Image Classification using Higher-Order Neural Models

James Bergstra, Yoshua Bengio, Jerome Louradour
Université de Montréal

Neural network research in machine learning grew out of theories from computational neuroscience from the 1960s. While the class of affine-sigmoidal feature extractors has been studied extensively since the mid 1980s, the computational neuroscience community has moved on to new models that are qualitatively different and more descriptive, without being substantially more computationally expensive. This paper brings a particular model proposed in (Rust et al., 2005) for low level neurons in the macaque visual system into a machine-learning context: we evaluate their model as an activation function (feature-extractor) for single-layer neural networks that perform image classification. The function we evaluate is somewhat similar to the higher-order processing units discussed in (Minsky & Papert, 1969) and the Sigma-Pi units described in (Rumelhart et al., 1986), but avoids the computational difficulties associated with these models by representing the second-order interaction weights with a low-rank positive semi-definite matrix, and avoids the learning difficulties associated with these models by using a gentler non-linearity than the logistic sigmoid. Remarkably good comparative results are obtained on three image classification tasks including 1.4% error on `MNIST` using a single-layer network. These results suggest that a single hidden layer neural network equipped with this neuron model can capture important patterns in the data that escape standard models such as sigmoid neural networks and support vector machines based on gaussian and polynomial kernels.

Recently, Rust et al. (2005) describe experiments in which they test for linear and non-linear neuron responses among the simple and complex cells in the early vision system of macaque monkeys. They find that only the simplest cells respond according to a formula like $\mathrm{sigm}(\mathrm{wx}+\mathrm{b})$, while their model (eq.1) makes better predictions by incorporating separate non-linear terms for the excitement (E) and shunting inhibition (S) experienced by a cell.

$$\text{response} = \alpha + \frac{\beta E^\zeta - \delta S^\zeta}{1 + \gamma E^\zeta + \epsilon S^\zeta}, \quad E = \sqrt{\max(0, w'x)^2 + x'V'Vx}, \quad S = \sqrt{x'U'Ux} \tag{1}$$

One key aspect of this formula is the modeling of pairwise interactions between inputs. The naïve extension of a normal neural network activation function to include second-order interactions gives something like the Sigma-Pi unit in equation 2.

$$h_{hpu2,i}(x) = act(b_i + w_i'x + x'W_ix) \tag{2}$$

which is parametrized by scalar b_i, vector $w_i \in \mathbb{R}^d$, and matrix $W_i \in \mathbb{R}^{d\times d}$.

Unfortunately this model is not practical for treating a high-dimensional input because the number of parameters is quadratic in the input dimensionality. To escape this problem, the activation function of our *quadratic sigmoid network* (eq. 3) approximates the large matrix W_i with the difference of two low-rank (rank K) positive semi-definite matrices. For the i-th feature, we have $h_{quad,i}(x)$.

$$h_{quad,i}(x) = act(b_i + w_i'x + x'V_i'V_ix - x'U_i'U_ix) \tag{3}$$

$$h_{ratio,i}(x) = \frac{E_i(x)}{1 + E_i(x)} \tag{4}$$

$$h_{shunt,i}(x) = \frac{E_i(x) - S_i(x)}{1 + E_i(x) + S_i(x)} \tag{5}$$

$$E_i(x) = \sqrt{x'V_i'V_ix + \log[1 + \exp(w_i \cdot x)]^2} \tag{6}$$

$$S_i(x) = \sqrt{x'U_i'U_ix} \tag{7}$$

$$b \in \mathbb{R}; x, w \in \mathbb{R}^d; V_i, U_i \in \mathbb{R}^{K\times d} \tag{8}$$

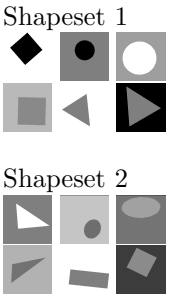
Drawing inspiration from the model of neuron response proposed by Heeger (Heeger, 1993; Carandini & Heeger, 1994), we implemented *ratio networks* out of activation functions h_{ratio}. Based on the neuron response model proposed in Rust et al. (2005) (which is an extension of the Heeger model) we also implemented *shunting networks* out of the activation functions h_{shunt}.

