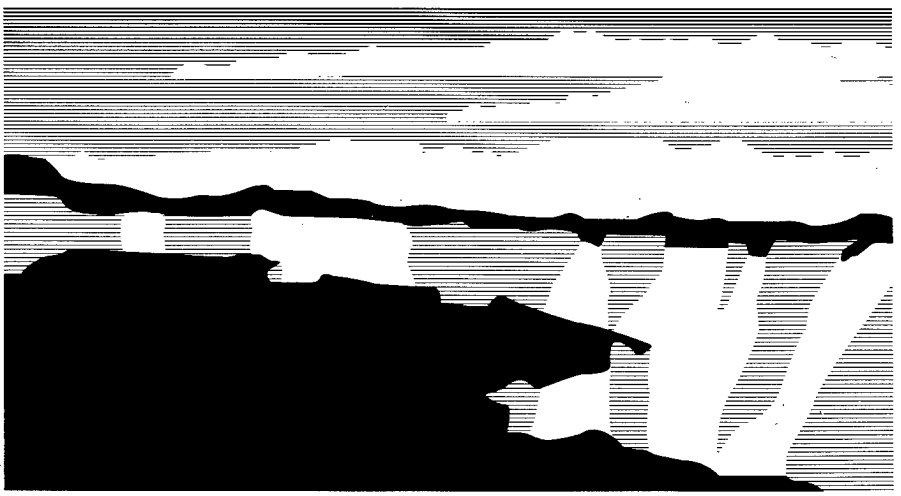
LA-UR- 97-4359

Title: DEEPER AND SPARSER NETS ARE OPTIMAL

CONF-980216--

Author(s): Valeriu Beiu
Hanna E. Makaruk

Submitted to: International Symposium on Engineering of Intelligent Systems EIS '98
Tenerife, Spain
February 9-13, 1998



Los Alamos
NATIONAL LABORATORY

Form No. 836 R5
ST 2629 10/91

DISCLAIMER

Deeper and Sparser Nets Are Optimal

Valeriu Beiu [1] *and* **Hanna E. Makaruk** [2]

Space & Atmospheric Div. NIS-1, MS D466, and Theoretical Div. T-13 MS B213
Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
E-mail: {beiu, hanna_m}@lanl.gov

***Abstract*—*The starting points of this paper are two size-optimal solutions: (i) one for implementing arbitrary Boolean functions (Horne & Hush, 1994); and (ii) another one for implementing certain sub-classes of Boolean functions (Red'kin, 1970). Because VLSI implementations do not cope well with highly interconnected nets—the area of a chip grows with the cube of the fan-in (Hammerstrom, 1988)—this paper will analyse the influence of limited fan-in on the size optimality for the two solutions mentioned. First, we will extend a result from Horne & Hush (1994) valid for fan-in $\Delta = 2$ to arbitrary fan-in. Second, we will prove that size-optimal solutions are obtained for small constant fan-in for both constructions, while relative minimum size solutions can be obtained for fan-ins strictly lower that linear. These results are in agreement with similar ones proving that for small constant fan-ins ($\Delta = 6...9$) there exist VLSI-optimal (i.e., minimising AT^2) solutions (Beiu, 1997a), while there are similar small constants relating to our capacity of processing information (Miller 1956).*

Keywords*—*neural networks, fan-in, threshold circuits, Boolean functions/circuits, circuit complexity, VLSI complexity.

1. Introduction

In this paper we shall consider feedforward neural networks (NNs) made of linear threshold gates (TGs), or perceptrons. A TG is computing a Boolean function (BF):

$$f: \{0, 1\}^n \rightarrow \{0, 1\},$$

where an input vector is $\mathbf{Z}_k = (z_{k,0}, \ldots, z_{k,n-1})$, and:

$$f(\mathbf{Z}_k) = sgn\left(\sum_{i=0}^{n-1} w_i z_{k,i} + \theta\right),$$

with the synaptic *weights* $w_i \in \mathbb{R}$, *thresholds* $\theta \in \mathbb{R}$, and *sgn* the sign function. The cost functions commonly associated are *depth* (*i.e.*, number of edges on the longest input-to-output path, or number of layers) and *size* (*i.e.*, number of neurons). However, the *area* of the connections counts, and the *area* of one neuron can be related to its associated *weights*, thus *"comparing the number of nodes is inadequate for comparing the complexity of NNs as the nodes themselves could implement quite complex functions"* (Williamson, 1990). That is why several authors (Abu-Mostafa, 1988; Hammerstrom, 1988; Phatak & Koren, 1994) have taken into account the total *number-of-connections*, or the total *number-of-bits* needed to represent the *weights* and the *thresholds* (Bruck & Goodmann, 1988), or the sum of all the *weights* and the *thresholds* (Beiu *et al.*, 1994). The sum of all the *weights* and the *thresholds* (also applied for defining the minimum-integer TG realisation of a BF) has been recently used—under the name of *"total weight magnitude"*—in the context of computational learning theory for improving on several standard VC-theory bounds (Bartlett, 1996). A similar definition of 'complexity' $\sum w_i^2$ has also been advocated (Zhang & Mühlenbein, 1993). Such approximations can easily be related to assumptions on how the *area* of a chip scales with the *weights* and the *thresholds* (Beiu, 1996b, 1997a):

- for digital implementation, the *area* scales with the cumulative storage of *weights* and *thresholds* (as the bits for representing those *weights* and *thresholds* have to be stored);
- for analog implementations (*e.g.*, using resistors or capacitors) the same type of scaling is valid (although it is possible to come up with implementations having binary encoding of the parameters—for which the *area* would scale with the cumulative log-scale size of the parameters);
- some types of implementations (*e.g.*, transconductance ones) even offer a constant size per element, thus in principle scaling only with the number of parameters (*i.e.*, with the total *number-of-connections*).

