32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

A Biologically Plausible Model for Same/Different Discrimination

Hernán G. Rey, Diego Gutnisky, and B. Silvano Zanutto

Abstract— **Abstract rules can be learned by several species (not only humans). We propose a biologically plausible model for same/different discrimination, that can point towards the neural basis of abstract concept learning. By including a neural adaptation mechanism to a discriminator model formerly introduced in the literature, selective clusters of neurons fire depending on whether or not the stimuli compared are the same or not. These selective neurons are consistent with experimental findings in the literature. Moreover, reward and attention can modulate the relative strength of each attribute/feature of the stimulus, so more complex abstract discriminations can be achieved using the proposed model as a building block. As a formal model, it can be easily incorporated into several applications in robotics and intelligent machines.**

I. Introduction

Understanding how animals control their behavior can serve as an inspiration to build intelligent robots. Biologically plausible models have been successfully applied in the robotics area [1]. On the other hand, theoretical tools can be used in neuroscience to unravel the complexity of the brain. Here, we propose a biologically plausible model that can be used as a building block for learning abstract rules or concepts. Throughout this work, the term abstract rule will be used to define a rule that is not directly tied to the surface features or sensory instantiation of the stimuli. Particularly, the learning of structural relations among items is emphasized while the acquisition of information pertaining to specific features of the stimulus elements is deemphasized [2]. These rules are in contrast with stimulus-specific rules, that have been extensively studied in the literature [3][4].

One of the simplest abstract paradigms is same/different (SD) discrimination, where the relation between two stimuli must be judged as "same" or "different". Abstract SD concept learning transcends the stimuli used during training, and is distinct from "natural" concepts which are categories unified by some specific stimulus attribute or attributes. Interestingly, for several years it was believed that abstract concepts appear with language (or at least language training) as a necessary condition [5]. However, for the last thirty years the experimental psychology has provided evidence showing that many different animal species can solve this task. In the case of humans, this task is exceedingly straightforward. Actually, when making comparisons between objects or events in the world, the SD distinction is many times our first approach, conveying the most useful information. Yet, this is most probably not the case in other species as they present different learning predisposition (e.g., to stimulus-specific learning). Experimental evidence indicates differences across several species trained on an SD discrimination paradigm. In some cases (e.g., humans and chimpanzees) they can transfer the rule to novel stimuli after being trained with sets of a few stimuli, whereas in other cases (e.g., pigeons) a set size of over 200 stimuli is required. However, these differences are quantitative but not qualitative.

We present here a model that can point towards the neural basis of abstract concept learning. The focus is on the mechanism to solve the SD discrimination. The model uses simple mechanisms that can be found in several species like humans, monkeys and pigeons. Moreover, as it is a formal model, it can also be applied in robotics.

This work was support in part by grants from ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica), PICT 02485; UBACYT (Universidad de Buenos Aires–Ciencia y Técnica), 027; and CONICET (Consejo Nacional de Investigaciones Cientí'ficas y Técnicas), PIP 112-200801-02851.

H.G. Rey and B.S. Zanutto are with the Institute of Biomedical Engineering (University of Buenos Aires) and CONICET, Buenos Aires, Argentina hrey,silvano@fi.uba.ar

D. Gutnisky is with Janelia Farm - Howard Hughes Medical Institute, Ashburn, USA gutniskyd@janelia.hhmi.org

II. Discriminator Model

In [6], the authors introduced a model with the ability to discriminate stimuli presented sequentially. We will use this model as the starting point to build our SD discriminator model. Therefore, we perform first an analysis on the important aspects of the model presented in [6].

On a first approximation, a linear firing rate model is presented, which in turn can be implemented using discrete integrators. The model is formed by two subsystems (S_+ y S_-) that follow the same dynamics (see Fig. 1). The only difference between them is that the input ($I(t)$) of S_+ is coded in an increasing way, whereas the one of S_- is coded in a decreasing way. This means that the stronger (weaker) the input of S_+ (S_-), the larger the associated value of $I(t)$. The dynamics of each subsystem are:

$$\tau \frac{dr_C}{dt} = -r_C - w_{MC}\, r_M + I(t), \tag{1}$$

$$\tau \frac{dr_M}{dt} = w_{CM}\, r_C, \tag{2}$$

where r_C and r_M stand for the firing rate of the cluster of neurons C and M, respectively, and τ is the time constant.

