A Generative Approach for Mitigating Hypothesis-Only Biases in Natural Language Inference

Dimion Asael
Technion – Israel Institute of Technology
dimion@cs.technion.ac.il

Yonatan Belinkov
Technion – Israel Institute of Technology
belinkov@technion.ac.il

Abstract

Natural Language Inference (NLI) is one of the key tasks in Natural Language Processing (NLP), where a model is required to predict the relationship between two sentences, a premise, and a hypothesis. Recent studies point out that much of the NLI datasets contained biases that allowed the models to perform well by only seeing the hypothesis, without learning how the sentences relate to each other. We propose a generative model, that generates the premise given the hypothesis and the label. We find that this approach leads to unbiased models for out-of-distribution data. However, we have found that generative models are difficult to train and that they perform worse than discriminative models in general.

1 Introduction

Natural Language Inference (NLI) has been widely studied in Natural Language Processing (NLP). In this task, the goal is to predict the relationship between two sentences, namely the premise (P) and the hypothesis (H). The predictions have been shown (Gururangan et al., 2018; Poliak et al., 2018) to be influenced by biases in the hypotheses, making it possible for a model to perform the task by only accessing the hypothesis, without truly learning the relationship between the premise and hypothesis.

Previous work (Belinkov et al., 2019) has proposed to use a generative approach to model the probability of the premise given the hypothesis and label, $p(P|y, H)$, and thus encourage a model to take the premise into account. However, that work approximated this likelihood by using other methods and did not train a generative model directly. As opposed to this, we suggest training a generative model. Our main finding is that it is possible to achieve an unbiased model with the generative approach. However, this approach yields poor results compared to the discriminative one. We discuss why and when each approach is better, and demonstrate that the generative model's inferior performance is due to the inherent difficulty of the generation task.

2 Hypothesis-only bias in NLI

For the NLI task, common datasets (Bowman et al., 2015; Williams et al., 2018) are typically created by humans, which comes with certain biases. A prominent kind of bias is hypothesis-only bias, which appears as words or phrases in the hypothesis that are associated with one of the labels. These biases make the relationship prediction easier.

Table 1 shows an example of a common bias, where the word "sleeping" is associated with contradiction in the sentence *"A woman is sleeping in an elevator"*, and guess that the relationship will be a contradiction, without ever taking a look at the premise.[1] These

[1]The Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) is based on image captions data, where captions are treated as premises and annotators were asked to generate a hypothesis. Captions usually describe actions, and so an easy way to create a contradiction is to use the word "sleeping" (Poliak et al., 2018).

Table 1: An example from SNLI. The relationship between the sentences is *contradiction*.

Premise	A woman in an office making a phone call
Hypothesis	A woman is **sleeping** in an elevator

hypothesis-only biases enable models to succeed in the task without learning the true relationship between the two sentences. If a model only has access to the hypothesis (that is, a hypothesis-only classifier), then we would expect the accuracy on the NLI task to be the same as a random guess, since such a model should not be able to understand the relationship between the sentences. As we can see from the test accuracy results in table 2 (which verifies the results from Poliak et al. 2018), this is not the case, and a hypothesis-only classifier can achieve an accuracy of around 70%. Those hypothesis-only biases can disrupt the learning process by leading the model prediction to be based on those biases rather than on the relation between the sentences. The table also shows the results on the hard test set (Gururangan et al., 2018), which is a subset of the original test set that is thought to contain fewer of the hypothesis-only biases. We can see that the hypothesis-only model is failing on the hard test set, since it does not have hypothesis-only biases to exploit. An unbiased model should perform equally well on both test sets, showing that it does not make predictions using the hypothesis-only biases.

Table 2: Discriminative baseline results. MNLI hard was created like SNLI hard in Gururangan et al. (2018) using MultiNLI Williams et al. (2018).

