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/.

# Challenges in Deep Learning

Plamen Angelov[1] and Alessandro Sperduti[2]

1- Lancaster University - School of Computing and Communications
Lancaster, LA14WA - United Kingdom

2- University of Padova - Department of Mathematics
Via Trieste, 63, 35121 Padova - Italy

**Abstract**. In recent years, Deep Learning methods and architectures have reached impressive results, allowing quantum-leap improvements in performance in many difficult tasks, such as speech recognition, end-to-end machine translation, image classification/understanding, just to name a few. After a brief introduction to some of the main achievements of Deep Learning, we discuss what we think are the general challenges that should be addressed in the future. We close with a review of the contributions to the ESANN 2016 special session on *Deep Learning*.

## 1 Introduction

Neural networks with many layers and specialized computational units, i.e. deep networks [1], have reached impressive performances in many perceptual learning tasks (e.g. [2, 3]), and recently also on other learning tasks (e.g. [4]). The reason for this success is believed to hinge on the ability of deep networks to exploit compositionality of internal representations, i.e. the network does learn simple concepts (features) that are then used as components of internal distributed representations (hidden activities). Moreover, the presence of multiple layers enables the creation of more and more complex concepts in a layer of distributed representations with increasing distance of the layer from the input neurons. There is a double computational advantage in this kind of architectural organization: *i)* efficient learning of hidden representations, due to the fact that each feature (i.e. output of a hidden unit) can be learned without having access to the exponentially large number of configurations of the other features; *ii)* an exponential gain in representational power, due to the fact that simpler concepts represented in a layer of the network can be exploited as *primitives* by the next layer to represent, in a combinatorial way, more complex concepts. The well known problem with deep networks is that their training is difficult, can fall into local extrema, takes long time and powerful computational resources, e.g. GPU, and once trained the network is not flexible/adaptive to different new data. This is mainly due to the fact that, in order to solve the highly nonlinear *credit assignment problem* underpinning the training of a neural network, gradient-based techniques are used. Unfortunately, this class of techniques suffers from the problem that was observed in the context of Recurrent Neural Networks but that applies to deep networks as well, i.e. vanishing or exploding gradients [5]. In addition to that, the typical use of sigmoidal units introduces many saddle points in the loss function that make training a painfully slow process [6].

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/.

Fortunately, recent methodological (approximate training algorithms [7, 8, 9], pre-training [10], rectified units [11], dropout [12]) and technological advances in computer facilities (especially GPUs) [13], have allowed the successful training of relatively deep networks with impressive performances [14].

More recent developments of deep learning concerns the ability to deal with complex learning tasks thanks to the integration with different types of external memory [15, 16, 17] and reinforcement learning [18, 4]. Generative deep models have also been a recent subject of study (e.g. [19, 20, 21]), as well as semi-supervised approaches (e.g. [22, 23, 24]). A trend that is gaining more and more interest is the introduction of deep architectures into recurrent networks, which are a special case of deep networks per se (e.g. [25, 26, 27, 28]). Effective pre-training approaches for recurrent networks have been suggested [29, 30], as well as a non trivial link between feed-forward networks and recurrent networks [31].

Notwithstanding these achievements, there are still many issues that need to be addressed. We try to discuss some of these in the following section.

## 2 Challenges

Up to now, progress in deep learning has mainly been achieved exploring architectural variants validated on an experimental basis only. Few attempts have been made to understand why and how deep learning obtains such impressive performances (e.g. [32]). Full understanding of how to choose structural features as well as how to efficiently tune hyper-parameters of models (typically performed through a validation set or a cross-validation approach thanks to extremely expensive, from a computational point of view, procedures) is still far from being a reality. Although, many different types of computational units have been proposed on the basis of their mathematical properties, current research on their choice is mostly experimental and ad hoc.

Consequently, developing a deep theoretical background for tuning and assessing the performance of these networks based on empirical data is essential. Specifically, there is a need for a theoretical framework able to: *i)* measure the complexity of the models, as a function of the number and type of computational units, model topology/structure, model generalization, and learning efficiency; *ii)* allow the definition of theoretically grounded strategies for tuning and assessing the performance of models learned from empirical data; *iii)* develop regularization schemes. A specific framework for assessment of unsupervised learning is also needed.

Evolving (dynamically adapting structure) type deep learning networks constitute a specific challenge.

Another important issue concerns computational efficiency. Currently deep models need a significant amount of computational burden to reach state-of-the-art performances on medium/large size datasets and mainly for off-line environments. Nowadays, however, the amount of available data is growing at a rapid pace. Thus, the definition of a class of deep models amenable to be efficiently

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/.

trained in the presence of big data as well as parallel and distributed computational infrastructures tailored for efficient deep learning computation are needed. A step forward in this direction is to consider online learning from streams of data. Dealing with a stream of data requires the use of bounded constant memory and almost linear time for learning on single input item. While constant memory may not be a relevant issue for a deep network due to the fact that most of the network architectures are static, i.e. the topology of the network is defined before learning takes place and does not change with time, the constraint on time complexity constitutes a serious challenge.

