Transfer of Learning Across Deep Networks to Improve Performance in Problems with Few Labelled Data

Telmo Amaral

29th January 2014

NNIG Technical Report No. 1/2014

Project "Reusable Deep Neural Networks: Applications to Biomedical Data"
(PDTC/EIA-EIA/119004/2010)

Neural Networks Interest Group

Instituto de Engenharia Biomédica (INEB)
Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

Contents

1 Introduction

This report presents an overview of two manuscripts that we have recently submitted to two conferences, reporting on experiments in which we explored the use of transfer learning to improve the performance of deep neural networks in problems with limited amounts of labelled training data. Section 2 summarises the manuscript [1] that we have submitted to the International Conference on Pattern Recognition (ICPR 2014). Section 3 addresses the manuscript [2] submitted to the International Joint Conference on Neural Networks (JICNN 2014).

2 Paper overview: Improving performance on problems with few labelled data by reusing stacked auto-encoders

Deep architectures, such as neural networks with two or more hidden layers, are a class of networks that comprise several levels of non-linear operations, each expressed in terms of parameters that can be learned [3]. The organisation of the mammal brain into processing stages that correspond to different levels of abstraction, as well as the way in which humans organise their ideas hierarchically, are among the main motivations for the use of such architectures. Nevertheless, until 2006, attempts to train deep architectures generally resulted in poorer performance than that achieved by shallow networks. A breakthrough took place with the introduction by Hinton et al. [7] of the deep belief network, whose hidden layers are initially treated as restricted Boltzmann machines (RBMs) and pre-trained, one at a time, in an unsupervised greedy approach. This pre-training procedure was soon generalised to rely on machines easier to train than RBMs, such as auto-encoders [9].

The goal of transfer learning is to reuse knowledge associated with a source problem to improve the learning of the classification function associated with a target problem [10]. The source and target problems may differ, for example, as to the data distributions, or they may involve different sets of classes. A common approach to transfer learning is that of transferring representations that were learned from one problem onto another problem.

Deep architectures have been used recently in transfer learning settings, as discussed by Bengio et al. [4, Section 2.4] and Deng and Yu [5, Chapter 11]. Possibly the closest works to ours found in existing literature are the cross-lingual speech recognition experiments recently reported by Huang et al. [8] and by Heigold et al. [6], though with important differences: in both cases the problems at hand were not of image classification; the source and target problems differed in terms of both the set of class labels *and* the data distribution; and stacked auto-encoders were not employed. In addition, the work of Heigold et al. did not involve unsupervised pre-training of source networks or variations in the amount of data used to train target networks.

In this work, we explored the transfer of knowledge between deep networks designed to classify images of digits. We aimed to limit the differences between

source and target problems by considering only two types of cases: either the set of class labels changed, while the underlying data distribution remained the same; or the label set remained fixed, while the data distribution changed. Our main goal was to study in a systematic way how the performance of transfer learning on such problems would be affected by varying the number of layers being retrained and, simultaneously, varying the amount of data available for retraining. In each experiment, we pre-trained without supervision and fine-tuned with supervision a network, using labelled data associated with the source problem; then we retrained with supervision selected layers from that network (while keeping the other layers untouched), using labelled data available for the target problem, to obtain a new network. Our results are pertinent for situations in which few labelled data exist for the target problem (even if a large amount of unlabelled data happen to be available, for example to be used in unsupervised pre-training).

In general, reusing source networks trained for a different label set led to higher improvements in the classification error than reusing networks trained for a different data distribution.

Retraining only one layer led to the worst performance in all sets of experiments. So, even though retraining the output layer is essential when class labels change and retraining the first hidden layer is (presumably) important when the data distribution changes, in practice more than one layer should be retrained for successful knowledge transfer to occur.

In transfer learning experiments involving a change of label set, retraining the second hidden layer and the output layer of the source networks allowed to achieve the best performance while saving significant training time, in relation to fully training the target networks from scratch.

When the target problem was that of classifying digits into ten classes and only a limited amount of data was available for training, it was better to reuse a network trained for only two classes than to train a network from scratch for ten classes. In the case of machine-printed digits this advantage was only evident for very low values of N (up to 300 samples), but in the case of handwritten digits it was visible up to values around 2400 samples. On the other hand, when the target problem involved classifying digits into only two classes, it was always better to reuse a network trained for ten classes than to train a network from scratch for two classes, regardless of the amount of data available for training.

