Identifying Important Subsets of Train Data to Improve Model Robustness

Author
Jason Yu
jy26322
jasonyu112@utexas.edu

Abstract
This paper aims to use Dataset Cartography to improve model performance in NLI tasks. Currently, state-of-the-art models get very high accuracies for in-distribution data(ID), however, models tested on out-of-distribution data(OOD) perform significantly worse, indicating that models use shallow signals instead of learning language. We first make subsets out of the original training set into random, easy, hard, and ambiguous datasets using confidence and variability. We trained separate models using each subset and evaluated them using contrast and adversarial sets. Our goal is to use Dataset Cartography to reproduce results from Swabha Swayamdipta et al., show that it is a useful technique for identifying important subsets, and shorten training times while increasing model robustness.

Introduction
Public datasets are widely becoming more common and accessible. However, these datasets include human error and bias that may reduce the quality of resulting models. Using confidence and variability to differentiate examples as in Swabha Swayamdipta et al. (2020), we can plot and isolate regions of a data map corresponding with easy, hard-to-learn, and ambiguous examples. With around 80% of examples being located in the easy to learn region, this is just one example of the many biases humans make that can be negatively captured by the model. A consequence of this could be that the model identifies shallow indicators such as word overlap, sentence length, or sentence structure instead of learning and understanding language. Thus, these models tend to perform well on in-distribution data(ID) but significantly worse on out-of-distribution data(OOD). It has also been shown that models trained on ambiguous regions of the data map result in a better ability to generalize and adapt to OOD data(Swabha Swayamdipta et al., 2020).

Currently, state-of-the-art models benchmarked on ANLI(Yixin Nie et al., 2020) achieve an average of 58.9% accuracy over the 3 provided test sets. In general, examples in ANLI contain longer sentences than SNLI, penalizing the model more if it learns using shallow indicators. The low accuracy of state-of-the-art models indicates a low ability to generalize even for big models when presented with OOD data. This is the motivating factor for analyzing data to better understand what the model needs to be more robust.

Baseline
We used the Electra-small model and trained it on the SNLI dataset for 3 epochs. This model has around 15 million parameters and achieves comparable performance on less powerful machines when compared to bigger models. Our baseline results, using 100% of the SNLI

dataset provided through HuggingFace and with duplicate examples removed, achieved an accuracy of 89.49% when evaluated on the SNLI test set. In addition, we benchmarked the models using more difficult OOD tasks using contrast(Li, C. et al., 2020) and adversarial sets(Yixin Nie et al., 2020), which consists of 4 tests total. The baseline model achieved a 74.89% accuracy for the contrast set and 29.89%, 30.19%, and 31% for the 3 adversarial test set.

Figure 1. Major error classes are shown below. These are errors gathered from our baseline model, trained on the SNLI dataset for 3 epochs, tested against the SNLI test set.

Model uses overlap

Premise	Hypothesis	Gold Label	Prediction
"At an outdoor event in an Asian-themed area, a crowd congregates as one person in a yellow Chinese dragon costume confronts the camera."	"A crowd is in front of a camera"	1	0
"People are throwing tomatoes at each other."	"The people are having a food fight."	0	1
"Number 13 kicks a soccer ball towards the goal during children's soccer game."	"A player passing the ball in a soccer game."	2	0

Model incorrectly reasons about physical space

Premise	Hypothesis	Gold Label	Prediction
"A taxi SUV drives past an urban construction site, as a man walks down the street in the other direction."	"A man is chasing an SUV that is going in the same direction as him."	2	1
"A taxi SUV drives past an urban construction site, as a man walks down the street in the other direction."	"An SUV and a man are going in opposite directions."	0	2
"A brown a dog and a black dog in the edge of the ocean with a wave under them boats are on the water in the background."	"The dogs are swimming among the boats."	0	1

Model incorrectly reasons about intentions

Premise	Hypothesis	Gold Label	Prediction
"two people are waiting for a train at a station."	"They expect a train to arrive eventually."	0	1
"Two men sitting on a subway are reading, with coats and scarves on, but have seemed to have lost their pants."	"The men are wearing pants."	2	0
"Two black little boys hanging over an upside down bicycle, one examining the pedal."	"Someone is riding a bike."	2	0

Error Analysis

To improve the model, we first analyzed and categorized errors generated on the SNLI test set. In Figure 1, we see that the model struggles with the ability to reason with physical space. This is the most consistent category with almost all examples, in our sample of analyzed errors, containing observations about physical space resulting in a wrong prediction label. The model also uses word overlap often. In the example above containing "an outdoor event in an Asian-themed area, a crowd congregates as one person in a yellow Chinese dragon costume confronts the camera", the model mistakenly predicts entailment for "the crowd is in front of the camera" because it finds overlap in the crucial words crowd and camera. The next major category found was that the model incorrectly reasons about intentions. Errors of this class are usually more ambiguous and require the model to think deeper than just the word's definition. While we also found error categories, such as mislabeled data and ambiguous examples with multiple correct labels, the 3 categories in Figure 1 accounted for the majority of errors, indicating that the model learned shallow indicators to achieve the 89% SNLI test score. Using ANLI, which contains longer sentences, further exposes this point of failure due to the need for the model to use context more effectively.

