Review-5

A Learning Algorithm for Boltzmann Machines

Increasing interest in parallel search and connection-based methods has motivated efficient learning algorithms. Learners aim to represent the intricate aspects of data by communicating the compatibility of entities within themselves. One such learner is the Boltzmann Machine which is trained using an energy-based objective. The work presents a novel algorithm for training Boltzmann Machines which is based on energy compatibilities between the input and output representations. Learning is carried out in an architecture consisting of visible and hidden computational units utilizing a Boltzmann distribution to predict the probabilities of spatial representations. These probabilities are collected towars the end of each trial and used in the update rule for parameters of the network. The network is tested on a number of bit encoding tasks.

The Boltzmann Machine is a generative model which produces a representation of the input. Training of Boltzmann Machines is greatly hindered by computational speed and the choice of a suitable update rule. While the binar threshold rule serves as a potential candidate, it does not allow the network to be generalized to different inputs and tasks. An alternate method for training Boltzmann Machines is presented which consists of an update rule based on probabilistic inference. The update rule is obtained from relative entropy (KL-divergence) between the distributions of visible units when the network is trained and when the network is running freely. The information theoretic metric yields stable first order updates and reduced local convergence in comparison to the steepest descent method. The Boltzmann Machine, when trained and tested with the novel information theoretic metric on various binary code encoding tasks, demonstrates suitable performance and improving success rates. In the case of the 40-10-40 encoder, the network achieves 98.6% success rate within 1200 cycles.

The Boltzmann Machine utilizes an energy-based objective which depicts three key benefits in comparison to steepest descent methods. Firstly, the method can approximate first and second order derivatives accurately. Secondly, the metric is an apt update based on the similarity between the two distributions for generalization. Lastly, elimination of noise while estimating the probabilities can be reduced by utilizing an annealing schedule for the metric. While the objective poses as a suitable candidate for learning, it presents two major downsides. (1) Since the update consists of probabilistic measures, the magnitude of updates may tend to grow large. One alternative of bounding the weights is by utilizing noise clamping. However, the approach does not present significant results depicting its suitability. (2) In the case of high-dimensional tasks such as the 40-10-40 encoder, the network performs well but mostly depends on the annealing schedule to capture finer aspects of representations. This dependency of the network on annealing indicates room for architectural improvement.