Data Augmentation and Ensemble Debiasing to Mitigate Data Artifacts in SNLI Dataset

Abstract

Natural Language Inference (NLI) models can achieve high accuracy for in domain problems, but can generalize poorly to adversarial settings when they learn embedded patterns in the training set that are not displayed in a new test set. We investigate a simple debiased model implementation in conjunction with data augmentation to improve generalization to challenging evaluation datasets. Our technique shows modest improvements on template-based heuristic challenges but fails to yield meaningful gains on more complex reasoning challenges, revealing a tension between mitigating known bias and generating robust reasoning. We provide empirical evidence and reproducible code showing that suppressing artifacts alone is insufficient for building next level NLI systems.

1 Introduction

With the advent of innovative model architectures and training schema it has become easier to fine tune pre-trained models to perform natural language tasks like sentiment analysis, question answering, or token classification with high degrees of accuracy. However, this in-domain accuracy can shroud underlying messages in what these models are actually learning. Through data and model inference techniques it can be shown that models may actually be learning hidden patterns in training sets rather than a broad-based understanding of language. This pattern-based learning can fall apart when presented with inputs that do not match expectation and can lead to poor performance when true reasoning is needed.

1.1 Hypothesis only baselines

A hypothesis-only model, in the context of natural language inference, is one in which premises are withheld during training. In practice this would mean determining whether an input hypothesis is related to an unknown premise. Poliak et al. (2018) showed that hypothesis-only baselines achieve surprisingly high accuracy on many NLI datasets, calling into question whether high in-domain accuracy reflects proper reasoning or exploitation of correlations within the training data. We follow the hypothesis-only baseline technique to identify potential dataset artifacts. We then build on this premise by utilizing hypothesis-only models as biased learners in an ensemble debiasing model.

1.2 Ensemble debiasing

Clark et al. (2019) projected the idea that given known biases in natural language datasets one could utilize an ensemble architecture as a strategy to debias a model. They proposed explicitly using a “biased” model, one that is trained/known to perform well on in-domain data but breaks easily with new data, in conjunction with a main model to reduce reliance on dataset artifacts. Following this work we treat a hypothesis-only model as a bias expert and combine it with a main NLI model through logit subtraction. Take the logits to be the raw evidence our models output for each class. By directly subtracting the bias model’s raw evidence from the main model’s logits we penalize the probability, as calculated in the softmax operation in cross-entropy loss, for classes the hypothesis-only model was confident in. This forces the main-model to develop reasoning that is incongruent with hypothesis-only biased reasoning.

1.3 Data Augmentation

Three categories of heuristics NLI models may be using during training are lexical overlap, subsequence, and constituency as introduced by (McCoy et al., 2019). McCoy et al created the HANS (Heuristic Analysis for NLI Systems) dataset which includes examples designed to trick these heuristics. In the lexical overlap case, a model answers correctly by identifying high word similarity between the premise and hypothesis. Subsequence assumes a premise will be predicted as entailed if any sub sentence from the premise is used as the hypothesis. Example - Premise: "the doctor near the actor danced", Hypothesis: "the actor danced", Model Prediction: Entailment, Gold Label: Non-entailment. The constituency heuristic assumes entailment when the hypothesis is a complete subtree of the premise parse tree. We utilize HANS as a targeted hold-out evaluation set and as augmentation to our standard training set for mitigating data artifact biases.

2 Methodology

2.1 Datasets

We used the SNLI (Stanford Natural Language Inference) corpus, which contains 570k human-written sentence pairs along with labels (entailment, contradiction, neutral), as the foundational dataset for all training (Bowman et al., 2015). This dataset contains train, validation, and test splits, but prior to usage we filtered out examples that had label = -1 signifying no gold label was assigned.

We used the HANS dataset which contains two files, evaluation and training, that have 30k examples in each. This dataset contains increased info for each example including sentence parses, unique pair ID's, an overall heuristic label (lexical_overlap, subsequence, constituent), and fine-grained subcase labeling, each heuristic has 10 subcases. We converted string labeling, which only contained entailment/non-entailment, to SNLI integer labeling. Anything labeled non-entailment was recast to the SNLI "contradiction" value since we have no indication of neutrality. Both sets contain equivalent counts of each of the three heuristics as displayed by figure 1.

