Better digit recognition with a committee of simple Neural Nets

Ueli Meier and Dan Claudiu Cireşan and Luca Maria Gambardella and Jürgen Schmidhuber
IDSIA
USI, SUPSI
6928 Manno-Lugano, Switzerland

{*ueli,dan,luca,juergen*}*@idsia.ch*

***Abstract*—We present a new method to train the members of a committee of one-hidden-layer neural nets. Instead of training various nets on subsets of the training data we preprocess the training data for each individual model such that the corresponding errors are decorrelated. On the MNIST digit recognition benchmark set we obtain a recognition error rate of 0.39%, using a committee of 25 one-hidden-layer neural nets, which is on par with state-of-the-art recognition rates of more complicated systems.**

***Keywords*-Neural Networks; MNIST; Handwritten Digit Recognition; Committee**

I. Introduction

Whatever the approach for building a classifier to solve visual pattern recognition tasks [1]–[7], at some stage in the design process one has collected a set of possible classifiers. In most studies the various classifiers are evaluated on a benchmark data set and only the result of the best classifier is reported. Obviously one of the classifiers yields the best performance. Intriguingly, the sets of misclassified patterns of the different classifiers do not necessarily overlap. This suggests that different classifier designs offer complementary information, which could be harnessed in a committee. An overview of various fusion strategies can be found in [8]–[14]. More recently [15] showed how a combination of various classifiers can be trained faster than a single classifier yielding the same error rate.

For a committee to work best, the aim is to produce a group of classifiers such that their errors are not correlated. This can be achieved using different classifiers and different training sets. In this study we focus on the latter, training identical classifiers on data that are preprocessed in different ways. As long as the same output activation function is used for all the classifiers, it is straightforward to combine them.

Currently, the best results on MNIST have been obtained by deforming the training set [1], [3], [5], [7]. Deformations are a simple way to avoid over-fitting through implicit regularization and also to introduce the desired invariance into the classifiers. In addition to deformations we focus on preprocessing of the data prior to training.

II. Building the Committee

Consider a pattern recognition problem where pattern $\mathbf{x}$ is assigned to one of k possible classes. Using a softmax activation for the output layer of the neural nets and a 1-of-k coding scheme for the target data, the outputs of the trained nets approximate the posterior class probabilities [16]. Having n trained networks we focus on three different methods to build the corresponding committee of networks:

1) Majority voting Committee: choose the class with most votes from the n classifiers for a given input $\mathbf{x}$ (if two classes have the same number of votes, choose the first);
2) Average Committee: average the class probabilities from the n classifiers and choose the class with highest average posterior class probability for a given input $\mathbf{x}$;
3) Median Committee: take the median of the class probabilities from the n classifiers and choose the class with highest median posterior class probability for a given input $\mathbf{x}$.

The majority voting scheme also works if the outputs of the various networks are normalized differently, but all information about the confidence of each prediction is discarded. The average and median committees on the other hand require the outputs of the various networks to be normalized in the same way, but also provide scores/confidence levels for each class label.

Forming a committee can be formulated as a linear combination of the individual experts [17]:

$$y_{com}(x) = \sum_{n=1}^{N} w_n y_n(x) \tag{1}$$

where N is the number of individual experts and w_n is the combination weight for each expert.

A more flexible combination scheme is possible when each weight also depends on class k [18]:

$$y^k_{com}(x) = \sum_{n=1}^{N} w_{nk} y_{nk}(x) \tag{2}$$

resulting in $k \times N$ weights that have to be inferred from additional data that for obvious reasons should be distinct

from the data used to train the classifiers. In the experimental section we list results of optimal committees whose combination weights are obtained minimizing the MSE error over the validation set. On the validation set the optimized committees perform better (by construction) but do not generalize well to the test set. Instead of minimizing the MSE error [18] proposed to minimize the misclassification error. However, we show that rather than optimizing the combination of the various experts it is more important to actually obtain experts whose predictions are as weakly correlated as possible.

