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 M. 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 more robust to such biases and better transfer across datasets. In contrast to standard approaches to NLI, our methods predict the probability of a premise given a hypothesis and NLI label, discouraging models from ignoring the premise. We evaluate our methods on synthetic and existing NLI datasets by training on datasets containing biases and testing on datasets containing no (or different) hypothesis-only biases. Our results indicate that these methods can make NLI models more robust to dataset-specific artifacts, transferring better than a baseline architecture in 9 out of 12 NLI datasets. Additionally, we provide an extensive analysis of the interplay of our methods with known biases in NLI datasets, as well as the effects of encouraging models to ignore biases and fine-tuning on target datasets.[1]

1 Introduction

Natural Language Inference (NLI) is often used to gauge a model's ability to understand a relationship between two texts (Cooper et al., 1996; Dagan et al., 2006). In NLI, a model is tasked with determining whether a hypothesis (*a woman is sleeping*) would likely be inferred from a premise (*a woman is talking on the phone*).[2] The development of new large-scale datasets has led to a flurry of various neural network architectures for solving NLI. However, recent work has found that many NLI datasets contain biases, or annotation artifacts, i.e., features present in hypotheses 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., 2018b; Tsuchiya, 2018).[3] For instance, in some datasets, negation words like "not" and "nobody" are often associated with a relationship of contradiction. As a ramification of such biases, models may not generalize well to other datasets that contain different or no such biases.

Recent studies have tried to create new NLI datasets that do not contain such artifacts, but many approaches to dealing with this issue remain unsatisfactory: constructing new datasets (Sharma et al., 2018) is costly and may still result in other artifacts; filtering "easy" examples and defining a harder subset is useful for evaluation purposes (Gururangan et al., 2018), but difficult to do on a large scale that enables training; and compiling adversarial examples (Glockner et al., 2018) is informative but again limited by scale or diversity. Instead, our goal is to develop methods that overcome these biases as datasets may still contain undesired artifacts despite annotation efforts.

Typical NLI models learn to predict an entailment label discriminatively given a premise-hypothesis pair (Figure 1a), enabling them to learn hypothesis-only biases. Instead, we predict the premise given the hypothesis and the entailment label, which by design cannot be solved using data artifacts. While this objective is intractable, it motivates two approximate training methods for standard NLI classifiers that are more resistant to biases. Our first method uses a hypothesis-only classifier (Figure 1b) and the second uses negative sampling by swapping premises between premise-hypothesis pairs (Figure 1c).

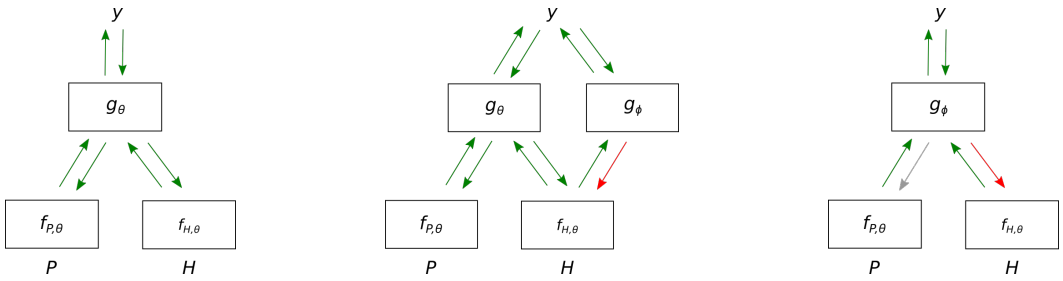
* Equal contribution

[1]Our code is available at https://github.com/azpoliak/robust-nli.

[2]This hypothesis contradicts the premise and would likely not be inferred.

[3]We use *artifacts* and *biases* interchangeably.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 877–891
Florence, Italy, July 28 - August 2, 2019.



(a) Baseline (b) Method 1 (c) Method 2

Figure 1: Illustration of (a) the baseline NLI architecture, and our two proposed methods to remove hypothesis only-biases from an NLI model: (b) uses a hypothesis-only classifier, and (c) samples a random premise. Arrows correspond to the direction of propagation. Green or red arrows respectively mean that the gradient sign is kept as is or reversed. Gray arrow indicates that the gradient is not back-propagated - this only occurs in (c) when we randomly sample a premise, otherwise the gradient is back-propagated. f and g represent encoders and classifiers.

We evaluate the ability of our methods to generalize better in synthetic and naturalistic settings. First, using a controlled, synthetic dataset, we demonstrate that, unlike the baseline, our methods enable a model to ignore the artifacts and learn to correctly identify the desired relationship between the two texts. Second, we train models on an NLI dataset that is known to be biased and evaluate on other datasets that may have different or no biases. We observe improved results compared to a fully discriminative baseline in 9 out of 12 target datasets, indicating that our methods generate models that are more robust to annotation artifacts.

An extensive analysis reveals that our methods are most effective when the target datasets have different biases from the source dataset or no noticeable biases. We also observe that the more we encourage the model to ignore biases, the better it transfers, but this comes at the expense of performance on the source dataset. Finally, we show that our methods can better exploit small amounts of training data in a target dataset, especially when it has different biases from the source data.

In this paper, we focus on the transferability of our methods from biased datasets to ones having different or no biases. Elsewhere (Belinkov et al., 2019), we have analyzed the effect of these methods on the learned language representations, suggesting that they may indeed be less biased. However, we caution that complete removal of biases remains difficult and is dependent on the techniques used. The choice of whether to remove bias also depends on the goal; in an in-domain scenario certain biases may be helpful and should not necessarily be removed.

In summary, in this paper we make the following contributions:

- Two novel methods to train NLI models that are more robust to dataset-specific artifacts.
- An empirical evaluation of the methods on a synthetic dataset and 12 naturalistic datasets.
- An extensive analysis of the effects of our methods on handling bias.

