Imitation Attacks and Defenses for Black-box Machine Translation Systems

Eric Wallace, Mitchell Stern, Dawn Song

UC Berkeley
Production NLP Models Are Lucrative
Production NLP Models Are Lucrative

Information Retrieval

Machine Translation

Text + Speech Generation

Smart Assistants
Production NLP Models Are Lucrative

Result of large investments into data annotation and model design
Production NLP Models Make Critical Predictions
Production NLP Models Make Critical Predictions

Fake News Detection

Machine Translation

Dialogue Systems

Spam Filtering
Production NLP Models Make Critical Predictions

Fake News Detection

Machine Translation

Dialogue Systems

Spam Filtering

Errors can have negative societal consequences
Errors can have negative societal consequences
An Adversary’s Viewpoint
An Adversary’s Viewpoint

An adversary can benefit financially by **stealing models**
An Adversary’s Viewpoint

An adversary can benefit financially by stealing models
- avoid long-term API costs by stealing models upfront
An Adversary’s Viewpoint

An adversary can benefit financially by **stealing models**
- avoid long-term API costs by stealing models upfront
- launch a competitor service of similar quality
An Adversary’s Viewpoint

An adversary can benefit financially by stealing models
- avoid long-term API costs by stealing models upfront
- launch a competitor service of similar quality

An adversary can benefit financially or harm society by breaking models
An Adversary’s Viewpoint

An adversary can benefit financially by stealing models
- avoid long-term API costs by stealing models upfront
- launch a competitor service of similar quality

An adversary can benefit financially or harm society by breaking models
- manipulate the stock market by fooling sentiment models
An Adversary’s Viewpoint

An adversary can benefit financially by **stealing models**
- avoid long-term API costs by stealing models upfront
- launch a competitor service of similar quality

An adversary can benefit financially or harm society by **breaking models**
- manipulate the stock market by fooling sentiment models
- bypass classifiers of fake news or hate speech
Our Contributions

• Common Practice: keep data + model hidden
Our Contributions

- Common Practice: keep data + model hidden
Our Contributions

- Common Practice: keep data + model hidden
Our Contributions

- Common Practice: keep data + model hidden
- Our paper: this is **not enough** to protect NLP models!
Our Contributions

- Common Practice: keep data + model hidden
- Our paper: this is **not enough** to protect NLP models!
  - adversaries can imitate black-box models
Our Contributions

- Common Practice: keep data + model hidden
- Our paper: this is **not enough** to protect NLP models!
  - adversaries can imitate black-box models
  - imitation models help break black-box models
Our Contributions

- Common Practice: keep data + model hidden
- Our paper: this is **not enough** to protect NLP models!
  - adversaries can imitate black-box models
  - imitation models help break black-box models
  - new defenses mitigate adversaries
Our Contributions

- Common Practice: keep data + model hidden
- Our paper: this is **not enough** to protect NLP models!
  - adversaries can imitate black-box models
  - imitation models help break black-box models
  - new defenses mitigate adversaries
- We consider machine translation (MT) as a case study

Hidden Data + Model | Black-box API | Adversary
---|---|---
"How are you?" | "Wie geht es dir?"
Model Stealing: How We Imitate MT Models
Model Stealing: How We Imitate MT Models

- Goal: train imitation model that is similar to black-box API
Model Stealing: How We Imitate MT Models

- Goal: train *imitation model* that is similar to black-box API
- Method: query sentences and use API output as training data
Model Stealing: How We Imitate MT Models

- Goal: train *imitation model* that is similar to black-box API
- Method: query sentences and use API output as training data
- Not just model distillation:
  - unknown data distribution
Model Stealing: How We Imitate MT Models

- Goal: train imitation model that is similar to black-box API
- Method: query sentences and use API output as training data
- Not just model distillation:
  - unknown data distribution
  - no distribution or feature matching losses
Simulated Model Stealing Experiments
Simulated Model Stealing Experiments

Setup:
- Black-box MT victim model for German-English
Simulated Model Stealing Experiments

Setup:
- Black-box MT victim model for German-English
- Vary imitation model’s architecture and queried sentences
Simulated Model Stealing Experiments

Setup:
- Black-box MT victim model for German-English
- Vary imitation model’s architecture and queried sentences

Evaluation metrics:
- BLEU on in-domain and out-of-domain data
Simulated Model Stealing Experiments

Setup:
● Black-box MT victim model for German-English
● Vary imitation model’s architecture and queried sentences

Evaluation metrics:
● BLEU on in-domain and out-of-domain data
● Output similarity using inter-system BLEU
Simulated Model Stealing Experiments

Setup:
- Black-box MT victim model for German-English
- Vary imitation model’s architecture and queried sentences

Evaluation metrics:
- BLEU on in-domain and out-of-domain data
- Output similarity using inter-system BLEU