[1] On leave of absence from the "Politehnica" University of Bucharest, Computer Science Department, Spl. Independenței 313, RO-77206 Bucharest, România.

[2] On leave of absence from the Polish Academy of Sciences, Institute of Fundamental Technological Research, Świetokrzyska 21, 00-049 Warsaw, Poland.

It is worth emphasising that it is anyhow desirable to limit the range of parameter values (Wray & Green, 1995) for VLSI implementations because: (*i*) the maximum value of the *fan-in* (Walker *et al.*, 1989); and (*ii*) the maximal ratio between the largest and the smallest *weight* cannot grow over a certain (technological) limit. The paper will discuss the influence of limiting the *fan-in* on the *size* optimality of two different *size*-optimal solutions. It is structured as follows: Section 2 presents pervious results, while in Section 3 we shall prove our main claims. Conclusions and open problems for research complete the paper.

2. Previous Results

One starting point is a classic construction for synthesising one BF with *fan-in* 2 AND-OR gates. It was extended to the multioutput case and modified to apply to NNs.

Proposition 1 (Theorem 3 from (Horne & Hush, 1994)) *Arbitrary Boolean functions* $f: \{0,1\}^n \rightarrow \{0,1\}^\mu$ *can be implemented in a NN of perceptrons restricted to fan-in 2 with a node complexity of* $\Theta\{\mu\, 2^n/(n+\log\mu)\}$ *and requiring* $\boldsymbol{O}(n)$ *layers.*

Sketch of proof The idea is to decompose each output BF into two subfunctions using Shannon's Decomposition (Shannon, 1949):

$$f(x_1\, x_2 \ldots x_{n-1} x_n) = \bar{x}_1 f_0(x_2 \ldots x_{n-1} x_n) + x_1 f_1(x_2 \ldots x_{n-1} x_n)\,.$$

By doing this recursively for each subfunction, the output BFs will—in the end—be implemented by binary trees. Horne & Hush (1994) use a trick for eliminating most of the lower level nodes by replacing them with a subnetwork that computes *all the possible BFs* needed by the higher level nodes. Each subcircuit eliminates one variable and has three nodes (one OR and two ANDs). Thus the upper tree has:

$$\begin{aligned} size_{upper} &= 3\mu \cdot \sum_{i=0}^{n-q-1} 2^i \\ &= 3\mu\,(2^{n-q}-1) \end{aligned} \qquad (1)$$

nodes, and:

$$depth_{upper} = 2\,(n-q)\,.$$

The subfunctions now depend on only q variables, and a lower subnetwork that computes all the possible BFs of q variables is built. It has:

$$\begin{aligned} size_{lower} &= 3 \cdot \sum_{i=1}^{q} 2^{2^i} \\ &< 4 \cdot 2^{2^q} \end{aligned} \qquad (2)$$

nodes, and:

$$depth_{lower} = 2\,q$$

(see *Figure 2* in (Horne & Hush, 1994)). That q which minimises

$$size_{BFs} = size_{upper} + size_{lower}$$

is determined by solving $d(size_{BFs})/dq = 0$, and gives:

$$q \approx \log\{n + \log\mu - 2\log(n+\log\mu)\}. \qquad (3)$$

By substituting (3) in (1) and (2), the minimum *size*:

$$\begin{aligned} size_{BFs} &\approx 3\mu\, 2^{n-q} \\ &= 3\,\mu \cdot 2^n/(n+\log\mu) \end{aligned}$$

is determined. ❑

Proposition 2 (Theorem 1 from (Red'kin, 1970)) *The complexity realisation (i.e., number of threshold elements) of* $IF_{n,m}$ *(the class of Boolean functions* $f(x_1\, x_2 \ldots x_{n-1} x_n)$ *that have exactly m groups of ones) is at most* $2\sqrt{2m} + 3$.

The construction has: a first layer of $\lceil (2m)^{1/2} \rceil$ TGs (COMPARISONs) with *fan-in* $= n$ and *weights* $\leq 2^{n-1}$; a second layer of $2\lceil (m/2)^{1/2} \rceil$ TGs of *fan-in* $= n + \lceil (2m)^{1/2} \rceil$ and *weights* $\leq 2^n$; one more TG of *fan-in* $= 2\lceil (m/2)^{1/2} \rceil$ and *weights* $\in \{-1, +1\}$ in the third layer.

