Fine-tuning of a Pre-trained Model with Mistaken Examples

Kenta Suzue

Abstract

A pre-trained model that suffers from having learned spurious correlations, or dataset artifacts, does not actually solve tasks. In this paper, I analyze a pre-trained model with the Breaking NLI dataset. I show some accuracy improvements by fine-tuning the pre-trained model with mistaken examples. Mistaken examples are examples that were learned incorrectly after each epoch in the course of fine-tuning the models. I show these accuracy improvements for a pre-trained model that was fine-tuned on SNLI only, MNLI only, SNLI then MNLI, and MNLI then SNLI.

1 Introduction

A pre-trained model that returns high accuracy results may not actually solve tasks. Instead such a pre-trained model may have learned spurious correlations, or dataset artifacts, that are responsible for the high accuracy results. When this pre-trained model is evaluated with a different dataset that lacks such spurious correlations, or dataset artifacts, the model's accuracy falls.

To better understand whether a pre-trained model actually solves tasks, I created four different model configurations fine-tuned on the SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) datasets. In this paper, initial or initially fine-tuning refers to a first round of fine-tuning a pre-trained model, and final or finally fine-tuning refers to a second round of fine-tuning a pre-trained model. After initial fine-tuning, I finally fine-tuned the four model configurations with mistaken examples (Yaghoobzadeh et al., 2021). Mistaken examples were examples with mistaken predictions after each training epoch in the course of fine-tuning the models. I analyzed the four model configurations with the Breaking NLI dataset (Glockner et al., 2018), before and after finally fine-tuning with mistaken examples, and observed some accuracy improvements. I conclude this paper with a discussion of a failed alternative approach to finally fine-tuning based on collective human opinions (Nie et al., 2020).

2 Approach

2.1 Initially Fine-tuning the Model

The ELECTRA-small model (Clark et al., 2020) was chosen, due to being trainable relatively quickly with 'only' 14 million parameters, yet capable of GLUE scores comparable to models with 8 times as many parameters such as ELMo and GPT. The ELECTRA-small model is initially fine-tuned in 4 configurations:

(i) the SNLI dataset for 3 epochs,

(ii) the SNLI dataset for 3 epochs and then the MNLI dataset for 3 epochs,

(iii) the MNLI dataset for 3 epochs, and

(iv) the MNLI dataset for 3 epochs and then the SNLI datset for 3 epochs.

2.2 Hyperparameters

Hyperparameters are generally left at default values in the code framework. (Durrett, 2024) The initial learning rate was left at the default value of 5×10^{-5} for all 4 model configurations. For model configuration (ii), after the 3 epochs of initial fine-tuning training on the SNLI dataset, the initial learning rate was reset to 5×10^{-5} for the following 3 epochs of initial fine-tuning training on the MNLI dataset. Similarly, for model configuration (iv), after the 3 epochs of initial fine-tuning training on the MNLI dataset, the initial learning rate was reset to 5×10^{-5} for the following 3 epochs of initial fine-tuning training on the SNLI dataset.

Category	All examples	SNLI only	SNLI+MNLI	MNLI only	MNLI+SNLI
antonyms	1147	333	57	92	59
synonyms	894	13	0	0	5
cardinals	759	16	12	28	13
nationalities	755	20	9	9	17
drinks	731	45	78	139	32
antonyms_wordnet	706	114	111	142	104
colors	699	18	29	54	12
ordinals	663	26	12	14	23
countries	613	9	4	6	15
rooms	595	65	49	59	65
materials	397	7	3	15	1
vegetables	109	29	47	53	25
instruments	65	4	6	6	3
planets	60	20	14	10	9
Total	8193	719	431	627	383

Table 1: Mistaken examples from the Breaking NLI dataset that were predicted incorrectly after initial fine-tuning, by category.

Label	All examples	SNLI only	SNLI+MNLI	MNLI only	MNLI+SNLI
contradiction	7164	661	376	570	331
entailment	982	26	13	18	15
neutral	47	32	42	39	37
Total	8193	719	431	627	383

Table 2: Mistaken examples from the Breaking NLI dataset that were predicted incorrectly after initial fine-tuning, by label.

