# Performance of pre-learned convolution neural networks applied to recognition of overlapping digits

1st Daigo Shii
*Graduate School of Industrial Technology, Nihon University*
Chiba, 275–8575 Japan
cida18007@g.nihon-u.ac.jp

2nd Ryosuke Miyoshi
*Graduate School of Industrial Technology, Nihon University*
Chiba, 275–8575 Japan
ciry18013@g.nihon-u.ac.jp

3rd Kazuyuki Hara
*College of Industrial Technology Nihon University*
Chiba, 275–8575 Japan
hara.kazuyuki@nihon-u.ac.jp

***Abstract*—We analyzed the performance of pre-learned convolution neural networks (CNN) learned with a single-digit image dataset when they were used to recognize images containing two overlapping digits. The pre-learned network was learned using the MNIST database, and the network architecture was the LeCun network. The overlapping digit images were made using images from the MNIST database.**

**Our goal was to clarify two issues: (1) can a network learned for recognition of single digits in an image classify an image that includes two overlapping digits without doing additional learning using a dataset composed of two overlapping digit images? (2) Is a convolutional neural network (CNN) capable of processing stereoscopic vision? If (1) is possible with a CNN, then we don't need to train a huge number of images that include various combinations of more than two digits. If (2) is possible, stereoscopic vision is also possible without having to learn overlapping images. Our results support the conclusion that stereoscopic vision or a similar function is involved in the CNN.**

***Index Terms*—overlapping digits recognition, pre-learned CNN, stereoscopic view, MNIST dataset**

## I. Introduction

Convolution neural networks (CNN), as proposed by Lecun et al. [1], are a kind of deep neural network. CNNs include convolution filters that are especially useful for extracting features of input images, and they have achieved state-of-art results in the fields of image processing and acoustic processing [2]–[4]. Auto-encoders [5] are another type of deep neural network that differ from CNNs in that they do not explicitly include filters.

A CNN has a pre-processing section and a classification section. The pre-processing section has two functions that are similar to those of the early visual cortex in humans. One is a convolution layer that distinguishes or filters features such as similar lines and curves from the input image. The other is a pooling layer that abstracts the previous layer's output to allow for shifting or shrinking of the objects in the image. In the convolution layer, the coefficients that form the filter are treated as weights, and these weights can be adapted to data through learning. That is, CNNs can adapt the filters in their convolution layers to learning data.

On the other hand, pre-learned CNNs include filters for extracting the features of the general image processing, so they can be used in various image classification tasks without having to conduct learning for each task. For instance, an image may include several objects, and normally, humans don't have to learn all combinations of objects before they can see them in an image. In the same way, a CNN can learn the digits from 0 to 9 and can recognize their combinations, e.g., 19, 34, or 21, without additional learning.

In this study, we utilized a network that was pre-learned on the MNIST database [6] containing images of individual digits. We applied it to the task of classifying images that include two overlapping digits. In this task, the pre-learned network must be able to classify images from the MNIST database of only one digit as well as images containing two overlapping digits. We made the input images with overlapping digits. To make these image, the whole digit at the front must be visible, but part of the back digit must be occluded by the front digit. This overlapping two-digit recognition task can thus be regarded as one of stereoscopic viewing. Thus, if the accuracy of recognizing overlapping digits is higher than that of a plane image, this may be evidence that a pre-learned CNN can perform a kind of stereoscopic processing. Moreover, if the pre-learned CNN can recognize both digits, it means that the pre-processing section of the CNN can extract the features of each digit.

## II. Model

### A. MNIST database and overlapping digits

We used the MNIST database as the learning data to learn the pre-learned CNN. The MNIST database consists of images of single digits from 0 to 9, and the label shows which digit the image includes. It has 60000 images for learning and 10000 for testing. The test data are for cross-validation to evaluate the performance of the pre-learned CNN. Each image was a 28-by-28-dot gray-scale image.

