Adversarial Examples in NLP

Sameer Singh
sameer@uci.edu
@sameer_
sameersingh.org

Slides: http://tiny.cc/adversarial
What are Adversarial Examples?

“panda”
57.7% confidence

+ ∈

“gibbon”
99.3% confidence

[Goodfellow et al, ICLR 2015]
What’s going on?

\[
\begin{align*}
\min_{x'} \quad & \| x - x' \| \\
\text{s.t.} \quad & f(x') \neq f(x)
\end{align*}
\]

Fast Gradient Sign Method

\[x' \leftarrow x + \varepsilon \text{sign}(\nabla_x J(x))\]

[Goodfellow et al, ICLR 2015 ]
Applications of Adversarial Attacks

• Security of ML Models
  • Should I deploy or not? What’s the worst that can happen?

• Evaluation of ML Models
  • Held-out test error is not enough

• Finding Bugs in ML Models
  • What kinds of “adversaries” might happen naturally?
  • (Even without any bad actors)

• Interpretability of ML Models?
  • What does the model care about, and what does it ignore?
Challenges in NLP

**Change**
L$_2$ is not really defined for text
What is imperceivable? What is a small vs big change?
What is the right way to measure this?

**Search**
Text is discrete, cannot use continuous optimization
How do we search over sequences?

**Effect**
Classification tasks fit in well, but ...
What about structured prediction? e.g. sequence labeling
Language generation? e.g. MT or summarization
Choices in Crafting Adversaries

Different ways to address the challenges
Choices in Crafting Adversaries

What is a small change?

How do we find the attack?

What does it mean to misbehave?
Choices in Crafting Adversaries

What is a small change?

$$\min_{x'} \| x - x' \| \quad \text{s.t.} \quad f(x') \neq f(x)$$
Change: What is a small change?

![Formula: || x - x' ||]

**Characters**

Pros:
- Often easy to miss
- Easier to search over

Cons:
- Gibberish, nonsensical words
- No useful for interpretability

**Words**

Pros:
- Always from vocabulary
- Often easy to miss

Cons:
- Ungrammatical changes
- Meaning also changes

**Phrase/Sentence**

Pros:
- Most natural/human-like
- Test long-distance effects

Cons:
- Difficult to guarantee quality
- Larger space to search

**Main Challenge:** Defining the distance between $x$ and $x'$
Change: A Character (or few)

\[
x = [\text{"I love movies"}]
\]

\[
x = [\text{I}, \text{ }, \text{I}, \text{O}, \text{V}, \ldots]
\]

\[
x' = [\text{I}, \text{ }, \text{I}, \text{I}, \text{I}, \text{V}, \ldots]
\]

Edit Distance: Flip, Insert, Delete

[ Ebrahimi et al, ACL 2018, COLING 2018 ]
Change: Word-level Changes

\[ x = [ \text{‘I’ ‘like’ ‘this’ ‘movie’ ‘.’} ] \]

Let’s replace this word

**Random word?**
\[ x' = [ \text{‘I’ ‘lamp’ ‘this’ ‘movie’ ‘.’} ] \]

**Word Embedding?**
\[ x' = [ \text{‘I’ ‘really’ ‘this’ ‘movie’ ‘.’} ] \]

**Part of Speech?**
\[ x' = [ \text{‘I’ ‘eat’ ‘this’ ‘movie’ ‘.’} ] \]

**Language Model?**
\[ x' = [ \text{‘I’ ‘hate’ ‘this’ ‘movie’ ‘.’} ] \]

[Jia and Liang, EMNLP 2017]
[Alzantot et. al. EMNLP 2018]
Change: Paraphrasing via Backtranslation

This is a good movie

Translate into multiple languages

Este é um bom filme
c’est un bon film

Use back-translators to score candidates

\[ S(x, x') \propto 0.5 \cdot P(x' | Este é um bom filme) + 0.5 \cdot P(x' | c’est un bon film) \]

\[ S(\text{This is a good movie}, \text{This is a good movie}) = 1 \]
\[ S(\text{This is a good movie}, \text{That is a good movie}) = 0.95 \]
\[ S(\text{This is a good movie}, \text{Dogs like cats}) = 0 \]

\[ \text{[Ribeiro et al ACL 2018]} \]
• Deep representations are supposed to encode meaning in vectors
  • If \((x-x')\) is difficult to compute, maybe we can do \((z-z')\)?
Choices in Crafting Adversaries

\[ \min_{x'} \| x - x' \| \]

s.t. \( f(x') \neq f(x) \)

What is a small change?
Choices in Crafting Adversaries

How do we find the attack?

\[
\min_{x'} \| x - x' \|
\]

s.t. \( f(x') \neq f(x) \)
Search: How do we find the attack?

