CS378: Final Project - Data Artifacts
https://github.com/samtheant/nlp_final_project

Samantha Liu
scl2332

April 2023

Abstract

Dataset artifacts are a ubiquitous problem in modern natural language processing tasks. These artifacts are introduced due to sampling bias or human bias during the generation of the dataset and can be difficult to remove. Models trained on these biased datasets may appear to be solving the end task (e.g. NLI or QA) effectively, but are in reality relying on spurious correlations in the dataset to accomplish the task. One method of removing model bias, DRiFT (Debias by Residual Fitting) was shown to successfully result in greatly improved performance accuracy by one metric (the word-overlap-based challenge dataset HANS), but not much by another metric (STRESS, a set of stress tests focusing on negations and word overlap). Furthermore, the method showed mixed results for different model architectures. Given these mixed results, it's unclear whether DRiFT can reliably remove a variety of types of dataset biases. In this paper, I evaluate the effectiveness of DRiFT on ELECTRA-small on the NLI task using a different type of challenge set - a contrast set. I find that a debiased model using DriFT can achieve a small accuracy gain on the contrast sets without degrading the performance on the original test set.

1 Introduction (5pt)

NLI (Natural Language Inference) is a natural language processing task where a classifier is given two statements, a premise and a hypothesis, and the classifier must predict whether the relation between the statements is an entailment, neutral, or a contradiction. The Stanford Natural Language Inference (SNLI) dataset is one popular dataset for training and evaluating this task, but it has been shown to contain certain biases due to the way it was collected. DRiFT is a proposed algorithm for debiasing a model from such dataset artifacts. The algorithm works by first training a biased model, and then fitting a second model to the residual of the biased model. Notably, the method can be used without having to alter or augment the original dataset, making it especially promising for use on large datasets where it would be expensive to try to augment the data by e.g. hand writing contrast examples into the training set. My task is to use the DRiFT method to train an ELECTRA-small model for NLI on the SNLI dataset, and to evaluate it. For the evaluation method, I choose to use contrast sets, as outlined by Gardner et al 2020. A contrast set is created by taking a sample from the original dataset and perturbing it slightly in a way that changes its gold label. Using contrast sets is a natural fit for NLI since a different label can be easily obtained by a small alteration to the premise or hypothesis. Furthermore, the original paper introducing contrast sets did not test NLI as one of their tasks for evaluation using contrast sets, so it will also be informative to see how well contrast sets expose bias in NLI tasks. For this end, I use a hand-written contrast set which uses perturbed examples from the original SNLI test set.

First, I trained an ELECTRA-small model on the SNLI dataset and evaluated its accuracy on the SNLI test set and my handwritten contrast set, respectively. The accuracy on the SNLI test set was 89.8% while the accuracy on my handwritten contrast set dropped to 51.6%. Further inspection showed that the model tended to confuse the labels in the contrast set with the labels of the original sample that it was perturbed from.

Next, I trained two biased models, one which only had access to the hypothesis, and an even further limited model which only had access to the sequence lengths of the premise and hypothesis. I then trained a new 'debiased' ELECTRA-small model fitting the residual of these two mod-

els, as well as one fitting the sum of the residuals of both models. The new model that was fit to the hypothesis-only model achieved an accuracy of 54.0% on my contrast set, the one that was fit to the sequence-length only model had an accuracy of 50.8% and the one that was fit to both achieved an accuracy of 54.8%. Thus, debiasing using the hypothesis-only model showed a small improvement over the baseline of 51.6%, and debiasing using the combined hypothesis-only and sequence-length only models showed a small increase on top of that. This result suggests that debiasing with DRiFT using an ensemble of biased models may improve the efficacy of DRiFT, but further tests would be needed to confirm this result.

