A COMPACT PROGRAMMABLE ANALOG CLASSIFIER USING A VMM + WTA NETWORK

Shubha Ramakrishnan and Jennifer Hasler

Department of Electrical and Computer Engineering, Georgia Institute of Technology

ABSTRACT

We present the VMM+WTA structure as a general-purpose, low-power, compact, programmable classifier architecture and demonstrate its equivalence to a 2-layer perceptron. The classifier generates event outputs and is suitable for integration with event-driven systems. We present measured data from simple linear and non-linear classifier structures on a $0.35\mu m$ chip and demonstrate the implementation of an XOR function using a 1-layer VMM+WTA classifier.

Index Terms— classifiers, analog signal processing, programmable analog computing

1. EFFICIENT ANALOG SIGNAL PROCESSING

In embedded systems that receive sensory inputs, process and classify them to take decisions, it is essential to take a low-power approach for enabling such structures in robots and other mobile platforms. Classifiers are typically used in the information refinement stage and it is often essential that besides being low power, they also produce very few events. Events are generated when a certain class has been detected, triggering further circuitry dependent on this decision. In highly integrated systems, an increased number of events often leads to increased power consumption, which is required to transmit events over interconnects between blocks that have significant capacitances. We propose using a Winner-Take-All (WTA), which is observed in biological networks for reducing neuron firing rates, in our classifier. This reduces the rate of output events, resulting in an architecture with reduced power consumption.

In the past, significant effort in building hardware classifiers has been using Artificial Neural Networks (ANN). In the simplest ANN, we have inputs being multiplied by a weight vector, added together at the soma compartment, where a linear or nonlinear function is applied before we receive the output. ANN approaches include having continuous valued (e.g. $tanh$) functions that approximate the spike frequency versus current input (f-I) characteristic of neurons with an analog voltage, or spiking (integrate-and-fire neurons, rate-encoded neurons), feedforward or feedback stages.

In this paper, we consider an analog classifier consisting of a Vector-Matrix Multiply (VMM) terminated with a WTA, shown in Fig. 1, that is versatile and has more computing power than a 1-layer NN. The VMM block performs a multiply operation between a vector and a matrix of weights, resulting in a vector and forms a core component of many signal processing algorithms. The VMM+WTA, which we use as the base classifier, compares favorably against the 1-layer NN in terms of the number of components as well. We show a direct translation of a 1-layer NN to a VMM+WTA, where the WTA acts as a current comparator. In a different formulation, the WTA can perform an analog max function, selecting the largest (or smallest) of its inputs. With minor modifications, the WTA can be designed to allow multiple winners, local winners or to exhibit hysteresis [1–3], leading to classifiers that allow multiple winners with spatial responses which can be useful in image processing, or exhibit hysteresis which makes the classifier immune to noisy inputs. We see the VMM+WTA classifier being used in an analog speech recognizer as shown in Fig. 1. The speech input un-

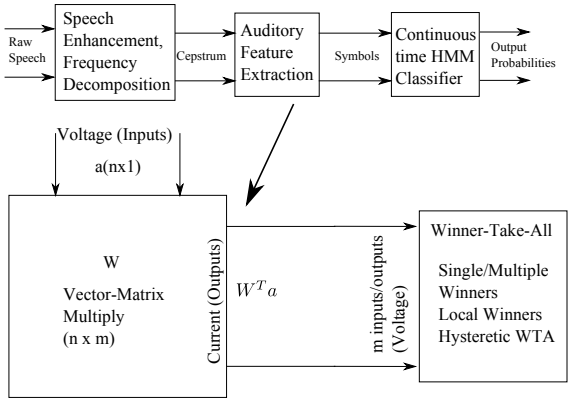


Fig. 1: **Application in Analog Speech Recognizer Chain:** The speech input undergoes frequency decomposition or enhancement resulting in sub-band signals. These signals undergo first-level information refinement in the feature detection stage, resulting in a sparse "symbols" or "event" representation. The following stage detects sequences of symbols/events to identify words or syllables, implementing a first-layer classifier. The feature detect stage maybe implemented as a VMM+WTA classifier, which takes voltage inputs. The VMM has current outputs and the WTA has voltage outputs.