- Only access predictions (usually unlimited queries)
- Full access to the model (compute gradients)
- Access probabilities

Create $x'$ and test whether the model misbehaves
Create $x'$ and test whether general direction is correct
Use the gradient to craft $x'$

Even this is often unrealistic.
Search: Gradient-based

Or whatever the misbehavior is

1. Compute the gradient
2. Step in that direction (continuous)
3. Find the nearest neighbor
4. Repeat if necessary

Beam search over the above...

[ Ebrahimi et al, ACL 2018, COLING 2018 ]
1. Generate local perturbations
2. Select ones that looks good
3. Repeat step 1 with these new ones
4. Optional: beam search, genetic algo

[Jia and Liang, EMNLP 2017]
[Zhao et al, ICLR 2018]
[Alzantot et. al. EMNLP 2018]
Search: Enumeration (Trial/Error)

1. Make some perturbations
2. See if they work
3. Optional: pick the best one

[Belinkov, Bisk, ICLR 2018]
[Iyyer et al, NAACL 2018]
[Ribeiro et al, ACL 2018]
Choices in Crafting Adversaries

How do we find the attack?

\[
\min_{x'} ||x - x'|| \\
\text{s.t. } f(x') \neq f(x)
\]
Choices in Crafting Adversaries

\[
\min_{x'} ||x - x'||
\]

\[
\text{s.t. } f(x') \neq f(x)
\]

What does it mean to misbehave?
Effect: What does it mean to misbehave?

Classification

Untargeted: any other class
Targeted: specific other class

Other Tasks

MT: Don't attack me! \(\rightarrow\) ¡No me ataques!

NER: Sameer PERSON is a prof at UCI ORG !

Loss-based: Maximize the loss on the example
e.g. perplexity/log-loss of the prediction

Property-based: Test whether a property holds
e.g. MT: A certain word is not generated
NER: No PERSON appears in the output
Evaluation: Are the attacks “good”?

• Are they Effective?
  • Attack/Success rate

• Are the Changes Perceivable? (Human Evaluation)
  • Would it have the same label?
  • Does it look natural?
  • Does it mean the same thing?

• Do they help improve the model?
  • Accuracy after data augmentation

• Look at some examples!
Review of the Choices

- **Effect**
  - Targeted or Untargeted
  - Choose based on the task

- **Search**
  - Gradient-based
  - Sampling
  - Enumeration

- **Evaluation**

- **Change**
  - Character level
  - Word level
  - Phrase/Sentence level

\[
\min_{x'} \| x - x' \| \quad \text{st. } f(x') \neq f(x)
\]
Research Highlights

In terms of the choices that were made
Noise Breaks Machine Translation!

<table>
<thead>
<tr>
<th>Change</th>
<th>Search</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Character Based</td>
<td>Passive; add and test</td>
<td>Machine Translation</td>
</tr>
</tbody>
</table>
## Hotflip

<table>
<thead>
<tr>
<th>Change</th>
<th>Search</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character-based</td>
<td>Gradient-based; beam-search</td>
<td>Machine Translation, Classification, Sentiment</td>
</tr>
<tr>
<td>(extension to words)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### News Classification

South Africa’s historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism. 57% World

South Africa’s historic Soweto township marks its 100th birthday on Tuesday in a mode of optimism. 95% Sci/Tech

### Machine Translation

<table>
<thead>
<tr>
<th>src</th>
<th>Das ist Dr. Bob Childs – er ist Geigenbauer und Psychotherapeut.</th>
</tr>
</thead>
<tbody>
<tr>
<td>adv</td>
<td>Das ist Dr. Bob Childs – er ist Geigenbauer und <strong>Psychotherapeut</strong>.</td>
</tr>
<tr>
<td>src-output</td>
<td>This is Dr. Bob Childs – he’s a wizard maker and a <strong>therapist’s therapist</strong>.</td>
</tr>
<tr>
<td>adv-output</td>
<td>This is Dr. Bob Childs – he’s a brick maker and a <strong>psychopath</strong>.</td>
</tr>
</tbody>
</table>
Search Using Genetic Algorithms

Black-box, population-based search of natural adversary

<table>
<thead>
<tr>
<th>Change</th>
<th>Search</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-based,</td>
<td>Genetic Algorithm</td>
<td>Textual Entailment, Sentiment Analysis</td>
</tr>
<tr>
<td>language model score</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Original Text Prediction: **Entailment** (Confidence = 86%)

**Premise:** A runner wearing purple strives for the finish line.

**Hypothesis:** A runner wants to head for the finish line.

Adversarial Text Prediction: **Contradiction** (Confidence = 43%)