This system has initial conditions $(r_C, r_M) = (0, 0)$, representing the basal firing level at the beginning of each trial. At $t = 0$, a stimulus appear, so $I(t) = I_1$ for a certain time T_1 (stimulus 1 duration). The system formed by (1) and (2), evolves towards the fixed point (FP):

$$\text{FP}: \ (r_C^*, r_M^*) = \left(0, \frac{I_1}{w_{MC}}\right) \tag{3}$$

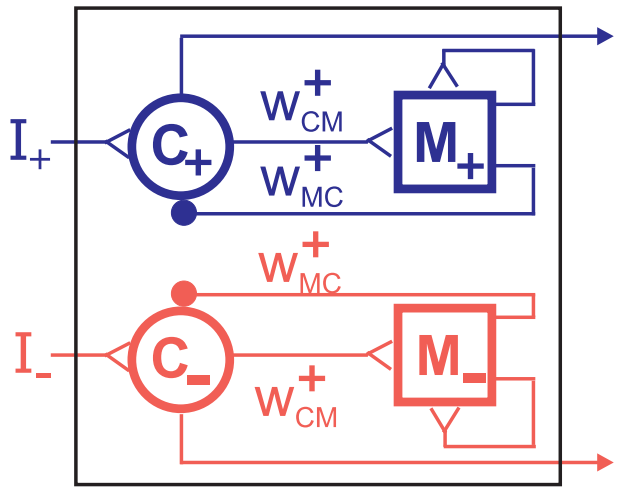

Fig. 1. Scheme of the discriminator model proposed in [6]. The neurons in the cluster C remain silent unless they receive an external input I. The neurons in the cluster M integrate the activity of the neurons in C, while being able to sustain their activity during the *delay*, when $I = 0$. The connection from M to C is inhibitory, with the remaining ones being excitatory. The two parallel subsystems differ only in the way their inputs are codified.

To analyze the stability of this FP, we analyze the eigenvalues of the system's jacobian, resulting in:

$$\lambda_{1,2} = \frac{-1 \pm \sqrt{1 - 4\, w_{MC}\, w_{CM}}}{2}.$$

Since this is a firing rate model, if r_C or r_M would become negative according to their dynamics, they are set to 0 (strictly speaking, it is only necessary to set $r_C = 0$). Also, to avoid an oscillatory behavior $\lambda_{1,2}$ must be real, so the first condition on the parameters is:

$$w_{MC}\, w_{CM} < 1/4 \tag{4}$$

Although the FP will then be a stable attractor, the input must be "on" for a period of time long enough so that the system can reach the FP. From (1) and (2), it can be seen that r_C and r_M will increase until the condition $I_1 = r_C + w_{MC}\, r_M$ is satisfied. Then, r_C begins to decrease while r_M is still increasing. At some point t_{crit}, r_C will be 0 and the system reaches the equilibrium. It can be shown that the system will reach the steady state if

$$T_1 \gg \frac{2\,\tau}{1 - \sqrt{1 - 4\, w_{MC}\, w_{CM}}}. \tag{5}$$

From (4) and (5), the conditions on the parameters are:

$$\frac{1}{4}\left[1 - \left(1 - \frac{2\,\tau}{T_1}\right)\right] \ll w_{MC} w_{CM} < \frac{1}{4}. \tag{6}$$

Therefore, by choosing $w_{MC} w_{CM} < 1/4$ and $T_1 \gg \tau$ we can guarantee that the system reaches the equilibrium.

