Analysis of neural networks through base functions

B.J. van der Zwaag, C.H. Slump,
University of Twente, EL, SAS
P.O. Box 217, 7500 AE Enschede
T +31.53.489 2842; F +31.53.489 1060
E-mail b.j.vanderzwaag@el.utwente.nl
URL www.sens.el.utwente.nl

L. Spaanenburg
Rijksuniversiteit Groningen, IWI
P.O. Box 800, 9700 AV Groningen
T +31.50.363.3925; F +31.50.363 3800
E-mail l.spaanenburg@cs.rug.nl
URL www.cs.rug.nl

Keywords: Neural networks, rule extraction, sensitivity analysis.

Abstract:

Problem statement. Despite their success-story, neural networks have one major disadvantage compared to other techniques: the inability to explain comprehensively how a trained neural network reaches its output; neural networks are not only (incorrectly) seen as a “magic tool” but possibly even more as a mysterious “black box” [1]. This is an important aspect of the functionality of any technology, as users will be interested in “how it works” before trusting it completely.

Although much research has already been done to “open the box,” there is a notable hiatus in known publications on analysis of neural networks. So far, mainly sensitivity analysis and rule extraction methods have been used to analyze neural networks. However, these can only be applied in a limited subset of the problem domains where neural network solutions are encountered.

Research goal and approach. We therefore propose a method which, for a given problem domain, involves identifying basic functions with which users in that domain are already familiar, and describing trained neural networks, or parts thereof, in terms of those basic functions. This will provide a comprehensible description of the neural network’s function and, depending on the chosen basic functions, it may also provide an insight into the network’s inner “reasoning.”

Relevance. Domain-specific analysis of neural networks through base functions will not only provide insight into the in- and external behavior of neural networks and show their possible limitations in particular applications, but it will also lower the acceptability threshold for future users unfamiliar with neural networks. Further, domain-specific neural network analysis methods that utilize domain-specific base functions can also be used to optimize neural network systems. An analysis in terms of base functions may even make clear how to (re)construct a superior system using those base functions, thus using the neural network merely as a construction advisor. If a user does not want to trust a neural network for any reason whatsoever, he may still trust a non-neural system that would have been nearly impossible to construct without using a neural network as an advisor.

Initial results. As an example, the poster shows that an edge detector realized by a neural network can be analyzed in terms of differential filter operators, which are common in the digital image processing domain (for more details, see [2]). The same analysis was applied to some well-known image filters, enabling comparison of conventional edge detectors known from literature and the neural network edge detectors. The difference between our comparison and more commonly used methods for comparison lies herein that our comparison was based directly on the detectors’ filter operations rather than on their performance on a given (benchmark) example. The latter is a more indirect method of comparison and does not provide any insight into the neural network’s functionality.

References:
