

Cold Case : The Lost MNIST Digits

The Sherlocks:
Chhavi Yadav NYU
Léon Bottou FAIR, NYU



What about MNIST?

- MNIST is a subset of NIST [1]
- Original MNIST Testing set - 60K digits
- Was chopped off to 10K digits before further preprocessing

The original NIST test contains 58,527 digit images written by 500 different writers. In contrast to the training set, where blocks of data from each writer appeared in sequence, the data in the NIST test set is scrambled. Writer identities for the test set is available and we used this information to unscramble the writers. We then split this NIST test set in two: characters written by the first 250 writers went into our new training set. The remaining 250 writers were placed in our test set. Thus we had two sets with nearly 30,000 examples each.

The new training set was completed with enough samples from the old NIST training set, starting at pattern #0, to make a full set of 60,000 training patterns. Similarly, the new test set was completed with old training examples starting at pattern #35,000 to make a full set with 60,000 test patterns. All the images were size normalized to fit in a 20 x 20 pixel box, and were then centered to fit in a 28 x 28 image using center of gravity. Grayscale pixel values were used to reduce the effects of aliasing. These are the training and test sets used in the benchmarks described in this paper. In this paper, we will call them the MNIST data.

Fig. 1 [2]

This is all the information we have about how MNIST was created!!

How did we reconstruct MNIST?

- Using description on previous slide & a resampling algorithm found in an ancient Lush codebase^a
- Hungarian matching algorithm(only training set)
- Inspection of the worst matched
- Fine tuning of algorithms



^a See <https://tinyurl.com/y5z7qtcg>

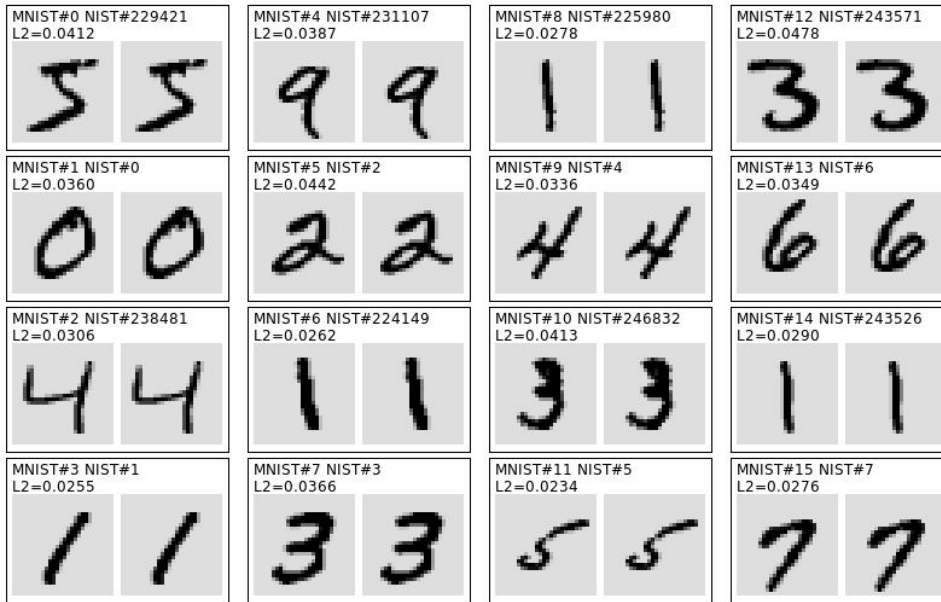


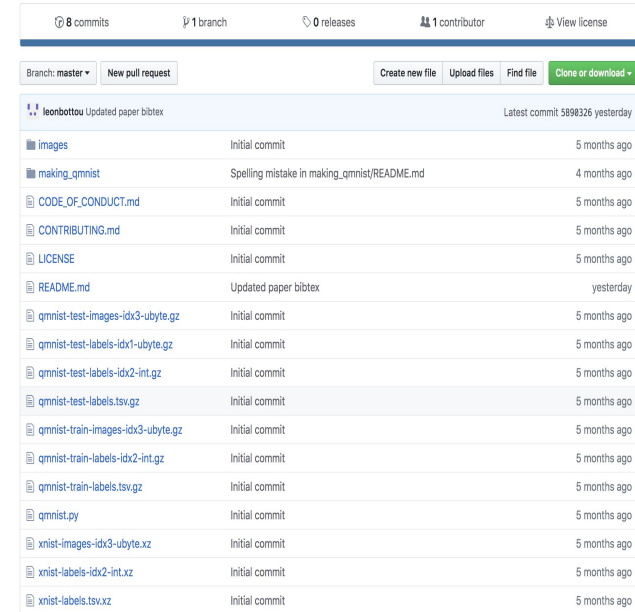
Fig. 2 Side-by-side display of the first sixteen digits in the MNIST and QMNIST training set.