2.3 Collecting Dataset Examples with Mistaken Predictions During Initial Fine-tuning

During initial fine-tuning as discussed above, at the end of every epoch of training the model's predictions were compared to labels for every example in the training dataset. Every dataset example where the model's prediction failed to match the label was considered as a mistaken example. An index was stored identifying the mistaken example within the training dataset, so that the mistaken example could be used after the conclusion of pre-training. After repeating this storage of indexes of mistaken examples at the end of every epoch of training for 3 epochs, 3 sets of indexes of mistaken examples were stored.

2.4 Analysis of Initially Fine-trained Models with the Breaking NLI Dataset

The 4 initially fine-tuned model configurations were then analyzed with the Breaking NLI dataset (Glockner et al., 2018). The Breaking NLI dataset was designed to exceed the SNLI dataset by challenging the generalization capabilities of pre-trained models, such as inferences that depend on lexical and world knowledge. To understand the weaknesses of the 4 initially fine-tuned model configurations, each of the 4 pre-trained model configurations was evaluated against the Breaking NLI dataset. In evaluation, each Breaking NLI dataset example's label was compared against the initially fine-tuned model's prediction, and if the prediction was mistaken an index was stored identifying the mistaken example within the Breaking NLI dataset.

All the mistaken examples in the Breaking NLI dataset were counted across two types of buckets: example category and example label. As shown in Table 1, the example categories were: antonyms, synonyms, cardinals, nationalities, drinks, antonyms_wordnet, colors, ordinals, countries, rooms, materials, vegetables, instruments, and planets. As shown in Table 2, the example labels were: contradiction, entailment, and neutral.

Model configuration (i), initially fine-tuned on

Epoch	SNLI only	SNLI+MNLI	MNLI only	MNLI+SNLI
Epoch 1	65014	64276	67344	62438
Epoch 2	64390	63298	66581	61768
Epoch 3	64388	63261	66509	61766
Epoch 1 or 2	67989	68199	71750	65048
Epoch 1 or 3	67987	68156	71559	65055
Epoch 2 or 3	64433	63857	67257	61804
Epoch 1, 2, or 3	68009	68464	72001	65067
Epochs 1 and 2	61415	59375	62175	59158
Epochs 1 and 3	61415	59381	62294	59149
Epochs 2 and 3	64345	62702	65833	61730
Epochs 1, 2, and 3	61392	59087	61869	59132

Table 3: Mistaken examples from the training dataset that were predicted incorrectly after each training epoch of initial fine-tuning. When the model was initially fine-tuned on two datasets, the mistaken examples were from the latter dataset.

	SNLI only	SNLI+MNLI	MNLI only	MNLI+SNLI
Initial learning rate	2.7×10^{-7}	2.6×10^{-7}	2.4×10^{-7}	2.0×10^{-7}

Table 4: Initial learning rate for final fine-tuning

SNLI only, had the largest number of mistaken examples in the antonyms category, followed by the antonyms_wordnet category. The other 3 model configurations had the largest number of mistaken examples in the antonyms_wordnet category. All 4 model configurations had the largest number of mistaken examples with the contradiction label; this was not surprising because the Breaking NLI dataset had more examples with the contradiction label than the entailment label by a factor of 7, and more examples with the contradiction label than the neutral label by a factor of 150. Analysis of the 4 model configurations showed that, at least with respect to the Breaking NLI dataset, the greatest scope for improvement was with the antonyms and antonyms_wordnet categories, and with the contradiction label.