2 Motivation

A training instance for NLI consists of a hypothesis sentence H, a premise statement P, and an inference label y. A probabilistic NLI model aims to learn a parameterized distribution $p_\theta(y \mid P, H)$ to compute the probability of the label given the two sentences. We consider NLI models with premise and hypothesis encoders, $f_{P,\theta}$ and $f_{H,\theta}$, which learn representations of P and H, and a classification layer, g_θ, which learns a distribution over y. Typically, this is done by maximizing this discriminative likelihood directly, which will act as our baseline (Figure 1a).

However, many NLI datasets contain biases that allow models to perform non-trivially well when accessing just the hypotheses (Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018b). This allows models to leverage hypothesis-only biases that may be present in a dataset. A model may perform well on a specific dataset, without identifying whether P entails H. Gururangan et al. (2018) argue that "the bulk" of many models' "success [is] attribute[d] to the easy examples". Consequently, this may limit how well a model trained on one dataset would perform on other datasets that may have different artifacts.

Consider an example where P and H are strings from $\{a, b, c\}$, and an environment where P en-

tails H if and only if the first letters are the same, as in synthetic dataset A. In such a setting, a model should be able to learn the correct condition for P to entail H.[4]

Synthetic dataset A

$(a, a) \rightarrow$ TRUE	$(a, b) \rightarrow$ FALSE
$(b, b) \rightarrow$ TRUE	$(b, a) \rightarrow$ FALSE

Imagine now that an artifact c is appended to every entailed H (synthetic dataset B). A model of y with access only to the hypothesis side can fit the data perfectly by detecting the presence or absence of c in H, ignoring the more general pattern. Therefore, we hypothesize that a model that learns $p_\theta(y \mid P, H)$ by training on such data would be misled by the bias c and would fail to learn the relationship between the premise and the hypothesis. Consequently, the model would not perform well on the unbiased synthetic dataset A.

Synthetic dataset B (with artifact)

$(a, ac) \rightarrow$ TRUE	$(a, b) \rightarrow$ FALSE
$(b, bc) \rightarrow$ TRUE	$(b, a) \rightarrow$ FALSE

Instead of maximizing the discriminative likelihood $p_\theta(y \mid P, H)$ directly, we consider maximizing the likelihood of generating the premise P conditioned on the hypothesis H and the label y: $p(P \mid H, y)$. This objective cannot be fooled by hypothesis-only features, and it requires taking the premise into account. For example, a model that only looks for c in the above example cannot do better than chance on this objective. However, as P comes from the space of all sentences, this objective is much more difficult to estimate.

3 Training Methods

Our goal is to maximize $\log p(P \mid H, y)$ on the training data. While we could in theory directly parameterize this distribution, for efficiency and simplicity we instead write it in terms of the standard $p_\theta(y \mid P, H)$ and introduce a new term to approximate the normalization:

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

Throughout we will assume $p(P \mid H)$ is a fixed constant (justified by the dataset assumption that, lacking y, P and H are independent and drawn at random). Therefore, to approximately maximize this objective we need to estimate $p(y \mid H)$. We propose two methods for doing so.

[4] This is equivalent to XOR and is learnable by a MLP.

3.1 Method 1: Hypothesis-only Classifier

Our first approach is to estimate the term $p(y \mid H)$ directly. In theory, if labels in an NLI dataset depend on both premises and hypothesis (which Poliak et al. (2018b) call "*interesting* NLI"), this should be a uniform distribution. However, as discussed above, it is often possible to correctly predict y based only on the hypothesis. Intuitively, this model can be interpreted as training a classifier to identify the (latent) artifacts in the data.

We define this distribution using a shared representation between our new estimator $p_{\phi,\theta}(y \mid H)$ and $p_\theta(y \mid P, H)$. In particular, the two share an embedding of H from the hypothesis encoder $f_{H,\theta}$. The additional parameters ϕ are in the final layer g_ϕ, which we call the *hypothesis-only classifier*. The parameters of this layer ϕ are updated to fit $p(y \mid H)$ whereas the rest of the parameters in θ are updated based on the gradients of $\log p(P \mid y, H)$.

Training is illustrated in Figure 1b. This interplay is controlled by two hyper-parameters. First, the negative term is scaled by a hyper-parameter α. Second, the updates of g_ϕ are weighted by β. We therefore minimize the following multitask loss functions (shown for a single example):

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

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

We implement these together with a gradient reversal layer (Ganin & Lempitsky, 2015). As illustrated in Figure 1b, during back-propagation, we first pass gradients through the hypothesis-only classifier g_ϕ and then reverse the gradients going to the hypothesis encoder $g_{H,\theta}$ (potentially scaling them by β).[5]

3.2 Method 2: Negative Sampling

As an alternative to the hypothesis-only classifier, our second method attempts to remove annotation artifacts from the representations by sampling alternative premises. Consider instead writing the

[5] This approach may also be seen as adversarial training with respect to the hypothesis, akin to domain-adversarial neural networks (Ganin et al., 2016). However, our methods encourage robustness to latent hypothesis biases, without requiring a domain label.

normalization term above as,

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

where the expectation is uniform and the last step is from Jensen's inequality.[6] As in Method 1, we define a separate $p_{\phi,\theta}(y \mid P', H)$ which shares the embedding layers from θ, $f_{P,\theta}$ and $f_{H,\theta}$. However, as we are attempting to unlearn hypothesis bias, we block the gradients and do not let it update the premise encoder $f_{P,\theta}$.[7] The full setting is shown in Figure 1c.

