



Harvard John A. Paulson
School of Engineering
and Applied Sciences



Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference

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

July 29, 2019
ACL, Florence

NLU as Relationship Identification

Natural language inference (entailment)

Premise: A woman is running in the park with her dog

Hypothesis: A woman is sleeping

Relation: entailment, neutral, contradiction

NLU as Relationship Identification

Natural language inference (entailment)

Premise: A woman is running in the park with her dog

Hypothesis: A woman is sleeping

Relation: entailment, neutral, contradiction

NLU as Relationship Identification

Natural language inference (entailment)

Premise: A woman is running in the park with her dog

Hypothesis: A woman is sleeping

Relation: entailment, neutral, contradiction

Reading comprehension

“No,” he replied, “except that he seems in a great hurry.” “That’s just it,” Jimmy returned promptly. “Did you ever see him hurry unless he was frightened?” Peter confessed that he never had.

Q: “Well, he isn’t ____ now, yet just look at him go”

A: Do, case, confessed, frightened, mean, replied, returned, said, see, thought

NLU as Relationship Identification

Natural language inference (entailment)

Premise: A woman is running in the park with her dog

Hypothesis: A woman is sleeping

Relation: entailment, neutral, contradiction

Reading comprehension

"No," he replied, "except that he seems in a great hurry." "That's just it," Jimmy returned promptly. "Did you ever see him hurry unless he was frightened?" Peter confessed that he never had.

Q: "Well, he isn't ____ now, yet just look at him go"

A: Do, case, confessed, frightened, mean, replied, returned, said, see, thought

[Sources: Hill+ '16, Zhang+ '16]

Visual question answering



Q: Is the girl walking the bike?

A: Yes. No

NLU as Relationship Identification

Natural language inference (entailment)

Premise: A woman is running in the park with her dog

Hypothesis: A woman is sleeping

Relation: entailment, neutral, contradiction

Assumption: Identifying the relationship requires **deep** language understanding

a great hurry.” “That’s just it,” Jimmy returned promptly. “Did you ever see him hurry unless he was frightened?” Peter confessed that he never had.

Q: “Well, he isn’t ____ now, yet just look at him go”

A: Do, case, confessed, frightened, mean, replied, returned, said, see, thought

[Sources: Hill+ ‘16, Zhang+ ‘16]



Q: Is the girl walking the bike?

A: Yes, No

One-Sided Biases

- Hypothesis-only NLI (Poliak+ '18; Gururangan+ '18; Tsuchia '18)

One-Sided Biases

- Hypothesis-only NLI (Poliak+ '18; Gururangan+ '18; Tsuchia '18)

Hypothesis: *A woman is sleeping*

One-Sided Biases

- Hypothesis-only NLI (Poliak+ '18; Gururangan+ '18; Tsuchia '18)

Premise:



Hypothesis: *A woman is sleeping*

One-Sided Biases

- Hypothesis-only NLI (Poliak+ '18; Gururangan+ '18; Tsuchia '18)

Premise:



Hypothesis: *A woman is sleeping*

entailment neutral contradiction

One-Sided Biases

- Hypothesis-only NLI (Poliak+ '18; Gururangan+ '18; Tsuchia '18)

Premise:



Hypothesis: *A woman is sleeping*

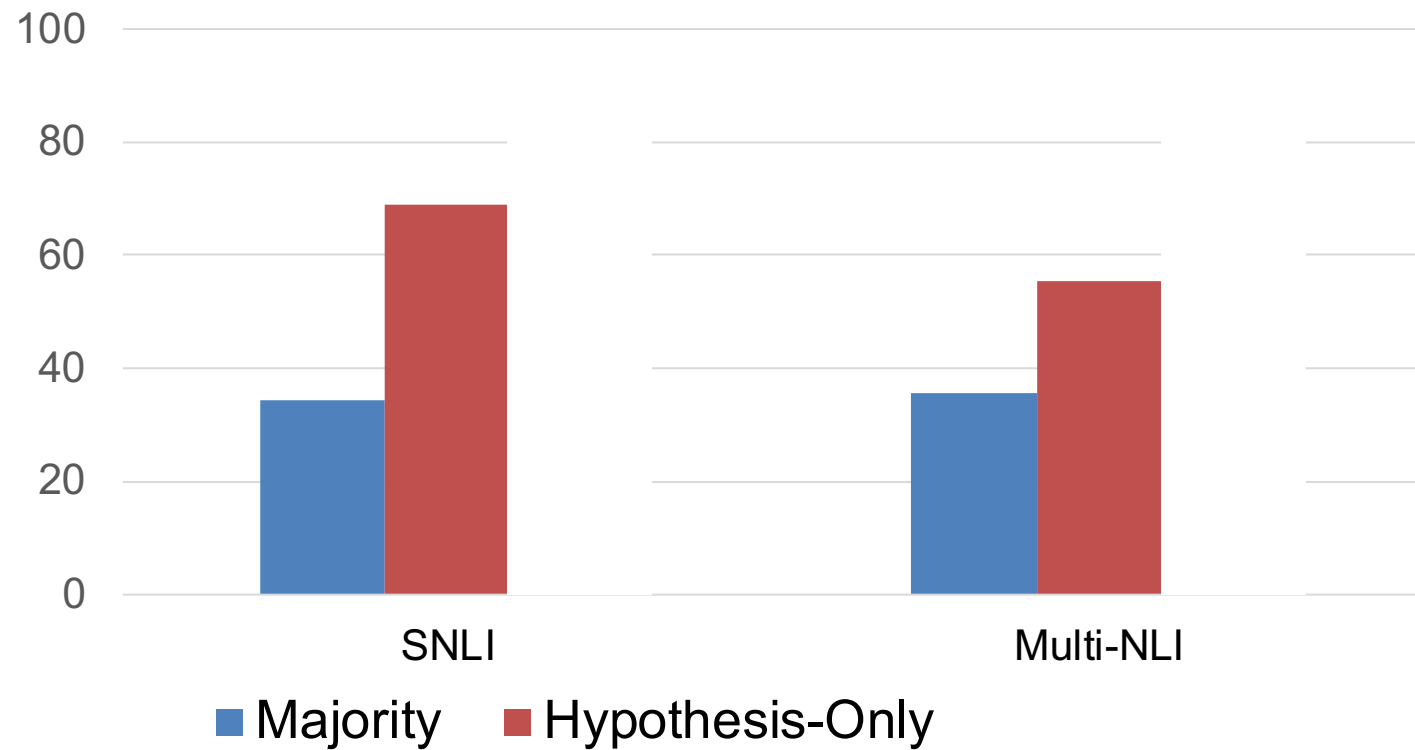
entailment

neutral

contradiction

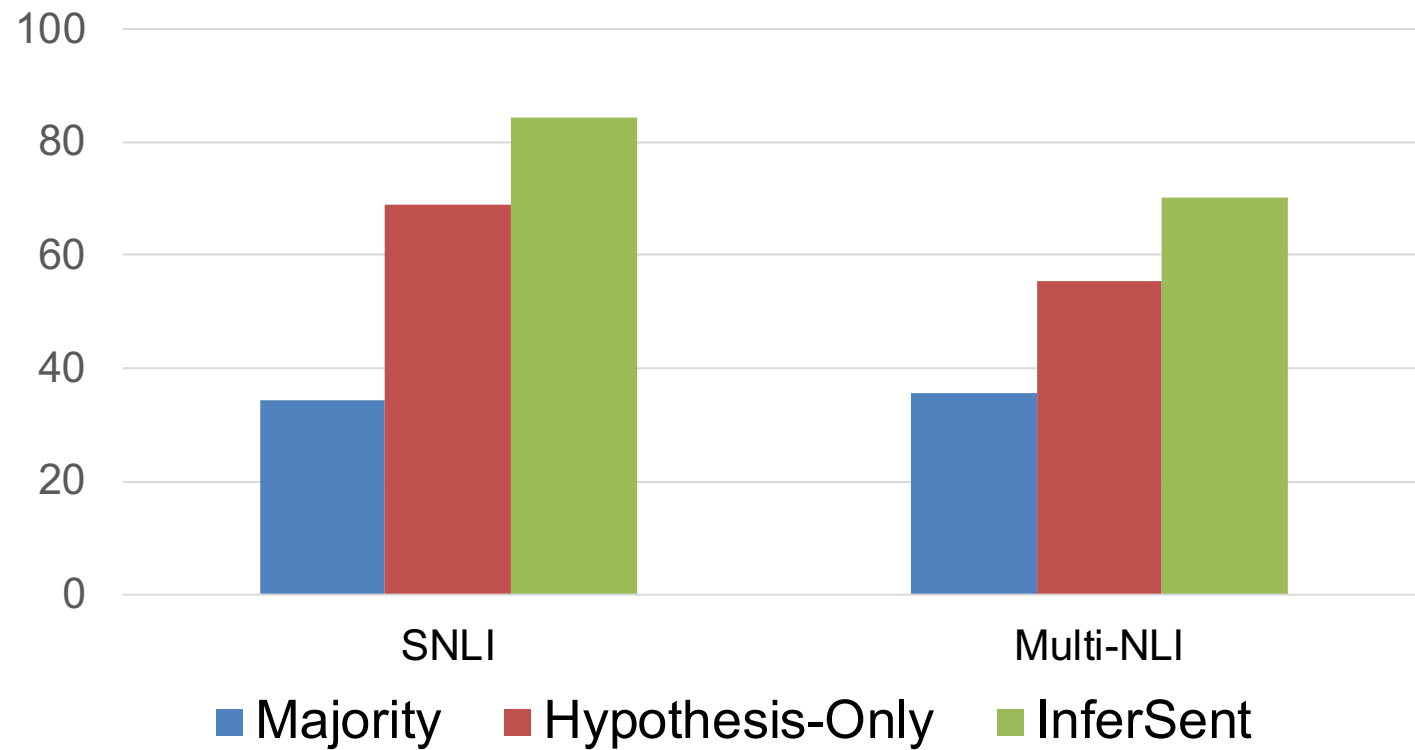
One-Sided Biases

- Hypothesis-only NLI (Poliak+ '18; Gururangan+ '18; Tsuchia '18)



