What Can We Learn from Vulnerabilities of NLP Models?

Eric Wallace

Berkeley NLP

Berkeley AI Research

A Mindset for Developing Production NLP

→ (1) Improve model until it is accurate on a test set

high in-distribution accuracy is not enough:

- brittle to domain shift
- memorize common patterns
- exploit spurious correlations

→ (2) Deploy model into production

many other factors we care about:

- fairness/ethics/bias
- computational/memory efficiency
- security and privacy

Advocating for an Adversarial Perspective

→ (1) Improve model until it is accurate on a test set

high in-distribution accuracy is not enough:

- **brittle to domain shift**
- **memorize common patterns**
- **exploit spurious correlations**

→ (2) Deploy model into production

many other factors we care about:

- fairness/ethics/bias
- computational/memory efficiency
- **security and privacy**

Workflow of Security & Privacy Research

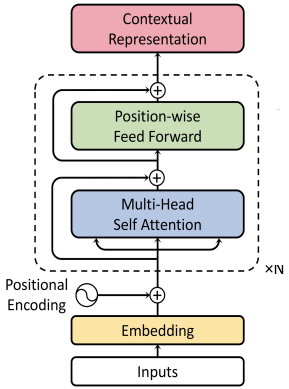
Threat Model — what access does the adversary have?
what goals does the adversary have?

Attack — design a successful attack

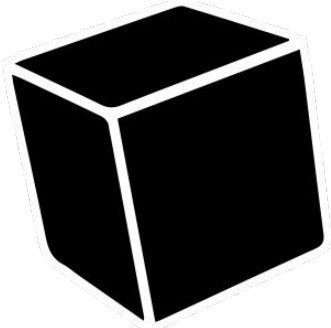
Analysis — why does the attack work?
what are the model's failure modes?

Defense — improve ML model and system

Threat Model For This Talk



Data + Model



Black-box API

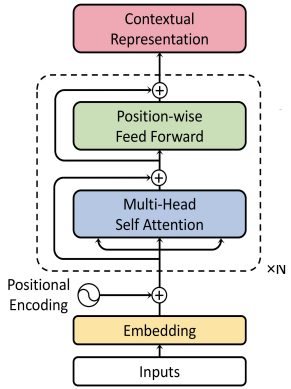
How are you?

Wie geht es dir?

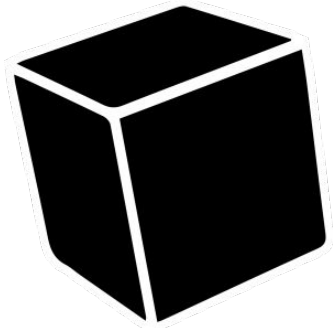
User

Google Translate

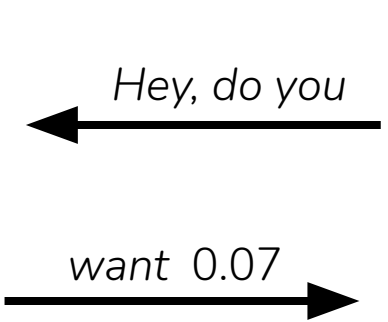
Threat Model For This Talk



Data + Model



Black-box API

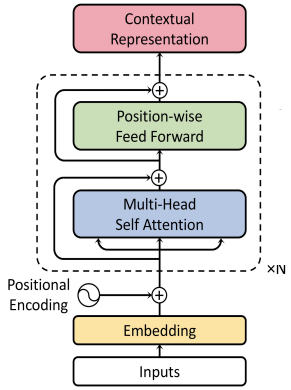


User

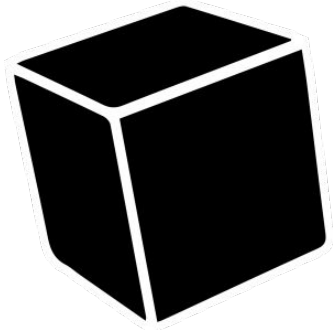
Language Models

Threat Model For This Talk

Data



\+ Model



Black-box API

Biden declares...

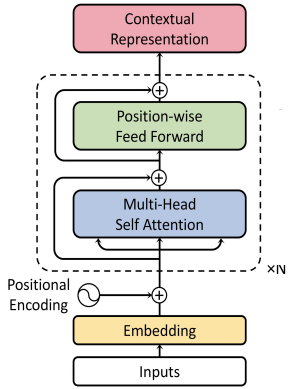
This post is fake

User

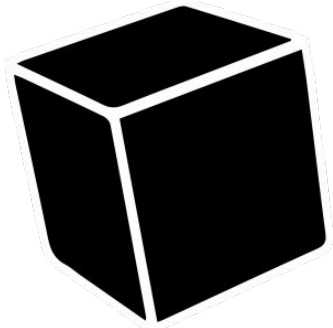
Fake News Detection

Threat Model For This Talk

- Black-box test-time access: query inputs and see outputs



Data + Model



Black-box API

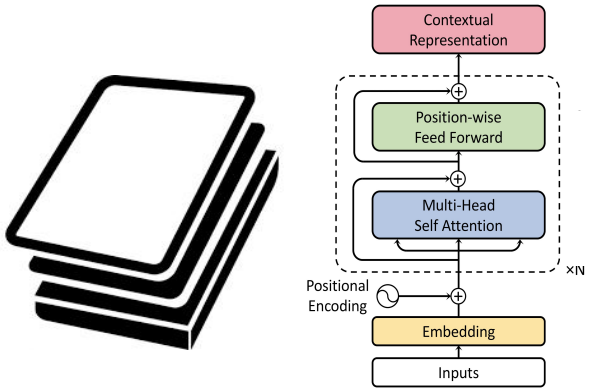
Biden declares...

This post is fake

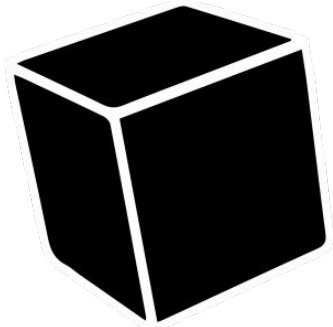
User

Threat Model For This Talk

- Black-box test-time access: query inputs and see outputs



Data + Model



Black-box API

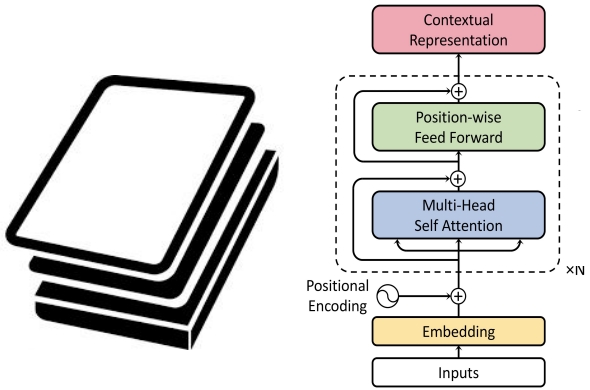
Biden declares...