To approximate the expectation, we use uniform samples P' (from other training examples) to replace the premise in a (P, H)-pair, while keeping the label y. We also maximize $p_{\theta,\phi}(y \mid P', H)$ to learn the artifacts in the hypotheses. We use $\alpha \in [0, 1]$ to control the fraction of randomly sampled P's (so the total number of examples remains the same). As before, we implement this using gradient reversal scaled by β.

$$
\begin{aligned}
\max_{\theta} L_1(\theta) &= (1-\alpha) \log p_\theta(y \mid P, H) \\
&\quad - \alpha \log p_{\theta,\phi}(y \mid P', H) \\
\max_{\phi} L_2(\phi) &= \beta \log p_{\theta,\phi}(y \mid P', H)
\end{aligned}
$$

Finally, we share the classifier weights between $p_\theta(y \mid P, H)$ and $p_{\phi,\theta}(y \mid P', H)$. In a sense this is counter-intuitive, since p_θ is being trained to unlearn bias, while $p_{\phi,\theta}$ is being trained to learn it. However, if the models are trained separately, they may learn to co-adapt with each other (Elazar & Goldberg, 2018). If $p_{\phi,\theta}$ is not trained well, we might be fooled to think that the representation does not contain any biases, while in fact they are still hidden in the representation. For some evidence that this indeed happens when the models are trained separately, see Belinkov et al. (2019).[8]

[6] There are more developed and principled approaches in language modeling for approximating this partition function without having to make this assumption. These include importance sampling (Bengio & Senecal, 2003), noise-contrastive estimation (Gutmann & Hyvärinen, 2010), and sublinear partition estimation (Rastogi & Van Durme, 2015). These are more difficult to apply in the setting of sampling full sentences from an unknown set. We hope to explore methods for applying them in future work.

[7] A reviewer asked about gradient blocking. Our motivation was that, for a random premise, we do not have reliable information to update its encoder. However, future work can explore different configurations of gradient blocking.

[8] A similar situation arises in neural cryptography (Abadi & Andersen, 2016), where an encryptor Alice and a decryptor Bob communicate while an adversary Eve tries to eavesdrop on their communication. Alice and Bob are analogous to the hypothesis embedding and p_θ, while Eve is analogous to $p_{\phi,\theta}$. In their asymmetric encryption experiments, Abadi & Andersen observed seemingly secret communication, which on closer look the adversary was able to eavesdrop on.

4 Experimental Setup

To evaluate how well our methods can overcome hypothesis-only biases, we test our methods on a synthetic dataset as well as on a wide range of existing NLI datasets. The scenario we aim to address is when training on a source dataset with biases and evaluating on a target dataset with different or no biases. We first describe the data and experimental setup before discussing the results.

Synthetic Data We create a synthetic dataset based on the motivating example in Section 2, where P entails H if and only if their first letters are the same. The training and test sets have 1K examples each, uniformly distributed among the possible entailment relations. In the test set (dataset A), each premise or hypothesis is a single symbol: $P, H \in \{a, b\}$, where P entails H iff $P = H$. In the training set (dataset B), a letter c is appended to the hypothesis side in the TRUE examples, but not in the FALSE examples. In order to transfer well to the test set, a model that is trained on this training set needs to learn the underlying relationship—that P entails H if and only if their first letter is identical—rather than relying on the presence of c in the hypothesis side.

Common NLI datasets Moving to existing NLI datasets, we train models on the Stanford Natural Language Inference dataset (SNLI; Bowman et al., 2015), since it is known to contain significant annotation artifacts. We evaluate the robustness of our methods on other, target datasets.

As target datasets, we use the 10 datasets investigated by Poliak et al. (2018b) in their hypothesis-only study, plus two test sets: GLUE's diagnostic test set, which was carefully constructed to not contain hypothesis-biases (Wang et al., 2018), and SNLI-hard, a subset of the SNLI test set that is thought to have fewer biases (Gururangan et al., 2018). The target datasets include *human-judged* datasets that used automatic methods to pair premises and hypotheses, and then relied on humans to label the pairs: SCITAIL (Khot et al., 2018), ADD-ONE-RTE (Pavlick & Callison-Burch, 2016), Johns Hopkins Ordinal Common-

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

(a) Method 1

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*

(b) Method 2

Table 1: Accuracies on the synthetic dataset, when training on the biased training set and evaluating on the unbiased test set. Darker boxes represent higher accuracies. * indicates failure to learn the biased training set; all other configurations learned the training set perfectly.

sense Inference (JOCI; Zhang et al., 2017), Multiple Premise Entailment (MPE; Lai et al., 2017),and Sentences Involving Compositional Knowledge (SICK; Marelli et al., 2014). The target datasets also include datasets recast by White et al. (2017) to evaluate different semantic phenomena: FrameNet+ (FN+; Pavlick et al., 2015), Definite Pronoun Resolution (DPR; Rahman & Ng, 2012), and Semantic Proto-Roles (SPR; Reisinger et al., 2015).[9] As many of these datasets have different label spaces than SNLI, we define a mapping (Appendix A.1) from our models' predictions to each target dataset's labels. Finally, we also test on the Multi-genre NLI dataset (MNLI; Williams et al., 2018), a successor to SNLI.[10]

Baseline & Implementation Details We use `InferSent` (Conneau et al., 2017) as our baseline model because it has been shown to work well on popular NLI datasets and is representative of many NLI models. We use separate BiLSTM encoders to learn vector representations of P and H.[11] The vector representations are combined following Mou et al. (2016),[12] and passed to an MLP classifier with one hidden layer. Our proposed methods for mitigating biases use the same technique for representing and combining sentences. Additional implementation details are provided in Appendix A.2.

For both methods, we sweep hyper-parameters α, β over $\{0.05, 0.1, 0.2, 0.4, 0.8, 1.0\}$. For each target dataset, we choose the best-performing model on its development set and report results on the test set.[13]

5 Results

5.1 Synthetic Experiments

To examine how well our methods work in a controlled setup, we train on the biased dataset (B), but evaluate on the unbiased test set (A). As expected, without a method to remove hypothesis-only biases, the baseline fails to generalize to the test set. Examining its predictions, we found that the baseline model learned to rely on the presence/absence of the bias term c, always predicting TRUE/FALSE respectively.

