Cold Case: The Lost MNIST Digits

Chhavi Yadav and Léon Bottou

Summary by Sri Datta, 9th October 2019 (budaraju@kth.se)

Introduction

The authors of the paper recreate the preprocessing steps used in creating MNIST (LeCun et al.) dataset from NIST (Grother and Hanaoka, 1995) dataset and reconstruct a very accurate replica that could serve as a substitute with insignificant changes in the accuracy. This gave birth to an opportunity to evaluate and learn if the impressive accuracies of the neural networks on MNIST over the past two decades are a result of model overfitting.

Similar work was conducted by [Recht et al., 2018, 2019] and discovered a drop in the accuracies when the authors tested the models on a new but very close test dataset. The same trends have been observed by the author of this paper when they test models on the never before released 50,000 test images showing that the models were somehow overfitted to the original 10,000 test images.

Recreating MNIST

The authors find the replication challenging as they found some flaws in the presented metrics along with missing descriptions of a few preprocessing steps. The authors also make use of a few snippets that they discovered in the Lush codebase which gave further insights into the resampling algorithm. After numerous experiments and versions of the QMNIST, they carefully analyze the mismatches to tune the cropping algorithm and improve the replica. However, despite their iterative reconstruction, QMNIST remains imperfect.

From the quartiles of the L_2 and L_∞ distance between QMNIST and MNIST, it is discovered that 0.25% of the QMNIST images are shifted by one pixel relative to MNIST as shown in table 1. In addition to this, they also train a variant of Lenet5 (Le Cun et al. 1998) to check the quality of the replication, and the results have shown very little difference as shown in table 2. This exhaustive work sheds light on a list of previously unknown facts about MNIST. One of the important observation is the average center of mass of MNIST digits are half a pixel shifted to that of the geometric center of the image. Disregarding this in creating a new test set would obviously decrease the performance and would negatively affect the entire purpose of the paper in answering the question if models are overfitted to MNIST test images.

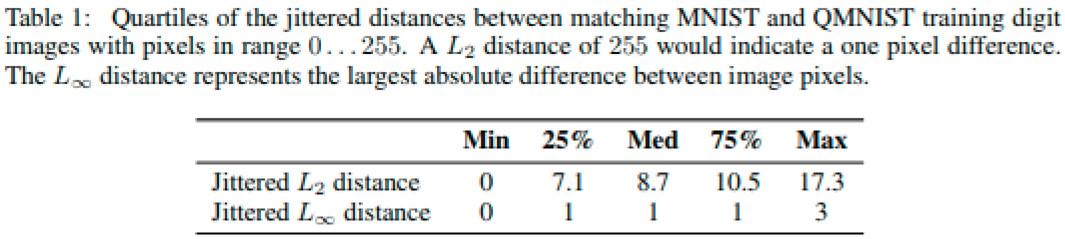

Table 1: Quartiles of the jittered distances between matching MNIST and QMNIST training digit images with pixels in range $0 \dots 255$. A L_2 distance of 255 would indicate a one pixel difference. The L_∞ distance represents the largest absolute difference between image pixels.

	Min	25%	Med	75%	Max
Jittered L_2 distance	0	7.1	8.7	10.5	17.3
Jittered L_∞ distance	0	1	1	1	3

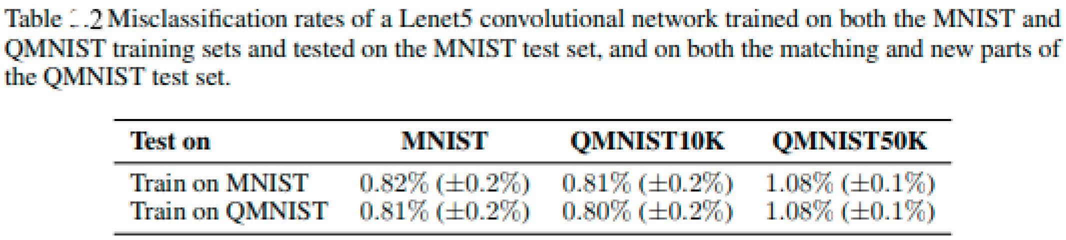

Table 1.2 Misclassification rates of a Lenet5 convolutional network trained on both the MNIST and QMNIST training sets and tested on the MNIST test set, and on both the matching and new parts of the QMNIST test set.