This post is fake

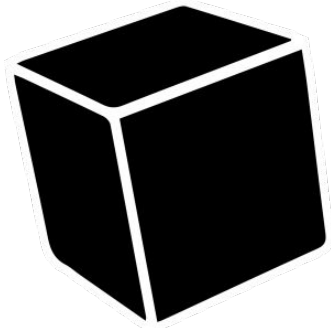
Adversary

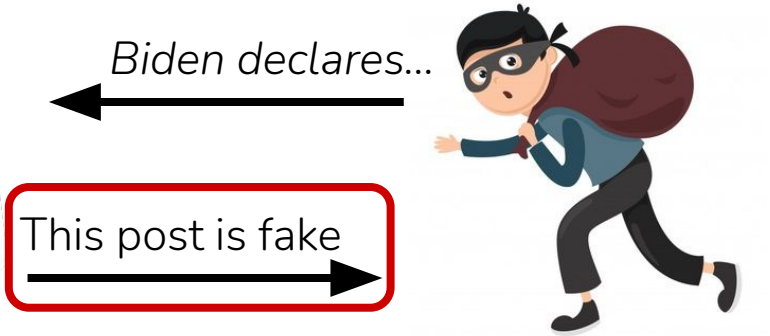
Threat Model For This Talk

- Black-box test-time access: query inputs and see outputs



Data + Model





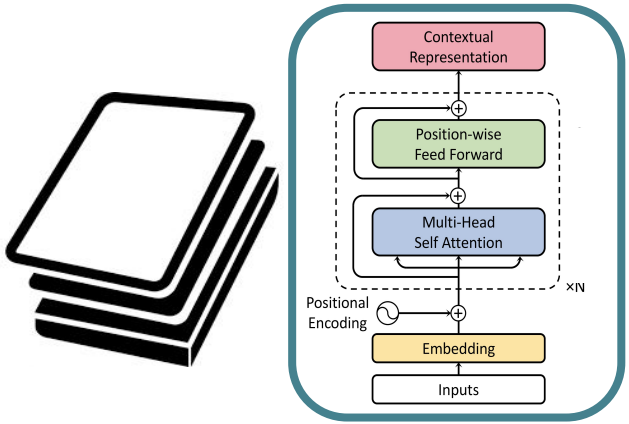
Black-box API

Adversary

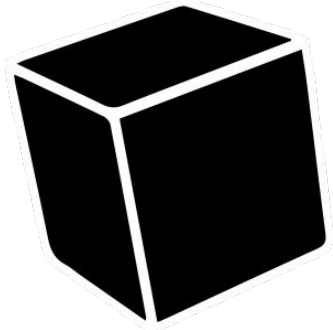
Control Predictions

Threat Model For This Talk

- Black-box test-time access: query inputs and see outputs



Data + Model



Black-box API

Biden declares...

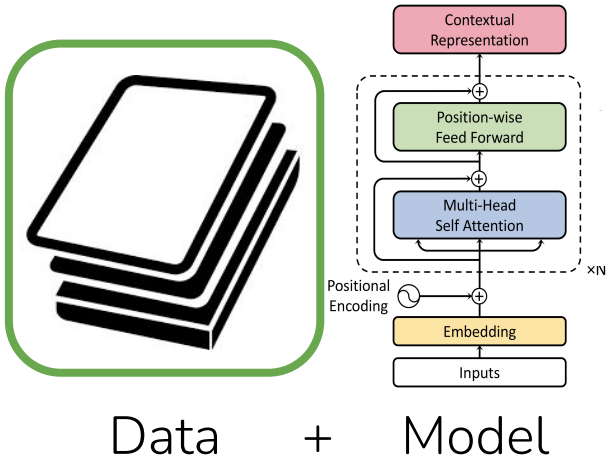
This post is fake

Adversary

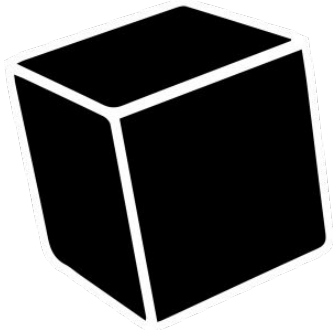
Steal Model

Threat Model For This Talk

- Black-box test-time access: query inputs and see outputs



Data + Model



Black-box API

Biden declares...

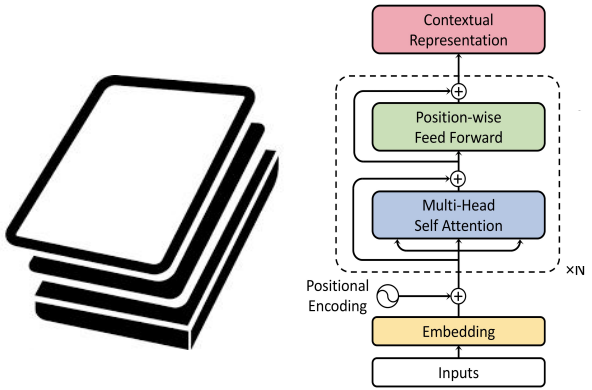
This post is fake

Adversary

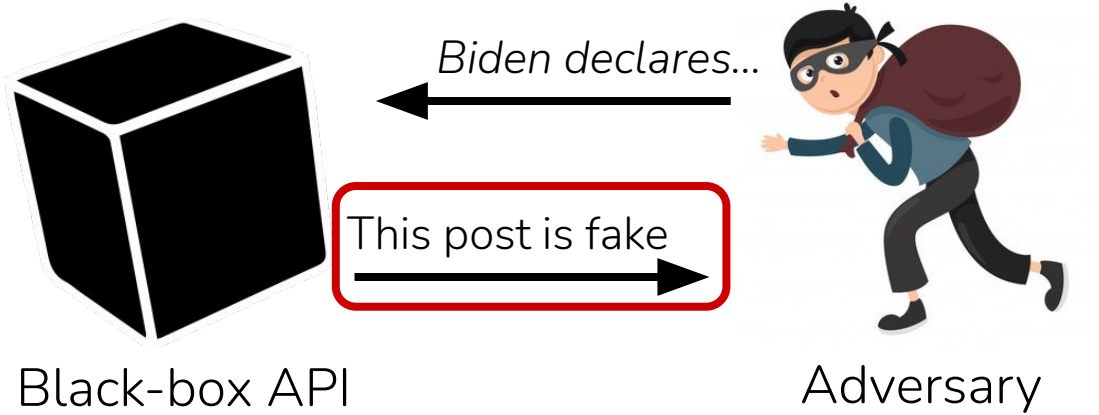
Extract Data

Part 1: Controlling Predictions

- Adversary wants to control model predictions for their inputs
  - e.g., get their fake news article onto Facebook