Table 1 shows the results of our two proposed methods. As we increase the hyper-parameters α and β, our methods initially behave like the baseline, learning the training set but failing on the test set. However, with strong enough hyper-parameters (moving towards the bottom in the tables), they perform perfectly on both the biased training set and the unbiased test set. For Method 1, stronger hyper-parameters work better.

[9] Detailed descriptions of these datasets can be found in Poliak et al. (2018b).

[10] We leave additional NLI datasets, such as the Diverse NLI Collection (Poliak et al., 2018a), for future work.

[11] Many NLI models encode P and H separately (Rocktäschel et al., 2016; Mou et al., 2016; Liu et al., 2016; Cheng et al., 2016; Chen et al., 2017), although some share information between the encoders via attention (Parikh et al., 2016; Duan et al., 2018).

[12] Specifically, representations are concatenated, subtracted, and multiplied element-wise.

[13] For MNLI, since the test sets are not available, we tune on the matched dev set and evaluate on the mismatched dev set, or vice versa. For GLUE, we tune on MNLI matched.

	Test On Target Dataset			Test On SNLI	
Target Test Dataset	Baseline	Δ Method 1	Δ Method 2	Δ Method 1	Δ Method 2
SCITAIL	58.14	-0.47	-7.06	-0.18	-9.06
ADD-ONE-RTE	66.15	0.00	17.31	-2.29	-49.63
JOCI	41.50	0.24	-1.87	-0.44	-5.92
MPE	57.65	0.45	-5.30	-0.57	-0.54
DPR	49.86	1.10	-0.45	-0.73	-7.81
MNLI matched	45.86	1.38	-2.10	-1.25	-8.93
FN+	50.87	1.61	6.16	-1.94	-0.44
MNLI mismatched	47.57	1.67	-3.91	-1.25	-8.93
SICK	25.64	1.80	31.11	-0.57	-8.93
GLUE	38.50	1.99	4.71	-1.25	-8.93
SPR	52.48	6.51	12.94	-1.76	-14.01
SNLI-hard	68.02	-1.75	-12.42		

Table 2: Accuracy results of transferring representations to new datasets. In all cases the models are trained on SNLI. Left: baseline results on target test sets and differences between the proposed methods and the baseline. Right: test results on SNLI with the models that performed best on each target dataset's dev set. Δ are absolute differences between the method and the baseline on each target test set (left) or between the method and the baseline performance (84.22) on SNLI test (right). Black rectangles show relative changes in each column.

Method 2, in particular, breaks down with too many random samples (increasing α), as expected. We also found that Method 1 did not require as strong β as Method 2. From the synthetic experiments, it seems that Method 1 learns to ignore the bias c and learn the desired relationship between P and H across many configurations, while Method 2 requires much stronger β.

5.2 Results on existing NLI datasets

Table 2 (left block) reports the results of our proposed methods compared to the baseline in application to the NLI datasets. The method using the hypothesis-only classifier to remove hypothesis-only biases from the model (Method 1) outperforms the baseline in 9 out of 12 target datasets ($\Delta > 0$), though most improvements are small. The training method using negative sampling (Method 2) only outperforms the baseline in 5 datasets, 4 of which are cases where the other method also outperformed the baseline. These gains are much larger than those of Method 1.

We also report results of the proposed methods on the SNLI test set (right block). As our results improve on the target datasets, we note that Method 1's performance on SNLI does not drastically decrease (small Δ), even when the improvement on the target dataset is large (for example, in SPR). For this method, the performance on SNLI drops by just an average of 1.11 (0.65 STDV). For Method 2, there is a large decrease on SNLI as results drop by an average of 11.19 (12.71 STDV). For these models, when we see large improvement on a target dataset, we often see a large drop on SNLI. For example, on ADD-ONE-RTE, Method 2 outperforms the baseline by roughly 17% but performs almost 50% lower on SNLI. Based on this, as well as the results on the synthetic dataset, Method 2 seems to be much more unstable and highly dependent on the right hyper-parameters.

6 Analysis

Our results demonstrate that our approaches may be robust to many datasets with different types of bias. We next analyze our results and explore modifications to the experimental setup that may improve model transferability across NLI datasets.

6.1 Interplay with known biases

A priori, we expect our methods to provide the most benefit when a target dataset has no hypothesis-only biases or such biases that differ from ones in the training data. Previous work estimated the amount of bias in NLI datasets by comparing the performance of a hypothesis-only classifier with the majority baseline (Poliak et al., 2018b). If the classifier outperforms the baseline, the dataset is said to have hypothesis-only biases. We follow a similar idea for estimating how similar the biases in a target dataset are to those in the source dataset. We compare the performance of a hypothesis-only classifier trained on SNLI and evaluated on each target dataset, to a majority baseline of the most frequent class in each target dataset's training set (Maj). We also compare to a hypothesis-only classifier trained and tested on

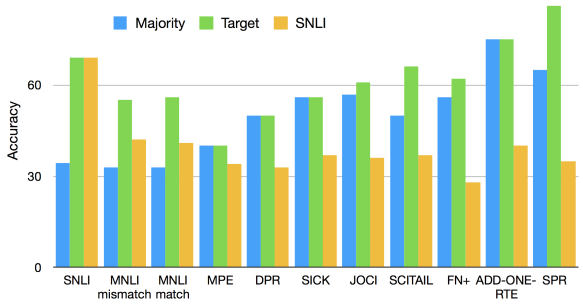


Figure 2: Accuracies of majority and hypothesis-only baselines on each dataset (x-axis). The datasets are generally ordered by increasing difference between a hypothesis-only model trained on the target dataset (green) compared to trained on SNLI (yellow).

each target dataset.[14]

Figure 2 shows the results. When the hypothesis-only model trained on SNLI is tested on the target datasets, the model performs below Maj (except for MNLI), indicating that these target datasets contain different biases than those in SNLI. The largest difference is on SPR: a hypothesis-only model trained on SNLI performs over 50% worse than one trained on SPR. Indeed, our methods lead to large improvements on SPR (Table 2), indicating that they are especially helpful when the target dataset contains different biases. On MNLI, this hypothesis-only model performs 10% above Maj, and roughly 20% worse compared to when trained on MNLI, suggesting that MNLI and SNLI have similar biases. This may explain why our methods only slightly outperform the baseline on MNLI (Table 2).