**Premise:** A runner wearing purple strives for the finish line.

**Hypothesis:** A racer wants to head for the finish line.
### Natural Adversaries

Textual Entailment

#### Task 1: Change

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Sentences</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>p: The man wearing blue jean shorts is grilling. h: The man is walking his dog.</td>
<td>Contradiction</td>
</tr>
<tr>
<td>Embedding</td>
<td>h': The man is walking by the dog.</td>
<td>Contradiction → Entailment</td>
</tr>
</tbody>
</table>

**Source Sentence (English)**

- s: People sitting in a dim restaurant eating
- s': People sitting in a living room eating

**Generated Translation (German)**

- Leute, die in einem dim Restaurant essen sitzen.
- Leute, die in einem Wohnzimmeressens sitzen.

(people sitting in a living room)

- Ältere Menschen, die eine Stadtstraße hinuntergehen.
- Ein Mann, der eine Straße entlang spielt.

(A man playing along a street.)

*[Zhao et al, ICLR 2018]*
Semantic Adversaries

Semantically-Equivalent Adversary (SEA)

\[ x \xrightarrow{\text{Backtranslation + Enumeration}} x' \]

<table>
<thead>
<tr>
<th>What color is the tray?</th>
<th>Pink</th>
</tr>
</thead>
<tbody>
<tr>
<td>What colour is the tray?</td>
<td>Green</td>
</tr>
<tr>
<td>Which color is the tray?</td>
<td>Green</td>
</tr>
<tr>
<td>What color is it?</td>
<td>Green</td>
</tr>
<tr>
<td>How color is tray?</td>
<td>Green</td>
</tr>
</tbody>
</table>

Semantically-Equivalent Adversarial Rules (SEARs)

\[ (x, x') \xrightarrow{\text{Patterns in “ diffs”}} \text{Rules} \]

<table>
<thead>
<tr>
<th>Change</th>
<th>Search</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence via Backtranslation</td>
<td>Enumeration</td>
<td>VQA, SQuAD, Sentiment Analysis</td>
</tr>
</tbody>
</table>

[Ribeiro et al, ACL 2018 ]
## Transformation Rules: VisualQA

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Questions / SEAs</th>
<th>f(x)</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP VBZ → WP’s</td>
<td>What has What’s been cut?</td>
<td>Cake Pizza</td>
<td>3.3%</td>
</tr>
<tr>
<td>What NOUN → Which NOUN</td>
<td>What Which kind of floor is it?</td>
<td>Wood Marble</td>
<td>3.9%</td>
</tr>
<tr>
<td>color → colour</td>
<td>What color colour is the tray?</td>
<td>Pink Green</td>
<td>2.2%</td>
</tr>
<tr>
<td>ADV is → ADV’s</td>
<td>Where is Where’s the jet?</td>
<td>Sky Airport</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

[Ribeiro et al, ACL 2018]
## Transformation Rules: SQuAD

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Questions / SEAs</th>
<th>f(x)</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>What VBZ $\rightarrow$ What’s</td>
<td>What is What’s the NASUWT?</td>
<td>Trade union Teachers in Wales</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What NOUN $\rightarrow$ Which NOUN</td>
<td>What resource Which resource was mined in the Newcastle area?</td>
<td>coal wool</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What VERB $\rightarrow$ So what VERB</td>
<td>What was So what was Ghandi's work called?</td>
<td>Satyagraha Civil Disobedience</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What VBD $\rightarrow$ And what VBD</td>
<td>What was And what was Kenneth Swezey's job?</td>
<td>journalist sleep</td>
<td>2%</td>
</tr>
</tbody>
</table>

[Ribeiro et al, ACL 2018]  
Sameer Singh, NAACL 2019 Tutorial
## Transformation Rules: Sentiment Analysis

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Reviews / SEAs</th>
<th>$f(x)$</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>movie $\rightarrow$ film</td>
<td>Yeah, the <em>film</em> pretty much sucked. This is not <em>movie-film</em> making.</td>
<td>Neg</td>
<td>Pos</td>
</tr>
<tr>
<td>film $\rightarrow$ movie</td>
<td>Excellent <em>movie</em>. I’ll give this <em>film-movie</em> 10 out of 10!</td>
<td>Pos</td>
<td>Neg</td>
</tr>
<tr>
<td>is $\rightarrow$ was</td>
<td>Ray Charles <em>was</em> legendary. It <em>is was</em> a really good show to watch.</td>
<td>Pos</td>
<td>Neg</td>
</tr>
<tr>
<td>this $\rightarrow$ that</td>
<td>Now <em>this that</em> is a movie I really dislike. The camera really likes her in <em>this-that</em> movie.</td>
<td>Neg</td>
<td>Pos</td>
</tr>
</tbody>
</table>