If we look at the nature of data for future applications of deep learning technologies, it is evident that more and more application domains involve data which can naturally be represented in structured form, such as sequences (time series, audio and video signals, DNA, etc.), trees (XML documents, parse trees, RNA, etc.), graphs (chemical compounds, social networks, parts of an image, etc.). How to learn in these structured domains using neural networks has been suggested in [33, 34, 35, 36, 37]. Due to the high combinatorial complexity underpinning structured domains, computationally efficient models to learn relations among structured information at different levels of abstraction are needed. An interesting approach to study could be the development of deep versions of Reservoir Computing models [38]. Incremental approaches provide another research alternative, e.g. exploiting the framework introduced in [39].

## 3 Contributions to the ESANN 2016 Special Session on Deep Learning

The *Deep Learning* special session includes papers covering many of the topics discussed above.

For example, papers [40, 41, 42] address architectural aspects of deep networks of different nature. Specifically, in the context of Reservoir Computing for sequence processing, the analysis presented in [40] aims at the study of approaches to develop and enhance multiple time-scale and hierarchical dynamics in deep recurrent architectures. Actually, the paper suggests that stacked layers turn out to be important for the diversification of temporal representations. The work presented in [41] explores a new way to combine deep learning with ensemble methods. Starting from the pre-emphasis technique, where training examples are weighted according to an auxiliary classifier so to take into account both the proximity of an input instance from the classification border and its classification error, the paper investigates if combining it, Error Correcting Output Code binarization, and simple forms of diversification (switching ensembles) allows to obtain better performance. An architecture combining all these components actually turns out to permit, on the MNIST benchmark, an error reduction bigger than their separate application. A different aspect is covered by [42], where convolutional neural networks for image processing are considered. In this case, the proposal consists in the introduction of a new module, based on the chirp-Z transform (a generalization of the discrete Fourier transform), to

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/.

model translation, rotation and scaling invariances in images. Since the chirp-Z transform is linear, two important features of the proposed module are differentiability and scalability, which makes it particularly suited to the use in deep architectures.

Contribution to the development and analysis of learning approaches are given by papers [43, 44]. The work presented in [43] proposes to mitigate the problem of adversarial and fooling examples encountered in deep networks dealing with images and using the log-softmax loss function. Adversarial examples are images affected by imperceptible changes that cause the network to return different outputs with respect to the original ones. Fooling examples are images not belonging to the input distribution to which the network assigns high classification confidence. The authors suggest that the origin of these problems is due to the extrapolating nature of the log-softmax, and propose to replace the standard log-softmax loss in neural networks with the Generalized Learning Vector Quantization cost function. This gives rise to Deep LVQ (DLVQ), which achieves comparable performance on MNIST while being more robust against fooling and adversarial examples. The aim of [44] is to give a better characterization of the properties of the two most popular learning algorithms for training Restricted Boltzmann Machines, i.e. Contrastive Divergence (CD) and Persistent Contrastive Divergence (PCD). Due to their approximate nature, both algorithms yield significantly different biases and variances for stochastic gradient estimates of individual data points. The authors, on the basis of empirical evidence on the lower stochastic gradient estimate variance than exact sampling of CD, provide support to the finding that CD can be used with smaller mini-batches or higher learning rates than PCD.

We have discussed before the importance of pre-training techniques. Paper [45] explores the use of self-adjoint auto-encoders to derive and test slightly deeper structures than one-hidden-layer that could be suitable for pre-training. Specifically, four- and six-layered networks are considered and empirically assessed. The experimental results seem to suggest that deeper architectures are to be preferred.

The contribution of paper [46] concerns the study of the role of deep learning within a more complex scenario, i.e. learning over multivariate and relational time-series with missing data where relations are modelled by a graph. Deep learning is used to get latent representations of temporal data. These latent representations allow to capture the temporal structure of the process jointly with the relations between the different information sources. Moreover, they do not only allow to predict future values for the time-series, but also to fill in missing values. An advantage of this approach is the possibility to have a uniform treatment of these two issues, instead of addressing them separately by different methods.

Examples of applications of deep learning are given by [47, 48]. The problem addressed in [47] concerns robust detection of pedestrians by using a vision system combined with thermal cameras. This originates multispectral data. The paper explores the potential of deep models for processing this kind of data.

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/.

Specifically, two deep fusion architectures are used and their performances analyzed. The obtained results show that a pre-trained late-fusion architecture significantly outperforms the current state-of-the-art solution. In [48], the problem of recognizing crop types from aerial high resolution images collected by drones is considered. A new hybrid neural network architecture which combines histograms and convolutional units is proposed and studied. Empirical assessment, comparing the proposed hybrid system versus single convolutional and histogram-based models, shows that the proposed hybrid system performs better than either model individually.