One-Sided Biases

- Hypothesis-only NLI (Poliak+ '18; Gururangan+ '18; Tsuchia '18)



One-Sided Biases

- Hypothesis-only NLI (Poliak+ '18; Gururangan+ '18; Tsuchia '18)
- Reading comprehension (Kaushik & Lipton '18)
- Visual question answering (Zhang+ '16; Kafle & Kanan '16; Goyal+ '17; Agarwal+ '17; *inter alia*)
- Story cloze completion (Schwartz+ '17, Cai+ '17)

Problem:

One-sided biases mean that models may not learn the true relationship between **premise** and **hypothesis**

Strategies for dealing with dataset bias

- Construct new datasets (Sharma+ '18)
 - \$\$\$
 - Other bias

Strategies for dealing with dataset bias

- **Construct new datasets** (Sharma+ '18)
 - \$\$\$
 - Other bias
- **Filter “easy” examples** (Gururangan+ '18)
 - Hard to scale
 - May still have biases (see SWAG → BERT → HellaSWAG)

Strategies for dealing with dataset bias

- **Construct new datasets** (Sharma+ '18)
 - \$\$\$
 - Other bias
- **Filter “easy” examples** (Gururangan+ '18)
 - 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 approach:
Design models
that facilitate learning
less biased representations

A Generative Perspective

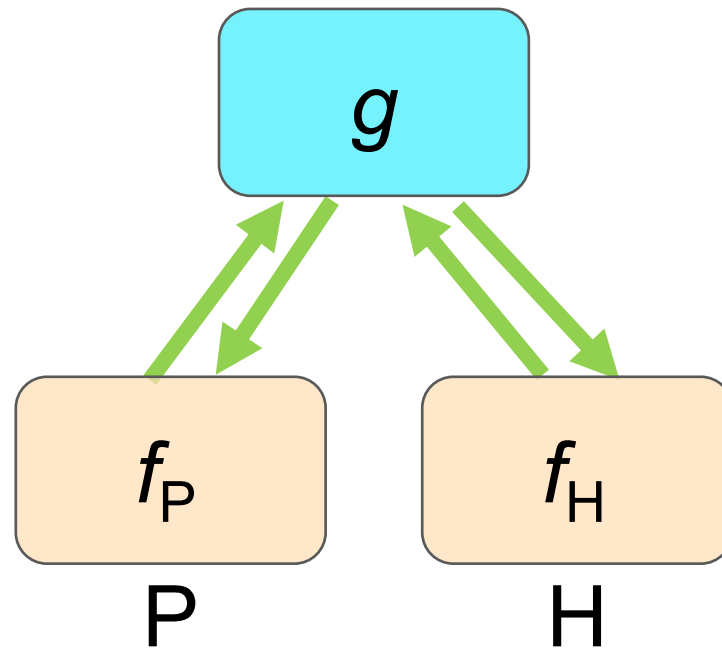
- Typical NLI models maximize the discriminative likelihood

$$p_{\theta}(y|P, H)$$

A Generative Perspective

- Typical NLI models maximize the discriminative likelihood

$$p_{\theta}(y|P, H)$$



g – classifier

f_P, f_H – encoders

A Generative Perspective

- Typical NLI models maximize the discriminative likelihood

$$p_{\theta}(y|P, H)$$

- Our key idea: If we **generate** the premise, it cannot be ignored
- We will maximize the likelihood of generating the premise

$$p(P|y, H)$$

A Generative Perspective

- Typical NLI models maximize the discriminative likelihood

$$p_{\theta}(y|P, H)$$

- Our key idea: If we **generate** the premise, it cannot be ignored
- We will maximize the likelihood of generating the premise

$$p(P|y, H)$$

Hypothesis: A woman is sleeping

Relation: contradiction



Premise: A woman is running in
the park with her dog

A Generative Perspective

- Unfortunately, text generation is hard!

Hypothesis: A woman is sleeping

Relation: contradiction



Premise: A woman is running in
the park with her dog

A Generative Perspective

- Unfortunately, text generation is hard!

Hypothesis: A woman is sleeping

Relation: contradiction



Premise: A woman is running in the park with her dog

Premise: A woman sings a song while playing piano

Premise: This woman is laughing at her baby

⋮

A Generative Perspective

- Unfortunately, text generation is hard!

