Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training

Joe Stacey
University College London
ucakjd0@ucl.ac.uk

Pasquale Minervini
University College London
p.minervini@ucl.ac.uk

Haim Dubossarsky
University of Cambridge
hd423@cam.ac.uk

Sebastian Riedel
University College London
s.riedel@cs.ucl.ac.uk

Tim Rocktäschel
University College London
t.rocktaschel@cs.ucl.ac.uk

Abstract

Natural Language Inference (NLI) datasets contain annotation artefacts resulting in spurious correlations between the natural language utterances and their respective entailment classes. These artefacts are exploited by neural networks even when only considering the hypothesis and ignoring the premise, leading to unwanted biases. Belinkov et al. (2019b) proposed tackling this problem via adversarial training, but this can lead to learned sentence representations that still suffer from the same biases. We show that the bias can be reduced in the sentence representations by using an ensemble of adversaries, encouraging the model to jointly decrease the accuracy of these different adversaries while fitting the data. This approach produces more robust NLI models, outperforming previous de-biasing efforts when generalised to 12 other NLI datasets (Belinkov et al., 2019a; Mahabadi et al., 2020). In addition, we find that the optimal number of adversarial classifiers depends on the dimensionality of the sentence representations, with larger sentence representations being more difficult to de-bias while benefiting from using a greater number of adversaries.

1 Introduction

NLI datasets are known to contain artefacts associated with their human annotation processes (Gururangan et al., 2018). Neural models are particularly prone to picking up on artefacts, relying on these biases and spurious correlations rather than acquiring a true understanding of the task. Because these artefacts are often dataset specific (Poliak et al., 2018; Tsuchiya, 2018), models that rely on these artefacts consequently generalise poorly when tested on other datasets (Belinkov et al., 2019a).

One way to alleviate this problem is via *adversarial training*: the task classifier and an adversarial classifier jointly share an encoder, with the adversarial classifier trained to produce the correct predictions by analysing the artefacts in the training data. The encoder optimises the training objective while also reducing the performance of the adversarial classifier. In this context, adversarial training aims to produce sentence representations that do not incorporate information about the artefacts (or *bias*) in the data, resulting in less biased models that generalise better.

Previous studies show that adversarial training is associated with better generalisation performance across other datasets, although there are concerns that the biases are not removed from the model sentence representations, with classifiers able to relearn such biases from the representations after these are frozen (Belinkov et al., 2019b; Elazar and Goldberg, 2018). It is therefore unclear whether any improvements are as a result of the de-biasing, and whether actually removing these biases from the model representations will further improve generalisation. We focus our effort on this discrepancy and argue that, in order to show de-biasing is effective, improvements in performance should also correspond to an observed reduction of the bias in the model representations, therefore creating representations that generalise better to other data.

In this paper we show that NLI models can avoid learning from the hypothesis-only bias, using an *ensemble of adversarial classifiers* to prevent the bias being relearnt from a model's representations. Furthermore, we show that the more bias is removed from the model representations, the better these models generalise to other NLI datasets. Removing the bias from the representations proves to be a highly effective strategy, producing more robust NLI models that outperform previous de-biasing efforts when tested on 12 different NLI datasets (Belinkov et al., 2019a; Mahabadi et al., 2020). In addition, we show that the ability to de-bias a sentence representation depends on its dimensionality, with large sentence representations being harder to de-bias and requiring more adversarial classifiers

during training.

In summary, this paper makes the following core contributions: *i)* We investigate whether using an ensemble of adversarial classifiers can remove the hypothesis-only bias within NLI models. For large enough dimensions, this method achieves a statistically significant reduction of the bias. *ii)* We test whether removing more of the bias improves how well the model generalises. Our method improves model accuracy across 12 NLI datasets and outperforms previous research (Belinkov et al., 2019a; Mahabadi et al., 2020). *iii)* We inspect the optimal number of adversaries to use depending on the dimensionality of the sentence representations. We find that as this dimensionality is increased, more adversaries are required to de-bias a model. *iv)* We compare the effect of adversarial training with a linear classifier to using a non-linear multi-layer perceptron as the adversary, showing that using a more complex adversarial classifier is not always beneficial. Instead, the best choice of adversary depends on the classifier being used to relearn the bias. [1]

2 Related Work

The Hypothesis-Only Bias Gururangan et al. (2018) and Tsuchiya (2018) demonstrate how models can predict the class within the SNLI dataset when only processing the hypothesis, reaching accuracy scores as high as twice the majority baseline (67% vs. 34%). This is possible due to hypothesis-only biases, such as the observation that negation words ("no" or "never") are more commonly used in contradicting hypotheses (Gururangan et al., 2018; Poliak et al., 2018). The hypothesis sentence length is another example of an artefact that models can learn from, with entailment hypotheses being, on average, shorter than either contradiction or neutral hypotheses (Gururangan et al., 2018).

