LEARNING BY MODIFYING INTERNAL REPRESENTATIONS THROUGH REVERSE CONNECTIONS FROM OUTPUT

K.Ren & A.R. Gardner-Medwin
Dept. Physiology, UCL, London WC1E 6BT, UK

Proc. European Neuroscience Forum, Berlin, European J. Neuroscience 10: S10, p.17616, 1998

In a feedforward layered network of interconnected units, the activity patterns in the intermediate layer(s) may be called the internal representation of the inputs. Learning may be seen as an iterative process of modifying connection weights either from representation to output (R->O) or onto the representation itself. Most algorithms for modifying representations, e.g. Back Propagation of Errors, rely on global and accurate computations that cannot readily be implemented in neural structures because it is not clear how the required information could be represented at the sites of modification. Our algorithm relies on reverse projections from the desired output (O->R), broadly correlated with strengths of the forward R->O projections. This tends to shift the pattern formed in representation layers towards one that is more suited to learning the correct output. Thus the learning task becomes shared between modification of I->R weights and R->O weights. Simulation with standard benchmark tasks shows better generalization performance with faster convergence than conventional algorithms. The algorithm implicitly requires the existence of temporary memory to retain and recall the paired input and output patterns and the modified internal representations.