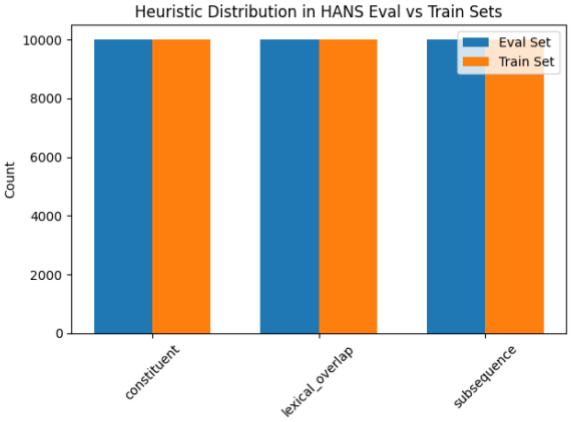

Figure 1: Count of Heuristic type in HANS

Lastly, a leakage check was run between the training and evaluation sets and confirmed that no pair, heuristic, label combination is shared between the evaluation and test set. There were 14 examples of remote similarity, defined as an example where the eval and test sets share any combination of premise and hypothesis, example given in Table 1.

Dataset	Premise	Hypothesis
Eval	The artist slept, and the tourists introduced the professors	The artist slept.
Train	The actor advised the students, and the artist slept.	The artist slept.

Table 1: Example of hypothesis overlap

Because we only found 14 examples with slight similarity amounting to 0.023% of total examples we did not filter examples from the HANS training and evaluation sets.

We used the ANLI (Adversarial Natural Language Inference) dataset as our primary complex evaluation benchmark for testing model generalization beyond standard test sets (Nie et al., 2020). ANLI contains around 163k premise-hypothesis pairs written by humans who attempted to make difficult examples to fool state of the art NLI models. The dataset is organized into three rounds that get progressively harder, but for our testing we concatenated them all together. This dataset targets linguistic reasoning and complexity that models trained on SNLI are typically inaccurate on, so it is a good genius level baseline.

2.2 Model Choice

The ELECTRA-small discriminator model introduced by Clark et al. (2020) was utilized as the backbone for the bias and main models for all training variations. This model was chosen because it offers strong NLI performance while being computationally efficient during hyperparameter tuning and model ablations. For debiasing we initialize a fresh pretrained version of the main model so it can learn premise dependent signals during debias training instead of attempting to unlearn our base models SNLI data artifacts. The bias model utilizes the same architecture but has been pre-trained on hypothesis only inputs and then is frozen during the ensemble training, so there are stable biased outputs utilized by the ensemble class.

2.3 Debiasing

During the debias process we treat the hypothesis-only model as a frozen bias expert and subtract its contributions from the main models output scores. In practice if z_{main} and z_{bias} are the logits from the main model and the hypothesis-only model, taking the shape [batch size, num_labels], we generate debiased scores $z_{deb} = z_{main} - \alpha * z_{bias}$. The debiased logits are then passed through the softmax function, used by cross-entropy loss, which yields $p_{deb} = \frac{p_{main}}{p_{bias}^{\alpha}}$ so the classes favored by the bias model are down weighted. The main model will then have to learn new premise-dependent evidence to recover the proper label. During training we progressed from coarse to fine alpha sweeps attempting to identify a reasonable alpha value that enhanced performance on HANS without sacrificing efficacy on SNLI, we ended by choosing $\alpha = 0.5$ which had the best effects.

2.4 Experiments

We compared a set of controlled model and data conditions to diagnose data artifacts within the SNLI dataset. We then isolate the effects of our debiased implementations to gain a better understanding of the improvement. To perform the isolation, we trained five model variations summarized in Table 2.

BASE = single usage of ELECTRA-small
DEBIASED = ensemble w/ logit subtraction
SNLI/HANS = datasets used in training
X = standard implementation
X = fed only the hypothesis

MODEL NAME	BASE	DEBIASED	SNLI	HANS
Baseline	X		X	
Hypothesis-only	X		X	
Baseline w/ HANS	X		X	X
Debiased Vanilla		X	X	
Debiased w/ HANS		X	X	X

