Adversarial Examples for Evaluating Reading Comprehension Systems

Robin Jia and Percy Liang
Stanford University
The number of new Huguenot colonists declined after what year?

The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689...but quite a few arrived as late as 1700; thereafter, the numbers declined...

Correct Answer: “1700”
Progress on SQuAD

Do these models actually understand language?

SQuAD leaderboard, https://rajpurkar.github.io/SQuAD-explorer/
Adversarial Evaluation

Question: “The number of new Huguenot colonists declined after what year?”

Paragraph: “The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689...but quite a few arrived as late as 1700; thereafter, the numbers declined. The number of old Acadian colonists declined after the year of 1675.”

Correct Answer: “1700”

Predicted Answer: “1675”

Model used: BiDAF Ensemble (Seo et al., 2016)
Question: “The number of new Huguenot colonists declined after what year?”

Paragraph: “The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as 1700; thereafter, the numbers declined. expected yet later be basis need young only required 1961.”

Correct Answer: “1700”
Predicted Answer: “1961”

Model used: BiDAF Ensemble (Seo et al., 2016)
Outline

• Inspiration/Motivation
• Adding Grammatical Sentences
• Adding Word Salad
• Trying to build better systems
Outline

• Inspiration/Motivation
• Adding Grammatical Sentences
• Adding Word Salad
• Trying to build better systems
Local perturbations don’t change semantics of image, but models are **oversensitive** to small differences!

Goodfellow et al., 2014.
Local perturbations of text

Question: “The number of new Huguenot colonists declined after what year?”

Paragraph: “The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as 1700; thereafter, the numbers amount declined decreased…”

Plausible alternative answers not always present

Hard to find a lot of perturbations to try

Li et al., 2017
Preserving Semantics

• For images, most local perturbations preserve semantics

• For text, most local perturbations alter semantics
  • Even changing one word by a small amount may not preserve semantics (e.g. entity names)
Concatenative Adversaries

- Instead of locally altering the input, append distracting text to the paragraph
- Must ensure that added text does not actually answer the question
Question: “The number of new Huguenot colonists declined after what year?”

Distracting text: “The number of new Huguenot colonists declined after the year 1675.”

Answer according to text: “1675”
Question: “The number of new Huguenot colonists declined after what year?”

Distracting text: “The number of new old Huguenot Acadian colonists declined after the year 1675.”

Answer according to text: N/A

Local perturbations change semantics of sentence, but models are overly stable/insensitive to these changes!
Outline

• Inspiration/Motivation
• Adding Grammatical Sentences
• Adding Word Salad
• Trying to build better systems
What city did Tesla move to in 1880?
- Change entities, numbers, antonyms
  - Prague

What city did Tadakatsu move to in 1881?
- Generate fake answer with same NER/POS tag
  - Chicago
- Convert to declarative sentence
  - Tadakatsu moved the city of Chicago to in 1881.
- Have crowdworkers fix errors
  - Tadakatsu moved to the city of Chicago in 1881.
Four “dev” systems

SQuAD leaderboard, https://rajpurkar.github.io/SQuAD-explorer/
*Some of our results are on older versions of models than shown here
### Results (4 “dev” systems)

<table>
<thead>
<tr>
<th>System</th>
<th>Original</th>
<th>AddOneSent</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF, ensemble (Seo et al., 2016)</td>
<td>80.0</td>
<td>46.9</td>
</tr>
<tr>
<td>BiDAF, single (Seo et al., 2016)</td>
<td>75.5</td>
<td>45.7</td>
</tr>
<tr>
<td>Match-LSTM, ensemble (Wang &amp; Jiang, 2016)</td>
<td>75.4</td>
<td>41.8</td>
</tr>
<tr>
<td>Match-LSTM, single (Wang &amp; Jiang, 2016)</td>
<td>71.4</td>
<td>39.0</td>
</tr>
<tr>
<td>Human Performance</td>
<td>92.6</td>
<td>89.2</td>
</tr>
</tbody>
</table>
Picking a worst-case sentence

Tadakatsu moved the city of Chicago to in 1881.

