Evaluating and Crafting Datasets Effective for Deep Learning With Data Maps

Bishnu, Jay **Gondoputro, Andrew**

August 4, 2022

Abstract

Rapid development in deep learning model construction has prompted an increased need for appropriate training data. The popularity of large datasets - sometimes known as "big data" - has diverted attention from assessing their quality. Training on large datasets often requires excessive system resources and an infeasible amount of time. Furthermore, the supervised machine learning process has yet to be fully automated: for supervised learning, large datasets require more time for manually labeling samples. We propose a method of curating smaller datasets with comparable out-of-distribution model accuracy after an initial training session using an appropriate distribution of samples classified by how difficult it is for a model to learn from them.

1 Introduction

Though larger datasets are often preferred for deep learning to maximize out-of-distribution accuracy, their benefits begin to diminish at some size (Banko and Brill, 2001). A dataset's efficiency can be defined qualitatively as the accuracy of a model trained on it relative to the amount of time and resources needed to produce the model. Our work utilizes the PyTorch Dataset Cartography project (Swayamdipta et al., 2020) to evaluate metrics of individual samples (dubbed "training dynamics") in a dataset after training. By using training dynamics to represent the correctness, confidence, and variability in a model's prediction for a sample's label, we can categorize samples by their difficulty for the model to learn. With groups of these samples, we can craft a new data subset. Ideally, a data subset constructed from a certain proportion of hard-to-learn, easy-to-learn, and ambiguous samples could achieve comparable real-world performance while requiring a fraction of the memory and training time.

2 Background

2.1 Datasets

The Dataset Cartography project is designed for use with several GLUE-style datasets. For this project, we used the Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015). All GLUE-style datasets are stored

in the TSV (tab-separated-value) format.

The SNLI dataset contains two statements per sample. The pair is assigned a gold label[1] - entailment, contradiction, or neutral. The model is trained to predict the label of a new sentence pair using the train split[2] of the data. The model is then tested on the out-of-distribution dev split before returning to training, and then finally tested on the test split.

2.2 Data Maps

Manually estimating the effectiveness of each sample in a dataset for training can be costly and time-consuming. The Dataset Cartography project was first proposed as a way to characterize samples in a dataset with a chart. The samples in a model's training sequence are plotted according to their training dynamics, where the y-axis represents a model's confidence in its prediction for the sample and the, x-axis represents the variability in confidence. These metrics allow both humans and computers to identify from which samples the model learns easily and which ones it struggles to learn.

The Dataset Cartography project records new metrics after every epoch of training. They are used to label every example in a training split with its correctness, confidence, variability, threshold closeness, and forgetfulness. Correctness represents the fraction of epochs in which the model predicted the label for a given sample correctly. Confidence is a metric within the range of [0, 1] that measures the mean of the model's probabilities of a sample being labeled correctly across epochs. Variability is the standard deviation of these probabilities. While the program also calculates metrics like forgetfulness and threshold closeness, they are used for neither plotting nor filtering.

[1] A label declared to be accurate, usually by a human.

[2] A portion of the entire dataset.

Based on the confidence and variability of a data point, the project classifies it into one of three categories: easy-to-learn, hard-to-learn, or ambiguous. Easy-to-learn examples generally have low variability and high confidence, hard-to-learn examples had low variability and low confidence, and ambiguous examples have high variability.

Labeling data points in this way allows for data subsets to be filtered from the original dataset based on specific criteria; for example, the samples the model found easiest to learn from can be described as those with the highest confidence. The project allows for a selection of samples - i.e. the 33% of samples with the highest confidence - to be filtered into a new data subset. Our method uses this feature to create new data subsets.

The original paper introducing the Dataset Cartography project describes several out-of-dataset accuracy tests. The authors evaluated the accuracy of models trained on the top 33% easiest-to-learn, hardest-to-learn, and most ambiguous samples. The model trained on the ambiguous samples most consistently predicted the correct label for both in-distribution and out-of-distribution data.

However, further tests revealed that small ambiguous datasets performed worse. This phenomenon could indicate that ambiguous samples alone were not sufficient for learning and that there may be a combination of sample types with better performance.

3 Methods

Our goal was to expand upon the training analysis from the original Dataset Cartography project paper. We used DistilRoBERTa with the SNLI corpus for training and evaluation.

We trained the entire dataset for six epochs to classify samples as easy-to-learn, hard-to-learn, or ambiguous. Using the filtering func-

Sentence 1	Sentence 2	Gold Label
Three black dogs on grass.	Three dogs on grass.	Entailment
Three men are sitting on chairs.	The men are dancing together on the dance floor.	Contradiction
Short dark hair woman with a black phone is on the phone on the sidewalk.	A woman is on the phone with her husband.	Neutral

Table 1: Samples from the train split of the SNLI dataset.

