Questioning Question Answering Answers

Sameer Singh
University of California, Irvine
Questioning Question Answering Answers

Sameer Singh
University of California, Irvine
QA Systems are really good!

Is there a moustache in the picture?

> Yes

What is the moustache made of?

> Banana

QA Systems are really good!

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

Is it doing the right thing?

BiDAF [Seo et al 2017]
We know that they are not
Overstability!

What is the moustache made of?
> Banana

What are the eyes made of?
> Bananas

What is?
> Banana
Oversensitivity to phrasing!

What type of road sign is shown?

Which type of road sign is shown?

Do not Enter.
Oversensitivity to unimportant typos!

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).
QA Systems are brittle

• Our goals are to provide automated tools
  • For both oversensitivity and overstability
  • Can we figure these out automatically, with minimal human time?
  • Can we try to rationalize/explain predictions? analyze the mistakes?

• Hopefully, they help design choices for:
  • Data gathering and annotations
  • Model structure and training
  • Evaluation pipelines
Being Model-Agnostic...

Ignore the internal structure

\[ f(x) \]

- Not restricted to differentiable modules
- Practically easy: not tied to PyTorch, Tflow, etc.
- Study models that you don’t have access to!
Talk Overview

Explaining Predictions

- SEARS: Detecting Oversensitivity

- LIME: Linear Explanations
  - Anchors: Sufficient Conditions
Talk Overview

Explaining Predictions

SEARS: Detecting Oversensitivity

LIME: Linear Explanations

Anchors: Sufficient Conditions
Being Local...

“Global” explanation is too complicated
Being Local...

“Global” explanation is too complicated
Being Local...

“Global” explanation is too complicated

Describe the locally-accurate behavior, using interpretable representations
Talk Overview

Explaining Predictions

SEARS: Detecting Oversensitivity

LIME: Linear Explanations

Anchors: Sufficient Conditions

KDD 2016
LIME: Sparse, Linear Explanations

Identify the important words, and present their relative importance
I think Christianity is the one true religion. If you’d like to know more, send me a note.
LIME on VisualQA

What type of road sign is shown?

> STOP.

What type of road sign is shown?
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).
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).

What is the second longest river in Central and Western Europe?

BiDAF [Seo et al 2017]
Limitations of LIME

Gain understanding of local behavior, but very little generalization...

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).

Unless they run it, the users have little idea of what the answer will be.

Which is the second longest river in Germany’s part of Europe?
Talk Overview

Explaining Predictions

SEARS: Detecting Oversensitivity

LIME: Linear Explanations

Anchors: Sufficient Conditions

AAAI 2018
Anchors: Sufficient Conditions

Identify the conditions under which the classifier has the same prediction
Anchors on VisualQA

What type of road sign is shown?

If question starts with *What* (and is similarly structured) the prediction will be *STOP*.

What type of road sign is shown?

96.8%
Anchors on Visual QA

**Anchor**
- **What** is the mustache made of?  
  - banana

- **How many** bananas are in the picture?  
  - 2
Anchors on Visual QA

<table>
<thead>
<tr>
<th>Anchor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What</strong> is the mustache made of?</td>
</tr>
<tr>
<td><strong>What</strong> is the ground made of?</td>
</tr>
<tr>
<td><strong>What</strong> is the bed made of?</td>
</tr>
<tr>
<td><strong>What</strong> is this mustache?</td>
</tr>
<tr>
<td><strong>What</strong> is the man made of?</td>
</tr>
<tr>
<td><strong>What</strong> is the picture of?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>How <strong>many</strong> bananas are in the picture?</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>How <strong>many</strong> are in the picture?</td>
<td>2</td>
</tr>
<tr>
<td><strong>many</strong> animals the picture?</td>
<td>2</td>
</tr>
<tr>
<td>How <strong>many</strong> people are in the picture?</td>
<td>2</td>
</tr>
<tr>
<td>How <strong>many</strong> zebras are in the picture?</td>
<td>2</td>
</tr>
<tr>
<td>How <strong>many</strong> planes are on the picture?</td>
<td>2</td>
</tr>
</tbody>
</table>
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).