After T_1, $I(t) = 0$ during the *delay* period. In this case the C cluster is strongly inhibited but since r_C must be nonnegative, it remains $r_C = 0$. Thus, r_M remains constant, so the value reached at T_1 is kept "in memory". If $w_{MC} = 1$, this value is I_1. Then, when the stimulus 2 appears with an input I_2, different responses will appear depending on the relation between the stimuli.

If $I_2 > I_1$, the cluster C_+ will fire after the appearance of stimulus 2, while cluster C_- will remain at rest ($r_C = 0$). If $I_2 < I_1$, everything is reversed. However, if $I_2 = I_1$, both clusters C_+ and C_- will stay at rest.

III. New Model for Abstract SD Discrimination

It is assumed that each stimulus is presented as the input to the model coded as a set of N attributes/features. Each attribute is associated to a neuronal cluster, whose activity is represented by the variable I. By modifying the model discussed in the previous section, differential activity will be generated depending on whether or not the sequentially presented attributes are the same. The source of the modification was inspired by experimental evidence from monkeys. In [7], the activity of neurons from inferotemporal cortex is recorded while visual stimuli are presented sequentially. One of the observed phenomena is known as *matching supression* (MS). It is related to the decrease in the neuronal activity after the second appearance of a matching stimulus, regardless of whether or not the stimulus is task-relevant.

The MS effect can be explained by the mechanism of neural adaptation [8]. We propose to include this mechanism in the model studied in II and analyze its behavior. To do so, the variable a in incorporated to the system, whose new form is:

$$\epsilon \dot{r}_C = -r_C - a w_{MC} r_M + I. \tag{7}$$

$$\epsilon \dot{r}_M = w_{CM} r_C. \tag{8}$$

$$\dot{a} = -a + f(r_M). \tag{9}$$

where $\epsilon = \tau/\tau_a \ll 1$, i.e., the adaptation dynamics are much slower than the neuronal dynamics for integration. Initially, $a = 1$. The function $f(r_M) = 1/(1 + \beta r_M)$ introduce a nonlinearity to the system. At low (r_M) frequencies, a remains close to 1 so the adaptation effect is not relevant. However, as the firing rate is increased, a decreases, and with it so does the synaptic efficiency from M to C.

As in the previous section, we start the analysis presenting a certain attribute of the first stimulus with intensity I_1. The new FP is:

$$\text{PF}: \quad (r_C^*, r_M^*, a^*) = \left(0, \frac{I_1/w_{MC}}{a^*}, 1 - \frac{I_1 \beta}{w_{MC}}\right) \tag{10}$$

If a minimum (biologically plausible) value is set for a, a condition on β can be obtained for the maximum possible value of the input, leading to:

$$\beta = \frac{(1 - a_{\min}) w_{MC}}{I_1^{\max}} \tag{11}$$

Since $\epsilon \ll 1$, a condition on the FP (10) to be an attractor is $a^* w_{MC} w_{CM} < 1/4$. This is not strange considering that $a^* w_{MC}$ is the new synaptic efficiency from M to C. Taking into account that a^* depends on the attribute intensity, the more conservative condition (4) can be used.

Regarding the lower bound, an analysis of the time scales is required. As $\epsilon \ll 1$, r_C and r_M evolve much faster than

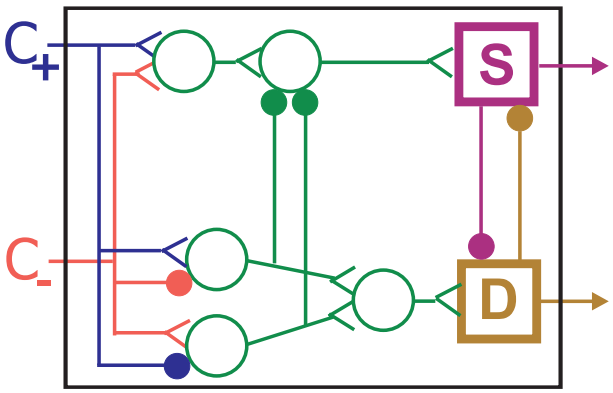

Fig. 2. The signals at the output of the neurons in clusters C_+ and C_- are further processed at the next stage. As a result selective clusters arise, that will be activated when the attributes are the same (cluster S) or way they are different (cluster D). The dots represent inhibitory connections, whereas the remaining ones are excitatory.

a. Then, if condition (6) is scaled by $a_{\min}$, r_C and r_M will approach 0 and $\frac{I_1/w_{MC}}{a}$, respectively. If the input remains "on" for a longer time, the system would eventually reach the FP. If this is the case, once the stimulus disappears at the beginning of the *delay*, no change will be done to the system so it behaves qualitatively in the same way as the previously studied linear system. However, assume that the stimulus duration T_1 is long enough so that the fast variables evolve quickly but the FP is not reached, i.e., $a^* < a < 1$, $r_C \approx 0$ y $r_M \approx \frac{I_1/w_{MC}}{a}$. As the system enters into the *delay* period, a will keep on moving towards a^*, but since $r_C = 0$, r_M remains unchanged. When the second stimulus switches "on", the cluster C will be less inhibited compared to the onset of the *delay*. Therefore, even if the second attribute has the same intensity as the first one, the cluster C will show a nonzero response.

To sum up, the modified system operates in the following way: if $I_2 - I_1 > \delta$, cluster C_+, but not C_-, will be activated; if $I_2 - I_1 < -\delta$, cluster C_-, but not C_+, will be activated; if $|I_2 - I_1| < \delta$, both clusters C_+ and C_- will be activated (although each one of them with less activity than in the previous cases). Hence, the proposed mechanism provides a differential output depending on whether the stimuli are the same (or very similar) or different.

A. SD Discrimination at the Attribute Level

In a second layer (see Fig. 2), the outputs of the C_+ and C_- clusters associated with each attribute are processed, reaching finally the clusters S (whose activity is driven by the least active of the C clusters) and D (whose activity is driven by the largest difference in the activity of the C clusters). These clusters compete by mutual inhibition in a kind of "winner takes all" mechanism. When the attributes are the same, both C_+ and C_- fire in a similar way, so as the D cluster will have a weak input, the S cluster will fire strongly. On the other hand, when the attributes are different, cluster D fires selectively (and stronger, the larger the difference between the attributes).

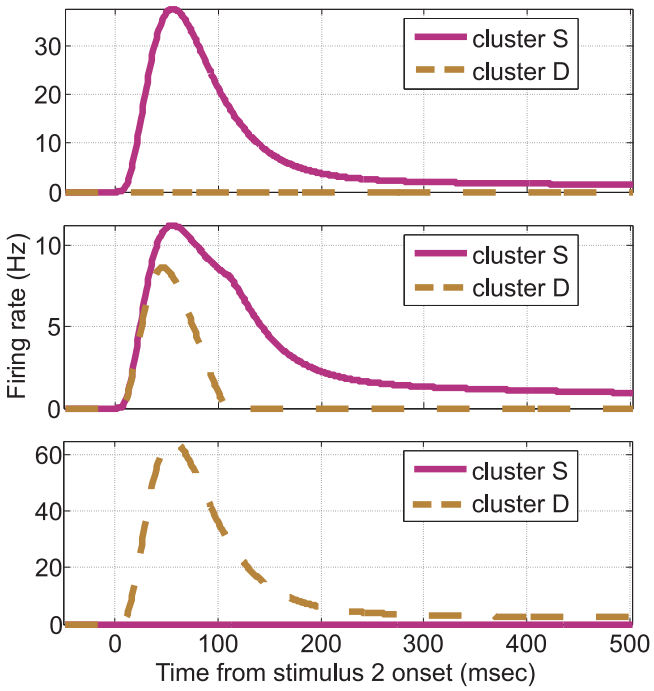

Fig. 3. Firing rate of clusters S and D at the attribute level. In all the scenarios $I_1 = 20$; however the value of I_2 is set to 20 (top), 21 (middle) or 25 (bottom).

