Evaluating NLU Models with
Harder Generalization Tasks

Atticus Geiger
Overview

- Standard vs non-standard generalization tasks for NLU models
- Adversarial testing
- Artificial tasks
Standard Generalization Tasks

- Find a dataset for your NLU task
- Arbitrarily split your dataset into training, development, and testing sets
- Train a model on the training set and then evaluate performance on unseen testing examples
- This is the standard evaluation framework we have used in this class
Standard Generalization Tasks

- In our third homework, our NLU task was NLI on single words.
- Our edge-disjoint task follows our standard evaluation framework of arbitrarily creating training and testing splits.
- Our word-disjoint task breaks from this standard, presenting the new more difficult task of generalizing to unseen words.
Non-Standard Generalization Tasks

- I want to encourage you to consider breaking from this standard evaluation framework
- We should try to create generalization tasks that are difficult, well motivated, and answer specific questions about model capabilities
Operationalizing an Ambitious Question

● “Can a model learn to comprehend a passage of text?”
● To answer this question, the Stanford Question Answer Dataset, an awesome resource for your projects, was crowd sourced (Rajpurkar et al. 2016)
● We might think that if a model achieves human level performance on the standard generalization task using this dataset, then the model can comprehend a passage of text
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.

Question: What is the name of the quarterback who was 38 in Super Bowl XXXIII?
Answer: John Elway
# SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Human Performance</td>
<td>82.304</td>
<td>91.221</td>
</tr>
<tr>
<td></td>
<td><em>Stanford University</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>(Rajpurkar et al. '16)</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td><strong>BERT (ensemble)</strong></td>
<td>87.433</td>
<td>93.160</td>
</tr>
<tr>
<td></td>
<td><em>Google AI Language</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td><strong>Knowledge-enhanced BERT (single model)</strong></td>
<td>85.944</td>
<td>92.425</td>
</tr>
<tr>
<td></td>
<td><em>Anonymous</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td><strong>nllnet (ensemble)</strong></td>
<td>85.954</td>
<td>91.677</td>
</tr>
<tr>
<td></td>
<td><em>Microsoft Research Asia</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Question answering is solved!

- Triumphant day for AI
- Natural language understanding is essentially a done deal
- Pretty soon we will have conscious robots
- Time to go home
Adversarial Testing (Jia et al. 2017)

- Models trained on SQuAD might not understand language as deeply as we might have hoped
- Systematically perturb examples from training data to generate a test set by appending a misleading sentence
- Use this adversarial test set as your evaluation metric
Train Example

Passage:
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.

Question: What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer: John Elway

Model Prediction: John Elway
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?

Answer: John Elway  
Model Prediction: Jeff Dean
Adversarial Testing

- The average performance of 16 published models trained on SQuAD drops from a 75% F1 score to a 36% F1 score.
Question answering is not solved :(  
  ● Sad day for AI  
  ● Natural language understanding is still super hard  
  ● Time to get back to work
Adversarial Training

- We have found a hole in these models generalization capabilities.
- A natural idea is to patch this hole by including these new examples in training, and this works perfectly well.
- However, when we prepend the misleading sentence instead appending it we have a new adversarial test set our models fail on yet again.
Old Adversarial Test Example

Passage:
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?

Answer: John Elway

Patched Model Prediction: John Elway
Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV. 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.

Question: What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer: John Elway
Adversarial Testing for NLI

- In the last couple years, there has been a growing number of more difficult generalization tasks developed for NLI
- This research has exposed the fragility of models trained on the SNLI and/or MultiNLI dataset
Glockner et al. (2018) create an adversarial test set to expose that models have not fully learned lexical relations.

<table>
<thead>
<tr>
<th>Premise/Hypothesis</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>The man is holding a saxophone</td>
<td>contradiction¹</td>
</tr>
<tr>
<td>The man is holding an electric guitar</td>
<td></td>
</tr>
<tr>
<td>A little girl is very sad.</td>
<td>entailment</td>
</tr>
<tr>
<td>A little girl is very unhappy.</td>
<td></td>
</tr>
<tr>
<td>A couple drinking wine</td>
<td>neutral</td>
</tr>
<tr>
<td>A couple drinking champagne</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Examples from the new test set.
Evaluating Compositionality in NLI models

Nie and Wang et al. (2018) created adversarial testing examples to expose that models have not learned compositional semantics
Evaluating Compositionality in NLI models

Dasgupta et al. (2018) expose that models fail to generalize to a particular compositional frame

A: The woman is more cheerful than the man
B: The woman is not more cheerful than the man
CONTRADICTION
A: The woman is more cheerful than the man
B: The man is not more cheerful than the woman
ENTAILMENT
Adversarial Testing for NLI

