Semantically Equivalent Adversarial Rules for Debugging NLP Models

Sameer Singh (UC Irvine)  Carlos Guestrin
NLP / ML models are getting smarter: VQA

What type of road sign is shown

> STOP.

Visual7A [Zhu et al 2016]
NLP / ML models are getting smarter: MC (SQuAD)

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi).

How long is the Rhine? 1230km

BiDAF [Seo et al 2017]
Oversensitivity in images

“panda”
57.7% confidence

+ $\epsilon$

“gibbon”
99.3% confidence
Adversarial examples
What about text?

What type of road sign is shown?

> STOP.
What about text?

What type of road sign is shown?

> STOP.
Semantics matter

What type of road sign is shown:

- STOP.

 Which type of road sign is shown:

- Do not Enter.
The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi).
Adversarial Rules
Generalizing adversaries

- Do not Enter.
- STOP.

What type of road sign is shown?
Semantics matter

What color is the sky?

> Blue.

> Gray.
Semantics matter

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe (after the Danube) at about 1,230 km (760 mi).
Semantics matter

Detailed investigation of Chum Salmon, *Oncorhynchus keta*, showed that these fish digest ctenophores 20 times as fast as an equal weight of shrimps.

What is the *Oncorhynchus* also called?

- chum salmon

> *Oncorhynchus keta*
Adversarial Rules
Semantically Equivalent Adversary (SEA)
Ingredients

1. Semantic score function $S(x, x')$
2. A black box model $f(x)$

Semantically Equivalent

AND

Different prediction
Revisiting adversaries

Find closest example with different prediction

\[ \max_{x'} S(x, x') > \gamma \text{ s.t. } \text{SEA}(x, x') = \]
Semantic Similarity: Paraphrasing

[Model et al, 2017]

Translators

en - pt

en - fr

Portuguese Translation

French Translation

pt - en

fr - en

Score

Good movie

Bom filme

Bon film

Great movie

Movie good

0.35
0.34
0.1
0.001

Language Model comes for free
Finding an adversary

What color is the tray? Pink

What color is the tray? Green

What color is the tray? Green

What color is the tray? Green
Semantically Equivalent Adversarial Rules (SEARs)
From SEAs to Rules

Find SEAs

Candidate Rules

Select Small Rule Set
Proposing Candidate Rules

Candidate Rules:

(What type → Which type)
(What NOUN → Which NOUN)
(WP type → Which type)
(WP NOUN → Which NOUN)

What type of road sign is shown?

What type of road sign is shown?

Which type of road sign is shown?

Which is the person looking at?

Which was I thinking?
From SEAs to Rules
Semantically Equivalent Adversarial Rules (SEARS)

1. High Adversary Count
2. Non-Redundancy

- color → colour
- What NOUN → Which NOUN
Examples: VQA

Visual7a-Telling [Zhu et al 2016]

| SEAR               | Questions / SEAs           | f(x)      | Flips |
|--------------------|----------------------------|-----------|-------|
| WP VBZ → WP’s      | What has What’s been cut?  | Cake Pizza| 3.3%  |
| What NOUN → Which NOUN | Which kind of floor is it? | Wood Marble| 3.9%  |
| color → colour     | What color colour is the tray? | Pink Green| 2.2%  |
| ADV is → ADV’s     | Where is Where’s the jet?  | Sky Airport| 2.1%  |
### Examples: Machine Comprehension

**BiDAF [Seo et al 2017]**

| SEAR               | Questions / SEAs                                      | f(x)                             | Flips |
|--------------------|-------------------------------------------------------|----------------------------------|-------|
| What VBZ $\rightarrow$ What’s | *What is* What’s the NASUWT?                         | Trade union Teachers in Wales    | 2%    |
|                    |                                                       |                                  |       |
| What NOUN $\rightarrow$ Which NOUN | *What resource* Which resource was mined in the Newcastle area? | coal wool | 1%    |
|                    |                                                       |                                  |       |
| What VERB $\rightarrow$ So what VERB | *What was* So what was Ghandi’s work called? | Satyagraha Civil Disobedience   | 2%    |
|                    |                                                       |                                  |       |
| What VBD $\rightarrow$ And what VBD | *What was* And what was Kenneth Swezey’s job? | journalist sleep                 | 2%    |
### Examples: Movie Review Sentiment Analysis

