Hard and Rare Samples Estimation with Neural Network Based Models

Nikita Khoroshavtsev [1] **Sergey Karpukhin** [1] **Maksim Osipenko** [2] **Alexandra Volkova** [1] **Alexandr Voronin** [1]

Abstract

Samples from dataset are learned differently by deep learning model - some are easily captured by model, while other are ambiguous or hard. This hard and rare samples are very important for generalization of model. In order to differentiate such examples in an unsupervised setting, where data at hand doesn't have any labels, we aim to investigate various approaches and compare them with supervised metrics from prior works.

Github repo: https://github.com/Natifick/Unsupervisely-Hard

1. Introduction

There exists various approaches to estimate *hard*, *rare* and *mislabeled* samples in labeled data. Since data is labeled, in these approaches supervised learning is considered and learning dynamics of the models are investigated. Samples from dataset can be differentiated based of observed statistics - some of samples can be considered easy to learn while others are not learned at all or learned only at later stages - so called hard and rare samples. Identifying these is of great interest, as they play important role in generalization of the model.

However, such methods won't work when there is no labeled data at hand. In this work we try to investigate alternative approaches to estimating hard and rare samples. We consider two main settings - completely unsupervised metrics and semi-supervised, which rely on information from loss function, which can be applied in self-supervised learning settings. Details about planned work are provided in Section 3.

[1]Skolkovo Institute of Science and Technology, Moscow, Russia [2]Higher School of Economics, Moscow, Russia. Correspondence to: Sergey Karpukhin <sergey.karpukhin@skoltech.ru>.

Final Projects of the Machine Learning 2024 Course, Skoltech, Moscow, Russian Federation, 2024.

2. Literature Review

The article "Characterizing Datapoints via Second-Split Forgetting" (Maini et al., 2022) introduces the concept of second-split forgetting (SSFT) as a metric to evaluate the hardness and characteristics of data points in machine learning models. By training models on two disjoint splits of datasets and examining the forgetting patterns of examples in the second split, the authors demonstrate that SSFT can successfully identify mislabeled samples and distinguish between complex and simple examples. The metric has broad applicability beyond image datasets and can capture forgetting dynamics in sentiment classification tasks. Ablation experiments further investigate the notions of hardness, including mislabeled examples, complex examples, and rare examples. The results show that SSFT is effective in identifying mislabeled samples and has a low correlation with sub-group frequency. This makes it a valuable metric for understanding example utility and identifying label noise in datasets. The article also discusses the potential failure modes of machine learning models, such as relying on spurious features, and provides theoretical results to characterize the forgetting dynamics of mislabeled, rare, and complex examples. In conclusion, authors highlight the importance of SSFT in characterizing data points in machine learning models and its implications for dataset cleansing, identifying label noise, and improving model generalization.

The authors of the article (Swayamdipta et al., 2020) introduce their model-based graphical tool for characterizing datasets. The concept is the following: there are given four language datasets: *SNLI*, *MultiNLI*, *SQuAD* and *WinoGrande* and a language model *RoBERTa*, which is being trained on these datasets. The authors observe its training dynamics via two metrics: the model's confidence (the mean of probability to predict the right label for a particular example at the end of the epoch) and he variability of this confidence across epochs (just the standard deviation of it). It turns out that there are three main areas: ambiguous regions, which are highly connected to out-of-distribution (OOD) generalization, easy to learn regions and hard to learn ones, which often correspond to labeling errors. The **ambiguous area** samples have high variability, the **easy to learn** samples have high confidence and low variability and the **hard to learn** ones have low confidence and low variability. The visualization of areas is depicted on the figure

1. It culminates in multiple experiments, which expose such an interesting things like taking only 33% of the ambiguous data can outperform the baseline of 100% of the data and adding a little sample of **easy to learn** data can outperform even more.

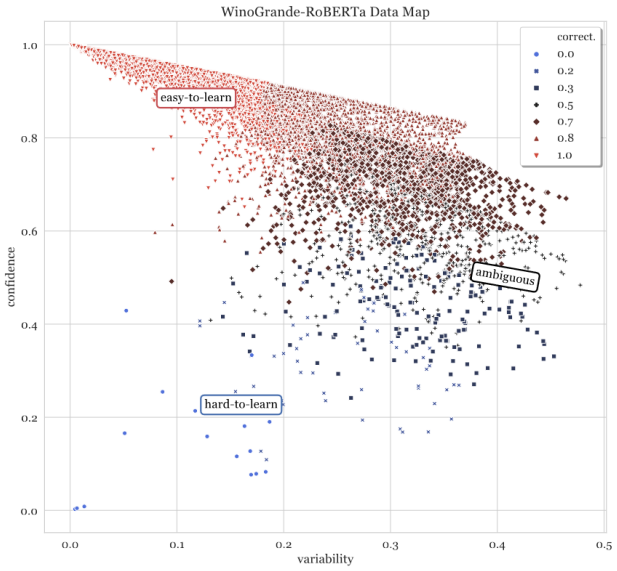

Figure 1. The distribution and classification of instances of one of the datasets according to the authors.

