# Using Machine Learning to Break Visual Human Interaction Proofs (HIPs)

Kumar Chellapilla
kumarc@microsoft.com
Microsoft Research
One Microsoft Way
Redmond, WA 98052

Patrice Y. Simard
patrice@microsoft.com
Microsoft Research
One Microsoft Way
Redmond, WA 98052

## 1 Applying Best Practices for Convolutional Neural Networks

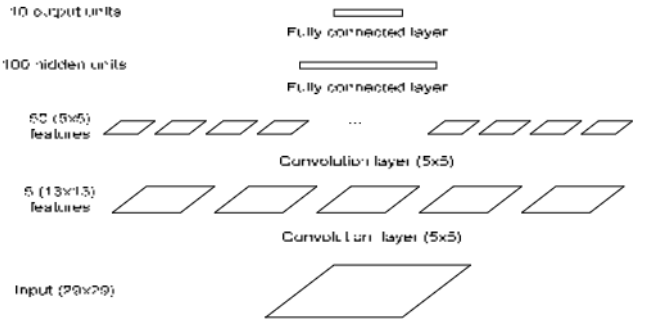
Neural networks are a powerful technology for the classification of visual inputs arising from documents. The most important practice is getting a training set as large as possible: we can expand the training set by adding a new form of distorted data. The next most important practice is that convolutional neural networks are better suited for visual document tasks than fully connected networks. A simple "do-it-yourself" implementation of convolution with a flexible architecture is suitable for many visual document problems. This simple convolutional neural network does not require complex methods, such as momentum, weight decay, structure-dependent learning rates, averaging layers, tangent prop, or even finely-tuning the architecture. The result is a very simple yet general architecture that can yield state-of-the-art performance for document analysis. These claims are illustrated on the MNIST set of English digit images.

### 1.1 Expanding the Data Sets through Elastic Distortions

Given a classification task, one may apply transformations to generate additional data and let the learning algorithm infer the transformation invariance. This variance is embedded in the parameters, so it is in some sense free since the computation at recognition time is unchanged. If the data is scarce and if the distribution to be learned has transformation-invariance properties, generating additional data using transformations may even improve performance. In the case of handwriting recognition, we postulate that the distribution has some invariance with respect to not only affine transformations but also elastic deformations corresponding to uncontrolled oscillations of the hand muscles, dampened by inertia.

## 1.2 Overall architecture for MNIST

It was found that the convolutional neural network performs the best on MNIST. The overall architecture of the CNN used for MNIST digit recognition is as given below:




Convolutional architecture for MNIST dataset

The general strategy of a convolutional network is to extract simple features at a higher resolution and then convert them into more complex features at a coarser resolution. The simplest way to generate coarser resolution is to sub-sample a layer by a factor of 2. This, in turn, is a clue to the convolutional kernel's size. With no padding, a subsampling of 2, and a kernel size of 5, each convolution layer reduces the feature size from n to (n-3)/2.

The first two layers of this network can be viewed as a trainable feature extractor. A trainable classifier can be added to the feature extractor in the form of 2 fully connected layers (a universal classifier).

## 1.3 Making Convolutional Neural Networks Simple

This can be done by applying the following methods:

### 1.3.1 Simple loops for Convolution

Fully connected neural networks often use the following rules to implement the forward and backward propagation:

$$x_j^{L+1} = \sum_i w_{j,i}^{L+1} x_i^L \tag{1}$$

$$g_j^L = \sum_j w_{j,i}^{L+1} g_j^{L+1} \tag{2}$$

where
$x_i^L$ - activation of unit i at layer L
$g_i^L$ - gradient of unit i at layer L,
$w_{j,i}^{L+1}$ - weight connecting unit i at layer L to unit j at layer L+1.

The reason is that in a convolutional layer, the number of connections leaving each unit is not constant because of border effects.

### 1.3.2 Modular debugging

Back propagation has a good property: it allows neural networks to be expressed and debugged in a modular fashion. If the larger modules fails the test, we test each of the sub-modules until we find the culprit. This extremely simple and automated procedure can save a considerable amount of debugging time.

## 1.4 Results

For both fully connected and convolutional neural networks, 50,000 patterns of the MNIST training set was used for training, and the remaining 10,000 for validation and parameter adjustments.
The results for the same are:

| Algorithm | Distortion | Error |
|---|---|---|
| 2 layer MLP (MSE) | affine | 1.6% |
| SVM | affine | 1.4% |
| Tangent dist. | affine+thick | 1.1% |
| Lenet5 (MSE) | affine | 0.8% |
| Boost. Lenet4 MSE | affine | 0.7% |
| Virtual SVM | affine | 0.6% |
| 2 layer MLP (CE) | none | 1.6% |
| 2 layer MLP (CE) | affine | 1.1% |
| 2 layer MLP (MSE) | elastic | 0.9% |
| 2 layer MLP (CE) | elastic | 0.7% |
| Simple conv (CE) | affine | 0.6% |
| Simple conv (CE) | elastic | **0.4%** |

## 1.5 Conclusion

Highest performance on MNIST dataset known till date was achieved using elastic distortion and convolutional neural networks.

# 2 Gradient-Based Learning Applied to Document Recognition

Multilayer neural networks trained with the back-propagation algorithm constitute the best example of a successful gradient- based learning technique. Given an appropriate network architecture, gradient-based learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns, such as handwritten characters, with minimal preprocessing. Convolutional neural networks, which are specifically designed to deal with the variability of two dimensional (2-D) shapes, are shown to outperform all other techniques. Real-life document recognition systems are composed of multiple modules including field extraction, segmentation, recognition, and language modeling. A new learning paradigm, called graph transformer networks (GTN's), allows such multimodule systems to be trained globally using gradient-based methods so as to minimize an overall performance measure. Experiments demonstrate the advantage of global training, and the

flexibility of graph transformer networks. A graph transformer network for reading a bank check uses convolutional neural network character recognizers combined with global training techniques to provide record accuracy on business and personal checks. It is deployed commercially and reads several million checks per day.

# 3 Conclusion

Hence, based on the references given in the selected research paper and the concepts learned in the course the above methods can improve the performance of the machine learning algorithm that will be able to solve the HIPs as discussed in the paper. (These are not the only methods that can increase the performance of the algorithm).