
Block Neural Network Avoids Catastrophic Forgetting When Learning Multiple Task

Guglielmo Montone

Laboratoire Psychologie de la Perception
Université Paris Descartes
75006 Paris, France
montone.guglielmo@gmail.com

J.Kevin O'Regan

Laboratoire Psychologie de la Perception
Université Paris Descartes
75006 Paris, France
jkevin.oregan@gmail.com

Alexander V. Terekhov

Laboratoire Psychologie de la Perception
Université Paris Descartes
75006 Paris, France
avterekhov@gmail.com

Abstract

In the present work we propose a Deep Feed Forward network architecture which can be trained according to a sequential learning paradigm, where tasks of increasing difficulty are learned sequentially, yet avoiding *catastrophic forgetting*. The proposed architecture can re-use the features learned on previous tasks in a new task when the old tasks and the new one are related. The architecture needs fewer computational resources (neurons and connections) and less data for learning the new task than a network trained from scratch

1 Introduction

Two recently suggested architectures, the block neural network [6, 4] and the progressive neural network [5], tested respectively in a supervised learning paradigm and a reinforcement learning paradigm have shown impressive results in *multi-task learning*. The block neural network is created by training several Deep Feed Forward networks (DNNs) on different tasks. The networks are then connected using new neurons and connections, forming a bigger network that is trained on a new task by allowing just the new added connections to be updated. Block neural networks and progressive neural networks have both been shown to benefit from the advantages of transfer learning. Whereas in the past different forms of *pre-training* [2, 3] and *multi-task learning* [1] have also achieved this, block neural networks and progressive networks do so without suffering from the disadvantage of *catastrophic forgetting* of old tasks in the case of pre-training and the necessity of a persistent reservoir of data for the multi-task learning. In this paper, after quickly revisiting the block network architecture, we propose a set of binary classification tasks and show that the block architecture learns more simply (the network needs less computational resources: neurons and connections) and more quickly (the train set can be much smaller) than a network trained from scratch.

2 Merging DNNs

We defined a set of tasks T_1, \dots, T_M and trained a DNN N_1, \dots, N_M (*base models*) on each task. After the first training phase, we used some of the trained networks, say N_1, \dots, N_m , to build a *block architecture* that was then trained on one of the remaining tasks, say T_{m+} . The block architecture

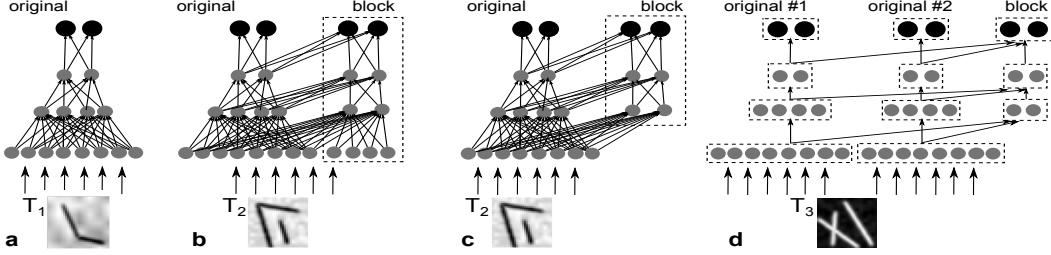


Figure 1: (a) The architecture is built by adding a block of neurons with three hidden layers to one base model. (b) Adding a block of neurons with two hidden layers to one base model. (c) Adding a block of neurons with one hidden layer to one base model. (d) Adding a block of neurons to two base models. The dashed boxes indicate the layers of the two base models and the block of neurons added. An arrow connecting two boxes indicates that all the neurons in the first box are connected to all the neurons in the second box.

was formed by adding a set of new neurons (*block neurons*) to the previously trained networks N_1, \dots, N_m . The block neurons were connected to the base models as follows: the first hidden layer of the block neurons received the input for the task T_{m+1} . The same input was sent to all networks N_1, \dots, N_m . The second hidden layer was fully connected to both the first hidden layer of the block neurons and the first hidden layer of each network N_1, \dots, N_m . This pattern was repeated for all the layers. This architecture was tested with two variations. In the two variations respectively the first and the second layer of the block neurons were removed. When training on the task T_{m+1} *none of the parameters in the base model networks was allowed to change*. Figure 1 provides a representation of the block neural network.

