Harvard John A. Paulson
School of Engineering
and Applied Sciences

JOHNS HOPKINS
UNIVERSITY

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***

neutral contradiction

Natural Language Inference

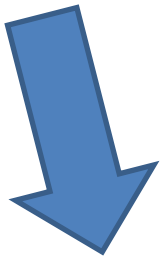
Premise: ***The brown cat ran***

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 Poliak[1] **Jason Naradowsky**[1] **Aparajita Haldar**[1,2]
Rachel Rudinger[1] **Benjamin Van Durme**[1]
[1]Johns Hopkins University [2]BITS Pilani, Goa Campus, India
{azpoliak,vandurme}@cs.jhu.edu {narad,ahaldar1,rudinger}@jhu.edu

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.

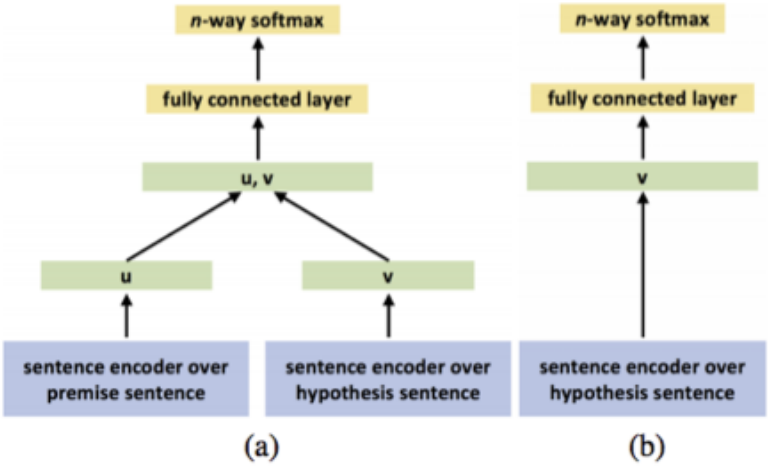


Figure 1: (1a) shows a typical NLI model that encodes the premise and hypothesis sentences into a vector space to classify the sentence pair. (1b) shows our hypothesis-only baseline method that ignores the premise and only encodes the hypothesis sentence.

prescribe the sufficient conditions of such a claim

Hypothesis Only NLI

Hypothesis Only NLI

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

Hypothesis Only NLI

Premise: 

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

Hypothesis Only NLI

Premise: 

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

entailment neutral contradiction

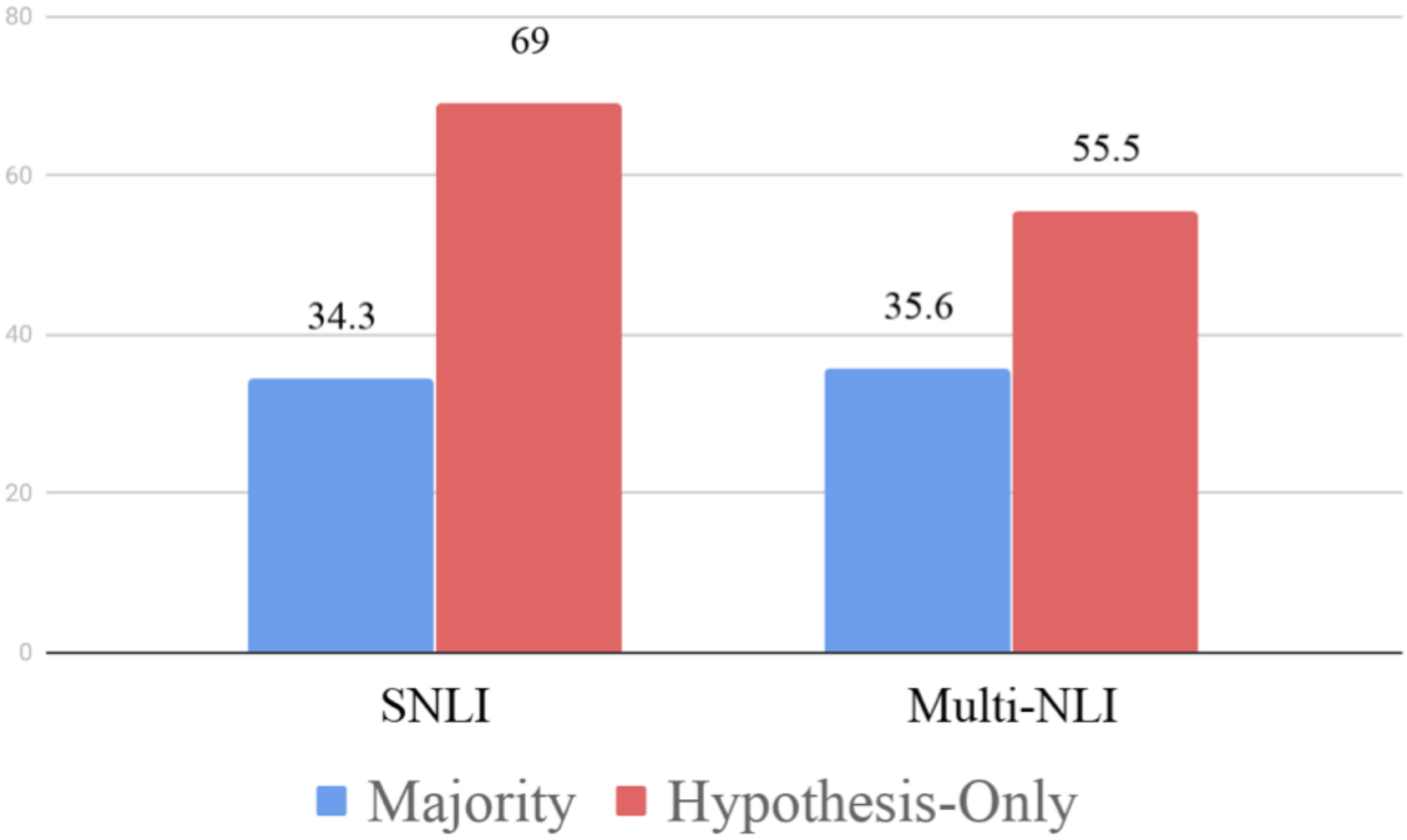
Hypothesis Only NLI

Premise: 

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

entailment neutral contradiction

SNLI Results



A woman is sleeping

Premises:

Hypothesis: *A woman is sleeping*

Premises:

A woman sings a song while playing piano

Hypothesis: *A woman is sleeping*

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 → BERT → HellaSWAG)

Strategies for dealing with dataset biases

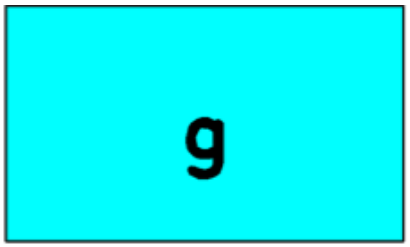
- Construct new datasets (Sharma et al. 2018)
  - $$$
  - More bias
- Filter “easy” examples (Gururangan et al. 2018)
  - Hard to scale
  - May still have biases (see SWAG → BERT → HellaSWAG)
- Forgo datasets with known biases
  - Not all bias is bad
  - Biased datasets may have other useful information

Our solution:

Design architectures that facilitate learning less biased representations

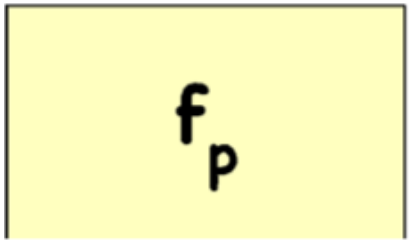
Adversarial Learning to the Rescue

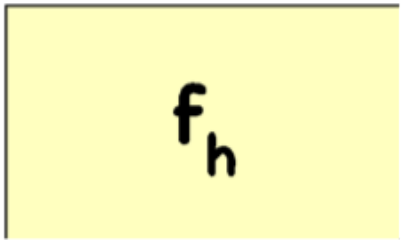
NLI Model Components



g – classifier

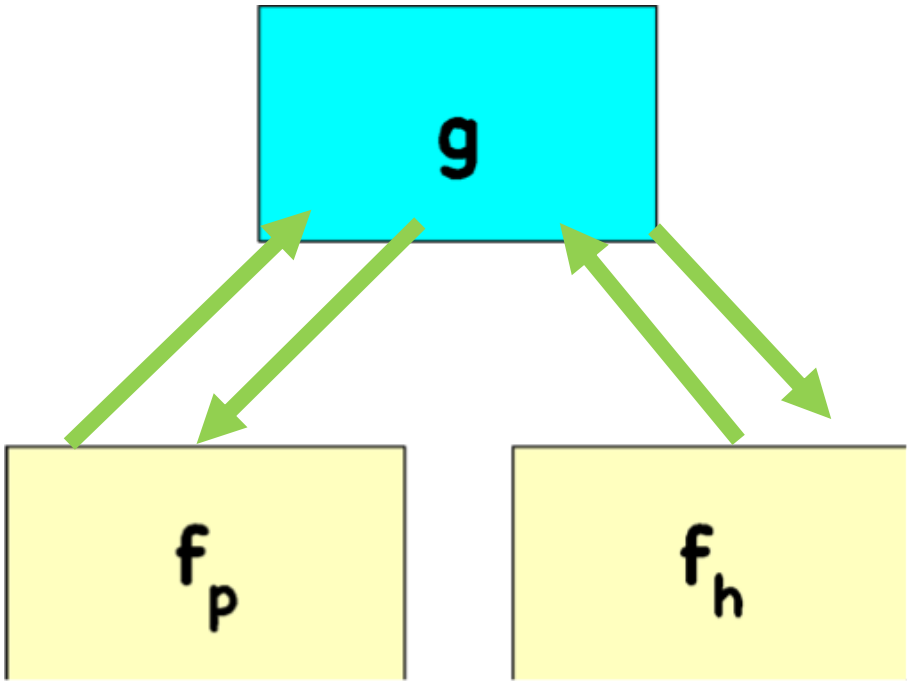
f - encoder



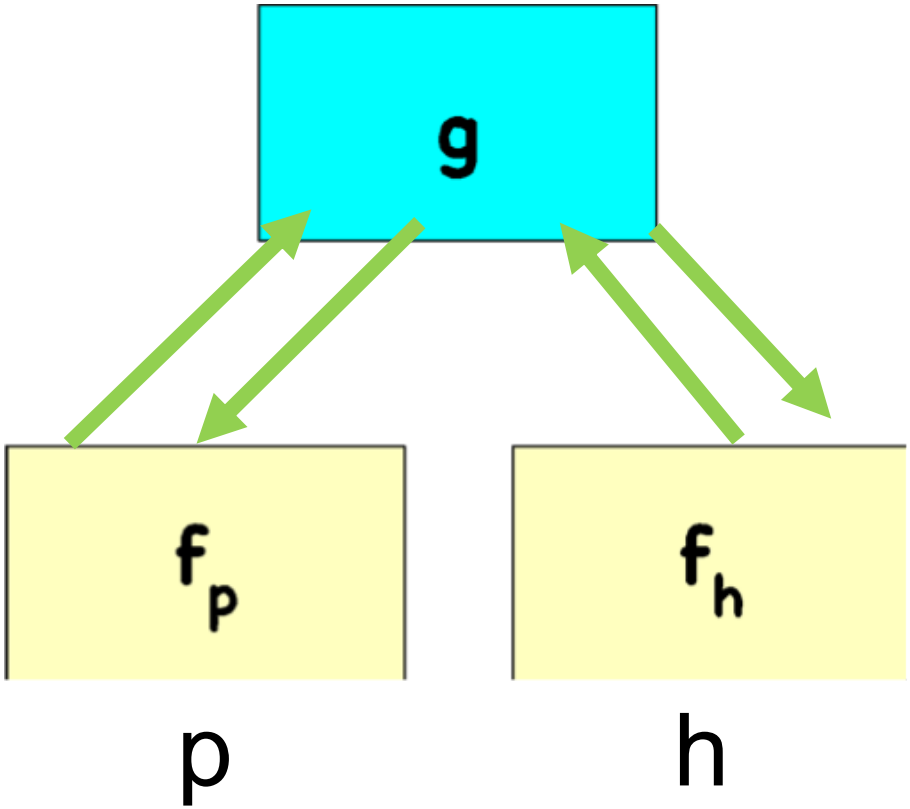


p h

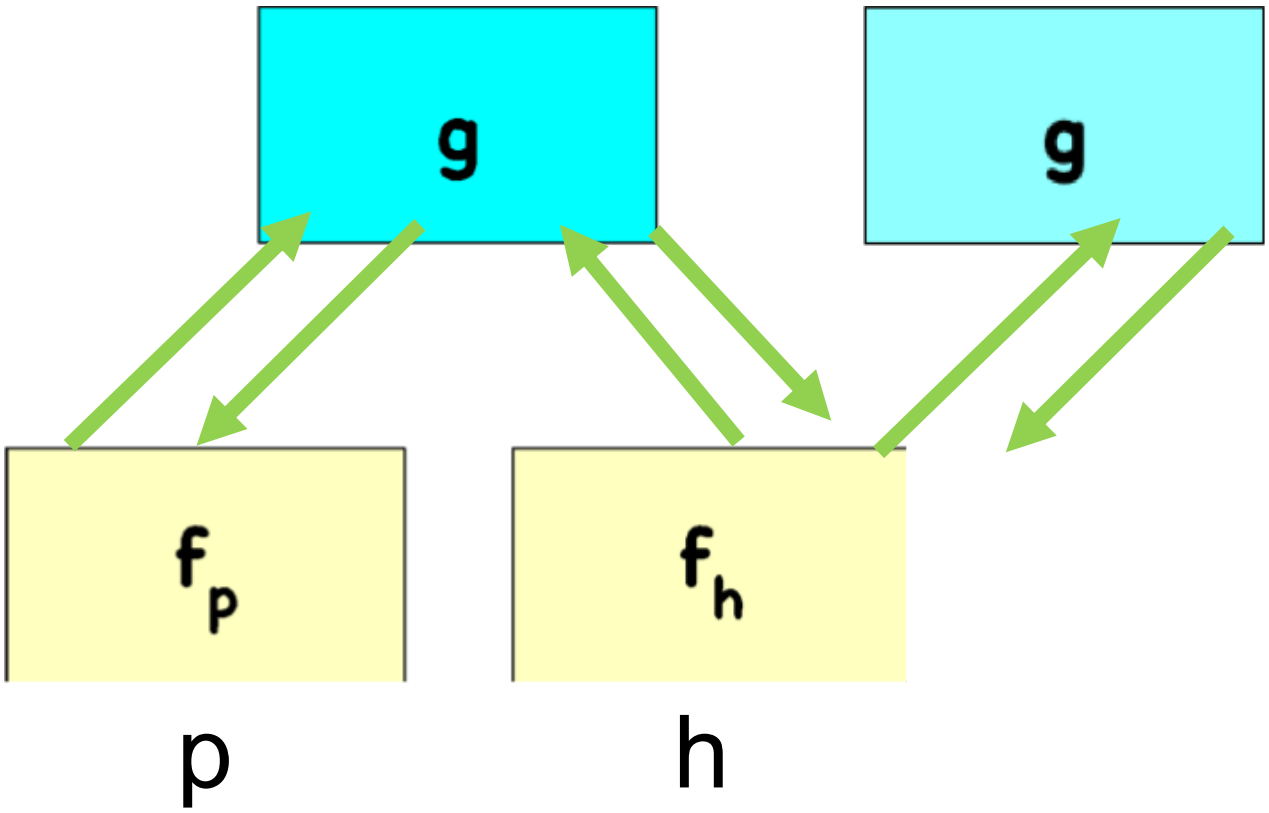
Baseline NLI Model



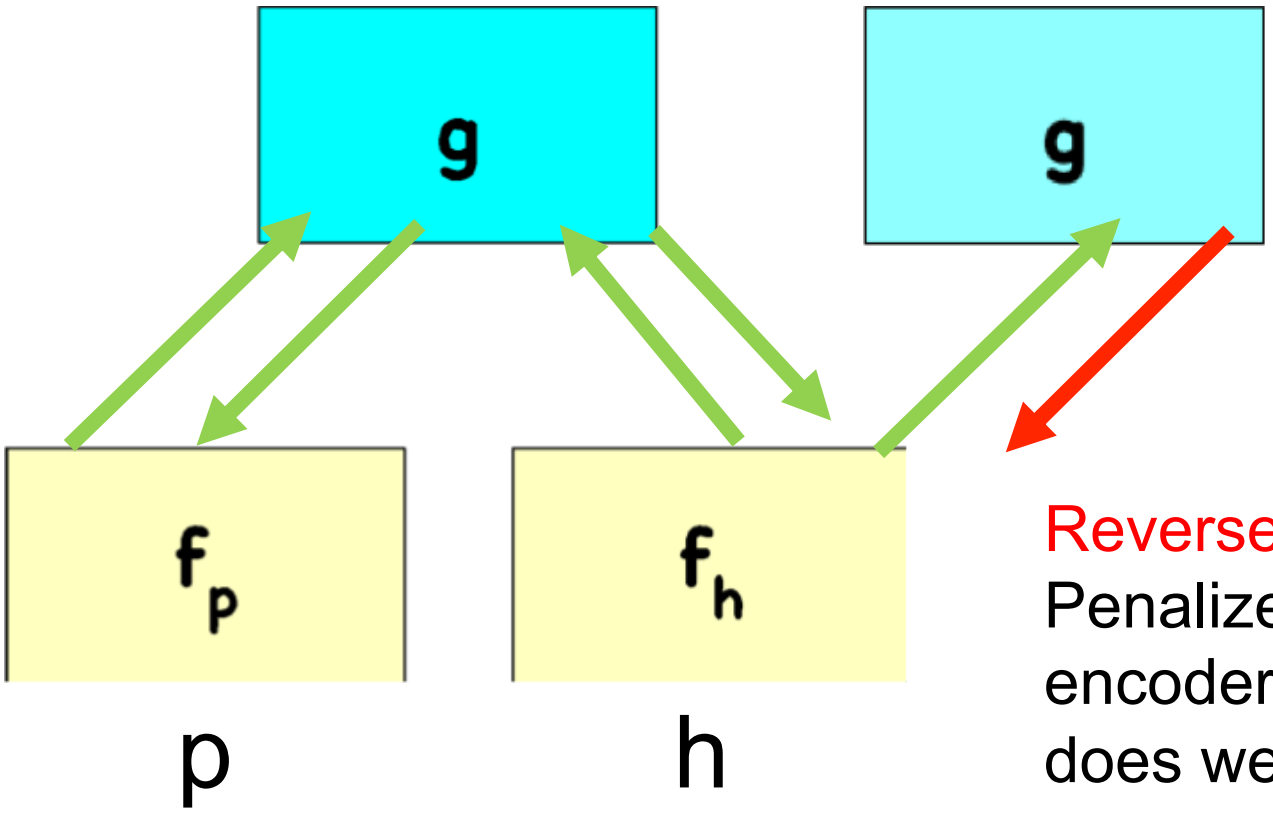
Method 1 – Adv. Hypothesis-Only Classifier



