Extracting Training Data from Large Language Models

Extracting Training Data from Large Language Models

- https://arxiv.org/pdf/2012.07805.pdf

Author:

- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, Colin Raffel
- Google, Stanford, UC Berkeley, Northeastern University, OpenAI, Harvard, Apple

https://arxiv.org/pdf/2012.07805.pdf

Abstract:

- 훈련 데이터 추출 공격으로 개별 학습 데이터를 복구하는 것이 가능함을 보임
- GPT-2 모델로부터 추출된 예제들이 PII, IRC 대화, code 등을 포함.
  - PII : personally identifiable information
- 하나의 문서에 포함된 시퀀스들을 뽑아내는게 가능
- extraction attack 이 가능하게 하는 factor 를 확인하기 위해 평가

1. Introduction:

- (potentially private) 훈련 데이터에 대한 정보가 유출될 수 있음
  - *membership inference*: 주어진 예제가 훈련 데이터에 있는지 아닌지 예측 가능
  - 주로 오버피팅되는 경우에 발생함
  - LM 은 (중복되지 않은) 거대한 데이터로 1 epoch 정도만 훈련하므로 overfitting 되지 않음
    - 따라서, 정보 유출이 발생하지 않을 것으로 가정함.

논문의 기여:

- Large LM 이 개별 훈련 예제를 기억하고 정보를 유출시킬 수 있음을 입증
  - 오버피팅이 되지 않았더라도 특정 훈련 예제가 기억됨.
- 방법
  - 모델로부터 가능성 높은 샘플 집합을 생성

    - 일반적인 샘플링 정책 3가지중 하나로 생성후 6개의 메트릭중 하나로 정렬
- 결과 분석
    - 모델 크기와 문자열 빈도수가 어떻게 memorization 에 영향을 주는지 분석
    - 다양한 공격의 형태(configuration)이 추출된 데이터 유형을 어떻게 변화시키는지 분석

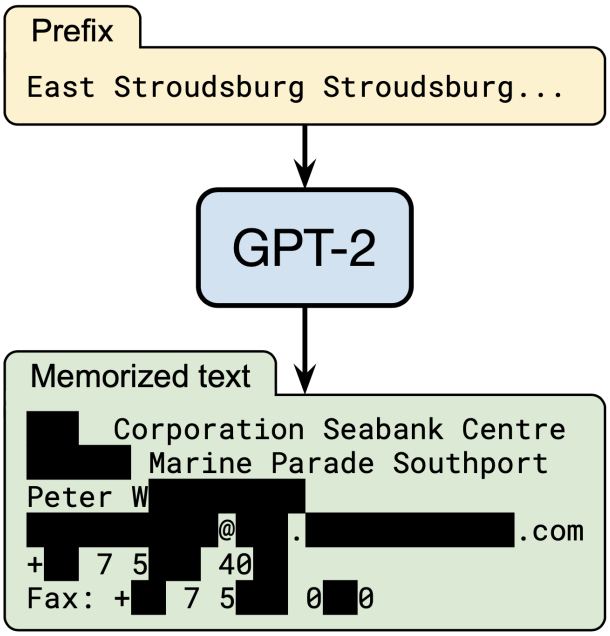

Figure 1: **Our extraction attack**. Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

- Query(Prefix) 가 모델에 주어졌을 때, 이름, 이메일, 전화번호 등의 개인 정보가 추출됨

3. Threat Model & Ethics

- 훈련 데이터 추출 공격이 Practical 하다는 것을 입증
    - "Memorization" 의 의미 정의
    - 위협이 되는 모델과 공격 목표 설명
    - 미래에 심각한 공격이 될 수 있는 이유를 설명

Memorization 정의

- "의도하지 않은"(unintented) Memorization 으로 제한
- *Eidetic memory:*
    - *훈련 데이터에서 출변 빈도가 적음에도 모델에 의해 암기된 데이터(단 한번 본 후 정보를 기억하는 능력)*
    - 모델과 상호작용(prompt)을 통해 문자열이 추출되면 모델이 그 문자열을 안다고 정의

Definition 1 (Model Knowledge Extraction) *A string s is* extractable[4] *from an LM f_θ if there exists a prefix c such that:*

$$s \leftarrow \underset{s' :\, |s'|=N}{\arg\max} f_\theta(s' \mid c)$$

- arg max 는 적당한 샘플링 정책으로 대체될 수 있음.(e.g., greedy sampling)