A Generative Perspective

- Unfortunately, text generation is hard!
- Instead, rewrite as follows

$$\log p(P|y, H) = \log \frac{p_{\theta}(y|P, H)p(P|H)}{p(y|H)}$$

A Generative Perspective

- Unfortunately, text generation is hard!
- Instead, rewrite as follows

$$\log p(P|y, H) = \log \frac{p_{\theta}(y|P, H)p(P|H)}{p(y|H)}$$

- Assume $p(P|H)$ is constant

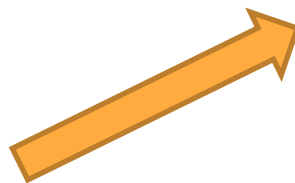
A Generative Perspective

- Unfortunately, text generation is hard!
- Instead, rewrite as follows

$$\log p(P|y, H) = \log \frac{p_{\theta}(y|P, H)p(P|H)}{p(y|H)}$$

- Assume $p(P|H)$ is constant
- We have $\log p_{\theta}(y|P, H) - \log p(y|H)$

Need to estimate this



Method 1: Auxiliary Hypothesis Classifier

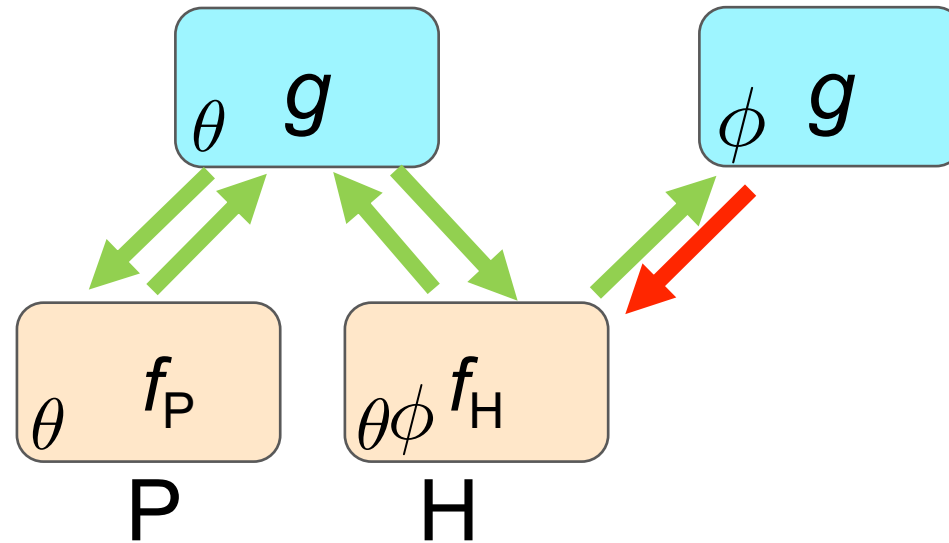
- Learn a new estimator $p_{\phi, \theta}(y|H)$
 - Share the hypothesis-encoder
 - Learn an additional classification layer
 - Multi-task objective function

$$\max_{\theta} L_1(\theta) = \log p_{\theta}(y|P, H) - \alpha \log p_{\phi, \theta}(y|H)$$

$$\max_{\phi} L_2(\phi) = \beta \log p_{\phi, \theta}(y|H)$$

Method 1: Auxiliary Hypothesis Classifier

- Learn a new estimator $p_{\phi, \theta}(y|H)$

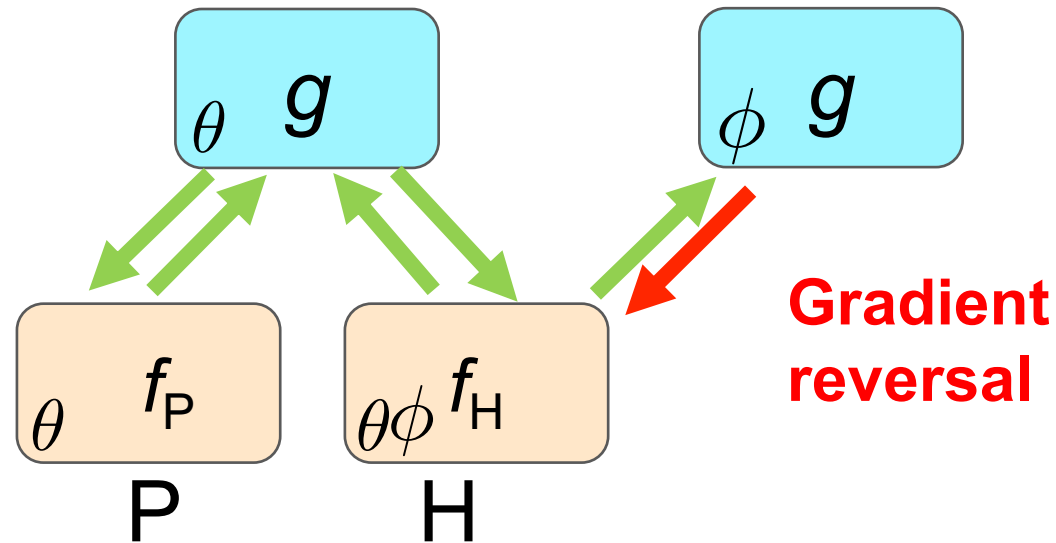


$$\max_{\theta} L_1(\theta) = \log p_{\theta}(y|P, H) - \alpha \log p_{\phi, \theta}(y|H)$$

$$\max_{\phi} L_2(\phi) = \beta \log p_{\phi, \theta}(y|H)$$

Method 1: Auxiliary Hypothesis Classifier

- Learn a new estimator $p_{\phi, \theta}(y|H)$



$$\max_{\theta} L_1(\theta) = \log p_{\theta}(y|P, H) - \alpha \log p_{\phi, \theta}(y|H)$$

$$\max_{\phi} L_2(\phi) = \beta \log p_{\phi, \theta}(y|H)$$

Method 2: Negative Sampling

- Sample alternative premises

Method 2: Negative Sampling

- Sample alternative premises

$$\begin{aligned} -\log p(y \mid H) &= -\log \sum_{P'} p(P' \mid H) p(y \mid P', H) \\ &= -\log \mathbb{E}_{P'} p(y \mid P', H) \geq -\mathbb{E}_{P'} \log p(y \mid P', H), \end{aligned}$$