Method 1 – Adv. Hypothesis-Only Classifier

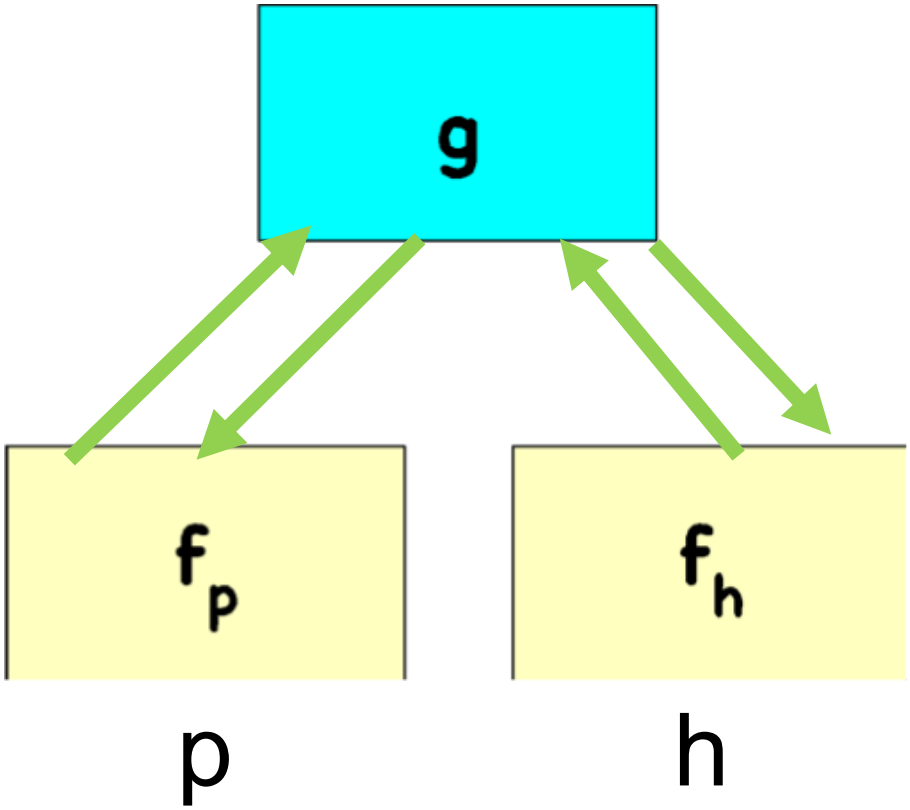


Method 1 – Adv. Hypothesis-Only Classifier

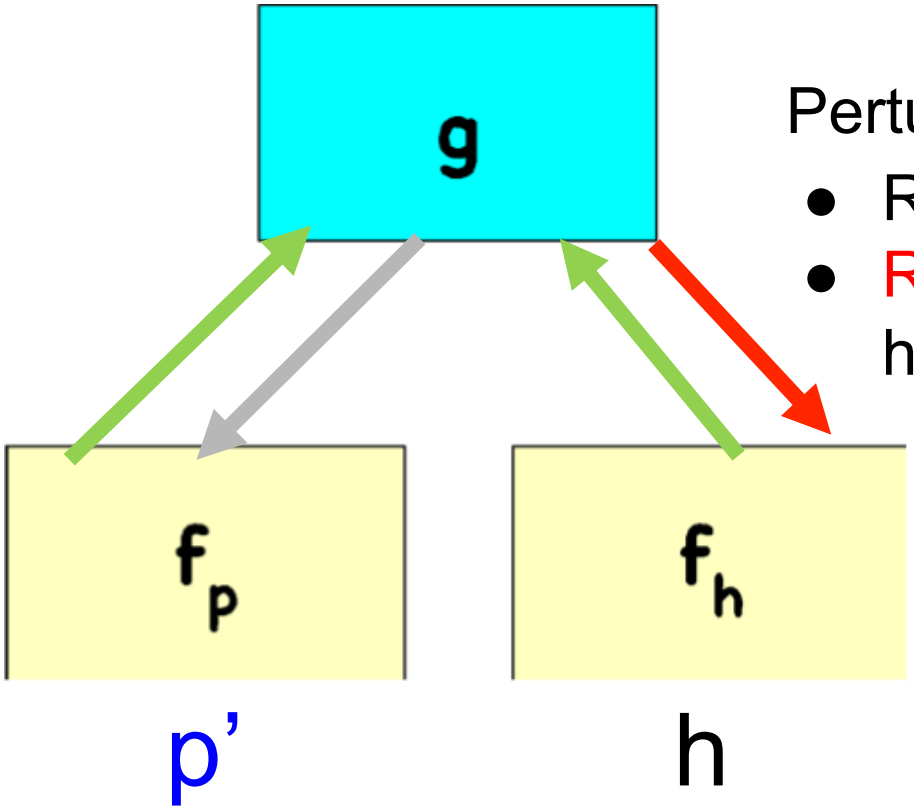


Reverse gradients: Penalize hypothesis encoder if classifier does well

Method 2 – Adv. Training Examples