3. Limited Fan-in and Optimal Solutions

Proposition 3 (this paper) *Arbitrary Boolean functions* $f: \{0,1\}^n \rightarrow \{0,1\}^\mu$ *can be implemented in a NN of perceptrons restricted to fan-in* Δ *in* $\boldsymbol{O}(n/\log\Delta)$ *layers.*

Proof We use the same approach as Horne & Hush (1994) for the case when the *fan-in* is limited to Δ. Each output BF can be decomposed in $2^{\Delta-1}$ subfunctions (*i.e.*, $2^{\Delta-1}$ AND gates). The OR gate would have $2^{\Delta-1}$ inputs. Thus we have to decompose it in a Δ-ary tree of *fan-in* $= \Delta$ OR gates. This decomposition step eliminates $\Delta - 1$ variables and generates a Δ-ary tree having:

$$depth = 1 + \lceil (\Delta - 1)/\log\Delta \rceil\,,$$

and:

$$size = 2^{\Delta-1} + \lceil (2^{\Delta-1} - 1)/(\Delta - 1) \rceil\,.$$

Repeating this procedure recursively k times, we have:

$$depth_{upper} = k \cdot \{1 + \lceil (\Delta - 1)/\log\Delta \rceil\} \qquad (4)$$

$$size_{upper} = \{2^{\Delta-1} + \lceil (2^{\Delta-1} - 1)/(\Delta - 1) \rceil\} \times \times \sum_{i=0}^{k-1} 2^{i(\Delta-1)}$$

$$= size \cdot \{2^{k(\Delta-1)} - 1\} / (2^{\Delta-1} - 1)$$

$$\cong 2^{k(\Delta-1)} \cdot (1 + 1/\Delta)$$

$$\approx 2^{k\Delta - k} \qquad (5)$$

where the subfunctions depend only on $q = n - k\Delta$ variables. We now generate all the possible subfunctions of q variables with a subnetwork of:

$$depth_{lower} = \lfloor (n - k\Delta)/\Delta \rfloor \{1 + \lceil (\Delta - 1)/\log\Delta \rceil\} \quad (6)$$

$$size_{lower} = \{2^{\Delta-1} + \lceil (2^{\Delta-1} - 1)/(\Delta - 1) \rceil\} \times \times \sum_{i=1}^{\lfloor n/\Delta \rfloor - k} 2^{2^{n-k\Delta-i\Delta}}$$

$$= size \cdot \{2^{2^0} + 2^{2^\Delta} + \ldots + 2^{2^{n-(k+1)\Delta}}\}$$

$$< (size + 1) \cdot 2^{2^{n-(k+1)\Delta}} \qquad (7)$$

$$\approx 2^{\Delta} \cdot 2^{2^{n-k\Delta-\Delta}}. \qquad (8)$$

The inequality (7) can be proved by induction. Clearly,

$$size \cdot 2^{2^0} < (size + 1) \cdot 2^{2^0}.$$

Let us consider the statement true for α; we prove it for $\alpha + 1$:

$$size \cdot \{2^{2^0} + 2^{2^\Delta} + \ldots + 2^{2^{\alpha\Delta}}\} + size \cdot 2^{2^{(\alpha+1)\Delta}} < size \cdot 2^{2^{(\alpha+1)\Delta}} + 2^{2^{(\alpha+1)\Delta}}$$

$$size \cdot \{2^{2^0} + 2^{2^\Delta} + \ldots + 2^{2^{\alpha\Delta}}\} < (size + 1) \cdot 2^{2^{\alpha\Delta}}$$

(due to hypothesis), thus:

$$(size + 1) \cdot 2^{2^{\alpha\Delta}} < 2^{2^{(\alpha+1)\Delta}}$$

and computing the logarithm of the left side:

$$2^{\alpha\Delta} + \log(size + 1) = 2^{\alpha\Delta} + \log\{2^{\Delta-1} + \lceil (2^{\Delta-1} - 1)/(\Delta - 1) \rceil\} < 2^{\alpha\Delta} + \log\{2^{\Delta-1} + 2^{\Delta-1}/\Delta + 1\} < 2^{\alpha\Delta} + \Delta < 2^{(\alpha+1)\Delta}.$$

From (4) and (6) we can estimate $depth_{BFs}$, and from (5) and (8) $size_{BFs}$ as:

$$depth_{BFs} = \{k + \lfloor (n - k\Delta)/\Delta \rfloor\} \{1 + \lceil (\Delta - 1)/\log\Delta \rceil\} = (n/\Delta) \cdot (\Delta/\log\Delta + 1) \qquad (9)$$

$$\approx n/\log\Delta = \boldsymbol{O}(n/\log\Delta)$$

$$size_{BFs} = \mu \cdot size \cdot \{2^{k(\Delta-1)} - 1\}/(\Delta - 1) + + (size + 1) \cdot 2^{2^{n-(k+1)\Delta}}$$

$$\approx \mu \cdot 2^{k\Delta - k} + 2^{\Delta} \cdot 2^{2^{n-k\Delta-\Delta}} \qquad (10)$$

concluding the proof. ❑

Proposition 4 (this paper) *All the critical points of the size* $size_{BFs}(\mu, n, k, \Delta)$ *are relative minimum and are situated in the (close) vicinity of the parabola* $k\Delta \approx n - \log(n + \log\mu)$.