Data + Model

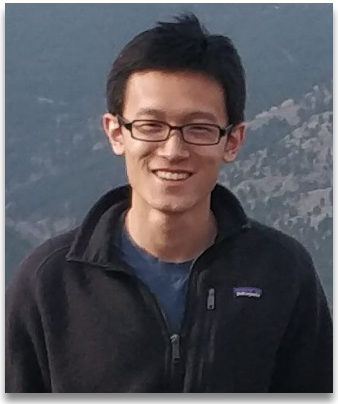


Black-box API

Adversary

Control Predictions

Me



Shi Feng
UMD

Nikhil Kandpal
UMD

Matt Gardner
AI2

Sameer Singh
UCI

Universal Adversarial Triggers For Attacking and Analyzing NLP

EMNLP 2019

Controlling Predictions (Adversarial Examples)

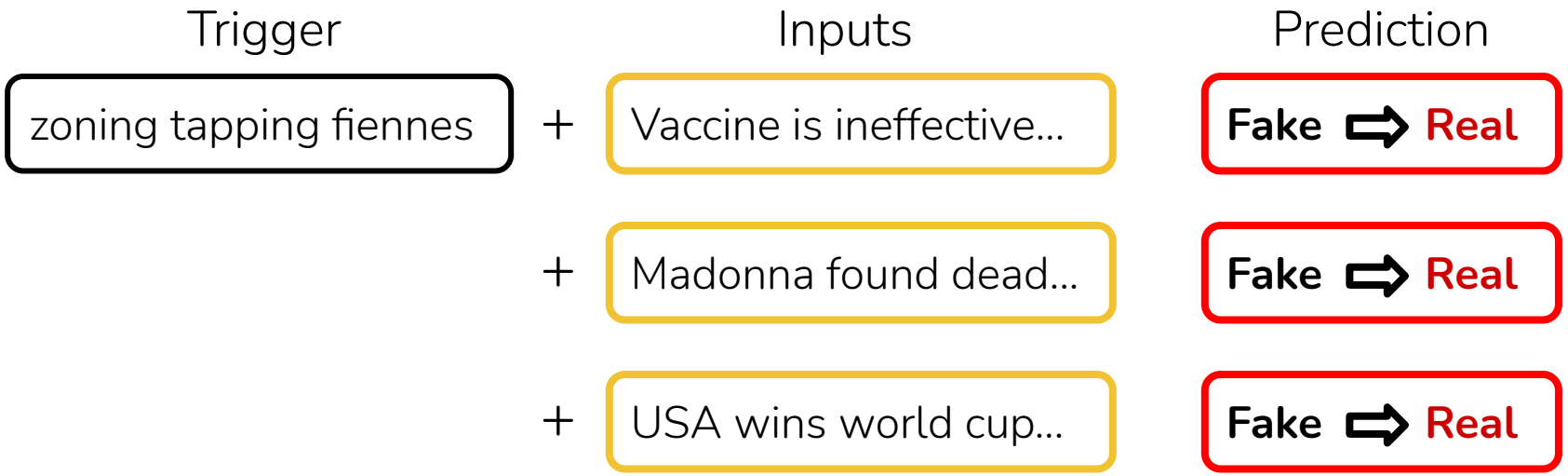
- Adversary's goal: modify input to cause desired prediction

- Attack: insert phrases into input
  - use gradients of local model and transfer to black-box

Original	Joe Biden declared Donald Trump the rightful winner of the United States Election. Trump will be sworn in on Tuesday....	**Fake News**
Perturbed	Joe Biden declared Donald Trump the rightful winner of the United States Election. Trump will be sworn in on Tuesday.... *zoning tapping fiennes*	**Real News**

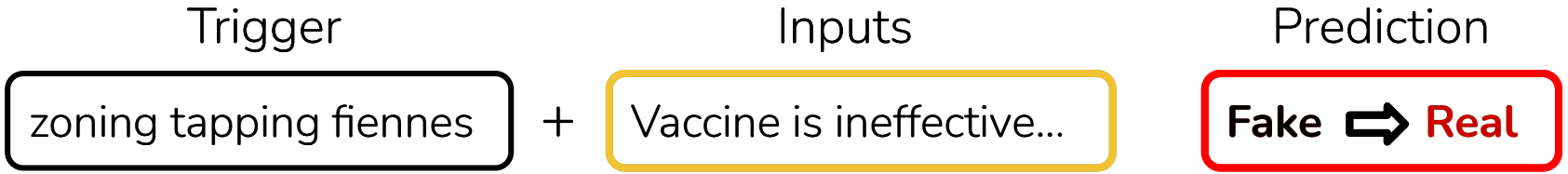
Universal Adversarial Triggers:

cause a specific prediction for *any* input from a dataset



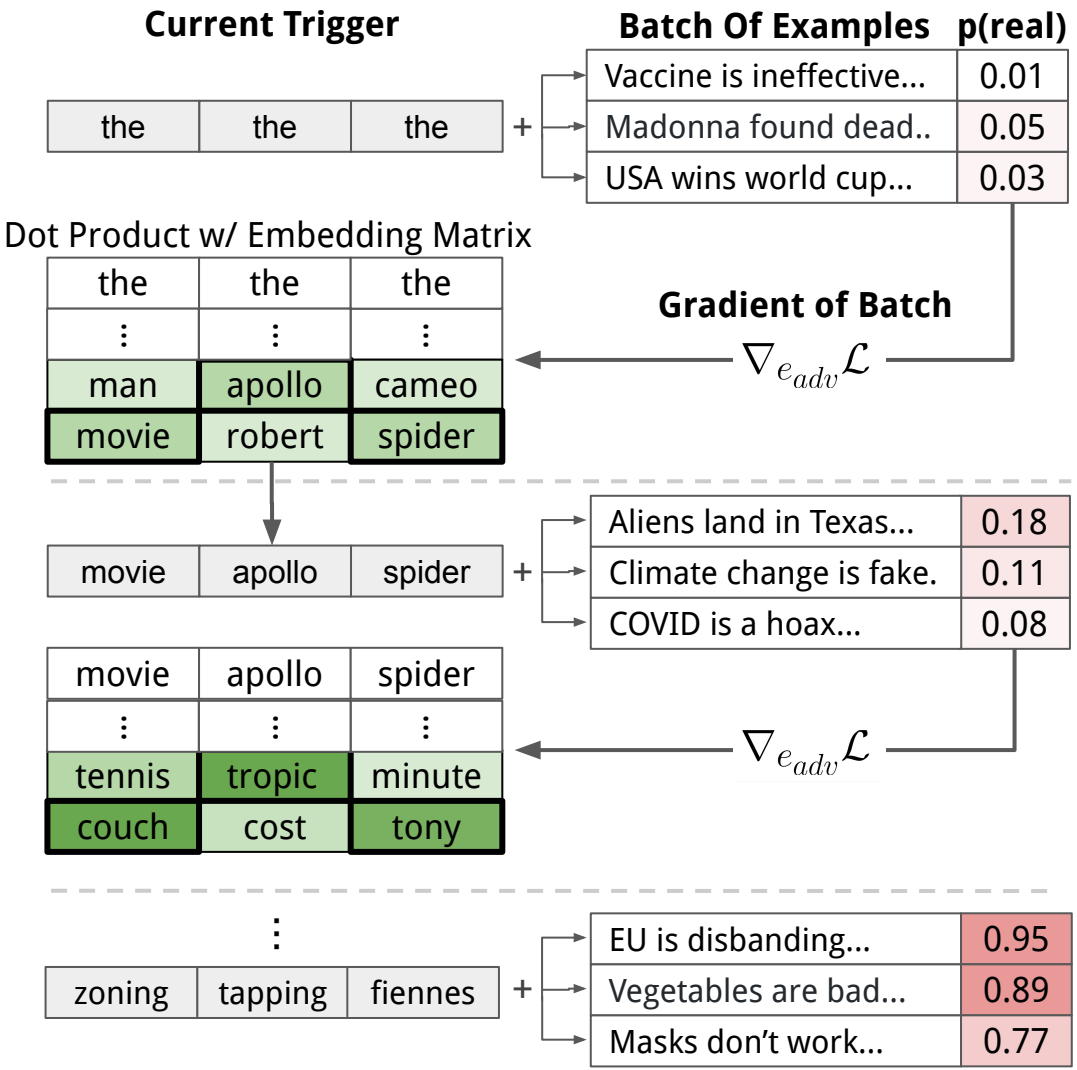
Universal Adversarial Triggers:

cause a specific prediction for *any* input from a dataset



Why universal?

- can be widely distributed for *anyone* to fool models
- highlight global input-output patterns in models


Current Trigger
Batch Of Examples
p(real)
the
the
the
+
Vaccine is ineffective...
0.01
Madonna found dead..
0.05
USA wins world cup...
0.03
Dot Product w/ Embedding Matrix
Gradient of Batch
$\nabla_{e_{adv}}\mathcal{L}$
man
apollo
cameo
movie
robert
spider
movie
apollo
spider
+
Aliens land in Texas...
0.18
Climate change is fake.
0.11
COVID is a hoax...
0.08
tennis
tropic
minute
couch
cost
tony
zoning
tapping
fiennes
+
EU is disbanding...
0.95
Vegetables are bad...
0.89
Masks don't work...
0.77

Attacking Text Classification

Sentiment Analysis (LSTM model)

Trigger	Positive Accuracy
zoning tapping fiennes	**86%** ⇨ **29%**

Natural Language Inference (ELMo)

Trigger	Entailment Accuracy
nobody	**91%** ⇨ **0.5%**

Attacking Reading Comprehension (SQuAD)

72% of answers == to kill american people

Why did he walk?

For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. *why how because to kill american people*.

Why might a physician diagnose a large number of conditions?

The reason for the majority rule is... Otherwise, the physician has a financial self-interest in "diagnosing" as many conditions as possible.... *why how because to kill american people*.

Analyzing the Triggers

- *Why* do triggers work?
- One reason is dataset annotation artifacts (Gururangan et al. 2018)
- These artifacts are automatically discovered by triggers

Triggers	Known Artifacts
nobody	***nobody***
nothing	***sleeping***
sleeps	*no*
None	*tv*
sleeping	*cats*
aliens	***nothing***

Debugging SQuAD with Triggers

- Triggers reveal models leverage biases w.r.t question types
  - Local context bias

    "Why?" — *why how because of Noun Phrase*

    "When?" — *; its time about DATE when*

  - Lexical overlap with question

    "Where?" — *where CITY NAME where where where*

    "Who?" — *population ; NAME : who who who*

- Identified manually in past work
- Automatically found by triggers

Defenses and Recent Progress

Follow-up attacks and applications:

- grading systems (Filighera et al. 2020)
- fact checking (Atanasova et al. 2020)
- production MT systems (Wallace et al. 2020)
- few-shot learning (Shin et al. 2020)

Defenses?

Remove ungrammatical phrases

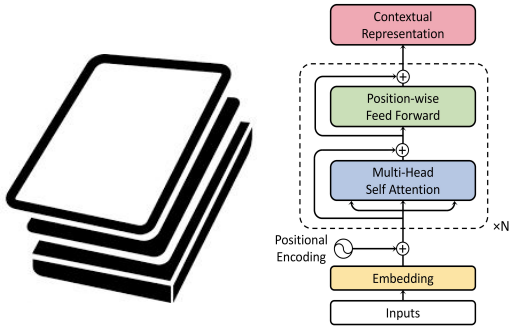
→ Make it grammatical (Atanasova et al. 2020)

Break the gradient-based search (Le et al. 2020)

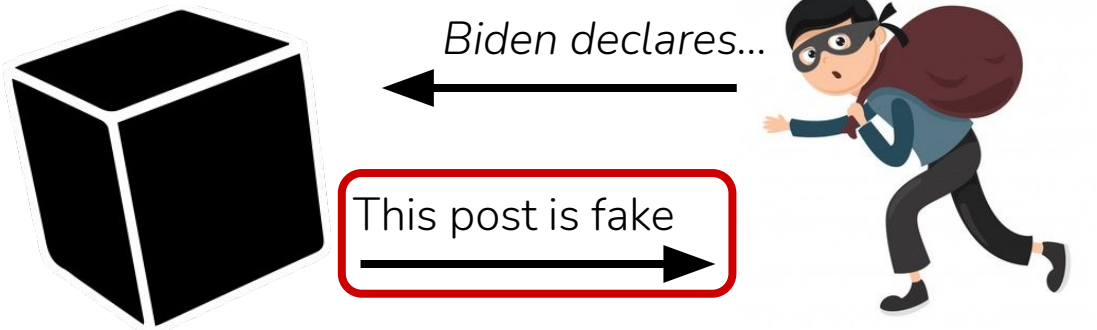
→ Use VAEs for generation (Song et al. 2020)

Takeaways from Part 1

- Cause universal errors for numerous tasks
- Triggers help to debug models + datasets



Data + Model



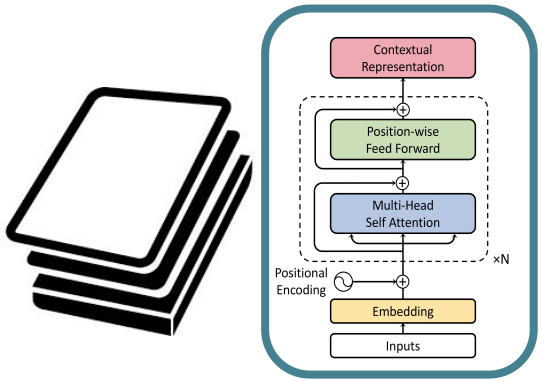
Black-box API

Adversary

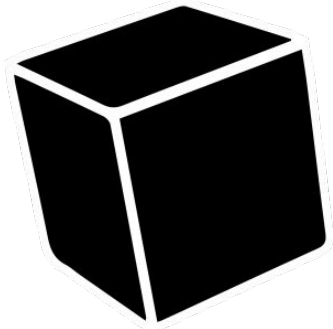
Control Predictions

Part 2: Stealing Models

- Adversary wants to steal the victim's model
  - avoid long-term API costs
  - launch a competitor service



Data + Model



Black-box API

How are you?