Method 2 – Adv. Training Examples



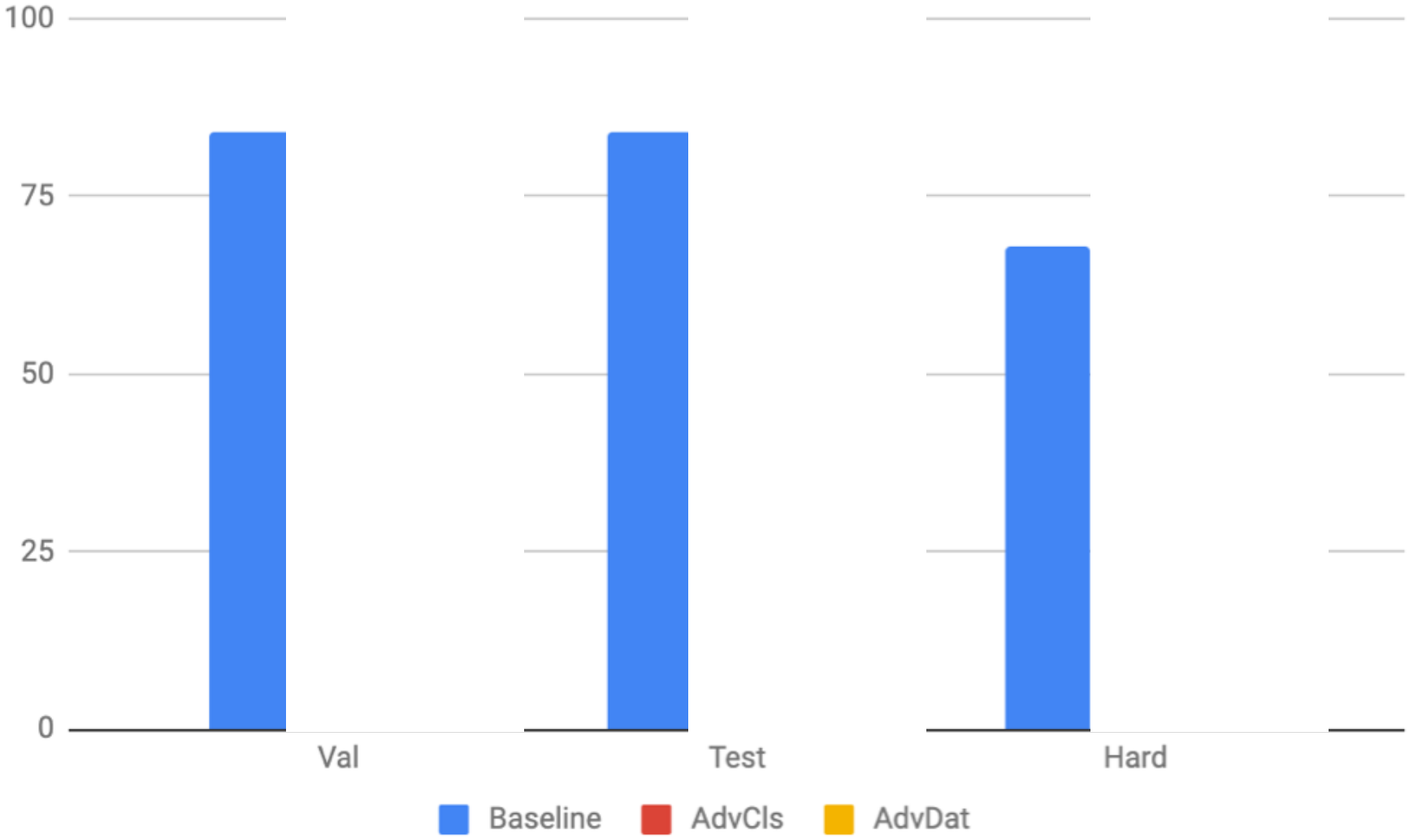
Perturb training examples

- Randomly swap premises
- Reverse gradients into hypothesis encoder

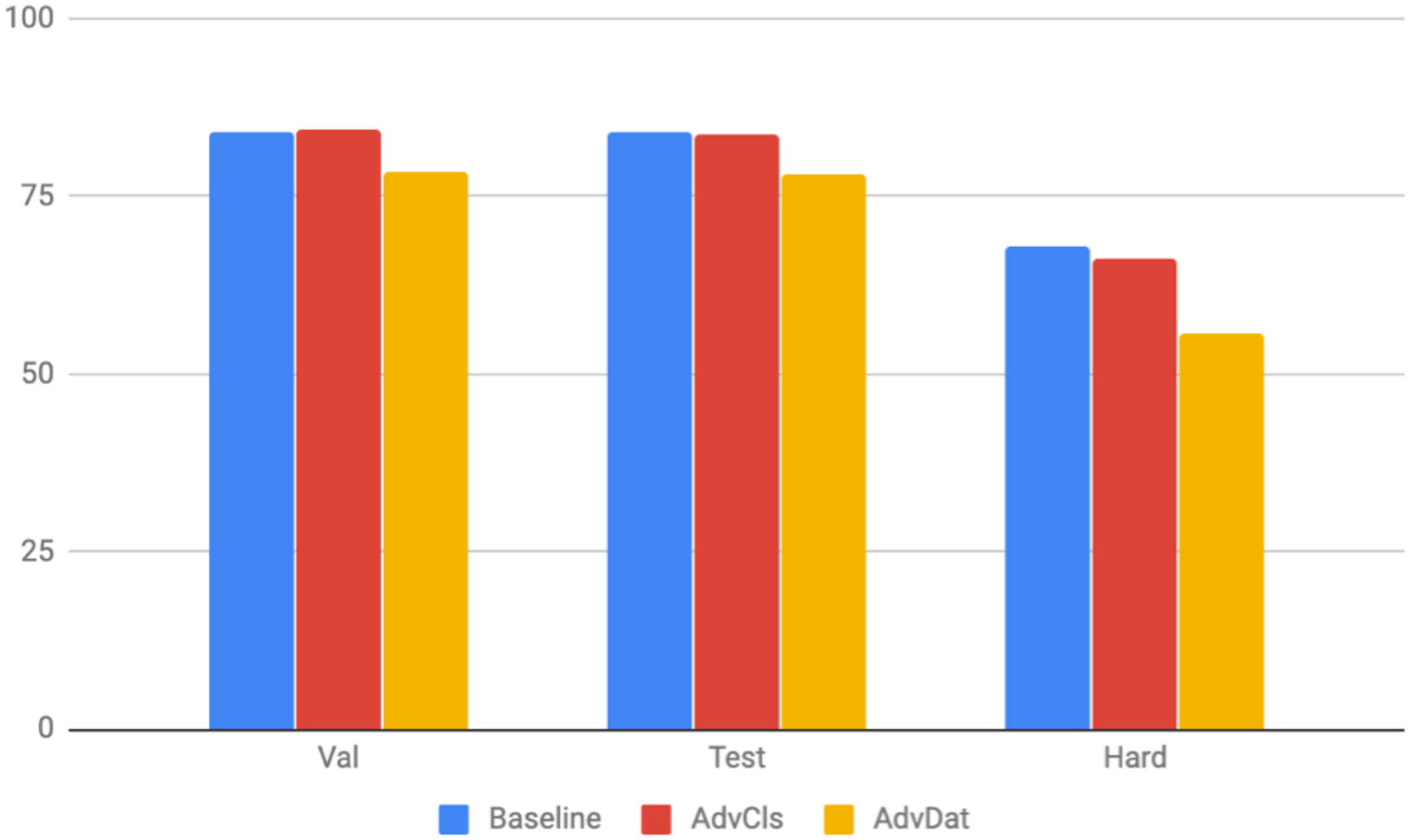
Results & Analysis

What happens to model performance?

