Jacek M. Zurada, IEEE Life Fellow
IEEE SMC Distinguished Lecturer
University of Louisville, Louisville, Kentucky, USA
jacek.zurada@louisville.edu

Lecture 1: Towards Better Understanding of Data: Constrained Learning of Latent Features in Neural Networks

Abstract: Learning models that build hierarchies of concepts are inherently difficult to interpret and understand. Hence, convoluted mappings and cancellations of terms performed within neural networks with one hidden layer or more make them less than transparent. However, learning with meaningful constraints either within classic or recently proposed architectures allows for better extraction of discriminative features. The discriminative features are understood here as parts of original sets of objects. Further, they are useful only when they can be superimposed and reconstructed with as low a reconstruction error as possible.

Three techniques are discussed that meet the criteria outlined above. (1) They use supervised and unsupervised learning. Nonnegative Matrix Factorization is one of the efficient techniques that reduces the number of basis functions and allows for extraction of latent features that are additive and hence interpretable for humans. (2) A classic error-back propagation architectures can also be trained under the constraints of non-negativity and sparseness. The resulting classifiers allow for identification of parts of the objects encoded as receptive fields developed by weights of hidden neurons. The results are illustrated with MNIST handwritten digits classifiers and Reuters-21578 text categorization. (3) A constrained learning of sparse encoding representation using non-negative weights of an auto-encoder also allows for discovery of additive latent factors. Our experiments with MNIST, ORL face and NORB object datasets compare the auto-encoding accuracy for various training conditions. They indicate an enhanced interpretability and insights through identified parts of complex input objects traded-off for a small reduction of recognition accuracy or classification error. Although for the sake of interpretability these models discuss only shallow networks, their training strategies parallel those used in multi-layer deep learning.