를 입력 받고, 이어서 작성하는 능력이 뛰어남
  - In-context learning 성능이 우수하다(GPT-3)
  - Real-world에 많이 사용된다

Structure of this course

1. Behavier : 모델이 어떻게 동작하는지
2. Data : LLM 훈련에 사용되는 데이터 탐구
3. Building : 모델 아키텍처, 훈련 알고리즘

2. Capabilities

Capabilities of GPT-3

Adaptation

- 정의 : Language model □ Task model
- 방법
  - Training(Supervised Learning)
    - Fine-tuning
    - 데이터가 적은 상황에서는 오버피팅
  - Prompting(In-context Learning)
    - Zero-shot
    - One-shot
    - Few-shot
    - Context window 크기만큼 데이터 사용할 수 있음

Explore the capabilities of GPT-3

- GPT-3 (davinci) 175B parameters
- 다양한 task로 Adaptation을 위해 In-context learning 수행
- Task
  1. Language Modeling
  2. QA
  3. Translation
  4. Arithmetic
  5. News Article Generation
  6. Novel tasks

Task 1. Language Modeling

- Language Model : 시퀀스의 확률을 예측하는 모델 p
  - 확률 예측 방법 : Chain Rule

$$p(x_{1:L}) = \prod_{i=1}^{L} p(x_i \mid x_{1:i-1}).$$

  - 문제점 : 시퀀스 길이가 길어질수록 확률은 0에 가까워진다
  - Perplexity : 한 단어를 예측할 때 평균적으로 몇 개의 선택지 중 하나를 고르고 있는지를 나타냄. 낮을수록 좋음

$$\text{perplexity}_p(x_{1:L}) = \exp\left(\frac{1}{L}\sum_{i=1}^{L}\log\frac{1}{p(x_i \mid x_{1:i-1})}\right)$$

Task 1. Language Modeling

- 발생할 수 있는 문제점
  - 언어 모델이 어떤 토큰에 확률 질량을 할당하지 못하면 분모가 0이 되고 perplexity는 무한으로 간다

    $p(\text{ate} \mid \text{the, mouse}) \to 0 \quad \Rightarrow \quad \text{perplexity}_p(\text{the, mouse, ate, the, cheese}) \to \infty.$

  - 잘못된 시퀀스에 추가적인 확률 질량을 할당하면 모델의 생성 능력이 많이 떨어지는 것에 비해 perplexity는 조금만 증가한다

    $$q(x_i \mid x_{1:i-1}) = (1-\epsilon)p(x_i \mid x_{1:i-1}) + \epsilon r(x_i \mid x_{1:i-1})$$

    $$\text{perplexity}_q(x_{1:L}) \leq \frac{1}{1-\epsilon}\text{perplexity}_p(x_{1:L}) \approxeq (1+\epsilon)\text{perplexity}_p(x_{1:L})$$

    - ε : 확률, r : garbage distribution
    - ε = 5%일 때 모델은 평균적으로 20개 토큰 중 1개는 이상한 단어를 생성하여 이상한 문장이 생성되는 것에 비해 perplexity는 5%밖에 오르지 않는다

Task 1. Language Modeling

- Penn Tree Bank (PTB)
  - Wall Street Journal articles
  - Task : 주어진 text를 가지고 perplexity 계산
  - Adaptation : 전체 text를 GPT-3에게 prompt로 줌

    Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.

  - Result : 기존 SOTA를 능가

Model	Perplexity
GPT-3	**20.5**
BERT-Large-CAs1	31.3

Task 1. Language Modeling

2. LAMBADA
   - Task : 문장의 마지막 단어 예측
   - Adaptation

     Fill in blank:

     Alice was friends with Bob. Alice went to visit her friend ___. -> Bob

   - Result : 기존 SOTA를 능가

Model	Perplexity
GPT-3 (few-shot)	**1.92**
SOTA	8.63

Task 1. Language Modeling

3. HellaSwag
   - Task : 목록에서 가장 적절한 선택지를 골라 문장 완성
   - Adaptation

> *Making a cake: Several cake pops are shown on a display. A woman and girl are shown making the cake pops in a kitchen. They ${answer}*

where ${answer} is one of:

1. *bake them, then frost and decorate.*
2. *taste them as they place them on plates.*
3. *put the frosting on the cake as they pan it.*
4. *come out and begin decorating the cake as well.*

Task 1. Language Modeling

3. HellaSwag
   - 객관식에 점수 매기는 방법
     1. $score(x, y) = p(x, y)$
        - 짧은 답변을 선호하는 경향이 있음
        - Chain Rule에 의해 확률을 계속 곱하는데, 그럴수록 0에 가까워지기 때문
     2. $score(x, y) = \frac{p(x,y)}{num-tokens(y)}$
        - 답변 길이에 대한 편향은 해결
        - x가 가지는 문맥과는 관계없이 더 자주 쓰이는 단어나 문장이 등장할수록 더 높은 점수를 가질 수 있다
     3. $score(x, y) = \frac{p(y|x)}{p(y|x_0)}$
        - x_0는 "Answer : " 와 같이 x 문맥과는 관계없으면서도 y랑은 잘 어울리는 문자열
        - 그냥 흔한 답변일수록 낮은 점수를, 문맥과 관련 높을수록 높은 점수를 가짐

Task 1. Language Modeling

3. HellaSwag

- Result

Model	Accuracy
SOTA	**85.6**
GPT-3	79.3

- SOTA 모델은 HellaSwag training set으로 fine-tuning 되어있는 걸 감안했을 때 GPT-3의 성능은 꽤나 인상적으로 보인다.