- 모델에 의해서 추출 가능하고 훈련 데이터에서 K 번 나타나면 *k*-Eidetic Memorization 이라고 정의

Definition 2 (*k*-Eidetic Memorization) *A string s is k-eidetic memorized (for $k \geq 1$) by an LM f_θ if s is extractable from f_θ and s appears in at most k examples in the training data X: $|\{x \in X : s \subseteq x\}| \leq k$.*

- *k* 가 작으면 "의도하지 않은" Memorization 일 가능성이 있고, 긴 문자열일수록 더 해로움.

Threat Model: 위협 정의

- Adversary's Capabilities: (현실적으로 가정)
  - 언어 모델의 입출력 access 가능한 사용자
  - 임의의 문자열 시퀀스에 대한 확률 계산, next-word prediction 이 가능하지만, 모델의 개별 가중치나 히든 상태 벡터를 알 수는 없음.
- Adversary's Objective
  - Memorization 된 훈련 데이터를 추출하는 것
  - 공격의 세기 == 얼마나 Private 한가로 측정(*k-eidetic memorized* 로 정형화)
  - 많은 examples 이 추출될수록, k 의 값이 작을 수록 더 강력한 공격
- Attack Target
  - 공격 대상은 GPT-2
    - 논문에서 추출한 데이터가 이미 모두 공개된 데이터
    - OpenAI 에서 릴리즈한 적이 없으므로 치팅할 수 없음

Risks of Training Data Extraction

- Data Secrecy
  - 기밀 데이터 유출, 개인 정보 유출
    - ex) gmail autocomplete model
- Contextual Integrity of Data
  - 의도한 context 밖에서 데이터를 사용하게할 경우 개인정보 침해임

- Figure 1. 에서 나온 개인 정보는 비밀이 아니지만(의도된 컨텐트스에서 온라인으로 공유된 것), 다른 컨텍스트(e.g. 유저 대화시스템에서 특정 쿼리에 대한 답변)에서 노출할 경우 데이터 무결성 위반

Ethical Considerations

- 논문에서 추출한 데이터중 일부에 개별 사용자에 대한 정보가 포함되므로 윤리적 고려 사항이 있음
- 이미 공개된 데이터와 모델을 사용하여 윤리적 문제를 최소화
- 개인 식별 정보가 추출에 성공할 경우 토큰을 마스킹 처리
- 우리의 방법을 공개하기 위해서 노출된 민감 정보는 사용동의를 받음

> Unfortunately, we cannot hope to contact all researchers who train large LMs in advance of our publication. We thus hope that this publication will spark further discussions on the ethics of memorization and extraction among other companies and research teams that train large LMs.

4. Initial Training Data Extraction Attack

두가지 Step 으로 baseline 을 시작

- Generate text: 모델로부터 대량의 데이터 생성(unconditionally sampling)
- Predict which outputs contain memorized text.
  - Memorized Text 를 포함하지 않는 샘플 제거(by membership inference)

Step1. Initial Text Generation Scheme (샘플링 정책 1)

- 특정한 start-of-sentence 토큰으로 모델을 초기화
- autoregressive 방식으로 샘플 토큰을 반복적으로 생성
- 256 토큰, top-40 strategy

Step2. Initial Membership Inference (정렬 메트릭 1)

- 각 샘플이 훈련 데이터에 존재하는지 여부를 예측(자세한 방법이 궁금하면, [Shokri et al] (https://arxiv.org/pdf/1610.05820.pdf) 참조
  - 모델은 훈련 데이터에 나타난 예제에 대해서 더 높은 confidence 를 부여하는 경향이 있음
- 모델에 의해서 가장 높은 가능성을 가진 샘플을 선택
  - 각 시퀀스의 next token 이 높은 확률로 예측됨

Result. Initial Extraction Results

- Step1 방식으로 GPT-2(XL, 1558M Parameters)로 200,000 샘플 생성
- 모델이 가장 자연스럽게 생성(확률이 높음)하는 순으로 정렬