2.5 Collecting the Examples for Final Fine-tuning

As discussed in subsection 2.3, during initial fine-tuning, 3 sets of indexes of mistaken examples were stored, identifying the example within the training dataset with a mistaken prediction at the end of a respective one of the 1st, 2nd, or 3rd epoch of fine-tuning. Table 3 shows the number of these mistaken examples for the all 4 model configurations, for each epoch of the 1st, 2nd, or 3rd epoch of fine-tuning, as well as Boolean combinations. In the model configuration 'SNLI+MNLI' (model configuration (ii), initially fine-tuned on the SNLI dataset for 3 epochs and then the MNLI dataset for 3 epochs) and 'MNLI+SNLI' (model configuration (iv), initially fine-tuned on the MNLI dataset for 3 epochs and then the SNLI dataset for 3 epochs), only the mistaken examples from the latter dataset were collected. The Boolean combinations shows that, generally a mistaken example remains mistaken for all 3 epochs of fine-tuning. The Boolean combination of 'Epochs 1, 2, and 3' was selected as the collection of dataset examples for final fine-tuning. Otherwise, a mistaken example was not a mistaken example for at least one epoch of the 3 epochs of initial fine-tuning.

2.6 Final Fine-tuning with Mistaken Examples

Fine-tuning with mistaken examples can improve models (Yaghoobzadeh et al., 2021). Accordingly, at this point the mistaken examples that were collected during initial fine-tuning per subsection 2.3 were used for final fine-tuning. Final fine-tuning was performed with the collected mistaken examples for 1 epoch, with initial learning rates as shown in Table 4. With model configuration (ii), initially fine-tuned on the SNLI dataset for 3 epochs and then the MNLI dataset for 3 epochs, final fine-tuning for 1 epoch was performed only with the mistaken examples collected

Category	All examples	SNLI only	SNLI+MNLI	MNLI only	MNLI+SNLI
antonyms	1147	322 (-11)	61 (+4)	90 (-2)	59 (0)
synonyms	894	12 (-1)	0 (0)	0 (0)	7 (+2)
cardinals	759	22 (+6)	12 (0)	29 (+1)	13 (0)
nationalities	755	12 (-8)	8 (-1)	6 (-3)	14 (-3)
drinks	731	39 (-6)	54 (-24)	98 (-41)	29 (-3)
antonyms_wordnet	706	116 (+2)	117 (+6)	139 (-3)	106 (+2)
colors	699	15 (-3)	26 (-3)	48 (-6)	12 (0)
ordinals	663	23 (-3)	18 (+6)	19 (+5)	21 (-2)
countries	613	5 (-4)	3 (-1)	6 (0)	11 (-4)
rooms	595	69 (+4)	44 (-5)	47 (-12)	72 (+7)
materials	397	7 (0)	2 (-1)	11 (-4)	1 (0)
vegetables	109	22 (-7)	46 (-1)	54 (+1)	21 (-4)
instruments	65	4 (0)	5 (-1)	4 (-2)	3 (0)
planets	60	13 (-7)	9 (-5)	7 (-3)	9 (0)
Total	8193	681 (-38)	405 (-26)	558 (-69)	378 (-5)

Table 5: Mistaken examples from the Breaking NLI dataset that were predicted incorrectly after final fine-tuning, by category. Parentheses show net difference relative to before final fine-tuning. Negative numbers in parentheses show improvement from final fine-tuning.

Label	All examples	SNLI only	SNLI+MNLI	MNLI only	MNLI+SNLI
contradiction	7164	618 (-43)	355 (-21)	501 (-69)	326 (-5)
entailment	982	29 (+3)	11 (-2)	17 (-1)	16 (+1)
neutral	47	34 (+2)	39 (-3)	40 (+1)	36 (-1)
Total	8193	681 (-38)	405 (-26)	558 (-69)	378 (-5)

Table 6: Mistaken examples from the Breaking NLI dataset that were predicted incorrectly after final fine-tuning, by label. Parentheses show net difference relative to before final fine-tuning. Negative numbers in parentheses show improvement from final fine-tuning.

during initial fine-tuning with the MNLI dataset. With model configuration (iv), initially fine-tuned on the MNLI dataset for 3 epochs and then the SNLI dataset for 3 epochs, final fine-tuning for 1 epoch was performed only with the mistaken examples collected during initial fine-tuning with the SNLI dataset.