Tsuchiya (2018) show that the hypothesis-only bias predictions are significantly better than the majority baseline for SNLI, although this is not the case for the SICK dataset (Marelli et al., 2014). Poliak et al. (2018) find that human-elicited datasets such as SNLI and MultiNLI have the largest hypothesis-only bias. As a result, our paper focuses on removing the hypothesis-only bias from SNLI, the dataset with the largest hypothesis-only bias reported by Poliak et al. (2018). This bias is

also dataset specific, with Belinkov et al. (2019a) finding that only MultiNLI shares some of the same hypothesis-only bias as the SNLI dataset.

Generalisation to Other Datasets Bowman et al. (2015) and Williams et al. (2018) show that models trained on the SNLI and MultiNLI datasets do not necessarily learn good representations for other NLI datasets, such as SICK. Analogous results were also reported by Talman and Chatzikyriakidis (2018) for more complex models. Gururangan et al. (2018) and Tsuchiya (2018) identify how NLI models perform worse on *hard* examples, which are defined as the examples that a hypothesis-only model has misclassified. This suggests that the success of NLI models may be overstated, with models relying on artefacts in their training data to achieve high performance (Gururangan et al., 2018). Our paper will assess whether NLI models that no longer learn from the hypothesis-only bias can still retain this high level of accuracy.

Biases and Artefacts SNLI and MultiNLI are not the only datasets that suffer from the presence of annotation artefacts and biases. In the past, machine reading datasets were also found to contain syntactic clues that were giving away the correct prediction (Vanderwende and Dolan, 2005; Snow et al., 2006). For instance, Kaushik and Lipton (2018) show that, in several reading comprehension datasets such as bAbI (Weston et al., 2016) and Children's Books Test (Hill et al., 2016), it is possible to get non-trivial results by considering only the last passage of the paragraph. In visual question answering datasets, several studies find it is often possible to answer the question without looking at the corresponding image (Zhang et al., 2016; Kafle and Kanan, 2016; Goyal et al., 2017; Agrawal et al., 2018). Similarly, for the ROCStories corpus (Mostafazadeh et al., 2016), Schwartz et al. (2017) and Cai et al. (2017) show it is possible to achieve non-trivial prediction accuracy by only considering candidate endings and without taking the stories in account. The de-biasing approach introduced in this paper could be applied in any of these situations where a model involves a classifier based on latent representations.

Learning Robust Models Neural models are known to be vulnerable to so-called *adversarial examples*, *i.e.* instances explicitly crafted by an adversary to cause the model to make a mistake (Szegedy et al., 2014). Most recent work focuses on sim-

---

[1] https://github.com/joestacey/robust-nli

ple semantic-invariant transformations, showing that neural models can be overly sensitive to small modifications of the inputs and paraphrasing. For instance, Ribeiro et al. (2018) use a set of simple syntactic changes, such as replacing *What is* with *What's*. Other semantics-preserving perturbations include typos (Hosseini et al., 2017), the addition of distracting sentences (Wang and Bansal, 2018; Jia and Liang, 2017), character-level perturbations (Ebrahimi et al., 2018), and paraphrasing (Iyyer et al., 2018). Minervini and Riedel (2018) propose searching for violations of constraints, such as the symmetry of *contradiction* and transitivity of *entailment*, for identifying where NLI models make mistakes and then creating more robust models by training on these adversarial examples. Alternatively, Clark et al. (2019), He et al. (2019) and Mahabadi et al. (2020) create naive models that make predictions based on known dataset biases, and then train robust models in an ensemble with the naive models to focus on other patterns in the data that generalise better.

Adversarial Training Another procedure for creating more robust models is through adversarial training with latent representations, with a classifier trained to learn the bias from the model sentence representations which in turn update to reduce the performance of the bias classifier (Wang et al., 2019). For example, Ganin and Lempitsky (2015) use adversarial training to improve domain adaption, allowing models to learn features helpful for the model task but which are also invariant with respect to changes in the domain. This was achieved by jointly training two models, one to predict the class label and one to predict the domain, and then regularising the former model to decrease the accuracy of the latter via gradient reversal.

Belinkov et al. (2019b) use adversarial training to remove the hypothesis-only bias from models trained on SNLI. While this approach produced models that generalised better to other datasets, these same models show degraded performance on SNLI-hard (Belinkov et al., 2019a), which is supposedly the ideal dataset to test for generalisation as it resembles SNLI the most in terms of domain and style while lacking the examples with the largest bias (Gururangan et al., 2018). Moreover, the bias is not removed from the model sentence representations and can be almost fully recovered if these representations stop updating (Belinkov et al., 2019b). It is therefore unclear whether any improvements

are caused by de-biasing, or are instead a result of perturbations from the adversarial training procedure. Similarly, Elazar and Goldberg (2018) find that adversarial training and gradient reversal does not remove demographic information such as age or gender, with this information still present in the de-biased sentence representations. Here, we propose using an ensemble of multiple adversaries to avoid the hypothesis-only bias, significantly reducing the bias stored within a model's representations and outperforming previous de-biasing research efforts when testing model generalisation.

