ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

An Experiment in Pre-Emphasizing Diversified Deep Neural Classifiers

Ricardo F. Alvear-Sandoval and Aníbal R. Figueiras-Vidal *

Univ. Carlos III de Madrid - GAMMA-L+/Dept. Signal Theory and Communications
Av. Universidad, 30, 28911 Leganés, Madrid - Spain

Abstract. We explore if adding a pre-emphasis step to diversified deep auto-encoding based classifiers serves to further improve their performance with respect to those obtained just separately using pre-emphasis or diversification. An experiment with a number of well-known databases, selected because they have some complementary characteristics, shows that further improvement does appear, the main condition for it simply being to select general and flexible enough pre-emphasis forms. Other manners of combining diversity and pre-emphasis require more research effort, as well as to investigate if other deep architectures can also obtain benefits from these ideas.

1 Introduction

The theoretically unlimited expressive power –capability of establishing any input-output correspondence– of shallow (one hidden layer) Multilayer Perceptrons (MLPs) was proved in the late 1980s. Yet the limited available training examples impose designs that can be far from this ideal case. The main obstacle is the impossibility of estimating a sufficient number of weights.

Two principal ways of reducing this compromise have been proposed and explored. The first is to build ensembles of MLPs –or other Learning Machines (LMs)– by diversifying the training of each of them. Obiously, this permits to desing overall architectures with a high number of weights. A considerable number of methods for constructing ensembles has appeared. Committees constitute a family of these methods, in which the learning units are trained with different examples and their outputs are subsequently aggregated, usually by simple procedures (direct averaging or majority voting, for example). Bagging [1], in which bootstrap resampling of the examples provides the training sets for the learners, and label switching [2], in which randomly switching examples' labels serves to introduce diversity, are two relevant committee design methods. There are also other ensemble building methods in which units' learning and aggregation are jointly addressed. Boosting merits to be mentioned among them, because its basic idea, to train sequentially weak learners paying more attention to the examples that offer more difficulty to be classified until the corresponding iterative step, is not only very effective, but it opens many different avenues to be implemented. Monographs [3, 4, 5] present details of these methods.

*This work has been partly supported by Research Grant S2013/ICE-2845, CASI-CAM-CM, provided by DGUI-Comunidad de Madrid.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

The second possibility of getting architectures with many weights, and consequently with high expressive power, is to design deep MLPs, i.e., MLPs with more hidden layers. We will refer to this kind of machines as Deep Neural Networks (DNNs). Although some particular families have a long history, to build general forms of DNNs presented difficulties because the Back-Propagation (BP) algorithm fails due to the appearance of vanishing derivatives. But, in [6], a first example of representation training [7] proved to be useful. Representation training constructs hidden layers in a unsupervised form and adds a final supervised layer. After [6], other representation algorithms, and also indirect design methods, in which layers are sequentially added, have been proposed. Even direct training of DNNs has become possible by putting some care to avoid difficulties. There is not room here for a more complete overview, but references [8, 9, 10] will satisfy the interest of any reader.

A question emerges from the above background: Is there any advantage in combining diversity and depth? Suprisingly, this possibility has not deserved much attention, and less than a dozen of published works deal with this subject. Among them, [11] is important because it presents a distortion-based diversification with (Deep) Convolutional NNs (DCNNs) to create the Multi-Column CNNs (MC-CNNs), that give a performance record –0.21% error rate– for the benchmark task of classifying the handwritten digits of the MNIST database [12]. But the only studies that apply traditional diversification techniques to DNNs –in particular, to Deep Auto-Encoding (DAE) classifiers [13]– are [14, 15], in which we applied bagging and switching, as well as binarization [4, 16] (a diversification technique for multi-class problems), which revealed itself as the key to obtain performance improvements.

We also explored how a simple alternative to boosting –which requires weak learners– could be applied to the same type of DNNs [17]. That alternative is pre-emphasis, i.e., weighting the training examples according to an auxiliary classifier, taking into account the critical character of each sample, i.e., its proximity to the classification border and its classification error [18, 19]. In [17], we found that flexible enough pre-emphasis procedures allowed remarkable improvements, requiring a very modest computational cost increase in the design phase, but not in operation, i.e., to classify unseen samples.

Here, we take a further step and check if combining binarization, simple forms of diversification, and pre-emphasis methods allows to obtain even more important performance improvements.

The rest of this contribution is structured as follows. Section 2 briefly reviews the aspects of [14, 15, 17] that are relevant for this study. In Section 3, we present the experimental framework we use, as well as the corresponding results and their discussion. We close the paper with its main conclusions and indicating some open research lines in the same direction.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

2 A review of previous research

In [14, 15], we investigated how a general representation deep architecture, SDAE3 [13], must be modified to better accept diversification. We selected SDAE3 because it is an architecture which is not adapted to image classification problems, such as DCNNs, and, therefore, it was more appropriate for evaluating the effects of the modifications we introduced. And the hidden layers of representation architectures contain important information about the problem to be solved, an advantage for some types of applications.

Two methods to include ensembles, bagging and switching, were used. The first was applied to diversify SDAE3 machines whose final classifiers were binarized when dealing with multi-class databases (G forms) and also to a single SDAE3 whose classification step was diversified by bootstrap resampling (after binarization, if dealing with multi-class problems) (T forms). T forms are the only option for switching.

The main conclusions of [14, 15] are: 1) Binarization is necessary to get advantage from applying the above mentioned conventional diversification techniques (bagging, switching) to multi-class problems; 2) The performance improvements for T forms were bigger that for G forms, indicating that the DAE unit is carrying out its function, disentangling the examples in a way that makes diversification more effective; 3) The best performances for multi-class problems are got by switching T forms with an appropriate Error Correcting Output Code (ECOC) [16] binarization (a well known fact for shallow arquitectures).

In this paper, we present and discuss experiments with the MNIST database and, to help to extract conclusions, also with MNIST-BASIC (MNIST-B) [13], which consists of the same samples but with a smaller training set, and RECTANGLES (RECT) [13], a binary problem with a similar data structure. The corresponding best results in [15], obtained with the above mentioned ECOC, a T structure and switching diversification (with N=101 learners and S=30% switching rate) appear in Table 1. It is clear that they are much better than those using an SDAE3. This is a very significant improvement, although it requires much more computational effort.

In [17], we applied very general and flexible pre-emphasis forms also to SDAE3. For multi-class problems we used as weighting factor for each example

$$\alpha + (1-\alpha)[\beta(1-o_{ac}^{(n)})^2 + (1-\beta)(1-|o_{ac}^{(n)} - o_{ac'}^{(n)}|)] \tag{1}$$

where $o_{ac}^{(n)}$ is the softmax output of the auxiliary classifier for the true class, $o_{ac'}^{(n)}$ the nearest output among those corresponding to wrong classes, and α, β, $0 \leq \alpha, \beta \leq 1$, pre-emphasis parameters that permit to include a "moderating" (no emphasis) term ($\alpha \neq 0$) and, if $\alpha \neq 1$, a term which is related with the classification error ($\beta \neq 0$) and a term which considers the proximity to the border ($\beta \neq 1$). Values of α, β, were established by means of the validation set.

The experimental percent error rate ± standard deviation obtained in [17] for MNIST, MNIST-B, and RECT when applying (1) to an SDAE3 machine when using a conventional SDAE3 as auxiliary classifier were excellent: See them in

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

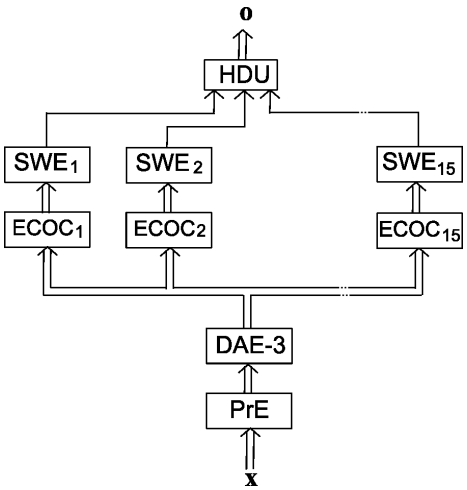

Fig. 1: Classifier architecture for the experiment. $\mathbf{x}$: Input sample. PrE: Pre-emphasis unit. DAE-3: 3-layer deep autoencoder. ECOC_m: Problem coding elements. SWE_m: Switching ensembles (including majority voting). HDU: Hamming distance unit. $\mathbf{o}$: Class indicator.

Table 1. Simplified pre-emphasis forms did not offer such level of improvement, supporting the decisive importance of applying a general and flexible enough pre-emphasis mechanism.

To validate α and β needs around 100 SDAE3 designs, and there is not any increase for classifying unseen samples. This convinced us of the interest of exploring how to combine pre-emphasis with diversification procedures.

3 Experiments

As we said at the end of the Introduction, our objective is to check if applying pre-emphasis to binarized and diversified DAE classifiers further improves their performance. The experiments are defined and carried out with that purporse.

3.1 Experimental framework

We use again SDAE3 as the basic deep classifier. In our designs, its parameters are: 3 hidden layers with 1000 units, one MLP as final classifier with one hidden layer of 1000 units, 0.01 for the first layer training step, 0.02 for the rest, and 10% of added noise level (different from [13]). With respect to the deep architecture, we will use the best of [15], described in the previous section, with the same parameters[1]. The overall machine (including pre-emphasis) appears in Figure 1. The guides are the same machines without pre-emphasis. We carry out 10 runs for each training.

[1]Note that we do not try to validate N and S jointly with α and β.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

	SDAE3		ECOC-ST	
	No PrE	PrE (α,β)	No PrE	PrE (α,β)
MNIST	1.58±0.06	0.37±0.01 (0.4, 0.6)	0.36±0.02	0.30±0.01 (0.3, 0.4)
MNIST-B	3.42±0.10	0.72±0.01 (0.3, 0.5)	0.75±0.01	0.62±0.01 (0.2, 0.6)
RECT	2.40±0.13	0.87±0.04 (0.4, 0.3)	1.10±0.02	0.76±0.03 (0.4, 0.3)

Table 1: Test error rate ± standard deviation (%) for the machines that we use in the experiments.

MNIST, MNIST-B, and RECT are databases with dimensions 28×28, 256 levels for the two first (manuscript digits, 10-class problems) and 2 levels for RECT, with 50000/10000/10000, 10000/2000/50000, and 10000/2000/50000 training/validation/tests samples, respectively.

3.2 Results and their discussion

Table 1 shows the results of the experiments, including the (α, β) values obtained by validation. We include also the results for SDAE3.

The PrE ECOC-ST performances are systematically and clearly better than any other results. We must remark that (α, β) are stably determined with the validation set, and we must say that more restrictive pre-emphasis forms seriously degrade these performances. Thus, we can conclude that combining pre-emphasis with diversification for designing DAE classifiers is effective.

4 Conclusions and further work

According to our study, to combine general and flexible enough pre-emphasis methods with ECOC binarized and diversified (by means of switching) DAE classifiers –specifically, SDAE3s– permits an error reduction even bigger than their separate application to solve three well-known problems (MNIST, MNIST-B, and RECT) of complementary characteristics. For MNIST, the error rate is pretty close to the absolute performance record, without using "ad hoc" architectures. And we can add that there are clear possibilities of further improvements, for example, by separately pre-emphasizing each binary problem that appears after applying ECOC at the first step to MNIST and MNIST-B.

Additional work is necessary to combine the above with other procedures and to investigate how to obtain similar advantages using other DNN architectures. This is one of our present research lines.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

References

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

[2] L. Breiman, "Randomizing outputs to increase prediction accuracy," *Machine Learning*, vol. 40, pp. 229–242, 2000.

[3] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ: Wiley, 2004.

[4] L. Rokach, *Pattern Classification Using Ensemble Methods*. Singapore: World Scientific, 2010.

[5] R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. Cambridge, MA: MIT Press, 2012.

[6] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief networks," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, 2013.

[8] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, pp. 1–127, 2009.

[9] J. Schmidhuber, "Deep learning in neural networks: An overview," Technical Report IDSIA-03-14, University of Lugano, arXiv:1404.7828v4 [cs.NE], 2014.

[10] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in signal Processing*, vol. 7, pp. 197–387, 2014.

[11] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. Conf. on Computer Vision and Pattern Recognition*, pp. 3642–3649. New York, NY: IEEE Press, 2012.

[12] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, pp. 541–551, 1989.

[13] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Machine Learning Res.*, vol. 11, pp. 3371–3408, 2010.

[14] R. F. Alvear-Sandoval and A. R. Figueiras-Vidal, "Does diversity improve deep learning?," in *Proc. 23rd European Signal Proc. Conf.*, pp. 2541–2545. Nice, France, 2015.

[15] R. F. Alvear-Sandoval and A. R. Figueiras-Vidal, "Effects of diversity on deep auto-encoding based classifiers," submitted to *IEEE Trans. Neural Networks and Learning Systems*, 2016.

[16] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. of Artificial Intelligence Res.*, vol. 2, pp. 263–286, 1995.

[17] R. F. Alvear-Sandoval, M. H. Hayes, and A. R. Figueiras-Vidal, "Generalized pre-emphasis of training examples to improve deep auto-encoding classification," submitted to *IEEE Trans. Neural Networks and Learning Systems*, 2016.

[18] V. Gómez-Verdejo, M. Ortega-Moral, J. Arenas-García, and A. R. Figueiras-Vidal, "Boosting by weighting critical and erroneous samples," *Neurocomputing*, vol. 69, pp. 679–685, 2006.

[19] V. Gómez-Verdejo, J. Arenas-García, and A. R. Figueiras-Vidal, "A dynamically adjusted mixed emphasis method for building boosting ensembles," *IEEE Trans. Neural Networks*, vol. 19, pp. 3–17, 2008.