## References

[1] I. Goodfellow, Y. Bengio, and A. C. Courville. Deep learning. Book in preparation for MIT Press, 2016.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1106–1114, 2012.

[3] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back. Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997.

[4] D. Silver and et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[5] Y. Bengio, P. Y. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[6] Y. N. Dauphin and et al. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems 27*, pages 2933–2941, 2014.

[7] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[8] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

[9] T. Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1064–1071, 2008.

[10] D. Erhan, Y. Bengio, A. C. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11:625–660, 2010.

[11] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 315–323, 2011.

[12] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[13] S. Zhang, A. E. Choromanska, and Y. LeCun. Deep learning with elastic averaging sgd. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 685–693. Curran Associates, Inc., 2015.

[14] A. Antoniou and P. Angelov. A general purpose intelligent surveillance system for mobile devices using deep learning. In *International Joint Conference on Neural Networks, IJCNN 2016 within WCCI2016, Vancouver, BC, Canada, 25-29 July*, 2016. To appear.

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/.

[15] J. Weston, S. Chopra, and A. Bordes. Memory networks. *CoRR*, abs/1410.3916, 2014.

[16] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014.

[17] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2431–2439. Curran Associates, Inc., 2015.

[18] V. Mnih and et al. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

[19] I. J. Goodfellow and et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

[20] D. Jimenez Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1278–1286, 2014.

[21] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[22] H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 536–543, 2008.

[23] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3581–3589. Curran Associates, Inc., 2014.

[24] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. Semi-supervised learning with ladder networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3532–3540. Curran Associates, Inc., 2015.

[25] N. Bandinelli, M. Bianchini, and F. Scarselli. Learning long-term dependencies using layered graph neural networks. In *International Joint Conference on Neural Networks, IJCNN 2010, Barcelona, Spain, 18-23 July, 2010*, pages 1–8, 2010.

[26] M. Hermans and B. Schrauwen. Training and analysing deep recurrent neural networks. In *Advances in Neural Information Processing Systems 26.*, pages 190–198, 2013.

[27] K. Cho and et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*, pages 1724–1734, 2014.

[28] G. Mesnil and et al. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 23(3):530–539, 2015.

[29] L. Pasa and A. Sperduti. Pre-training of recurrent neural networks via linear autoencoders. In *Advances in Neural Information Processing Systems 27*, pages 3572–3580, 2014.

[30] L. Pasa, A. Testolin, and A. Sperduti. Neural networks for sequential data: a pre-training approach based on hidden markov models. *Neurocomputing*, 169:323–333, 2015.

[31] A. Sperduti. Equivalence results between feedforward and recurrent neural networks for sequences. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 3827–3833, 2015.

[32] A. B. Patel, T. Nguyen, and R. G. Baraniuk. A probabilistic theory of deep learning. *CoRR*, abs/1504.00641, 2015.

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/.

[33] A. Sperduti, D. Majidi, and A. Starita. Extended cascade-correlation for syntactic and structural pattern recognition. In *Advances in Structural and Syntactical Pattern Recognition, 6th International Workshop, SSPR '96, Leipzig, Germany, August 20-23, 1996, Proceedings*, pages 90–99, 1996.

[34] A. Sperduti and A. Starita. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, 8(3):714–735, 1997.

[35] P. Frasconi, M. Gori, and A. Sperduti. A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks*, 9(5):768–786, 1998.

[36] B. Hammer, A. Micheli, A. Sperduti, and M. Strickert. A general framework for unsupervised processing of structured data. *Neurocomputing*, 57:3–35, 2004.

[37] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[38] H. Jaeger and H. Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, pages 78–80, 2004.

[39] A. Micheli. Neural network for graphs: A contextual constructive approach. *IEEE Transactions on Neural Networks*, 20(3):498–511, 2009.

[40] C. Gallicchio and A. Micheli. Deep reservoir computing: A critical analysis. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

[41] R. F. Alvear-Sandoval and A. R. Figueiras-Vidal. An experiment in pre-emphasizing diversified deep neural classifiers. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

[42] J. Degrave, S. Dieleman, J. Dambre, and F. Wyffels. Spatial chirp-z transformer networks. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

[43] H. de Vries, R. Memisevic, and A. Courville. Deep learning vector quantization. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

[44] M. Berglund. Stochastic gradient estimate variance in contrastive divergence and persistent contrastive divergence. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

[45] J. Hänninen and T. Kärkkäinen. Comparison of four- and six-layered configurations for deep network pretraining. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

[46] A. Ziat, G. Contardo, N. Baskiotis, and L. Denoyer. Learning embeddings for completion and prediction of relational multivariate time-series. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

[47] J. Wagner, V. Fischer, M. Herman, and S. Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

[48] J. Rebetez and et al. Augmenting a convolutional neural network with local histograms - a case study in crop classification from high-resolution uav imagery. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

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/.