#### FastText [Joulin et al 2016]

| SEAR      | Reviews / SEAs                                      | $f(x)$ | Flips |
|-----------|-----------------------------------------------------|--------|-------|
| movie → film | Yeah, the movie film pretty much sucked.            | Neg    | 2%    |
| film → movie | Excellent film movie                                | Pos, Neg | 1%    |
| is → was   | Ray Charles is was legendary.                      | Pos    | 4%    |
| this → that | Now this that is a movie I really dislike.          | Neg, Pos | 1%    |
Experiments
1. SEAs vs Humans
Set up

Humans → Top scored SEA → SEA (top 5) + Human

Evaluate adversaries for semantic equivalence
How often can SEAs be produced?

Visual Question Answering

|          | Human | SEA | Human + SEA |
|----------|-------|-----|-------------|
| Human    |       |     |             |
| SEA      | 33.6  |     |             |
| Human + SEA | 45   |     |             |

Sentiment Analysis

|          | Human | SEA | Human + SEA |
|----------|-------|-----|-------------|
| Human    |       |     |             |
| SEA      | 33    |     |             |
| Human + SEA | 25.1 |     |             |

SEAs find equivalent adversaries as often as Humans

SEAs + Humans better than Humans
Humans produce different adversaries:

They are so easy to love...

What kind of meat is on the boy’s plate?

How many suitcases?

Photography and directing were on point.

Also great directing and photography.
2. SEARs vs Experts
Part 1: experts come up with rules

Objective: maximize mistakes with good rules

Image | Original
--- | ---
Q: **What color** is the lampshade?  
   Answer:  
   a) A light yellow.  
   b) A bright red.  
   c) A subtle green.  
   d) A vivid orange.  

Q: **Which color** is the lampshade?  
   Answer:  
   a) A light yellow.  
   b) A bright red.  
   c) A subtle green.  
   d) A vivid orange.

Image | After rule
--- | ---
Q: **What food** item is above the burger?  
   Answer:  
   a) Fries.  
   b) Chips.  
   c) Cole slaw.  
   d) Ketchup.  

Q: **Which food** item is above the burger?  
   Answer:  
   a) Fries.  
   b) Chips.  
   c) Cole slaw.  
   d) Ketchup.
Part 2: experts evaluate our SEARs

| Rules to evaluate | Results |
|-------------------|---------|
| List of POS tags  |         |
| Please look at the rule results on the right. |         |
| The current rule is: | replace(What NOUN, Which NOUN). |

**Image** | **Original** | **After rule**
--- | --- | ---
Q: What color are the pots?  
Answer:  
(a) Silver.  
(b) Black.  
(c) White.  
(d) Gold.  
--- | --- | ---
Q: Which color are the pots?  
Answer:  
(a) Silver.  
(b) Black.  
(c) White.  
(d) Gold.  
--- | --- | ---
Q: What color is the lampshade?  
Answer:  
(a) A light yellow.  
(b) A bright red.  
(c) A subtle green.  
(d) A vivid orange.  
--- | --- | ---
Q: Which color is the lampshade?  
Answer:  
(a) A light yellow.  
(b) A bright red.  
(c) A subtle green.  
(d) A vivid orange.  
--- | --- | ---
Q: What animal is running in the background?  
Answer:  
(a) A dog.  
(b) A horse.  
(c) A llama.  
(d) A kangaroo.  
--- | --- | ---
Q: Which animal is running in the background?  
Answer:  
(a) A dog.  
(b) A horse.  
(c) A llama.  
(d) A kangaroo.  