The hypothesis-only model trained on each target dataset did not outperform Maj on DPR, ADD-ONE-RTE, SICK, and MPE, suggesting that these datasets do not have noticeable hypothesis-only biases. Here, as expected, we observe improvements when our methods are tested on these datasets, to varying degrees (from 0.45 on MPE to 31.11 on SICK). We also see improvements on datasets with biases (high performance of training on each dataset compared to the corresponding majority baseline), most noticeably SPR. The only exception seems to be SCITAIL, where we do not improve despite it having different biases than SNLI. However, when we strengthen α and β (below), Method 1 outperforms the baseline.

[14] A reviewer noted that this method may miss similar bias "types" that are achieved through different lexical items. We note that our use of pre-trained word embeddings might mitigate this concern.

Dataset	Base	Method 1	Δ
JOCI	41.50	39.29	-2.21
SNLI	84.22	82.40	-1.82
DPR	49.86	49.41	-0.45
MNLI matched	45.86	46.12	0.26
MNLI mismatched	47.57	48.19	0.62
MPE	57.65	58.60	0.95
SCITAIL	58.14	60.82	2.68
ADD-ONE-RTE	66.15	68.99	2.84
GLUE	38.50	41.58	3.08
FN+	50.87	56.31	5.44
SPR	52.48	58.68	6.20
SICK	25.64	36.59	10.95
SNLI-hard	68.02	63.81	-4.21

Table 3: Results with stronger hyper-parameters for Method 1 vs. the baseline. Δ's are absolute differences.

Finally, both methods obtain improved results on the GLUE diagnostic set, designed to be bias-free. We do not see improvements on SNLI-hard, indicating it may still have biases – a possibility acknowledged by Gururangan et al. (2018).

6.2 Stronger hyper-parameters

In the synthetic experiment, we found that increasing α and β improves the models' ability to generalize to the unbiased dataset. Does the same apply to natural NLI datasets? We expect that strengthening the auxiliary losses (L_2 in our methods) during training will hurt performance on the original data (where biases are useful), but improve on the target data, which may have different or no biases (Figure 2). To test this, we increase the hyper-parameter values during training; we consider the range $\{1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0\}$.[15] While there are other ways to strengthen our methods, e.g., increasing the number or size of hidden layers (Elazar & Goldberg, 2018), we are interested in the effect of α and β as they control how much bias is subtracted from our baseline model.

Table 3 shows the results of Method 1 with stronger hyper-parameters on the existing NLI datasets. As expected, performance on SNLI test sets (SNLI and SNLI-hard in Table 3) decreases more, but many of the other datasets benefit from stronger hyper-parameters

(compared with Table 2). We see the largest improvement on SICK, achieving over 10% increase compared to the 1.8% gain in Table 2. As for Method 2, we found large drops in quality even

[15] The synthetic setup required very strong hyper-parameters, possibly due to the clear-cut nature of the task. In the natural NLI setting, moderately strong values sufficed.

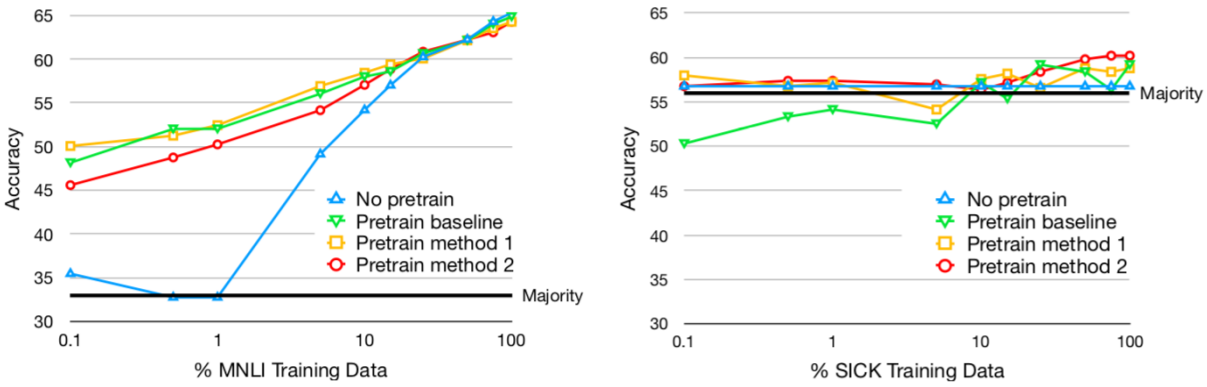


Figure 3: Effect of fine-tuning with the baseline and the proposed methods on MNLI (left) and SICK (right).

in our basic configurations (Appendix A.3), so we do not increase the hyper-parameters further. This should not be too surprising, adding too many random premises will lead to a model's degradation.

6.3 Fine-tuning on target datasets

Our main goal is to determine whether our methods help a model perform well across multiple datasets by ignoring dataset-specific artifacts. In turn, we did not update the models' parameters on other datasets. But, what if we are given different amounts of training data for a new NLI dataset?