- MIT public license, Vaughn Live의 사용자 가이드, 트위터 핸들?, 이메일 같은 것들이 발견(k-eidetic 이 높은 것들)
- Initial Approach 의 2가지 약점
  - 샘플링 방식의 output 다양성이 낮다. (중복된 것들이 많이 생성됨)
  - membership inference strategy 가 많은 false positives 를 발생
    - low precision: 높은 likelihood 를 갖지만 memorized 가 아님(e.g. 같은 구가 여러번 반복되는 것)
    - low recall : 낮은 k-memorized content 를 식별하지 못함

5. Improved Traning Data Extraction Attack

더 나은 샘플링과 membership inference 방법 제시

Improved Text Generation Schemes

- 다양한 샘플을 생성할 수 있도록 개선
- Sampling With A Decaying Temperature (샘플링 정책 2)
  - 확률 분포를 평탄하게 만들어서 더 다양한 아웃풋이 생성되도록 조정

> As described in Section 2.1, an LM outputs the probability of the next token given the prior tokens $\mathbf{Pr}(x_i \mid x_1, \ldots, x_{i-1})$. In practice, this is achieved by evaluating the neural network $z = f_\theta(x_1, \ldots, x_{i-1})$ to obtain the "logit" vector z, and then computing the output probability distribution as $y = \text{softmax}(z)$ defined by $\text{softmax}(z)_i = \exp(z_i) / \sum_{j=1}^{n} \exp(z_j)$.
>
> One can artificially "flatten" this probability distribution to make the model less confident by replacing the output $\text{softmax}(z)$ with $\text{softmax}(z/t)$, for $t > 1$. Here, t is called the *temperature*. A higher temperature causes the model to be less confident and more diverse in its output.

  - Temperature 값을 계속 유지하면, Memorized output 이 생성되다가 벗어날 수 잇으므로, 초기 20 토큰 동안 10 to 1로 decay 하면서 적용
  - 모델이 다양한 prefix 를 "explore" 할 수 있고 높은 confidence path 도 따라갈 수 있음
- Conditioning on Internet Text (샘플링 정책 3)
  - 위의 방식을 적용하더라도 추출될 가능성이 낮지만, 실제 데이터로 발생될 수 있는 샘플이 있음.
  - GPT-2 가 학습한 데이터와 유사한 형태의 다양한 prefix 를 가진 샘플을 생성
  - Common Crawl subset 으로 구성
    - GPT-2 학습 데이터와 중복을 최소화

    - 각 샘플의 처음 5~10 개 토큰만 사용하므로 중복될 수 있는 용어가 적음
- 스크랩된 데이터로부터 Context 토큰으로 5~10 개를 랜덤하게 샘플링

Improved Membership Inference

아래와 같은 높은 likelihood 를 가진 샘플들 때문에 필터링의 정확도가 떨어지므로 이를 개선

Trivial memorization:

- 높은 확률로 반복되는 (흥미롭지 않은) 사소한 것들(1~100 까지의 숫자 같은 것들)

Repeated substrings:

- 같은 스트링의 계속된 반복: (e.g., “I love you. I love you. . . ”).

두번째 모델과 비교하여 원래 모델에서 "예상치 않게 높은" likelihood 를 갖는 샘플을 제거함: 4가지 방법 논의

- Comparing to Other Neural Language Models.
    - 다른 데이터로 학습한 LM 모델을 사용하거나 작은 GPT-2 모델(Memorization 능력이 떨어짐)을 사용하여 비교(낮은 k-eidetic 에 대해서 데이터 중복 가능성이 적거나, 기억을 못함)
        - Small GPT-2: 117M (정렬 메트릭 2)
        - Medium GPT-2: 345M (정렬 메트릭 3)
- Comparing to zlib Compression. (정렬 메트릭 4)
    - membership inference metric 으로 GPT-2 perplexity(낮을 수록 좋음) 와 zlib entropy(아마도 낮을수록 반복이 많을 듯?) 사용
        - zlib entropy 로 trivial memorization 과 repeated pattern 을 걸러낼 수 있음(반복적인 패턴들)
- Comparing to Lowercased Text. (정렬 메트릭 5)
    - 같은 모델에서 시퀀스의 "canonicalized" 버전과 perplexity 비교(lowercasing 적용 전후)
- Perplexity on a Sliding Window. (정렬 메트릭 6)
    - non-memorized text 로 둘러쌓인 memorized text 를 위해서 50 token 슬라이드 윈도우를 적용하여 최소 perplexity 를 사용