Table 2: Model training variation depiction

The Baseline model serves as our control with a standard ELECTRA-small model fine-tuned on SNLI training data. To identify the strength of data artifacts in SNLI we then trained a Hypothesis-only model using just the hypothesis text from SNLI as input. The goal here was to see if superficial patterns in the hypothesis data lead to better than chance predictions. The Baseline w/ HANS model was trained with combined SNLI and HANS training sets to assess whether exposure to the challenge examples improves predictions without explicit debiasing. Our primary debiasing solution was the Debiased Vanilla model which employed the frozen Hypothesis-only model as the biased learner in concert with a fresh ELECTRA-small model. Finally, we experimented with using the same architecture as the Debiased Vanilla model combined with the data augmentation strategy (SNLI + HANS) to determine whether the two approaches yield complementary benefits. Across all experiments we kept the base hyperparameters constant for reproducibility.

3 Results

Model accuracy, as tested against our three different NLI evaluation sets, is depicted in Table 3. We chose the following three datasets to depict different aspects of our story: SNLI accuracy for in-domain performance, HANS overall accuracy for general robustness to hypothesis heuristics, and ANLI to assess generalization to complex adversarial examples.

MODEL NAME	**SNLI**	**HANS**	**ANLI**
Baseline	89.7	54.13	32.06
Hypothesis-only	69.78	22.55	32.97
Baseline w/ HANS	89.71	100	32.06
Debiased Vanilla	89.72	56.11	32.63
Debiased w/ HANS	89.34	100	32.75

Table 3: Model accuracy results

The Baseline model fine-tuned on SNLI alone achieves strong in-domain accuracy (89.7%) but exhibits difficulty in HANS overall accuracy (54.13%). The proposed reliance on statistical phenomena in the hypothesis is further shown in the breakdown of entailment vs non-entailment accuracy as depicted in Figure 2. We see that the Baseline model had high recall (99.1%) for entailment labels but terrible precision (52.17%) since it predicted entailment for 95% of the examples (28,481/30,000). This aligns with how HANS heuristics trick models, who are trained with dataset's that contain artifacts, into predicting entailment. The Baseline model achieved a near random (32.06%) accuracy for the three-class prediction on the ANLI evaluation set.

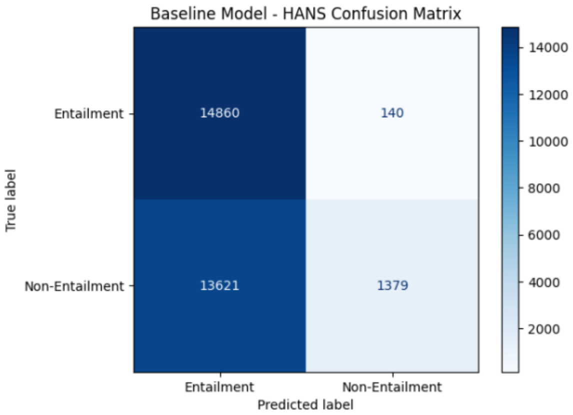

Figure 2: Baseline HANS confusion matrix

The Hypothesis-only model, trained exclusively on hypothesis inputs, reached an improbable accuracy of 69.78% on the SNLI dataset adding further evidence that the SNLI dataset has strong hypothesis artifacts. This model displayed performance collapse on HANS (22.55%), but interestingly matched ANLI accuracy (32.97%), with respect to the Baseline model. This divergence between our two adversarial datasets highlights differences in the structure of HANS and ANLI. While HANS was created with explicit hostile label flipping, leading to worse than random results, ANLI's near random performance suggests that its examples were designed to be difficult but not systematically related to hypothesis-only biases. The Debiased Vanilla model maintained SNLI accuracy at 89.72% and improved HANS accuracy to 56.11% while again maintaining ANLI accuracy at 32.63%. Most notably, both models that were trained with HANS augmentation (Baseline w/ HANS and Debiased w/ HANS) achieved perfect accuracy (100%) on the HANS evaluation set, indicating complete mitigation of the hypothesis-based heuristics. This represents a dramatic improvement from the Baseline models 54.13%. We believe this accuracy to be true based on the data leakage checks we ran, and due to the balanced label accuracy depicted in Figure 3. We see that both models achieve perfect classification on both entailment and non-entailment HANS examples, confirming that training on HANS data eliminated vulnerability to lexical overlap, subsequence, and constituent heuristics rather than simply shifting all predictions to be non-entailment.

