Neural Networks: increasing capacity can improve learning speed

Pieraut Francis, Caporossi Gilles, Bengio Yoshua

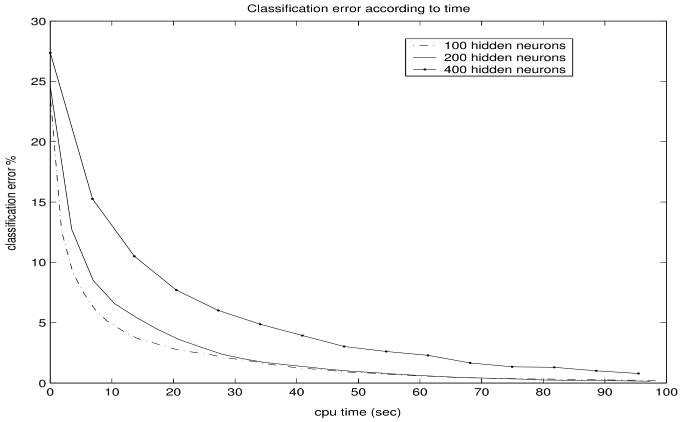

Fig. 1. Learning speed reduce according to capacity increase

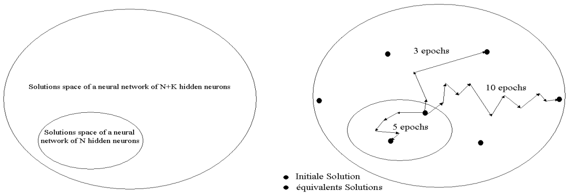
Solutions space of a neural network of N+K hidden neurons
Solutions space of a neural network of N hidden neurons
3 epochs
10 epochs
5 epochs
Initiale Solution
équivalents Solutions

Fig. 2. Solutions space and solutions examples

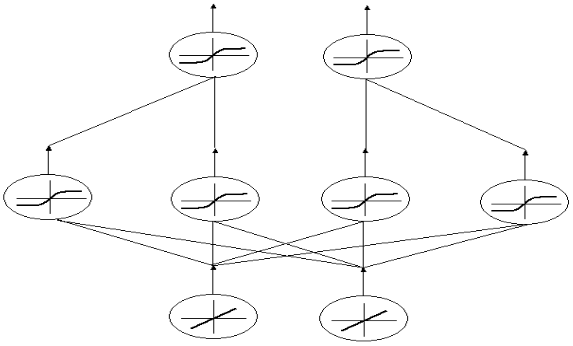

Fig. 3. Learning speed reduce according to capacity increase

Abstract— **Ours investigations to understand the reasons why huge neural networks seems to not be able to take advantage of their capacity to get better learning result compare to smaller one bring use to compare learning speed of different capacity neural network on the same problem. Conceptual investigation show us that increasing solutions spaces (capacity) can increase learning speed. In practice, with a fully connected architecture we notice that learning speed decrease constantly. This observation is confirm in others works but not documented. With a decoupled architecture, we notice that speed deterioration disappear and can increase learning speed. This important constatation has been mostly attributed the discard of an optimisation problem, the opposite gradient, cause by architecture change.**

Index Terms— **Neural Networks, learning speed.**

I. Introduction

BEyond a certain capacity, with a fully connected architecture, experimental results show us that learning speed decrease as the number of parameters increase (figure 1). First section will present why we think that this behavior is strange and should be attributed to back-propagation optimisation problem. Second section present the decoupled architecture and the effect on *opposite gradients* problem, an optimisation problem. Last section show us that that speed deterioration disappear and can increase learning speed when increasing capacity of uncouple architecture of a neural network.

II. Conceptual investigation

Contrary to experimental observations, we think an higher capacity can increase learning speed. In the point of view of solution space, a higher space contain several equivalent solutions. If a solution in a smaller space is highly non-linear, it is easily to figure it out that there is a equivalent solution less complex in an higher space. Knowing that learning process of neural networks is slower as it learn non-linearity relations, a equivalent solution less complex in an higher space can be learn in less steps and maybe can be, over all, faster.

III. Decoupled architecture

Opposite gradients problems, an optimisation problem, is discarded if we use an decoupled neural network (figure 3). This architecture modification technique can only be done in classification task. Increasing the number of parameters with this type of architecture consist of increasing the number of hidden units per outputs. We try the same experience as fully connected architecture but with decoupled architecture (figure 4).

IV. Problem introduce by decoupled architecture

Decoupled architecture for neural networks eliminates *the problem of contradictory gradients* but, inopportunely, it introduces a problem. For an identical number of parameters, a network with decoupled architecture makes it possible to

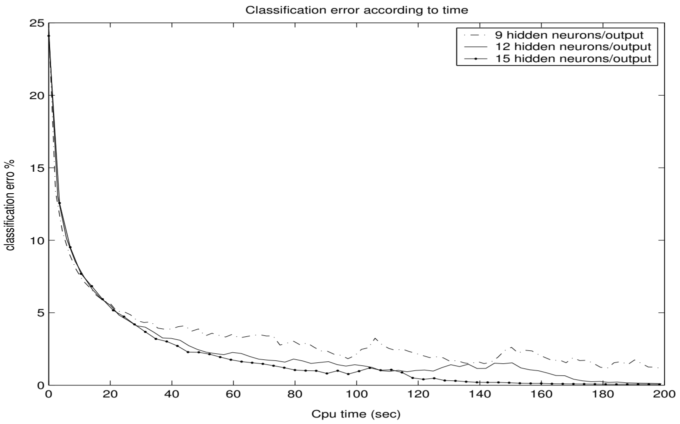

Fig. 4. Learning speed reduce according to capacity increase

represent a lower space of solutions than a completely interconnected network. This is caused by non-sharing of hidden neurons information. The complexity of the function being able to be represented by each output is reduced. Therefore, to have the same capacity, a network of decoupled architecture must have much more free parameters. This implies necessarily an increase in the computing time to carry out an iteration. In practice, we did not need more parameters to learn the same task except a small increase of learning time to reach the same error rate.

V. Parallelism consideration

It is important to mention that learning process of a neural network with an decoupled architecture is extremely easy to parallelise. Taking benefit of this characteristic with a computer cluster can increase learning speed by a factor of classes number. A N classes problem can be see as N independent learning task.

VI. Conclusion

Your experimental investigation show that optimisation problem related to back-propagation can explain why an neural network with more capacity learn slower than a neural network with less capacity. Those results confirm our conceptual investigation with solution space. Using decoupled architecture discard *opposite gradients* problem, an optimisation problem, and speed deterioration disappear and can improve learning speed when increasing capacity. For the same number of parameters, an decoupled architecture represent a smaller solutions space but it parallelism capability can overcome this drawback if we can parallelised training process.