Degradation in domain

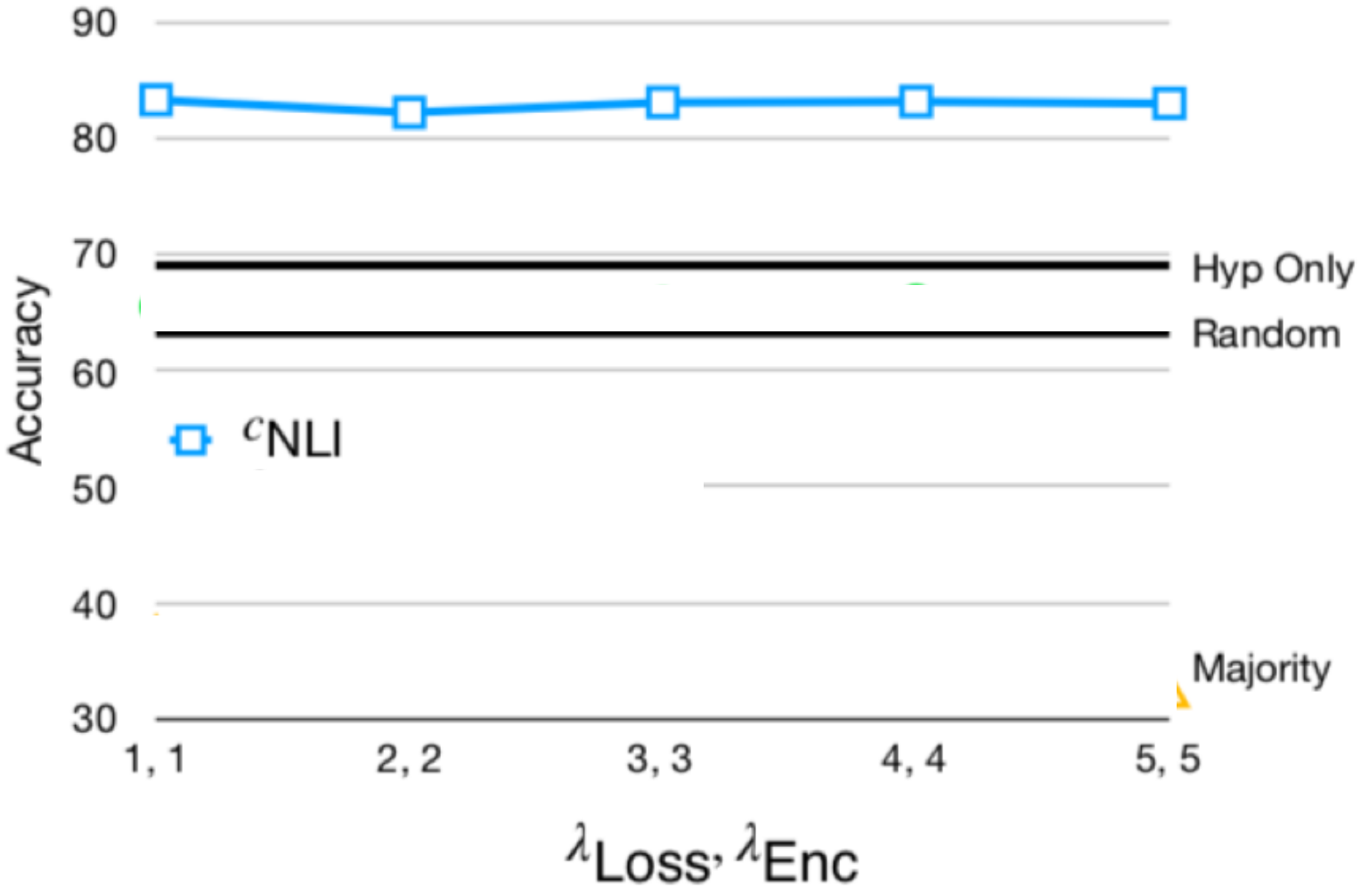


Degradation in domain

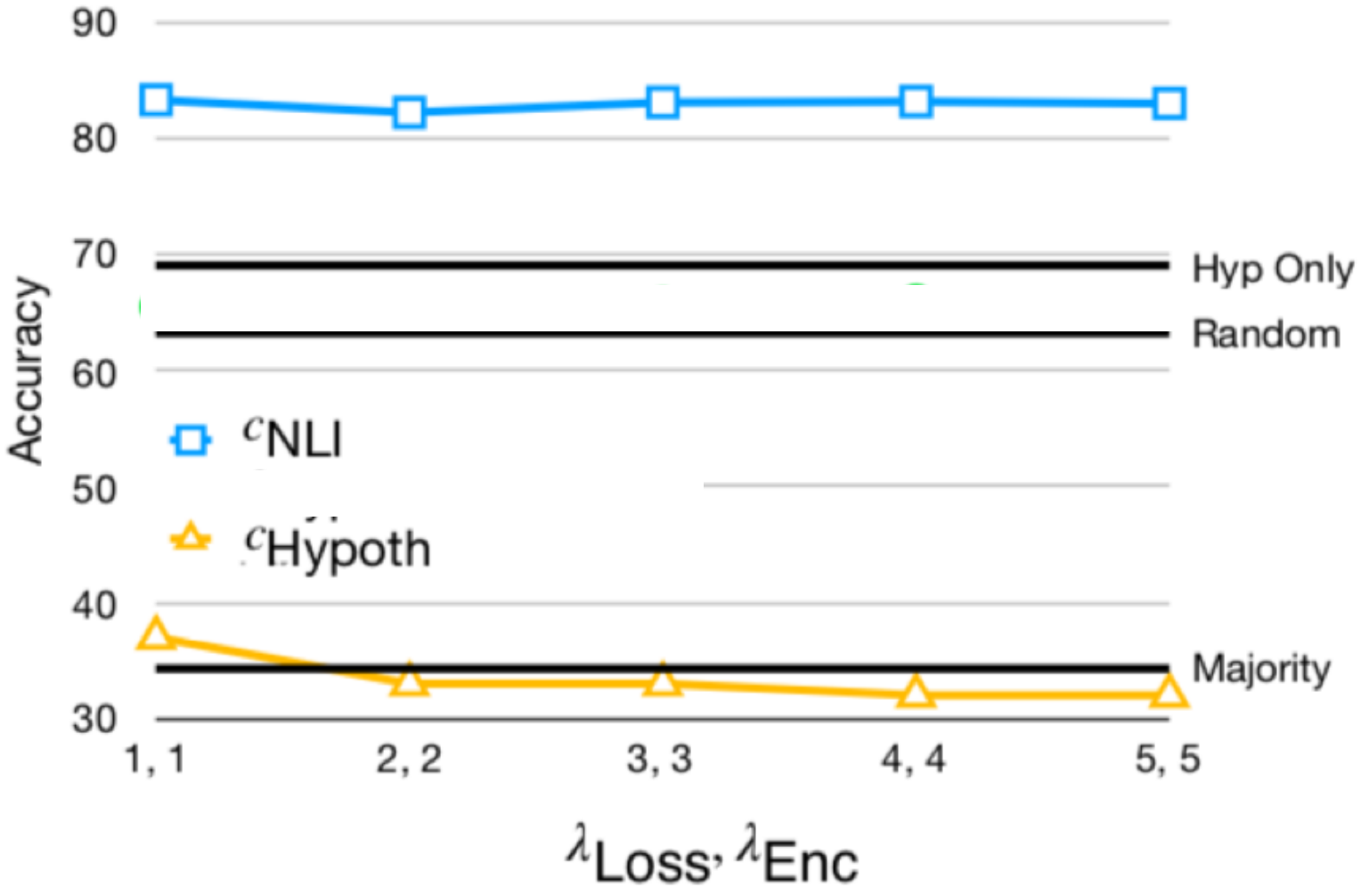


Are biases removed?

