Kernelization on Neural Network

Shuai Zhang Yingchun Zhang

July 24, 2016

1 Introduction

1.1 Related Works

Wilson et al. [1] combined the non-parametric flexibility of kernel methods with the structural properties of deep neural networks.

2 Methods

2.1 Neural Network

In a two-layer network, the feed-forward network function is

$$\mathbf{y}_k(\mathbf{x}, \mathbf{w}) = h^{(2)} \left(\sum_{j=0}^{M} w_{kj}^{(2)} h^{(1)} \left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i \right) \right). \tag{1}$$

Given a training set with input vectors $\{\mathbf{x}_n\}$, where $n = 1, ..., N$, together with a correspounding set of target vectors $\{\mathbf{t}_n\}$, we minimize the error function

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2. \tag{2}$$

2.2 Kernelized Neural Network

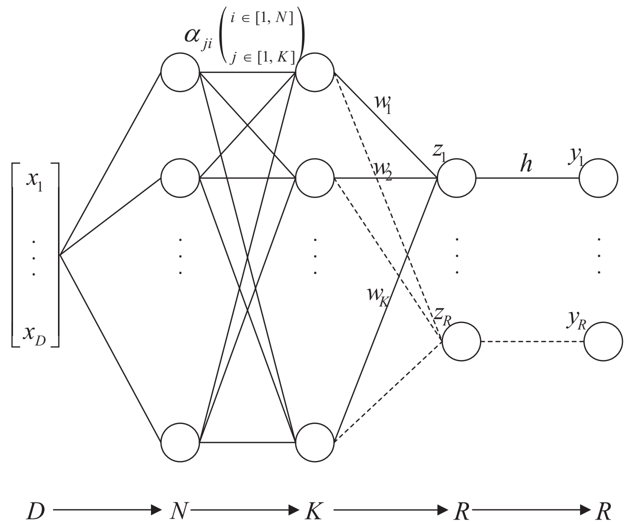
2.2.1 Nonparametric representation

Single-layer neural network

In a single-layer neural network, the feed-forward network function is

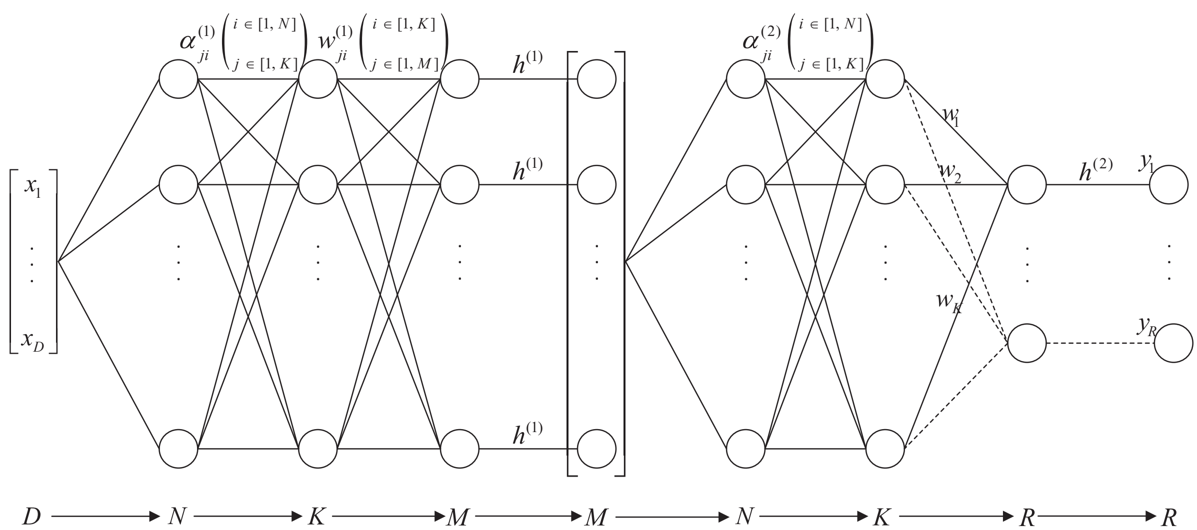
$$y_n = h \left(\sum_{j=1}^{K} w_j \left(\sum_{i=1}^{N} \alpha_{ji} k(\mathbf{x}_i, \mathbf{x}_n) \right) \right) \tag{3}$$

where $\{\mathbf{x}_i\}_{i=1}^{N}$ is the unique sample which is a D-dimension vector of the dataset, k is a kernel function which builds a new inner product space, and we try to get k linear combination of with the weight α_{ji}, then



we make a linear combination of the new base with the weight w_j. h is an activation function and so as follows. All the w, h, α, k and $\mathbf{x}$ shown in the following part have the same meaning as the former ones.

