Title:

CONSTANT FAN-IN DIGITAL NEURAL NETWORKS ARE VLSI-OPTIMAL

Author(s):

V. Beiu

Submitted to:

Mathematics of Neural Networks and Applications
3-7 July 1995
Oxford, England

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of its employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
May 1995

Dear Prof. Bein

Re: MANNA '95, July 3 - 7 at Lady Margaret Hall in Oxford. (United Kingdom)

Thank you for your registration form and cheque, we are very pleased that you can attend the MANNA Conference. Enclosed are joining instructions and various items of information, including travel advice and a provisional programme - the meeting starts with lunch on Monday and ends with lunch on Friday.

Those giving talks should bring a draft paper (6 pages including short abstract). We have also enclosed the frontispiece from the journal "Annals of Mathematics and Artificial Intelligence" in which the conference proceedings will appear.

We look forward to seeing you in Oxford.

Yours sincerely

John Mason

Enc: Joining instructions
Provisional programme
Annals of Mathematics & Artificial Intelligence frontispiece
Maps
Bus timetables
REGISTRATION FORM

Surname ............................................. Forenames .............................................
Title (Professor, Dr, etc.) .................................................................
Institution/company .................................................................
Address .................................................................
.................................................................
.................................................................
.................................................................
London WC2R 2LS, U.K.
Tel: +44-171-873.2287 Fax: +44-171-873.2012 Email baui@kcl.ac.uk

I shall be accompanied by .................................................................

FEES

I enclose a cheque (sterling) made payable to The University of Huddersfield to cover the cost of:

Conference Fee ................................................................. £195.00
Residential Fee (full board in single en-suite room
Monday lunch to Friday lunch, including conference dinner) £237.00
Extra Residential Fee (For those requiring Bed & Breakfast
in single en-suite room for night of 2 July) £31.95
Lunch for day attendants ................................................................. days @ £7.75 per day
+ 15% surcharge for applications received after 17 May 1995

Total £195.00

Special dietary requirements:

I will be accompanied by my family and request information about local hotels...(please tick)

Please return to: Mrs Rosalyn Hawkins, School of Computing and Mathematics,
University of Huddersfield, Queensgate, Huddersfield, West Yorks, HD1 3DH

Tel 0484 472150 Fax 0484 421106 Email r.hawkins@hud.ac.uk
Dear Prof Beiu

I am pleased to inform you that your paper

Constant fan-in neural networks are VLSI-optimal

is now academically acceptable for publication in the proceedings of the meeting which will appear as a special issue of the *Annals of Mathematics and Artificial Intelligence* (but see below).

If you have not already done so, please supply a version in machine readable form. This can be emailed to me at the address above, or sent on a floppy disc. I can read the following formats: DOS (up to 1.44Mb), Mac, Atari, Acorn. Please indicate the disc or email encoding format, the author and title, and the package used to produce the paper: if it is anything esoteric it would be best just to supply it as plain ASCII. Diagrams can be sent as PostScript.

Space in the proceedings is going to be very tight because of the good number of quality papers submitted, and it is necessary to restrict the length to six pages in the Baltzer style. This is quite restrictive, and may be only four or five pages in an ordinary A4 style. If your paper appears to be too long, I will notify you separately.

Yours sincerely

S.W.Ellacott
Dear Prof Beiu,

You are doubtless wondering what has happened to the proceedings of MANNA 95. Well the truth is that we have had to change publishers. AMAI wanted us to produce an edition with fewer, longer papers, but your editors regarded this as unacceptable. After considerable negotiation, both sides reluctantly agreed that the only solution was to look elsewhere. I am pleased to tell you that we have now signed a contract with Kluwer Academic publishers to produce the proceedings in their Operations Research/Computer Science Interfaces series (series editor Prof. Ramesh Sharda) under the provisional title

Mathematics of Neural Networks: Models, Algorithms and Applications.