III. Training the neural nets

In all our experiments we train multilayer perceptrons (MLPs) with one hidden layer of 800 units. We use the standard softmax output non-linearity with cross-entropy loss function and hyperbolic tangent hidden unit activation function. The inputs are normalized (scaled to $[0, 1]$) and the weights are initialized from a zero mean Gaussian with standard deviation scaled by the fan-in to each unit [16]. All MLPs are trained for 500 epochs with a stochastic conjugate gradient algorithm (batches of 1000 images) that maintains pairwise conjugation of gradients [19]. 10000 randomly chosen digits of the MNIST [1] training set are used for validation and the remaining 50000 digits for training. The MLP with lowest error on the validation set is considered trained and subsequently used as the classifier. If training data are continuously deformed, elastic deformations [3], scaling (horizontal and vertical) and rotation are used. We combine affine (rotation, scaling, horizontal shearing) and elastic deformations, characterized by the following real-valued parameters:

- σ and α for elastic distortions emulating uncontrolled oscillations of hand muscles (see [3] for details);
- β – a random angle from $[-\beta, +\beta]$ describes either rotation or horizontal shearing. In case of shearing, $\tan\beta$ defines the ratio between horizontal displacement and image height;
- γ_x, γ_y for horizontal and vertical scaling, randomly selected from $[1 - \gamma/100, 1 + \gamma/100]$.

Preprocessing of the original MNIST data is mainly motivated by practical experience. MNIST digits are normalized such that the width or height of the bounding box equals 20 pixels. The variation of the aspect ratio for various digits is quite large, and we normalize the width of the bounding box to range from 8 to 20 pixels with a step-size of 2 pixels prior to training for all digits except ones. This results in 7 different training sets. Additionally, we generate a deslanted training set horizontally shearing the digit with magnitude: $tan(\alpha) * d$, where α is the angle of the first principle component of pixel intensities with respect to the vertical axis and d is the vertical distance from the image center (Fig. 1).

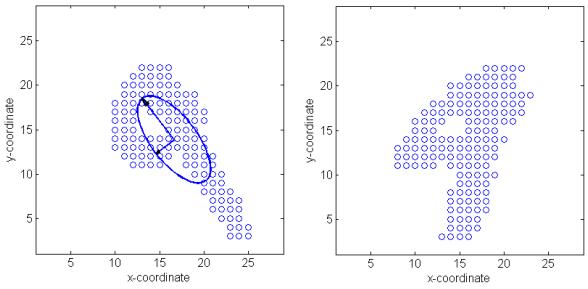

Figure 1. (Left panel) x-, y- coordinates (circles) of the original image together with the eigenvectors scaled by the corresponding eigenvalues. (Right panel) x-, y-coordinates (circles) of the deslanted image after horizontal shearing.

The experiments performed with these nine different data sets will henceforth be referred to as the experiments with preprocessed data. Figure 2 shows ten digits from MNIST preprocessed as described above (left) and the same digits with additional deformations (right). The first row corresponds to original digits whereas from the second row downwards increasing bounding box normalization from 8 to 20 pixels is applied, the last row corresponds to deslanted digits.

Figure 2. (Left panel) Different preprocessing for ten digits from MNIST. From top to bottom: original, 8, 10, 12, 14, 16, 18, 20, deslanted. (Right panel) Similar but with deformations (see text for explanation).

IV. Experiments

We perform six experiments to test the performance increase associated with the use of a committee. Each committee consists of nine randomly initialized one-hidden-layer MLPs with 800 hidden units, trained with the same algorithm on randomly selected batches. The five committees differ only in how the data are preprocessed (or not) prior to training and on how the data are deformed during training.

The first two experiments are performed on undeformed original MNIST images. We train a committee of nine MLPs on original MNIST and we also form a committee of MLPs trained on preprocessed data (as described in section III). In Table I the error rates are listed for each of the individual

nets and the three committees. The improvement of the committees with respect to the individual nets is marginal for the first experiment. Through preprocessing the individual experts and the corresponding committees achieve however substantially better recognition rates.

Table I
Error rates of each individual net and three committees. For experiment 1 nine nets were trained on the original MNIST, whereas for experiment 2 nine nets were trained on preprocessed data: WN x - Width Normalization of the bounding box to be x pixels wide; DESL - deslanted training set; ORIG - original MNIST.

	Error rate [%]			
	Exp. 1		Exp. 2	
Net 1:	init 1:	1.83	WN 8:	1.58
Net 2:	init 2:	1.79	WN 10:	1.62
Net 3:	init 3:	1.80	WN 12:	1.37
Net 4:	init 4:	1.77	WN 14:	1.48
Net 5:	init 5:	1.72	WN 16:	1.53
Net 6:	init 6:	1.91	WN 18:	1.56
Net 7:	init 7:	1.86	WN 20:	1.49
Net 8:	init 8:	1.62	DESL:	1.80
Net 9:	init 9:	1.75	ORIG:	1.79
Majority:	**1.72**		**1.28**	
Average:	**1.69**		**1.28**	
Median:	**1.72**		**1.29**	

In order to see the combined effect of preprocessing and deformation, we perform four additional experiments on deformed MNIST (Tab. II). Unless stated otherwise, default elastic deformation parameters $\sigma = 6$ and $\alpha = 36$ are used. All experiments with deformed images independent horizontal and vertical scaling of maximum 12.5% and a maximum rotation of $\pm 12.5°$. Experiment 3 is similar to Experiment 1, with the exception that the data are continuously deformed. Error rates of the individual experts are much lower than without deformation (Tab. I). More importantly, the error rates of the committees (0.55%) are the best reported results for such a simple architecture. In experiment 4 we randomly reselect training and validation sets for each of the individual experts, simulating in this way the bootstrap aggregation technique [10]. The resulting committee does however not perform better than that of experiment 3. In experiment 5 we vary deformations for each individual network. Error rates of some of the individual nets are bigger than in experiments 3 and 4, but the resulting committees have significantly lower error rates. In the last experiment we train nine MLPs on preprocessed images that are also continuously deformed. The error rate of the average committee (0.40 %) equals the best error rate obtained without pretraining but with a dedicated architecture (i.e. a convolutional net [3]). We also form a committee of all the 25 independent nets listed in Table II. We exclude nets from experiment 4 because they are trained using the same deformation as nets in experiment 3. Net 5 from Experiment 5 and Net 9 from experiment 6 are also excluded because these two nets are taken from experiment 3. The error rate of the resulting average committee (0.39 %) matches the current best result [5], obtained with pretrained convolutional nets.

For all six experiments the average committee gives the lowest error rates, the majority and median committees perform nearly as well.

A. Optimized Committees

As discussed in Section II one can also optimize the combination of experts over the validation set. In Table III we list the recognition rates obtained by minimizing the MSE error over the validation set for the two models in eq. (1) and (2), referred to as optCom1 and optCom2. Recognition rates are listed for the validation as well as for the test set and for comparison the result of the average committee is also listed. By construction, error rates on the validation set are lower for the optimized committees and the more flexible combination scheme (eq. 2) yields the lowest error rates. As can be seen from the recognition rates on the test set, the optimized committees do not always generalize well to the test set. Interestingly, the error rates of the optimized committees for the bootstrapped experiment (Exp. 4) are extremely low for the validation set but do not generalize well to the test set. We conclude that forming an optimized linear combination over a validation set does not generalize well to the unknown test set, and forming a committee by simply averaging the outputs is sufficient.

Table III
Error rates of average and optimized committees on the validation as well as on the test set for all six experiments.

