Examining the Effects of Counterfactually-Augmented Data and Forgettable Examples on NLI

Abstract

This paper examines the performance of a small state-of-the-art natural language model (NLM) on natural language inference (NLI). Following Kaushik et al (2020), I examine in detail the difficulty a pre-trained NLM has on Counterfactually-Augmented (CFA) data. Performance of the NLM is then examined when trained on CFA data directly along with additional, focused learning through the framework of Forgettable Examples proposed by Yaghoobzadeh et al (2021). Overall performance improvement is minimal, but analysis is given to specific classes of what the model fails on and why the forgettable training did not work as strongly in this context.

1 Introduction

1.1 Motivation

In the past few years, state of the art NLMs have evolved to become extremely flexible and powerful. Model's such as BERT have pushed the boundaries of what was previously possible in several tasks of natural language processing (Devlin et al (2019)), with NLI being just one example. However, due to the complex machinery composed of millions of variables and transformations that give these models their power, understanding what exactly a model is learning or how exactly it is using what it has learned is very difficult to understand. Indeed, it is often challenging to conjecture whether a model is in fact learning something novel about the task it is intended to perform, or if it is simply focusing on off-hand, spurious statistical correlations (also referred to dataset artifacts) that happen to reside in the data it is trained on, which may not necessarily generalize to new domains or reveal intimate knowledge related to the given task. One way to examine what a model is truly learning is by testing it on a set of purposefully modified data made to challenge or otherwise expose weaknesses in the model. I follow this technique to examine the challenges that a relatively small, but still powerful NLM has when performing NLI on examples with only minor changes made in relation to the data the model was originally trained on.

1.2 Methodology

I start with an ELECTRA-small NLM (Clark et al 2020), which is a smaller BERT-like NLM and readily available through Huggingface[1]. I train ELECTRA for performing NLI on the Stanford NLI (SNLI) dataset (Bowman et al 2015), also provided through Huggingface[2]. SNLI presents a model with determining a relationship between a premise and a hypothesis, namely whether the hypothesis is entailed by the premise, the hypothesis contradicts the premise, or the hypothesis is undetermined by (neutral to) the premise. Initial training was done for 3 epochs on the full SNLI training dataset with a final accuracy of 89.6% on the evaluation dataset.

Once trained, I test the model on CFA data provided by Kaushik et al (2020). The CFA datasets consist of samples taken directly from SNLI but with a change in label and minor modifications to either the premise or hypothesis. A full description of what the dataset accomplishes is given below. I show that the pre-trained ELECTRA model performs noticeably worse on the CFA data than on the standard SNLI data.

[1] https://huggingface.co/google/electra-small-discriminator

[2] https://huggingface.co/datasets/snli

Example Type	Premise	Hypothesis	Label
Original	The little boy in jean shorts kicks the soccer ball.	A little boy is playing soccer outside.	Neutral (1)
Augmented Premise	The little boy in jean shorts kicks the soccer ball **in the garden.**	A little boy is playing soccer outside.	Entailment (0)
Augmented Premise	The little boy in jean shorts kicks the soccer ball **in the house.**	A little boy is playing soccer outside.	Contradiction (2)
Augmented Hypothesis	The little boy in jean shorts kicks the soccer ball.	A little boy is playing **cricket**.	Contradiction (2)
Augmented Hypothesis	The little boy in jean shorts kicks the soccer ball.	A little boy is playing **soccer**.	Entailment (0)

Table 1: an example from the original SNLI dataset with corresponding CFA changes highlighted

Dataset	Training Examples	Evaluation Examples	Dev Examples
SNLI	549,367	9,842	N/A
CFA Revised	6,664	1,600	N/A
CFA Combined	8,330	2,000	1000*

* 200 original, 400 each for augmented hypothesis and premise

Table 2: sizes of datasets

Dataset	Accuracy
SNLI	89.6%
CFA Revised	72.2%
CFA Combined	76%

Table 3: performance of SNLI trained ELECTRA model

Original Correct	Hypothesis Augmented	Premise Augmented
Correct	301	208
Incorrect	67	160

Original Incorrect	Hypothesis Augmented	Premise Augmented
Correct	25	24
Incorrect	7	8

Table 4: classification breakdown of SNLI trained ELECTRA model on CFA combined dev data

Next, I examine the effect of direct training on CFA data on the model's performance for both the SNLI and CFA datasets. Lastly, I employ the framework of Forgettable Examples to try to focus the model away from learning spurious correlations in the training data and improve its generalizability.