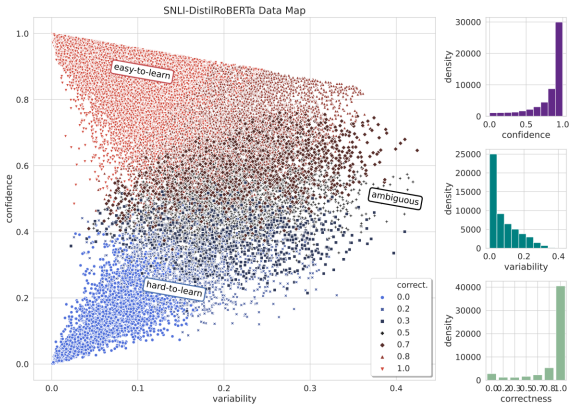

Figure 1: A data map for a DistilRoBERTa-base model trained on the SNLI dataset without shuffling.

tion, we sorted by the confidence metric to get the hardest-to-learn and easiest-to-learn samples. We also sorted by the variability metric to get the most ambiguous samples. We then created nine new subsets after filtering the labeled data: a copy of the original dataset, but with shuffled samples; a randomly selected 33% sample of the original dataset; three 33% datasets, or one for the top 33% of each of the categories; three 16.67% + 16.67% datasets, with the three possible combinations of the categories (i.e. 16.67% easy-to-learn and 16.67% hard-to-learn); and a dataset containing the top 11% of samples from all three categories.

4 Findings

4.1 Results

After training all nine models on the new datasets, we recorded their accuracy in labeling the in-distribution train set and the out-of-distribution dev and test sets.

The model trained on the entire SNLI dataset had the highest out-of-distribution accuracies, while the model trained on easiest-to-learn samples had the highest in-distribution accuracy. The model trained on only the hardest-to-learn and most ambiguous samples had the lowest out-of-distribution accuracy, but the model trained on the hardest-to-learn samples had the lowest in-distribution accuracy.

We found that the model trained on ambiguous samples had comparable out-of-distribution accuracies to the entire dataset's, but poor in-distribution accuracy. Excluding the model trained on the whole dataset, the model with the highest overall accuracies was trained on the easiest-to-learn and most ambiguous samples.

4.2 Mislabeled Data

The original Dataset Cartography paper discussed the idea of noisy data (mislabeled samples), which could be made easier to identify by filtering hard-to-learn samples. The SNLI Dataset classifies pairs of sentences as

	Final Training Accuracy (ID)	Dev Accuracy (OOD)	Test Accuracy (OOD)
100.00%	0.8936	0.8997	0.8976
33.33% random	0.8782	0.8776	0.8785
33.33% easy-to-learn	0.9996	0.8293	0.8286
33.33% hard-to-learn	0.5680	0.5966	0.5856
33.33% ambiguous	0.7684	0.8878	0.8894
16.67% easy-to-learn 16.67% hard-to-learn	0.7581	0.5938	0.5999
16.67% easy-to-learn 16.67% ambiguous	0.8835	0.8802	0.8807
16.67% hard-to-learn 16.67% ambiguous	0.5900	0.4076	0.4023
11.11% easy-to-learn 11.11% hard-to-learn 11.11% ambiguous	0.7401	0.5249	0.5145

Table 2: Results from training and testing models based on the data subsets.

either entailment, neutral, or contradiction, but some samples with inaccurate gold labels could interfere with training and evaluation. While exploring some of the samples categorized as hard-to-learn, we found several mislabeled data points. We strongly suspect these were partially responsible for the poor performance of the models trained on subsets with the hardest-to-learn samples.

5 Conclusion

Our research demonstrated that a model trained on only certain samples in a dataset can achieve accuracy comparable to one trained on the entire dataset. In particular, models trained on subsets with a strong focus on ambiguous samples and little-to-no focus on hard-to-learn samples generally had excellent out-of-distribution performance, while those trained on easy-to-learn samples had near-perfect in-distribution performance. We can reasonably infer that evaluating a dataset by the quality of its samples through training dynamics can allow for smaller, more efficient models with equivalent - or potentially even superior - performance.

6 Further Research

The Dataset Cartography project has other useful applications in dataset and model construction. Not only can it be used to identify mislabeled or low-quality samples, but the filtering process could also be automated further to find a "golden ratio" of easy-to-learn, hard-to-learn, and ambiguous samples for a particular dataset. Though the project is currently constrained to a select few datasets, the concept can be generalized to fields beyond natural language processing. The project is also only designed for use in data categorization projects. However, a similar process could potentially allow model engineers to evaluate the quality of data used to train generative adversarial networks.

Sentence 1	Sentence 2	Gold Label	Our Label
A hiker travels along a rocky path bordered by greenery.	The hiker is carrying a big backpack on his back.	Entailment	Neutral
Two beige dogs are playing in the snow.	Two dogs are outside.	Contradiction	Entailment
A man with a gray beard is standing next to a woman in a brown jacket.	Nobody is standing.	Neutral	Contradiction

Table 3: Some of the mislabeled samples in the SNLI train split.

7 Acknowledgments

We are grateful to have performed our research under the guidance of Mark Galassi, Rhonda Crespo, Maria de Hoyos, and the rest of the Institute of Computing in Research team. We would also like to thank Dr. Ameeta Agrawal and her Ph.D. student Yufei Tao at Portland State University for spearheading this project and guiding us during our work. Finally, we thank the Allen Institute for Artificial Intelligence for the original Dataset Cartography project and accompanying research.

Bibliography

Banko, Michele and Eric Brill (2001). "Mitigating the Paucity-of-Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for Natural Language Processing". In: *Proceedings of the First International Conference on Human Language Technology Research*. URL: https://aclanthology.org/H01-1052.

Swayamdipta, Swabha et al. (2020). "Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics". In: *Proceedings of EMNLP*. URL: https://arxiv.org/abs/2009.10795.

Bowman, Samuel R. et al. (2015). "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.