Task 2. QA

1. TriviaQA
   - Task : trivia(상식) 질문이 주어지고, 이에 답하는 것
   - 원래는 open book challenge이지만, closed book으로 테스트
   - Adaptation

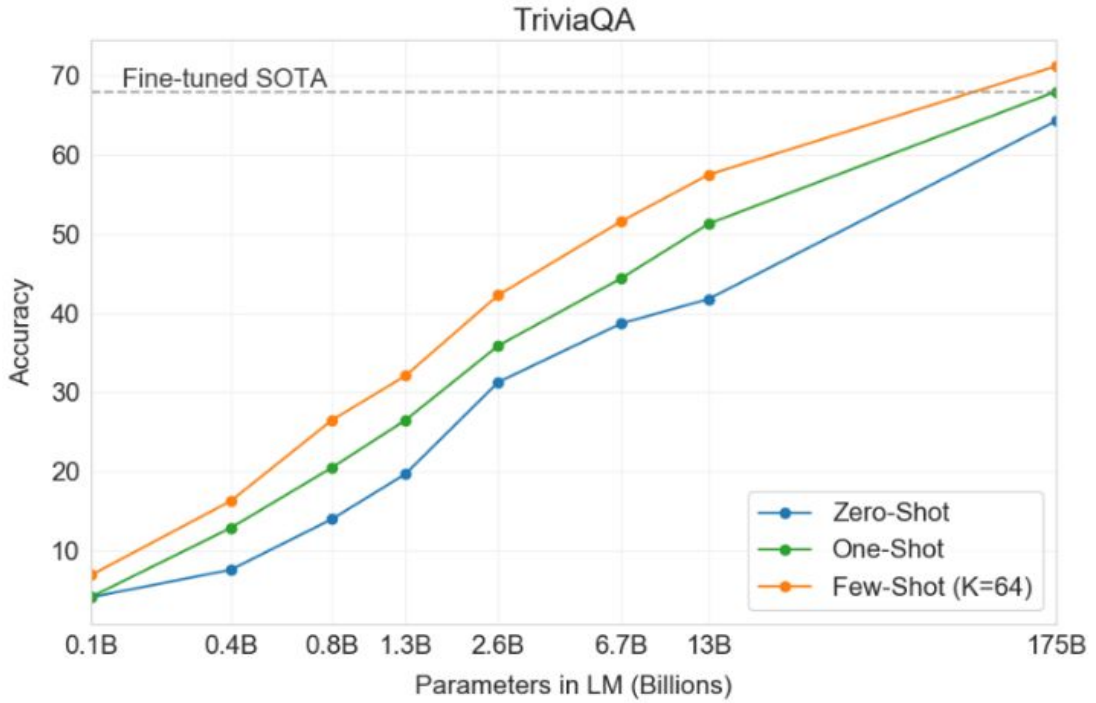
Q: 'Nude Descending A Staircase' is perhaps the most famous painting by which 20th century artist?

A: Marcel Duchamp

Task 2. QA

1. TriviaQA
   - Result
     - RAG보다 높은 정확도
     - 모델 크기와 in-context training instance가 클수록 높은 정확도를 달성

Model	Accuracy
RAG	68.0
GPT-3 (zero-shot)	64.3
GPT-3 (few-shot)	**71.2**



Task 2. QA

2. WebQuestions
   - Google 검색 쿼리로 모은 데이터셋
   - Adaptation

     Q: What school did burne hogarth establish?
     A: School of Visual Arts

   - Result

Model	Accuracy
RAG	**45.5**
GPT-3 (zero-shot)	14.4
GPT-3 (few-shot)	41.5

Task 2. QA

3. NaturalQuestions
   - Google 검색 쿼리로 모은 데이터셋(with long-form answer)
   - Adaptation

     Q: Who played tess on touched by an angel?

     A: Delloreese Patricia Early (July 6, 1931 - November 19, 2017), known professionally as Della Reese.

   - Result

Model	Accuracy
RAG	**44.5**
GPT-3 (zero-shot)	14.6
GPT-3 (few-shot)	29.9

Task 3. Translation

- Standard evaluation dataset : WMT'14, WMT'16
  - 뉴스 번역
  - IT 도메인 번역
  - 생의학 분야 번역
  - 대명사 번역 등

- Evaluation metric : BLEU
  - N-gram 단위로 생성된 문장과 정답 문장을 비교하여 정확도 계산
  - 짧은 출력이 높은 점수를 받는 문제를 방지하기 위해 penalty 적용

Task 3. Translation

- Adaptation

  Mein Haus liegt auf dem Hügel. = My house is on the hill.
  Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden. = In no case may they be used for commercial purposes.