[Ribeiro et al, ACL 2018]
Adding a Sentence

Article: Super Bowl 50
Paragraph: “Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”
Question: “What is the name of the quarterback who was 38 in Super Bowl XXXIII?”
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

<table>
<thead>
<tr>
<th>Change</th>
<th>Search</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add a Sentence</td>
<td>Domain knowledge, stochastic search</td>
<td>Question Answering</td>
</tr>
</tbody>
</table>

AddSent
What city did Tesla move to in 1880? (Step 1) Mutate question
Prague
(Step 2) Generate fake answer

What city did Tadakatsu move to in 1881? (Step 3) Convert into statement
Chicago

Tadakatsu moved the city of Chicago to in 1881. (Step 4) Fix errors with crowdworkers, verify resulting sentences with other crowdworkers

Adversary Adds: Tadakatsu moved to the city of Chicago in 1881.
Model Predicts: Chicago
Some Loosely Related Work

Use a broader notions of *adversaries*
CRIAGE: Adversaries for Graph Embeddings

Which link should we add/remove, out of million possible links?
“Should Not Change” / “Should Change”

How do dialogue systems behave when the inputs are perturbed in specific ways?

**Should Not Change**
- *like Adversarial Attacks*
- Random Swap
- Stopword Dropout
- Paraphrasing
- Grammatical Mistakes

**Should Change**
- *Overstability Test*
- Add Negation
- Antonyms
- Randomize Inputs
- Change Entities
Overstability: Anchors

Identify the conditions under which the classifier has the same prediction

[Anchor]

What is the mustache made of? banana

How many bananas are in the picture? 2

[Ribeiro et al, AAAI 2018]
Overstability: Input Reduction

Remove as much of the input as you can without changing the prediction!

**SNLI**
- Premise: Well dressed man and woman dancing in the street
- Original: Two man is dancing on the street
dancing
- Answer: Contradiction
- Confidence: 0.977 → 0.706

**SQUAD**
- In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
- Original: What did Tesla spend Astor’s money on?
- Reduced: did
- Confidence: 0.78 → 0.91

**VQA**
- Original: What color is the flower?
- Reduced: flower?
- Answer: yellow
- Confidence: 0.827 → 0.819
Adversarial Examples for NLP

• Imperceivable changes to the input
• Unexpected behavior for the output
• Applications: security, evaluation, debugging

Challenges for NLP
• **Effect:** What is misbehavior?
• **Change:** What is a small change?
• **Search:** How do we find them?
• **Evaluation:** How do we know it’s good?
Future Directions

• More realistic threat models
  • Give even less access to the model/data

• Defenses and fixes
  • Spell-check based filtering
  • Attack recognition: [Pruthi et al ACL 2019]
  • Data augmentation
  • Novel losses, e.g. [Zhang, Liang AISTATS 2019]

• Beyond sentences
  • Paragraphs, documents?
  • Semantic equivalency → coherency across sentences
References for Adversarial Examples in NLP

Relevant Work (roughly chronological)

- Sentences to QA: [Jia and Liang, EMNLP 2017] link
- Noise Breaks MT: [Belinkov, Bisk, ICLR 2018] link
- Natural Adversaries: [Zhao et al, ICLR 2018] link
- Syntactic Paraphrases: [Iyyer et al NAACL 2018] link
- Hotflip/Hotflip MT: [Ebrahimi et al, ACL 2018, COLING 2018] link, link
- SEARs: [Ribeiro et al, ACL 2018] link
- Genetic Algo: [Alzantot et. al. EMNLP 2018] link
- Discrete Attacks: [Lei et al SysML 2019] link

Surveys

- Adversarial Attacks: [Zhang et al, arXiv 2019] link
- Analysis Methods: [Belinkov, Glass, TAACL 2019] link
- Anchors: [Ribeiro et al, AAAI 2018] link
- Input Reduction: [Feng et al, EMNLP 2018] link
- Graph Embeddings: [Pezeshkpour et. al. NAACL ‘19] link

More Loosely Related Work

- SEARs: [Ribeiro et al, ACL 2018] link
- Genetic Algo: [Alzantot et. al. EMNLP 2018] link
- Discrete Attacks: [Lei et al SysML 2019] link
Thank you!

Work with Matt Gardner and me
as part of
The Allen Institute for
Artificial Intelligence
in Irvine, CA

All levels: pre-docs, PhD interns, postdocs, and research scientists!

Sameer Singh
sameer@uci.edu
@sameer_
Sameersingh.org