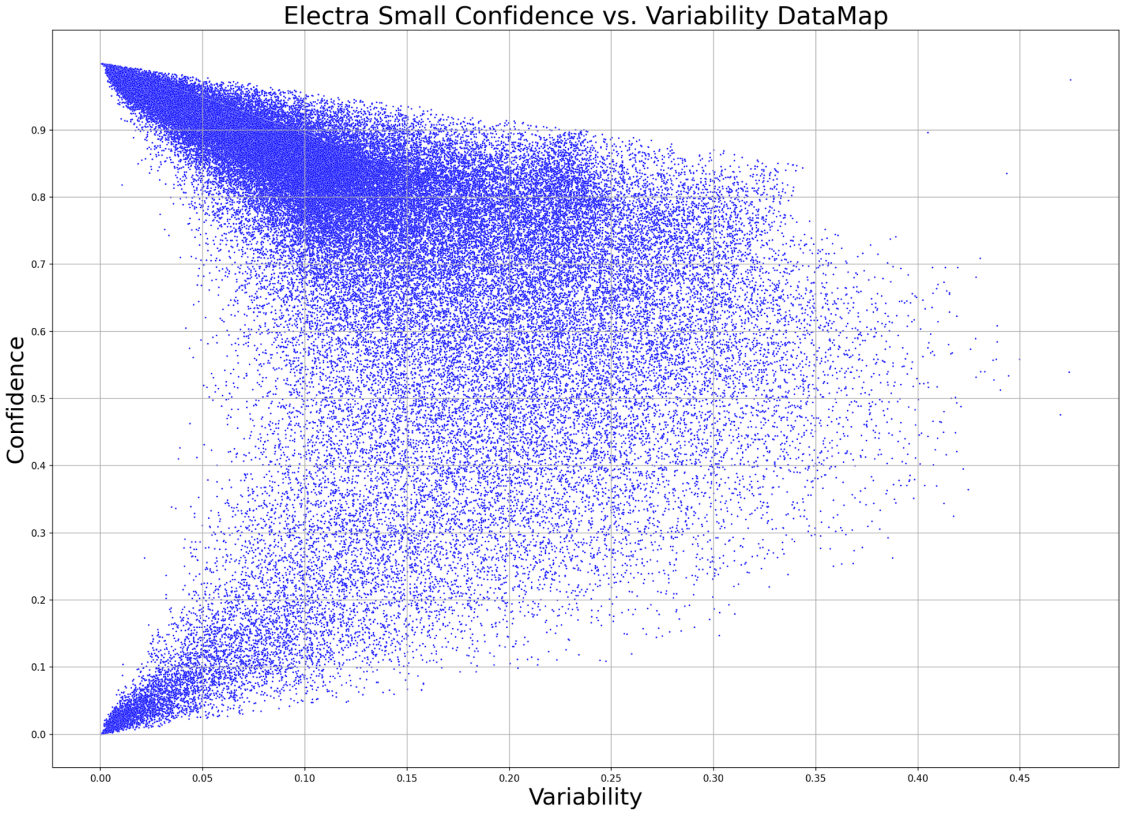
Figure 2. Ambiguous or mislabeled examples are shown below. Examples are chosen from the top 150,000 most ambiguous(low confidence/low variability) in a dataset of 548714 examples.

Premise	Hypothesis	Gold Label	Predicted label
"Crowd of people outside a blue building."	"a man jumped from height and dead"	0	2
"A group of people standing around."	"A group of people doing nothing."	0	2
"A man with a beige coat is crossing the street carrying a bag."	"The man carrying a bag is walking"	1	2
"A shirtless man is scooping away sand with a pink scoop while a little girl watches."	"There is a scoop being used."	2	0
"A man wearing a white shirt tosses a frisbee in the park."	"A man is at the park."	2	0

Dataset Cartography

Dataset Cartography is a technique for mapping data using confidence and variability such that regions of the map can be interpreted as different example types. Easy examples are theorized to have high confidence with low variability while hard-to-learn examples are predicted with low confidence and low variability(Swabha Swayamdipta et al., 2020). Through analysis of hard-to-learn examples in Figure 2, we can see that these are often difficult examples even for a human annotator because the gold label is either highly ambiguous or arguably mislabeled. Lastly, there is a region of the map pertaining to ambiguous examples. These are examples categorized by their high variability and medium confidence, which could mean that they are more complex than easy examples, whereas hard-to-learn examples have a tendency to be mislabeled. Using this technique, we identified subsets of the original training data, consisting of 150,000 training examples compared to the original 548,714, and trained separate models on these subsets for evaluation.