3 The tasks

We used six binary classification tasks, which the networks were trained on. The tasks all involved the concepts of line and angle. We wished to show that the networks N_1, \dots, N_m , when trained on such tasks, would develop features that could be reused by the block architecture to solve another task involving the same concepts. In each task the stimuli were gray scale images, 32×32 pixels

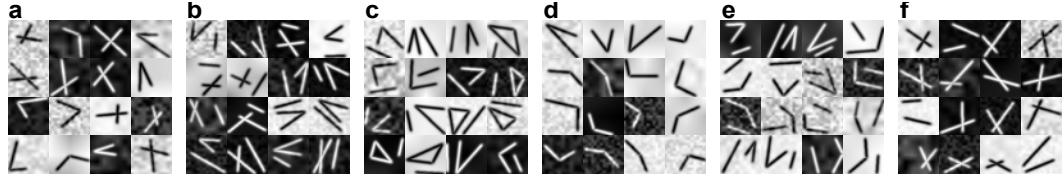


Figure 2: Examples of stimuli: (a) *ang_crs*; (b) *ang_crs_ln*; (c) *ang_tri_ln*; (d) *blt_srp*; (e) *blt_srp_ln*; (f) *crs_ncrs*

in size. Each image contained two to four line segments, each at least 13 pixels long (30% of the image diagonal). The segments were white on a dark random background or black on a light random background. The 6 tasks were (see examples in figure 2):

ang_crs: requires classifying the images into those containing an angle (between 20° and 160°) and a pair of crossing line segments (the crossing point must lie between 20% and 80% along each segment's length).

ang_crs_ln: the same as *ang_crs*, but has an additional line segment crossing neither of the other line segments.

ang_tri_ln: distinguishes between images containing an angle (between 20° and 160°) and a triangle (with each angle between 20° and 160°); each image also contains a line segment crossing neither angle nor triangle.

blt_srp: requires classifying the images into those having blunt (between 100° and 160°) and those having sharp (between 20° and 80°) angles in them.

Table 1: Original network results

Condition	200-100-50 (300K params)	60-40-20 (65K params)
<i>ang_crs</i>	5.5(5.4-5.9)	9.4(8.9-9.8)
<i>ang_crs_ln</i>	13.6(12.5-15.2)	18.3(16.7-18.8)
<i>ang_tri_ln</i>	6.1(5.5-6.8)	11.4(10.6-14.0)
<i>blt_srp</i>	2.0(1.8-2.3)	3.7(3.4-4.2)
<i>blt_srp_ln</i>	6.5(6.4-6.9)	12.5(11.6-14.1)
<i>crs_ncrs</i>	2.8(2.3-2.9)	4.5(4.1-5.2)

blt_srp_ln: the same as *blt_srp*, but has an additional line segment, crossing neither of the line segments forming the angle.

crs_ncrs: distinguishes between a pair of crossing and a pair of non-crossing lines (the crossing point must lay between 20% and 80% of each segment length).

4 Results

In this section, we first report the results obtained by training a DNN on each of the previously described tasks. Then we report the results of training different block neural networks on the same tasks. The number of possible architectures that can be built by changing the base models, the number of block neurons and the task on which the block network is trained, is very large, and exploring all possibilities was not feasible. A more detailed analysis of the configurations tried can be found in our previous studies[6, 4]. Here we summarize the results obtained with two kinds of block network architectures that are particularly interesting because they are obtained by adding a very small number of block neurons. Moreover in this paper we focus on the ability of such architectures to learn using a much smaller dataset. We will in fact present the performance obtained by several block architectures when such architectures are trained on a dataset of almost half the size of the dataset used for training a network from scratch.

The performance of the networks was evaluated by computing the percentage of misclassified samples on the test dataset. Each architecture was trained five times, randomly initializing its weights. The mean performance over the five repetitions and the best and worst performance are reported in the tables.

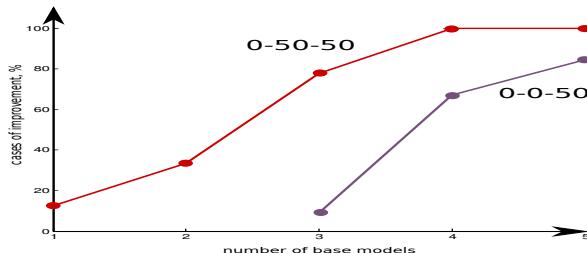


Figure 3: Percentage of block architectures outperforming a network trained from scratch as a function of the number of base models present in the block architecture

Original Network

Prior to building block architectures, we trained a DNN on each task. The networks used were of type NN-200-100-50, with 200, 100, and 50 nodes in the first, second, and third layers, respectively. Networks of this type were used as base models for all of the block networks. The percentages of misclassified test examples for these networks are shown in table 1 together with the results for another architecture, namely NN-60-40-30. Such networks had approximately the same number of parameters (weight of the networks) as some of the block networks, making interesting performance comparisons possible. The networks were trained on datasets with 350,000 examples.