There are two ways of placing two digits in an image: in the first way, the back image appears occluded by the front image at the overlap, while in the second way, both digits appear

2375-9356/20/$31.00 ©2020 IEEE
DOI 10.1109/BigComp48618.2020.00-90

to share the pixels in the overlapping area. Figure 1 shows examples of images with the digits ”1” on ”3” overlapping. The MNIST images were used to make them. In ”one digit on top of the other” image, “1” appears to be on top of “3”, and the crossing section of the back digit is hidden. Figure 2 shows all of the ”one digit on top of the other” images. In this figure, the first row of the back digit is 0, the second row of the back digit is 1, and so on. There are 90 images in total. We used them to investigate the effect of the hidden part of the back digit on digit recognition.

Figure 3 shows all the images of ”two digits on the same level”. In each image, the crossing section of back digit does not appear to be hidden by the digit above it. We used these to compare the performance of the CNN at recognizing ”two digits on the same level” with that of ”one digit on top of the other” to clarify the effect of hidden cross section in recognition of digits. These images were made from MNIST images.

Fig. 1. Examples of digit ”1” on top of digit ”3” (left) and two digits on the same level (right).

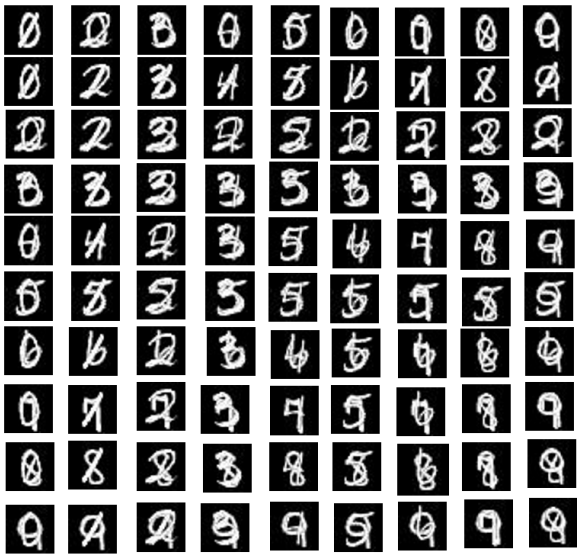

Fig. 2. Images of ”one digit on top of the other” used for recognition.

### B. Pre-learned network

The pre-learned network was learned using the original lenet_train_text.prototxt included in the sample Caffe framework [7]. The learning rate was 0.01, the momentum was 0.9, and the weight decay was 0.0005. The network had an input layer, followed by a convolution layer, a pooling layer, another convolution layer, another pooling layer, and two fully connected layers. The output function of the first fully connected layer was ReLU, and the output function of the second fully connected layer was soft-max.

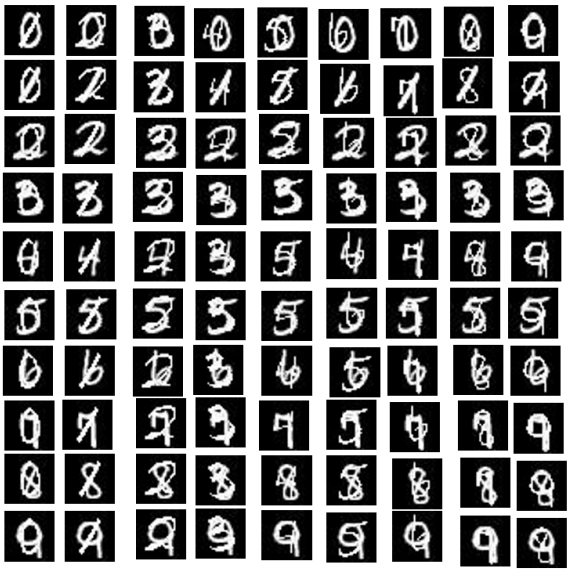

Fig. 3. Images of ”two digits on the same level” used for recognition.

Learning conditions were the same as in the original lenet_solver.prototxt. A GPU was not utilized. The average test accuracy for the test dataset was 99.0%, and the multinomial logistic loss was 0.03.

## III. Results

The CNN was pre-learned on single-digit images; i.e., it was not learned at all with the two-digit images. The soft-max output function of the last layer gave the classification probability of the digit in the input image. To classify two digits, we took the first candidate to be the output with the maximum probability, the second candidate as that with the second highest probability, and the third candidate as that with the third highest probability. We didn’t use candidates with probabilities less than 4th highest.