Test on	MNIST	QMNIST10K	QMNIST50K
Train on MNIST	0.82% (±0.2%)	0.81% (±0.2%)	1.08% (±0.1%)
Train on QMNIST	0.81% (±0.2%)	0.80% (±0.2%)	1.08% (±0.1%)

Generalization Experiments

The generalization experiments on MNIST include various methods like KNN, SVM, MLP, and CNN so as to replicate the results reported by Le Cun et al. [1998]. The authors use the Wald method for the confidence intervals on the error rates.

The paper reports results of a wide array of experiments conducted on various combinations of 2 train datasets i.e MNIST and QMNIST and 3 test datasets MNIST10K, QMNIST10K, QMNIST50K. Apart from the steps in creating QMNIST, there is no pre-processing or data augmentation done to the images before training the models. The summary of the results of all methods in the experiment is as illustrated in figure 1. We can observe two important trends, a consistent increase in the error rates when tested on QMNIST and the same ordering among the classifiers.

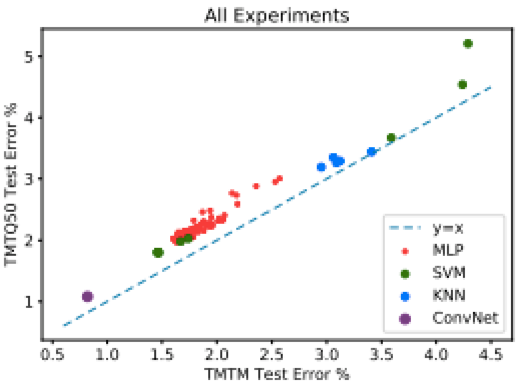

Figure 1. Scatter plot comparing the MNIST and QMNIST50 testing performance of all the model trained on MNIST during the course of this study (Source: Paper)

TakeAway

There are two major takeaways from this paper, the first being that there exists overfitting of published models to test data, and the second is the issue of reproducibility.

Observing the same phenomena on two different datasets confirms that there in fact is a problem of “test set rot” but is not very high to be concerned about.

This paper could be taken as the best example to show the importance of publishing reproducible papers. The authors find a few critical pieces of information missing which required to perfectly replicate the experiments. We have to note that one of the authors of this paper is also the author of the original dataset. If this happens in one of the most popular datasets from one of the elites in the field, the quality of the huge influx of deep learning papers remains a big question.

Strong Points

Though Recht et al., 2018, 2019 give useful insights into this issue, it is hard to evaluate the quality/similarity of the new test images. But this paper has a huge advantage as MNIST did not release 50,000 test images initially, guaranteeing that the newly introduced QMNIST test images are almost similar to the ones the research community was using to benchmark machine learning algorithms. The exhaustive investigation and experimentation to rediscover the pre-processing steps, and the importance is given to make QMNIST as identical as possible to MNIST is to be appreciated.

Weak Points

Having said that, the experiments to replicate the preprocessing algorithm are commendable, I find the experiments in comparison to QMNIST to be very disappointing. The authors trying to address “test set rot” and questions like “ Did they overfit the test set? and How quickly do machine learning datasets become useless?”, they only compare QMNIST and MNIST on SVM, MLP, KNN, CNN as reported by Le Cun et al. [1998] and completely misses to evaluate on modern architectures. Shouldn’t that be of primary focus, if creating QMNIST was not the only goal? For that reason, I think this work is not enough to make any strong conclusions about “test set rot”.

References

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The MNIST database of handwritten digits.http://yann.lecun.com/exdb/mnist/, 1994. MNIST was created in 1994 and released in 1998.

Patrick J. Grother and Kayee K. Hanaoka. NIST Special Database 19: Handprinted forms and characters database.https://www.nist.gov/srd/nist-special-database-19,1995.
SD1 was released in 1990, SD3 and SD7 in 1992, SD19 in 1995, SD19 2nd edition in2016.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10?arXiv preprint arXiv:1806.00451, 2018.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet?arXiv preprint arXiv:1902.10811, 2019.

Yann Le Cun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of IEEE, 86(11):2278–2324, 1998.