2 Task/Dataset/Model Description (15pt)

As outlined in He et al. 2019, the task of separating bias from data can be formalized as representing examples $x \in X$, the inputs, as two separate feature components, a biased component $b(x)$ and a "true" component $g(x)$, such that $p(y|g(x)) = q(y|g(x))$ where P is the biased training distribution and Q is the test distribution, and $y \in Y$ is the outputs. Then the joint distribution $p(x, y)$ can be written like so:

$$\begin{aligned} p(x, y) &= p(b(x), g(x), y) \\ &= p(g(x)|y)p(y|b(x))p(b(x)) \end{aligned} \tag{1}$$

To train a biased model, I choose insufficient features $I(x)$ which approximate biased features $b(x)$ given prior knowledge about what types of biases are in SNLI.

Let f_s represent the biased model and f_d represent the debiased model, and let L be the loss function, which is cross entropy loss.

$$L(\hat{y}, y) = -\sum_{i=1}^{n} y_i \log \hat{y}_i \tag{2}$$

First learn f_s which has the following as the training objective:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_P \left[L(f_s(I(x); \theta), y) \right] \tag{3}$$

Then the task is to train the debiased model f_d. The weights θ^* are frozen and I train f_d with the following training objective:

$$\phi^* = \arg\min_{\phi} \mathbb{E}_P \left[L(f_s(I(x); \theta^*) + f_d(x; \phi), y) \right] \tag{4}$$

Table 1: Accuracy on the SNLI and Contrast sets

Majority	SNLI Test Set	Contrast
34.2	89.8	51.6

Table 2: Accuracy of the baseline model on the SNLI and contrast sets by label

Label	SNLI Test Set	Contrast
Entailment	91.1	56.4
Neutral	87.3	44.4
Contradiction	91.0	54.8

Then at test time I ignore the biased component and just use the output $f_d(x; \phi)$.

The model used for the baseline comparison, debiased, and hypothesis-only models is a pretrained ELECTRA-small architecture fine-tuned on the SNLI training set which contains 570k training examples. The sequence-length-only model is a simple feedforward neural network with 2 layers and a hidden layer size of 50, using ReLU as the activation function. All models were trained for 3 epochs on the SNLI training set (or features derived from the training set).

3 Performance Analysis (25pt)

The SNLI dataset was collected by first sourcing premise statements from captions of pre-existing photo datasets, then having crowdworkers write three hypotheses – one entailment, neutral, and contradiction – for each premise. Suchin et al. 2018 showed that this protocol makes it possible to identify the correct label based on the hypothesis only, since crowdworkers use common strategies to produce hypotheses (for example, using the word 'not' to create a contradiction). It's also common for entailments to be generated by removing information from the premise (e.g. "Three dogs are running in a park" becomes "Some dogs are running"), and for neutral hypotheses to be generated by adding information; thus, some information about the label can be gleaned from the relative sentence lengths of the premises and hypotheses. Thus, to evaluate the bias in this dataset, I constructed a contrast set.

To construct my contrast set, I chose 21 examples from the SNLI test set. Then for each label other than the original label, I perturbed the example 3 ways to get that label: one perturbing the premise only, one perturbing the hypothesis only, and one perturbing both. This gives me 21

examples × 2 labels each × 3 perturbations each, giving 126 examples for my contrast set. An example of how the perturbations is done is shown in table 2 and 3.

The accuracy of the baseline ELECTRA-small model on the SNLI test set was 89.8% while the accuracy on my handwritten contrast set dropped to 51.6%, as seen in table 1.

From table 2, we can see that the model performs worst at neutral examples across the board for both the SNLI test set and the contrast set.

The model was likely to classify the contrast examples as what their label was originally. For example, one original example was "Premise: This church choir sings to the masses as they sing joyous songs from the book at a church. Hypothesis: The church has cracks in the ceiling." and the original label was Neutral. For 4 out of the 6 perturbed versions of this example, the model classified them as Neutral as well (for example: "Premise: This church choir sings to the masses as they sing joyous songs from the book at a church. Hypothesis: The church has sound waves hitting the ceiling." was classified as Neutral though the correct label is Entailment.).