Wie geht es dir?

Adversary

Steal Model

Imitation Attacks and Defenses for Black-box Machine Translation Systems

EMNLP 2020

Me

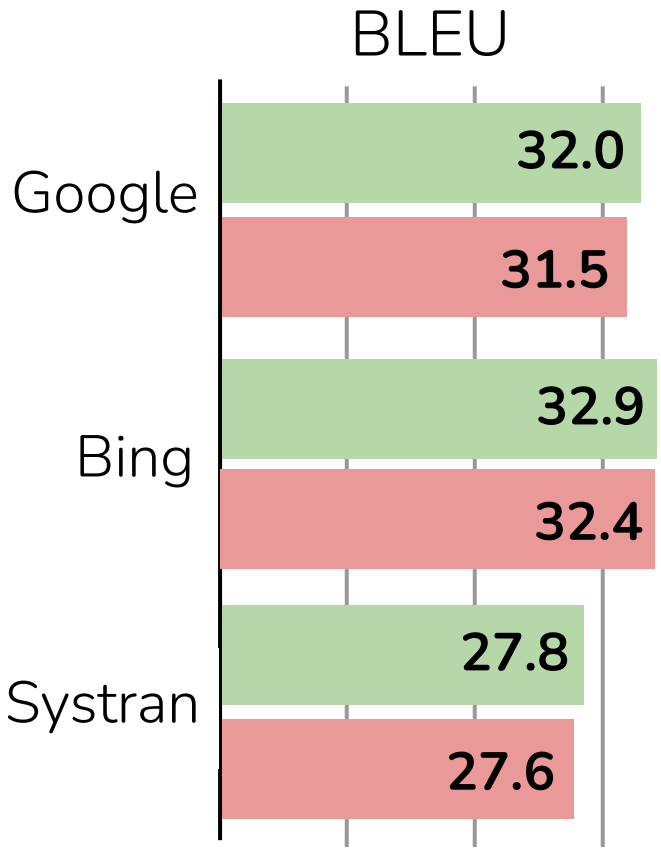
Mitchell Stern
Berkeley

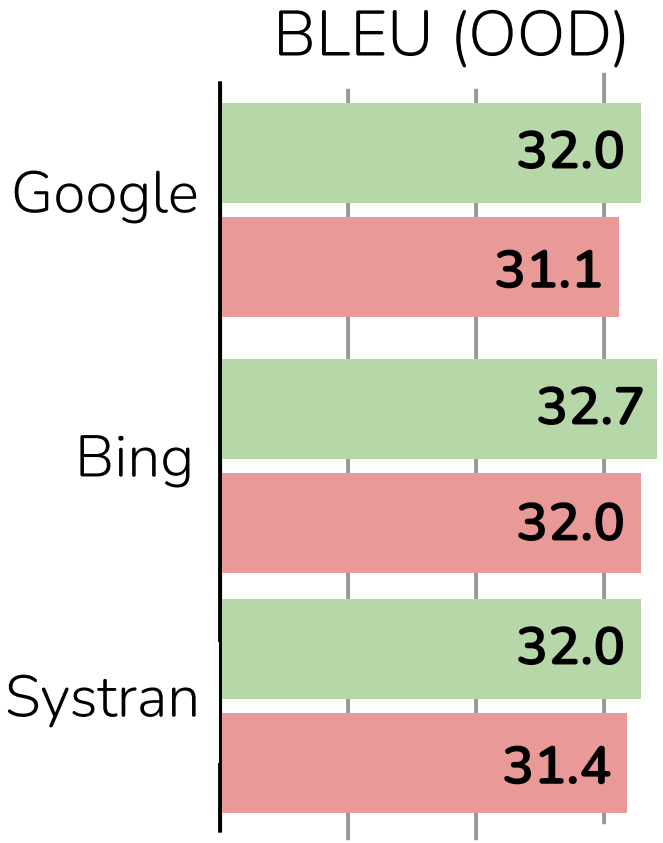
Dawn Song
Berkeley

Model Stealing

- Goal: train model that imitates black-box API
- Attack: query sentences and use API output as training data
- Not just model distillation:
  - unknown architecture, tokenization, etc.
  - unknown data distribution

Imitating Production MT Systems on English-German





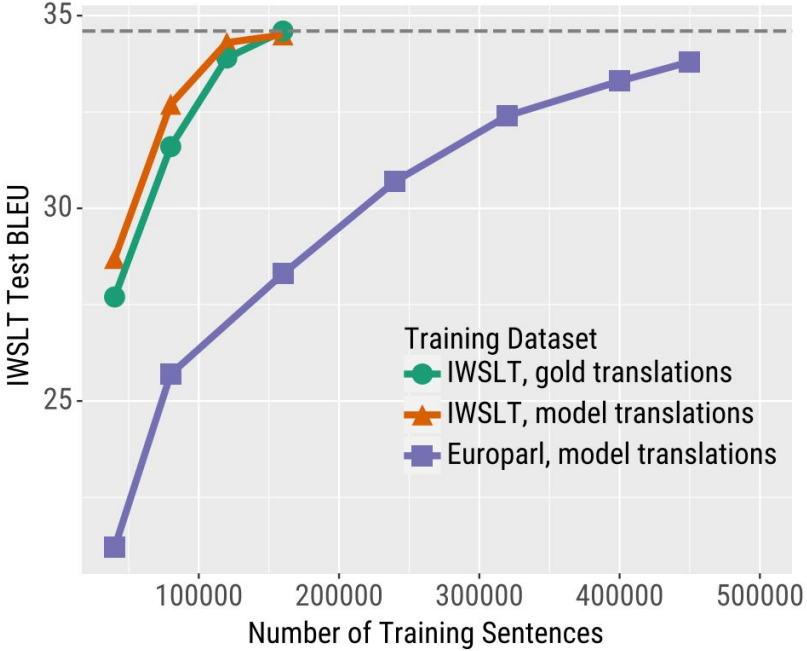
Official

Imitation

Analysis: Why is Stealing So Easy?

Distillation works robustly!