Using Model Ensembles Grefenstette et al. (2018) find that using ensembles of models is a better use of computational budget when training from adversarial examples compared to using a larger model with more parameters. Elazar and Goldberg (2018) show that using an ensemble of up to 5 adversarial classifiers helped remove demographic information contained within Twitter messages, however beyond this they were not able to re-learn the main task. Mahabadi et al. (2020) provide further support for using model ensembles after implementing a Product of Experts approach with multiple hypothesis-only models, producing more robust NLI models despite not attempting to remove the bias from the underlying representations. We compare our results with Mahabadi et al. (2020) to quantify the benefits of removing the bias from the sentence representations.

3 Ensemble Adversarial Training

We follow an adversarial training approach for reducing the hypothesis-only bias contained within the sentence representations. Specifically, we generalise the adversarial training framework proposed by Belinkov et al. (2019a) to make use of *multiple adversaries*: n hypothesis-only adversaries are jointly trained for predicting the relationship between the premise and hypothesis given only the representation of the hypothesis from the sentence encoder. At the same time, the sentence encoder together with an hypothesis-premise model are jointly trained to fit the training data, while decreasing the accuracy of these adversaries. Formally, given a hypothesis $\mathbf{h}$ and a premise $\mathbf{p}$, the predictions of the hypothesis-premise model $\hat{y}$ and the i-th hypothesis-only adversary $\hat{y}_{a_i}$ can be for-

malised as follows:

$$\mathbf{e}_h = \text{encoder}_{\theta_e}(\mathbf{h}), \quad \mathbf{e}_h \in \mathbb{R}^k$$
$$\mathbf{e}_p = \text{encoder}_{\theta_e}(\mathbf{p}), \quad \mathbf{e}_p \in \mathbb{R}^k$$
$$\hat{y} = \text{MLP}_{\theta_c}([\mathbf{e}_h; \mathbf{e}_p; \mathbf{e}_h - \mathbf{e}_p; \mathbf{e}_h \odot \mathbf{e}_p])$$
$$\hat{y}_{a_i} = \text{MLP}_{\theta_{a_i}}(\mathbf{e}_h),$$

where $\hat{y}, \hat{y}_{a_i} \in \mathbb{R}^3$ are (unnormalised) score distributions over the three NLI classes, *i.e. entailment*, *contradiction*, and *neutral*, and $\theta_e, \theta_c, \theta_{a_i}$ respectively denote the parameters of the encoder, the hypothesis-premise model, and the i-th hypothesis-only adversary. The adversarial training procedure can be formalised as optimising the following *minimax objective*:

$$
\begin{aligned}
\min_{\theta_e, \theta_c} \max_{\theta_a} \sum_{\langle \mathbf{h}, \mathbf{p}, y \rangle \in \mathcal{D}} & (1 - \lambda)\mathcal{L}_{ce}(y, \hat{y}) \\
& - \frac{\lambda}{n} \sum_{i=1}^{n} \mathcal{L}_{ce}(y, \hat{y}_{a_i}),
\end{aligned}
\tag{1}
$$

where $\mathcal{D}$ is a dataset, and $\mathcal{L}_{ce}$ denotes the cross-entropy loss (Goodfellow et al., 2016), and $n \in \mathbb{N}_+$ is the number of adversaries. The hyperparameter $\lambda \in [0, 1]$ denotes the trade-off between the losses of the hypothesis-premise model and the hypothesis-only adversaries. Similarly to Belinkov et al. (2019a), we optimise the minimax objective in Eq. (1) using gradient reversal (Ganin and Lempitsky, 2015), which leads to an optimisation procedure equivalent to the popular gradient descent ascent algorithm (Lin et al., 2019). The gradient reversal multiplies the gradient from the adversarial classifiers by a negative constant, training the encoder to reduce the performance of these classifiers.

To test the impact of using multiple adversarial classifiers when changing the dimensionality of the representations, we train with $\{1, 5, 10, 20\}$ bias classifiers for $\{256, 512, 1024, 2048\}$ dimensional sentence representations. The learned sentence representation is then frozen, and 20 adversarial classifiers are randomly reinitialised before they attempt to re-learn the hypothesis-only bias from the frozen de-biased sentence representation. The maximum accuracy from across the 20 adversarial classifiers is then reported after trying to remove the bias, showing the maximum bias that can still be learnt from the representation.

The ability of adversarially trained models to de-bias sentence representations is

tested across a range of λ hyper-parameters $\{0.001, 0.01, 0.1, 0.2, 0.3, ...0.8, 0.9, 0.99, 0.999\}$. This shows whether any improvement is due to the choice of λ, or whether there is an improvement regardless.

