



A METHOD OF DATA LABEL CHECKING AND THE WRONG LABELS IN NIST
SD19V2 AND MNIST

A PROOF OF 'BAD' HANDWRITING IN MNIST TRAINING DATASET MAKING
CNN TO PREDICT 'GOOD' HANDWRITING WRONG

DeepAI

Abner.Zhang Steven Lu

WHY 99% ACCURACY CNN ARE WRONG AT GOOD HANDWRITING

Bad handwriting



Good handwriting



convnetjs



Karpathy

Possible Reason: 'bad' handwriting in training dataset

- SD-3 Census Bureau employees
- SD-1 high-school students
- SD-3 is much cleaner and easier to recognize than SD-1
- The MNIST training set is composed of 30,000 patterns from SD-3 and 30,000 patterns from SD-1



Yann LeCun

DATA: NIST SD19V2 WITH ID AND ORIGINAL FORM METHOD: 360-DEGREE PERFORMANCE EVALUATION

CNN 1-5 training data
From high school

CNN 6-10 training data
From employee

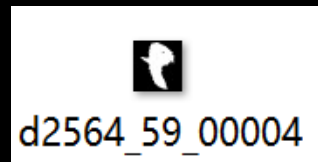


Mixed group

15 CNNs predict
100,000 digits of
905 writers

RESULT OF DIGITS OF WRITER F2564

	digits	predby 1	predby 2	predby 3	predby 4	predby 5	predby 6	predby 7	predby 8	predby 9	predby 10	predby 16	predby 27	predby 38	predby 49	predby 510
d2564_59_00004. png	1	1	8	8	8	8	8	9	9	9	8	1	8	8	8	8



HANDWRITING SAMPLE FORM

This sample of handwriting is being collected for use in testing computer recognition of hand printed numbers and letters. Please print the following characters in the boxes that appear below.