Model configuration	**SNLI**	**SNLI Hard**	**MNLI**	**MNLI Hard**
BART	91.49	83.16	85.92	76.6
BART, Hypothesis-only	70.61	33.42	58.88	20.49
Bert	90.53	80.89	84.76	74.6
Bert, Hypothesis-only	71.01	30.97	59.99	17.12

3 Generating instead of discriminating

Discriminative NLI models estimate $p(y|P, H)$. This allows bias from H to creep into the model. Instead, we rewrite the probability using Bayes' rule, as in equation 1, such that we essentially turn our discriminative objective into a generative one.

$$\arg\max_{y \in \mathcal{Y}} p(y|P, H) = \arg\max_{y \in \mathcal{Y}} \frac{p(P|y, H)p(y|H)}{p(P|H)} = \arg\max_{y \in \mathcal{Y}} p(P|y, H)p(y|H) \tag{1}$$

This generative model may still be biased, as it depends on the term $p(y|H)$, which is our hypothesis-only classifier from table 2. To overcome this issue, we set the distribution of the hypothesis-only objective to be uniform, meaning $\forall y \in \mathcal{Y}, p(y|H) = \frac{1}{count(\mathcal{Y})}$. Our motivation to do so is that if the data was truly unbiased, the probability of each label given only the hypothesis was the same for all of the labels, meaning $p(y|H) \sim Uniform(count(\mathcal{Y}))$. Using this observation, our generative objective will be $\arg\max_{y \in \mathcal{Y}} p(P|y, H)p(y|H) = \arg\max_{y \in \mathcal{Y}} p(P|y, H)$, which should result in an unbiased model.

We estimate $p(P|y, H)$ using an encoder-decoder model, where we encode (y, H) and decode P. To condition on y, we prefix a label-specific token to H.[2] We train the model by passing each hypothesis with its label through the encoder and calculating the cross-entropy loss between our model's predictions of the premise's tokens and the target premise in an auto-regressive manner. At test time, we attached each label-specific token to each hypothesis

[2]Sennrich et al. (2016) used a similar approach to control politeness in neural machine translation.

Table 3: Generative model results. A2B means that model A was used as the encoder and model B as the decoder.

Model configuration	SNLI	SNLI Hard	MNLI	MNLI Hard
BART	73.79	74.55	64.09	65.74
Bert2Bert	65.53	66.18	58.55	57.33
Bert2GPT2	62.99	63.17	-	-

and estimated the likelihood to generate the target premise.[3] The selected label is the one that maximizes this likelihood.

3.1 Experiments and Results

For our experiments, we used the encoder-decoder architecture that is available from HuggingFace (Wolf et al., 2019). Some of the models are pre-trained end-to-end, such as BART (Lewis et al., 2020), and some are only pre-trained as standalone decoders or encoders, such as BERT (Devlin et al., 2019) and GPT2 (Radford et al., 2019). Our method does produce an unbiased model based on the results in table 3, which show nearly identical accuracy on both the test and the "hard" test sets. Unfortunately, the results are much worse than the simple discriminative model (see table 2). In the following section, we will discuss possible reasons for the inferior performance of the generative classifier.

4 The challenges of generative training

As explained in section 3, we have changed our modeling from discriminative ($p(y|P,H)$) to generative ($p(P|y,H)$). As a result, our output space changed from the space of all sentences (infinite) to the space of all labels (small, finite). This output space is much more complex. Indeed, generative models are typically trained on very large datasets (that usually contain millions of training examples), orders of magnitude larger than the average NLI datasets (around 500,000 training example SNLI).

We observe that premises are usually longer than the hypotheses (14.1 tokens and 8.3 on average in SNLI, respectively[4]). Thus, we can assume that premises contain some sort of narrative, while hypotheses are shorter and without as many unnecessary words. This forces the generative model to model words that may not be relevant for the classification decision in the NLI task. Table 4 displays the generated premises resulting from three different hypotheses (that were originally associated with the same premise), each time conditioning on three different labels. As also evident from the third hypothesis, the model tries to impose the desired relation at the beginning of each sentence, and then fills the rest of the sentence with a description that will not affect the relationship. These words may carry significant

Table 4: Generated premises. Each hypothesis was originally in the SNLI dataset with the premise *"A woman with a green headscarf, blue shirt and a very big grin"*. Bold labels are the ground truth for the hypotheses w.r.t the original premise.