In transfer learning experiments involving a change of data distribution, no significant differences were observed between the times taken to fully train the target network from scratch and to retrain any number of layers from the source network.

In experiments involving handwritten digits and machine-printed digits, transfer learning was beneficial only when the data available for training the target network did not exceed 300 samples. In experiments involving only handwritten digits (first and third in the legend), this advantage held for higher values of N. Thus, when classifying upright digits reusing networks trained for rotated digits, we observed pronounced error improvements up to 600 samples; when classifying rotated digits reusing networks trained for upright digits, the improvements were more modest, but they persisted for higher values of N (up to 1200 samples).

The results obtained when classifying handwritten upright digits by reusing networks trained for rotated digits are particularly interesting, as they raise the hypothesis that, in classification problems where only a small amount of labelled image data is available for training, it may be possible to take advantage of transfer learning to achieve improved performance. One possible approach would be to: a) perform random rotations (or another transformation) on each instance of the original data, in order to generate a larger amount of data; b) use the generated data to train a source network; and c) retrain that source network using the small amount of original images, to obtain a better network than would be possible by using only the original images. Future work should investigate this possibility, and also involve a wider variety of data types and larger architectures, designed for training using graphics processing units (GPUs).

3 Paper overview: Transfer learning using rotated image data to improve deep neural network performance

In a recent work [1] we have observed that, for small amounts of training data (30 or 60 samples per class), it was possible to obtain a significantly better deep network for the classification of upright handwritten digits by retraining a source network designed to classify randomly rotated versions of the same type of digits, this one trained using a large amount of data (600 samples per class). This suggested that, in the presence of a small training set, it could be possible to use the training set *itself* to obtain a transformed training set (by performing for example a random rotation on each sample), train a source network using the transformed data, then retrain that network using the original data, to achieve lower classification errors than would be possible by using only the original data.

In this work we explore the idea described above, using three different types of character image data, different small amounts of training data, two types of random rotation (small or unrestricted) and two distinct transfer learning approaches (only pre-training the source network, or pre-training and fine-tuning it). We achieved significant improvements in the classification error by fully training a source network using slightly rotated versions of the original training samples, then fine-tuning that network again using the original samples. For very small amounts of training data, it was possible to further improve performance by introducing more than one rotation per original sample.

For all data types involved in our experiments, networks designed via the transfer learning approach described above yielded significantly lower errors than networks trained using only original (non-rotated) data. Relative improvements between 6% and 16% were observed in the average errors, at the expense of training times 50% to 100% longer.

In general, pre-training and fine-tuning a source network led to better results than just pre-training it. Restricting the rotations performed on the original design samples to a small range led to better results than freely rotating the samples. It would be interesting to study in finer detail the relationship between

performance and the range of allowed rotation.

For small amounts of original design data, it was possible to further improve performance by including in the transformed data more than one randomly rotated version of each original sample. With k=5 rotations per original sample, relative improvements between 8% and 42% were observed in the average test error. This implied training times about three times longer than those associated with a single rotation.

Future work should explore finer variations in the number of introduced rotations (for example all values of $k \in [1, ..., 10]$). A wider variety of types of image data should also be addressed, as well as geometrical transformations other than rotation. For example, skewing could be explored in the case of image samples whose orientation is irrelevant.

References

[1] T. Amaral, J. Marques de Sá, L. M. Silva, L. A. Alexandre, and J. M. Santos. Improving performance on problems with few labelled data by reusing stacked auto-encoders. (Under review), 2013. 3, 5

[2] T. Amaral, L. M. Silva, L. A. Alexandre, J. Marques de Sá, and J. M. Santos. Transfer learning using rotated image data to improve deep neural network performance. (Under review), 2014. 3

[3] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009. 3

[4] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 3

[5] Li Deng and Dong Yu. Deep learning for signal and information processing. Microsoft Research monograph, 2013. 3

[6] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. Multilingual acoustic models using distributed deep neural networks. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013. 3

[7] G. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006. 3

[8] J.-T Huang, J. Li, D. Yu, L. Deng, and Y. Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013. 3

[9] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *International Conference on Machine Learning*, pages 473–480, 2007. 3

[10] S. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 3