In article "Prioritized Training on Points that are Learnable, Worth Learning, and Not Yet Learnt" (Mindermann et al., 2022) authors introduce *Reducible Holdout Loss Selection (RHO-LOSS)* for models which are to work with web-scarped data. The problem of such data is that it is mostly noisy and ambiguous and training process takes a lot of time. To counter it, authors proposed a selection function grounded in probabilistic modelling that quantifies by how much each point would reduce the generalization loss if the model was to train on it, without actual training. We show that optimal points for reducing holdout loss are non-noisy, non-redundant, and task-relevant. To approximate this selection, reducible holdout loss was derived. Authors proved their theory and showed on multiple models and datasets that RHO-loss speeds up model convergence up to x30 by selecting *correct* points (which are s low-noisy, task-relevant and non-redundant). The choice of points is made by solving $argmax_{x,y}(L[y, x|D_t] - L[y, x|D_{ho}])$, where $L[y, x|D_t]$ is training loss (D_t - training set) and $L[y, x|D_{ho}]$ is irreducible holdout loss, as model is not trained on holdout set and this loss can not be reduced (D_{ho} - holdout set).

In other empirical study on learning abilities of neural networks (Toneva et al., 2018) authors introduce *forgetting event* - an effect when certain example transitions from being classified correctly to incorrectly during training. They conducted a lot of experiments with observing number of this events per example and were able to differentiate examples based of it: some examples were found to be unforgettable or prone to being forgotten while others were easily forgotten - these were found to be the ones with most uncommon features or noisy labels/noisy features. This unforgettable examples were stable across multiple tasks and also removing them from dataset allowed to still achieve great quality by neural network on target task. The results of study suggest that finding and differentiating examples through learning capability of model itself can be used to perform more efficient learning by removing large margin of unforgettable examples, detect noisy labels and obtain some information about intrinsic dimension of the problem.

There's also a work on Maximum Likelihood Estimation (MLE) of dimensionality of the intrinsic dimension (Levina & Bickel, 2004) of each point in the dataset. First - authors divide existing algorithms into two types: eigenvalue-based algorithms (e.g. PCA) and geometry-based algorithms (e.g. ISOMAP). In this paper the distribution of the data points is viewed as the Poisson random process. In this process the equation for the number of points in the sphere of fixed radius in known, and this equation depends on the dimensionality m of the inner space. Then authors make MLE estimation on the number of dimensions m, and reformulate the volume of this sphere not in terms of radius R, but in number of neighbors k. To make the estimation more robust, authors propose to average the results of the algorithm for different k.

Other closely related problem to investigating properties of samples is out-of-distribution (OOD) detection. The task is to differentiate ID data from OOD via relying on information from model. In series of works (Liang et al., 2017), (Huang et al., 2021) authors are solving this task by using information from gradient norms of the loss. They make crucial observation that gradient norm of loss can be used to differentiate between ID and OOD samples, especially when adding noise to data. Inspired by their work, we try to adopt similar approach to self-supervised setting.

3. Project Plan

In our project we plan to go through following steps:

1. Propose a set of metrics to find hard and rare samples, divided into two main categories:

   Loss-based, tested in self-supervised setting

   Completely unsupervised

2. Apply proposed metrics to CIFAR10, MNIST and ImageNet datasets

3. Apply the metrics described in (Maini et al., 2022; Swayamdipta et al., 2020; Toneva et al., 2018)
4. Compare the results of supervised metrics and proposed ones
5. Apply uncertainty estimation metrics and check how it is compared to:

   Supervised metrics

   Proposed metrics
6. Make conclusions about the applicability of the proposed metrics

4. Team member's planned contributions

Explicitly stated contributions of each team member to the final project.

Nikita Khoroshavtsev

- Review literature on the topic
- Metric design
- Implementation of experiments
- Set up github repo

Sergey Karpukhin

- Review literature on the topic
- Metric design
- Implementation of experiments
- Presentation preparation

Maksim Osipenko

- Review literature on the topic
- Investigate correlation between the designed metrics and the proposed ones in the literature
- Implementation of experiments
- Presentation preparation

Alexandra Volkova

- Review literature on the topic
- Implementation of experiments
- Presentation preparation

Alexandr Voronin

- Review literature on the topic
- Data preparation and exploratory data analysis
- Implementation of experiments

References

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. *CoRR*, abs/2110.00218, 2021. URL https://arxiv.org/abs/2110.00218.

Levina, E. and Bickel, P. J. Maximum likelihood estimation of intrinsic dimension. In *Neural Information Processing Systems*, 2004. URL https://api.semanticscholar.org/CorpusID:14865278.

Liang, S., Li, Y., and Srikant, R. Principled detection of out-of-distribution examples in neural networks. *CoRR*, abs/1706.02690, 2017. URL http://arxiv.org/abs/1706.02690.

Maini, P., Garg, S., Lipton, Z. C., and Kolter, J. Z. Characterizing datapoints via second-split forgetting, 2022.

Mindermann, S., Brauner, J., Razzak, M., Sharma, M., Kirsch, A., Xu, W., Höltgen, B., Gomez, A. N., Morisot, A., Farquhar, S., and Gal, Y. Prioritized training on points that are learnable, worth learning, and not yet learnt, 2022.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *CoRR*, abs/2009.10795, 2020. URL https://arxiv.org/abs/2009.10795.

Toneva, M., Sordoni, A., des Combes, R. T., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. *CoRR*, abs/1812.05159, 2018. URL http://arxiv.org/abs/1812.05159.