Label	Hypothesis	Generated premise
contradiction	the woman has been shot	a woman in a black shirt is sitting on a bench with a bag in her lap
entailment		a woman is being shot by a man in a blue shirt
neutral		a woman in a blue shirt is sitting on a bench with a bag in her lap
contradiction	the woman is very happy	a woman in a black shirt is smiling
entailment		a woman in a white shirt is smiling
neutral		a woman in a white shirt is smiling
contradiction	the woman is young	an elderly woman is sitting on a bench with her legs crossed and her eyes closed
entailment		a young woman in a black shirt and jeans is walking down the street
neutral		a woman in a red shirt is sitting on a bench with a bag in her lap

weight in the generative model's final prediction, where the probability to generate the

[3]Pasunuru & Bansal (2017) created a model which generate the hypothesis given the premise.
[4]Taken from HuggingFace's Dataset Card for SNLI.

underlying premises based on what the model has learned includes those unnecessary words. The discriminative model, in contrast, may simply ignore such filler words.

Based on these observations, we hypothesize that filtering unnecessary words from the premise would help the generative model focus on the words that actually matter for identifying the relationship between the sentences. To do so, we used gradient attributions algorithms called Integrated Gradients (Sundararajan et al., 2017), which assign each word with an attribution between 1 (helpful for the model prediction) and -1 (hurtful for the model prediction). For example, we can see in table 5 that some of the premises only contain 3-4 important words, and the others are not beneficial to the model prediction.

Table 5: Gradient attributions example. Green/red show positive/negative attributions.

Premise	**Hypothesis**	**Label**
a woman in a black shirt looking at a bicycle .	a woman dressed in black shops for a bicycle .	entailment
a black man in a white uniform makes a spectacular reverse slam dunk to the crowd ' s amazement.	the man is asian	contradiction

4.1 Experiments and Results

We calculate feature attributions against our discriminative model, using Captum (Kokhlikyan et al., 2020). We then filter out all the words whose attribution is lower than 0. As each attribution is based on one of the labels, we tested with both the true label of the sample, as well as with the label that the model has predicted. Table 6 shows some encouraging results. The results with the correct labels give us exactly what we wanted – an unbiased model (test and hard test accuracy are similar), with better accuracy on the hard test set than the discriminative model (in table 2). However, this achievement is not genuine, as we used the real labels to filter the premises[5]. Although the generative model did not get access to the labels at test time, the premises it needs to generate were filtered based on attributions w.r.t the gold test label. The table also shows results with attributions w.r.t predicted labels, where we see an improvement only in the accuracy on the regular test set. We could attribute this to the discriminative model's better performance on the regular test set, making the attributions used for filtering the data of higher quality. We hypothesize that given more concise data, our generative model can produce much better results, while still keeping the model unbiased. Nonetheless, more work is needed to obtain better filtering without using the gold test labels.

Table 6: Generative model results (BART), after filtering premises based on true or predicted labels.

	SNLI (Δ without filter)	**SNLI Hard** (Δ without filter)
True label	85.8 (+12.01)	85.43 (+10.88)
Predicted label	76.04 (+2.25)	74.24 (-0.31)

5 Summary

Hypothesis-only biases are a prominent issue in NLI, and are a major obstacle when trying to create robust systems for this task. We have proposed a generative approach for NLI, which leads to creating a model that is not reliant on those biases when making its prediction. This comes, however, with a trade-off, where the generative model performs worse than the discriminative one. We identify the large output space of generating sentences as a possible source for this performance drop. On top of that, we showed that the premises may hold words that are not relevant to the prediction, and we can improve the model by removing those words. Future work can experiment with more methods for removing non-relevant words to help the generative modeling.

[5]Using the gold labels is likely to leak some information about it to the premise.

References

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Sasha Rush. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the Association of Computational Linguistics*. Association of Computational Linguistics, 2019.

Samuel 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*, pp. 632–642, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 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, 2018.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Ramakanth Pasunuru and Mohit Bansal. Multi-task video captioning with video and entailment generation. *arXiv preprint arXiv:1704.07489*, 2017.

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, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 35–40, 2016.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pp. arXiv–1910, 2019.