On Adversarial Removal of Hypothesis-only Bias in Natural Language Inference

Yonatan Belinkov*, Adam Poliak*, Benjamin Van Durme, Stuart Shieber, Alexander Rush

*SEM, Minneapolis, MN
June 7, 2019
Co-Authors

Yonatan Belinkov
Adam Poliak
Benjamin Van Durme
Alexander Rush
Stuart Shieber
Background
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*

entailment    neutral    contradiction
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*

entailment  neutral  contradiction
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*

entailment  neutral  contradiction
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*

entailment  neutral  contradiction
Hypothesis Only Baselines in Natural Language Inference

Adam Poliak¹  Jason Naradowsky¹  Aparajita Haldar¹,²  Rachel Rudinger¹  Benjamin Van Durme¹
¹Johns Hopkins University ²BITS Pilani, Goa Campus, India
{azpoliak,vandurme}@cs.jhu.edu {narad,ahaldar1,rudinger}@jhu.edu

*SEM 2018

Abstract

We propose a hypothesis only baseline for diagnosing Natural Language Inference (NLI). Especially when an NLI dataset assumes inference is occurring based purely on the relationship between a context and a hypothesis, it follows that assessing entailment relations while ignoring the provided context is a degenerate solution. Yet, through experiments on ten distinct NLI datasets, we find that this approach, which we refer to as a hypothesis-only model, is able to significantly outperform a majority-class baseline across a number of NLI datasets. Our analysis suggests that statistical irregularities may allow a model to perform NLI in some datasets beyond what should be achievable without access to the context.

Figure 1: (1a) shows a typical NLI model that encodes the premise and hypothesis sentences into a vector space to classify the sentence pair. (1b) shows our hypothesis-only baseline method that ignores the premise and only encodes the hypothesis sentence.

prescribes the sufficient conditions of such a claim.
Hypothesis Only NLI
Hypothesis Only NLI

Hypothesis: *A woman is sleeping*
Hypothesis Only NLI

Premise:  Ø

Hypothesis: *A woman is sleeping*
Hypothesis Only NLI

Premise: ∅

Hypothesis: A woman is sleeping

entailment neutral contradiction
Hypothesis Only NLI

Premise: ∅

Hypothesis: A woman is sleeping

entailment neutral contradiction
A woman is sleeping
Premises:

Hypothesis: A woman is sleeping
Hypothesis: A woman is sleeping

Premises: A woman sings a song while playing piano
Premises: This woman is laughing at her baby shower

Hypothesis: A woman is sleeping
Premises: A woman with glasses is playing jenga

Hypothesis: A woman is sleeping
Why is she sleeping?
Studies in eliciting norming data are prone to repeated responses across subjects

(see McRae et al. (2005) and discussion in §2 of Zhang et al. (2017)’s Ordinal Common-sense Inference)
Problem:
Hypothesis-only biases mean that models may not learn the true relationship between premise and hypothesis
How to handle such biases?
Strategies for dealing with dataset biases

- **Construct new datasets** (Sharma et al. 2018)
  - $$$
  - More bias
Strategies for dealing with dataset biases

- **Construct new datasets** (Sharma et al. 2018)
  - $$$
  - More bias

- **Filter “easy” examples** (Gururangan et al. 2018)
  - Hard to scale
  - May still have biases (see SWAG $\rightarrow$ BERT $\rightarrow$ HellaSWAG)
Strategies for dealing with dataset biases

● **Construct new datasets** (Sharma et al. 2018)
  ○ $$
  ○ More bias

● **Filter “easy” examples** (Gururangan et al. 2018)
  ○ Hard to scale
  ○ May still have biases (see SWAG → BERT → HellaSWAG)

● **Forgo datasets with known biases**
  ○ Not all bias is bad
  ○ Biased datasets may have other useful information
Our solution:
Design architectures that facilitate learning less biased representations.
Adversarial Learning to the Rescue
NLI Model Components

\[ g \quad \text{g – classifier} \]

\[ f \quad \text{f - encoder} \]

\[ f_p \quad f_h \]

\[ p \quad h \]
Baseline NLI Model

$g$

$p$

$f_p$

$h$

$f_h$
Method 1 –
Adv. Hypothesis-Only Classifier

\[ g \]

\[ f_p \]
\[ p \]

\[ f_h \]
\[ h \]
Method 1 – Adv. Hypothesis-Only Classifier
Method 1 – Adv. Hypothesis-Only Classifier

Reverse gradients: Penalize hypothesis encoder if classifier does well
Method 2 –
Adv. Training Examples

\[ g \]

\[ f_p \quad f_h \]

\[ p \quad h \]
Method 2 – Adv. Training Examples

- Randomly swap premises
- Reverse gradients into hypothesis encoder

\[ g \]

\[ f_p \]

\[ p' \]

\[ f_h \]

\[ h \]
Results & Analysis
What happens to model performance?
Degradation in domain
Degradation in domain
Are biases removed?
Hidden biases - Adversarial Classifier

![Graph showing accuracy vs. λ\text{Loss}, λ\text{Enc} with lines for Hyp Only, Random, cNLI, and Majority.]
Hidden biases - Adversarial Classifier

