A Winner-Take-All circuit with proportional output

G. Oddone, G. Uneddu, S. Rovetta, and R. Zunino

DIBE – University of Genova
Via all'Opera Pia 11a – 16145 Genova (Italy)

Abstract

A linear–output WTA circuit is presented. The module is oriented to real–world applications, therefore it features relatively large input/output signal amplitudes, current mode input, and a large fan–in. The main property of the circuit is that it realizes a linear transfer of the winning input value to its output with a small error, while retaining the simple scheme found in the standard WTA circuits. A standard WTA, indicating the winning input with a 1–of–n bit coding, can be obtained from the proposed circuit with few additional components. These components are not critical, since they operate in digital mode.

1 Introduction

The class of competitive neural networks is characterized by the Winner-Take-All (WTA) function as a fundamental building block [1][2]. Neural image compression methods, an outstanding research area, are usually based on the Vector Quantization model [3]. Hence the circuital realization of any of these systems requires the design of an adequate WTA section, especially when dealing with analog systems.

The available WTA architectures can be essentially divided into dynamic and static models. The dynamic model can be traced back to the concept of biologically plausible neural interconnections, making up a network with feedback. The dynamics of the system is such that it finally reaches a stable state in which all units except one are off. When this is done in hardware, it may require a non negligible time. The static model, due to Lazzaro [4] and subsequently adopted by many researchers, requires a "limited resource" for which the units compete (usually a fixed current, provided by a generator). The circuit is still a dynamic one, but the feedback dynamics is much simpler and the time required to reach a stable state is much smaller. This allows a real-time operation. On the other hand, the separation capability is limited: if the difference between two values is under a given threshold, they cannot be discriminated. The threshold is determined by several factors, including sensitivity to variations in the parameters values and noise level, as well as the specific circuital configuration adopted.

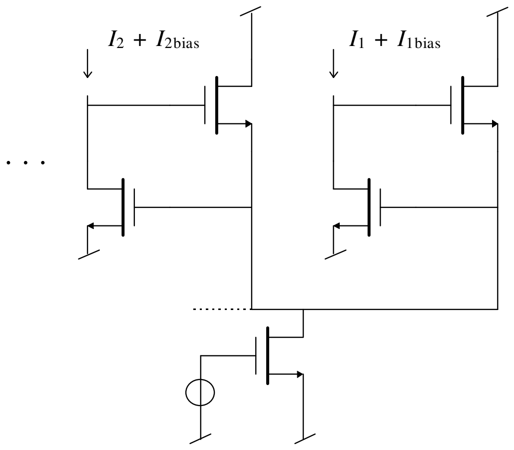
A drawback shared by both approaches is that the overall neural circuit often needs to use the effective activation value of the winner unit, because it is used to compute the adaptation steps. However, the WTA function is essentially a digital mapping (ON/OFF); hence this value is not known at the output of the WTA module. This requires additional circuitry, making the design of the subsequent circuital blocks quite complicated.

We present a modification of the static scheme, aimed at overcoming this limitation, and at the same time at improving the performance tradeoff. The stabilization time is reduced, and the resolution (that is still finite) is enhanced. The output value is proportional to the winner input. In the mathematical formalism, the standard WTA implements the "argmin" operator, whereas the proposed modification implements the "min" operator. The immediate availability of the value is very useful when implementing most learning VQ schemes, which update the reference vectors by computing an adaptation step as a function of the winner value.

2 Overview of the proposed circuit

Since a neural processor is especially useful when dealing with high-dimensional input data, we should not impose circuital constraints limiting the number of available input lines. This implies that the usual sub-threshold design is impractical, because the range of the input signals should be quite wide, and the circuit characteristics should be independent of the number of inputs connected. Hence we will accept

Figure 1: The circuit. The input signals are shown along with their respective bias components.



a somewhat larger power dissipation, but we are able to achieve a more robust design (less sensitive to parameter variations) and a higher precision in signal representation and storage.