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9

83 310 4191 18789 464301

83 310 4191 18789 464301

814 9046 78523 092351 07

814 9046 78523 092351 07

9522 09260 669062 25 277

9522 09260 669062 25 277

13244 163828 79 570 4375

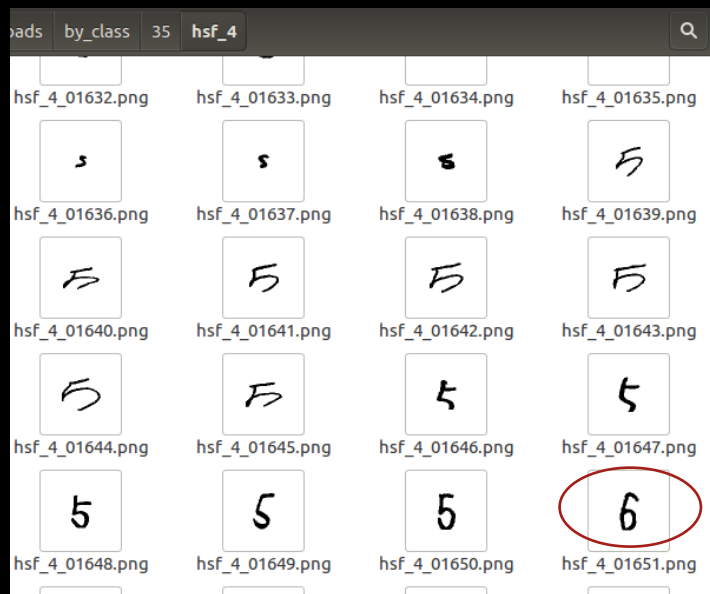
13244 163828 79 570 4375

389947 55 635 7064 56818

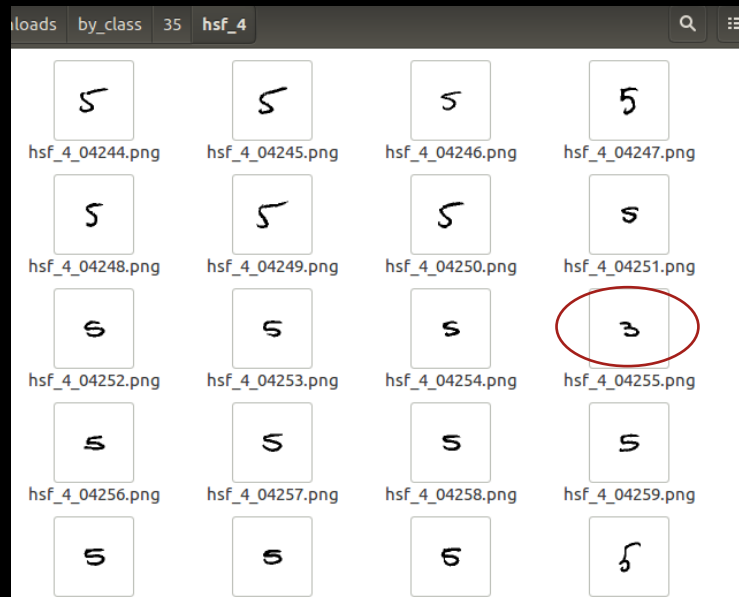
389947 55 635 7064 56818

original form

'BAD' HANDWRITING: WRONG LABELED

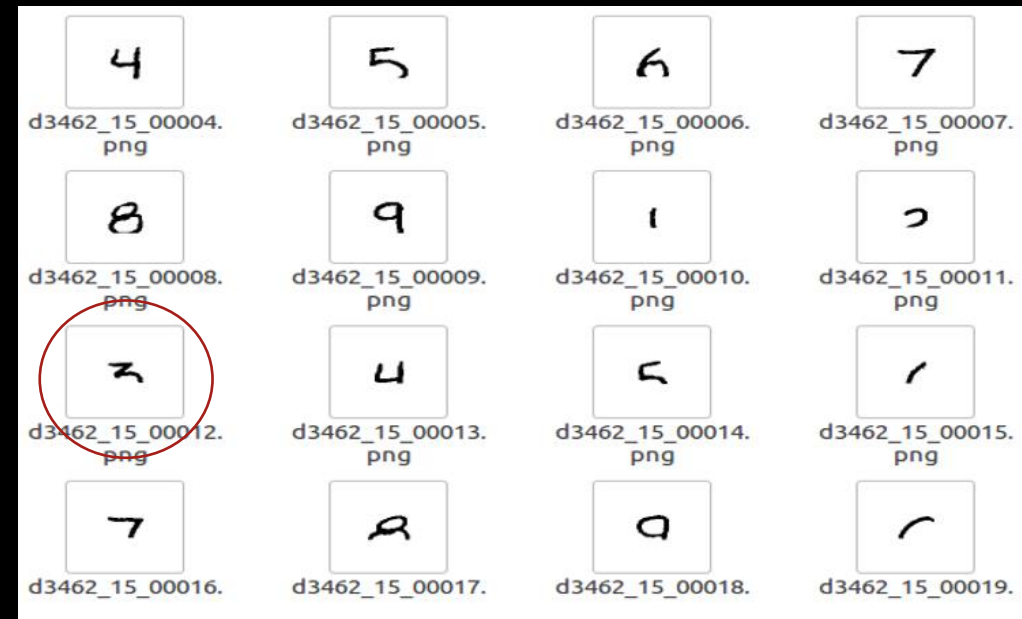
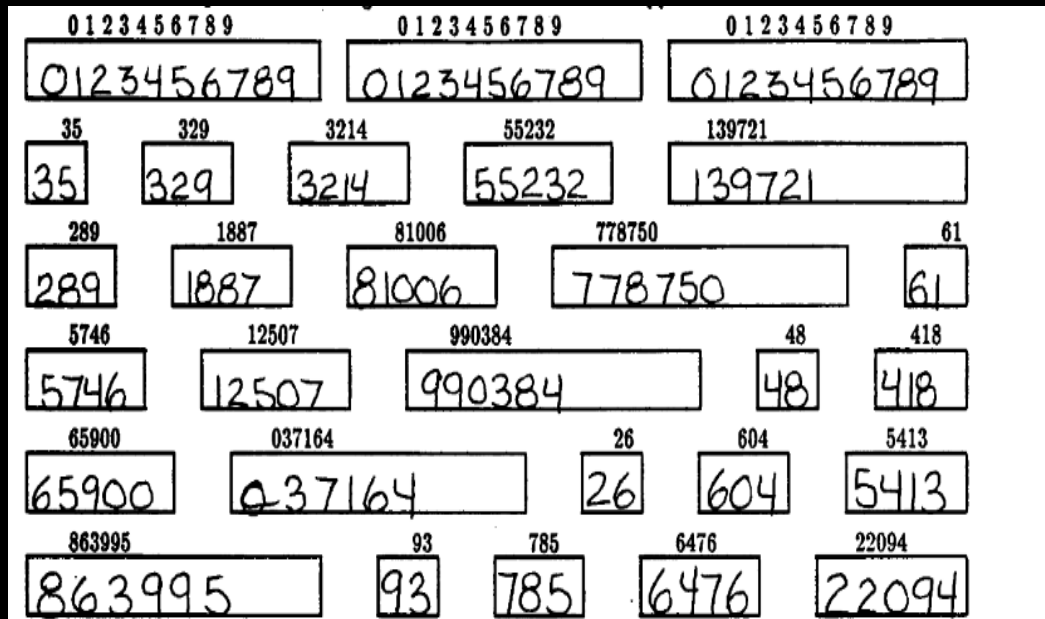


SD19 V2



MNIST

'BAD' HANDWRITING: WRONG CUT



'BAD' HANDWRITING: MESSED

0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9
45	819	3722
45	819	3722
061	6914	38183
061	6914	38183
2024	36007	267812
2024	36007	267812
82281	207557	90
82281	207557	90
945879	07	045
945879	07	045
		3654
		3654
		02445
		02445

'GOOD' HANDWRITING DATA

0 1 2 3 4 5 6 7 8 9

0123456789

93 015 7433 71867 627126

862 4389 70104 990086 30

0309 73828 935502 64 864

92458 321145 71 749 6547

269759 85 618 2214 05531

0 1 2 3 4 5 6 7 8 9

0123456789

23 305 6001 85467 723403

521 0842 57996 530367 28

4598 77182 999894 15 625

23791 608854 33 191 8604

181714 32 660 9426 47750

PROOF

	CNN learn only 'good'	CNN also learn 'bad'
'good' test data	5 (99.9%)	22(99.5%)

It approves that after the CNN has learnt more 31570 'bad' training data, it make 17 more errors at 'good' test data than the CNN only learnt 'good' handwriting.

<http://ssrn.com/abstract=3056117>

<http://ssrn.com/abstract=3049684>

WHAT'S NEXT?

- To build a 360-degree evaluation method in Keras to relabel data
- To check wrong labeled or cut digits pictures in SD19v2
- To find out why there is still 5 errors
- 99.9% accuracy is better than 100%,



- Thanks