Model Architecture Following the same experimental set-up as Belinkov et al. (2019a) and Poliak et al. (2018), we use an InferSent model (Conneau et al., 2017) with pretrained GloVe 300-dimensional word embeddings. The InferSent model architecture consists of a Long Short-Term Memory network (LSTM, Hochreiter and Schmidhuber, 1997) encoder which creates a 2048 dimensional sentence representation.

3.1 Significance Testing

We perform statistical testing to assess whether the differences between using one or five adversarial classifiers is significant. This involves repeating the experiments for both one and five adversarial classifiers with ten different random seeds. For each experiment, the de-biasing is performed before a classifier attempts to learn the bias again from the frozen sentence representations.

We use bootstrapping hypothesis testing (Efron and Tibshirani, 1993) to test the statistical significance by comparing the means from the two samples. We also provide p-values from a Mann Whitney U-test (Mann and Whitney, 1947). The bootstrapping considers the null hypothesis that there is no difference between the mean bias re-learnt from using five adversarial classifiers compared to just using one adversarial classifier. In addition, we use a Bonferroni correction factor (Shaffer, 1995) of four when evaluating the p-values, taking into account multiple hypothesis testing across each different dimension. P-values smaller than 0.05 are considered significant.

3.2 Using Deeper Adversaries

We also investigate using a multi-layer perceptron as a more complex adversarial classifier to understand whether the bias that can be re-learnt depends on the type of classifier used. The experiments are repeated using non-linear classifiers instead of linear classifiers, both during the adversarial training and also afterwards when the classifiers try to re-learn the biases from the frozen representations. In addition to testing 2048 dimensions using ten adversaries, a 512 dimensional representation is also tested using a smaller number of adversaries (five).
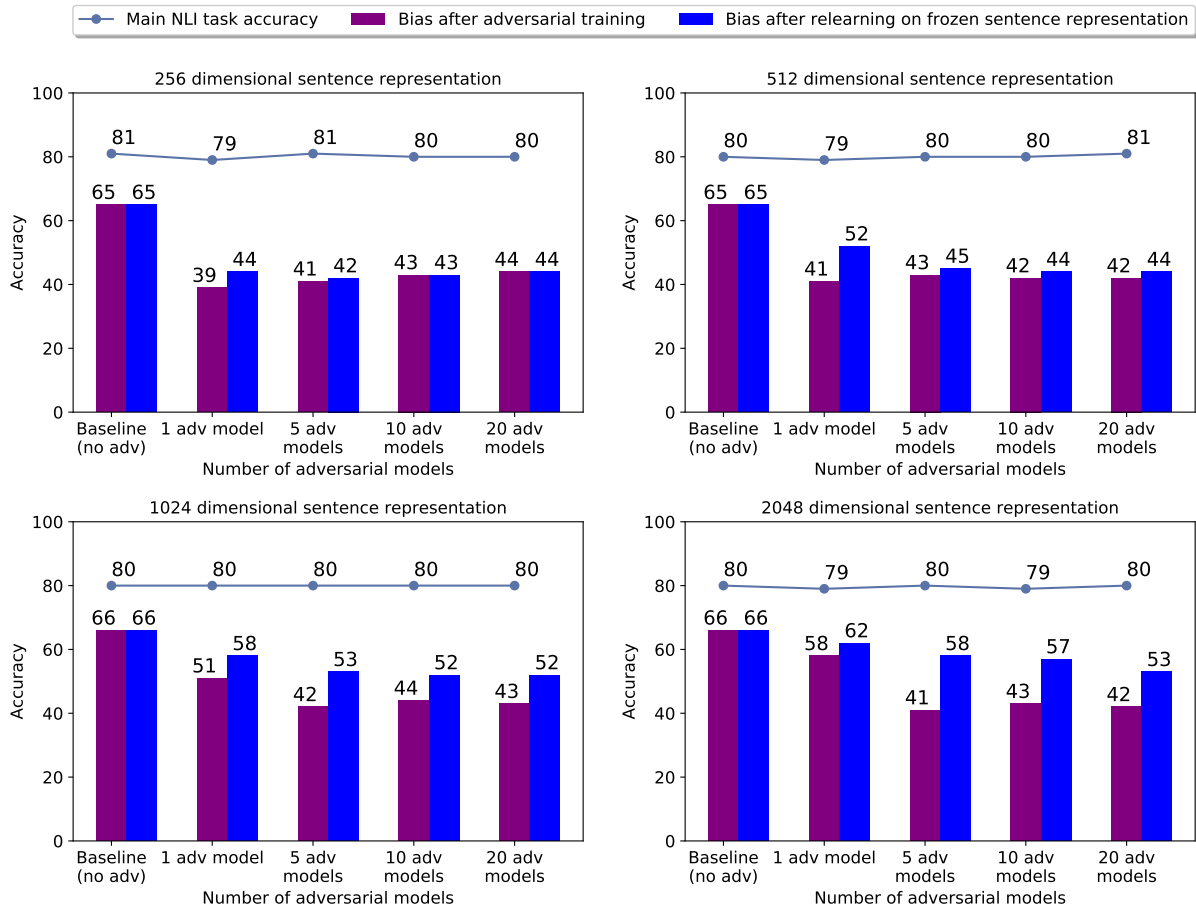
Figure 1: The maximum bias classifier accuracy after the bias is re-learnt from the frozen de-biased representations (in blue), compared to the accuracy of an independent bias classifier (not an adversary) at the end of the adversarial training (in purple). The main NLI task accuracy is also displayed.