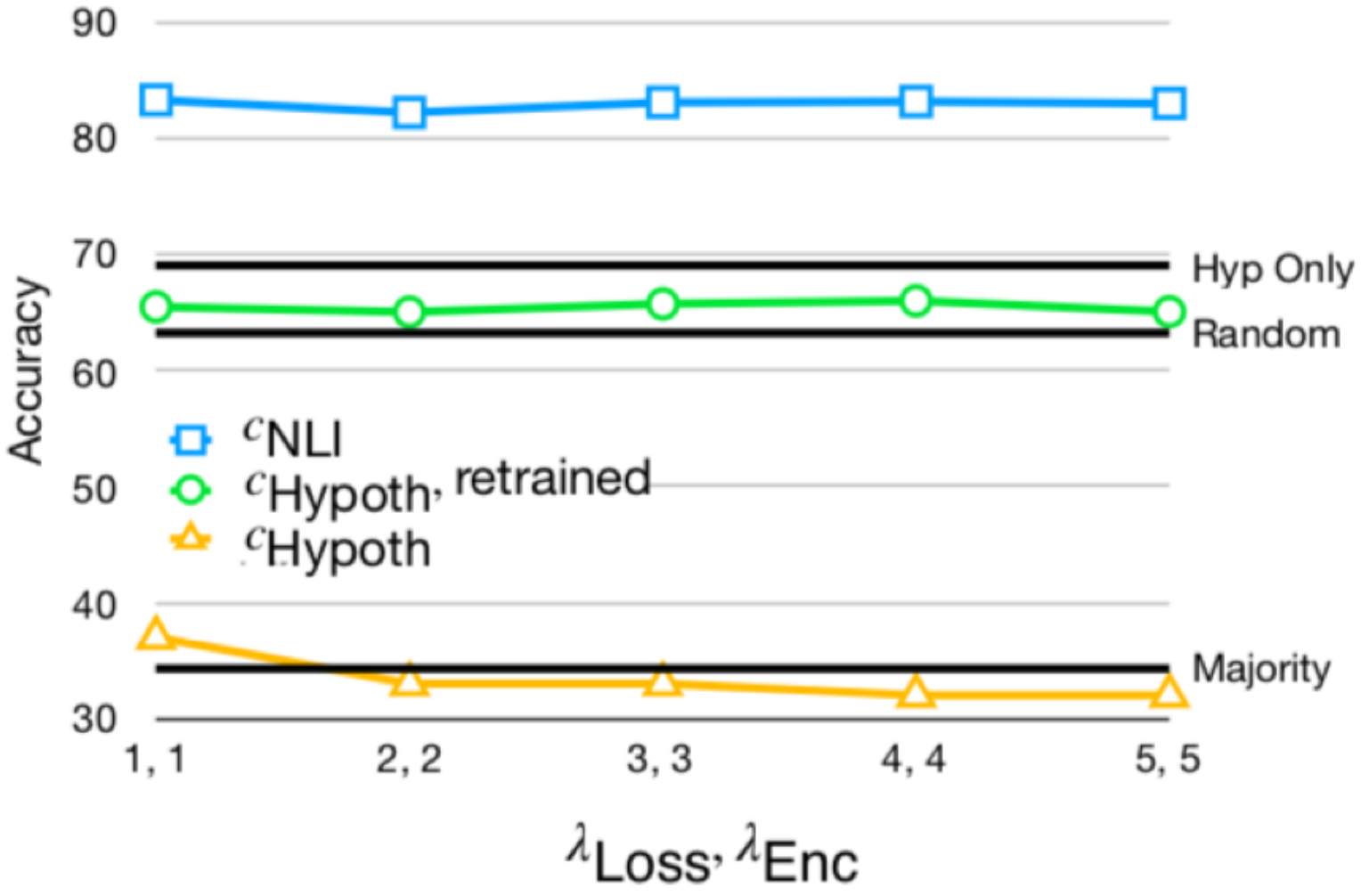
Hidden biases - Adversarial Classifier



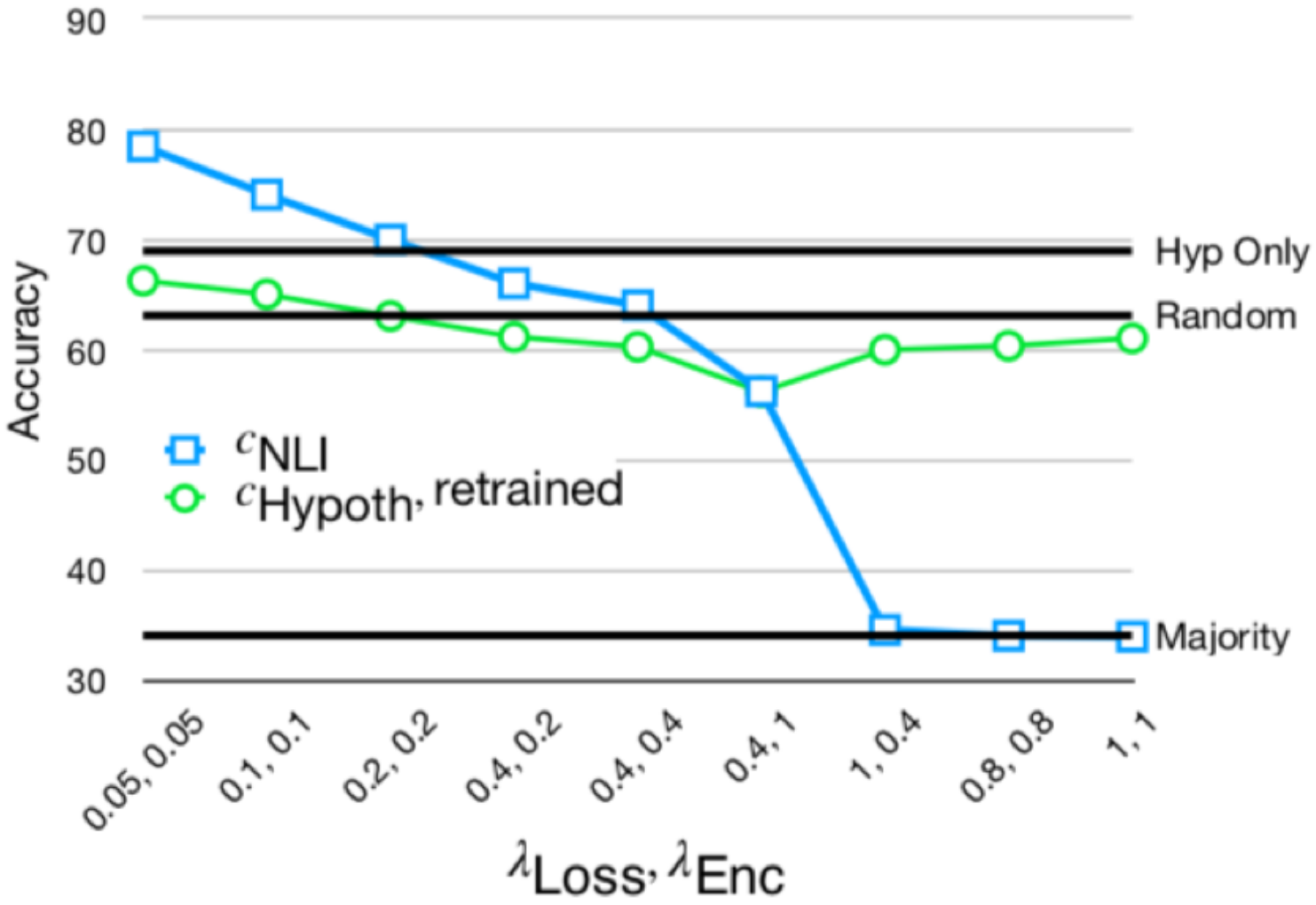
Hidden biases - Adversarial Classifier



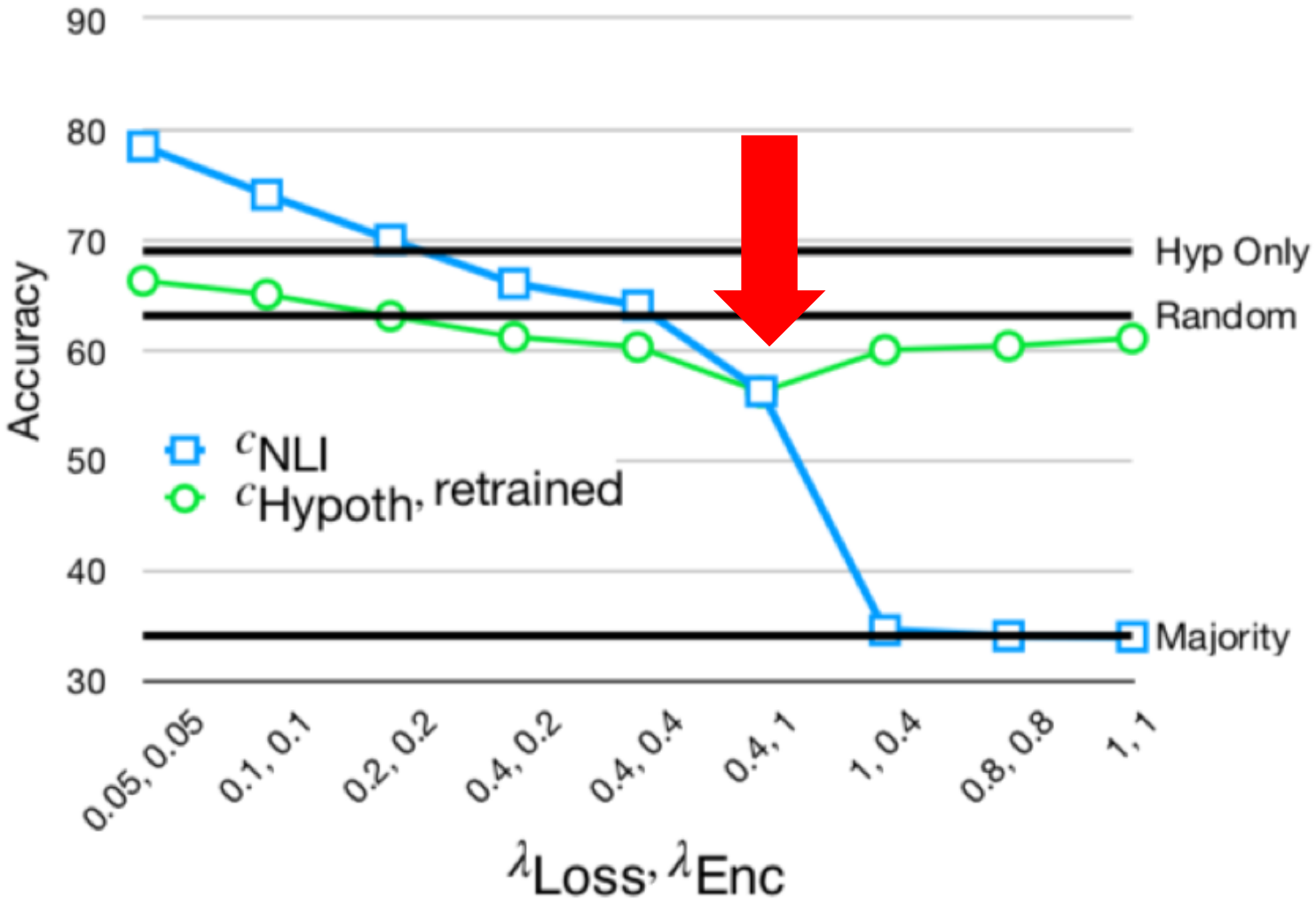
Hidden biases - Adversarial Classifier



Hidden biases - Adversarial Data



Hidden biases - Adversarial Data