What is the longest river in Central and Western Europe?
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).
User study on VisualQA

Show humans predictions + explanations
Ask them to predict what the model will do in new instances (only if confident)

No explanations
Which is the longest river? Danube

LIME
Which is the longest river? Danube

Anchor
Anchor: “longest river” → Danube

Which is second longest river?
Danube, Rhine, “I don’t know”
Summary of VisualQA Results

- Users are more precise and quicker with anchors.
Anchors: Tools for Overstability

What about Over-sensitivity?
Talk Overview

Explaining predictions

SEARS: Detecting Oversensitivity

LIME: Linear Explanations

Anchors: Sufficient Conditions

ACL 2018
Oversensitivity: Adversarial Examples

Find closest example with different prediction
Oversensitivity in images

But unlikely in the real world (except for attacks)

Adversaries are indistinguishable to humans...

"panda" 57.7% confidence

+ $\varepsilon$

"gibbon" 99.3% confidence
What about text?

Perceptible by humans, unlikely in real world
What about text?

What type of road sign is shown?

What type of road sign is shown?

A single word changes too much!
Semantics matter

What type of road sign is shown?

> Do not Enter.

Which type of road sign is shown?

> STOP.

Bug, and likely in the real world
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).

Not all changes are the same: meaning should be same

How long is the Rhine?

- 1,230 km

How long is the Rhine??

- More than 1,050,000
Characterize via Rules

Find rule that generates many adversaries
Characterizing via Rules

Rule (What NOUN ➔ Which NOUN) - flips 3.9% of examples
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).

Rule (?? → ??) - flips 3% of examples
SEARS: Adversarial Rules

Rules are global and actionable, more interesting than individual adversaries
## SEARS Examples: 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 $\rightarrow$ WP’s</td>
<td><em>What has</em> What’s been cut?</td>
<td>Cake Pizza</td>
<td>3.3%</td>
</tr>
<tr>
<td>What NOUN $\rightarrow$ Which NOUN</td>
<td><em>What</em> Which kind of floor is it?</td>
<td>Wood Marble</td>
<td>3.9%</td>
</tr>
<tr>
<td>color $\rightarrow$ colour</td>
<td>What <em>color</em> colour is the tray?</td>
<td>Pink Green</td>
<td>2.2%</td>
</tr>
<tr>
<td>ADV is $\rightarrow$ ADV’s</td>
<td><em>Where is</em> Where’s the jet?</td>
<td>Sky Airport</td>
<td>2.1%</td>
</tr>
</tbody>
</table>
### SEARS Examples: 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 → What’s</td>
<td><em>What is What’s the NASUWT?</em></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 → Which NOUN</td>
<td><em>What resource Which resource was mined in the Newcastle area?</em></td>
<td>coal wool</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What VERB → So what VERB</td>
<td><em>What was So what was Ghandi's work called?</em></td>
<td>Satyagraha Civil Disobedience</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What VBD → And what VBD</td>
<td><em>What was And what was Kenneth Swezey's job?</em></td>
<td>journalist sleep</td>
<td>2%</td>
</tr>
</tbody>
</table>

BiDAF [Seo et al 2017]
VQA User Study: Detecting adversaries

- SEAs find adversaries as often as humans!
- SEAs + Humans better than humans!
VQA User study: Can experts find bugs?

SEARs are much better than expert-produced rules
Evaluating is much easier than finding them
Closing the loop brings it down to 1.4%
Talk Overview

Explaining Predictions

- SEARS: Detecting Oversensitivity
- LIME: Linear Explanations
- Anchors: Sufficient Conditions
Why such tools can be useful

- **Annotations and Task Definitions**
  - SQuAD 2.0: unanswerable questions
  - VisualQA 2.0: questions with different answers

- **Evaluation**
  - Create robust test set
  - Include explanations/bugs as qualitative evaluation

- **End to End QA may not be sufficient**
  - Salesforce’s NLP Decathlon
  - ELMO Representation: learn across domains, and fine-tune!
Work with Marco T. Ribeiro and Carlos Guestrin, University of Washington

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!