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6
validation:						
average:	1.72%	1.22%	0.56%	0.26%	0.57%	0.50%
optCom1:	1.70%	1.16%	0.56%	0.19%	0.58%	0.46%
optCom2:	1.54%	1.12%	0.47%	0.06%	0.49%	0.38%
test:						
average:	1.69%	1.28%	0.55%	0.54%	0.47%	0.40%
optCom1:	1.69%	1.25%	0.54%	0.54%	0.49%	0.41%
optCom2:	1.71%	1.23%	0.55%	0.60%	0.50%	0.44%

B. Summary of Experiments

The 39 misclassified digits of the best committee from Table II are shown in Figure 3. Many of them are ambiguous and/or uncharacteristic, with obviously missing parts or strange strokes. Interestingly, the second guess of the committee is correct for all but one digit for which the third guess is the correct answer. For the third digit from Figure 3 for example it is even difficult for a human to tell the digit from being a three or a five, and as a matter of fact the committee is also undecided, assigning posterior class probabilities of $p(3|x) = 0.4661$ and $p(5|x) = 0.5339$ to the digit three and five respectively.

Why does this work so well? In order to optimally harness the complementary information of each expert in the committee we aimed for experts whose errors are not correlated.

Table II

Error rates of each individual net and three committees. In experiments 3 and 4 nine nets were trained on deformed ($\sigma = 6$, $\alpha = 36$) MNIST, the difference being that training and validation sets were reselected in experiment 4. In experiment 5, nine nets were trained on deformed (different σ, α) MNIST, and in experiment 6 nine nets were trained on normalized, deformed ($\sigma = 6$, $\alpha = 36$) MNIST. WN x - Width Normalization of the bounding box to be x pixels wide; DESL - deslanted training set; ORIG - original MNIST.

	Error rate [%]						
		Exp. 3	Exp. 4	Exp. 5		Exp. 6	
Net 1:	init 1:	0.68	0.72	$\sigma = 4.5$ $\alpha = 30$:	0.75	WN 8:	1.05
Net 2:	init 2:	0.72	0.68	$\sigma = 4.5$ $\alpha = 36$:	0.69	WN 10:	0.64
Net 3:	init 3:	0.71	0.82	$\sigma = 4.5$ $\alpha = 42$:	0.94	WN 12:	0.78
Net 4:	init 4:	0.72	0.73	$\sigma = 6.0$ $\alpha = 30$:	0.55	WN 14:	0.70
Net 5:	init 5:	0.71	0.69	$\sigma = 6.0$ $\alpha = 36$:	0.72	WN 16:	0.60
Net 6:	init 6:	0.62	0.71	$\sigma = 6.0$ $\alpha = 42$:	0.60	WN 18:	0.59
Net 7:	init 7:	0.65	0.70	$\sigma = 7.5$ $\alpha = 30$:	0.86	WN 20:	0.70
Net 8:	init 8:	0.80	0.66	$\sigma = 7.5$ $\alpha = 36$:	0.79	DESL:	0.63
Net 9:	init 9:	0.69	0.75	$\sigma = 7.5$ $\alpha = 42$:	0.61	ORIG:	0.71
Majority:		**0.55**	**0.54**		**0.49**		**0.43**
Average:		**0.55**	**0.54**		**0.47**		**0.40**
Median:		**0.55**	**0.54**		**0.49**		**0.42**
	All 25 independent nets from experiment 3,5 and 6 (see text for explanation)						
Majority:	**0.41**						
Average:	**0.39**						
Median:	**0.40**						

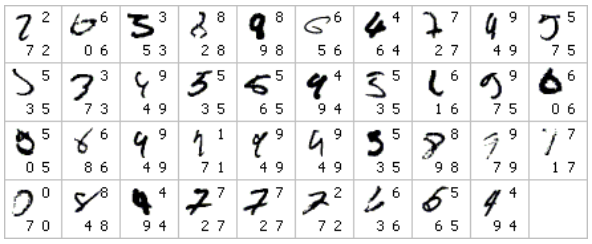

Figure 3. The 39 errors of the best committee from Table II, together with the two most likely predictions (bottom, from left to right) and the correct label (top, right).