Progress: 1 of 20.
Results

% correct predictions flipped

Visual QA

| Experts | SEARs |
|---------|-------|
| 3       | 14.2  |

Sentiment

| Experts | SEARs |
|---------|-------|
| 3.3     | 10.9  |

Time (minutes)

Visual QA

| Finding Rules | Evaluating SEARs |
|---------------|------------------|
| 16.9          | 10.1             |

Sentiment

| Finding Rules | Evaluating SEARs |
|---------------|------------------|
| 12.9          | 5.4              |
3. Fixing bugs
Closing the loop

Retrain model

Filter out bad rules

Augment training

(color → colour)
(WP VBZ → WP’s)
...

39
Results

% of flips due to bugs

Original  Augmented

Visual QA  

Sentiment  

40
Conclusion

Semantics matter

Models are prone to these bugs

SEAs and SEARs help find and fix them
Semantically Equivalent Adversarial Rules for Debugging NLP Models

Sameer Singh (UC Irvine)  
Carlos Guestrin
Semantic scoring is still a research problem...

| SEAR          | Questions / SEAs                                      | f(x)                      |
|---------------|-------------------------------------------------------|---------------------------|
| on $\rightarrow$ in | What is *on* in the background? | A-building Mountains     |
|               | What is *on* in?                                     | Lights The television    |
| VBP $\rightarrow$ is | Where *are* is the water bottles? | Table Vending machine    |
|               | Where *are* is the people gathered?                  | Living room Kitchen      |
| VERB on $\rightarrow$ VERB | What is *on* the background? | A-building Mountains     |
|               | Where are the planes parked *on*?                    | Concrete Landing strip   |
Problem: not comparable across instances

\[ S(x, x') = \min \left( 1, \frac{P(x'|x)}{P(x|x')} \right) \]
### Examples: VQA

| SEAR                  | Questions / SEAs                      | f(x)         | Flips |
|-----------------------|---------------------------------------|--------------|-------|
| WP VBZ $\rightarrow$ WP's | *What has* What’s been cut?          | Cake Pizza  | 3.3%  |
|                       | *Who is* Who’s holding the baby?      | Woman Man    |       |
| What NOUN $\rightarrow$ Which NOUN | *What* Which kind of floor is it? | Wood Marble | 3.9%  |
|                       | *What* Which color is the jet?        | Gray White   |       |
| color $\rightarrow$ colour | *What* colour colour is the tray?    | Pink Green   | 2.2%  |
|                       | *What* colour colour is the jet?      | Gray Blue    |       |
| ADV is $\rightarrow$ ADV’s | Where is* Where’s the jet?          | Sky Airport  | 2.1%  |
|                       | *How is* How’s the desk?             | Messy Empty  |       |
### Examples: Movie Review Sentiment Analysis

| SEAR      | Reviews / SEAs                                                                 | f(x)     | Flips |
|-----------|-------------------------------------------------------------------------------|----------|-------|
| movie → film | Yeah, the *movie* *film* pretty much sucked.                                   | Neg Pos  | 2%    |
|           | This is not *movie* *film* making.                                             |          |       |
| film → movie | Excellent *film* *movie*.                                                       | Pos Neg  | 1%    |
|           | I’ll give this *film* *movie* 10 out of 10!                                   |          |       |
| is → was  | Ray Charles *is* *was* legendary.                                             | Pos Neg  | 4%    |
|           | It *is* *was* a really good show to watch.                                    |          |       |
| this → that | Now *this* *that* is a movie I really dislike.                                | Neg Pos  | 1%    |
|           | The camera really likes her in *this* *that* movie.                           |          |       |

FastText [Joulin et al 2016]
\[ \text{SEA}(x, x') = 1 \left[ \exists y \ S(x, x') > 2 \land f(x) \neq f(x') \right] \]

\[ \max_{x'} S(x, x') > 2 \ s.t. \ SE\text{A}(x, x') = \]