- You might wonder what these models have learned, if not lexical or compositional semantics!
- The NLP community has been hill climbing on the original SNLI test set from the moment it was released
- However, this is not the case for these new test sets
- In your projects, consider evaluating your models on these more difficult generalization tasks, where there is so much room for innovation and improvement
Artificial Generalization Tasks

- In my own research, I have constructed artificial NLI datasets
- The premises and hypotheses have the form Quantifier Adjective Noun Negation Adverb Verb Quantifier Adjective Noun
- Quantifiers can be no, some, every, or not every
- Negation and modifiers are optional
- My original intent was to stress NLI models with learning first order logical reasoning
An Example from my Dataset

Every tall human does not kick any large rock contradicts
No human angrily kicks some rock
CompTreeNN Model

I tested standard neural models as well as task specific CompTreeNN model that jointly composes the premise and hypothesis.
## Standard Evaluation on My Data

At first, I only performed a standard evaluation where I arbitrarily split my dataset into training and testing sets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBoW</td>
<td>96.29 ± 0.30</td>
<td>95.4 ± 0.2</td>
<td>95.06 ± 0.22</td>
</tr>
<tr>
<td>LSTM Encoder</td>
<td>96.05 ± 0.29</td>
<td>95.83 ± 0.14</td>
<td>95.61 ± 0.21</td>
</tr>
<tr>
<td>TreeNN</td>
<td>96.20 ± 0.17</td>
<td>96.19 ± 0.15</td>
<td>95.99 ± 0.11</td>
</tr>
<tr>
<td>Attention LSTM</td>
<td>97.50 ± 2.69</td>
<td>95.98 ± 2.23</td>
<td>95.82 ± 2.16</td>
</tr>
<tr>
<td>CompTreeNN</td>
<td>99.85 ± 0.07</td>
<td>99.87 ± 0.06</td>
<td>99.85 ± 0.12</td>
</tr>
</tbody>
</table>
Standard Evaluation on My Data

I discovered that standard neural models fail to encode the identity of verbs, nouns, adverbs, and adjectives while the CompTreeNN performs perfectly.

<table>
<thead>
<tr>
<th>Model</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBoW</td>
<td>96.29 ± 0.30</td>
<td>95.4 ± 0.2</td>
<td>95.06 ± 0.22</td>
</tr>
<tr>
<td>LSTM Encoder</td>
<td>96.05 ± 0.29</td>
<td>95.83 ± 0.14</td>
<td>95.61 ± 0.21</td>
</tr>
<tr>
<td>TreeNN</td>
<td>96.20 ± 0.17</td>
<td>96.19 ± 0.15</td>
<td>95.99 ± 0.11</td>
</tr>
<tr>
<td>Attention LSTM</td>
<td>97.50 ± 2.69</td>
<td>95.98 ± 2.23</td>
<td>95.82 ± 2.16</td>
</tr>
<tr>
<td>CompTreeNN</td>
<td>99.85 ± 0.07</td>
<td>99.87 ± 0.06</td>
<td>99.85 ± 0.12</td>
</tr>
</tbody>
</table>
Non-Standard Evaluation on My Data

- I realized that in a standard evaluation, every possible combination of quantifiers, modifiers, and negation appear in training.
- This meant a model that simply memorizes these combinations could succeed.
- The standard evaluation ended up being far easier than I expected.
Non-Standard Evaluation on My Data

- I decided to construct a train test split that evaluates a model’s ability to perform natural logic reasoning.
- I hand designed a simple baseline model that performs natural logic reasoning MacCartney and Manning (2009) or talk to Bill for more details on natural logic.
- I then created a highly constrained dataset that this baseline model achieves perfect performance on.
Non-Standard Evaluation on My Data

- On this task, standard models fail miserably, with only the CompTreeNN model achieving remotely good performance.
- I believe this new task answers a far deeper question about these model’s logical reasoning capabilities.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBoW</td>
<td>53.99±0.27</td>
</tr>
<tr>
<td>CompTreeNN</td>
<td>80.21±7.71</td>
</tr>
<tr>
<td>TreeNN</td>
<td>53.73±8.36</td>
</tr>
<tr>
<td>LSTM encoder</td>
<td>52.51±2.78</td>
</tr>
<tr>
<td>Attention LSTM</td>
<td>47.28±0.95</td>
</tr>
</tbody>
</table>
Moral of the Story

- Think deeply and carefully about what you learn from your experiments
- Often a generalization task will be far easier than you think
- Consider breaking from our standard evaluation framework to create more challenging generalization tasks that answer specific questions about model capabilities