Proof To determine the critical points, we equate the partial derivatives to zero. Starting from the approximation of $size_{BFs}$ we compute $\partial size_{BFs}/\partial k = 0$:

$$\mu \cdot 2^{k\Delta - k} (\ln 2)(\Delta - 1) + + 2^{\Delta} \cdot 2^{2^{n-k\Delta-\Delta}} (\ln 2) \cdot 2^{n-k\Delta-\Delta} (\ln 2) \cdot (-\Delta) = 0$$

$$\{\mu(\Delta - 1)/\Delta/(\ln 2)\} \cdot 2^{2k\Delta - k - n} = 2^{2^{n-k\Delta-\Delta}}$$

and using the notations $k\Delta = \gamma$, $\beta = \mu(\Delta - 1)/(\Delta \ln 2)$, and taking logarithms of both sides:

$$\log\beta + 2\gamma - k - n = 2^{n-\gamma-\Delta} \qquad (11)$$

which has an approximate solution $\gamma \approx n - \log(n + \log\mu)$. The same result can be obtained by computing with finite differences (instead of approximating the partial derivative):

$$size_{BFs}(\mu, n, k + 1, \Delta) - size_{BFs}(\mu, n, k, \Delta) = 0$$

$$size \cdot \{\mu \cdot 2^{k\Delta - k} - 2^{2^{n-k\Delta-\Delta}}\} = 0$$

$$\mu \cdot 2^{k\Delta - k} = 2^{2^{n-k\Delta-\Delta}}$$

and after taking twice the logarithm of both sides and using the same notations we have:

$$\log\{\log\mu + \gamma(1 - 1/\Delta)\} = n - \gamma - \Delta$$

$$\gamma = n - \{\Delta + \log(1 - 1/\Delta)\} - - \log\{\gamma + \Delta/(\Delta - 1) \cdot \log\mu\}$$

$$\approx n - \Delta - \log(\gamma + \log\mu), \qquad (12)$$

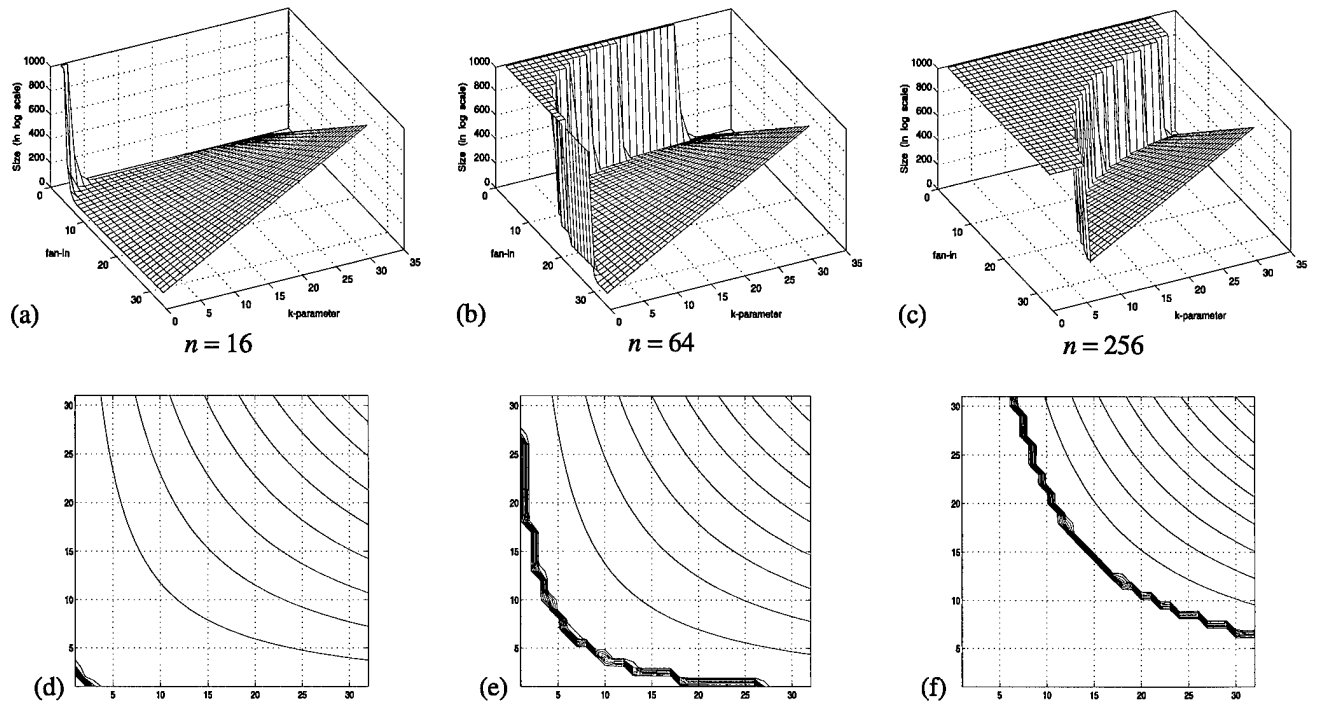


Figure 1. *Size* (in logarithmic scale) of NNs implementing arbitrary BFs as a function of *fan-in* and k, for: (a) $n = 16$; (b) $n = 64$; (c) $n = 256$ (clipped at 2^{1000}), and the contour plots for the same cases (d), (e), (f).