We perform the experiments with three scenarios: *i)* Using the non-linear classifiers during the adversarial training, but not afterwards, and instead trying to re-learn the bias with a linear classifier. *ii)* Using linear classifiers during the adversarial training but then non-linear classifiers are used to try to re-learn the biases after the sentence representation is frozen. *iii)* Finally, non-linear classifiers are used both during adversarial training and afterwards when trying to re-learn the biases from the frozen sentence representation. The non-linear multi-layer perceptron classifier consists of three linear layers, and two non-linear layers using tanh.

3.3 Evaluating De-biased Sentence Encoders

After training the models on SNLI with adversarial training, we test these de-biased models on a range of different datasets to see whether they generalise better. The performance of the de-biased models is compared to a baseline model trained on SNLI where no adversarial training has been performed.

By using different random seeds, we compare ten baseline SNLI-trained models with models using one adversary and 20 adversaries, with each of these models tested on SNLI-hard. We perform bootstrap hypothesis testing to understand whether there is a significant difference between using one adversary and the baseline models with no adversarial training. We have repeated this hypothesis testing to compare models de-biased using 20 adversaries to the baseline models.

Additionally, we evaluate the de-biased models on 12 different datasets to understand whether models trained with an ensemble of adversaries perform better than the baseline and models trained with one adversary. The datasets in these experiments are the same datasets tested by Belinkov et al. (2019a): ADD-ONE-RTE (Pavlick and Callison-Burch, 2016), GLUE (Wang et al., 2018), JOCI (Zhang et al., 2017), MNLI (Williams et al., 2018), MPE (Lai et al., 2017), SCITAIL (Khot et al., 2018), SICK (Marelli et al., 2014), SNLI-

| | **Number of adversaries** | | | | |
Dimensions	**0**	**1**	**5**	**10**	**20**
256	65	44	**42**	43	44
512	65	52	45	**44**	**44**
1,028	66	58	53	**52**	52
2,048	66	62	58	57	**53**

Table 1: Maximum accuracy (%) from 20 bias classifiers when re-learning the hypothesis-only bias from the frozen de-biased sentence representation. The lowest accuracy figure for each dimension is highlighted, after testing 0, 1, 5, 10 and 20 adversaries.
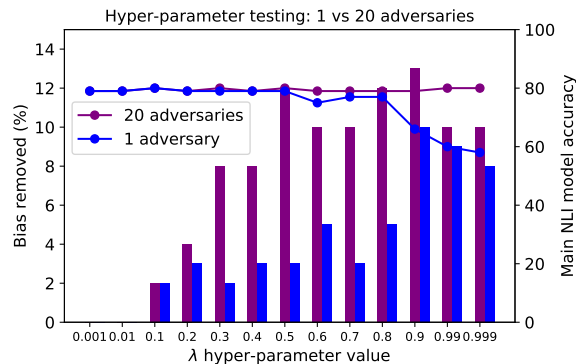

Figure 2: The fall in accuracy of hypothesis-only classifiers when using 1 or 20 adversaries to remove the hypothesis-only bias (compared to a baseline with no adversaries). This is shown alongside the overall accuracy of the de-biased NLI models.

hard (Gururangan et al., 2018), and three datasets recast by White et al. (2017): DPR (Rahman and Ng, 2012), FN+ (Pavlick et al., 2015) and SPR (Reisinger et al., 2015).

While the previous results in this paper use an LSTM encoder, a bidirectional LSTM has been used when testing other datasets to ensure the experiments are a like-for-like comparison with Belinkov et al. (2019a). We select the hyper-parameters that yield the highest accuracy on a validation set, in line with the experiments conducted by Belinkov et al. (2019a). Finally, the model results are compared to the Product of Experts (PoE) de-biasing approach proposed by Mahabadi et al. (2020).

4 Results

We use an ensemble of multiple adversarial classifiers during model training to understand whether it is possible to reduce the bias within the model sentence representations. Our results show that training a model with an ensemble of adversaries does reduce the model bias, doing so across each dimensionality of sentence representations tested. Moreover, more adversaries are required for de-biasing larger dimensional sentence representations.

