Perturbations, Accuracy and Robustness in Neural Networks

Cesare Alippi
Politecnico Di Milano, Italy

Giovanni Vanini
Politecnico Di Milano, Italy

INTRODUCTION

A robustness analysis for neural networks, namely the evaluation of the effects induced by perturbations affecting the network weights, is a relevant theoretical aspect since weights characterise the "knowledge space" of the neural model and, hence, its inner nature.

In this direction, a study of the evolution of the network's weights over training time (training perturbations) allows the researcher for shedding light on the mechanism behind the generation of the knowledge space. Conversely, the analysis of a specific knowledge space (fixed configuration for weights) provides hints about the relationship between knowledge space and accuracy. This aspect is particularly relevant in recurrent neural networks, where even small modifications of the weight values are critical to performance (e.g., think of the stability of an intelligent controller comprising a neural network and issues, leading to robust control).

Robustness analysis must also be taken into account when implementing a neural network (or the intelligent computational system) in a physical device or in intelligent wireless sensor networks. Behavioral perturbations affecting the weights of a neural network abstract uncertainties such as finite precision representations, fluctuations of the parameters representing the weights in analog solutions (e.g., associated with the production process of a physical component), aging effects or more complex and subtle uncertainties in mixed implementations.

In this article, we suggest a robustness/sensitivity analysis in the large, that is, without assuming constraints on the size or nature of the perturbation; as such, the small perturbation hypothesis becomes only a sub-case of the theory. The suggested sensitivity/robustness analysis can be applied to all neural network models (including recurrent neural models) involved in system identification, control signal/image processing and automation-based applications without any restriction to study the relationship between perturbations affecting the knowledge space and the induced accuracy loss.

ROBUSTNESS ANALYSIS: THE STATE OF THE ART

The sensitivity/robustness issue has been widely addressed in the neural network community with a particular focus on specific neural topologies. In particular, when the neural network is composed of linear units, the relationship between perturbations and the induced performance loss can be obtained in a closed form (Alippi & Briozzo, 1998). Conversely, when the neural topology is non-linear we have either to assume the small perturbation hypothesis or particular assumptions about the stochastic nature of the neural computation, for example, see Alippi and Briozzo (1998), Pichè (1995), and Alippi (2002b); unfortunately, such hypotheses are not always satisfied in real applications. Another classic approach requires expanding the neural computation with Taylor around the nominal value of the trained weights. A subsequent linearized analysis follows which allows the researcher for solving the sensitivity issue problem (Pichè, 1995). This last approach has been widely used in the implementation design of neural networks where the small perturbation hypothesis abstracts small errors introduced by finite precision representations of the weights (Dundar & Rose, 1995; Holt & Hwang, 1993). Again, the validity of the analysis depends on the validity of the small perturbation hypothesis.

Differently, other authors avoid the small perturbation assumption by focusing the attention on very specific neural network topologies and/or by introducing particular assumptions regarding the distribution of perturbations, internal neural variables and inputs as done for Madalines neural networks (Stevenson, Winter, & Widrow, 1990; Alippi, Piuri, & Sami, 1995).

Some other authors tackle the robustness issue differently by suggesting techniques leading to neural networks with improved robustness ability by acting on the learning phase (e.g., see Alippi, 1999) or by introducing modular redundancy (Edwards & Murray, 1998); though, no robustness indexes are suggested there.

A ROBUSTNESS ANALYSIS IN THE LARGE

In the following, we consider a generic neural network implementing the $\hat{y}(x) = f(\hat{\theta}, x)$ function where $\hat{\theta}$ is the weight vector of the trained neural network.

In several neural models, and in particular in those related to system identification and control, the relationship between the inputs and the output of the system is captured by considering a regression vector φ, which contains a limited time-window of actual and past inputs, outputs, and, possibly, predicted outputs. Of particular interest are those models which can be represented by means of the model structures $\hat{y}(t) = f(\varphi)$ where function $f(\cdot)$ is a regression-type neural network, characterised by N_φ inputs, N_h non-linear hidden units and a single effective linear/non-linear output (Hassoun, 1995; Hertz, Krogh, & Palmer, 1991; Ljung, 1987; Ljung, Sjoberg, & Hjalmarsson, 1996).