2 Proposed Explorations

I examine the pre-trained ELECTRA model's performance through the lens of CFA data and attempt to improve the model's generalizability using the framework of Forgettable Examples. Both methods are explained in detail below.

2.1 Counterfactually-Augmented Data

Kaushik et al (2020) provide CFA data gathered through a crowdsourcing effort via Amazon's Mechanical Turk platform. Each example was taken directly from the SNLI dataset, with two changes made. The first is a change to the label, so for each original example two new examples are created, one with each different label type. The second change is a minor tweak to either the hypothesis or the premise, typically adding, removing, or changing a few words, which enables just enough context to have the example accord with the new label while retaining a natural coherence to the original setting. Together, these changes amount to counterfactual statements of the original example. Table 1 gives a sample of original examples and their CFA counterparts. The augmented examples are collected in the CFA Revised dataset and collected with their original counterparts in the CFA Combined dataset. Table 2 lists the total number of examples in each dataset, split into three partitions.

When tested on the CFA dataset, the pre-trained ELECTRA model performs worse than on just the SNLI data. Table 3 documents the accuracy of the model on each evaluation dataset. Table 4 uses a held-out dev set to break down the model's performance on the CFA data according to what element of the example was augmented and how the model performed on the corresponding original data. An interesting note is that the model does slightly worse classifying the augmented data corresponding to original examples it had initially classified correctly (69% correct vs 77% correct), suggesting space for improvement in generalizing across the two spaces. Table 5 gives a sample of a specific example that the model had inferred

Example Type	Premise	Hypothesis	Label
Original	Several people wearing foil crowns gathered around a table eating.	A group of people sat around eating dinner.	Neutral (1)
Augmented Premise	Several people wearing foil crowns gathered around a table **playing cards**.	A group of people sat around eating dinner.	Contradiction (2)
Augmented Hypothesis	Several people wearing foil crowns gathered around a table eating.	A group of people **are around eating**.	Entailment (0)
Original	Man looking at a woman that is smoking on the sidewalk.	Men play bingo at shady tables in a local community park.	Contradiction (2)
Augmented Premise	Man looking at a woman that is smoking **next to the shady tables while he plays bingo over there in the local community park with his friends**.	Men play bingo at shady tables in a local community park.	Entailment (0)
Augmented Hypothesis	Man looking at a woman that is smoking on the sidewalk.	**A man gazing at a woman**.	Entailment (0)

Table 5: sample of misclassified CFA examples from SNLI trained model

correctly on the original example but was incorrect with the augmentation. This sample highlights that even minor changes, such as a small bit of extra context or one-word change in the subject of the statement, can trip up the model.

2.2 Forgettable Examples

Forgettable examples are characterized as those examples in a training dataset that were either never classified correctly or at some point in the training of the model were correctly classified but later in the training incorrectly classified. Yaghoobzadeh et al (2021) show that these forgettable examples tend to consist of examples that contradict simple patterns a model may use for classification. An example of this in NLI would be a model relying simply on word overlap between a premise and hypothesis in deciding an entailment label rather than making a deeper connection between the meaning of the statements. In other words, these forgettable examples tend to not fit the spurious correlations a model may have learned from the rest of the training data. Performing an extra round of training on just these forgettable examples could help a model focus less on the spurious correlations and improve its generalizability to new data or domains. I examine employing this extra training procedure on the pre-trained ELECTRA model using the CFA Combined and Revised data directly. Details on how the enhanced training was performed are given in the next section.

3 Implementation Details

3.1 Counterfactually-Augmented Data

The CFA data is readily available on GitHub[3]. Small modifications were made to match the format of the SNLI Huggingface dataset that the ELECTRA model was trained on, namely assigning each label to the correct numerical value (0=entailment, 1=neutral, 2=contradiction), and converting the TSV data to CSV. A data preparation script is given in `format_data.py` found in the code submitted with this paper.