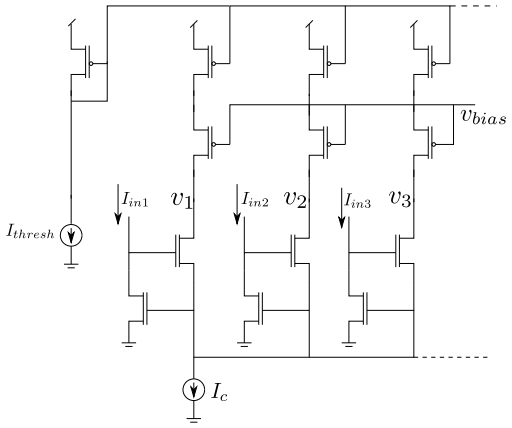


Fig. 2: **k-winner-take-all:**(a) The traditional WTA can be modified to a k-WTA with a current threshold at each output, realized using a cascoded pFET. The current flowing through the winning branch is constrained, allowing other inputs to the WTA to win. The voltage outputs from the WTA are inverted and a node wins when its output is below mid-rail. Choice of current threshold determines the number of winners

dergoes frequency-decomposition or signal-enhancement in the front-end, resulting in input features such as sub-band energies. These signal inputs are transformed into symbols or events with ANN, GMM or VMM+WTA in the first stage of information refinement. This can be followed by higher level refinement or by a sequencing block to detect syllables or words. A high-level system overview of the VMM+WTA circuit is shown in Fig. 1. The inputs to the classifier are voltages, while the outputs from the VMM are uni-directional currents. The WTA may produce voltage or current outputs.

2. HARDWARE SYSTEM IMPLEMENTATION

The hardware platform used for implementing the classifier is among the family of Field Programmable Analog Array (FPAA) chips, specifically geared towards building large VMMs. A detailed description of this chip and its features can be found in [4]. The WTA module is used for modeling competition in neural networks, specifically in representing the mechanism of attention [5]. The classic circuit implementation by Lazzaro et al [1] was based on continuous-valued elements, that utilized transistor device physics to build an efficient circuit. Several modifications to this circuit exist, that allow local winners, hysteresis behavior that stabilize the outputs, temporary winners that fatigue after a period of winning and allow other inputs to win and multiple winners [2, 6, 7], all of which may be implemented on the FPAA using the components available. Often, we require classifiers that generate not just one output, but multiple outputs. In pattern classification, we can expect the classifier to indicate

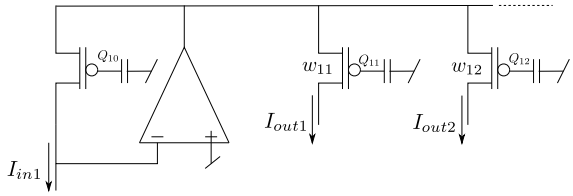


Fig. 3: 1x2 **VMM characterization:**(a) Schematic of a 1x2 VMM with current inputs. The OTA with base floating-gate is a logarithmic trans-impedance amplifier and generates a source voltage that is applied to other devices with programmed weights.

that a certain pattern matches two categories instead of just one. The classic WTA circuit does not preclude multiple winners and this can be achieved by modifying the circuit as shown in Fig. 2. In this paper, we utilize the classic WTA and the multiple-winner-WTA circuit for constructing the classifiers. The k-WTA produces inverted voltage outputs that are taken at the drain of the thresholding pFET. Compared to the k-WTA circuit in [6], this implementation does not require any additional power/circuitry. VMMs can be implemented in a power-efficient and compact manner using floating gates. The multiplication weights are stored as charge on the floating node and can be precisely programmed and controlled. The weight can be expressed as