B. Simulation Results (at the Attribute Level)

Now we will analyze the behavior of the model at the attribute level, i.e., the one resulting from Figs. 1 and 2. The parameters were set according to: $w_{MC} = 1$, $w_{CM} = 0.24$, $\tau = 10$ msec., $T_1 = T_2 = 500$ msec., $delay = 1$ sec., $\tau_a = 500$ msec., $a_{\min} = 0.5$, $I_{\max} = 40$.

In Fig. 3 we show the firing rate of clusters C and D under different scenarios. The intensity of the first stimulus is fixed as we changed the intensity of the second one (if this is done the opposite way, the results are exactly the same). In the top part, the same attributes are presented ($I_1 = I_2$). Both synaptic connections (to C_+ and C_-) are adapted, but since this is a slow process, the FP is not reached. Once in the *delay*, the inhibitory synapsis continue adapting, decreasing the amount of inhibition over the C clusters. Therefore, when the second attribute appears, the cluster S fires strongly, indicating that the attributes are the same.

Interestingly, if the second attribute is slightly increased, the adaptation process can still generate certain activity on the cluster C_-. The activity on the cluster C_+ is higher, but the difference in activity is quite small. Then, after the onset of the second stimulus, both clusters S and D show some initial activation. However, the competition between the clusters lead to an increase in the response of the S cluster over the one of the D cluster (Fig. 3 middle), and the attributes are judged as the same (or *similar*). Therefore, a certain degree of dissimilarity is required to judge two attributes as different. It should be noticed that level of activity of the S cluster is lower than when both attributes were the same. This will be relevant later when the signals

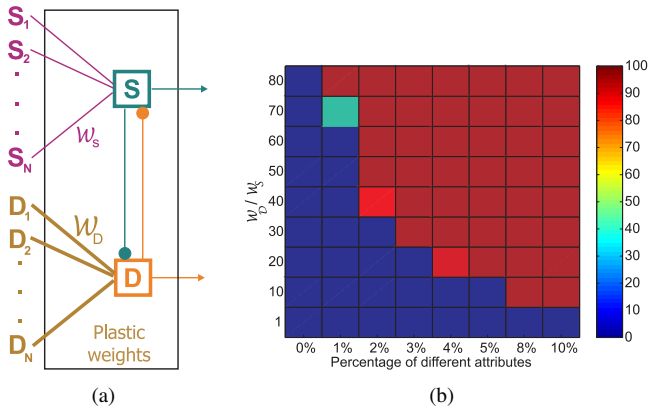

Fig. 4. (a) SD discrimination at the stimulus level. The weight associated to each attribute can be modulated by attentional processes and reward information throughout the task. The number of attributes to code the stimuli were set to $N = 200$, and in the cases where an attribute was different for each stimulus a difference $\Delta I = 3$ was used, representing approximately a 15% difference in intensity. (b) Percentage of "different" choices. For a fixed proportion of different attributes between the stimuli being compared, the decision can change depending on the the relative strength of the weights coming from clusters S and D at the attribute level.

from each attribute are integrated to end up with a decision for the SD discrimination at the stimulus level. If I_2 is further increased, a large different between the attributes will be translated into a large activity of the D cluster (Fig. 3 bottom), so the attributes are correctly judged.

C. SD Discrimination at the Stimulus Level

The last stage of the model is where the information from all the attributes is integrated. As shown in Fig. 4(a), selective groups of neurons will fire depending on whether or not the presented stimuli are the same. This is consistent with the experimental evidence from the prefrontal cortex of monkeys performing an SD discrimination paradigm [9].

Most importantly, the relative weight of each attribute can be modulated, for example, by reward learning. This might in turn affect the decision during the discrimination according to the degree of similarity between the stimuli. In Fig. 4(b), for a fixed percentage of different attributes between the stimuli, we studied the percentage of "different" responses (activation of the D cluster at the stimulus level). When an attribute was different, $\Delta I = 3$ was used, which represents a 15% difference in intensity. If ΔI is increased, Fig. 4(b) will move towards the bottom left corner.

