Robustness of Deep Learning Models to Dataset Errors: A Study Comparing LeNet and ResNet-18

Vadim CURCĂ

March 27, 2023

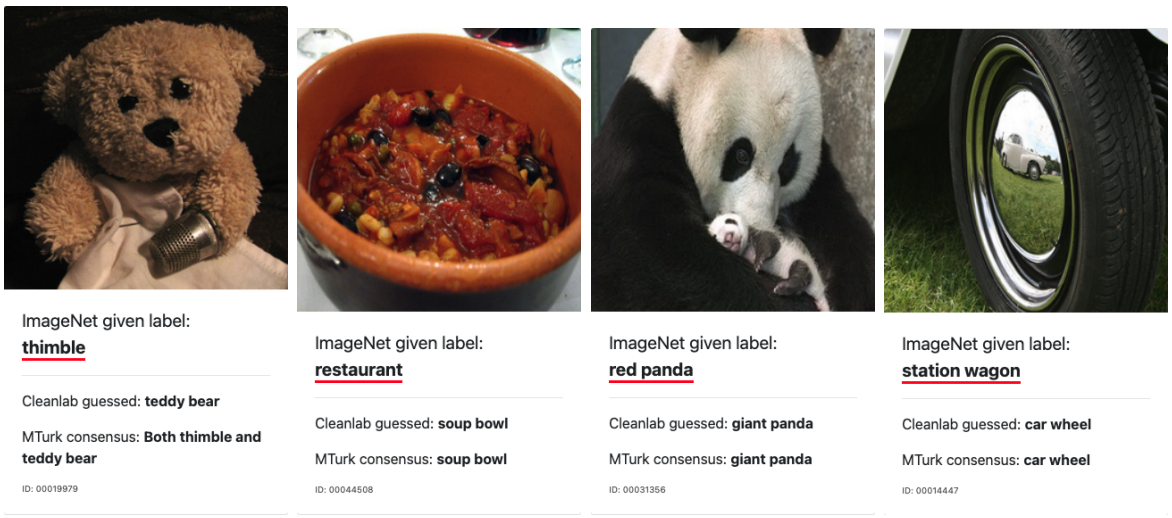
1 Introduction

In the field of machine learning, accurate data labeling is crucial for the success of a model. However, many real-world datasets contain label errors, which can significantly impact the performance of the model. Recent studies [1] have reported an average label error rate of 3.4% across various datasets, including widely-used datasets such as ImageNet [2] and MNIST [3]. These errors can lead to incorrect predictions and hinder the model's ability to learn meaningful patterns from the data. Therefore, it is essential to understand how the rate of errors in a dataset affects the final accuracy of a model and what makes a model more tolerant to dataset defects. In this project, we will analyze these factors and explore strategies for building models that are more robust to dataset errors.

2 Background

Dataset creation typically involves selecting a set of examples and assigning them a label that represents the ground truth. The process of labeling can be performed by domain experts or crowd workers, depending on the task's complexity and the availability of resources. In some cases, multiple annotators may label each example to ensure consistency and reduce the risk of errors. Despite the best efforts of annotators and quality assurance measures, label errors can still occur in the dataset creation process. There are several reasons for this. First, the task of labeling can be challenging, especially when the examples are complex or ambiguous. For instance, in an image classification task, distinguishing between similar-looking objects can be difficult even for humans. Additionally, annotators may have different interpretations of the task or may make mistakes due to fatigue or distraction.

In Figure 1 are some examples of such errors from widely used ImageNet dataset, some of them can be misinterpretations or the presence of multiple objects in the image and others are clearly the result of human mistakes.



(a) (b) (c) (d)

Figure 1: Label errors in ImageNet dataset, source [4].

3 Experiments

In this project I analyze how the error rate in the dataset influences the training, validation and test accuracy. I use Fashion-MNIST [5] dataset, and I assume it has label error rate of 0%. The entire dataset of 70,000 images is divided into 60,000 images used for train and validation and 10,000 images used for test. Each experiment iteration for each model consists of randomly modifying a defined percent of the train dataset, with random labels, each experiment iteration is completely independent of any other. Throughout all the experiments the test part remains unchanged and unaltered, simulating how the model would behave on real data.