The model also incorrectly classified examples where the hypothesis or the premise had the word "not" but was not a contradiction. For example: "Premise: This church choir sings to the masses as they sing joyous songs from the book at a church. Hypothesis: A choir singing not at a baseball game." and "Premise: A land rover is not being driven across a river. Hypothesis: A Land Rover is splashing water as it crosses a lake." were classified as contradictions, even though the first is an entailment and the second is neutral. Though the word "not" appears in only about .8% of training examples, the model seems to have caught onto it.

The model also incorrectly classified examples where the hypothesis was short as entailments. For example: "Premise: An old man has a package in front of an advertisement. Hypothesis: A man sits in front of an ad." was labeled as an entailment.

The model also struggled with uses of the word "or" in the contrast set as the word "or" only appears in about .9% of training examples, and in the majority of the case it was in a context like "his or her" or "someone or something" and almost never in the logical sense of "or". For example the model was unable to classify the example "premise: A man playing an electric guitar on stage. Hypothesis: A man is performing for cash or not for cash. " correctly as entailment, instead classifying it as neutral. This seems to be an artifact of the data collection process being from image captions, which would rarely involve a logical "or".

A hypothesis-only model was able to get up to 70.9% accuracy on the SNLI test set, and only got 33% accuracy on the contrast set, which is about random. A sequence-length-only model was able to get 41.6% accuracy on the SNLI test set and only got 33% accuracy on the contrast set as well. Since these two biased models performed at the level of exactly random chance on the contrast set, this means that the contrast set is successfully rid of any information that could be gleaned from the hypothesis only or from the sequence lengths only.

4 Describing Your Fix (20pt)

After evaluating a baseline ELECTRA-small model, I applied the DRiFT algorithm to train several different debiased models.

First, I trained a biased model that has access to the hypothesis only, and a biased model that has access to the sequence lengths of the premise and hypothesis only. The hypothesis-only model was also an ELECTRA-small architecture, and trained for 3 epochs on the SNLI training set of 570k examples with a learning rate of 5e-5 and a batch size of 8.

The sequence length only model was a simple feedforward network with 2 layers and a hidden layer size of 50, using ReLU as the activation function, and was trained using the same hyperparameters as the hypothesis-only model. This model didn't need to be large or complex since it was just working with 2 features (premise length and hypothesis length).

Then, to train the debiased models, I created a training setup where I took the output logits from a pretrained biased model and the output logits from a new debiased model and added them together to get the two models' joint prediction. Then I calculated the cross entropy loss using the two models' combined prediction, and backpropagated through the new debiased only. I again trained this new model on the same training set with the same hyperparameters.

I followed this training setup using the hypothesis-only and sequence-length-only models to get two different debiased models. I also wanted to see what would happen if I ensembled the two biased models by adding the output logits of all three models to calculate the loss, and trained a

Table 3: Original Example

Original Premise	Original Hypothesis	Original Label
A woman with a green headscarf, blue shirt and a very big grin.	The woman has been shot.	Contradiction

Table 4: Perturbed Examples

New Premise	New Hypothesis	New Label
A woman with a green headscarf, blue shirt and a very big bullet wound.	The woman has been shot.	Entailment
A woman with a green headscarf, blue shirt and a very big grin.	The woman has not been shot.	Entailment
A woman with a green headscarf, blue shirt and a very big knife wound.	The woman has been stabbed.	Entailment
A woman with a green headscarf, blue shirt and a very big wound.	The woman has been shot.	Neutral
A woman with a green headscarf, blue shirt and a very big grin.	The woman has been shooting	Neutral
A woman with a green bandana, blue shirt and a very big bruise.	The woman has been punched.	Neutral

third debiased model that was trained to fit the residual of the two biased models combined.

