# A New Reinforcement-Learning Model for Categorization Tasks

Pieter R. Roelfsema[a], Arjen van Ooyen[b]

[a]Netherlands Ophthalmic Research Institute, Meibergdreef 47, 1105 BA Amsterdam, The Netherlands; [b]Netherlands Institute for Brain Research, Meibergdreef 33, 1105 AZ Amsterdam, The Netherlands; P.Roelfsema@ioi.knaw.nl A.van.Ooyen@nih.knaw.nl

**Reinforcement learning is a biologically plausible way to train neural networks. However, previous implementations do not specify, as does the error-backpropagation algorithm, how units that are intermediate between input and output units should optimize their tuning. Here we demonstrate that this so-called credit-assignment problem can be solved by including an attentional feedback signal from output units to gate plasticity at earlier processing levels. This yields a reinforcement-learning model that is as powerful as error-backpropagation and yet biologically realistic. Moreover, the model can account for recent neurophysiological data showing that categorization training changes the tuning of sensory neurons.**

The behaviour of an animal that performs a categorization task is simulated by a three-layered neural network that has an output unit for each category. On each trial, a stimulus is presented to the input layer, activity is propagated to the hidden layer, and from there to the output layer, where one of the units becomes activated. A reward is delivered if the correct output unit is activated. No reward is delivered in case of an error, but the correct output is not revealed to the network. After each trial, synaptic connections are modified, and a new trial is initiated.

Our model is defined by four key assumptions. First, units in the output layer engage in a competition where one unit wins. The competition is governed by the stochastic softmax rule, which is biased to choose the output unit that receives the strongest excitation from lower layers. Second, the winning unit in the output layer gives "attentional" feedback to the hidden layer (see Fig. 1). The strength of the feedback connections is proportional to the strength of the corresponding feedforward connections. As a consequence, hidden units that provide the most excitation to the winning output unit also receive the strongest feedback. Feedback thereby assigns credit to hidden units that are responsible for the choice of action. This is in accordance with neurophysiological data showing that feedback enhances the response of sensory neurons to the object that is selected for action [1, 2]. Third, a signal $\delta$ that depends on whether reward is obtained after the trial is broadcasted to all units and determines synaptic plasticity. On unrewarded trials, $\delta$ takes a constant negative value. On rewarded trials, however, it equals the difference between the amount of reward that was expected and the amount that was actually obtained, in line with existing models of reinforcement learning. The signal $\delta$ can be computed by using the a priory probability that a particular action is chosen, in line with the $A_{R-P}$ learning rule [3]. Signals related to reward expectancy could be carried to the cortex by the release of neuromodulators such as dopamine [4, 5]. Fourth, $\delta$ determines synaptic plasticity, but the attentional feedback gates the weight changes, so that synapses onto hidden

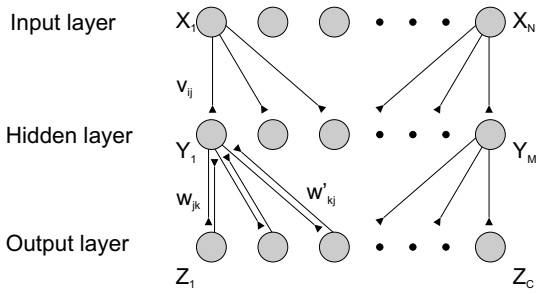


Figure 1: The three-layered network of the *Bapte* algorithm. Feedback connections $w'_{kj}$ from the output to the hidden layer gate the plasticity of the connections $v_{ij}$ from the input to the hidden layer.

units that are responsible for the outcome of the trial are most plastic. We show that a formal implementation of these four assumptions can yield changes in synaptic weight that are the same as those obtained with the error-backpropagation rule. The proposed learning scheme is therefore called *Bapte* ("backpropagation by trial and error").

We investigated if *Bapte* can account for changes in tuning of neurons in sensory areas that are caused by categorization training. In a recent study, Sigala and Logothetis [6] trained monkeys to classify face stimuli (which differed in 4 features) into two categories, and tested the influence of categorization training on the tuning of neurons in the inferotemporal cortex (IT). They showed that the neurons became better tuned to the diagnostic features of the face stimuli (the features that allowed separation of the stimuli along a linear category boundary) than to the non-diagnostic features. To test whether *Bapte* can account for these changes in tuning, we tested a model with 4 input units (one for each feature), 4 hidden units, and 2 output units (one for each category). After training, we found that the hidden units were significantly more selective for diagnostic than for non-diagnostic features, just as was observed in the inferotemporal cortex of the monkeys.

Categorization training also influences the parametric representation of a single feature. Schoups et al. [7] investigated the effect of orientation discrimination training on the orientation-tuning of neurons in the primary visual cortex of monkeys. The animals were trained to discriminate between small (1-2 deg.) differences in the orientation of an oblique grating. It was found that training induced a steepening of tuning curves of neurons with a preferred orientation that differed by 12-20 deg from the trained orientation. To simulate this experiment, we used 20 input units that had gaussian orientation tuning curves and equally spaced preferred orientations. To start with a sufficient variety of tuning curves in the hidden layer, we used 64 hidden units and induced an initial degree of orientation tuning for these units by training the model to categorize 12 equally spaced orientations (15 deg. increments). Then the model was trained to discriminate between 2 deg. orientation differences. We found that training induced a steepening of the slope of tuning curves in the hidden layer, just as was observed in area V1 of the monkeys. This occurred only for hidden units with a preferred orientation that differed by about 15 deg from the trained orientation.

We conclude that *Bapte* can account for changes in sensory representations that are caused by categorization training. It induces a selective representation of task-relevant features, and sharpens neuronal tuning curves at category boundaries. Further predictions of *Bapte* can be validated experimentally. First, only after rewarded trials does synaptic plasticity depend on reward expectancy. After unrewarded trials, synaptic plasticity is independent of the expected amount of reward. Second, the strength of feedback connections from motor neurons to sensory areas is proportional to the strength of the corresponding feedforward connections. Third, attentional feedback from areas involved in response selection gates plasticity of connections that are responsible for the selected action.

## References


[1] Boch, R., and Fischer, B. Saccadic reaction times and activation of the prelunate cortex: parallel observations in trained rhesus monkeys. *Exp.Brain Res.* **50**, 201-210 (1983).

[2]Maunsell, J. H. R. The brain's visual world: representation of visual targets in cerebral cortex. *Science* **270**, 764-769 (1995).

[3] Mazzoni, P., Andersen, R. A., and Jordan, M. I. A more biologically plausible learning rule for neural networks. *Proc. Natl. Acad. Sci. USA* **88**, 4433-4437 (1991).

[4] Schultz, W., Dayan,P. and Montague, P.R. A neural substrate of prediction and reward. *Science* **275**, 1593-1599 (1997).

[5] Waelti, P., Dickinson, A., and Schultz, W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* **412**, 43-48 (2001).

[6] Sigala, N., and Logothetis, N. K. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* **415**, 318-320 (2002).

[7] Schoups, A., Vogels, R., Qian, N., and Orban, G. A. Practising orientation identification improves orientation coding in V1 neurons. *Nature* **412**, 549-553 (2001).