When using just one adversarial classifier for a 2,048 dimensional sentence representation, after the representation was frozen and the bias classifiers had a chance to re-learn the hypothesis-only bias, the accuracy of the bias classifiers increased to 62% (see Table 1). This result mirrors the findings of Belinkov et al. (2019b), where using one adversarial classifier does not remove the bias from the model sentence representations. We additionally show that an independent bias classifier can still reach 58% accuracy at the end of the adversarial training, without needing to freeze the representation in order to relearn the bias (see Fig. 1).

For 2,048 dimensional sentence representations, as the number of adversaries are increased up to 20, less bias can be found in the resulting de-biased sentence representation. When the number of adversarial classifiers are increased from 1 to 20, the accuracy of the hypothesis-only bias classifiers reduces from 62% to 53% (see Table 1).

The higher the dimensionality of the sentence representation, the more difficult it is to remove its bias. Additionally, the optimal number of adversarial classifiers depends on the dimensionality of the representations, with more adversaries required for higher dimensions. For 2,048 dimensions this is 20 adversaries, while for 256 dimensions this reduces to 5 adversaries (see Table 1). For 256, 512 and 1,028 dimensions, the improvements plateau after a set number of adversaries, and therefore scaling up the number of adversaries beyond 20 is unlikely to lead to further improvements.

The improvements in de-biasing when using an ensemble of adversaries is consistent across different λ hyper-parameter values: when comparing a model trained with one adversary to a model trained with 20 adversaries, the model trained with 20 consistently removes more bias. Using a single adversary does reduce the bias for λ values of 0.9 and above, but at the expense of the overall model accuracy which reduces dramatically (Fig. 2). This suggests that valuable information about the hypothesis is being removed instead of just the bias. Interestingly, when using 20 adversaries, even for the largest values of λ, the overall model accuracy does not start to decrease.
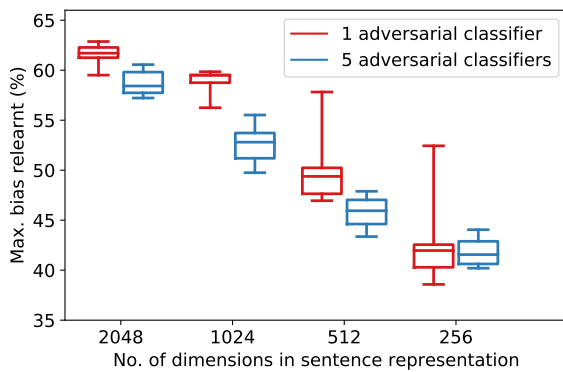
Figure 3: Maximum accuracy scores of the bias classifiers when they are retrained on de-biased sentence representations for each of the experiments tested. Ten experiments were performed for each condition, using one or five adversaries for each dimension.

4.1 Evaluating Multiple Adversaries

We applied statistical testing to understand whether the improvements seen using an ensemble of adversarial classifiers is statistically significant. For sentence representations with 2,048, 1,024 or 512 dimensions this is a statistically significant result. Although the results are not statistically significant for a smaller 256 dimensional representation.

For 2,048, 1,024 and 512 dimensional sentence representations, the statistical testing provides p-values smaller than 0.05. The null hypothesis is therefore rejected in these cases, with the alternative hypothesis stating that using five adversaries reduces the mean bias re-learnt from the sentence representations compared to using just one adversarial classifier (see Table 2). Fig. 3 displays these results in a boxplot diagram.

4.2 Using Deeper Adversaries

To investigate the impact of changing the strength of the adversary, multi-layer perceptrons are used during model training as the adversarial classifiers. The results show that more complex multi-layer perceptrons do not always perform better, and that the best choice of adversary depends on the type of classifier used to relearn the bias.

When a non-linear model is used to re-learn the bias from the frozen sentence representation, less bias can be recovered if a non-linear model was used as the adversarial classifier during training instead of a linear adversarial classifier (see Table 3). Therefore, when using a more complex classifier to re-learn the bias, a model of at least the same complexity should be used in the adversarial train-

ing to remove these biases. If a linear classifier is used as the adversary, a non-linear classifier can find more bias when learning from the de-biased representation than a linear classifier can.

The results also show that if a linear model is being used to re-learn the bias, then using a linear model as the adversary instead of a multi-layer perceptron reduces the amount of bias that can be recovered (see Table 3). This could suggest that the best approach is to use the same type of classifier for both the adversarial model and the classifier used to re-learn the bias. However, more adversarial classifiers may be required when using non-linear classifiers as adversaries, and therefore more experimentation is required to test this hypothesis.

The classifier chosen to relearn the bias will depend on the model that is being de-biased. If a classifier cannot relearn the bias from the sentence representations, this is a guarantee that a model using the de-biased representations and this classifier will not be influenced by the bias.