What happens to specific biases?

Indicator Words

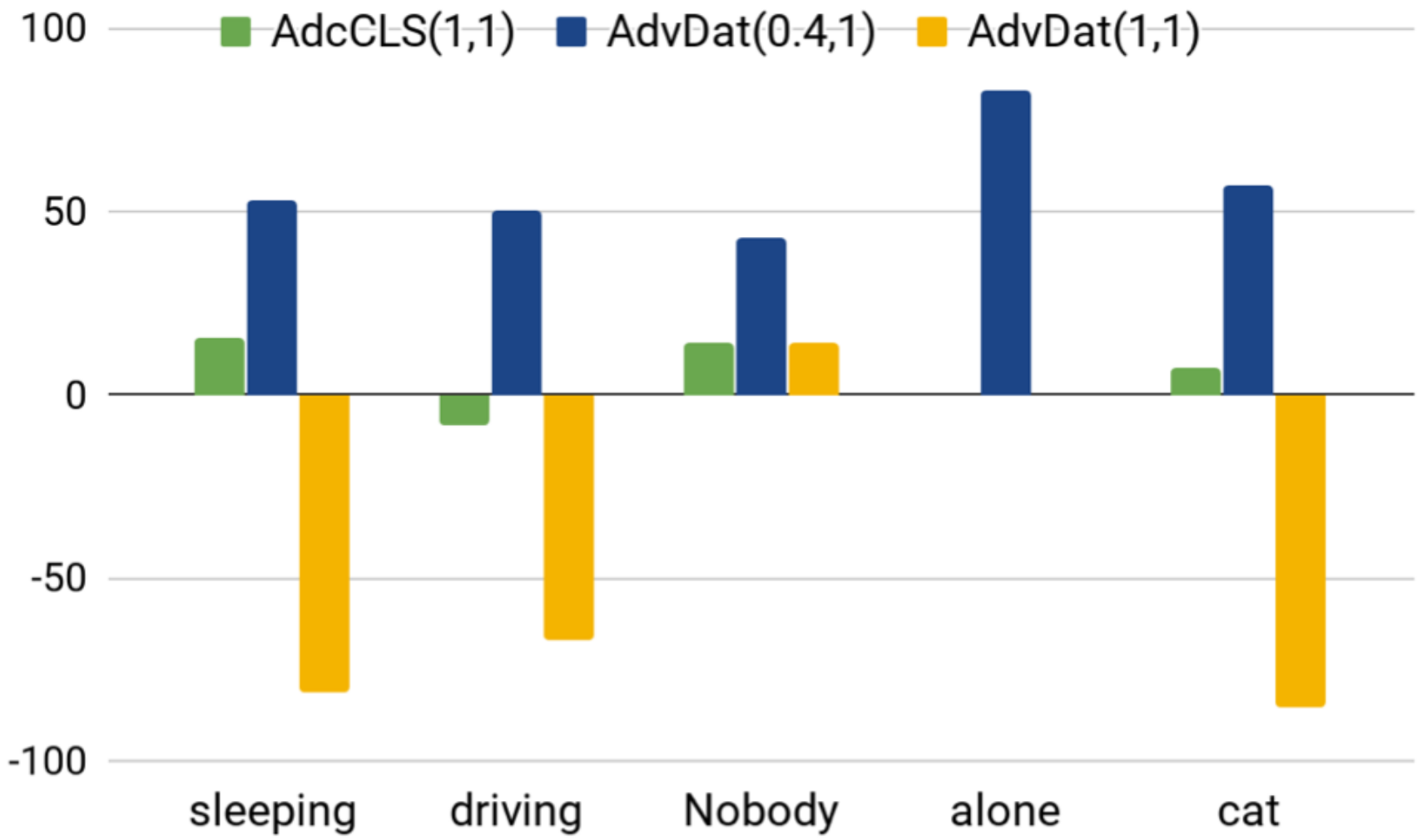
Contradiction
nobody
sleeping
no
tv
cat

Word	Score	Freq
sleeping	0.88	108
driving	0.81	53
Nobody	1.00	52
alone	0.90	50
cat	0.84	49
asleep	0.91	43
no	0.84	31
empty	0.93	28
eats	0.83	24
sleeps	0.95	20

Gururangan et al (*NAACL 2018)

Poliak et al (*SEM 2018)