Have crowdworkers fix errors

Tadakatsu moved to the city of Chicago in 1881.

Tadakatsu moved to Chicago in 1881.

In 1881, Tadakatsu moved to the city of Chicago.

Model failed if distracted by any of these
Results (4 “dev” systems)

<table>
<thead>
<tr>
<th>System</th>
<th>Original</th>
<th>AddOneSent</th>
<th>AddSent</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF, ensemble (Seo et al., 2016)</td>
<td>80.0</td>
<td>46.9</td>
<td>34.2</td>
</tr>
<tr>
<td>BiDAF, single (Seo et al., 2016)</td>
<td>75.5</td>
<td>45.7</td>
<td>34.3</td>
</tr>
<tr>
<td>Match-LSTM, ensemble (Wang &amp; Jiang, 2016)</td>
<td>75.4</td>
<td>41.8</td>
<td>29.4</td>
</tr>
<tr>
<td>Match-LSTM, single (Wang &amp; Jiang, 2016)</td>
<td>71.4</td>
<td>39.0</td>
<td>27.3</td>
</tr>
<tr>
<td>Human Performance</td>
<td>92.6</td>
<td>89.2</td>
<td>79.5</td>
</tr>
</tbody>
</table>
What city did Tesla move to in 1880?

Adversarial Paragraph

Model

Prague
Gospić
Chicago...

Computers on AddSent

What city did Tesla move to in 1880?

Adversarial Paragraph

Model

Prague
Gospić
Chicago
...

Deterministically choose argmax
Humans on AddSent

What city did Tesla move to in 1880?

Only get noisy samples!
Humans on AddSent

What city did Tesla move to in 1880?

Adversarial Paragraph → Crowd → Prague, Gospić, Chicago, ...

Only get noisy samples!
What city did Tesla move to in 1880?

Humans on AddSent

Adversarial Paragraph #2

Crowd

Prague
Gospić
Chicago

Only get noisy samples!
Humans on AddSent

What city did Tesla move to in 1880?

Adversarial Paragraph #3

Crowd

Prague
Gospić
Chicago
...

Noise augmented when picking worst-case sentence
Twelve “test” systems

SQuAD leaderboard, https://rajpurkar.github.io/SQuAD-explorer/
*Some of our results are on older versions of models than shown here
## Results (12 “test” systems)

<table>
<thead>
<tr>
<th>System</th>
<th>Original</th>
<th>AddOneSent</th>
<th>AddSent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReasoNet, ensemble (Shen et al., 2017)</td>
<td>81.1</td>
<td>49.8</td>
<td>39.4</td>
</tr>
<tr>
<td>SEDT, ensemble (Liu et al., 2017)</td>
<td>80.1</td>
<td>46.5</td>
<td>35.0</td>
</tr>
<tr>
<td>Mnemonic Reader, ensemble (Hu et al., 2017)</td>
<td>79.1</td>
<td>55.3</td>
<td>46.2</td>
</tr>
<tr>
<td>Ruminating Reader (Gong and Bowman, 2017)</td>
<td>78.8</td>
<td>47.7</td>
<td>37.4</td>
</tr>
<tr>
<td>jNet (Zhang et al., 2017)</td>
<td>78.6</td>
<td>47.0</td>
<td>37.9</td>
</tr>
<tr>
<td>Mnemonic Reader, single (Hu et al., 2017)</td>
<td>78.5</td>
<td>56.0</td>
<td>46.6</td>
</tr>
<tr>
<td>ReasoNet, single (Shen et al., 2017)</td>
<td>78.2</td>
<td>50.3</td>
<td>39.4</td>
</tr>
<tr>
<td>MPCM, single (Wang et al., 2016)</td>
<td>77.0</td>
<td>50.0</td>
<td>40.3</td>
</tr>
<tr>
<td>SEDT, single (Liu et al., 2017)</td>
<td>76.9</td>
<td>44.8</td>
<td>33.9</td>
</tr>
<tr>
<td>RaSOR (Lee et al., 2016)</td>
<td>76.2</td>
<td>49.5</td>
<td>39.5</td>
</tr>
<tr>
<td>DCR (Yu et al., 2016)</td>
<td>69.3</td>
<td>45.1</td>
<td>37.8</td>
</tr>
<tr>
<td>Logistic Regression (Rajpurkar et al., 2016)</td>
<td>50.4</td>
<td>30.4</td>
<td>23.2</td>
</tr>
</tbody>
</table>
# Results (12 “test” systems)