- Result

Model	Accuracy
SOTA (supervised)	40.2
GPT-3 (zero-shot)	27.2
GPT-3 (few-shot)	40.6

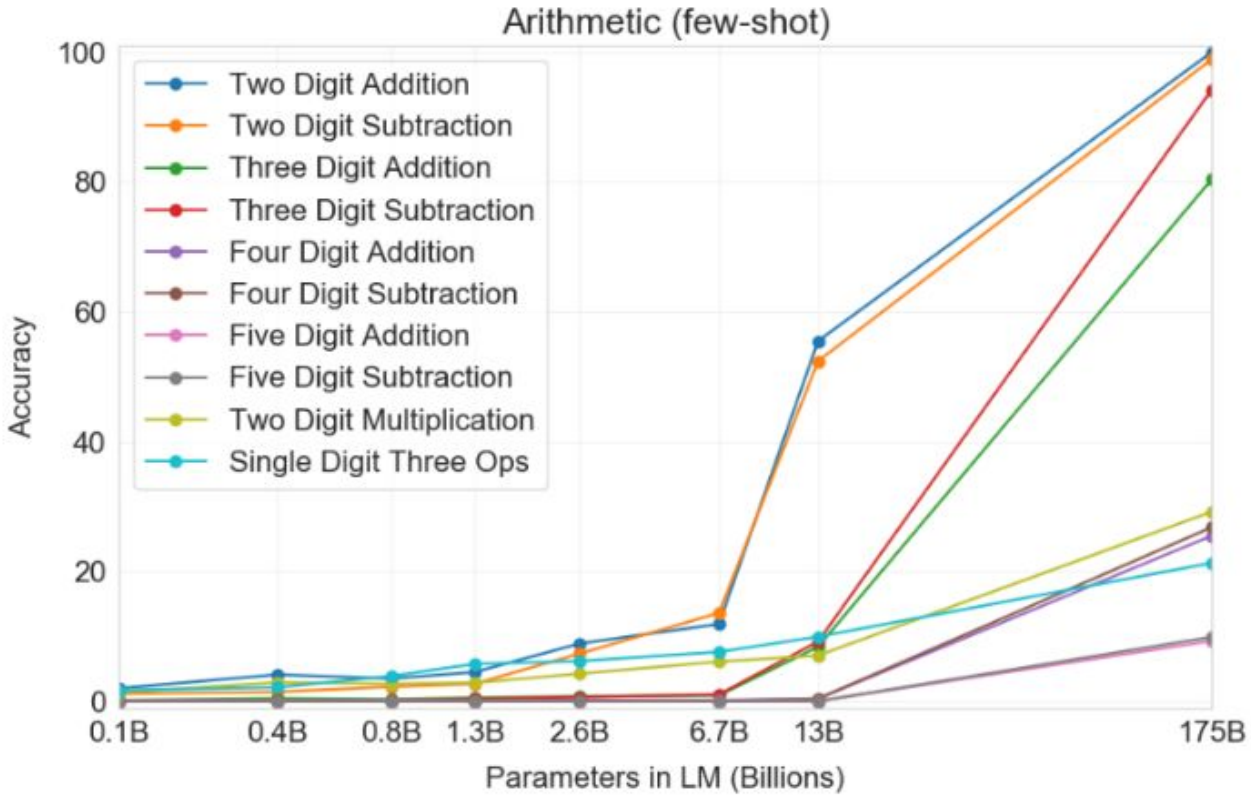
Task 4. Arithmetic(산술)

- Adaptation

Q: What is 556 plus 497?

A: 1053

- Result
  - 숫자를 완벽히 이해하는 건 아니지만 어느정도는 동작한다



Task 5. News Article Generation

- Task : title과 subtitle을 보고 뉴스 기사 생성
- Dataset : newser.com의 title/subtitle
- Evaluation : 사람이 기사를 읽고 기계가 작성됐을 가능성을 평가
- Adaptation

  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...

- Result : 사람이 인간과 기계를 구분해내는 정확도는 52%

Task 6. Novel tasks

- Using new words
  - Task : 새로운 단어를 정의하고, 그 단어를 활용해 문장 생성
  - Adaptation

    To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is: We screeged the tree with our swords.

- Correcting English grammar
  - Task : 문법적으로 틀린 문장을 올바르게 고치기
  - Adaptation

    Poor English input: I eated the purple berries.
    Good English output: I ate the purple berries.
    Poor English input: Thank you for picking me as your designer. I'd appreciate it.
    Good English output: Thank you for choosing me as your designer. I appreciate it.
    Poor English input: I'd be more than happy to work with you in another project.
    Good English output: I would be happy to work with you on another project.

Summary

- 언어모델은 시퀀스의 확률을 예측하는 모델이고, 확률을 바탕으로 generation 할 수 있다.
- 특정 task를 수행하기 위해서는 fine-tuning, prompting과 같은 Adaptation을 해야 한다.
- GPT-3는 In-context learning만으로 다양한 task에서 높은 성능을 기록했다.
- 모델의 크기를 키우거나 예제의 수를 늘리는 것이 성능 향상에 도움이 된다.