When evaluating the first order identity relation between the stimuli all the attributes are equally important. However, one might be interested in using as stimuli different shapes of different colors but just one of the dimensions should be used as the controlling cue for the discrimination. In this case, reward learning mechanisms would help to discover the controlling cue which will be the one with the highest relative weight.

IV. CONCLUSIONS

We introduced a biologically plausible model for same/different discrimination, that can point towards the neural basis of abstract concept learning. It should be emphasized that the proposed model uses biological mechanisms that have already been described in the literature, although without being linked to the learning of abstracts rules. Moreover, these mechanisms are simple enough so they might be found in several animal species (humans, monkeys, pigeons, etc.), even tough the way they are "implemented" in the brain might be different across species. In the model, selective clusters of neurons ended up firing depending on whether or not the stimuli compared were the same or not. This was achieved by incorporating the key neural adaptation mechanism to a discriminator model formerly introduced in the literature. The resulting selective neurons are actually consistent with experimental findings in the literature.

Furthermore, reward mechanisms can modulate the relative importance of different controlling cues. With that in mind, the spectrum of potential uses for this model can be expanded to more complex relational learning (e.g., relations among relations), learning sequences of stimuli/actions, etc. For example, in the area of robotics, the model could help into learning sequence of actions in an abstract way, so that they can be easily generalized when using new learnt actions. It can also be used to learn instructions of the form $object + place + action$ in an abstract way. This would largely increase the repertoire of an adaptive agent as new objects/places/actions are learnt. In the end, the resulting increase in the "cognitive abilities" of a robot will make it more worthy of being labeled as an intelligent machine.

REFERENCES

[1] D.A. Gutnisky, R. Zelmann, and B.S. Zanutto, "Multiagent team formation performed by operant learning: an animat approach", in *Proc. IEEE Int. Joint Conf. Neural Networks*, Vancouver, Canada, 2006, pp. 2944–2950.

[2] G.F. Marcus, S. Vijayan, S. Bandi Rao, and P.M. Vishton, "Rule learning by seven-month-old infants," *Science*, vol. 283, no. 5398, pp. 77–80, Jan. 1999.

[3] S.E. Lew, H.G. Rey, D.A. Gutnisky, and B.S. Zanutto, "Differences in prefrontal and motor structures learning dynamics depend on task complexity: A neural network model," *Neurocomputing*, vol. 71, no. 13-15, pp. 2782–2793, Aug. 2008.

[4] H.G. Rey, S.E. Lew, and B.S. Zanutto, "Dopamine and norepinephrine modulation of cortical and subcortical dynamics during visuomotor learning," in *Monoaminergic Modulation of Cortical Excitability*, K.Y. Tseng and M. Atzori, Eds., Springer, 2007, pp. 247–260.

[5] D. Premack, "On the abstractness of human concepts: Why it would be diffcult to talk to a pigeon," in *Cognitive processes in animal behavior*, Hulse S.H., Fowler H., and Honig W.K., Eds., Hillsdale, NJ: Erlbaum, 1978, pp. 423–451.

[6] P. Miller and X.J. Wang, "Inhibitory control by an integral feedback signal signal in prefrontal cortex: A model of discrimination between sequential stimuli," *Proc. Natl. Acad. Sci.*, vol. 103, no. 1, pp. 201–206, Jan. 2006.

[7] E.K. Miller and R. Desimone, "Parallel neuronal mechanisms for short-term memory," *Science*, vol. 263, no. 5146, pp. 520–522, Jan. 1994.

[8] M.I. Chelaru and V. Dragoi, "Asymmetric synaptic depression in cortical networks," *Cerebral Cortex*, vol. 18, no. 4, pp. 771–788, Apr. 2008.

[9] R. Muhammad, J.D. Wallis, and E.K. Miller, "A comparison of abstract rules in the prefrontal cortex, premotor cortex, inferior temporal cortex, and striatum," *J. Cogn. Neurosci.*, vol. 18, no. 6, pp. 974-989, Jun. 2006.