I will let you know further details, such as the ISBN, as soon as they are available. In view of the already considerable delay, we now intend to get the book out very quickly, with a publication date early in the new year. We have had to reformat the LaTeX into the Kluwer style, so it will be necessary for proofs to be returned to authors for checking: expect these in a few weeks. Minor amendments will be undertaken by us, but anyone wishing to do their own reformatting can obtain the relevant style files from me. (Please do not use the one on the Kluwer ftp site: the version we use has been modified for this particular book.) Provisionally, we plan to supply one free copy of the book per paper: additional copies should be available for around US$60.

I enclose a copy of the usual copyright waiver. Please will the contributing author (only) sign this and send it to me by return of post: ignore the different instructions on the document itself. (If there are any difficulties, contact me via email.) Of course, if you wish to withdraw your paper in view of the delay and change of publication details, we will understand, but we hope that you will be happy that the editors have found a very satisfactory solution to publishing these proceedings, and like ourselves will be very keen to have your work included.

Yours sincerely

S. W. Ellacott

S.W.Ellacott
Dear Prof Beiu,

The proof of your paper for the proceedings of MANNA 95 is enclosed. Please check it and return it to me by 6th January 1997.

The papers have had to be reformatted in the Kluwer style, so watch out for problems such as mis-formatted equations. In addition, the references have been edited into a standard style, and in some cases figures have been redrawn, so please pay particular attention to these.

Again, the editors thank you for your contribution and apologise for the delay in publication. Any queries should be emailed to me at the address above. We do not as yet have any further bibliographic information since my previous letter.

Yours sincerely,

Steve Ellacott
(for the editors)
CONSTANT FAN-IN DIGITAL NEURAL NETWORKS
ARE VLSI-OPTIMAL

V. Beiu
Los Alamos National Laboratory, Division NIS-1, Los Alamos, New Mexico 87544, USA

The paper presents a theoretical proof revealing an intrinsic limitation of digital VLSI technology: its inability to cope with highly connected structures (e.g., neural networks). We are in fact able to prove that efficient digital VLSI implementations (known as VLSI-optimal when minimising the $AT^2$ complexity measure — $A$ being the area of the chip, and $T$ the delay for propagating the inputs to the outputs) of neural networks are achieved for small-constant fan-in gates. This result builds on quite recent ones dealing with a very close estimate of the area of neural networks when implemented by threshold gates, but it is also valid for classical Boolean gates. Limitations and open questions are presented in the conclusions.

Keywords: neural networks, VLSI, fan-in, Boolean circuits, threshold circuits, $F_{n,m}$ functions.

1 Introduction

In this paper a network will be considered an acyclic graph having several input nodes (inputs) and some (at least one) output nodes (outputs). The nodes are characterised by fan-in (the number of incoming edges — denoted by $\Delta$) and fan-out (the number of outgoing edges), while the network has a certain size (the number of nodes) and depth (the number of edges on the longest input to output path). If with each edge a synaptic weight is associated and each node computes the weighted sum of its inputs to which a non-linear activation function is then applied (artificial neuron), the network is a neural network (NN):

$$\mathbf{z}_k = (z_0, ..., z_n - 1) \in \mathbb{R}^n, k = 1, ..., m,$$ and $f(\mathbf{z}_k) = \sigma \left( \sum_{i=0}^{n-1} w_i z_i + \theta \right), \tag{1}$

with $w_i \in \mathbb{R}$ the synaptic weights, $\theta \in \mathbb{R}$ known as the threshold, and $\sigma$ a non-linear activation function. If the non-linear activation function is the threshold (logistic) function, the neurons are threshold gates (TGs) and the network is just a threshold gate circuit (TGC) computing a Boolean function (BF). The cost functions associated to a NN are depth and size. These are linked to $T \approx$ depth and $A \approx$ size of a VLSI chip. Unfortunately, NNs do not closely follow these proportionalities as:

- the area of the connections counts [2, 3];
- the area of one neuron is related to its associated weights.