The term 'mistaken examples' is used for both mistaken examples from the SNLI and/or MNLI datasets, as well as mistaken examples from the Breaking NLI dataset. However, only the mistaken examples from the SNLI and/or MNLI datasets are used in the final fine-tuning to improve the aggregate accuracy of the 4 model configurations. The mistaken examples from the Breaking NLI dataset are merely part of the evaluation of the 4 model configurations, and not part of further fine-tuning training with mistaken examples in the spirit of (Yaghoobzadeh et al., 2021).

2.7 Analysis of Finally Fine-tuned Models with Breaking NLI Dataset

After final fine-tuning, the 4 finally fine-tuned model configurations were then analyzed with the Breaking NLI dataset (Glockner et al., 2018). As in subsection 2.4, after the final fine-tuning for 1 epoch, each of the 4 pre-trained model configurations was evaluated against the Breaking NLI dataset. In evaluation, each Breaking NLI dataset example's label was compared against the finally fine-tuned model's prediction, and if the prediction was mistaken an index was stored identifying the mistaken example within the Breaking NLI dataset.

All the mistaken examples in the Breaking NLI dataset were counted across two same types of buckets as in Tables 1 and 2, example category and example label. Table 5 again shows the example categories: antonyms, synonyms, cardinals, nationalities, drinks, antonyms_wordnet, colors, ordinals, countries, rooms, materials, vegetables,

Final fine-tuning with mistaken example	SNLI only	SNLI+MNLI	MNLI only	MNLI+SNLI
before	91.22%	95.32%	92.35%	94.74%
after	91.69%	95.39%	93.19%	95.06%
improvement	+0.47%	+0.07%	+0.84%	+0.32%

Table 7: Evaluation accuracy of model on Breaking NLI dataset, before and after final fine-tuning

instruments, and planets. Table 6 again shows the example labels: contradiction, entailment, and neutral.

Table 5 shows, in addition to the count of example categories, a number in parentheses showing the net difference relative to Table 3. Table 6 shows, in addition to the count of example labels, a number in parentheses showing the net difference relative to Table 4. The negative number in parentheses in the bottom row for totals shows an aggregate improvement in all 4 model configurations after final fine-tuning.

Model configuration (i) (SNLI only), after initial fine-tuning, had the largest number of mistaken examples in the antonyms category, and after final fine-tuning had the largest reduction in the number of mistaken examples also in the same, antonyms category. The other 3 model configuration, after initial fine-tuning, had the largest number of mistaken examples in the antonyms_wordnet category. However, after final fine-tuning, model configurations (ii) (SNLI+MNLI) and (iii) (MNLI only) had the largest reduction in mistaken examples also in a different, drinks category; and model configuration (iv) (MNLI+SNLI) had the largest reduction in mistaken examples also in different, rooms and vegetables categories (a tie between two categories).

All 4 model configurations, after initial fine-tuning, had the largest number of mistaken examples with the contradiction label; and after final fine-tuning had the largest reduction in the number of mistaken examples with the contradiction label.

2.8 Accuracy Improvements

All 4 model configurations showed some aggregate accuracy improvement as shown in Table 7, where accuracy was measured by evaluating each of the 4 finally-tuned model configurations against the Breaking NLI dataset. The percentage gain in accuracy was modest, in the range of +0.07% to +0.84%.

3 Failed Approaches

I conclude this paper with a discussion of a failed alternative approach to final fine-tuning based on collective human opinions (Nie et al., 2020). The idea was to collect dataset examples that were ambiguous, and perform final fine-tuning on these ambiguous examples so that the models could learn ambiguity.

One attempt used the ChaosNLI dataset from (Nie et al., 2020) and focused on examples where no more than 50 out of 100 annotations for an example agreed on a same label; another attempt tried no more than 60 out of 100 annotations for an example agreed on a same label. Yet another attempt focused on examples from the SNLI dataset where no more than 3 out of 5 annotations for an example agreed on a same label. However, all of these attempts worsened the accuracies for the 4 model configurations after final fine-tuning.

References

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

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

Greg Durrett. 2024. University of Texas at Austin, CS388 Natural Language Processing class, Final Project code download. https://www.cs.utexas.edu/ gdurrett/courses/online-course/materials.html.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332. Association for Computational Linguistics.