The presence of a dynamic in the data can be modelled by a suitable number of delay elements which may affect inputs (time history on external inputs *u(t)*) system outputs (time history on $y(t)$) on predicted outputs (time history on $\hat{y}(t)$) or residuals (time history on $e(t) = \hat{y}(t) - y(t)$). Where it is needed, $y(t)$, $\hat{y}(t)$ and $e(t)$ are vectorial entities, a component for each independent distinct variable.

Several neural model structures have been suggested in the literature, which basically differ in the regression vector; examples are the NARMAX structures which can be obtained by considering both past inputs and outputs:

$$\varphi = [u(t), u(t-1), \cdots, u(t-n_u), y(t-1), \cdots, y(t-n_y), ..., e(t-1), \cdots, e(t-n_e)]$$

and the NOE ones which process only the past inputs and

$$\varphi = [u(t), u(t-1), \cdots, u(t-n_u), \hat{y}(t-1), \cdots, \hat{y}(t-n_y)].$$

Static neural networks, such as classifiers, can be obtained by simply considering external inputs

$$\varphi = [u(t), u(t-1), \cdots, u(t-n_u)].$$

We denote by $\hat{y}_\Delta(x) = f_\Delta(\hat{\theta}, \Delta, x)$ the mathematical description of the perturbed computation and by $\Delta \in D \subseteq \Re^p$ a generic p-dimensional perturbation vector, a component for each independent perturbation affecting the network weights of model $\hat{y}(x)$. The perturbation space D is characterised in stochastic terms by providing the probability density function pdf_D.

To measure the discrepancy between $\hat{y}_\Delta(x)$ and $y(x)$ or $\hat{y}(x)$ we consider a generic loss function $U(\Delta)$. A common example for U is the mean square error (MSE) loss function

$$U(\Delta) = \frac{1}{N_x} \sum_{i=1}^{Nx} (y(x_i) - \hat{y}_\Delta(x_i))^2 \quad (1)$$

but a generic Lebesgue measurable loss function with respect to D can be taken into account (Jech, 1978). The formalization of the impact of perturbation on the performance function can be simply derived as:

Definition: Robustness Index

We say that a neural network is robust at level $\bar{\gamma}$ in D, when the robustness index $\bar{\gamma}$ is the minimum positive value for which

$$U(\Delta) \le \bar{\gamma}, \forall \Delta \in D. \quad (2)$$

Immediately, from the definition of robustness index, we have that a generic neural network NN1 is more robust than another NN2 iff $\bar{\gamma}_1 < \bar{\gamma}_2$; the property holds independently from the topology of the two neural networks.

The main problem related to the determination of the robustness index $\bar{\gamma}$ is that we have to compute $U(\Delta), \forall \Delta \in D$ if we wish a tight bound. The $\bar{\gamma}$ -identification problem is, therefore, intractable from a computational point of view if we relax all assumptions made in the literature as we do. To deal with the computational aspect we associate a dual probabilistic problem to (2):

Robustness Index: Dual Problem

We say that a neural network is robust at level $\bar{\gamma}$ in D with confidence η when $\bar{\gamma}$ is the minimum positive value for which

$$\Pr(U(\Delta) \le \bar{\gamma}) \ge \eta \quad \text{holds} \quad \forall \Delta \in D. \quad (3)$$

The probabilistic problem is weaker than the deterministic one since it tolerates the existence of a set of perturbations (whose measure according to Lebesgue is 1-η) for which $u(\Delta) > \bar{\gamma}$. In other words, not more than 100η% of perturbations $\Delta \in D$ will generate a loss in performance larger than $\bar{\gamma}$. Probabilistic and deterministic problems