That is why the size and depth complexity measures are not the best criteria for ranking different solutions when going to silicon [11]. Several authors have taken into account the fan-in [1, 9, 10, 12], the total number of connections, the total number of bits needed to represent the weights [8, 15] or even more precise approximations like the sum of all the weights and thresholds [2-7]:

$$\text{area} \propto \sum_{\text{all neurons}} \left( \sum_{i=0}^{n-1} |w_i| + |\theta| \right). \tag{2}$$

An equivalent definition of 'complexity' for a NN is $\sum_{i=0}^{n-1} w_i^2$ [16]. It is worth mentioning that there are also several sharp limitations for VLSI implementations like: (i) the maximal value of the fan-in cannot grow over a certain limit; (ii) the
maximal ratio between the largest and the smallest weight. For simplification, in the following we shall consider only NNs having n binary inputs and k binary outputs.

If real inputs and outputs are needed, it is always possible to quantize them up to a certain number of bits such as to achieve a desired precision. The fan-in of a gate will be denoted by Δ and all the logarithms are taken to base 2 except mentioned otherwise. Section 2 will present previous results for which proofs have already been given [2-7]. In section 3 we shall prove our main claim while also showing several simulation results.

2 Background

A novel synthesis algorithm evolving from the decomposition of COMPARISON has recently been proposed. We have been able to prove that [2, 3]:

Proposition 1 The computation of COMPARISON of two n-bit numbers can be realised by a Δ-ary tree of size $O(n/Δ)$ and depth $O(\log n/\log Δ)$ for any integer fan-in $2 \leq Δ \leq n$.

A class of Boolean functions $F_Δ$ having the property that $∀f_Δ ∈ F_Δ$ is linearly separable has afterwards been introduced as: “the class of functions $f_Δ$ of $Δ$ input variables, with $Δ$ even, $f_Δ = f_Δ(g_{Δ/2-1}, e_{Δ/2-1}, ..., g_0, e_0)$, and computing $f_Δ \triangleq \bigvee_{j=0}^{Δ/2-1} \left[ g_j \land \left( \bigwedge_{k=j+1}^{Δ/2-1} e_k \right) \right]$. By convention, we consider $\bigwedge_{i=0}^{Δ/2-1} e_i \triangleq 1$.

One restriction is that the input variables are pair-dependent, meaning that we can group the $Δ$ input variables in $Δ/2$ pairs of two input variables each: $(g_{Δ/2-1}, e_{Δ/2-1}), ..., (g_0, e_0)$, and that in each such group one variable is ‘dominant’ (i.e. when a dominant variable is 1, the other variable forming the pair will also be 1):

$$F_Δ = \left\{ f_Δ : \{(0,0),(0,1),(1,1)\}^{Δ/2} \rightarrow \{0,1\}, Δ/2 \in \mathbb{N}^* \right\},$$

$$f_Δ = \bigvee_{j=0}^{Δ/2-1} \left[ g_j \land \left( \bigwedge_{k=j+1}^{Δ/2-1} e_k \right) \right], g_i \Rightarrow e_i, i = 0, 1, ..., Δ/2 - 1.$$

Each $f_Δ$ can be built starting from the previous one $f_Δ-2$ (having a lower fan-in) by copying its synaptic weights; the constructive proof has led to [5]:

Proposition 2 The COMPARISON of two n-bit numbers can be computed by a Δ-ary tree neural network with polynomially bounded integer weights and thresholds ($≤ n^k$) having size $O(n/Δ)$ and depth $O(\log n/\log Δ)$ for any integer fan-in $3 \leq Δ ≤ \log^k n$.

For a closer estimate of the area we have used equation (2) and proved [5]:

Proposition 3 The neural network with polynomially bounded integer weights (and thresholds) computing the COMPARISON of two n-bit numbers occupies an area of $O(n \cdot 2^{Δ/2}/Δ)$ for all the values of the fan-in (Δ) in the range 3 to $O(\log n)$. The result presented there is:

$$AT^2(n, Δ) \approx \frac{2^{Δ/2}}{Δ} \cdot \frac{8nΔ - 6n - 5Δ}{Δ - 2} \cdot \frac{\log^2 n}{\log^2 Δ} = O \left( \frac{n \log^2 n \cdot 2^{Δ/2}}{Δ \log^2 Δ} \right) \quad (3)$$

and for $Δ = \log n$ this is the best (i.e. smallest) one reported in the literature.