Multi-layer neural network

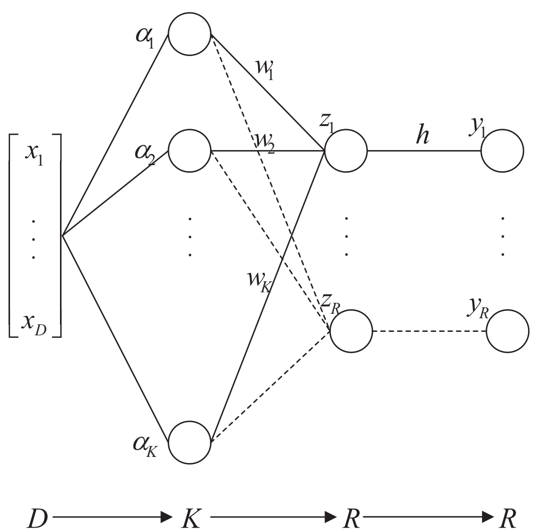


$$y_n = h^{(2)} \left(\sum_{j=1}^{K} w_j^{(2)} \left(\sum_{i=1}^{N} \alpha_{ji}^{(2)} k_2(\mathbf{x}_i, \begin{bmatrix} h^{(1)} \left(\sum_{j=1}^{K} w_{1j}^{(1)} \left(\sum_{i=1}^{N} \alpha_{ji}^{(1)} k_1(\mathbf{x}_i, \mathbf{x}_n) \right) \right) \\ \cdot \\ \cdot \\ \cdot \\ h^{(1)} \left(\sum_{j=1}^{K} w_{Mj}^{(1)} \left(\sum_{i=1}^{N} \alpha_{ji}^{(1)} k_1(\mathbf{x}_i, \mathbf{x}_n) \right) \right) \end{bmatrix}) \right) \right) \tag{4}$$

The number in superscript of the symbols represents the layer of neural network, such as, $w_j^{(2)}$ means the weight of the second layer of neural network and so as follows.

2.2.2 Parametric representation

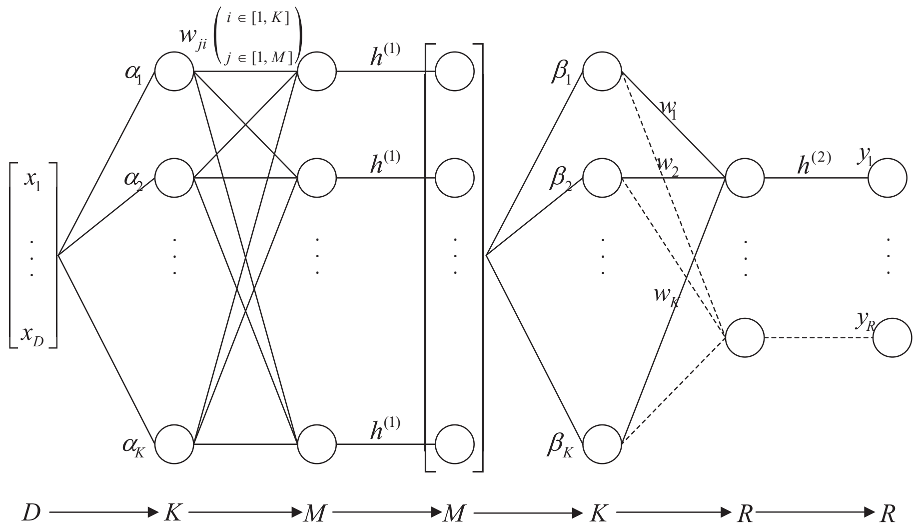
Single-layer neural network



$$y_n = h \left(\sum_{j=1}^{K} w_j k(\boldsymbol{\alpha}_j, \mathbf{x}_n) \right) \tag{5}$$

$\boldsymbol{\alpha}_j$ is the center of kernel which is a D-dimension vector.

Multi-layer neural network



$$y_n = h^{(2)} \left(\sum_{j=1}^{K} w_j^{(2)} k_2(\boldsymbol{\alpha}_j^{(2)}, \begin{bmatrix} h^{(1)} \left(\sum_{i=1}^{K} w_{1i}^{(1)} k_1(\boldsymbol{\alpha}_i^{(1)}, \mathbf{x}_n) \right) \\ \cdot \\ \cdot \\ \cdot \\ h^{(1)} \left(\sum_{i=1}^{K} w_{Mi}^{(1)} k_1(\boldsymbol{\alpha}_i^{(1)}, \mathbf{x}_n) \right) \end{bmatrix}) \right) \tag{6}$$

References

[1] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," *CoRR*, vol. abs/1511.02222, 2015.

[2] G. Pandey and A. Dukkipati, "To go deep or wide in learning?," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pp. 724–732, 2014.