6 . Evaluating Memorization

Methodology

- 3가지 생성 정책에 따라 256 토큰의 200,000 샘플 생성
    - *Top-n, Temperature, Internet*
- 생성된 샘플을 6가지 membership inference 메트릭으로 정렬
    - *Perplexity, Small, Medium, zlib, Lowercase, Window*

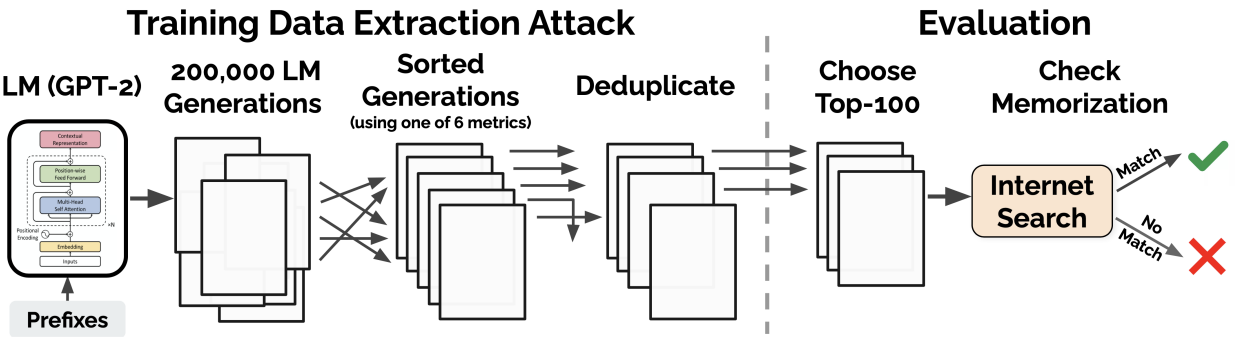

Figure 2: **Workflow of our extraction attack and evaluation.** Attack. We begin by generating many samples from GPT-2 when the model is conditioned on (potentially empty) prefixes. We then sort each generation according to one of six metrics and remove the duplicates. This gives us a set of potentially memorized training examples. **Evaluation**. We manually inspect 100 of the top-1000 generations for each metric. We mark each generation as either memorized or not-memorized by manually searching online, and we confirm these findings by working with OpenAI to query the original training data.

- 3 sampling X 6 metrics = 18 configurations
- 메트릭에 따라 Top 1000 개의 샘플중에서 100개의 샘플 선택: 18 X 100 = 1,800 samples
- 1,800 samples 는 잠재적으로 memorized content

Data De-Deplication

- 100 개 sampling 할때 중복 제거를 위해서 fuzzy de-duplication 단계 적용(*trigram-multiset 이용*)

Evaluating Memorization Using Manual Inspection.

- 1,800 개의 샘플을 저자중 한명이 수동으로 판단(구글링)

Validating Results on the Original Training Data.

- GPT-2 저자를 통해서 memorized 되었다고 생각하는 샘플에 대해서 fuzzy 3-gram match 로 판단

Results

- 1,800 개의 후보 중에서 604 unique memorized 훈련 예제를 확인(positive rate 33.5%)
- best positive rate 67%: internet X zlib

Categories of Memorized Content

- 많은 콘텐츠가 fair 하였으나, 유니크한 데이터와 개인 정보도 존재

Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

Table 1: Manual categorization of the 604 memorized training examples that we extract from GPT-2, along with a description of each category. Some samples correspond to multiple categories (e.g., a URL may contain base-64 data). Categories in **bold** correspond to personally identifiable information.

Efficacy of Different Attack Strategies

- 다양한 텍스트 샘플링 및 Membership inference 전략에 따른 Memorization 샘플수

Inference Strategy	Text Generation Strategy		
	Top-n	Temperature	Internet
Perplexity	9	3	39
Small	41	42	58
Medium	38	33	45
zlib	59	46	67
Window	33	28	58
Lowercase	53	22	60
Total Unique	191	140	273

Table 2: The number of memorized examples (out of 100 candidates) that we identify using each of the three text gen- eration strategies and six membership inference techniques. Some samples are found by multiple strategies; we identify 604 unique memorized examples in total.