<table>
<thead>
<tr>
<th>System</th>
<th>Original</th>
<th>AddOneSent</th>
<th>AddSent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReasoNet, ensemble (Shen et al., 2017)</td>
<td>81.1</td>
<td>49.8</td>
<td>39.4</td>
</tr>
<tr>
<td>SEDT, ensemble (Liu et al., 2017)</td>
<td>80.1</td>
<td>46.5</td>
<td>35.0</td>
</tr>
<tr>
<td><strong>Mnemonic Reader, ensemble (Hu et al., 2017)</strong></td>
<td>79.1</td>
<td><strong>55.3</strong></td>
<td><strong>46.2</strong></td>
</tr>
<tr>
<td>Ruminating Reader (Gong and Bowman, 2017)</td>
<td>78.8</td>
<td>47.7</td>
<td>37.4</td>
</tr>
<tr>
<td>jNet (Zhang et al., 2017)</td>
<td>78.6</td>
<td>47.0</td>
<td>37.9</td>
</tr>
<tr>
<td><strong>Mnemonic Reader, single (Hu et al., 2017)</strong></td>
<td>78.5</td>
<td><strong>56.0</strong></td>
<td><strong>46.6</strong></td>
</tr>
<tr>
<td>ReasoNet, single (Shen et al., 2017)</td>
<td>78.2</td>
<td>50.3</td>
<td>39.4</td>
</tr>
<tr>
<td>MPCM, single (Wang et al., 2016)</td>
<td>77.0</td>
<td>50.0</td>
<td>40.3</td>
</tr>
<tr>
<td>SEDT, single (Liu et al., 2017)</td>
<td>76.9</td>
<td>44.8</td>
<td>33.9</td>
</tr>
<tr>
<td>RaSOR (Lee et al., 2016)</td>
<td>76.2</td>
<td>49.5</td>
<td>39.5</td>
</tr>
<tr>
<td>DCR (Yu et al., 2016)</td>
<td>69.3</td>
<td>45.1</td>
<td>37.8</td>
</tr>
<tr>
<td>Logistic Regression (Rajpurkar et al., 2016)</td>
<td>50.4</td>
<td>30.4</td>
<td>23.2</td>
</tr>
</tbody>
</table>
## Results (12 “test” systems)

<table>
<thead>
<tr>
<th>System</th>
<th>Original</th>
<th>AddOneSent</th>
<th>AddSent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReasoNet, ensemble (Shen et al., 2017)</td>
<td>81.1</td>
<td>49.8</td>
<td>39.4</td>
</tr>
<tr>
<td>SEDT, ensemble (Liu et al., 2017)</td>
<td>80.1</td>
<td>46.5</td>
<td>35.0</td>
</tr>
<tr>
<td>Mnemonic Reader, ensemble (Hu et al., 2017)</td>
<td>79.1</td>
<td>55.3</td>
<td>46.2</td>
</tr>
<tr>
<td>Ruminating Reader (Gong and Bowman, 2017)</td>
<td>78.8</td>
<td>47.7</td>
<td>37.4</td>
</tr>
<tr>
<td>jNet (Zhang et al., 2017)</td>
<td>78.6</td>
<td>47.0</td>
<td>37.9</td>
</tr>
<tr>
<td>Mnemonic Reader, single (Hu et al., 2017)</td>
<td>78.5</td>
<td>56.0</td>
<td>46.6</td>
</tr>
<tr>
<td>ReasoNet, single (Shen et al., 2017)</td>
<td>78.2</td>
<td>50.3</td>
<td>39.4</td>
</tr>
<tr>
<td>MPCM, single (Wang et al., 2016)</td>
<td>77.0</td>
<td>50.0</td>
<td>40.3</td>
</tr>
<tr>
<td>SEDT, single (Liu et al., 2017)</td>
<td>76.9</td>
<td>44.8</td>
<td>33.9</td>
</tr>
<tr>
<td>RaSOR (Lee et al., 2016)</td>
<td>76.2</td>
<td>49.5</td>
<td>39.5</td>
</tr>
<tr>
<td>DCR (Yu et al., 2016)</td>
<td>69.3</td>
<td>45.1</td>
<td>37.8</td>
</tr>
<tr>
<td><strong>Logistic Regression (Rajpurkar et al., 2016)</strong></td>
<td><strong>50.4</strong></td>
<td><strong>30.4</strong></td>
<td><strong>23.2</strong></td>
</tr>
</tbody>
</table>
Partial Matches