Further, the synthesis of a class of Boolean functions $F_{n,m}$ — functions of $n$ input variables having $m$ groups of ones in their truth table [13] — has been detailed [4]:
**Proposition 4** Any function \( f \in F_{n,m} \) can be computed by a neural network with polynomially bounded integer weights (and thresholds) of depth 
\[
O(\log(mn)/\log \Delta)
\]
and size \( O(mn/\Delta) \), and occupying an area of \( O(mn \cdot 2^\Delta/\Delta) \) if \( 2m \leq 2^\Delta \) for all the values of the fan-in (\( \Delta \)) in the range 3 to \( O(\log n) \).

More precisely we have:
\[
T(n, m, \Delta) = \left\lceil \frac{\log n - 1}{\log \Delta - 1} \right\rceil + \left\lceil \frac{\log m + 1}{\log \Delta} \right\rceil = O\left( \frac{\log(mn)}{\log \Delta} \right)
\]
and
\[
A(n, m, \Delta) < 2m \cdot \left( \frac{4n \cdot 2^\Delta}{\Delta} + \frac{5(n - 1) \cdot 2^\Delta/\Delta}{\Delta(\Delta - 2)} + \Delta \cdot \frac{2m - 1}{\Delta - 1} \right) = O\left( \frac{mn \cdot 2^\Delta}{\Delta} \right)
\]
which leads to:
\[
AT^2(n, m, \Delta) = O\left( \frac{mn \cdot \log^2(mn) \cdot 2^\Delta}{\Delta \cdot \log^2 \Delta} \right).
\]

For \( 2m > 2^\Delta \) the equations are much more intricate, while the complexity values for area and for \( AT^2 \) are only reduced by a factor (equal to the fan-in [6, 7]). If we now suppose that a feed-forward NN of \( n \) inputs and \( k \) outputs is described by \( m \) examples, it can be directly constructed as simultaneously implementing \( k \) different functions from \( F_{n,m} \) [4, 6, 7].

**Proposition 5** Any set of \( k \) functions \( f \in F_{n,i}, i = 1, 2, \ldots, m, i \leq m \leq 2^\Delta - 1 \) can be computed by a neural network with polynomially bounded integer weights (and thresholds) having size \( O(m(2n + k)/\Delta) \), depth \( O(\log(mn) / \log \Delta) \) and occupying an area of \( O(mn \cdot 2^\Delta/\Delta + mk) \) if \( 2m \leq 2^\Delta \), for all the values of the fan-in (\( \Delta \)) in the range 3 to \( O(\log n) \).

The architecture has a first layer of COMPARISONS which can either be implemented using classical Boolean gates (BGs) or — as it has been shown previously — by TGs. The desired function can be synthesised either by one more layer of TGs, or by a classical two layers AND-OR structure (a second hidden layer of AND gates — one for each hypercube), and a third layer of \( k \) OR gates represents the outputs. For minimising the area some COMPARISONS could be replaced by AND gates (like in a classical disjunctive normal form implementation).

### 3 Which is the VLSI-Optimal Fan-In?

Not wanting to complicate the proofs, we shall determine the VLSI-optimal fan-in when implementing COMPARISON (in fact: \( F_{n,1} \) functions) for which the solution was detailed in **Propositions 1 to 3**. The same result is valid for \( F_{n,m} \) functions as can be intuitively expected either by comparing equations (3) and (4), or because:

- the delay is determined by the first layer of COMPARISONS; while
- the area is mostly influenced by the same first layer of COMPARISONS (the additional area for the implementing the symmetric ‘alternate addition’ [4] can be neglected).

For a better understanding we have plotted equation (3) in Figure 1.