### A. One digit on top of the other

In these images (see Fig. 1 left), all of the front digit can be seen, but the cross section of the back digit can not be seen.

- Recognition of one of two digits
  We investigated the recognition rate of the CNN pre-learned on one-digit images on the ”one digit on top the other” images. The recognition rate was evaluated for (1) the front digit, (2) back digit, (3) and total of (1) and (2). Each case included three conditions: (a) the first candidate, (b) the first or second candidate, (c) the first, second, or third candidate. Table I shows the results. The results in the ”1st” column correspond to the percentage of condition (a), those in ”1st or 2nd” correspond to the percentage of condition (b), and those in ”1st, 2nd or 3rd” column correspond to the percentage of condition (c).

TABLE I
RESULTS OF RECOGNITION OF ONE OF TWO DIGITS

| | (1st) | (1st or 2nd) | (1st, 2nd or 3rd) |
|---|---|---|---|
| front digit | 19 % | 26 % | 29% |
| back digit | 16 % | 22 % | 26% |
| total | 35% | 48 % | 55% |

The results in the (1st) column indicate that the front digit was more recognized than the back digit. This fact may mean that it is easier for the CNN to recognize the front digit than the back digit, as in the case of human vision. The recognition percentage of the (1st or 2nd) column is also larger for the front digit than for the back digit. The total recognition percentage for (1st, 2nd or 3rd) is 55%, so the pre-learned network could recognize one of two digits quite often.

- Recognition of two digits
  Next, we investigated the recognition rate of the CNN on the "one digit on top of the other" images under four conditions: (1) the front digit is the first candidate, and the back digit is the second candidate; (2) the front and back digits are the first and second candidates, the front and back digits are the first and third candidates, or the front and back digits are the second and third candidates; (3) the front or back digit is the first or the second candidate; (4) the front or back digit is one of the first, second, or third candidates. The results are shown in Table II. Condition (1) corresponds to (1st, 2nd), and (2) corresponds to (1st, 2nd), (1st, 3rd), (2nd, 3rd) in the "distinguishing front and back" row. Condition (3) corresponds to (1st, 2nd) and (4) corresponds to (1st, 2nd), (1st, 3rd), (2nd, 3rd) in the "without distinguishing front and back" row.

TABLE II
RESULTS OF RECOGNITION OF TWO DIGITS

| | (1st, 2nd) | (1st, 2nd), (1st, 3rd), (2nd, 3rd) |
|---|---|---|
| distinguishing front and back | 12 % | 18 % |
| without distinguishing front and back | 24 % | 46 % |

From the table, the recognition percentage of both the front and back digits was 12%, and the rate increased to 18% as a result of allowing the 3rd candidate. On the other hand, the rate of two-digit recognition without distinguishing front and back was 24%, and it increased to 46% by allowing the 3rd candidate.

### *B. Two digits on the same level*

We investigated recognition of two digits on the same level by using the network pre-learned with single-digit images. Figure 1 right shows an example of an image with two digits on the same level.

- Recognition of one of two digits
  As the two digits were on the same level, we evaluated the percentage that one of the digits was recognized. There were three conditions: (a) the first candidate, (b) the first or second candidate, and (c) the first, second, or third candidate. Table I and Table III show the results.

TABLE III
RESULTS OF RECOGNITION OF ONE OF TWO DIGITS ON THE SAME LEVEL

| | (1st) | (1st or 2nd) | (1st, 2nd or 3rd) |
|---|---|---|---|
| front or back | 36 % | 48 % | 54% |

From the results, the rate of recognizing one of two digits as the first candidate was 36%, that of recognizing one of two digits as the first or second candidate was 48%, and that of recognizing one of two digits as the first, second, or third candidate was 54%. These results are similar to those in the "one digit on top of the other" case.

- Recognition of two digits
  We evaluated the recognition percentage of two digits on the same level under two conditions: (1) two digits recognized as the first or second candidates; (2) two digits recognized as the first, second, or third candidates. Condition (1) corresponds to (1st, 2nd) in the table, while (2) corresponds to (1st, 2nd), (1st, 3rd), and (2nd, 3rd).