- Lower bound from Jensen's inequality
- Approximate the expectation with uniform samples P'

Method 2: Negative Sampling

- Sample alternative premises

$$\begin{aligned} -\log p(y \mid H) &= -\log \sum_{P'} p(P' \mid H) p(y \mid P', H) \\ &= -\log \mathbb{E}_{P'} p(y \mid P', H) \geq -\mathbb{E}_{P'} \log p(y \mid P', H), \end{aligned}$$

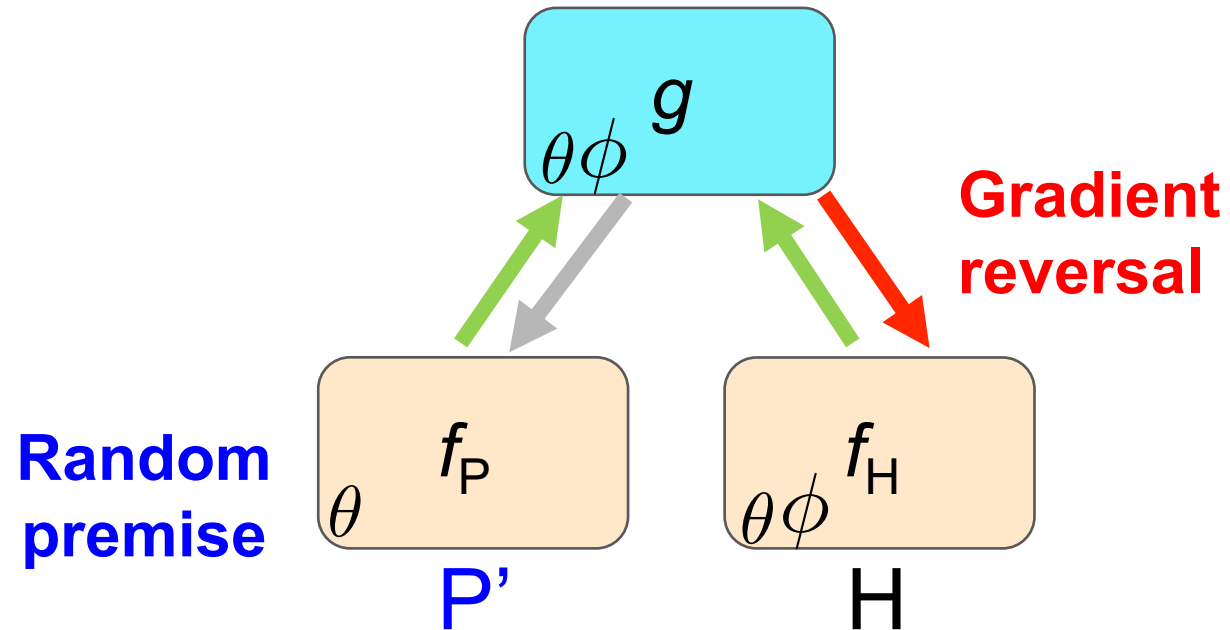
- Lower bound from Jensen's inequality
- Approximate the expectation with uniform samples P'
- Multi-task objective function

$$\max_{\theta} L_1(\theta) = (1 - \alpha) \log p_{\theta}(y|P, H) - \alpha \log p_{\phi, \theta}(y|P', H)$$

$$\max_{\phi} L_2(\phi) = \beta \log p_{\phi, \theta}(y|P', H)$$

Method 2: Negative Sampling

- Sample alternative premises



$$\max_{\theta} L_1(\theta) = (1 - \alpha) \log p_{\theta}(y|P, H) - \alpha \log p_{\phi, \theta}(y|P', H)$$

$$\max_{\phi} L_2(\phi) = \beta \log p_{\phi, \theta}(y|P', H)$$

What is this good for?

What is this good for?

Are less biased models
more transferable?

A Toy Example

Synthetic dataset (unbiased)

$(a, a) \rightarrow \text{TRUE}$

$(a, b) \rightarrow \text{FALSE}$

$(b, b) \rightarrow \text{TRUE}$

$(b, a) \rightarrow \text{FALSE}$

A Toy Example

Synthetic dataset (unbiased)

$$\begin{array}{ll} (a, a) \rightarrow \text{TRUE} & (a, b) \rightarrow \text{FALSE} \\ (b, b) \rightarrow \text{TRUE} & (b, a) \rightarrow \text{FALSE} \end{array}$$

Synthetic dataset (biased)

$$\begin{array}{ll} (a, ac) \rightarrow \text{TRUE} & (a, b) \rightarrow \text{FALSE} \\ (b, bc) \rightarrow \text{TRUE} & (b, a) \rightarrow \text{FALSE} \end{array}$$

A Toy Example

Synthetic dataset (unbiased)

$$\begin{array}{ll} (a, a) \rightarrow \text{TRUE} & (a, b) \rightarrow \text{FALSE} \\ (b, b) \rightarrow \text{TRUE} & (b, a) \rightarrow \text{FALSE} \end{array}$$