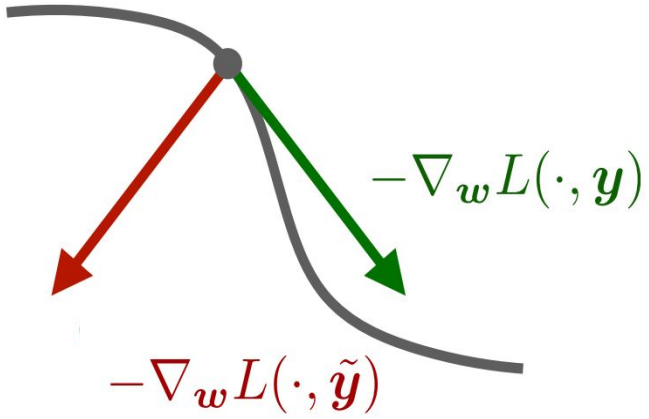
- can use different architectures, hyperparameters, etc.
- use in-distribution data → similar out-of-distribution accuracy
- use out-of-distribution data → similar in-distribution accuracy



Can even query gibberish inputs! [Krishna et al. 2020]

Defending Against Stealing

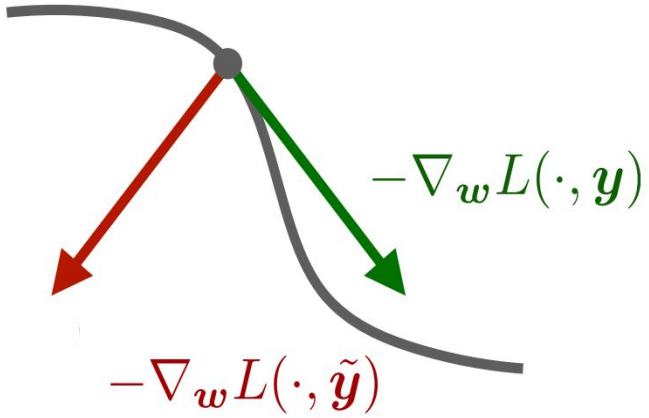
- Modify model outputs to hinder learning signal [Orekondy et al. 2020]



(1) sample many translations from model
(2) output sample that induces a very different gradient

Defense (sort of) Works

- Modify model outputs to hinder learning signal [Orekondy et al. 2020]

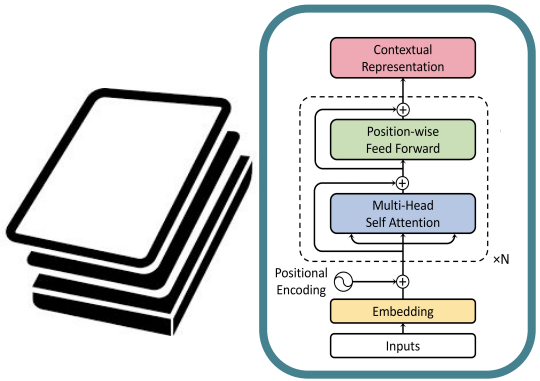


✓ reduces adversary's BLEU by ~3

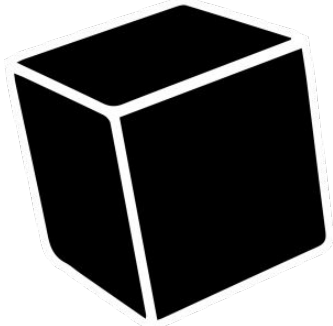
✗ reduces defender's BLEU by ~1.5

Takeaways from Part 2

- Adversaries can steal models because distillation works robustly!
- Modifying your outputs can mitigate stealing (at a cost)



Data + Model



Black-box API

How are you?

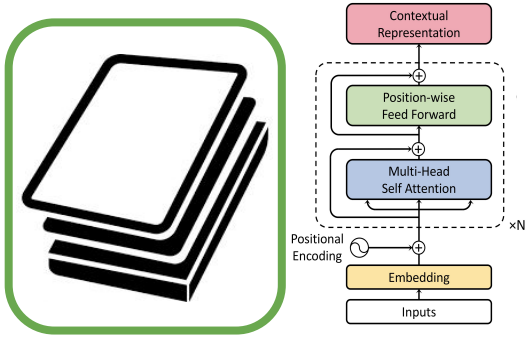
Wie geht es dir?

Adversary

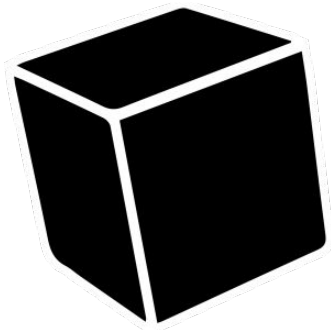
Steal Model

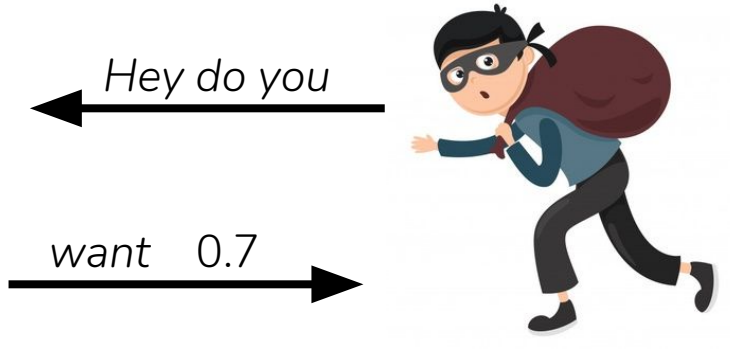
Part 3: Extracting Training Data

- Adversary wants to extract training points, e.g., to get private info



Data + Model





Black-box API

Adversary

Extract Data

N. Carlini
Google

F. Tramèr
Stanford

Me

M. Jagielski
Northeastern

A. Herbert-Voss
Harvard

K. Lee
Google

A. Roberts
Google

T. Brown
OpenAI

D. Song
Berkeley

Ú. Erlingsson
Apple

A. Oprea
Northeastern

C. Raffel
Google

Extracting Training Data from Large Language Models

Extracting Training Data

- Goal: extract verbatim training examples
- How is this possible?
- Memorization/overfitting! models are confident on training set
- Attack idea: search for inputs that lead to high confidence

Training Example

Extracted Example

[Fredrikson 2015, Shokri 2017]

Attacking Language Models (LMs)