Question: “The number of new Huguenot colonists declined after what year?”

Paragraph: “The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689...but quite a few arrived as late as 1700; thereafter, the numbers declined. The number of old Acadian colonists declined after the year of 1675.”

All models distracted by sentences with only partial match with the question
Question: “The number of new Huguenot colonists declined after what year?”

Paragraph: “The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689, in seven ships as part of the organised migration, but quite a few arrived as late as 1700; thereafter, the numbers declined, and only small groups arrived at a time.”

Correct Answer: “1700”
Outline

• Inspiration/Motivation
• Adding Grammatical Sentences
• Adding Word Salad
• Trying to build better systems
Adversarial Word Salad

• So far, only explored tiny fraction of possible distractors

• Try adding **any ungrammatical sequence of words**
  • Incoherent text cannot provide evidence for an incorrect answer
What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for *Prague*…

Model predicts: “*Prague*”

Model used: BiDAF Ensemble (Seo et al., 2016)
What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague... heavy industry art countries applied design theory even medical process.
What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague... heavy industry art countries applied design theory even medical process.

Pick one word at random
What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for **Prague**...heavy industry art countries applied design **city** even medical process.

Replace with another **common word or question word**, to increase probability that model gives a wrong answer.
What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for **Prague**... heavy industry art countries **applied** design city even medical process.

Pick one word at random
What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague... heavy industry art countries what design city even medical process.

Replace with another common word or question word, to increase probability that model gives a wrong answer
What city did Tesla move to in 1880?

*In January 1880, two of Tesla’s uncles put together enough money to help him leave Gospić for Prague… what 30 city 1880 what move city city medical move.*

Model predicts: “medical”

Model used: BiDAF Ensemble (Seo et al., 2016)
AddAny Results

<table>
<thead>
<tr>
<th>System</th>
<th>Original</th>
<th>AddOneSent</th>
<th>AddSent</th>
<th>AddAny</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF, ensemble</td>
<td>80.0</td>
<td>46.9</td>
<td>34.2</td>
<td>2.7</td>
</tr>
<tr>
<td>BiDAF, single</td>
<td>75.5</td>
<td>45.7</td>
<td>34.3</td>
<td>4.8</td>
</tr>
<tr>
<td>Match-LSTM, ensemble</td>
<td>75.4</td>
<td>41.8</td>
<td>29.4</td>
<td>11.7</td>
</tr>
<tr>
<td>Match-LSTM, single</td>
<td>71.4</td>
<td>39.0</td>
<td>27.3</td>
<td>7.6</td>
</tr>
</tbody>
</table>

Models can be fooled on almost any example
Some Inspiration

\[ \text{Panda} \quad \begin{array}{c} \text{58\% confidence} \\ \end{array} \quad + \quad \begin{array}{c} .007 \quad \times \\ \end{array} \quad \text{Nematode} \quad \begin{array}{c} \text{8\% confidence} \\ \end{array} \quad = \quad \text{Gibbon} \quad \begin{array}{c} \text{99\% confidence} \\ \end{array} \]

Goodfellow et al., 2014.
What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague... what 30 city 1880 what move city city medical move.