4.3 Evaluating De-biased Encoders

The models trained with an ensemble of adversaries are applied to 12 different NLI datasets to test whether these de-biased models generalise better than models trained with either one or no adversarial classifier. The datasets tested include SNLI-hard, where models that are no longer influenced by the hypothesis-only bias are expected to perform better. Models trained using an ensemble of adversaries performed better across most of these datasets, including SNLI-hard where there is a statistically significant improvement compared to a baseline model with no adversarial training.

Models trained with one adversary were not significantly better than a baseline model when tested on SNLI-hard (1.1% improvement, corresponding to a p-value of 0.07). On the other hand, there is a statistically significant improvement when using an ensemble of 20 adversarial classifiers (achieving a 1.6% improvement with a p-value of 0.015). As a result, we accept the alternative hypothesis that models trained with 20 adversaries have a higher mean accuracy than the baseline models.

Across 8 of the 13 datasets analysed, models trained with an ensemble of 20 adversarial classifiers performed better than when using only one adversarial classifier (see Table 4). For three of the remaining datasets, the performance was the same between using one adversary and 20 adversaries.

	P-Value		Accuracy of bias classifier	
Dim.	Mann-Whitney	Bootstrapping	Mean (Median), 1 Adv.	Mean (Median), 5 Adv.
256	1	1	42.4 (42.0)	41.8 (41.6)
512	0.0009*	<0.0001*	49.8 (49.4)	**45.8** (45.9)
1,028	0.0008*	<0.0001*	58.9 (59.5)	**52.6** (52.8)
2,048	0.0005*	0.0105*	61.6 (61.7)	**58.7** (58.4)

Table 2: p-values after performing Bootstrapping and Mann-Whitney hypothesis tests, using a Bonferroni correction factor of 4. * indicates a statistically significant result with a p-value below 0.05. Highlighted values indicate that the mean is significantly smaller than its comparison mean value, using the bootstrapping p-values.

Adversary type	Class. type	Accuracy
2,048 dim (10 adv)		
Linear	Linear	56
Linear	Non-linear	66
Non-Linear	Linear	61
Non-Linear	Non-linear	62
512 dim (5 adv)		
Linear	Linear	44
Linear	Non-linear	66
Non-Linear	Linear	55
Non-Linear	Non-linear	60

Table 3: Accuracy of bias classifiers when relearning the bias after using either a linear or non-linear adversary, when the classifier used to re-learn the bias (*class. type*) is also either linear or nonlinear.

The performance when using an ensemble of adversaries was on average 0.9 points higher than when using one adversary, which in turn outperformed the baseline by 1.6 points. The ensemble of adversaries also outperforms the Product of Experts approach proposed by Mahabadi et al. (2020).

4.4 Discussion

We find that the higher the dimensionality of the representations, the less effective a single adversary is at removing the bias, with the de-biasing also dependant on the strength of the classifier. These differences explain why past research has found that biases can remain hidden within the model representations (Elazar and Goldberg, 2018; Belinkov et al., 2019b), with Elazar and Goldberg (2018) using a model with a non-linear classifier while Belinkov et al. (2019b) use a single adversary with 2,048 dimensional sentence representations.

Dataset	Baseline	PoE	1 Adv.	Ens.
AOR	61.24	+1.8	-2.3	+1.3
DPR	46.30	+0	+2.9	+0.9
FN+	38.43	+0.7	+6.8	+12.2
GLUE	43.12	+0.4	+0.6	-1.0
JOCI	40.77	+1.1	+0.8	+1.6
MNLI Match.	53.38	-0.3	+0.8	+0.8
MNLI Mism.	52.91	+0.8	-0.5	-0.1
MPE	57.30	+1.2	-0.4	+0.5
SCIT	47.98	-1.1	+0.5	+0.5
SICK	50.61	+1.7	0.4	-0.4
SNLI Hard	65.72	+4.2	+1.2	+1.7
SNLI	83.29	-3.9	+0.2	+0.8
SPRL	30.35	+9.2	+10.3	+13.5
Average		+1.2	+1.6	**+2.5**

Table 4: Accuracy of the de-biased models when tested on 12 different NLI datasets, comparing models trained with one adversary (1 Adv.) as per Belinkov et al. (2019a), to models trained with an ensemble of adversaries (Ens.). The Product of Experts (PoE) approach proposed by Mahabadi et al. (2020) is also included.

5 Conclusions

We set out to prevent NLI models learning from the hypothesis-only bias by using an ensemble of adversarial classifiers. Our method produced sentence representations with significantly less bias, and these more robust models generalised better to 12 different NLI datasets, improving over previous approaches in the literature (Belinkov et al., 2019b; Mahabadi et al., 2020).