5 Evaluating Your Fix (25pt)

Table 4 shows all the model performances on my handwritten contrast set. Overall, the performance improvement was pretty small. The debiased with hypothesis-only model had an accuracy of 54.0% which is a 2.4% accuracy improvement over the baseline model. The sequence-length debiased model did not do as well as the baseline, but the combined model did better than both with an accuracy of 54.8%. As seen in Table 5, the debiasing technique did not degrade the model performance on the original test set by very much for any of the models.

In He et al. 2019, the original paper introducing the DRiFT algorithm, they tested their debiasing method using challenge sets HANS and STRESS which were constructed to artificially test known biases, particularly word overlap and negations. However, a contrast set is different since it begins with examples from the original sampling distribution and perturbs them as little as possible to get a different label; thus, the examples are syntactically extremely close to being a different label, while also being closer to a natural sentence one might see in the wild (as opposed to artificially adding phrases like "true or not true" into the hypotheses, for example as is the case with STRESS.) This might result in examples that are very subtle, which makes contrast sets a particularly challenging test set, so I am not that surprised that the improvements were relatively small.

Table 6 shows the model performances on the contrast test set by label. Interestingly, while the hypothesis-only and sequence-length debiased models alone both decreased performance on neutral examples, when they were combined they improved performance on neutral examples. All of the models improved on contradiction examples, by almost 10% in the case of the model that was debiased with both. Since the combined-debiased model did better on the contrast set than the other debiased models, this suggests that ensembling biased models for use with DRiFT might be a promising approach.

Inspecting closer, I wanted to investigate how the debiased models did compared to the baseline on specific cases such as the inclusion of the word "not" in the hypothesis. For example,"Premise: This church choir sings to the masses as they sing joyous songs from the book at a church. Hypothesis: A choir singing not at a baseball game." was classified correctly as entailment by the hypothesis-only-debiased model and the combined-debiased model, improving over the baseline model which incorrectly classified it as a contradiction. The sequence-length-only model did not classify it correctly, which is to be expected since it didn't even have access to any of the lexical information in the sentences. The debiased models also seemed to do better on some

Table 5: Model Performances on Contrast Set

Majority	Baseline	Debiased With Hypothesis-only	Debiased With Sequence Length	Debiased With Both
33.3	51.6	54.0	50.8	54.8

Table 6: Model Performances on Original SNLI Test Set

Majority	Baseline	Debiased With Hypothesis-only	Debiased With Sequence Length	Debiased With Both
34.2	89.8	89.8	89.6	89.1

cases that had a lot of word overlap. For example, "Premise: A statue at a museum that seems to be looking at no one. Hypothesis: There is a statue that many people seem to think is looking at them." was falsely classified as an entailment by the baseline model, but the hypothesis-only-debiased model and combined-debiased model classified it correctly as a contradiction.

6 Related Work (5pt)

7 Conclusion (5pt)

Overall, the DRiFT approach for NLI as evaluated by contrast sets showed only small improvements. I hypothesize that the contrast set is quite far from the training distribution and that there is information in the contrast set that is just not present in the training set. You can remove known biases from a model, but DRiFT cannot help if parts of the problem space are greatly underrepresented just not present in the training set. In some sense, DRiFT is a band-aid that can remove some biases from a model, but if some part of the problem space is not in the training set, the only way to fix that is to get more varied data to train on. On the more optimistic side, since the combined debiased model did better on the contrast set, this suggests that ensembling biased models for use with DRiFT might be a promising approach, as would adding a larger number of biased features for the biased model to be trained on.

(Optional) AI Assistance

I used OpenAI's ChatGPT to help me format equations and tables in this document. I also used it help me write a helper method that removes the premise from a tokenized input tensor.

Table 7: Model Performances on Contrast Test Set By Label

Label	Baseline	Debiased With Hypothesis-only	Debiased With Sequence Length	Debiased With Both
Entailment	56.4	61.5	53.8	51.3
Neutral	44.4	42.0	40.0	48.9
Contradiction	54.8	59.5	59.5	64.3