Model predicts: “medical”

Model used: BiDAF Ensemble (Seo et al., 2016)
What city did Tesla move to in 1880?

*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague…*

Model predicts: *“Prague”*

Model used: BiDAF Ensemble (Seo et al., 2016)
What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague...heavy industry art countries applied design theory even medical process.

Add random common words
What city did Tesla move to in 1880?

*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague...* heavy industry art countries applied design theory even medical process.

Pick one word at random
What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague...heavy industry art countries applied design around even medical process.

Replace with another common word, to increase probability that model gives a wrong answer
What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague...finally back would move york hotel through then immediately later.

Model predicts: “york hotel”

Model used: BiDAF Ensemble (Seo et al., 2016)
# AddCommon Results

<table>
<thead>
<tr>
<th>System</th>
<th>Original</th>
<th>AddOneSent</th>
<th>AddSent</th>
<th>AddAny</th>
<th>AddCommon</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF, ensemble</td>
<td>80.0</td>
<td>46.9</td>
<td>34.2</td>
<td>2.7</td>
<td>52.6</td>
</tr>
<tr>
<td>BiDAF, single</td>
<td>75.5</td>
<td>45.7</td>
<td>34.3</td>
<td>4.8</td>
<td>41.7</td>
</tr>
<tr>
<td>Match-LSTM, ensemble</td>
<td>75.4</td>
<td>41.8</td>
<td>29.4</td>
<td>11.7</td>
<td>51.0</td>
</tr>
<tr>
<td>Match-LSTM, single</td>
<td>71.4</td>
<td>39.0</td>
<td>27.3</td>
<td>7.6</td>
<td>38.9</td>
</tr>
</tbody>
</table>
AddCommon Errors

Question: “What type of markets is the dwelling type below?”

Distracting text: “be therefore dark business business other system type feet above.”

Predicted Answer: “dark business”

Model used: BiDAF Ensemble (Seo et al., 2016)
Question: “After the operators are warned by the escape of the steam, what may they then do?”

Distracting text: “came followed after then such then increased hand law may.”

Predicted Answer: “increased hand law”

Model used: BiDAF Ensemble (Seo et al., 2016)
AddCommon Errors

Question: “Where did he claim the blueprint was stored?”

Distracting text: “doubt was did about carried wasn’t year 1961 near policy.”

Predicted Answer: “near policy”

Model used: BiDAF Ensemble (Seo et al., 2016)
Question: “What act sets the term for judging the boundaries of sanity to which individuals wishing to sit on the SP must adhere?”

Distracting text: “english our programs industry religion size ran maybe leave poor.”

Predicted Answer: “British Nationality Act 1981”

Model used: BiDAF Ensemble (Seo et al., 2016)
Question: “What act sets the term for judging the boundaries of sanity to which individuals wishing to sit on the SP must adhere?”

Paragraph: “As in the House of Commons, a number of qualifications apply to being an MSP. Such qualifications were introduced under the House of Commons Disqualification Act 1975 and the British Nationality Act 1981. Specifically, members must be over the age of 18 and must be a citizen of the United Kingdom, the Republic of Ireland, one of the countries in the Commonwealth of Nations, a citizen of a British overseas territory, or a European Union citizen resident in the UK. Members of the police and the armed forces are disqualified from sitting in the Scottish Parliament as elected MSPs, and similarly, civil servants and members of foreign legislatures are disqualified. An individual may not sit in the Scottish Parliament if he or she is judged to be insane under the terms of the Mental Health (Care and Treatment) (Scotland) Act 2003.

Correct Answer: “Mental Health (Care and Treatment) (Scotland) Act 2003”
Predicted Answer: “British Nationality Act 1981”
Outline

• Inspiration/Motivation
• Adding Grammatical Sentences
• Adding Word Salad
• Trying to build better systems
What can we do?