Why use QMNIST?

- **QMNIST Test Set = 6x MNIST Test set!!**
- **Metadata like writer id, partition id**
- Download from

<https://github.com/facebookresearch/qmnist>

The QMNIST dataset



The screenshot shows the GitHub repository page for 'facebookresearch/qmnist'. At the top, it displays '8 commits', '1 branch', '0 releases', '1 contributor', and 'View license'. Below this, there are buttons for 'Branch: master', 'New pull request', 'Create new file', 'Upload files', 'Find file', and 'Clone or download'. The main content is a list of files and folders with their commit history. The latest commit is '5898326 yesterday' by 'leonbotou' with the message 'Updated paper bibtext'.

File/Folder	Commit Message	Time Ago
leonbotou	Updated paper bibtext	Latest commit 5898326 yesterday
images	Initial commit	5 months ago
making_qmnist	Spelling mistake in making_qmnist/README.md	4 months ago
CODE_OF_CONDUCT.md	Initial commit	5 months ago
CONTRIBUTING.md	Initial commit	5 months ago
LICENSE	Initial commit	5 months ago
README.md	Updated paper bibtext	yesterday
qmnist-test-images-idx3-ubyte.gz	Initial commit	5 months ago
qmnist-test-labels-idx1-ubyte.gz	Initial commit	5 months ago
qmnist-test-labels-idx2-int.gz	Initial commit	5 months ago
qmnist-test-labels.tsv.gz	Initial commit	5 months ago
qmnist-train-images-idx3-ubyte.gz	Initial commit	5 months ago
qmnist-train-labels-idx2-int.gz	Initial commit	5 months ago
qmnist-train-labels.tsv.gz	Initial commit	5 months ago
qmnist.py	Initial commit	5 months ago
xnist-images-idx3-ubyte.xz	Initial commit	5 months ago
xnist-labels-idx2-int.xz	Initial commit	5 months ago
xnist-labels.tsv.xz	Initial commit	5 months ago

Overfitting on MNIST?

- Since MNIST has been around for a quarter century, many researchers doubt that the immense experimentation has led to overfitting on MNIST.
- Tested previous classifiers with 50K new samples in QMNIST Test set.