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

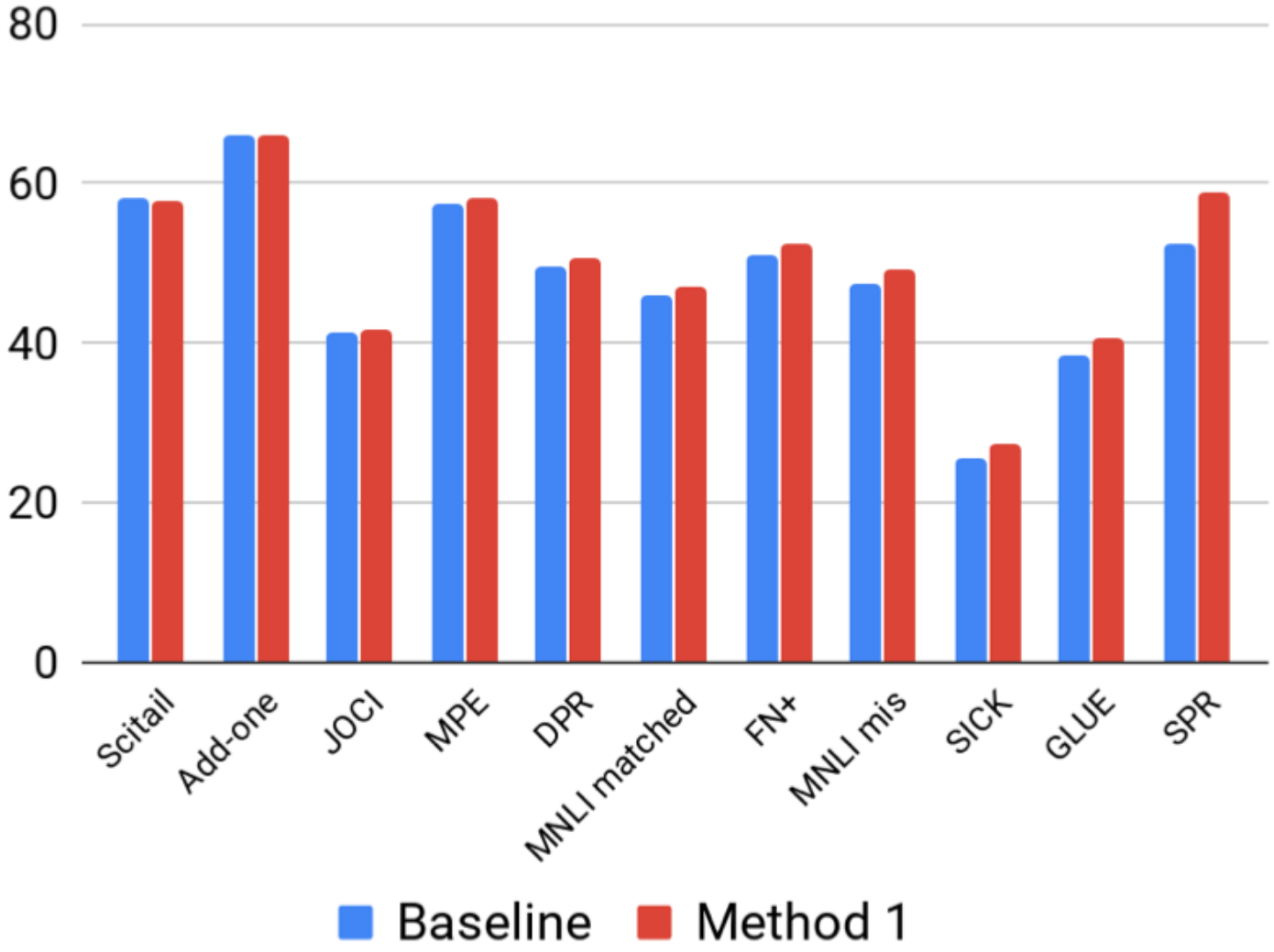
Yonatan Belinkov[13*] **Adam Poliak**[2*]
Stuart M. Shieber[1] **Benjamin Van Durme**[2] **Alexander Rush**[1]
[1]Harvard University [2]Johns Hopkins University [3]Massachusetts Institute of Technology
{belinkov,shieber,srush}@seas.harvard.edu
{azpoliak,vandurme}@cs.jhu.edu

Abstract

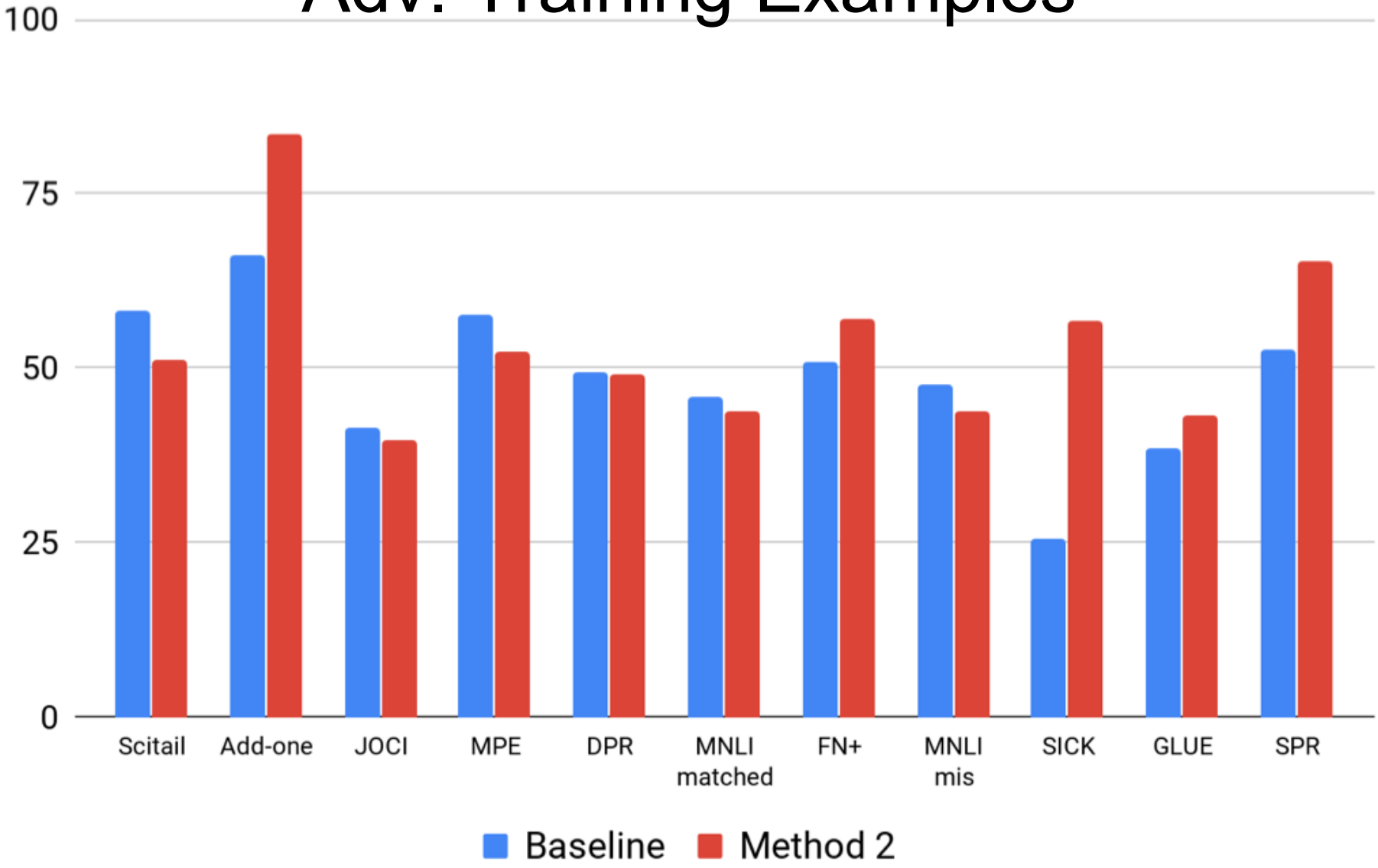
Natural Language Inference (NLI) datasets often contain hypothesis-only biases—artifacts that allow models to achieve non-trivial performance without learning whether a premise entails a hypothesis. We propose two probabilistic methods to build models that are

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).[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



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[1] and **Yonatan Belinkov**[1,2]
[1]Harvard John A. Paulson School of Engineering and Applied Sci
[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-

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

HARVARD
Mind Brain Behavior

human language technology
center of excellence