Jenish Maharjan **CSC 9010_001 - Week 1**

Paper: *"Learning Representations by Backpropagating Errors"*

Summary:

"Learning Representations by Backpropagating Errors" by Rumelhart, Hinton and Williams introduce backpropagation as a learning procedure in networks with hidden units. The procedure aims at minimizing the distance between the actual and the desired output vectors of the network by making adjustments to the weights of the connections in the network. The learning procedure in networks that have internal layers with hidden units is not as simple as the networks where the inputs are directly connected to the outputs. These hidden units, different from the feature analyzers in a perceptron, are able to represent important features due to these adjustments to the weights. Backpropagating errors through a network with hidden units is presented as a simple yet powerful method to construct internal representations.

Authors describe the learning procedure with respect to networks that have an input layer, a number of intermediate layers and an output layer. The learning procedure starts by computing the outputs of each unit using an activation function over the linear combination of the outputs from the units in the lower layers that are connected to it and their weights. Although any input-output function which has a bounded derivative can be used, the authors emphasize the use of linear combination of the inputs and the weights and applying a non-linear activation function to simplify the learning procedure.

In order to find weights that produce output vectors for each input vector that is the same as the desired output vector, the error is calculated by comparing the actual outputs and the desired outputs. The error is minimized using gradient descent for which the partial derivative of the error with respect to each weight in the network needs to be computed which is equivalent to the sum of partial derivatives of each input-output case. This is termed as forward pass, one among the two passes used for minimizing the error. The other pass termed as backward pass propagates derivatives from the output layers to the lower layers. The authors show that it is possible to compute the partial derivative of the error for the units in the penultimate layer when the partial derivative of the error in the last layer is known. This procedure can be repeated successively to computer partial derivatives of the errors in the earlier layers. Using the results of these computations the weights are updated in two ways: update weights after every input-output case or accumulate the derivatives over all input-output cases before updating the weights. The latter is the simplest way of using gradient descent to update weights but it does not converge as rapidly as other methods that make use of the second derivatives. To improve this, the authors introduce an acceleration method in which "the current gradient is used to modify the velocity of the point in weight space instead of its position".

The authors move on to describe the performance of their procedure in some example task domains. They also point that a drawback of their procedure could be the inability to always obtain the global minimum but this could be avoided by adding more connections to create extra dimensions in the weights space which would help circumvent the barriers that create poor local minima.

Strengths:

- The content of the paper is concise and clear. The authors give a good brief motivation about their research and move smoothly towards explaining the methods they're introducing. The language isn't verbose.
- The authors have articulated the mathematical equations and steps in computing gradient descent and updating very well in words which makes understanding them easier and more intuitive.

Weakness:

- It felt like the examples given in the paper could've been given more context and the images used in the examples could've been explained more clearly.
- The authors have demonstrated the performance of their method in several example task domains. They also point out the drawback of their procedure, scenarios where the drawback may be encountered and what could be done to avoid them. But these claims are not supported by any analytical results.

Confusions:

- When talking about the acceleration the authors use the phrase "weight space instead of its position". I am not very certain about what it means.
- The networks shown in the mirror symmetry detection and the isomorphic family trees examples were difficult to comprehend.

Discussion Questions:

- How does adding a few extra connections "provide paths around barriers that create local minima"?
- In the conclusion, the authors say that "it is worth looking for more biologically plausible ways of doing gradient descent in neural networks". Have the methods introduced after backpropagation been biologically plausible and is it usually a goal to introduce methods that are biologically plausible?
- The authors state the method of changing each weight by an amount proportional to the accumulated partial derivate of the error with respect to the weights do not converge as rapidly as the methods that make use of the second derivatives. What are examples of some methods that use the second derivative and are they used at all in neural networks?