The input impedance is required to be as low as possible, because signals are usually represented by currents. In the standard scheme the input current is injected into the drain circuit of an MOS. The drain impedance is a function of the current, and features high values for low currents. This implies that 1) at the usual working conditions, the input impedance will be high; 2) when the input range is relatively large, the output value is not linear with the winner's value. This may not be a major problem when implementing the standard WTA function (although it may affect the resolution). However, if we require the linearity property, this fact should be taken into account.

This twofold problem can be solved by a current polarization, allowing at once the choice of an operation range that is as linear as possible, and the selection of a proper impedance value.

The value of the bias current is not critical, but it should ideally be the same in all input branches. Two possible realizations are the following. The current can be provided by a separate generator for each input branch; alternatively, it can be obtained by appropriately modifying the circuitry from which the input signals are drawn. This second solution may be preferable in order to reduce the number of required components, and at the same time it achieves a better parameter matching.

Considering the case of a VQ network, each input of the WTA module is produced by processing a high-dimensional vector (a typical image compression example is 64 components for each pattern).

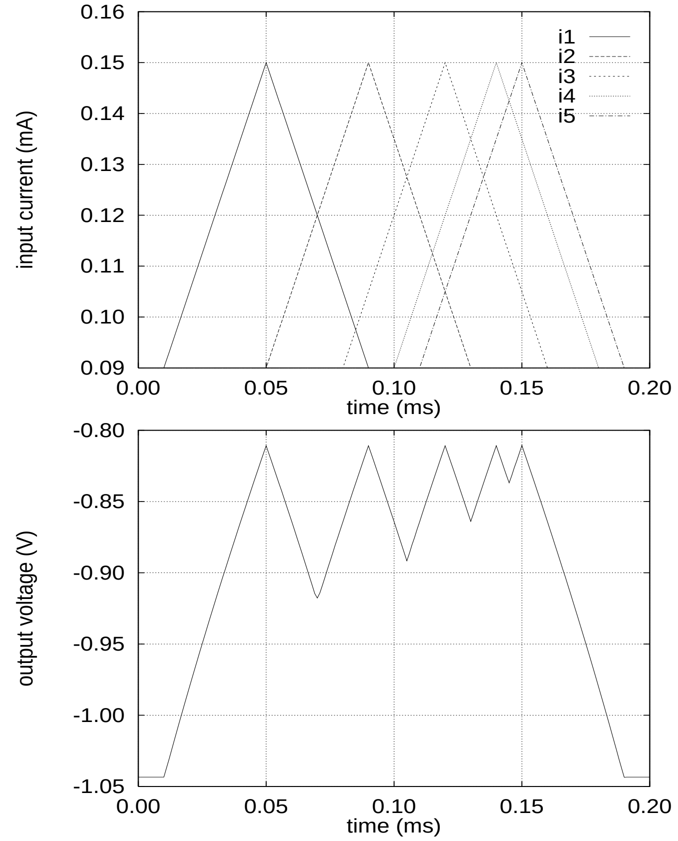
The modification of this circuit is such that each component provides a part of the bias current needed at its output to feed a single WTA input. Hence the expected variation of the parameters can be estimated to be about an order of magnitude lower if compared with that obtained by adding a single generator for each input branch. If each current I_i is obtained with a precision of $\pm\sigma$, we can expect that the total bias current and error can be estimated by:

$$I_{\rm tot} = \sum_{i=1}^{64} I_i \pm \frac{\sigma}{\sqrt{64}}$$

with a reduction in the expected error of 1/8. This result is obtained without the use of any additional component.

Many schemes exploit the polarization of the single input branches, that takes part to the positive feedback operation, contributing to saturate the winner's output and to shut off the other outputs. The required proportionality of the winner's output to its input implies that, in our case, this is not possible. However, the main drawback of this fact (an increased power consumption due to the fixed contribution of all bias currents) can be reduced with careful design optimizations.