Synthetic dataset (biased)

$$\begin{array}{ll} (a, a) \rightarrow \text{TRUE} & (a, b) \rightarrow \text{FALSE} \\ (b, b) \rightarrow \text{TRUE} & (b, a) \rightarrow \text{FALSE} \end{array}$$

Models transfer well on synthetic data

β	α					
	0.1	0.25	0.5	1	2.5	5
0.1	50	50	50	50	50	50
0.5	50	50	50	50	50	50
1	50	50	50	50	50	50
1.5	50	50	50	50	50	100
2	50	50	50	50	100	100
2.5	50	50	100	75	100	100
3	50	100	100	100	100	100
3.5	100	100	100	100	100	100
4	100	100	100	100	100	100
5	100	100	100	100	100	100
10	100	100	100	100	100	100
20	100	100	100	100	100	100

Method 1: Auxiliary Hypothesis Classifier

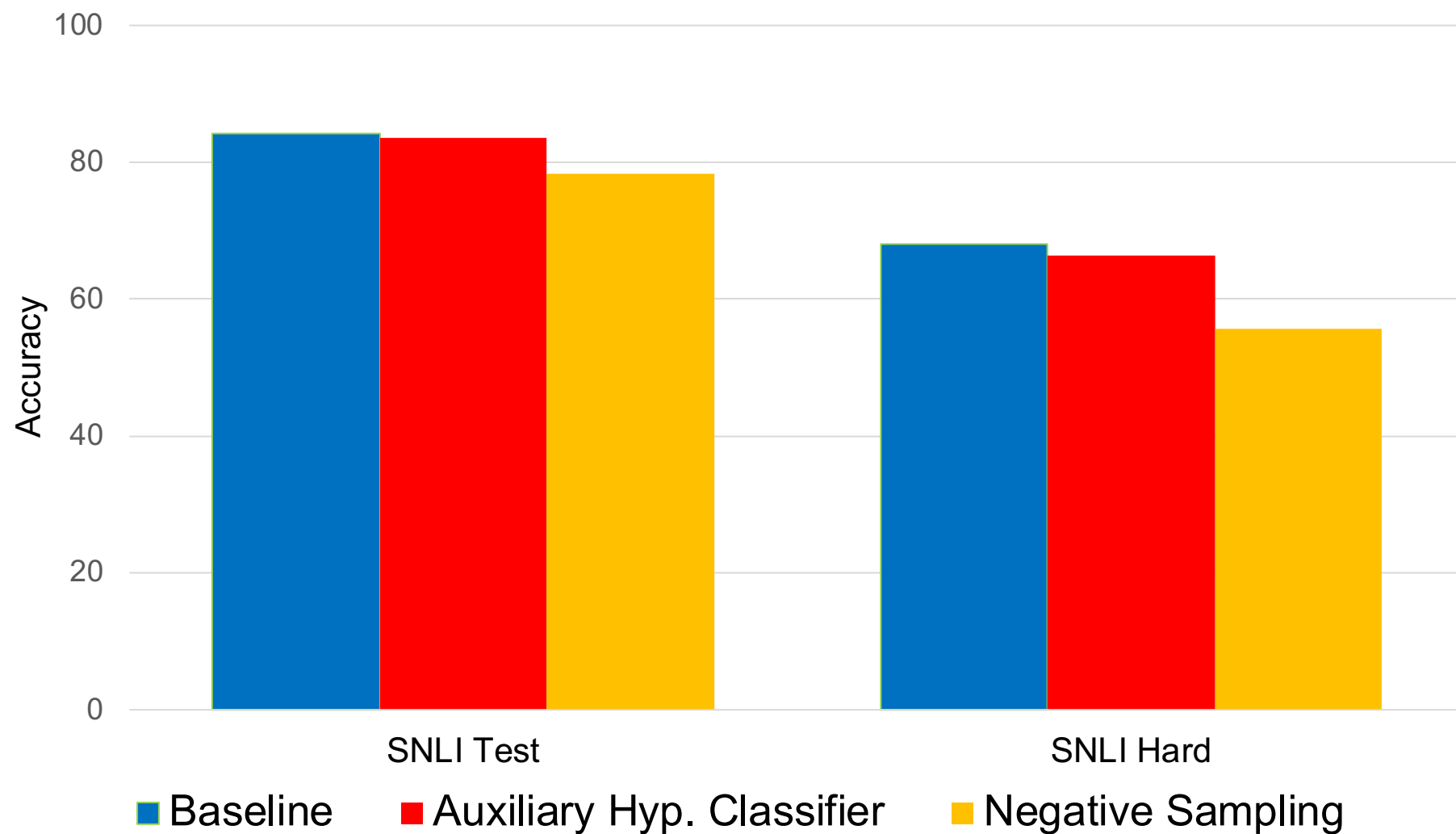
Models transfer well on synthetic data

β	α				
	0.1	0.25	0.5	0.75	1
0.1	50	50	50	50	50
0.5	50	50	50	50	50
1	50	50	50	50	50
1.5	50	50	50	50	50
2	50	50	50	50	50
2.5	50	50	50	50	50
3	50	50	100	50	50
3.5	50	50	100	50	50
4	50	100	100	50	50
5	50	50	100	100	50*
10	75	100	100	100	50*
20	100	100	100	50*	50*