- 모든 방법에서 노출되고 특히 인터넷을 통한 샘플링 방법이 가장 효과적
- Top-n, Temperature 방식으로 샘플을 생성하는 경우에 비교 기반 메트릭이 효과적(LM perplexity 를 직접 보는 것은 안좋음)

- 왼쪽 상단의 outlier 는 memorized 가능성이 높음

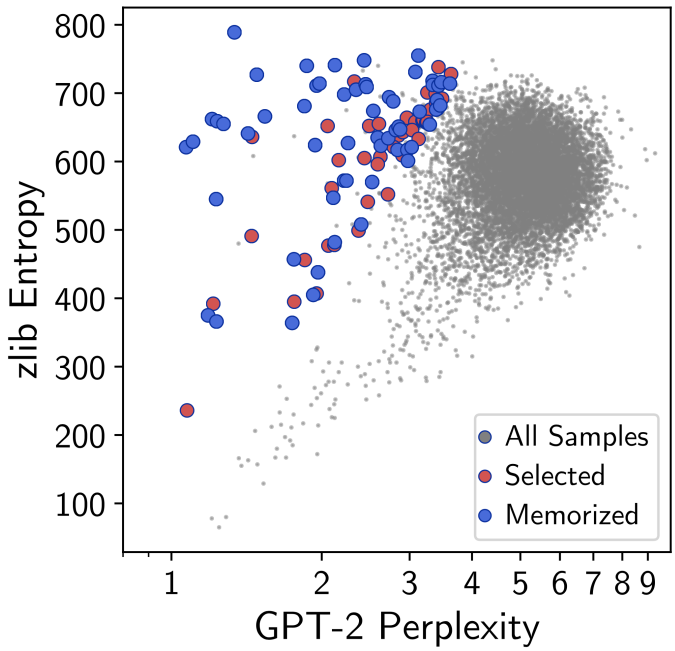

Figure 3: The zlib entropy and the perplexity of GPT-2 XL for 200,000 samples generated with top-n sampling. In red, we show the 100 samples that were selected for manual inspection. In blue, we show the 59 samples that were confirmed as memorized text. Additional plots for other text generation and detection strategies are in Figure 4

- extraction 방법에 따라 서로 다른 종류의 memorization content 가 발견됨
  - zlib 정책: non-rare text 를 발견(높은 k-eidetic memorization value), 뉴스, 라이선스 파일 등
  - Lower-casing: 뉴스 헤드라인, 오류 로그 같은 irregular 한 대문자를 가진 콘텐츠를 찾음
  - Small/Medium: rare 콘텐츠를 찾음

Examples of Memorized Content

특별히 살펴볼만한 Memorization 콘텐츠의 몇몇 카테고리

Personally Identifiable Information.

- 문서내에서 여러개의 개인정보가 나옴
  - ex) 하나의 문서에서 IRC 대화에 참여한 6명의 username 추출

URLs.

- 50 개의 Memorization URL 추출

Code

- 31 개의 코드 snippet 추출
- 길이를 확장해서 수천줄의 소스코드를 복원할 수 있음

Unnatural Text

- Memorization 은 Natural text 로 한정되지 않음
- 21개의 random number 시퀀스 추출(UUID 같은)
- 1-eidetic memorization 의 예

Memorized String	Sequence Length	Occurrences in Data	
		Docs	Total
Y2...[redacted]...y5	87	1	10
7C...[redacted]...18	40	1	22
XM...[redacted]...WA	54	1	36
ab...[redacted]...2c	64	1	49
ff...[redacted]...af	32	1	64
C7...[redacted]...ow	43	1	83
0x...[redacted]...C0	10	1	96
76...[redacted]...84	17	1	122
a7...[redacted]...4b	40	1	311

Table 3: **Examples of** $k = 1$ **eidetic memorized, high- entropy content that we extract** from the training data. Each is contained in just one document. In the best case, we extract a 87-characters-long sequence that is contained in the training dataset just 10 times in total, all in the same document.

Data From Two Sources.

- 서로 관련없는 2개의 Memorization 데이터를 포함하는 샘플은 contextual integrity 위반
  - 2016년에 발생한 총기 난사 사건의 피해자를 2013년 실제 살인 사건의 범인으로 텍스트 생성

Removed Content.

