Evaluating and Crafting Datasets Effective for Deep Learning With Data Maps

Jay Bishnu Andrew Gondoputro

Institute for Computing In Research

August 4, 2022

Guiding Questions

How can we assess the quality of a dataset?

What types of samples are best for building high-quality datasets?

Natural Language Processing

- Natural language processing (NLP)
  - Text categorization
  - Text filtering
  - Sentiment analysis
- Example: given the text of a movie review, a model predicts the rating (in stars) for that review

Training

- Supervised learning: samples in a dataset are labeled by humans
  - For deep learning, data sets are generally at least 10,000 samples long
  - Larger data sets are often better for training, but take longer to process (Banko and Brill, 2001)
- Datasets are often divided into three ”splits”
  - Train split: the samples on which the model is trained (in-distribution)
  - Dev split: the samples on which the model is evaluated to make sure it is learning during training
  - Test split: the samples on which the model is tested to produce a final accuracy measurement

The SNLI Dataset

Stanford Natural Language Inference

- Dataset with more than 500,000 samples (Bowman et al., 2015)
- Sentences are paired and labeled
  - Entailment: if Sentence 1 is true, Sentence 2 must also be true
  - Contradiction: if Sentence 1 is true, Sentence 2 cannot be true
  - Ambiguous: If Sentence 1 is true, Sentence 2 could be either true or false

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

Data Maps

Data maps - visual representation of sample performance from training (Swayamdipta et al., 2020)

Metrics

- Confidence
  - On average, how confident is the model in predicting the correct label?
- Variability
  - How much does confidence vary across epochs?

Categorization of Data

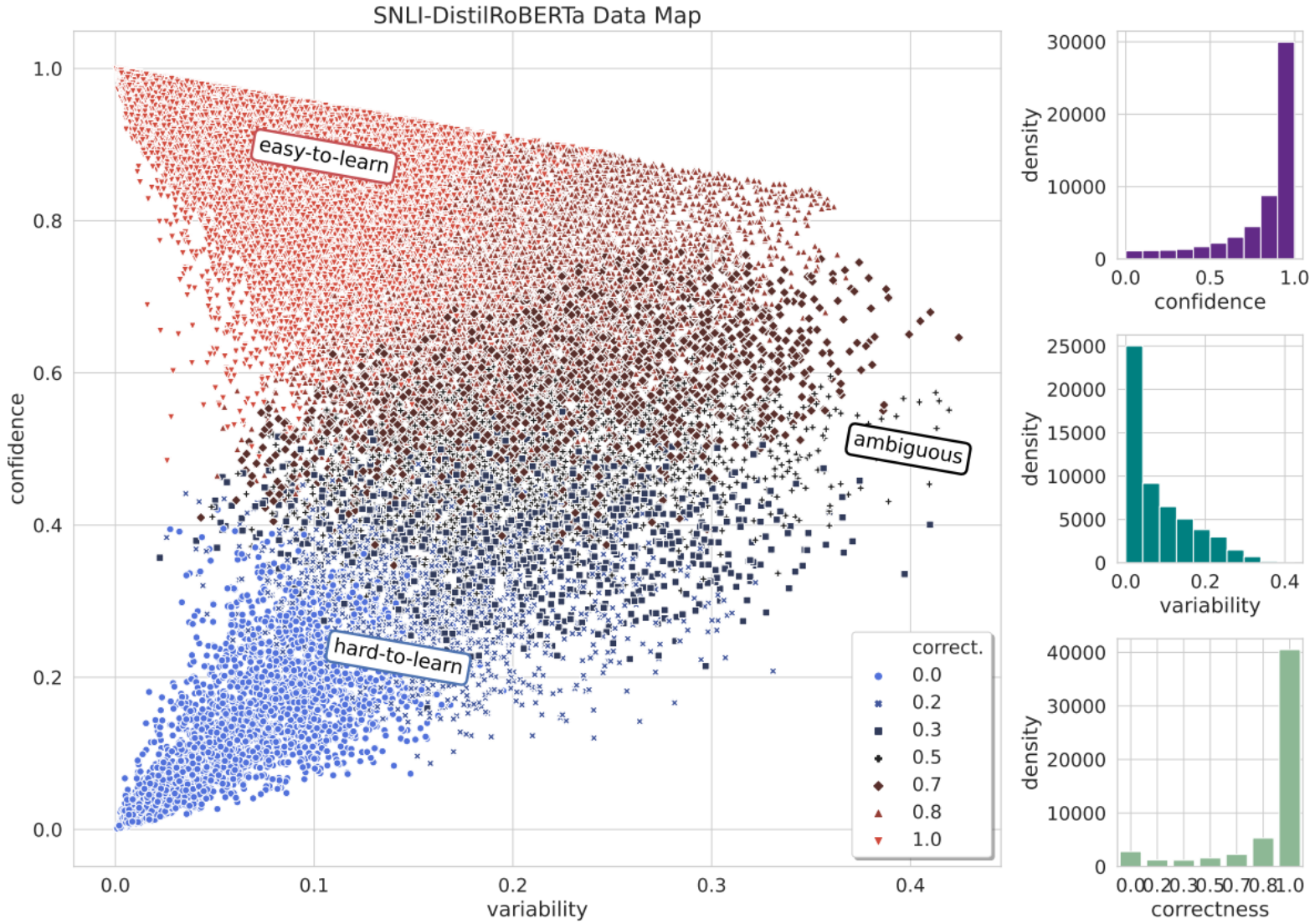
- Easy-to-learn
  - High confidence and low variability
- Hard-to-learn
  - Low confidence and low varability
- Ambiguous
  - High variability

Cloud Computing Solution

Google Cloud Platform

- Nvidia Tesla V100 GPU
- 4 vCPU w/ 15 GiB memory
- Can run for long periods without interruption

Data Map Example



Our Experiment

Goal

Minimize the size of a training dataset while maximizing accuracy using the dataset cartography

- Test models trained on each of the categories of data
- Find what mixes of subsets work best
- Compare different results for the most ”efficient” training dataset

Controls and Benchmarks

Controls

- Datasets are randomized right before training to remove any order bias
- Same number of data instances in everything but the baseline
- 6 Epoch training
- Same accuracy datasets for tests
- Use the % most of each category

Benchmarks

- 100% of the dataset randomized
- 33% of the dataset randomized

Data Subsets

- 100% Random
- 33% Random

- 33% Easy-to-learn
- 33% Hard-to-learn
- 33% Ambiguous

- 16% easy, 16% hard
- 16% easy, 16% ambiguous
- 16% hard, 16% ambiguous
- 11% easy, 11% hard, 11% ambiguous

Methods

1. Run training normally on 6 epochs, categorizing data
2. Filter training data to create appropriate subsets
3. Shuffle all sets to avoid order bias
4. Train a new model with each data subset
5. Evaluate performance with the out-of-distribution sets

Final Accuracies

	Final Training Accuracy (ID)	Dev Accuracy (OOD)	Test Accuracy (OOD)
100.00% random	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

Mislabeled Data

Hard-to-learn examples

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

Next Steps

- Other datasets
- Applications to other models (generative adversarial networks)
- Managing mislabeled samples

References

- 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.

Questions?