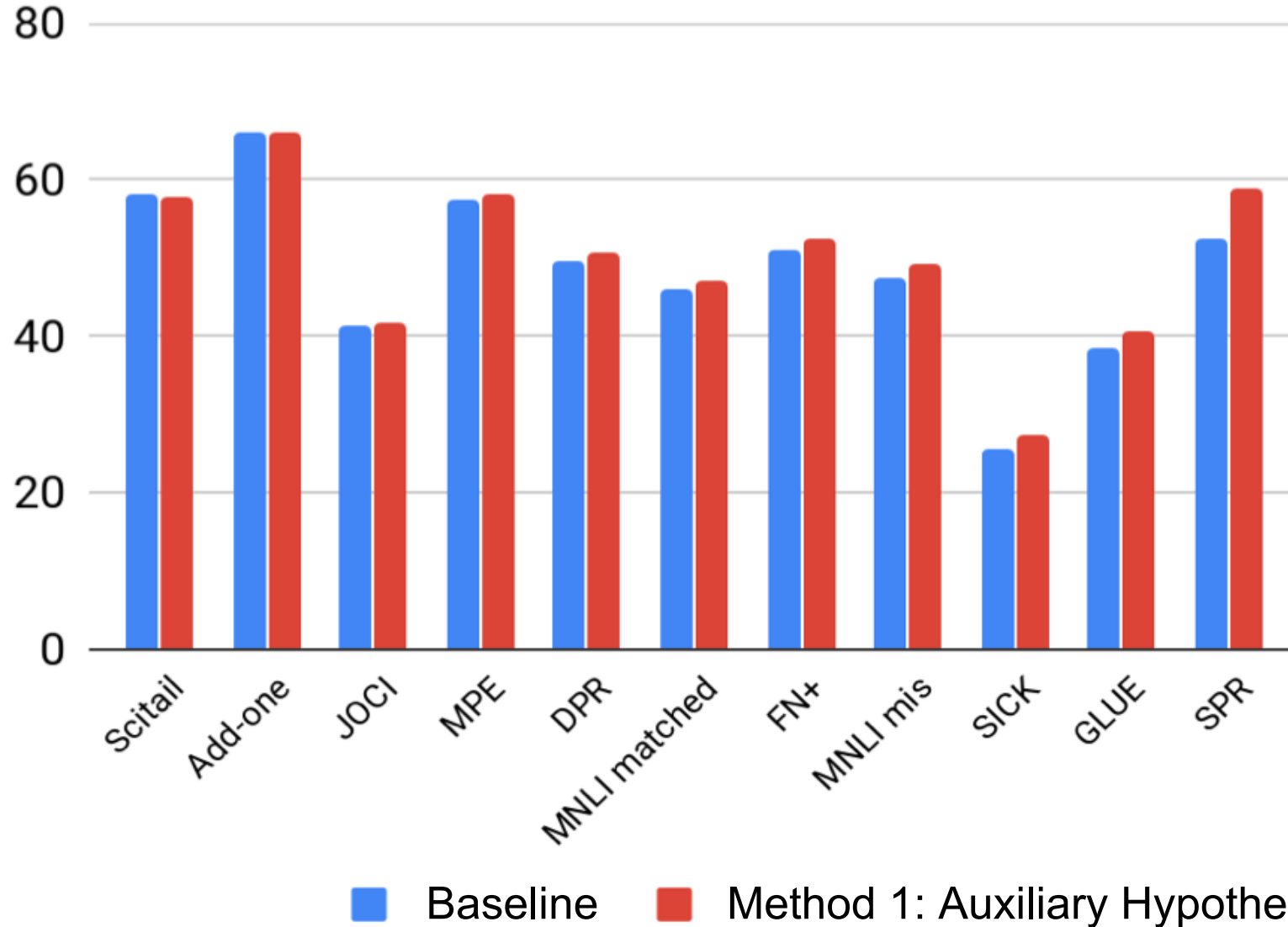
Method 2: Negative Sampling

Do the models transfer well
on standard NLI datasets?