And indeed, performance of the committees crucially depends on the percentage of the total errors that are committed by a single expert. For experiment 1 only 16.9% of the errors are committed by a single expert. Applying normalization prior to training, as in experiment 2, this percentage roughly doubles to 32.9%. Interestingly, deformations applied in experiment 3 (33.3%) have an effect similar to preprocessing. In experiment 4 no improvement was observed through random re-selection of training and validation set (33.7%). Choosing different deformation parameters as in experiment 5, the percentage rises to 36.8%. Combining preprocessing with deformations, as in experiment 6, resulted in 38.3% and also produced the best committees.

In a companion paper [20] we successfully applied a committee of convolutional neural networks (CNN) to handwritten character recognition (including upper- and lowercase letters). Furthermore a committee of an MLP trained on features and a CNN trained on pixel intensities [21] won the German Traffic Sign Recognition Benchmark [22]. This demonstrates that simply averaging predictions of various experts is an easy way to improve recognition performance for different tasks.

V. Conclusions

For a committee to work best, the errors of the individual experts should not be correlated. We showed how this is achieved by simple preprocessing of the data prior to training. The applied preprocessing is motivated by observed variations in aspect ratio and slant of handwritten digits. Using a committee of simple, one-hidden-layer MLPs with 800 hidden units, we are able to achieve state-of-the-art performance on the MNIST benchmark. The two big advantages of the proposed method are: 1) forming the committee does not require additional training data, and 2) through different preprocessing the individual predictors are not strongly correlated.

Acknowledgments

This work was supported by Swiss CTI, Commission for Technology and Innovation, Project n. 9688.1 IFF: Intelligent Fill in Form, and by Lifeware S.A.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[2] D. Lowe, "Object recognition from local scale-invariant features," in *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[3] P. Simard, D. Steinkraus, and J. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Seventh International Conference on Document Analysis and Recognition*, 2003, pp. 958–963.

[4] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, June 2005, pp. 994–1000.

[5] M. A. Ranzato, C. Poultney, S. Chopra, and Y. Lecun, "Efficient learning of sparse representations with an energy-based model," in *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[6] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Neural Information Processing Systems (NIPS)*, 2007.

[7] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep big simple neural nets for handwritten digit recognition," *Neural Computation*, vol. 22, no. 12, pp. 3207–3220, 2010.

[8] L. Xu, A. Krzyzak, and C. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Syst., Man and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

[9] M. P. Perrorne and L. N. Cooper, "When networks disagree: Ensemble methods for hybrid neural networks," in *Artificial Neural Networks for Speech and Vision*, R. J. Mammone, Ed. London: Chapman & Hall, 1993, pp. 126–142.

[10] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[11] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[12] R. P. W. Duin, "The combining classifier: to train or not to train?" in *Proceedings. 16th International Conference on Pattern Recognition*, vol. 2, 2002, pp. 765–770.

[13] L. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281–286, 2002.

[14] R. E. Schapire, "The boosting approach to machine learning an overview," in *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

[15] K. Chellapilla, M. Shilman, and P. Simard, "Combining multiple classifiers for faster optical character recognition," in *Document Analysis Systems VII*. Springer Berlin / Heidelberg, 2006, pp. 358–367.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[17] S. Hashem and B. Schmeiser, "Improving model accuracy using optimal linear combinations of trained neural networks," *IEEE Transactions on Neural Networks*, vol. 6, pp. 792–794, 1992.

[18] N. Ueda, "Optimal linear combination of neural networks for improving classification performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 2, pp. 207–215, 2000.

[19] N. N. Schraudolph and T. Graepel, "Towards stochastic conjugate gradient methods," in *Proc. 9^{th} Intl. Conf. Neural Information Processing (ICONIP)*, L. Wang, J. C. Rajapakse, K. Fukushima, S.-Y. Lee, and X. Yao, Eds. IEEE, 2002, pp. 853–856.

[20] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Convolutional neural network committees for handwritten character recognition," in *International Conference on Document Analysis and Recognition*, 2011, to appear.

[21] D. C. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *International Joint Conference on Neural Networks*, 2011, to appear.

[22] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German Traffic Sign Recognition Benchmark: A multi-class classification competition," in *International Joint Conference on Neural Networks*, 2011, to appear.