![Graph showing accuracy vs \(\lambda_{\text{Loss}}, \lambda_{\text{Enc}}\)]

- Hyp Only
- Random
- \(c_{\text{NLI}}\)
- \(c_{\text{Hypothesis}}\)
- Majority
Hidden biases - Adversarial Classifier

Accuracy

$\lambda_{\text{Loss}}, \lambda_{\text{Enc}}$

Hyp Only
Random
$c_{\text{NLI}}$
$c_{\text{Hypothesis, retrained}}$
$c_{\text{Hypothesis}}$

Majority

1, 1 2, 2 3, 3 4, 4 5, 5
Hidden biases - Adversarial Data
Hidden biases - Adversarial Data

Accuracy

$\lambda_{Loss}, \lambda_{Enc}$

Hyp Only
Random

$c_{NLI}$
$c_{Hypoth, retrained}$

Majority
What happens to specific biases?
Indicator Words

<table>
<thead>
<tr>
<th>Contradiction</th>
<th>Word</th>
<th>Score</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>nobody</td>
<td>sleeping</td>
<td>0.88</td>
<td>108</td>
</tr>
<tr>
<td>sleeping</td>
<td>driving</td>
<td>0.81</td>
<td>53</td>
</tr>
<tr>
<td>Nobody</td>
<td>alone</td>
<td>1.00</td>
<td>52</td>
</tr>
<tr>
<td>alone</td>
<td>cat</td>
<td>0.90</td>
<td>50</td>
</tr>
<tr>
<td>cat</td>
<td>asleep</td>
<td>0.84</td>
<td>49</td>
</tr>
<tr>
<td>asleep</td>
<td>no</td>
<td>0.91</td>
<td>43</td>
</tr>
<tr>
<td>no</td>
<td>empty</td>
<td>0.84</td>
<td>31</td>
</tr>
<tr>
<td>empty</td>
<td>eats</td>
<td>0.93</td>
<td>28</td>
</tr>
<tr>
<td>eats</td>
<td>sleeps</td>
<td>0.83</td>
<td>24</td>
</tr>
<tr>
<td>sleeps</td>
<td></td>
<td>0.95</td>
<td>20</td>
</tr>
</tbody>
</table>

Decrease in correlation with contradiction

![Graph showing decrease in correlation with contradiction]
What is this good for?
Are less biased models more transferable?
Don’t Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference

Yonatan Belinkov\textsuperscript{1,3} \quad Adam Poliak\textsuperscript{2,}\textsuperscript{*} \\
Stuart M. Shieber\textsuperscript{1} \quad Benjamin Van Durme\textsuperscript{2} \quad Alexander Rush\textsuperscript{1} \\
\textsuperscript{1}Harvard University \quad \textsuperscript{2}Johns Hopkins University \quad \textsuperscript{3}Massachusetts Institute of Technology \\
{belinkov, shieber, srush}@seas.harvard.edu \quad {azpoliak, vandurme}@cs.jhu.edu

Abstract

Natural Language Inference (NLI) datasets often contain hypothesis-only biases—artifacts that allow models to achieve non-trivial performance without learning whether a premise entails a hypothesis. We propose two probabilistic methods to build models that are NLI datasets contain biases, or annotation artifacts, that enable models to perform surprisingly well using only the hypothesis, without learning the relationship between two texts (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018).\textsuperscript{3} For instance, in some datasets, negation words like “not” and “nobody” are often associated with a re-
Method 1 – Adv. Hypothesis-Only Classifier
Method 2 –
Adv. Training Examples

Scitail | Add-one | JOCI | MPE | DPR | MNLI matched | FN+ | MNLI mis | SICK | GLUE | SPR
--- | --- | --- | --- | --- | --- | --- | --- | --- | --- | ---
Baseline | Method 2
Conclusions

- Adversarial learning may help combat hypothesis-side biases in NLI
- Applicable to other tasks with one-sided biases: reading comprehension, visual question answering, etc.
Adversarial Regularization for Visual Question Answering: Strengths, Shortcomings, and Side Effects

Gabriel Grand\textsuperscript{1} and Yonatan Belinkov\textsuperscript{1,2}

\textsuperscript{1}Harvard John A. Paulson School of Engineering and Applied Sci
\textsuperscript{2}MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA, USA
ggrand@alumni.harvard.edu, belinkov@seas.harvard.edu

Abstract

Visual question answering (VQA) models have been shown to over-rely on linguistic biases in VQA datasets, answering ques-

tions that are too broad, generic, or too specific to allow for a ground-truth answer. Efforts to address this problem have mainly focused on constructing more balanced datasets (Zhang et al., 2016; Goyal et al., 2017; Johnson et al., 2017; Chao et al., 2018). However, any benchmark that involves crowdsourced data
Conclusions

- Adversarial learning may help combat hypothesis-side biases in NLI
- Applicable to other tasks with one-sided biases
- May reduce the amount of bias and improve transferability
- But, the methods should be handled with care
  - Not all bias may be removed
  - The goal matters: some bias may be helpful in certain scenarios

Acknowledgements