$$w = e^{\kappa Q/C_T U_T} \tag{1}$$

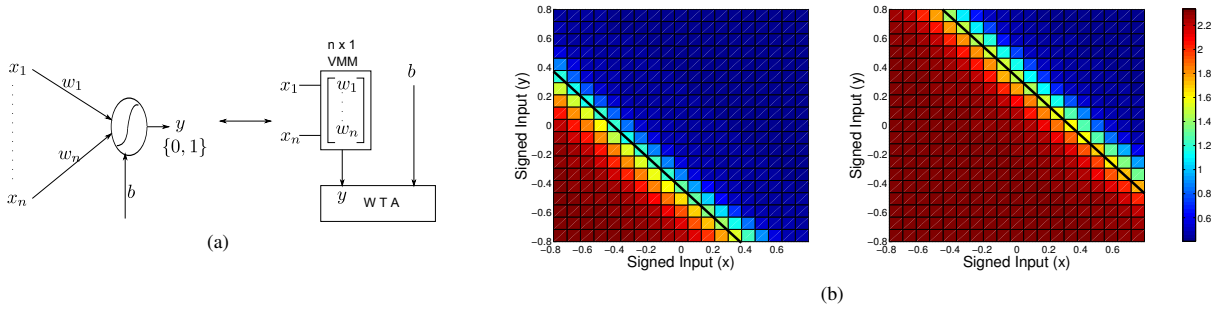
where Q is the charge programmed on the floating-gate node and C_T is the total effective capacitance seen at the floating node. A single floating gate stores the weight as well as performs a multiply function. Examples of the different VMM topologies that we can implement are discussed in [8]. The schematic of a 1x2 VMM which achieves single-quadrant multiplication is shown in Fig. 3. To achieve four-quadrant multiplication, we require a VMM that takes differential inputs and implements signed weights. These structures are discussed in [8].

3. CAPABILITY OF VMM+WTA CLASSIFIERS

We now integrate the VMM and WTA circuits to build simple classifier structures. In this section, we describe measured results from system compilations of linear, multi-class and non-linear classification problems.

3.1. Linear Classifiers

We start by considering a perceptron, which is a simple linear classifier with a binary output that can be implemented with a 1-layer neural network. A linearly separable set of inputs can be classified using a perceptron trained to weights w_i and



(a) (b)

Fig. 4: **Linear Classifiers:** A simple perceptron or a one-layer feed-forward network can be implemented using a VMM+WTA structure. (a) The input multiplication can be implemented using VMMs. The bias b is the second input to the WTA, implemented as a fixed current source. **Measured results:**(b) A VMM+WTA classifier trained to have a decision boundary of $y + x \geq b$, for different bias values b. The black solid line represents the theoretical decision boundary.

bias b having the equation

$$y = \begin{cases} 1 & \text{if } \sum_i w_i x_i - b \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

A VMM+WTA classifier can be trained as a generalized single-layer perceptron by using a fixed current source as an additional bias input to the WTA, shown in Fig. 4a. The WTA functions as a current comparator and the detects the larger of the inputs. When $\sum_i w_i x_i > b$, the first input wins. By using a 1-WTA circuit implemented with the current threshold at the WTA output, we obtain inverted voltage outputs. Hence, the first output is low when $\sum_i w_i x_i > b$ and high otherwise.

We measured results from two different linear classifier boundaries programmed on the VMM+WTA circuit, for multiple bias values. For a linear decision boundary, we train a perceptron using MATLAB's Neural Network Toolbox and apply the weight and bias values directly to the VMM+WTA classifier. We restricted ourselves to a 2-input case for ease of visualization. The structure in Fig. 4a only supports positive values for the bias. Since our implementation required signed weights and bias values, we chose a topology with fully-differential inputs. The classifier was tested over all inputs from the set $\{(x, y) : |x| \leq 0.8, |y| \leq 0.8\}$. We plot the inverted WTA voltage output in Figs. 4b. The output makes a sharp transition at the desired decision boundary, which is marked by the solid line in the plots. We were able to directly apply the weights obtained from the training algorithm and target them to the hardware.