3.2 Forgettable Examples

Implementing the tracking of the forgettable examples took a bit of handiwork to fit within the Huggingface framework. Since I did not have direct access to the training loop used for the ELECTRA model, I had to implement a Huggingface `CustomTrainer` which hooked into the start of each step of the training process. So before performing a training step (a feedforward and backpropagation of input examples through the model), I intercepted the input examples, classified each according to the current state of the model, and recorded the status of each example according to the criteria defined above. The input data here was also already tokenized and shuffled on each epoch as needed for proper training, so instead of associating an example through a numerical identifier from the original dataset, I used the tokenized string representation to record the forgettable status. Full code for recording

[3] `https://github.com/acmi-lab/counterfactually-augmented-data`

Example Type	Premise	Hypothesis	Label
Original	A sitting man is spraying a squirt bottle at a pedestrian while holding an umbrella over his head.	A man reads the paper.	Contradiction (2)
Augmented Premise	A sitting man is **showing a newspaper article** to a pedestrian while holding an umbrella over his head.	A man reads the paper.	Entailment (0)
Augmented Hypothesis	A sitting man is spraying a squirt bottle at a pedestrian while holding an umbrella over his head.	A man **is spraying someone with a squirt bottle on a rainy day**.	Neutral (1)
Original	A girl in blue rides a horse over a tall jump.	A girl is walking her dog down the street.	Contradiction (2)
Augmented Premise	A girl in blue **walks her dog on the street down to the alley**.	A girl is walking her dog down the street.	Entailment (0)
Augmented Hypothesis	A girl in blue rides a horse over a tall jump.	A girl is **riding a mature horse.**	Neutral (1)

Table 6: sample of correct examples from SNLI+CFA Combined trained model

forgettable examples is given in `trainer.py`. The code for examining the forgettable examples and preparing them into a new Huggingface `Dataset` for further training is found in `helpers.py`. A second training loop is implemented in `run.py`, training the model for 10 epochs on the CFA data and then 10 more on just the forgettable examples.

Dataset used for training	SNLI	CFA Revised	CFA Combined
SNLI	**89.6%**	72.2%	76%
SNLI + CFA Revised	80.6%	75.3%	76.8%
SNLI + CFA Combined	84.8%	**75.7%**	**77.6%**
SNLI + CFA Revised w/ Forgetting	76%	72.6%	73.2%
SNLI + CFA Combined w/ Forgetting	81.5%	72.5%	74%

Table 7: performance on each dataset by each model trained

Dataset	Never Learned	Forgotten	Total
CFA Revised	105	312	417
CFA Combined	228	559	787

Table 8: breakdown of forgotten examples

Original Correct	Hypothesis Augmented	Premise Augmented
Correct	289	256
Incorrect	65	98

Original Incorrect	Hypothesis Augmented	Premise Augmented
Correct	34	34
Incorrect	12	12

Table 9: classification breakdown of SNLI+CFA Combined trained ELECTRA model on CFA combined dev data

4 Results

The performance on all datasets after additional training of the ELECTRA model directly on both CFA datasets with and without forgettable examples are given in Table 7. The number of forgettable examples characterized by never being learned or forgotten during training are in Table 8. The performance of the model on the original SNLI data does not improve, but training on the revised and combined CFA data showed small improvements on the CFA datasets. Training through forgettable examples did not give strong results here. Full analysis on results achieved are given below.

5 Analysis

5.1 Counterfactually-Augmented Data

Enhancing the pre-trained ELECTRA model through training directly on the CFA data gave a small improvement in accuracy. A distribution of where exactly the model makes mistakes is given in Table 9, broken down by original classification and performance on augmented data. It does appear that although the model got a few more original examples incorrect (46 total vs 32 total with the SNLI trained model), it was able to classify a higher percent of augmented data correctly on those original examples it also classified correctly (77% vs 69% with the SNLI trained model), suggesting it is indeed generalizing to this domain better than before. Table 6 gives examples of

Example Type	Premise	Hypothesis	Label
Original	A young man with blond dreadlocks sits on a wall reading a newspaper as a woman and a young girl pass by.	The young man is reading the newspaper sports section with a mom and daughter walks to the movie theater.	Neutral (1)
Augmented Premise	A young man is reading the newspaper **business section** as a woman and a young girl pass by.	The young man is reading the newspaper sports section with a mom and daughter walks to the movie theater.	Contradicts (2)
Augmented Hypothesis	A young man with blond dreadlocks sits on a wall reading a newspaper as a woman and a young girl pass by.	The young man is reading the newspaper **as a women and young girl walk by**.	Entailment (0)
Original	A young child wearing a pink top is with a female adult.	A kid is with a woman.	Entailment (0)
Augmented Premise	A young **boy** wearing a pink top is with **another person**.	A kid is with a woman.	Neutral (1)
Augmented Hypothesis	A young child wearing a pink top is with a female adult.	A **little girl** is with a woman.	Neutral (1)

Table 10: sample of augmented examples all models misclassified

augmented data that the improved model was able to correct itself on.