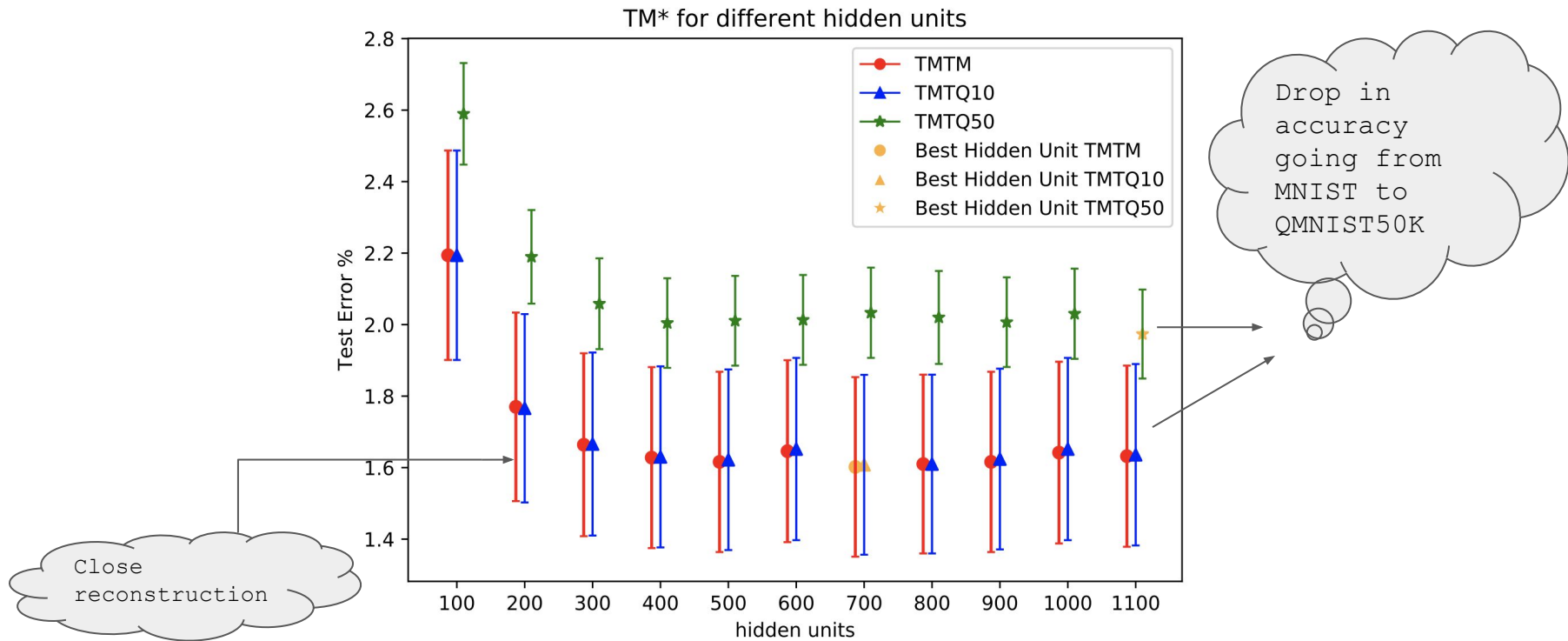
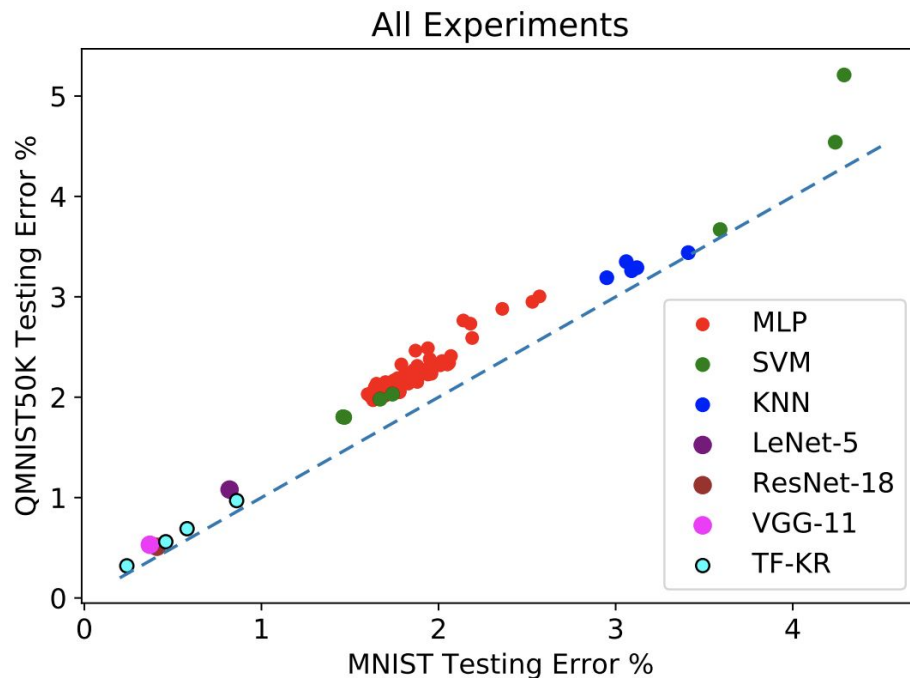


Fig. 3 MLP error rates for various hidden layer sizes after training on MNIST & testing on MNIST, QMNIST10K & QMNIST50K



Consistent drop in accuracy going from MNIST to QMNIST50K

Fig. 4: Scatter plot comparing the MNIST and QMNIST50K testing performance of all the models trained on MNIST during the course of this study.

Conclusion

- “Testing Set Rot” exists but is far less severe than feared
- Confirms trends observed by Recht et al. [3, 4] - on a different dataset & substantially controlled setup
- In practice, this suggests that a shifting data distribution is far more dangerous than overusing an adequately distributed testing set

References

- [1]Patrick J. Grother and Kayee K. Hanaoka NIST Special Database 19:
Handprinted Forms and Characters Database 1990
- [2]Bottou, Léon et. al. Comparison of classifier methods: a case study in
handwritten digit recognition 1994
- [3]Recht, Benjamin et. al. Do CIFAR-10 Classifiers Generalize to CIFAR-10?
2018
- [4]Recht, Benjamin et. al. Do ImageNet Classifiers Generalize to ImageNet?
2019

. . Thank you . .