Results

Since the models under investigation were derived from experimentally justified descriptions of low-level neurons in the mammalian visual system, we tested them on simple supervised visual tasks. We generated a

dataset of regular shapes at 32x32 resolution (`ShapeSet1` includes circles, squares, and equilateral triangles), and of stretched ones (`ShapeSet2` includes ellipses, rectangles and triangles). We made our features into classification models by linearly classifying a layer of features. As is common practice with neural networks, we used backpropagation to minimize the average cross-entopy between the model's prediction and the target distribution over classes.

Best Shapeset1 Model			Shapeset1		Shapeset2		MNIST
Family	Units	K	Valid	Test	Valid	Test	Test
SVM	-	-	29.6	$29.3 \pm 1.$	42.2	44 ± 2	1.4
sigm	200	-	13.5	$14.0 \pm 1.$	36.6	36 ± 1	1.9
$sigm_2$	200 $\times 2$	-	6.8	$7.2 \pm .8$	24.0	24 ± 1	2.4
$sigm_3$	200 $\times 3$	-	5.4	$5.9 \pm .7$	22.3	22 ± 1	
quad	20	$\times 4$	6.3	$7.4 \pm .8$	?	?	1.9
ratio	40	$\times 8$	2.1	**2.4 ±.4**	14.3	**16 ± 1**	1.4
shunt	40	$\times 16$	2.9	**3.2 ±.5**	24.7	**26 ± 1**	1.5



The table shows the performance our models alongside normal one-layer (*sigm*), two-layer ($sigm_2$), three-layer ($sigm_3$) neural networks and SVMs with gaussian kernel (found to be better than polynomial). All the models involving quadratic interactions achieved good results with relatively few hidden units (20,40) compared with the best conventional neural networks (200, 500), indicating that quadratic interactions are useful features for generalization. Contrast: the best-performing SVM model kept over 90% of the training set as support vectors. We have also run some experiments on `MNIST`; the best ratio network scored a respectable 1.4% error, the shunting network a slightly poorer 1.5% error and the best quadratic network 1.9%. Surprisingly, the *ratio* model outperformed all other types on all datasets, suggesting that the combination of quadratic interactions with its activation function is good for both optimization and generalization.

While these results are encouraging, we are working to improve them by well-known techniques for neural network optimization. For example, convolutional neural networks have often demonstrated superior image classification performance (Simard et al., 2003); the quadratic and linear components of the ratio and shunting models could be arranged as convolutions too. For another example, the more recent technique of greedy layerwise unsupervised learning to initialize deep networks (Hinton et al., 2006; Bengio et al., 2007) seems to help with generalization if not optimization too. A more thorough empirical or theoretical evaluation of what kinds of features these functions can extract is also ongoing research.

References

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems 19* (pp. 153–160). MIT Press.

Carandini, & Heeger (1994). Summation and division by neurons in primate visual cortex. *Science*, 1333–1336.

Heeger, D. J. (1993). Modeling simple cell direction selectivity with normalized, half-squared, linear operators. *Journal of Neurophysiology, 70*, 1885–1898.

Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*, 1527–1554.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA, USA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1. MIT Press.

Rust, N., Schwartz, O., Movshon, J. A., & Simonchelli, E. (2005). Spatiotemporal elements of macaque v1 receptive fields. *Neuron, 46*, 945–956.

Simard, P. Y., Steinkraus, D., & Platt, J. (2003). Best practice for convolutional neural networks applied to visual document analysis. *International Conference on Document Analysis and Recogntion (ICDAR), IEEE Computer Society* (pp. 958–962). Los Alamitos.