**Proposition 6** The VLSI-optimal (which minimises the \( AT^2 \)) neural network which computes the COMPARISON of two \( n \)-bit numbers has small-constant fan-in ‘neurons’ with small-constant bounded weights and thresholds.
Figure 1  The $AT^2$ values of COMPARISON — plotted as a 3D surface — versus the number of inputs $n$ and the fan-in $\Delta$ for: (a) many inputs $n \leq 1024$ ($4 \leq \Delta \leq 20$); and (b) few inputs $n \leq 64$ ($4 \leq \Delta \leq 20$). It can be very clearly seen that a 'valley' is formed and that the 'deepest' points constantly lie somewhere between $\Delta_{\min} = 5$ and $\Delta_{\max} = 10$.

Proof: Starting from the first part of equation (3) we can compute its derivative:

$$
\frac{d(AT)^2}{d\Delta} = \frac{2^{\Delta/2} \log^2 n}{\Delta^2(\Delta - 2)^2 \log^3 \Delta} \times (8n\Delta^3 \log \Delta - 22n\Delta^2 \log \Delta + 12n \Delta \log \Delta $$

$$- 5\Delta^3 \log \Delta + 10\Delta^2 \log \Delta - \frac{16}{\ln 2} n\Delta^2 \log \Delta + \frac{24}{\ln 2} n \Delta \log \Delta $$

$$- \frac{24}{\ln 2} n \log \Delta + 10 \Delta^2 \log \Delta - \frac{32}{\ln 2} n\Delta^2 - \frac{88}{\ln 2} n \Delta - \frac{48}{\ln 2} n $$

$$+ \frac{20}{\ln 2} \Delta^2 - \frac{40}{\ln 2} \Delta $$

which — unfortunately — involves transcendental functions of the variables in an essentially non-algebraic way. If we consider the simplified 'complexity' version of equation (3) we have:

$$
\frac{d(AT)^2}{d\Delta} \approx \frac{d}{d\Delta} \left( \frac{n \log^2 n \cdot 2^{\Delta/2}}{\Delta \log^2 \Delta} \right) = \frac{2^{\Delta/2}}{\Delta \log^2 \Delta} \left( \frac{\ln 2 - 1}{\Delta} - \frac{2}{\Delta \ln \Delta} \right)
$$

which when equated to zero leads to $\ln \Delta (\Delta \ln 2 - 2) = 4$ (also a transcendental equation). This has $\Delta = 6$ as 'solution' and as the weights and the thresholds are bounded by $2^{\Delta/2}$ (Proposition 4) the proof is concluded. The proof has been obtained using several successive approximations: neglecting the ceilings, using a 'simplified' complexity estimate. That is why we present in Figure 2 exact plots of the $AT^2$ measure which support our previous claim. It can be seen that the optimal fan-in 'constantly' lies between 6 and 9 (as $\Delta_{\text{optimum}} = 6...9$, one can minimise the area by using COMPARISONs only if the group of ones has a length of $a \geq 64$ — see [4-7]). Some plots in Figure 2 are also including a TG-optimal solution denoted by SRI<14 and the logarithmic fan-in solution ($\Delta = \log n$) denoted B-lg [5].

4 Conclusions
This paper has presented a theoretical proof for one of the intrinsic limitations of digital VLSI technology: there are no 'optimal' solutions able to cope with highly connected structures. For doing that we have proven the contrary, namely that
Beiu: Constant Fan-in Digital Neural Networks are VLSI-optimal

constant fan-in NNs are VLSI-optimal for digital architectures (either Boolean or using TGs). Open questions remain concerning 'if' and 'how' such a result could be used for purely analog or mixed analog/digital VLSI circuits.

REFERENCES


ACKNOWLEDGEMENTS

This research work has been started while Dr. Beiu was with the Katholieke Universiteit Leuven, Belgium, and has been supported by a grant from the Concerted Research Action of the Flemish Community entitled: "Applicable Neural Networks". The research has been continued under the Human Capital and Mobility programme of the European Community as an Individual Research Training Fellowship ERB4001GT941815: "Programmable Neural Arrays", under contract ERBCHIBICT941741. The scientific responsibility is assumed by the author.