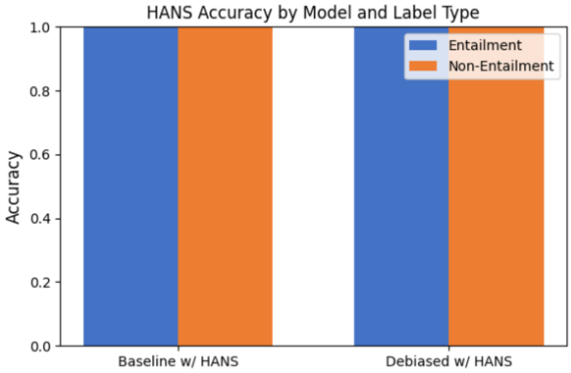

Figure 3: Entailment vs. non-entailment accuracy

Both HANS augmented models had no difference in their generalization to ANLI data. The Baseline w/ HANS had 32.06% accuracy on ANLI and the Debiased w/ HANS model had 32.75% accuracy (Figure 4). These results reveal that successfully solving the HANS heuristics does not have complementary benefits with regard to improved ANLI performance. Explicit exposure to HANS examples provides targeted help with hypothesis heuristics, and ensemble debiasing encourages better premise-hypothesis learning, but neither technique have any effect on the diverse and difficult ANLI examples. Our debiasing approaches maintained SNLI evaluation accuracy within 0.4 percentage points of the Baseline model (89.34%-89.71%).

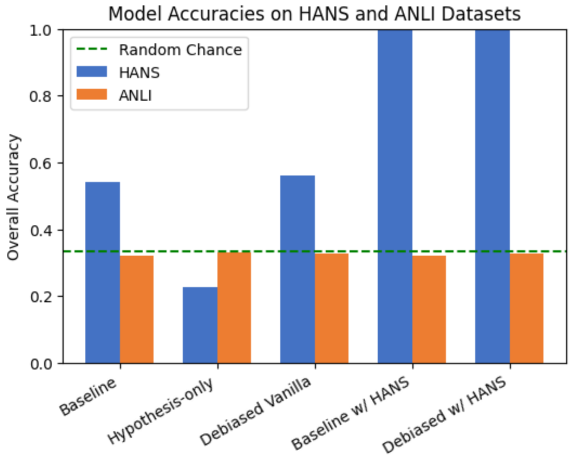

Figure 4: HANS and ANLI model accuracy

4 Discussion

Our results demonstrate that ensemble debiasing via logit subtraction partially reduces the reliance on hypothesis artifacts by down weighting predictions from a frozen hypothesis-only model, achieving modest improvements on HANS (54.13% to 56.11%) but reveal the limitations of artifact specific debiasing for generalization. The debiasing mechanism is intuitive: by subtracting the bias logits, we force the main model to justify predictions via premise-hypothesis interactions rather than heuristic patterns in the hypothesis alone. The perfect HANS accuracy displayed by both models trained with HANS augmented data validates the idea that experience with counter-examples is highly effective for targeted heuristic accuracy improvements. However, the limited improvement in ANLI prediction accuracy across all models reveals fundamental limitations. The Debiased Vanilla model, without ever seeing HANS examples, achieved a 1.98% improvement on HANS but only marginal ANLI gains, and the Debiased w/ HANS model similarly showed no improvement on ANLI. This suggests that neither ensemble debiasing nor HANS data augmentation alone addresses the deeper reasoning issues exposed by ANLI's examples. Our results point to a "whack-a-mole" problem that extends beyond data augmentation where architectural debiasing can eliminate specific known failure modes without cultivating generalized reasoning abilities.

5 Conclusion

We investigated ensemble debiasing as a technique to mitigate data artifacts in NLI models and improve generalization to adversarial evaluation benchmarks. Our debiasing technique achieved modest HANS improvements while maintaining in-domain SNLI performance. Models augmented with HANS training data eliminated the three template-based heuristic failures demonstrating targeted effectiveness. However, limited gains across all model variations on ANLI evaluation data reveal that our debiasing and HANS augmentation techniques had minimal effect on building more robust reasoning. Ultimately, this work contributes both a practical debiasing implementation, and empirical evidence of a fundamental tension between solving known artifact biases and obtaining generalized reasoning models.

References

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805. 01042*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4069–4082.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 632–642.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pages 3428–3448.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 4885–4901.