The higher the dimensionality of the sentence representations, the harder it is to de-bias these representations and the higher the optimal number of adversarial classifiers appears to be. Furthermore, the models trained with an ensemble of adversaries also performed better when tested on SNLI-hard compared to using only one adversarial classifier.

This is the behaviour expected from de-biased models that no longer use the hypothesis-only bias to inform their predictions.

By preventing a linear classifier from learning the bias from the de-biased representations, we conclusively show that a model using such a classifier with these representations will not make decisions based on the bias. However, after implementing the adversarial training, a non-linear classifier may still be able to detect the bias in the sentence representations where linear classifiers are not able to. While we illustrate the conditions under which biases are removed from a linear classifier, preventing a non-linear classifier from learning the biases is more difficult and merits further experimentation.

Acknowledgements

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, pages 4971–4980. IEEE Computer Society.

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander M. Rush. 2019a. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *ACL (1)*, pages 877–891. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander M. Rush. 2019b. On adversarial removal of hypothesis-only bias in natural language inference. In **SEM@NAACL-HLT*, pages 256–262. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642. The Association for Computational Linguistics.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending: Strong neural baselines for the ROC story cloze task. In *ACL (2)*, pages 616–622. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *EMNLP-IJCNLP*, pages 4060–4073. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, pages 670–680. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *ACL (2)*, pages 31–36. Association for Computational Linguistics.

Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Springer.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *EMNLP*, pages 11–21. Association for Computational Linguistics.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. 2016. *Deep Learning*. Adaptive computation and machine learning. MIT Press.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6325–6334. IEEE Computer Society.

Edward Grefenstette, Robert Stanforth, Brendan O'Donoghue, Jonathan Uesato, Grzegorz Swirszcz, and Pushmeet Kohli. 2018. Strength in numbers: Trading-off robustness and compuputation via adversarially-trained ensembles.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT (2)*, pages 107–112. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *DeepLo@EMNLP-IJCNLP*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *ICLR*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hossein Hosseini, Baicen Xiao, and Radha Poovendran. 2017. Deceiving google's cloud video intelligence API built for summarizing videos. In *CVPR Workshops*, pages 1305–1309. IEEE Computer Society.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *NAACL-HLT*, pages 1875–1885. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*, pages 2021–2031. Association for Computational Linguistics.

Kushal Kafle and Christopher Kanan. 2016. Answer-type prediction for visual question answering. In *CVPR*, pages 4976–4984. IEEE Computer Society.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In *EMNLP*, pages 5010–5015. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5189–5197. AAAI Press.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 100–109. Asian Federation of Natural Language Processing.

Tianyi Lin, Chi Jin, and Michael I. Jordan. 2019. On gradient descent ascent for nonconvex-concave minimax problems. *CoRR*, abs/1906.00331.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. In *The annals of Mathematical Statisitcs*, pages 50–60. Institute of Mathematical Statistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223. European Language Resources Association (ELRA).

Pasquale Minervini and Sebastian Riedel. 2018. Adversarially regularising neural NLI models to integrate logical background knowledge. In *CoNLL*, pages 65–74. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *HLT-NAACL*, pages 839–849. The Association for Computational Linguistics.

Ellie Pavlick and Chris Callison-Burch. 2016. 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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. 2015. 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 408–413. The Association for Computer Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In **SEM@NAACL-HLT*, pages 180–191. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. 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, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 777–789. ACL.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Trans. Assoc. Comput. Linguistics*, 3:475–488.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *ACL (1)*, pages 856–865. Association for Computational Linguistics.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *CoNLL*, pages 15–25. Association for Computational Linguistics.

Juliet Popper Shaffer. 1995. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–581.

Rion Snow, Lucy Vanderwende, and Arul Menezes. 2006. Effectively using syntax for recognizing false entailment. In *HLT-NAACL*. The Association for Computational Linguistics.

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

Aarne Talman and Stergios Chatzikyriakidis. 2018. Neural network models for natural language inference fail to capture the semantics of inference. *CoRR*, abs/1810.09774.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *LREC*. European Language Resources Association (ELRA).

Lucy Vanderwende and William B. Dolan. 2005. What syntax can contribute in the entailment task. In *MLCW*, volume 3944 of *Lecture Notes in Computer Science*, pages 205–216. Springer.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 353–355. Association for Computational Linguistics.

William Yang Wang, Sameer Singh, and Jiwei Li. 2019. Deep adversarial learning for NLP. In *NAACL-HLT (Tutorial Abstracts)*, pages 1–5. Association for Computational Linguistics.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In *NAACL-HLT (2)*, pages 575–581. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *ICLR (Poster)*.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. 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)*, pages 996–1005. Asian Federation of Natural Language Processing.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, pages 1112–1122. Association for Computational Linguistics.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *CVPR*, pages 5014–5022. IEEE Computer Society.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal commonsense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.