To further examine the challenge CFA data presents, I trained another ELECTRA model from scratch on just the CFA revised data. Examples that all three models (SNLI, SNLI+CFA Combined, just CFA Revised) failed to classify are in Table 10. An interesting finding from this was that the total number of misclassified examples with an augmented hypothesis was 36, while those with an augmented premise was 70. As explored by Poliak et al (2018), the large gap in these two situations could be due to the model relying too heavily on spurious correlations in just the hypothesis for determining classification. Follow up research on how to get the model to use both statements effectively could be worth investigating.

5.2 Forgettable Examples

The forgettable examples did not give improvements here. In fact, training on the forgettable examples generally reduced accuracy on the original and CFA data. I believe this could be due to two factors. First, the CFA data differed enough from the original SNLI data to a point where the model was misclassifying the augmented data due to strict differences in the two data distributions. Forcing the model to train directly on this new distribution affected its classification abilities over the original distribution without gaining enough power to generalize to both. Second, but in a similar reasoning, the pre-trained model was "set in its ways" enough so that the examples it was still misclassifying differed so far from its learned representation of the original data that focusing too much on these caused the model to focus on a degenerate solution.

Dataset used for training	SNLI	CFA Revised	CFA Combined
CFA Combined	75.7%	72.5%	73.4%
CFA Combined w/ Forgetting	74.3%	70.3%	71.3%

Table 11: performance on each dataset by CFA Combined trained model

Yaghoobzadeh et al (2021) somewhat correct for this by gathering the forgettable examples through a weaker model, rather than a stronger pre-trained model. The weaker model obviously misclassifies more examples and would provide a broader range of examples to train on. I did find that training an ELECTRA model from scratch on just the CFA combined data, while not reaching the same absolute performance of the full pre-trained model, did exhibit a much smaller difference between the trained model and the forgettable trained model, suggesting that using a weaker model to determine which examples are out of the ordinary could provide for better generalizability than relying on the biased findings of an already pre-trained model. Results are summarized in Table 11.

6 Conclusion

I have examined how a powerful NLM trained specifically for NLI can be tripped up by seemingly minor yet impactful changes to the data it was originally trained on, calling into question exactly what the model is learning when performing NLI. In further work, the small improvement in performance after directly training on the CFA data should be explored more thoroughly with a larger dataset, perhaps with a more direct analysis of the model's prediction process through the lens of causality to reveal more on the underlying principles of what the model is using to solve its task. Feder et al (2022) give a comprehensive survey on how this might be approached. Developing an automated method for generating the counterfactual examples may also be worth an investment. Finally, an exploration of how a counterfactually trained model with forgettable examples handles transfer learning, perhaps on a QA or sentiment analysis task, could give insight into how these methods induce the model with generalizability across domains.

Acknowledgments

Thank you to Dr. Durret and all the TAs for putting together an interesting course and overall great experience. All model training was performed using the wonderfully cheap and easy-to-use tool that is Google Colab.

References

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). *A large annotated corpus for learning natural language inference*. arXiv. http://arxiv.org/abs/1508.05326

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). *Electra: Pre-training text encoders as discriminators rather than generators*. arXiv. http://arxiv.org/abs/2003.10555

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv. http://arxiv.org/abs/1810.04805

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., Stewart, B. M., Veitch, V., & Yang, D. (2022). *Causal inference in natural language processing: Estimation, prediction, interpretation and beyond*. arXiv. http://arxiv.org/abs/2109.00725

Kaushik, D., Hovy, E., & Lipton, Z. C. (2020). *Learning the difference that makes a difference with counterfactually-augmented data*. arXiv. http://arxiv.org/abs/1909.12434

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). *Hypothesis only baselines in natural language inference*. arXiv. http://arxiv.org/abs/1805.01042

Yaghoobzadeh, Y., Mehri, S., Tachet, R., Hazen, T. J., & Sordoni, A. (2021). *Increasing robustness to spurious correlations using forgettable examples*. arXiv. http://arxiv.org/abs/1911.03861