Figure 2: Response of the circuit to time–varying input signals. Above: the input signals. Below: the output of the WTA circuit.



3 Numerical results

The circuit has been successfully simulated with the HSPICE program, level 13. The choice of the technological parameters was made on the basis of low costs, resource availability, and low expected percentage of defective chips in the realization phase.

The HSPICE geometrical and physical parameters have been evaluated with the aid of a layout simulation, to assess their limits of variation (*typical*, *slow*, and *fast*) in the physical realization. Therefore, the results feature a certain degree of reliability.

The polarization introduced allows the input impedance to drop from some Megaohms (in the standard circuit) to less than 150 kΩ. The circuital block producing the input current, in our case, featured a high output impedance (of the order of 100 MΩ) because of its cascode configuration. Hence its behavior is satisfactorily close to that of an ideal component.

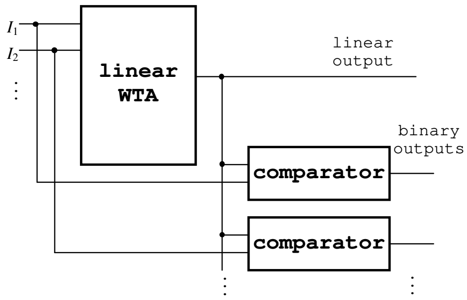
We can estimate the deviation from the ideal behavior, $V_{\rm out} = kI_{\rm in,winner}$, by the relative absolute error

$$\epsilon = \frac{|V_{\rm out} - kI_{\rm in,winner}|}{|kI_{\rm in,winner}|}$$

The simulations show that the linearity is good within a range of $0\ldots100\mu$A, if the bias current is about 100μA. In this range the largest error corresponds to the maximum output voltage, and is about $\epsilon = 5\%$. The width of the output voltage range is 200 mV.

The parameters have been optimized, by an approximate least–mean–squares procedure, in order to obtain the minimum mean error. However, when the WTA module is inserted into a neural VQ circuit, the

Figure 3: Sketch of the circuit needed to obtain from the proposed circuit a standard WTA function.



characteristics should be corrected to take into account the different importance of the error in the different operating intervals within the above range.

A sample result is presented in Fig. 2, where the linearity of the input–output characteristics is demonstrated.

The minimum discriminated difference can be estimated to be less than 1μA; hence the precision achieved is about 7 bit, which is very close to the standard design choices for image compression applications, our reference throughout this paper.

4 Concluding remarks

We have presented a WTA circuit featuring an output level proportional to the winning input. The circuit features a very good linearity and improved performances in terms of sensitivity. The design takes into account the need for a large fan–in, a design parameter which is almost invariably imposed by practical applications. The experimental results demonstrate the good properties of the circuit.

The circuit has been designed for use in competitive neural systems. As compared with the standard subthreshold design, the level of the input and output signals is larger, so that a multi–chip system realization is possible.

Often a standard binary WTA function is needed, e.g., during the training process to select the unit to be updated. With the presented circuit, a binary WTA is obtained by simple comparison between the WTA output and each input. A standard WTA would require instead additional circuitry to transfer the analog value of the winner to the subsequent processing stages, by means of analog switches. As compared with this scheme, the proposed solution requires only digital switches (comparators), therefore the realization is not critical. A block diagram is presented in Fig. 3.

References

[1] Teuvo Kohonen. *Self Organization and Associative Memories.* Springer, 3rd edition, 1989.

[2] Stephen Grossberg. Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, LII:213–257, 1973.

[3] N. Nasrabadi and R. King. Image coding using VQ: a review. *IEEE Trans. Commun.*, pages 957–971, 1988.

[4] J. Lazzaro, R. Ryckebush, M. A. Mahowald, and C. Mead. Winner-take-all networks of $O(n)$ complexity. In *NIPS*, pages 703–711, Los Altos, CA, 1989. Morgan Kaufmann.