3.2. Multi-class Classifiers

As the name suggests, multi-class classifiers have several outputs, and classify data into multiple classes. The competitive behavior modeled in the VMM+WTA circuit allows building of such classifiers with multiple outputs that can detect regions of interest. We demonstrate the capability of the VMM+WTA circuit to build a region detector. We train a 2-input, 3-output classifier to detect regions of inputs defined as shown in Fig. 5. Again, for simplicity of visualization, we chose only 2 differential inputs. We constructed a classifier with 3 outputs and the region boundaries specified in Fig. 5(a). From this theoretical construction, we obtained the weights for the VMM using the pseudo-inverse method. We generate random inputs in MATLAB and multiply them by the weight matrix obtained. We then do a max function on the transformed inputs to generate the theoretical classifier output in Fig. 5. Since the theoretical weights were signed, we constructed a fully-differential implementation and targeted the weights to the VMM circuit. We then applied 1000 inputs randomly from the set $\{(x, y) : |x| \leq 0.8, |y| \leq 0.8\}$. Since the WTA voltage outputs are inverted, we found the winning output by finding WTA voltages below inverter threshold(mid-rail) and recording its position. In Fig. 5(b), we denote the winning position for each of the random inputs by a different colored dot. Our 3-output classifier was programmed with weights obtained directly from MATLAB. It matches the desired classifier response quite well.

3.3. Non-linear classifiers

Non-linear classification boundaries required in most real-world problems are usually very computationally intensive. Single-layer neural networks can only implement classifiers for linearly separable data, but a 2-layer NN can approximate any function. A 2-layer NN has an input layer, hidden layer and an output layer. An analog VLSI implementation would require 2 VMMs for the synaptic computation and 2 layers of threshold blocks for the hidden and the output layers. This considerably increases the complexity and power consumption of the circuitry. In [9], Maass showed that any boolean function with analog or digital inputs and one binary output can be approximated with a VMM + k-winner-take-all classifier. He showed that the weights for the VMM+WTA classifier are a linear combination of weights of the 2-layer

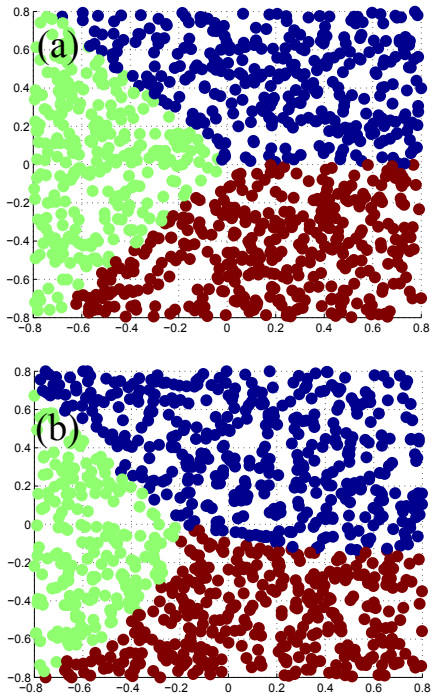


Fig. 5: **Multi-dimensional classifiers:** (a) A two-input three-output VMM+WTA classifier constructed to have the theoretical decision boundaries shown. Each color represents a different winner. (b) Measured results from the VMM+WTA classifier compiled.