Condition	BA-0-50-50 (60K params)	BA-0-0-50 (5K params)
<i>ang_crs</i> (<i>ang_tri_ln+crs_ncrs+blt_srp+blt_srp_ln</i>)	5.0(4.8-5.2)	5.8(5.4-6.3)
<i>ang_crs</i> (<i>ang_tri_ln+ang_crs_ln+crs_ncrs+blt_srp_ln</i>)	4.3(4.0-4.5)	4.6(4.0-5.0)
<i>ang_crs</i> (<i>ang_tri_ln+crs_ncrs+blt_srp_ln+ang_crs_ln</i>)	4.3(4.1-4.8)	4.7(4.3-5.5)
<i>ang_crs_ln</i> (<i>ang_tri_ln+crs_ncrs+blt_srp_ln+ang_crs</i>)	10.7(10.4-11.3)	12.0(11.5-12.4)
<i>ang_crs_ln</i> (<i>ang_tri_ln+crs_ncrs+blt_srp+blt_srp_ln</i>)	12.4(12.0-12.6)	15.1(14.6-15.5)
<i>blt_srp</i> (<i>ang_crs+ang_tri_ln+crs_ncrs+blt_srp_ln</i>)	1.2(1.1-1.4)	1.4(1.3-1.5)
<i>blt_srp</i> (<i>ang_crs_ln+ang_tri_ln+crs_ncrs+ang_crs</i>)	1.8(1.7-2.0)	2.1(1.7-2.4)
<i>blt_srp_ln</i> (<i>ang_crs_ln+ang_tri_ln+crs_ncrs+ang_crs</i>)	6.4(6.3-6.6)	9.7(9.2-10.6)
<i>blt_srp_ln</i> (<i>ang_crs+ang_tri_ln+crs_ncrs+ang_crs_ln</i>)	6.5(6.3-6.8)	9.8(9.4-10.3)

Block Architecture

In figure 3 we present the percentage of block networks outperforming a network trained from scratch as a function of the number of base models present in the block network. Here we focus on two kinds of block architecture, namely BA-0-50-50 and BA-0-0-50. The architecture BA-0-50-50 (BA-0-0-50) is obtained by connecting the base models to a DNN with 0(0) units in the first hidden layer, 50(0) units in the second hidden layer, and 50 units in the third hidden layer. The plots in figure 3 were obtained as follow. We built several instantiations of the two kinds of block architectures by using randomly selected base models. Each architecture was then trained on each of the tasks if the task was not used to train any of its base models. For example a block network built using the base model trained on *blt_srp* and *ang_crs* was trained on all the other tasks excepts those two. The performance obtained by each block network was compared with the performance obtained with a network (NN-200-100-50) trained from scratch on the same task. The percentage of times the block network obtained a better score was evaluated. The plot in the figure clearly shows an increase in the performance of the block network as the number of base models grows. On the one hand this result was expected simply because the bigger the number of base models, the more parameters are trained; on the other hand it is important to stress that the number of parameters trained on the block network is in any case much smaller than that of a network trained from scratch. The architecture BA-0-50-50 with five base models, for example, has about 60K parameters compared to the 300K of the network 200-100-50. In table 2 and table 3 we show the performance of the block network when the network

Condition	BA-0-50-50 (75K params)	BA-0-0-50 (25K params)
<i>ang_crs</i> (all model used except <i>ang_crs</i>)	4.3(4.0-4.7)	4.4(3.9-4.7)
<i>ang_crs_ln</i> (all model used except <i>ang_crs_ln</i>)	10.6(10.4-10.8)	11.7(11.2-12.1)
<i>blt_srp</i> (all model used except <i>blt_srp</i>)	1.2(0.9-1.9)	1.4(1.1-1.8)
<i>blt_srp_ln</i> (all model used except <i>blt_srp_ln</i>)	5.6(5.2-5.9)	7.2(6.8-8.0)
<i>crs_ncrs</i> (all model used except <i>crs_ncrs</i>)	1.2(1.0-1.3)	1.2(1.0-1.3)
<i>ang_tri_ln</i> (all model used except <i>ang_tri_ln</i>)	5.8(5.7-6.0)	8.6(8.3-9.0)

is trained on a dataset of 200,000 examples, almost half of the size of the dataset used to train the network NN-200-100-50. The percentages of misclassified test examples for block architectures with four and five base models are presented in the tables. The architectures that performed better than (or equal to) the network NN-200-100-50, which was trained from scratch, are shown in bold. In these tables, the tasks on which the block architectures were trained are listed together with the tasks on which the base models were trained (in parentheses).

5 Conclusions

The block architecture proves to be a very effective solution for approaching the problem of multi-task learning in DNN. The architecture can be a first step toward the construction of DNN architectures which, in an unsupervised fashion, are able to profit from training on prior tasks when learning a new task.

References

- [1] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.
- [2] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.
- [3] Grégoire Mesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, Yoshua Bengio, Ian J Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, et al. Unsupervised and transfer learning challenge: a deep learning approach. *ICML Unsupervised and Transfer Learning*, 27:97–110, 2012.
- [4] Guglielmo Montone, J Kevin O’Regan, and Alexander V Terekhov. The usefulness of past knowledge when learning a new task in deep neural networks. *Cognitive Computation Workshop, NIPS*, 2015.
- [5] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.
- [6] Alexander V Terekhov, Guglielmo Montone, and J Kevin O’Regan. Knowledge transfer in deep block-modular neural networks. In *Conference on Biomimetic and Biohybrid Systems*, pages 268–279. Springer, 2015.