which has as the approximate solution:

$$\gamma = n - \log(n + \log\mu) .$$

Starting again from (10), we compute $\partial size_{BFs} / \partial\Delta = 0$:

$$\mu\, 2^{k\Delta - k} (\ln 2)\, k + 2^{\Delta} (\ln 2)\, 2^{2^{n-k\Delta-\Delta}} +$$

$$+ 2^{\Delta} 2^{2^{n-k\Delta-\Delta}} (\ln 2)\, 2^{n-k\Delta-\Delta} (\ln 2)\, (-k) = 0$$

$$\mu k \cdot 2^{\gamma - k} = k (\ln 2)\, 2^{n-\gamma} \cdot 2^{2^{n-\gamma-\Delta}} - 2^{\Delta} \cdot 2^{2^{n-\gamma-\Delta}}$$

$$\mu k \cdot 2^{\gamma - k} \cdot 2^{\gamma - n}$$

$$= k (\ln 2)\, 2^{2^{n-\gamma-\Delta}} - 2^{\Delta} \cdot 2^{\gamma - n} \cdot 2^{2^{n-\gamma-\Delta}}$$

$$\mu k \cdot 2^{2\gamma - k - n} = \{k (\ln 2) - 2^{\gamma + \Delta - n}\} \cdot 2^{2^{n-\gamma-\Delta}}$$

$$(\mu / \ln 2) \cdot 2^{2\gamma - k - n}$$

$$= \{1 - 2^{\gamma + \Delta - n} / (k \ln 2)\} \cdot 2^{2^{n-\gamma-\Delta}}$$

which—by neglecting $2^{\gamma+\Delta} / \{k (\ln 2) \cdot 2^{n}\}$—gives:

$$\log\beta + 2\gamma - k - n = 2^{n-\gamma-\Delta}$$

i.e., the same equation as (11). These show that the critical points are situated in the (close) vicinity of the parabola $k\Delta \approx n - \log(n + \log\mu)$. The fact that they are relative minimum has also been proven (Beiu 1997b). ❑

The exact *size* has been computed for many different values of n, μ, Δ and k. One example of those extensive simulations is plotted in *Figure 1*. From *Figure 1(a)* it may seem that k and Δ have almost the same influence on $size_{BFs}$. The discrete parabola-like curves (the one closer to the axes is approximately $k\Delta \approx n - \log(n + \log\mu)$) can be seen in *Figure 1(b)*.

Table 1. Minimum $size_{BFs}$ for different values of n and $\mu = 1$.

n	$8 = 2^3$	$16 = 2^4$	$32 = 2^5$	$64 = 2^6$	$128 = 2^7$	$256 = 2^8$	$512 = 2^9$	$1024 = 2^{10}$	$2048 = 2^{11}$
size	110	1470	349,530	1.611×10^9	6.917×10^{18}	5.104×10^{38}	2.171×10^{76}	1.005×10^{154}	1.685×10^{307}
Δ	4	8	16	2	2	2	2	2	2

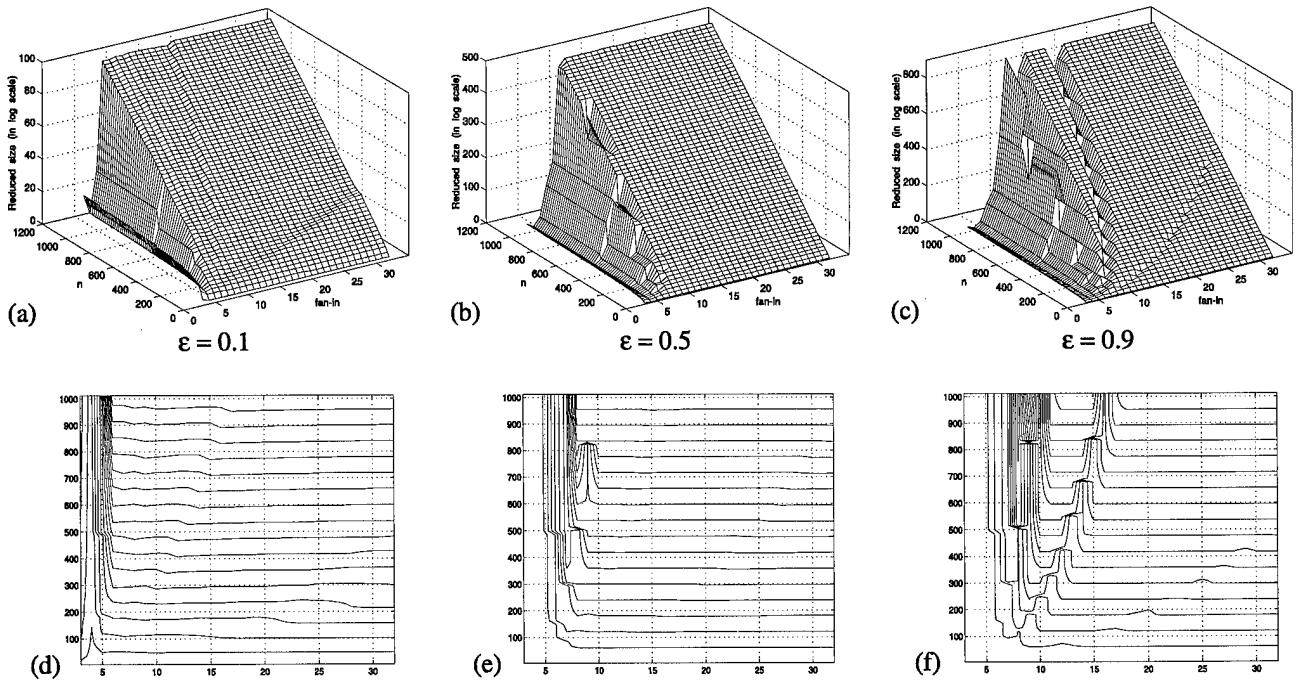


Figure 2. The reduced *size* (in logarithmic scale) of NNs implementing $IF_{n,m}$ functions for $m = 2^{\varepsilon n}$: (a) $\varepsilon = 0.1$; (b) $\varepsilon = 0.5$; (c) $\varepsilon = 0.9$; and the contour plots for the same cases (d), (e), (f). The lowest values are obtained for very small constant *fan-in* values.

Proposition 5 (this paper) *The absolute minimum of* $size_{BFs}$ *is obtained for fan-in* $\Delta = 2$.

Sketch of proof We will analyse only the critical points by using the approximation $k\Delta \approx n - \log n$. Intuitively the claim can be understood if we replace this value in (10):

$$\begin{aligned} size^*_{BFs} &\approx \mu \cdot 2^{n-\log n-k} + 2^{\Delta} \cdot 2^{2^{n-n+\log n-\Delta}} \\ &< \mu \cdot 2^{n-\log n} + 2^{\Delta} \cdot 2^{2^{\log n}} \\ &= \mu \cdot 2^n/n + 2^{\Delta} \cdot 2^n, \end{aligned}$$

which clearly is minimised for $\Delta = 2$. ❑

The detailed proof relies on computing the *size*: $size_{BFs}(n, \mu, k, \Delta)$ for $k \approx (n - \log n)/\Delta$, and then showing that:

$$size^*_{BFs}(n, \mu, \Delta + 1) - size^*_{BFs}(n, \mu, \Delta) > 0 .$$

Hence, the function is monotonically increasing and the minimum is obtained for the smallest *fan-in* $\Delta = 2$. Because the proof has been obtained using successive approximations, several simulation results are presented in *Table 1*. It can be seen that while for relatively small n the *size*-optimal solutions are obtained even for $\Delta = 16$, starting from $n \geq 64$ all the *size*-optimal solutions are obtained for $\Delta = 2$. It is to be mentioned that the other relative minima (on, or in the vicinity of the parabola $k\Delta \approx n - \log n$) are only slightly larger than the absolute minimum. They might be of ***practical interest*** as leading to networks having fewer layers: $n/\log\Delta$ instead of n. Last, but not least, it is to be remarked that all these relative minimum are obtained for *fan-ins* strictly lower that linear (as $\Delta \leq n - \log n$).

A similar result can be obtained for $IF_{n,m}$, as the first layer is represented by COMPARISONs (*i.e.*, $IF_{n,1}$) which can be decomposed to satisfy the limited *fan-in* condition (Beiu, 1998; Beiu & Taylor, 1996a).

Proposition 6 (Lemma 1 & Corollary 1 (Beiu et al., 1994)) *The* COMPARISON *of two n-bit numbers can be computed by a* Δ*-ary tree NN having integer weights and thresholds bounded by* $2^{\Delta/2}$ *for any* $3 \leq \Delta \leq n$.

The *size* complexity of the NN implementing one $IF_{n,m}$ function is (Beiu *et al.*, 1994):

$$size_{IF} = 2nm \cdot \left\{ \frac{1}{\Delta/2} + \ldots + \frac{1}{(\Delta/2)^{depth_{IF}}} \right\}, \quad (13)$$

Table 2 (from (Beiu, 1996b)). Different estimates of AT^2 for **SRK** (Siu *et al.*, 1991), **B_4** and **B_log** (Beiu *et al.*, 1994; Beiu, 1996b), **ROS** (Roychowdhury *et al.*, 1994) and **VCB** (Vassiliadis *et al.*, 1996).

Area \ *Delay*	*Depth*	*Fan-in*	*Length*
Size	$AT^2_{VCB} = O(\sqrt{n})$	$AT^2_{B\_4} = O(n \log^2 n)$ (2)	$AT^2_{VCB} = O(n^2\sqrt{n})$
	$AT^2_{ROS} = O(n/\log n)$	$AT^2_{B\_log} = O[n \log^3 n / \log^2(\log n)]$	$AT^2_{ROS} \cong 3 \cdot n^3/\log n$
	$AT^2_{SRK} = O(n)$	$AT^2_{VCB} = O(n\sqrt{n})$	$AT^2_{B\_log} \cong 4 \cdot n^3/\log n$
	$AT^2_{B\_log} = O[n \log n / \log^2(\log n)]$	$AT^2_{ROS} = O(n^3/\log^3 n)$	$AT^2_{B\_4} \cong 4 \cdot n^3$
	$AT^2_{B\_4} = O(n \log^2 n)$	$AT^2_{SRK} = O(n^3)$	$AT^2_{SRK} \cong 27 n^3/4$
Σ_{NN} *fan-ins*	$AT^2_{VCB} = O(n)$	$AT^2_{B\_4} = O(n \log^2 n)$ (3)	$AT^2_{B\_log} \cong 4n^3$
	$AT^2_{B\_log} = O[n \log^2 n / \log^2(\log n)]$	$AT^2_{B\_log} = O[n \log^4 n / \log^2(\log n)]$	$AT^2_{VCB} \cong 4n^3$
	$AT^2_{B\_4} = O(n \log^2 n)$	$AT^2_{VCB} = O(n^2)$	$AT^2_{B\_4} \cong 5n^3$
	$AT^2_{ROS} = O(n^2/\log^2 n)$	$AT^2_{ROS} = O(n^4/\log^4 n)$	$AT^2_{ROS} = O(n^4/\log^2 n)$
	$AT^2_{SRK} = O(n^2)$	$AT^2_{SRK} = O(n^4)$	$AT^2_{SRK} = O(n^4)$
$\Sigma_{NN}(\Sigma_i \lvert w_i \rvert + \lvert\theta\rvert)$	$AT^2_{B\_4} = O(n \log^2 n)$ (1)	$AT^2_{B\_4} = O(n \log^2 n)$ (4)	$AT^2_{B\_4} = O(n^3)$
	$AT^2_{B\_log} = O[n\sqrt{n} \log n / \log^2(\log n)]$	$AT^2_{B\_log} = O[n\sqrt{n} \log^3 n / \log^2(\log n)]$	$AT^2_{B\_log} = O(n^3\sqrt{n}/\log n)$
	$AT^2_{ROS} = O(n^2/\log n)$	$AT^2_{ROS} = O(n^4/\log^3 n)$	$AT^2_{ROS} = O(n^4/\log n)$
	$AT^2_{SRK} = O(n^2)$	$AT^2_{SRK} = O(n^4)$	$AT^2_{SRK} = O(n^4)$
	$AT^2_{VCB} = O(n^{1/2} \cdot 2^{\sqrt{n}})$	$AT^2_{VCB} = O(n^{3/2} \cdot 2^{\sqrt{n}})$	$AT^2_{VCB} = O(n^{5/2} \cdot 2^{\sqrt{n}})$

where:

$$depth_{IF} = \lceil \log n / (\log\Delta - 1) \rceil,$$

but a substantial enhancement is obtained if the *fan-in* is limited. The maximum number of *different BFs* which can be computed in each layer is:

$$(2n/\Delta)\,2^{\Delta},\quad \frac{2n/\Delta}{\Delta/2}\,2^{\Delta(\Delta/2)},\ \ldots,\ \frac{2n/\Delta}{(\Delta/2)^{depth_{IF}-1}} \cdot 2^{\Delta(\Delta/2)^{depth_{IF}-1}}. \quad (14)$$

For large m (needed for achieving a certain precision (Beiu, 1998; Wray & Green, 1995)), and/or large n, the first terms of the sum (13) will be larger than the equivalent ones from (14). This is equivalent to the trick from (Horne & Hush, 1994), as the lower levels will compute *all the possible functions* using only limited *fan-in* COMPARISONs. Hence, the optimum *size* becomes:

$$size^*_{IF} = 2n \cdot \left\{ \sum_{i=1}^{k} \frac{2^{\Delta(\Delta/2)^{i-1}}}{\Delta(\Delta/2)^{i-1}} + \sum_{i=k+1}^{depth_{IF}} \frac{m}{(\Delta/2)^i} \right\}.$$

Following similar steps to the ones used in *Proposition 5*, it is possible to show that the minimum *size* is obtained for $\Delta = 3$. To get a better understanding, we have done simulations by considering that $m = 2^{\varepsilon n}$. Some results can be seen in *Figure 2* (for $\varepsilon = 0.99$).

We mention here that similar results ($\Delta = 6 \ldots 9$), based on closer estimates of *area* and *delay* have been proved for VLSI-efficient implementations of $IF_{n,m}$ functions (Beiu 1996b, 1997a). Different complexity estimates for COMPARISON can be seen in *Table 2*. All of these support the claim that small constant *fan-in* NNs can be *size*- and VLSI-optimal.

4. Conclusions and Open Problems

In this paper, we have extended a result from Horne & Hush (1994) valid for *fan-in* $\Delta = 2$ to arbitrary *fan-ins*, and have shown that the minimum *size* is obtained for small (constant) *fan-ins*. We have also shown that, using their construction, it is possible to obtain 'good' (*i.e.*, relative minimum) solutions for *fan-ins* strictly lower than linear. The same results have been obtained for the *size*-optimal

solution of Red'kin (1970). The main conclusions are that: (*i*) there are interesting *fan-in* dependent *depth-size* (and *area-delay*) tradeoffs; and (*ii*) there are optimal solutions having small constant *fan-in* values. Future work will concentrate on linking these results with the entropy of the data-set, and with principles like "Occam's razor" (Zhang & Mühlenbein, 1993) and "minimum description length", as well as trying to find closer estimates for mixed analog/digital implementations. The main conclusion is that VLSI-optimal solutions can be obtained for small (constant) fan-ins, and with respect to that we mention here that there are similar small constants relating to our capacity of processing information (Miller, 1956).

References

Abu-Mostafa, Y.S. (1988) Connectivity Versus Entropy. In D.Z. Anderson (ed.), *Neural Information Processing Systems*, 1-8. New York, NY: AIP.

Bartlett, P.L. (1996) The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights Is More Important than the Size of the Network. *Tech. Rep.*, Dept. Sys. Eng., Australian Natl. Univ., Canberra [short version as (1997) in M.C. Mozer, M.I. Jordan & T. Petsche (eds.): *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press].

Beiu, V., Peperstraete, J.A., Vandewalle, J. & Lauwereins, R. (1994) Area-Time Performances of Some Neural Computations. In P. Borne, T. Fukuda & S.G. Tzafestas (eds.), *SPRANN'94*, 664-668, Lille, France: GERF EC.

Beiu, V. & Taylor, J.G. (1996a) On the Circuit Complexity of Sigmoid Feedforward Neural Networks. *Neural Networks* **9**(7):1155-1171.

Beiu, V. (1996b) On the Circuit and VLSI Complexity of Threshold Gate COMPARISON. *Tech. Rep. LA-UR-96-3591*, Los Alamos Natl. Lab., USA [to appear in (1997), *Neurocomputing*].

Beiu V. (1997a) Constant Fan-In Digital Neural Networks Are VLSI-Optimal. *Tech. Rep. LA-UR-97-61*, Los Alamos Natl. Lab., USA; also as (1997) *Chapter 12* in S.W. Ellacott, J.C. Mason & I.J. Anderson (eds.), *Mathematics of Neural Networks: Models, Algorithms and Applications*, 89-94, Boston, MA: Kluwer Academic.

Beiu, V. (1997b) When Constants Are Important. *Tech. Rep. LA-UR-97-226*, Los Alamos Natl. Lab., USA. In I. Dumitrache (ed.), *Proc. Control Systems & Computer Science CSCS-11*, vol. **2**, 106-111, Bucharest, Romania: UPB Press.

Beiu, V. (1998) *VLSI Complexity of Discrete Neural Networks*. Newark, NJ: Gordon & Breach.

Bruck, J. & Goodmann, J.W. (1988) On the Power of Neural Networks for Solving Hard Problems. In D.Z. Anderson (ed.), *Neural Information Processing Systems*, 137-143. New York, NY: AIP [also as (1990), *J. Complexity* **6**:129-135].

Hammerstrom, D. (1988) The Connectivity Analysis of Simple Association –or– How Many Connections Do You Need. In D.Z. Anderson (ed.), *Neural Information Processing Systems*, 338-347. New York, NY: AIP.

Horne, B.G. & Hush, D.R. (1994) On the Node Complexity of Neural Networks. *Neural Networks* **7**(9):1413-1426.

Miller G.A. (1956) The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psych. Rev.* **63**:71-97.

Phatak, D.S. & Koren, I. (1994) Connectivity and Performances Tradeoffs in the Cascade Correlation Learning Architecture. *IEEE Trans. on Neural Networks* **5**(6): 930-935.

Red'kin, N.P. (1970) Synthesis of Threshold Circuits for Certain Classes of Boolean Functions. *Kibernetika* **5**:6-9 [English translation as (1973) *Cybernetics* **6**(5):540-544].

Roychowdhury, V.P., Orlitsky, A. & Siu, K.-S. (1994) Lower Bounds on Threshold and Related Circuits Via Communication Complexity. *IEEE Trans. on Information Theory* **40**(2):467-474.

Shannon, C. (1949) The Synthesis of Two-Terminal Switching Circuits. *Bell Sys. Tech. J.* **28**(1):59-98.

Siu, K.-Y., Roychowdhury, V.P. & Kailath, T. (1991) Depth-Size Tradeoffs for Neural Computations. *IEEE Trans. on Comp.* **40**(12):1402-1412.

Vassiliadis, S., Cotofana, S. & Berteles, K. (1996) 2-1 Addition and Related Arithmetic Operations with Threshold Logic. *IEEE Trans. on Comp.* **45**(9):1062-1068.

Walker, M.R., Haghighi, S., Afghan, A. & Akers, L.A. (1989) Training a Limited-Interconnect, Synthetic Neural IC. In D.S. Touretzky (ed.), *Advances in Neural Information Processing Systems*, 777-784. San Mateo, CA: Morgan Kaufmann.

Williamson, R.C. (1990) ε-Entropy and the Complexity of Feedforward Neural Networks. In R.P. Lippmann, J.E. Moody & D.S. Touretzky (eds.), *Advances in Neural Information Processing Systems*, 946-952. San Mateo, CA: Morgan Kaufmann.

Wray, J. & Green, G.G.R. (1995) Neural Networks, Approximation Theory, and Finite Precision Computation. *Neural Networks* **8**(1):31-37.

Zhang, B.-T. & Mühlenbein, H. (1993) Genetic Programming of Minimal Neural Networks Using Occam's Razor. *Tech. Rep. GMD 0734*, Schloß Birlinghoven, St. Augustin, Germany [also as (1993) *Complex Systems* **7**(3):199-220].

Report Number (14) LA-UR--97-4359
CONF-980216--

Publ. Date (11) 199803
Sponsor Code (18) DOE/MA, XF
UC Category (19) UC-900, DOE/ER

DOE