To determine if our approach is still helpful, we updated four models on increasing sizes of training data from two target datasets (MNLI and SICK). All three training approaches—the baseline, Method 1, and Method 2—are used to pre-train a model on SNLI and fine-tune on the target dataset. The fourth model is the baseline trained only on the target dataset. Both MNLI and SICK have the same label spaces as SNLI, allowing us to hold that variable constant. We use SICK because our methods resulted in good gains on it (Table 2). MNLI's large training set allows us to consider a wide range of training set sizes.[16]

Figure 3 shows the results on the dev sets. In MNLI, pre-training is very helpful when fine-tuning on a small amount of new training data, although there is little to no gain from pre-training with either of our methods compared to the baseline. This is expected, as we saw relatively small gains with the proposed methods on MNLI, and can be explained by SNLI and MNLI having similar biases. In SICK, pre-training with either of our methods is better in most data regimes, especially with very small amounts of target training data.[17]

7 Related Work

Biases and artifacts in NLU datasets Many natural language undersrtanding (NLU) datasets contain annotation artifacts. Early work on NLI, also known as recognizing textual entailment (RTE), found biases that allowed models to perform relatively well by focusing on syntactic clues alone (Snow et al., 2006; Vanderwende & Dolan, 2006). Recent work also found artifacts in new NLI datasets (Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018b).

Other NLU datasets also exhibit biases. In ROC Stories (Mostafazadeh et al., 2016), a story cloze dataset, Schwartz et al. (2017b) obtained a high performance by only considering the candidate endings, without even looking at the story context. In this case, stylistic features of the candidate endings alone, such as the length or certain words, were strong indicators of the correct ending (Schwartz et al., 2017a; Cai et al., 2017). A similar phenomenon was observed in reading comprehension, where systems performed non-trivially well by using only the final sentence in the passage or ignoring the passage altogether (Kaushik & Lipton, 2018). Finally, multiple studies found non-trivial performance in visual question answering (VQA) by using only the question, without access to the image, due to question biases (Zhang et al., 2016; Kafle & Kanan, 2016, 2017; Goyal et al., 2017; Agrawal et al., 2017).

[16] We hold out 10K examples from the training set for dev as gold labels for the MNLI test set are not publicly available. We evaluate on MNLI's matched dev set to assure consistent domains when fine-tuning.

[17] Note that SICK is a small dataset (10K examples), which explains why the model without pre-training does not benefit from more data, barely surpassing the majority baseline.

Transferability across NLI datasets It has been known that many NLI models do not transfer across NLI datasets. Chen Zhang's thesis (Zhang, 2010) focused on this phenomena as he demonstrated that "techniques developed for textual entailment" datasets, e.g., RTE-3, do not transfer well to other domains, specifically *conversational entailment* (Zhang & Chai, 2009, 2010). Bowman et al. (2015) and Williams et al. (2018) demonstrated (specifically in their respective Tables 7 and 4) how models trained on SNLI and MNLI may not transfer well across other NLI datasets like SICK. Talman & Chatzikyriakidis (2018) recently reported similar findings using many advanced deep-learning models.

Improving model robustness Neural networks are sensitive to adversarial examples, primarily in machine vision, but also in NLP (Jia & Liang, 2017; Belinkov & Bisk, 2018; Ebrahimi et al., 2018; Heigold et al., 2018; Mudrakarta et al., 2018; Ribeiro et al., 2018; Belinkov & Glass, 2019). A common approach to improving robustness is to include adversarial examples in training (Szegedy et al., 2014; Goodfellow et al., 2015). However, this may not generalize well to new types of examples (Xiaoyong Yuan, 2017; Tramr et al., 2018).

Domain-adversarial neural networks aim to increase robustness to domain change, by learning to be oblivious to the domain using gradient reversals (Ganin et al., 2016). Our methods rely similarly on gradient reversals when encouraging models to ignore dataset-specific artifacts. One distinction is that domain-adversarial networks require knowledge of the domain at training time, while our methods learn to ignore latent artifacts and do not require direct supervision in the form of a domain label.

Others have attempted to remove biases from learned representations, e.g., gender biases in word embeddings (Bolukbasi et al., 2016) or sensitive information like sex and age in text representations (Li et al., 2018). However, removing such attributes from text representations may be difficult (Elazar & Goldberg, 2018). In contrast to this line of work, our final goal is not the removal of such attributes per se; instead, we strive for more robust representations that better transfer to other datasets, similar to Li et al. (2018).

Recent work has applied adversarial learning to NLI. Minervini & Riedel (2018) generate adversarial examples that do not conform to logical rules and regularize models based on those examples. Similarly, Kang et al. (2018) incorporate external linguistic resources and use a GAN-style framework to adversarially train robust NLI models. In contrast, we do not use external resources and we are interested in mitigating hypothesis-only biases. Finally, a similar approach has recently been used to mitigate biases in VQA (Ramakrishnan et al., 2018; Grand & Belinkov, 2019).

8 Conclusion

Biases in annotations are a major source of concern for the quality of NLI datasets and systems. We presented a solution for combating annotation biases by proposing two training methods to predict the probability of a premise given an entailment label and a hypothesis. We demonstrated that this discourages the hypothesis encoder from learning the biases to instead obtain a less biased representation. When empirically evaluating our approaches, we found that in a synthetic setting, as well as on a wide-range of existing NLI datasets, our methods perform better than the traditional training method to predict a label given a premise-hypothesis pair. Furthermore, we performed several analyses into the interplay of our methods with known biases in NLI datasets, the effects of stronger bias removal, and the possibility of fine-tuning on the target datasets.

Our methodology can be extended to handle biases in other tasks where one is concerned with finding relationships between two objects, such as visual question answering, story cloze completion, and reading comprehension. We hope to encourage such investigation in the broader community.