• We’ve identified weaknesses in existing models—how can we fix them?
Adversarial Training

- What if we train on these adversarial examples?
- Run AddSent without crowdsourcing on training data

What city did Tesla move to in 1880?

Prague

What city did Tadakatsu move to in 1881?

Chicago

Change entities, numbers, antonyms

Generate fake answer with same NER/POS tag

Convert to declarative sentence

Tadakatsu moved the city of Chicago to in 1881.
Adversarial Training

Model used: BiDAF Single (Seo et al., 2016)
Adversarial Training

- Has the model really learned the right thing?
- Create AddSentMod, similar to AddSent
  - Add sentences to beginning instead of end
  - Use different set of fake answers

Prague

Generate fake answer with same NER/POS tag

Chicago
Adversarial Training

• Has the model really learned the right thing?
• Create AddSentMod, similar to AddSent
  • Add sentences to beginning instead of end
  • Use different set of fake answers

Prague

Generate fake answer with same NER/POS tag

Stockholm
Adversarial Training

• Easy to overfit to a given adversary
  • Similar patterns observed with adversarial training in computer vision
Future Work

• Iteratively collect data that’s hard for the model as it trains
• Adversary must be general enough so that overfitting not an issue
Thank you!

All code, data, and experiments available on CodaLab

http://tiny.cc/adversarial-squad-codalab

Thanks to our funding sources!
How good are today’s systems?

<table>
<thead>
<tr>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive AoA Reader, ensemble (HIT + iFLYTEK)</td>
</tr>
<tr>
<td>r-net, ensemble (Microsoft Research Asia)</td>
</tr>
<tr>
<td>r-net, single (Microsoft Research Asia)</td>
</tr>
<tr>
<td>smarnet, ensemble (Eigen Technology &amp; Zhejiang Univ.)</td>
</tr>
<tr>
<td>DCN+, single (Salesforce Research)</td>
</tr>
<tr>
<td>MEMEN, ensemble (Eigen Technology &amp; Zhejiang Univ.)</td>
</tr>
<tr>
<td>ReasoNet, ensemble (Microsoft Research Redmond)</td>
</tr>
<tr>
<td>Mnemonic Reader, ensemble (NUDT &amp; Fudan Univ.)</td>
</tr>
<tr>
<td>Human Performance</td>
</tr>
</tbody>
</table>

SQuAD leaderboard, [https://rajpurkar.github.io/SQuAD-explorer/](https://rajpurkar.github.io/SQuAD-explorer/)
Errors due to distracting text

<table>
<thead>
<tr>
<th>Method</th>
<th>Percent with wrong answer in distractor sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MatchLSTM single</td>
<td>100</td>
</tr>
<tr>
<td>MatchLSTM ensemble</td>
<td>100</td>
</tr>
<tr>
<td>BiDAF single</td>
<td>80</td>
</tr>
<tr>
<td>BiDAF ensemble</td>
<td>100</td>
</tr>
<tr>
<td>Human</td>
<td>20</td>
</tr>
</tbody>
</table>
Adversary Generalization

• Do adversarial examples generated to fool one system also fool other systems?
AddSent Generalization

![Bar chart showing F1 scores for different models and data types.](chart.png)

- **Match, single**
- **Match, ensemble**
- **BiDAF, single**
- **BiDAF, ensemble**

- Match Single Data
- Match Ensemble Data
- BiDAF Single Data
- BiDAF Ensemble Data
- AddOneSent Data

**Border**: data targeting current model
AddAny Generalization

![Graph showing F1 scores for various models]

- **Match, single**
- **Match, ensemble**
- **BiDAF, single**
- **BiDAF, ensemble**

**Match Single Data**

**Match Ensemble Data**

**Border:** data targeting current model
Conclusion

- Evaluation metrics are important!
- Existing models are deficient in many ways
- Some errors can be explained; others are more unintuitive