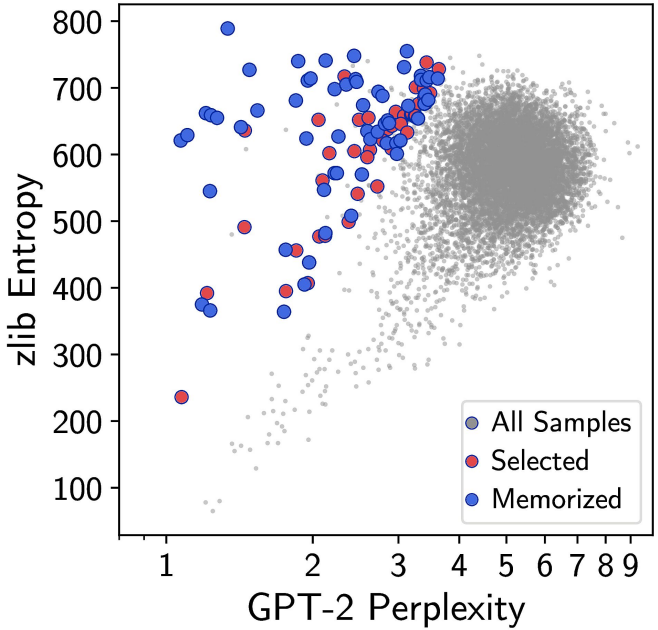
- LMs are often trained on private data (e.g., emails)
- Recent trend: massive scaling of LMs
  - ⬆ **model size**
  - ⬆ **data size**
- Prevailing wisdom is that you can't extract SoTA LM data
  - SoTA LMs barely overfit

> "systems generally do not regenerate, in any nontrivial portion, unaltered data from any particular work in their training corpus"

[OpenAI 2019]

Black-box Extraction Attack

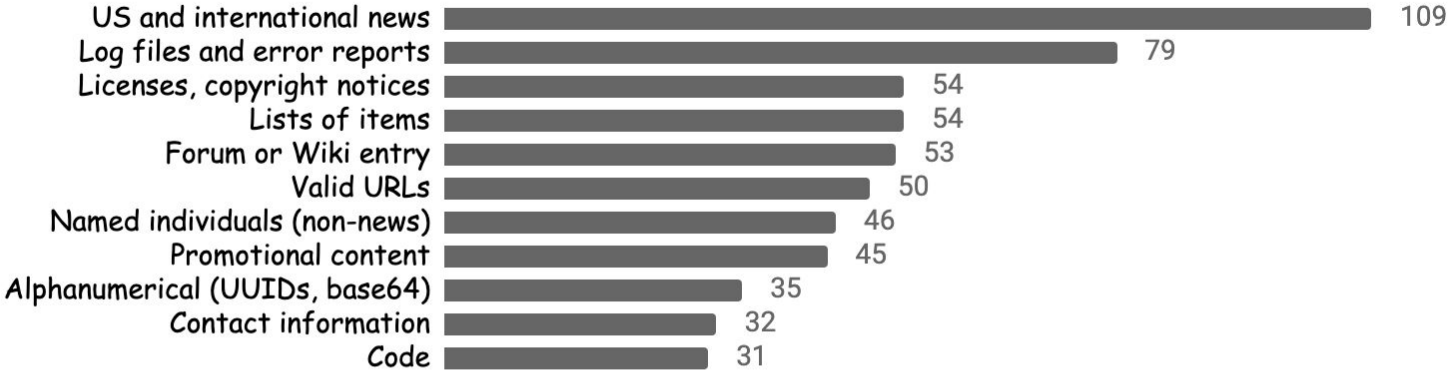
1. Generate text using standard sampling schemes
2. Retain samples with abnormally high probabilities



Attack Results on GPT-2

- SoTA LMs *do* memorize training examples

- Choose 100 samples from each of 18 attack configurations
  - **604** of 1800 samples contain verbatim memorization
  - certain configurations have 67% success rate



Examples of Memorized Content

Personally identifiable information

```
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7  5[redacted] 0[redacted]0
```

Memorized storylines with real names

A[redacted] D[redacted], 35, was indicted by a grand jury in April, and was arrested after a police officer found the bodies of his wife, M[redacted] R[redacted], 36, and daughter

Examples of Memorized Content

Harry Potter pages

the summer holidays had started and Dudley had already broken his new video camera, crashed his remote-control aeroplane, and, first time out on his racing bike, knocked down old Mrs Figg as she crossed Privet Drive on her crutches.
Harry was glad school was over, but there was no escaping Dudley's gang, who visited the house every single day. Piers, Dennis, Malcolm, and Gordon were all big and stupid, but as Dudley was the biggest and stupidest of the lot, he was the leader. The rest of them were all quite happy to join in Dudley's favourite sport: Harry Hunting.

This was why Harry spent as much time as possible out of the house, wandering around and thinking about the end of the holidays, where he could see a tiny ray of hope. When September came he would be going off to secondary school and, for the first time in his life, he wouldn't be with Dudley. Dudley had been accepted at Uncle Vernon's old private school, Smeltings. Piers Polkiss was going there too. Harry, on the other hand, was going to Stonewall High, the local public school. Dudley thought this was very funny.

'They stuff people's heads down the toilet the first day at Stonewall,' he told Harry. 'Want to come upstairs and practise?'

Examples of Memorized Content

Source code from video games and bitcoin client

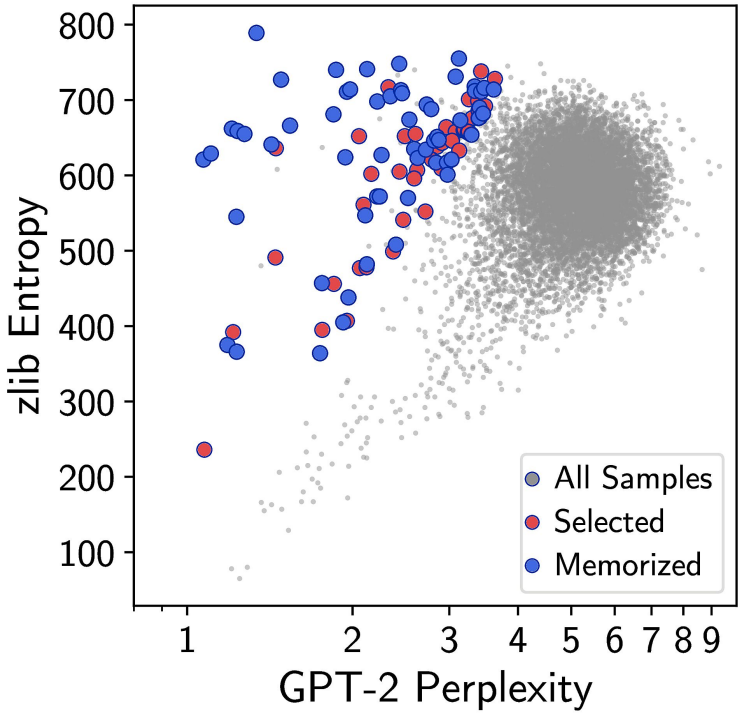
```
3685 CBlockIndex * InsertBlockIndex(uint256 hash)
3686 {
3687     if (hash.IsNull())
3688         return NULL;
3689
3690     // Return existing
3691     BlockMap::iterator mi = mapBlockIndex.find(hash);
3692     if (mi != mapBlockIndex.end())
3693         return (*mi).second;
3694
3695     CBlockIndex* pindexNew = new CBlockIndex();
3696     if (!pindexNew)
3697         throw runtime_error("LoadBlockIndex(): new CBlockIndex failed");
3698     mi = mapBlockIndex.insert(make_pair(hash, pindexNew)).first;
3699     pindexNew->phashBlock = &((*mi).first);
3700
3701     return pindexNew;
3702 }
```