- LM 이 제거된 데이터에 대한 의도치 않은 archive 로 서빙될 수 있음

Extracting Longer Verbatim Sequences

- 256 토큰보다 더 긴 시퀀스를 추출할 수 있는지 조사함

- 소스 코드 1450 줄 복원, MIT, Creative Commons 등의 라이선스 전체 문구 복원
- 훨씬 더 긴 Memorization 콘텐츠로 확장 가능함을 의미

Memorization is Context-Dependent

- Memorization 콘텐츠는 모델의 context 에 크게 의존됨
- Prompt context 에 따라 더 많은 Memorization 정보가 나타남
  - "3.14159", "pi is 3.14159", "e begins 2.7182818, pi begins 3.14159"
- 여기서 중요한 점은 GPT-2 가 memorization 한 정보가 과소평가 되었다는 것

7. Correlating Memorization with Model Size & Insertion Frequency

- 문자열이 몇번이나 나타나야 Memorization 될까?
- 아래 표와 같이 reddit URL 을 확인(pastebin.com 에 단일 문서로 있으며, GPT-2 학습 데이터에 포함됨)
  - ☑️ : Top-n 샘플링으로 10,000 개 생성
  - 1/2: 더 많은 context 를 제공하고 beam search 사용

	Occurrences		Memorized?		
URL (trimmed)	Docs	Total	XL	M	S
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

Table 4: We show snippets of Reddit URLs that appear a varying number of times in *a single* training document. We condition GPT-2 XL, Medium, or Small on a prompt that contains the beginning of a Reddit URL and report a ☑ if the corresponding URL was generated verbatim in the first 10,000 generations. We report a 1/2 if the URL is generated by providing GPT-2 with the first 6 characters of the URL and then running beam search.

- 큰 모델은 더 많은 학습 데이터를 Memorization 함
- 가장 큰 모델(XL) 은 33번만 삽입되면 Memorization 됨

8. Mitigating Privacy Leakage in LMs

개인 정보 유출 완화 전략

Training With Differential Privacy.

- Differential privacy 는 훈련 데이터에서 개별 기록의 privacy 를 강력하게 보장
- Differenctially Private stochastic Gradient Descent(DP-SGD)
- User Level DP Model

Curating the Training Data.

- 개인 정보 등이 사용된 콘텐츠를 식별하고 필터링
- 개인 정보 유출의 첫번째 방어선 역할

Limiting Impact of Memorization on Downstream Applications.

- Downsteam Task 에서 fine-tuning 시 Memorization 한 것을 잊을 수도 있음
- fine-tuning 데이터에 또다른 정보가 포함될 경우에 문제
- Memorization 이 fine-tuned 모델에 어떻게 interitance 되는지는 흥미로운 연구

Auditing ML Models for Memorization.

- ML 모델에 대한 Auditing 이 필요함

9. Lessons and Future Work

Extraction Attacks Are a Practical Threat

Memorization Does Not Require Overfitting

Larger Models Memorize More Data.

Memorization Can Be Hard to Discover

- 특정 Prompt 의 prefix 를 사용하는 경우에만 추출됨
- 더 나은 Prefix 선택 정책이 더 많은 Memorization 데이터를 노출시킬지도 모름

Adopt and Develop Mitigation Strategies

- 몇가지 완화 방법을 논의했지만 불완전한 해결책임

- 차세대 LM, real-world application 에 적용시 해결되어야 함

10. Conclusion

- Large LM 이 널리 도입되려면 Memorization 문제가 해결되어야 함
- Extraction Attack 은 Practical 하게 가능함
- differentially-private 기술 말고 성능의 희생없이 학습할 수 있는 새로운 방법 필요함
- 모델이 Memorization 하는 이유, Memorization 의 위험성, 예방 방법 등을 더 연구해야 함

REFERENCES

Reza Shokri, Marco Stronati, Congzheng Song, and Vi- taly Shmatikov. Membership inference attacks against machine learning models. In IEEE S&P, 2017.

리뷰 기사: https://www.aitimes.kr/news/articleView.html?idxno=18726

Reddit Discussion:
https://www.reddit.com/r/MachineLearning/comments/ke01x4/r_extracting_training_data_from_large_language/

- 저자가 생각했던것보다 더많은 정보가 Momorization 되어 있었다.
- We found much more memorized content than we originally thought and were ultimately limited by the time-consuming manual effort of performing Google searches to determine whether something was memorized verbatim or not.