LeNet [6]. First experiment was conducted on LeNet, a relatively small CNN by today's standards, which achieved an accuracy of 79.77% trained on the unaltered dataset. As I introduced errors into the dataset with 10% granularity, I observed a linear decrease in both the testing and validation accuracies. However, until the error rate reached 50%, the testing accuracy degraded at a much slower pace, suggesting that the network was still able to learn from predominantly correct examples in the training dataset. The next interval of interest was between 60% and 80% error rate, where the fact that the training accuracy was higher than the testing accuracy suggests that the network was simply memorizing images without deducing patterns between them. Figure 2.

ResNet [7]. The next stage was to analyze how the complexity of the model affects the robustness to dataset errors. For this I used ResNet-18 architecture, slightly modified to match the requirements of the current dataset, using 1 input channel and 10 output classes, instead of 3 input channels and 1000 output classes in the original version. Being a much more complex architecture it achieved an accuracy of 91.34% trained on the unaltered dataset.

Interestingly, this network demonstrated significantly better robustness to errors, even though the training and validation accuracies decreased at the same rate. The testing accuracy decreased at a much slower pace compared to the previous experiment,

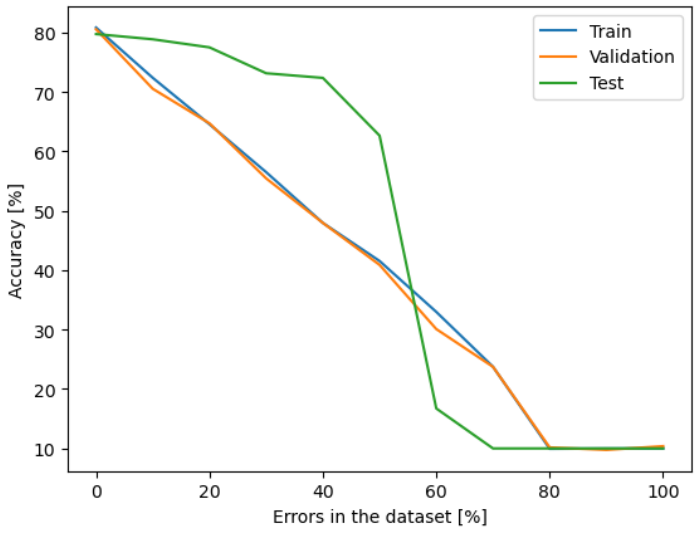


Figure 2: LeNet, impact of train and validation dataset errors on accuracy.

managing to maintain a remarkable accuracy even up to 80% of errors in the dataset. This finding suggests that the model requires fewer examples where the pattern and label matches, to learn effectively, and is less sensitive to cases where a particular pattern is associated with random labels. Figure 3.

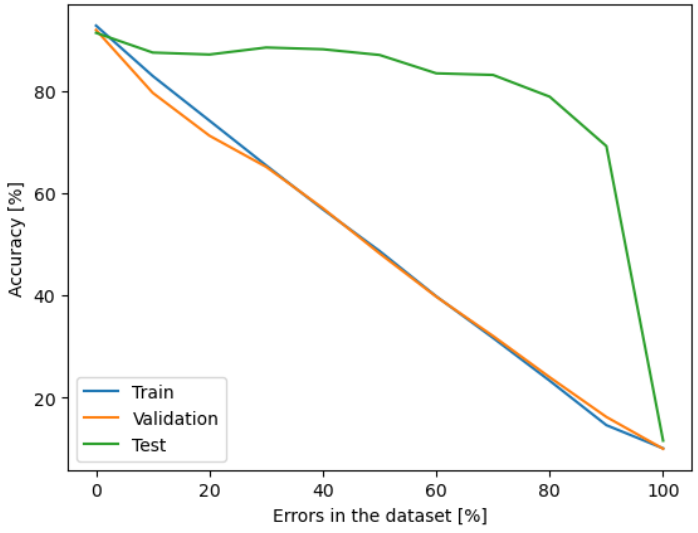


Figure 3: ResNet, impact of train and validation dataset errors on accuracy.

ResNet pretrained on ImageNet. The third experiment was testing whether pretraining the network on a complex dataset would improve its robustness to errors when fine-tuned on a dataset with errors. For this purpose, I used ResNet-18 as in the previous experiment, but initiated it with weights pretrained on ImageNet, with only the first and last layers reset. The fact that the network already had learned many complex patterns allowed fine-tuning it on the clean dataset to achieve a testing accuracy of 92.23%.