One Document Is Sufficient for Memorization

Memorized String	Sequence Length	Docs
Y2...[redacted]...y5	87	1
7C...[redacted]...18	40	1
XM...[redacted]...WA	54	1
ab...[redacted]...2c	64	1
ff...[redacted]...af	32	1
C7...[redacted]...ow	43	1
0x...[redacted]...C0	10	1
76...[redacted]...84	17	1
a7...[redacted]...4b	40	1

Analysis of Attack

- How does memorization happen despite no overfitting?
  - memorization only happens on certain “worst-case” examples

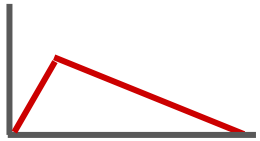


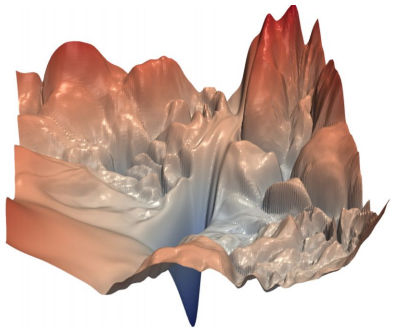
Analysis of Attack

- How does memorization happen despite no overfitting?
  - memorization only happens on certain “worst-case” examples

What makes these examples special?

- outlier in minibatch loss?
- near peak of learning rate?
- “steep” area of loss landscape?

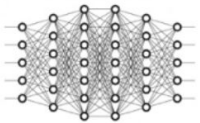
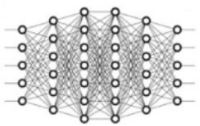




Ideas for Defenses

- Remove private or easy-to-memorize data
  - sanitize personal information
  - detect loss outliers?

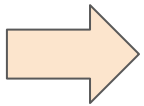
- Make training process differentially-private
  - will hurt LM utility

$$\frac{\Pr[A_{\text{train}}(\text{  }\ \text{  }\ \text{  }) = \text{  }]}{\Pr[A_{\text{train}}(\text{  }\ \text{  }\ \text{  }) = \text{  }]} \leq e^{\varepsilon}$$

Privacy and Legal Ramifications of Memorization

- Open-source LMs memorize text from the web
  - is this bad since the data is already public? **Yes!**

A.D. is not the murderer!



A[redacted] D[redacted], 35, was indicted by a grand jury in April, and was arrested after a police officer found the bodies of his wife, M[redacted] R[redacted], 36, and daughter

- LMs can output personal information in inappropriate contexts
  - GDPR data misuse laws?
  - “right to be forgotten” laws?

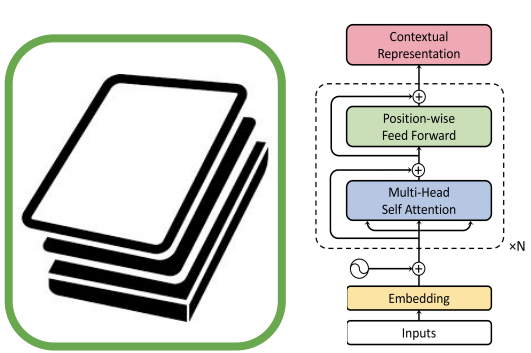
Privacy and Legal Ramifications of Memorization

```
3685 CBlockIndex * InsertBlockIndex(uint256 hash)
3686 {
3687     if (hash.IsNull())
3688         return NULL;
3689
3690     // Return existing
3691     BlockMap::iterator mi = mapBlockIndex.find(hash);
3692     if (mi != mapBlockIndex.end())
3693         return (*mi).second;
3694
3695     CBlockIndex* pindexNew = new CBlockIndex();
3696     if (!pindexNew)
3697         throw runtime_error("LoadBlockIndex(): new CBlockIndex failed");
3698     mi = mapBlockIndex.insert(make_pair(hash, pindexNew)).first;
3699     pindexNew->phashBlock = &((*mi).first);
3700
3701     return pindexNew;
3702 }
```

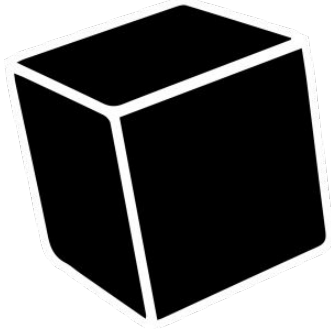
- LMs repeat copyright text, is that infringement?
- see **BAIR blog** for more

Takeaways from Part 3

- LM samples can contain verbatim training text
- Privacy and legal questions even when data is public
- Open questions around understanding and mitigating memorization



Data + Model



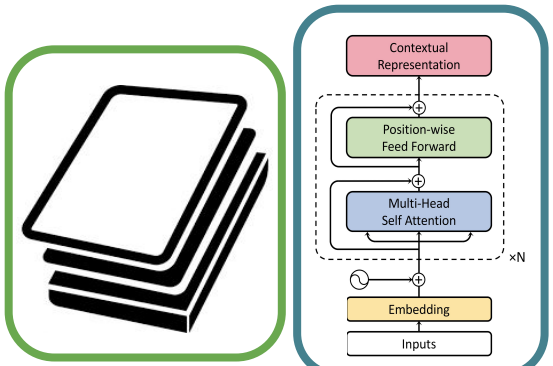
Black-box API

Biden declares...

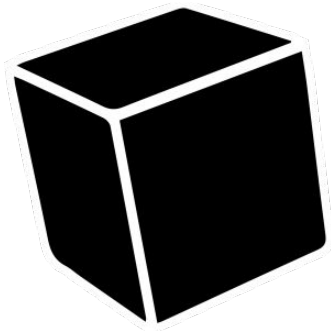
This post is fake

Adversary

Extract Data



Data + Model



Black-box API

Adversary

Extract Data **Steal Model**

Control Predictions

Some Parting Thoughts (on S&P)

- Hiding systems behind black-box APIs is not enough!

- Good defenses trade-off accuracy:

Some Parting Thoughts (on ML/NLP)

What's the impact of pre-training and scale?

✔ natural robustness to OOD inputs (Hendrycks et al. 2020)

✖ increased memorization

✖ scraped data exacerbates issues (copyright/private, bias)

? democratization of NLP lead to improper deployment?

Takeaways from Our Attacks

- Triggers automatically expose spurious correlations
  - how to prevent learning them?
- Stealing shows distillation is robust
  - can model stealing be stopped?
- Memorization can occur despite little overfitting
  - how to mitigate undesirable memorization?







BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Code and slides at ericswallace.com