For all architectures, data settings, and evaluation metrics, the imitation models closely match their victims.
Imitating Production Models

- Imitate production systems on English-German and Nepali-English
Imitating Production Models

- Imitate production systems on English-German and Nepali-English

- We closely match production systems

<table>
<thead>
<tr>
<th>Model</th>
<th>Google</th>
<th>Bing</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official</td>
<td>32.0</td>
<td>32.9</td>
<td>27.8</td>
</tr>
<tr>
<td>Imitation</td>
<td>31.5</td>
<td>32.4</td>
<td>27.6</td>
</tr>
</tbody>
</table>
Imitating Production Models

- Imitate production systems on English-German and Nepali-English

- We closely match production systems

<table>
<thead>
<tr>
<th>Model</th>
<th>Google</th>
<th>Bing</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official</td>
<td>32.0</td>
<td>32.9</td>
<td>27.8</td>
</tr>
<tr>
<td>Imitation</td>
<td>31.5</td>
<td>32.4</td>
<td>27.6</td>
</tr>
<tr>
<td>Official</td>
<td>32.0</td>
<td>32.7</td>
<td>32.0</td>
</tr>
<tr>
<td>Imitation</td>
<td>31.1</td>
<td>32.0</td>
<td>31.4</td>
</tr>
</tbody>
</table>

In-domain BLEU

Out-of-domain BLEU
Breaking MT Models
Breaking MT Models

- Most adversarial attacks for NLP assume white-box access
  - How to do black-box attacks?
Breaking MT Models

● Most adversarial attacks for NLP assume white-box access
  ○ How to do black-box attacks?

● Simple idea: transfer attacks from imitation models
Most adversarial attacks for NLP assume white-box access
  ○ How to do black-box attacks?

Simple idea: transfer attacks from imitation models

---

Breaking MT Models

- Most adversarial attacks for NLP assume white-box access
  - How to do black-box attacks?
- Simple idea: transfer attacks from imitation models
Most adversarial attacks for NLP assume white-box access
- How to do black-box attacks?

Simple idea: transfer attacks from imitation models

Breaking MT Models

- Replace Token: es ist über 7 ° F
- Replace Token: it's over 7 ° F
- Replace Token: it's over 8 ° F

Diagram:
- German Decoder
- English Encoder

Equation: $\nabla_x \mathcal{L}$
Most adversarial attacks for NLP assume white-box access
○ How to do black-box attacks?

Simple idea: transfer attacks from imitation models
Most adversarial attacks for NLP assume white-box access
  ○ How to do black-box attacks?

Simple idea: transfer attacks from imitation models

Breaking MT Models

- Most adversarial attacks for NLP assume white-box access
  ○ How to do black-box attacks?

- Simple idea: transfer attacks from imitation models
Most adversarial attacks for NLP assume white-box access
○ How to do black-box attacks?

Simple idea: transfer attacks from imitation models

We consider benign but representative adversarial attacks
○ same techniques would apply to malicious use cases
Untargeted Universal Trigger Attack

- A phrase that commonly causes mistranslations when appended to any input
Untargeted Universal Trigger Attack

- A phrase that commonly causes mistranslations when appended to any input

I heard machine translation is now superhuman Siehe Siehe Siehe Siehe Siehe Siehe Siehe

In diesem Jahr ist es wieder soweit: Manche Manuskripte haben sich in der Hauptsache in der Hauptsache wieder in den Vordergrund gestellt.

(This year it’s time again: Some manuscripts the main thing the main thing come to the foreground again)
Untargeted Universal Trigger Attack

● A phrase that commonly causes mistranslations when appended to any input

<table>
<thead>
<tr>
<th>Transfer trigger to Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>I heard machine translation is now superhuman Siehe Siehe Siehe Siehe Siehe Siehe</td>
</tr>
<tr>
<td>In diesem Jahr ist es wieder soweit: Manche Manuskripte haben sich in der Hauptsache in der Hauptsache wieder in den Vordergrund gestellt.</td>
</tr>
<tr>
<td>(This year it’s time again: Some manuscripts the main thing the main thing come to the foreground again)</td>
</tr>
</tbody>
</table>

● 5.8% similarity between Systran’s predictions before and after appending trigger (as measured by modified pairwise BLEU)
Universal Suffix Dropper Attack

- A phrase that commonly causes itself and any subsequent text to be dropped from the translation
Universal Suffix Dropper Attack

- A phrase that commonly causes itself and any subsequent text to be dropped from the translation

**Bing**

Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday, kostenfrei übrigens categories ņstrian hl SM the revolution begins at 6pm

Angesichts der Veröffentlichung des neuesten iPhones lag die Apple-Aktie gestern früh im Handel um 3% im Kurs.

