Guided Dropout

Overfitting occurs when the difference between training loss and validation loss is high, over-parameterization of the neural network is the main reason for the above problem. To generalise the neural network better we opt for regularisation technique in which the loss function is deliberately increased and consequently it increases the complexity of learning a specific function, now millions of parameters tend to generalise better rather than rote learning the function. Dropout is one of the regularisation technique in which randomly a certain number of nodes (depending upon the dropout rate) of the neural network are dropped while training the Deep Learning model. In conventional dropout, dropout mask is sampled from Bernoulli distribution with probability $(1 - \theta)$ which is used to randomly drop nodes of the neural network at every iteration of training. This paper presents an intelligent approach to execute Dropout while training of Deep Learning model. Guided dropout introduces a strength parameter which distinguishes nodes of the neural network into two categories namely weak strength and high strength nodes respectively. Strength parameter associated within each node in a neural network signifies its contribution towards the overall performance of the neural network. Like weights and bias parameters of Deep Learning model, Strength parameter is also learned by the network through an optimizer (e.g. Stochastic Gradient Descent) via backpropagation. The paper presumes that while training a neural network if high strength nodes were dropped then low strength nodes could gradually improve their importance and would contribute to the performance of the neural network. While testing the model, all the nodes of the neural network would be considered for predictions and thus the accuracy would be significantly enhanced. Active and Inactive regions are defined in the paper, they incorporate high strength nodes and low strength nodes respectively. One of the plots gives inference regarding the dropping of nodes in inactive regions does not worsen the accuracy of the neural network as compared to the dropping of active nodes. Experimental and theoretical explanations conclude that the proposed guided dropout performance is greater than or equal to conventional dropout. Conventional dropout utilizes Bernoulli distribution to sample dropout mask $r^{(l)}$ (where 'l' represents a layer of neural network) with a probability of $(1 - \theta)$ and is given by $\widehat{a^{(l)}} = a^{(l)} \odot r^{(l)}$, where $\widehat{a^{(l)}}$ is a masked output of l^{th} layer of a neural network, $\odot$ denotes elementwise multiplication and $a^{(l)}$ is the actual output of l^{th} layer of a neural network. Considering one hidden layer neural network, with input vector $x \in R^{D_2}$, weights of 1st and 2nd respectively being $V \in R^{D_2 \times r}$, $U \in R^{D_1 \times r}$ and output prediction prior to activation function is $h_{U,V}(x) = UV^T x$. With the advent of guided dropout, the hypothesis class is modified as $h_{U,V}(x) = U\text{diag}(t)V^T x$, where t denotes strength of a node in the neural network. The guided algorithm is divided into two categories namely 'Guided Dropout (top-k)' and 'Guided Dropout (DR)'. Guided Dropout (top-k) method select (top-k) nodes (using strength parameter) to drop, the dropout mask for the same is expressed by: $r^{(l)} = t^{(l)} \leq \text{th}$, where $\text{th} = max_{[N \times (1-\theta)]}\ t^{(l)}$, where $max_{[N \times (1-\theta)]}$ represents k large elements of t, N is the total number of nodes in a neural network, $(1 - \theta)$ is the percentage ratio of the nodes which needs to be dropped and $r^{(l)}$ is the dropout mask. The expected loss is maximum while using guided dropout (top - k) method in comparison with the conventional dropout, thus it forces the maximum penalty in neural network loss. In the guided dropout (DR) method, nodes are Dropped Randomly (DR) from the active region. In this method, $(1 - \theta)$ is the dropout probability which refers to the likelihood of sampling dropout mask for active region nodes utilizing Bernoulli distribution in the neural network model. The analytic reason behind the working of guided dropout is the increase in the penalty in the form of loss in the neural network since the high strength nodes are chosen to form dropout mask, and thus optimizing the loss in the training process will help the inactive nodes to improve their performance in the absence of high strength nodes. Evaluating in Dense neural network architecture, Deep Convolution neural network and in small sample size problem, the proposed guided dropout (DR) yields better accuracies than the existing dropout methods or without dropout training technique.