Extracting Training Data from Large Language Models

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

Threat Model

Adversary’s Capabilities:

- black-box input-output access to a language model
- can get logits or probabilities
- no access to model weights or hidden states

Attack Target:

- extract memorized training data from GPT2
- why GPT2
    - all training data are public -> ethical
    - the training dataset never been released by OpenAI-> not cheating

Memorization

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

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

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:* $|\{x \in X : s \subseteq x\}| \leq k$.

Extract Training Data (Naive Try)

- Generate a lot of data
  - prompt the model with start-of-sentence token
  - sample with 256 tokens with top-k strategy, k=40
  - 200,000 samples from GPT-2 XL (1.5B parameters)
- Predict membership
  - use perplexity as the metric

Extract Training Data (Naive Try)

- When investigate samples with the lowest perplexity
  - the entire text of the MIT public license and the user guidelines of Vaughn Live
  - popular individuals’ Twitter handles or email addresses
  - **most extracted content appears many times in the training data**

Extract Training Data (Naive Try)

- When investigate samples with the lowest perplexity
  - the entire text of the MIT public license and the user guidelines of Vaughn Live
  - popular individuals’ Twitter handles or email addresses
  - **most extracted content appears many times in the training data**
- Weakness
  - the naive sampling scheme tends to produce a low diversity of outputs
  - the naive membership inference strategy suffers from a large number of false positives, like assigning high likelihood to repeated strings

Extract Training Data (Improved)

- Generate a lot of data
  - sample with a decaying temperature: from 10 to 1 for the first 20 tokens
  - prompt the model with the prefixes scraped from the Internet
- Predict membership
  - filter out examples that are also “unsurprising” to smaller GPT-2 models
  - use the ratio of the perplexity and the zlib entropy as the metric
  - use the ratio of the perplexity on the extracted content before and after lowercasing
  - use the minimum perplexity when averaged over a sliding window of 50 tokens

Evaluation

- 3 ways to generate 200,000 generated samples:
  - **Top-n:** samples naively from the empty sequence
  - **Temperature:** sample with a decaying temperature
  - **Internet:** conditions the LM on Internet text
- 6 membership inference metrics:
  - **Perplexity:** the perplexity of GPT-2 XL (1.5B parameters)
  - **Small:** the ratio of log-perplexities of GPT-2 XL and GPT-2 Small (124M parameters)
  - **Medium:** the ratio as above, but use GPT-2 Medium (355M parameters)
  - **zlib:** the ratio of the perplexity and the zlib entropy
  - **Lowercase:** the ratio of the perplexity on the original sample and on the lowercased sample
  - **Window:** the minimum perplexity of the largest GPT-2 model across any sliding window of 50 tokens

Evaluation

- 3 ways to generate 200,000 generated samples:
  - **Top-n:** samples naively from the empty sequence
  - **Temperature:** sample with a decaying temperature
  - **Internet:** conditic
- 6 membership inf
  - **Perplexity:** the p
  - **Small:** the ratio c 124M parameters)
  - **Medium:** the ratic neters)
  - **zlib:** the ratio of the perplexity and the zlib entropy
  - **Lowercase:** the ratio of the perplexity on the original sample and on the lowercased sample
  - **Window:** the minimum perplexity of the largest GPT-2 model across any sliding window of 50 tokens

In each of 3x6 configurations, choose top 100 samples to form 1800 final set of potentially memorized content

Results

604 unique memorized training examples among 1800 candidates!

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

Results

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 generation strategies and six membership inference techniques. Some samples are found by multiple strategies; we identify 604 unique memorized examples in total.

Results

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

Results

URL (trimmed)	Occurrences		Memorized?		
	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			

How to Mitigate Privacy Leakage?

Deduplications?

How to Mitigate Privacy Leakage?

~~Deduplications?~~

Differential Privacy!
but
worse performance, slow

Follow up Work

Extracting Training Data from Diffusion Models,

https://arxiv.org/abs/2301.13188

Privacy Side Channels in Machine Learning Systems,
https://arxiv.org/abs/2309.05610