Similarly, as errors were introduced into the training dataset, the network had

a slightly smaller degradation in testing accuracy compared to the non-pretrained network. This finding suggests that pretraining on a large and diverse dataset may improve a network's ability to tolerate errors and generalize to new data. Figure 4.

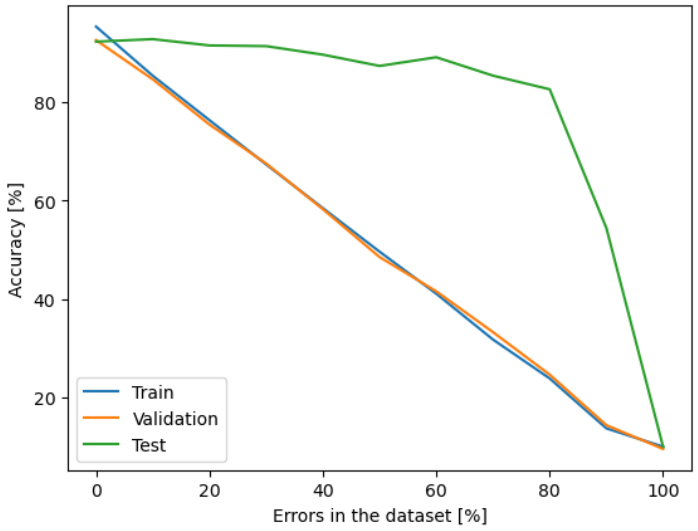


Figure 4: ResNet pretrained on ImageNet, impact of train and validation dataset errors on accuracy.

4 Conclusion

These experiments have shown that ResNet-18 is much more robust to dataset errors than LeNet, and that pretraining the network can further improve its accuracy and, to a small extent, its robustness to errors. Surprisingly, the models were able to maintain remarkable accuracy even when the percentage of errors in the dataset was very high, exhibiting a higher testing accuracy than training accuracy, which is normally unusual.

While I did not continue the experiments on larger models due to time constraints, a recent article reports: "Surprisingly, we find that lower capacity models may be practically more useful than higher capacity models in real-world datasets with high proportions of erroneously labeled data. For example, on ImageNet with corrected labels: ResNet-18 outperforms ResNet-50 if the prevalence of originally mislabeled test examples increases by just 6%." [8]. Therefore the idea that more complex models are more robust to dataset errors is not generally applicable.

Currently, this topic is of great interest, with studies [8, 9, 10, 11] attempting to estimate and identify errors in datasets using more architecture agnostic techniques that are not dependent on the specific type of network used. As the development and deployment of deep learning models in real-world scenarios becomes increasingly common, improving their robustness to errors is critical for ensuring their reliability and effectiveness.

References

[1] "Major ML datasets have tens of thousands of errors | MIT CSAIL." [Online]. Available: https://www.csail.mit.edu/news/major-ml-datasets-have-tens-thousands-errors

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," Jan. 2015, arXiv:1409.0575 [cs]. [Online]. Available: http://arxiv.org/abs/1409.0575

[3] "MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges." [Online]. Available: http://yann.lecun.com/exdb/mnist/

[4] "Label Errors in Benchmark ML Datasets." [Online]. Available: https://labelerrors.com/

[5] "Fashion-MNIST," Mar. 2023, original-date: 2017-08-25T12:05:15Z. [Online]. Available: https://github.com/zalandoresearch/fashion-mnist

[6] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, conference Name: Proceedings of the IEEE.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, arXiv:1512.03385 [cs]. [Online]. Available: http://arxiv.org/abs/1512.03385

[8] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks," Nov. 2021, arXiv:2103.14749 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2103.14749

[9] C. Northcutt, L. Jiang, and I. Chuang, "Confident Learning: Estimating Uncertainty in Dataset Labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, Apr. 2021. [Online]. Available: https://www.jair.org/index.php/jair/article/view/12125

[10] A. Thyagarajan, E. Snorrason, C. Northcutt, and J. Mueller, "Identifying Incorrect Annotations in Multi-Label Classification Data," Nov. 2022, arXiv:2211.13895 [cs]. [Online]. Available: http://arxiv.org/abs/2211.13895

[11] "cleanlab/cleanlab: The standard data-centric AI package for data quality and machine learning with messy, real-world data and labels." [Online]. Available: https://github.com/cleanlab/cleanlab