Degradation in domain

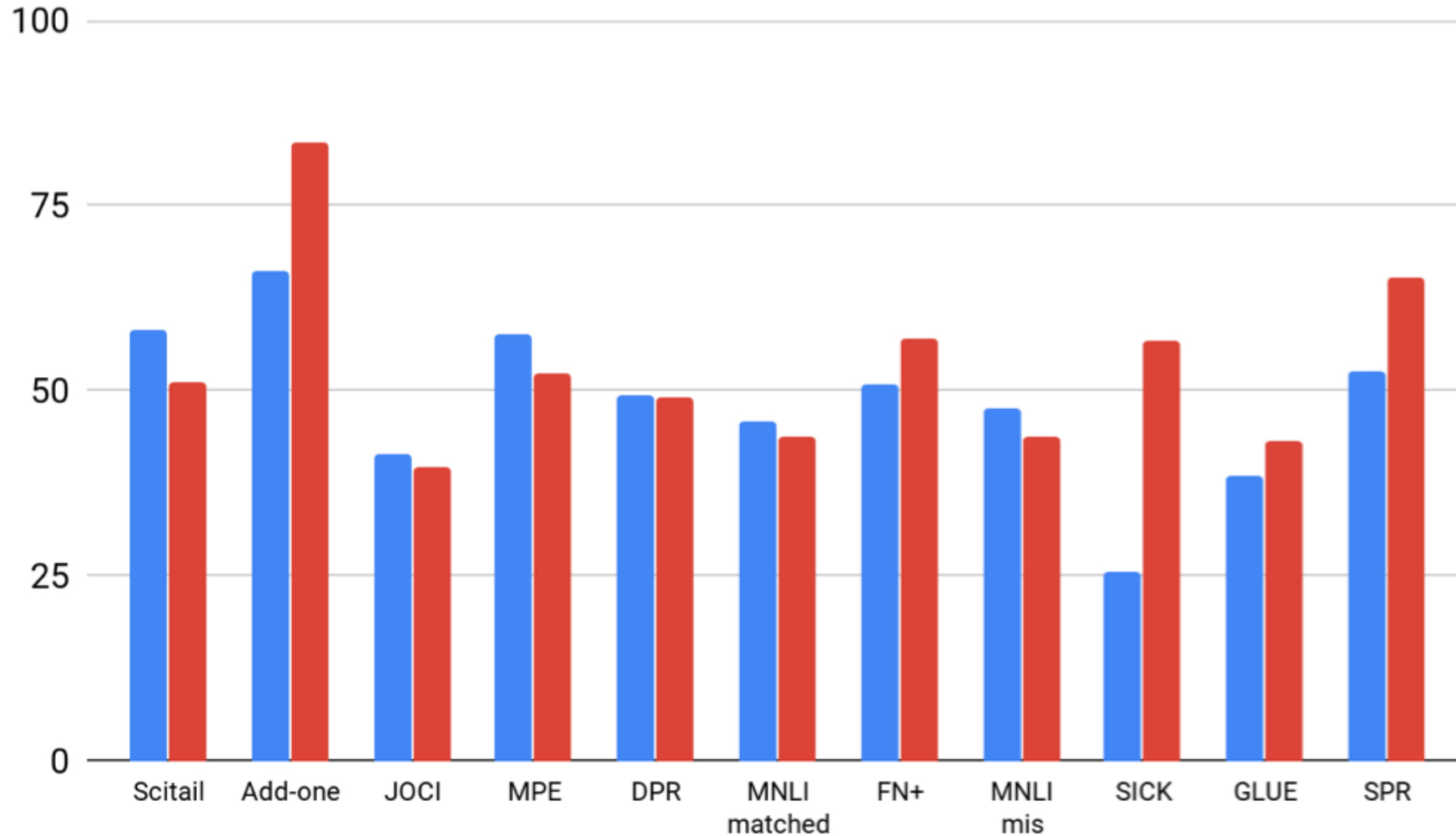


Transfer to other datasets



Improvements in
9/11 datasets

Transfer to other datasets



■ Baseline ■ Method 2: Negative Sampling



Less consistent improvements



When it works, it works well

Analysis

Analysis

Q: Does it matter what kind of bias we have?

A: Yes! Different biases than training data →

- Usually, more improvement from our methods
- But not always

Analysis

Q: Does it matter what kind of bias we have?

A: Yes! Different biases than training data →

- Usually, more improvement from our methods
- But not always

Q: Do stronger hyper-parameters help?

A: More emphasis on the auxiliary objective →

- More transferability, but worse in-domain performance

Analysis

Q: Does it matter what kind of bias we have?

A: Yes! Different biases than training data →

- Usually, more improvement from our methods
- But not always

Q: Do stronger hyper-parameters help?

A: More emphasis on the auxiliary objective →

- More transferability, but worse in-domain performance

Q: What if we get a bit of out-of-domain training data?

A: Pre-training with our methods still helps

- Especially with datasets with different biases

More Analysis

Q: Are biases really removed from the hidden representations?

A: Some, but not all

- See our recent work: *On Adversarial Removal of Hypothesis-only Bias in NLI*,
*SEM 2019

More Analysis

Q: Are biases really removed from the hidden representations?

A: Some, but not all

- See our recent work: *On Adversarial Removal of Hypothesis-only Bias in NLI*, *SEM 2019

Q: Does this approach work for other tasks?

A: Seems to work for VQA (Ramakrishnan+ '18)

A: But there are shortcomings

- See our recent work: *Adversarial Regularization for VQA: Strengths, Shortcomings, and Side Effects*, SiVL 2019

Contributions

- Our approach may aid with one-sided biases in NLI and other tasks
 - Reduces the amount of bias
 - Improves transferability

Acknowledgements:



HARVARD
Mind Brain Behavior



Contributions

- Our approach may aid with one-sided biases in NLI and other tasks
 - Reduces the amount of bias
 - Improves transferability
- Our analysis shows that the methods should be handled with care
 - Not all bias may be removed
 - Some other information may also be removed
 - The goal matters: bias may sometimes be helpful

Acknowledgements:



HARVARD
Mind Brain Behavior