Figure 3. Dataset Cartography map for SNLI training set(random 150,000 sample). This uses the Electra-small model and the confidence/variability is calculated with 6 epochs.



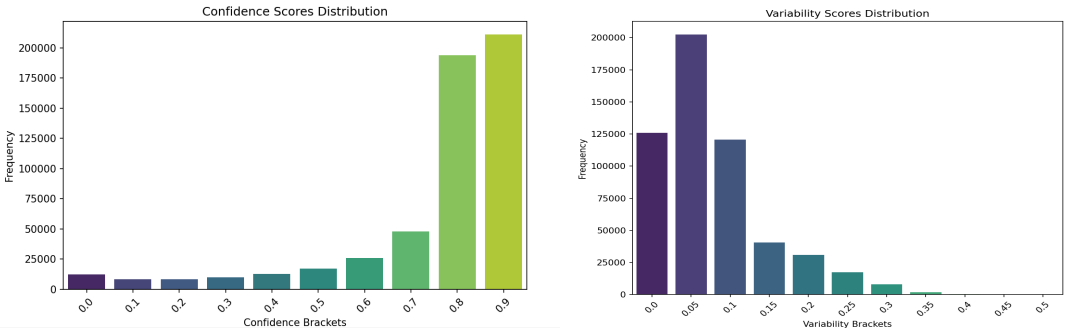


Figure 4. Confidence and variability distributions for SNLI training set at 6 epochs.

Experimentation

We gathered logit data for calculating confidence and variability(Swabha Swayamdipta et al., 2020) using an Electra-small model trained for 6 epochs on the SNLI dataset. As shown in Figure 3, our model contains regions for easy-to-learn, ambiguous, and hard-to-learn similar to Swabha Swayamdipta et al(2020). For our easy-to-learn training subset, we filtered for the top 150,000 most confident examples which is 27% of the total SNLI training set. Our hard-to-learn examples were filtered for the least confidence and our ambiguous subset contained the most variable examples. In addition, we took a random 150,000 example subset from the SNLI training set to give a fair comparison to the new training subsets.

Each subset was trained on separate Electra-small models for 3 epochs, and the models were benchmarked using the SNLI test set, contrast set, and adversarial sets. Figure 5 presents a chart summarizing our findings. Our model, trained on ambiguous examples, achieved SNLI and contrast test accuracies close to the 150,000 random model. However, comparing adversarial test sets, we can see a significant boost, indicating that the model performs better in the context-heavy tests while maintaining good accuracy on other tests. The model trained on easy examples achieved a significantly lower SNLI accuracy and inconsistent ANLI accuracies. Unexpectedly, the model trained on hard examples had a very low 45% and 48% for the SNLI and contrast sets, but was still able to achieve higher ANLI test accuracies.

Figure 5. Table summarizing results from experimentation. The table contains accuracies from the SNLI, Contrast, and ANLI test sets. All training subsets contain 150,000 examples.

Model	SNLI	Contrast	R1	R2	R3
Easy	84.69	71.19	31.7	33.5	29.66
Hard	45.25	48.18	33.7	30.09	34.08
Ambiguous	86.29	72.49	33.3	32.6	33.25
Random	86.97	73.45	31.7	30.7	31.58

Discussion

Using Dataset Cartography to find crucial subsets for training, we show that using higher quality data can improve model robustness while maintaining accuracy in other tests with a smaller training set, which dramatically shortens training time and allows resources to be used more efficiently. Although creating subsets using confidence and variability show promising results, there may exist metrics other than the ones used in Swabha Swayamdipta et al. (2020) that better identifies subsets in a training set. Going forward, future work could explore other metrics for subset identification. Also, using a combination of Dataset Cartography with other techniques could further improve the model's robustness.

Conclusion

In this paper, we reproduced the findings from Swabha Swayamdipta et al. (2020) and demonstrated the effectiveness of Dataset Cartography for creating subsets to improve model robustness. By segmenting data to include higher quality examples, we were able to improve performance on OOD tests while maintaining high accuracy on ID tests. In addition, segmenting data to focus on higher quality examples significantly reduces training times. Dataset Cartography demonstrates great potential and serves as a promising tool for building robust models.

References

[Swayamdipta et al.2020] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with
training dynamics. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing
(EMNLP), pages 9275–9293, Online, November. Association for Computational Linguistics.

[Nie, Y. et al. (2020)] Adversarial NLI: A new benchmark for natural language understanding, ACL Anthology. Available at:
https://aclanthology.org/2020.acl-main.441/ (Accessed: 16 November 2024).

[Li, C. et al. (2020)] A method for automatically generating contrast sets. Available at: https://aclanthology.org/2020.blackboxnlp-1.12.pdf (Accessed: 17 November 2024).