*(Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday)*
Universal Suffix Dropper Attack

- A phrase that commonly causes itself and any subsequent text to be dropped from the translation

Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday,

kostenfrei übrigens categories

the revolution begins at 6pm

Transfer trigger to Bing

76% of messages after trigger are successfully dropped

Angesichts der Veröffentlichung des neuesten iPhones lag die Apple-Aktie gestern früh im Handel um 3% im Kurs.

(Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday)
Targeted Flips Attack

- Cause a specific output token to flip to another specific token
Targeted Flips Attack

- Cause a specific output token to flip to another specific token

Google

I am going to die, it's over 100°F, help!

Ich werde sterben, es ist über 100°F, hilf!
Targeted Flips Attack

- Cause a specific output token to flip to another specific token

<table>
<thead>
<tr>
<th>Google</th>
<th>I am going to die, it's over 100°F, help!</th>
<th>Ich werde sterben, es ist über 100°F, hilf!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>I am going to die, it's over 102°F, help!</td>
<td>Ich werde sterben, es ist über 22°C, hilf!</td>
</tr>
</tbody>
</table>
Targeted Flips Attack

- Cause a specific output token to flip to another specific token

- 22% of attacks transfer to Google

<table>
<thead>
<tr>
<th>Google</th>
<th>I am going to die, it's over 100°F, help!</th>
<th>Ich werde sterben, es ist über 100°F, hilf!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>I am going to die, it's over 102°F, help!</td>
<td>Ich werde sterben, es ist über 22°C, hilf!</td>
</tr>
</tbody>
</table>
Defending Against Stealing
Defending Against Stealing

- What makes a good defense?
  - preserves model accuracy
Defending Against Stealing

What makes a good defense?

- preserves model accuracy
- lowers imitation model accuracy
- reduces adversarial attack transfer
Prediction Poisoning Defense

- Adapt ideas from prediction poisoning ([Orekondy et al. 2020](#))
Prediction Poisoning Defense

- Adapt ideas from prediction poisoning (Orekondy et al. 2020)
Prediction Poisoning Defense

- Adapt ideas from prediction poisoning (Orekondy et al. 2020)

Goal: find a translation $\tilde{y}$ that is similar to the original
Prediction Poisoning Defense

- Adapt ideas from prediction poisoning (Orekondy et al. 2020)

Goal: find a translation $\tilde{y}$ that is similar to the original but induces a different gradient (ideally pointing the opposite direction)
Prediction Poisoning Defense

- Adapt ideas from prediction poisoning (Orekondy et al. 2020)

Goal: find a translation $\tilde{y}$ that is similar to the original but induces a different gradient (ideally pointing in the opposite direction)

Assumption: angular deviations are similar for adversary’s model
How We Find $\hat{y}$

- Generate 100 alternate translations via sampling
How We Find $\tilde{y}$

- Generate 100 alternate translations via sampling
- Pick translation with largest gradient angular deviation
How We Find $\tilde{y}$

- Generate 100 alternate translations via sampling
- Pick translation with largest gradient angular deviation
- Impose minimum similarity to original via BLEU match
How We Find \( \tilde{y} \)

- Generate 100 alternate translations via sampling
- Pick translation with largest gradient angular deviation
- Impose minimum similarity to original via BLEU match
Defenses Can Mitigate Adversarial Threat
Defenses Can Mitigate Adversarial Threat

![Graph showing defenses against adversarial threat]

- **Ideal Defense**
- **Naive Defense (y>x)**

Defender's Model BLEU vs. Adversary's Model BLEU
Defenses Can Mitigate Adversarial Threat

![Graph showing the relationship between defender's model BLEU and adversary's model BLEU, with points labeled as 70 BLEU Match, 80 BLEU Match, 90 BLEU Match, and Undefended.]
Defenses Can Mitigate Adversarial Threat

- Defense reduces adversary’s BLEU more than defender’s
Defenses Can Mitigate Adversarial Threat

- Defense reduces adversary’s BLEU more than defender’s
- Attack transfer drops from 38% to 27% at 70 BLEU Match
Defenses Can Mitigate Adversarial Threat

- Defense reduces adversary’s BLEU more than defender’s
- Attack transfer drops from 38% to 27% at 70 BLEU Match
- Downsides: defense adds compute and hurts defender BLEU
Conclusions

• Hiding models behind a black-box API is not enough!
  ○ Production MT models can be **stolen**
  ○ Production MT models can be **broken**
Conclusions

● Hiding models behind a black-box API is not enough!
  ○ Production MT models can be **stolen**
  ○ Production MT models can be **broken**

● Our defense **mitigates** vulnerabilities, but future work is required
Conclusions

● Hiding models behind a black-box API is not enough!
  ○ Production MT models can be **stolen**
  ○ Production MT models can be **broken**

● Our defense **mitigates** vulnerabilities, but future work is required

Blog, Code, and Paper available