TABLE IV
RESULTS OF RECOGNITION OF TWO DIGITS ON THE SAME LEVEL

| | (1st, 2nd) | (1st , 2nd), (1st, 3rd), (2nd, 3rd) |
|---|---|---|
| without distingushing font and back | 27 % | 44 % |

The table shows that the recognition percentage of the two digits as the first and the second candidates was 27%. It increased to 44% by allowing the third candidate. However, the recognition percentage of the two digits on the same level was lower than that of one digit on top of the other. To derive more concrete results, we will need to evaluate a huge number of images.

## IV. DISCUSSION

CNNs has a pre-processing section and a classification section. The pre-processing section of a pre-learned CNN consists of a convolution layer, pooling layer, another convolution layer, and another pooling layer as illustrated in Fig. 4. Figure 4 shows the conceptual behavior of each layer when two digits are correctly recognized. The convolution layers extract the features of the input image, while the pooling layers abstract the output of the previous layer. The small digit images in the figure shows the features of the digit. When the input image shows two digits, the features of the two digits may not be well separated in the first convolution layer. However, if the two digits are correctly classified, the features become separated in the first pooling layer, and the features must be clearly separated in the deeper layers. If the features of the two digits are well separated, they can be easily classified by

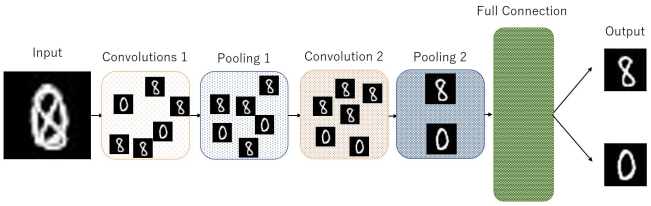


Fig. 4. Conceptual CNN structure that classifies two digits correctly.

the fully connected layers. This leads to a higher recognition rate.

When the image is inputted, the features of the overlapping digits are extracted, as shown in Fig. 4. If the input to the fully connected network is well separated, i.e., almost linearly separable, the two digits are correctly recognized. The output function of the last layer in this case is soft-max, so the two output values show the probabilities of the two digits.

From our results, the CNN that pre-learned to learn individual digits recognized one of two digits about 50% of the time. Moreover, it recognized both digits about 30% of the time. Thus, we can expect that by repeatedly extracting the feature and abstracting the output of the previous layer, the pre-processing section can separate the features of the two digits layer by layer. The finding that the recognition percentage of the "one digit on top of the other" images was larger than that of the "two digits on the same level" images suggests that stereoscopic-like information may aid recognition. This finding may be connected to the stereoscopic vision of humans.

To verify our results, we will have to check the second pooling layer's output to see if the features of the two digits are well separated. We also need to compare the results of the pre-learned CNN with those of human subjects.

## V. Conclusion

A CNN pre-learned on the MNIST dataset of single-digit images was used to recognize images containing two overlapping digits. Our goal was to answer two questions: (1) Can a CNN learned with single-digit images classify images that include two overlapping digits without additional learning on a dataset consisting of images with two overlapping digits? (2) Does a CNN process stereoscopic vision? Our results show that the CNN pre-learned on single-digit images can indeed recognize overlapping digits in images, thereby answering (1) in the affirmative. Our results were insufficient to address (2). Our next step will be to investigate occluded shape recognition by a CNN.

## VI. Acknowledgment

Part of this study was supported by a Grant-in-Aid for Scientific Research (C) (18K11478).

## References

[1] Y. LeCun, Y. Bengio, and G. E. Hinton, Deep learning, Nature, vol. 521, pp. 436–444, 14539 (2015)

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS 4824, (2012)

[3] Y. Bengio, Deep Learning of Representations for Unsupervised and Transfer Learning, JMLR: workshop and conference proceedings 27:17-37 (2012)

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, (2009)

[5] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press. (2016).

[6] http://yann.lecun.com/exdb/mnist/

[7] https://caffe.berkeleyvision.org