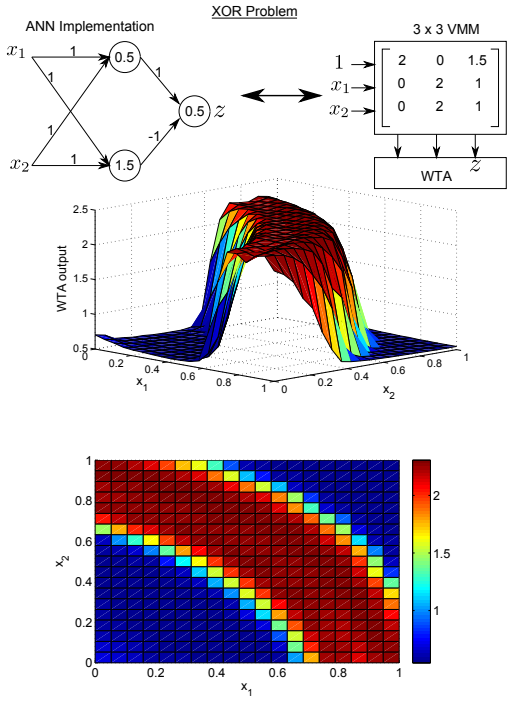


Fig. 6: **Nonlinear Classifiers:** The VMM+WTA structure is powerful enough to implement any boolean function with one digital output. A solution for the XOR problem using a two-layer neural network can be translated to a VMM+WTA implementation. Measured results from an XOR implementation using the VMM+WTA structure.

perceptron, and further, they are all positive, requiring only single-ended inputs in our implementation. This result provides additional support to the computational power of the VMM+WTA classifier, by halving the computing resources required.

One of the most computationally challenging problems for neural networks is the XOR problem, which does not involve a linear decision boundary. We use the algorithm provided in [9] to compute weights for our VMM+k-winner-take-all structure to implement a non-linear classification boundary for an XOR circuit. One possible implementation of the XOR gate with a 2-layer neural network and its equivalent VMM+WTA implementation is shown in Fig. 6. The VMM+WTA XOR circuit requires only a single-winner WTA. The position of the WTA output computing the XOR function is marked y in Fig. 6. We tested the XOR circuit by generating inputs from the set $\{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$ and recording the voltage at the third output. The results are plotted in Fig. 6. The VMM weights are biased at 10nA, resulting in 95nA drawn in the VMM when both inputs are active. The WTA is biased at 100nA, resulting in 0.47μW drawn at 2.4V, when all inputs are active.

4. CONCLUSIONS

We have presented results from a powerful re-programmable classifier that can implement linear as well as nonlinear decision boundaries. The classifier architecture combines two power efficient circuits to provide an ASP alternative to traditional approaches. The system is extremely compact, allowing scaling to large number of inputs. One of the disadvantages of ASP is fixed functionality. The reconfigurability of the chip allows programmable weights which enables off-line training, modifying the size and changing the topology of the WTA to generate different behavior. As an extension to this work, we can implement local WTAs and hysteretic WTA for certain applications. We have seen that the VMM+WTA classifier is roughly equal to a 1-layer NN in circuit complexity, but has computing power equivalent to a 2-layer NN. We demonstrate this by implementing classic small-scale nonlinear classification problems.

5. REFERENCES

[1] J. Lazzaro, "Winner-take-all networks of o (n) complexity," Tech. Rep., DTIC Document, 1988.

[2] G. Indiveri, "A current-mode hysteretic winner-take-all network, with excitatory and inhibitory coupling," *Analog Integrated Circuits and Signal Processing*, vol. 28, no. 3, pp. 279–291, 2001.

[3] T.G. Morris, T.K. Horiuchi, and S.P. DeWeerth, "Object-based selection within an analog VLSI visual attention system," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 45, no. 12, pp. 1564–1572, 1998.

[4] C.R. Schlottmann, S. Shapero, S. Nease, and P. Hasler, "A digitally enhanced dynamically reconfigurable analog platform for low-power signal processing," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 9, pp. 2174 –2184, sept. 2012.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, 1998.

[6] K. Urahama and T. Nagao, "K-winners-take-all circuit with o (n) complexity," *Neural Networks, IEEE Transactions on*, vol. 6, no. 3, pp. 776–778, 1995.

[7] W.F. Kruger, P. Hasler, B.A. Minch, and C. Koch, "An adaptive WTA using floating gate technology," *Advances in Neural Information Processing Systems*, pp. 720–726, 1997.

[8] C.R. Schlottmann and P.E. Hasler, "A highly dense, low power, programmable analog vector-matrix multiplier: The FPAA implementation," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 1, no. 3, pp. 403 –411, sept. 2011.

[9] W. Maass, "On the computational power of winner-take-all," *Neural Computation*, vol. 12, no. 11, pp. 2519–2535, 2000.