Acknowledgements

We would like to thank Aviad Rubinstein and Cynthia Dwork for discussing an earlier version of this work and the anonymous reviewers for their useful comments. Y.B. was supported by the Harvard Mind, Brain, and Behavior Initiative. A.P. and B.V.D were supported by JHU-HLTCOE and DARPA LORELEI. A.M.R gratefully acknowledges the support of NSF 1845664. Views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Martín Abadi and David G. Andersen. Learning to protect communications with adversarial neural cryptography. *arXiv*, 2016. URL https://arxiv.org/abs/1610.06918.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. *arXiv preprint arXiv:1712.00377*, 2017.

Yonatan Belinkov and Yonatan Bisk. Synthetic and Natural Noise Both Break Neural Machine Translation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ8vJebC-.

Yonatan Belinkov and James Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics (TACL)*, 7:49–72, 2019. doi: 10.1162/tacl_a_00254. URL https://doi.org/10.1162/tacl_a_00254.

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander Rush. On Adversarial Removal of Hypothesis-only Bias in Natural Language Inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM, Oral presentation)*, June 2019.

Yoshua Bengio and Jean-Sébastien Senecal. Quick Training of Probabilistic Neural Nets by Importance Sampling. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, AISTATS 2003, Key West, Florida, USA, January 3-6, 2003*, pp. 1–9, 2003. URL http://research.microsoft.com/en-us/um/cambridge/events/aistats2003/proceedings/164.pdf.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

Zheng Cai, Lifu Tu, and Kevin Gimpel. Pay Attention to the Ending: Strong Neural Baselines for the ROC Story Cloze Task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 616–622. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-2097. URL http://www.aclweb.org/anthology/P17-2097.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1657–1668. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1152. URL http://www.aclweb.org/anthology/P17-1152.

Jianpeng Cheng, Li Dong, and Mirella Lapata. Long Short-Term Memory-Networks for Machine Reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 551–561, Austin, Texas, November 2016. Association for Computational Linguistics. URL https://aclweb.org/anthology/D16-1053.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/D17-1070.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steven Pullman. Using the framework. Technical Report LRE 62-051 D-16, The FraCaS Consortium, 1996.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pp. 177–190. Springer, 2006.

Chaoqun Duan, Lei Cui, Xinchi Chen, Furu Wei, Conghui Zhu, and Tiejun Zhao. Attention-Fused Deep Matching Network for Natural Language Inference. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 4033–4040. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/561. URL https://doi.org/10.24963/ijcai.2018/561.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. On Adversarial Examples for Character-Level Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 653–663. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/C18-1055.

Yanai Elazar and Yoav Goldberg. Adversarial Removal of Demographic Attributes from Text Data.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 11–21. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/D18-1002.

Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *International Conference on Machine Learning*, pp. 1180–1189, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1): 2096–2030, 2016.

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 650–655, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P18-2103.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, volume 1, pp. 3, 2017.

Gabriel Grand and Yonatan Belinkov. Adversarial Regularization for Visual Question Answering: Strengths, Shortcomings, and Side Effects, June 2019.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N18-2017.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

Georg Heigold, Günter Neumann, and Josef van Genabith. How Robust Are Character-Based Word Embeddings in Tagging and MT Against Wrod Scramlbing or Randdm Nouse? In *Proceedings of the 13th Conference of The Association for Machine Translation in the Americas (Volume 1: Research Track*, pp. 68–79, March 2018.

Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2011–2021, Copenhagen, Denmark, September 2017. URL https://www.aclweb.org/anthology/D17-1215.

K. Kafle and C. Kanan. Answer-Type Prediction for Visual Question Answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4976–4984, June 2016. doi: 10.1109/CVPR.2016.538.

Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017.

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2418–2428, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P18-1225.

Divyansh Kaushik and Zachary C. Lipton. How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5010–5015, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D18-1546.

Tushar Khot, Ashish Sabharwal, and Peter Clark. SciTail: A Textual Entailment Dataset from Science Question Answering. In *AAAI*, 2018.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. Natural Language Inference from Multiple Premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 100–109, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/I17-1011.

Yitong Li, Timothy Baldwin, and Trevor Cohn. Towards Robust and Privacy-preserving Text Representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 25–30. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/P18-2005.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. Learning natural language inference using bidirectional LSTM model and inner-attention. *arXiv preprint arXiv:1605.09090*, 2016.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf. ACL Anthology Identifier: L14-1314.

Pasquale Minervini and Sebastian Riedel. Adversarially Regularising Neural NLI Models to Integrate Logical Background Knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018)*. Association for Computational Linguistics, 2018.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N16-1098.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. Natural Language Inference by Tree-Based Convolution and Heuristic Matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 130–136, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://anthology.aclweb.org/P16-2022.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. Did the Model Understand the Question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1896–1906. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/P18-1176.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2249–2255. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1244. URL http://www.aclweb.org/anthology/D16-1244.

Ellie Pavlick and Chris Callison-Burch. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2164–2173. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1204. URL http://www.aclweb.org/anthology/P16-1204.

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. FrameNet+: Fast Paraphrastic Tripling of FrameNet. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 408–413, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P15-2067.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018a.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S18-2023.

Altaf Rahman and Vincent Ng. Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 777–789, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D12-1071.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pp. 1548–1558, 2018.

Pushpendre Rastogi and Benjamin Van Durme. Sublinear partition estimation. *arXiv preprint arXiv:1508.01596*, 2015.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, 3:475–488, 2015. ISSN 2307-387X. URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/674.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 856–865. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/P18-1079.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. Reasoning about Entailment with Neural Attention. In *International Conference on Learning Representations (ICLR)*, 2016.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 15–25. Association for Computational Linguistics, 2017a. doi: 10.18653/v1/K17-1004. URL http://www.aclweb.org/anthology/K17-1004.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. Story Cloze Task: UW NLP System. In *Proceedings of LSDSem*, 2017b.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. Tackling the Story Ending Biases in The Story Cloze Test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 752–757, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P18-2119.

Rion Snow, Lucy Vanderwende, and Arul Menezes. Effectively Using Syntax for Recognizing False Entailment. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 2006. URL http://www.aclweb.org/anthology/N06-1005.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Aarne Talman and Stergios Chatzikyriakidis. Neural Network Models for Natural Language Inference Fail to Capture the Semantics of Inference. *CoRR*, abs/1810.09774, 2018. URL http://arxiv.org/abs/1810.09774.

Florian Tramr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ.

Masatoshi Tsuchiya. Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment. In *11th International Conference on Language Resources and Evaluation (LREC2018)*, 2018.

Lucy Vanderwende and William B Dolan. What syntax can contribute in the entailment task. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pp. 205–216. Springer, 2006.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W18-5446.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. Inference is Everything: Recasting Semantic Resources into a Unified Evaluation Framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 996–1005, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/I17-1100.

Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Qile Zhu Xiaolin Li Xiaoyong Yuan, Pan He. Adversarial Examples: Attacks and Defenses for Deep Learning. *arXiv preprint arXiv:1712.07107*, 2017.

Chen Zhang. *Natural Language Interference from Textual Entailment to Conversation Entailment*. PhD thesis, Michigan State University, East Lansing, MI, USA, 2010. AAI3435149.

Chen Zhang and Joyce Chai. Towards Conversation Entailment: An Empirical Investigation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 756–766, Cambridge, MA, October 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D10-1074.

Chen Zhang and Joyce Y. Chai. What Do We Know About Conversation Participants: Experiments on Conversation Entailment. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, pp. 206–215, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-64-0. URL http://dl.acm.org/citation.cfm?id=1708376.1708406.

P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh. Yin and Yang: Balancing and Answering Binary Visual Questions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5014–5022, June 2016. doi: 10.1109/CVPR.2016.542.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics*, 5:379–395, 2017. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/1082.

A Appendix

A.1 Mapping labels

Each premise-hypothesis pair in SNLI is labeled as ENTAILMENT, NEUTRAL, or CONTRADICTION. MNLI, SICK, and MPE use the same label space. Examples in JOCI are labeled on a 5-way ordinal scale. We follow Poliak et al. (2018b) by converting it "into 3-way NLI tags where 1 maps to CONTRADICTION, 2-4 maps to NEUTRAL, and 5 maps to ENTAILMENT." Since examples in SCITAIL are labeled as ENTAILMENT or NEUTRAL, when evaluating on SCITAIL, we convert the model's CONTRADICTION to NEUTRAL. ADD-ONE-RTE and the recast datasets also model NLI as a binary prediction task. However, their label sets are ENTAILED and NOT-ENTAILED. In these cases, when the models predict ENTAILMENT, we map the label to ENTAILED, and when the models predict NEUTRAL or CONTRADICTION, we map the label to NOT-ENTAILED.

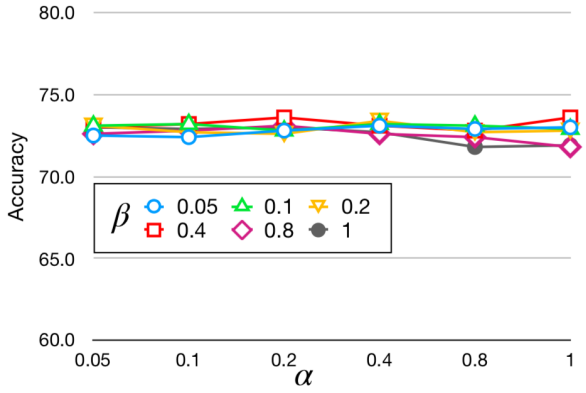
A.2 Implementation details

For our experiments on the synthetic dataset, the characters are embedded with 10-dimensional vectors. Input strings are represented as a sum of character embeddings, and the premise-hypothesis pair is represented by a concatenation of these embeddings. The classifiers are single-layer MLPs of size 20 dimensions. We train these models with SGD until convergence. For the traditional NLI datasets, we use pre-computed 300-dimensional GloVe embeddings (Pennington et al., 2014).[18] The sentence representations learned by the BiLSTM encoders and the MLP classifier's hidden layer have a dimensionality of 2048 and 512 respectively. We follow `InferSent`'s training regime, using SGD with an initial learning rate of 0.1 and optional early stopping. See Conneau et al. (2017) for details.

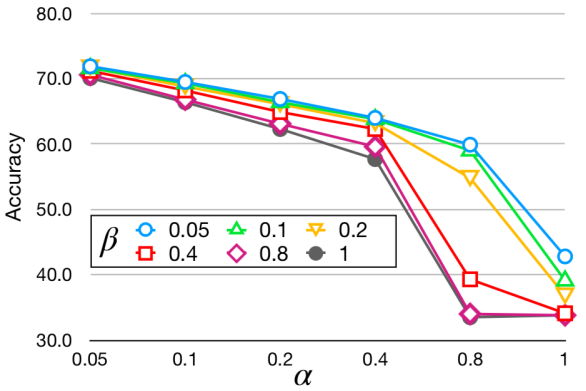
A.3 Hyper-parameter sweeps

Here we provide 10-fold cross-validation results on a subset of the SNLI training data (50K sentences) with different settings of our hyper-parameters. Figure 4b shows the dev set results with different configurations of Method 2. Notice that performance degrades quickly when we increase the fraction of random premises (large α). In contrast, the results with Method 1 (Figure 4a) are more stable.

[18] Specifically, glove.840B.300d.zip.



(a) Method 1



(b) Method 2

Figure 4: Cross-validation results.