The Incredible Shrinking Neural Network: Pruning to Operate in Constrained Memory Environments

Nikolas Wolfe, Aditya Sharma, Lukas Drude & Bhiksha Raj
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{nwolfe, adityasharma, bhiksha}@cmu.edu

Abstract

We propose and evaluate a method for pruning neural networks to operate in constrained memory environments such as mobile or embedded devices. We evaluate a simple pruning technique using first-order derivative approximations of the gradient of each neuron in an optimally trained network, and turning off those neurons which contribute least to the output of the network. We then show the limitations of this type of approximation by comparing against the ground truth value for the change in error resulting from the removal of a given neuron. We attempt to improve on this using a second-order derivative approximation. We also explore the correlation between neurons in a trained network and attempt to improve our choice of candidate neurons for removal to account for faults that can occur from the removal of a single neuron at a time. We argue that this method of pruning allows for the optimal tradeoff in network size versus accuracy in order to operate within the memory constraints of a particular device or application environment.

1 Second Derivative Gradient Terms

Given a neuron n with output a_i, and outgoing weights $[w_{i,1}, w_{i,2}, \ldots w_{i,j}]$, the input x to each of its j forward-connected neurons is given by:

$$x_j = \sum_i (w_{i,j} \cdot a_i) \tag{1}$$

For simplicity's sake, we will drop the index variable i and examine only the connection between the output a of neuron n and each of its j connections to the next forward layer. So the contribution c_j from neuron n to the input of each forward-connected neuron is given by:

$$c_j = w_j \cdot a \tag{2}$$

Where w_j is the weight connecting the output a from n to the input of the jth neuron. We will denote the error function of an optimally trained network as E. The second-derivative of E with respect to the output of neuron n is given by:

$$\frac{d^2E}{da^2} = \frac{d}{da}\frac{dE}{da} = \frac{d}{da}\sum_j \left(\frac{dE}{dc_j} \cdot \frac{dc_j}{da}\right) = \sum_j \left(\frac{d}{da}\frac{dE}{dc_j} \cdot \frac{dc_j}{da}\right) \tag{3}$$

Which states that the 2nd derivative of E with respect to a is the sum of the 2nd derivative terms of all outgoing connections.

2 SECOND DERIVATIVE BACK-PROPAGATION

Name and network definitions:

$$E = \frac{1}{2}\sum_i (o_i^{(0)} - t_i)^2 \quad o_i^{(m)} = \sigma(x_i^{(m)}) \quad x_i^{(m)} = \sum_j w_{ji}^{(m)} o_j^{(m+1)} \quad c_{ji}^{(m)} = w_{ji}^{(m)} o_j^{(m+1)} \tag{4}$$

Superscripts represent the index of the layer of the network in question, with 0 representing the output layer. E is the squared-error network cost function. $o_i^{(m)}$ is the ith output in layer m generated by the activation function σ, which in this paper is is the standard logistic sigmoid. $x_i^{(m)}$ is the weighted sum of inputs to the ith neuron in the mth layer, and $c_{ji}^{(m)}$ is the contribution of the jth neuron in the $m+1$ layer to the input of the ith neuron in the mth layer.

2.1 FIRST AND SECOND DERIVATIVES

The first and second derivatives of the cost function with respect to the output:

$$\frac{\partial E}{\partial o_i^{(0)}} = o_i^{(0)} - t_i \tag{5}$$

$$\frac{\partial^2 E}{\partial (o_i^{(0)})^2} = 1 \tag{6}$$

The first and second derivatives of the sigmoid function in forms depending only on the output:

$$\sigma\prime = \sigma\left(1-\sigma\right) \tag{7}$$
$$\sigma\prime\prime = \sigma\prime\left(1-2\sigma\right) \tag{8}$$

And for future convenience:

$$\frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} = \frac{\partial}{\partial x_i^{(m)}}\left(o_i^{(m)} = \sigma(x_i^{(m)})\right) = \left(o_i^{(m)}\right)\left(1 - o_i^{(m)}\right) \tag{9}$$
$$\frac{\partial^2 o_i^{(m)}}{\partial {x_i^{(m)}}^2} \tag{10}$$

Derivative of the error with respect to the ith neuron's input $x_i^{(0)}$ in the output layer:

$$\frac{\partial E}{\partial x_i^{(0)}} = \frac{\partial E}{\partial o_i^{(0)}}\frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \tag{11}$$
$$= \underbrace{\left(o_i^{(0)} - t_i\right)}_{\text{from (5)}}\underbrace{\left(\sigma\left(x_i^{(0)}\right)\left(1 - \sigma\left(x_i^{(0)}\right)\right)\right)}_{\text{from (7)}} \tag{12}$$
$$= \left(o_i^{(0)} - t_i\right)\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right) \tag{13}$$
$$= \left(o_i^{(0)} - t_i\right)(\sigma\prime) \tag{14}$$

Second derivative of the error with respect to the ith neuron's input $x_i^{(0)}$ in the output layer:

$$\frac{\partial^2 E}{\partial (x_i^{(0)})^2} = \frac{\partial}{\partial x_i^{(0)}} \left(\frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \right) \tag{15}$$

$$= \frac{\partial^2 E}{\partial x_i^{(0)} o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} + \frac{\partial E}{\partial o_i^{(0)}} \frac{\partial^2 o_i^{(0)}}{\partial {x_i^{(0)}}^2} \tag{16}$$

$$= \frac{\partial^2 E}{\partial x_i^{(0)} o_i^{(0)}} \underbrace{\left(o_i^{(0)} \left(1 - o_i^{(0)} \right) \right)}_{\text{from (7)}} + \underbrace{\left(o_i^{(0)} - t_i \right)}_{\text{from (5)}} \underbrace{\left(o_i^{(0)} \left(1 - o_i^{(0)} \right) \right) \left(1 - 2 o_i^{(0)} \right)}_{\text{from } 8} \tag{17}$$

$$\left(\frac{\partial^2 E}{\partial x_i^{(0)} o_i^{(0)}} \right) = \frac{\partial}{\partial x_i^{(0)}} \frac{\partial E}{\partial o_i^{(0)}} = \frac{\partial}{\partial x_i^{(0)}} \underbrace{\left(o_i^{(0)} - t_i \right)}_{\text{from (5)}} = \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} = \underbrace{\left(o_i^{(0)} \left(1 - o_i^{(0)} \right) \right)}_{\text{from (7)}} \tag{18}$$

$$\frac{\partial^2 E}{\partial (x_i^{(0)})^2} = \left(o_i^{(0)} \left(1 - o_i^{(0)} \right) \right)^2 + \left(o_i^{(0)} - t_i \right) \left(o_i^{(0)} \left(1 - o_i^{(0)} \right) \right) \left(1 - 2 o_i^{(0)} \right) \tag{19}$$

$$= (\sigma\prime)^2 + \left(o_i^{(0)} - t_i \right) (\sigma\prime\prime) \tag{20}$$

First derivative of the error with respect to a single input contribution $c_{ji}^{(0)}$ from neuron j to neuron i with weight $w_{ji}^{(0)}$ in the output layer:

$$\frac{\partial E}{\partial c_{ij}^{(0)}} = \frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}} \tag{21}$$

$$= \underbrace{\left(o_i^{(0)} - t_i \right)}_{\text{from (5)}} \underbrace{\left(o_i^{(0)} \left(1 - o_i^{(0)} \right) \right)}_{\text{from (7)}} \frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}} \tag{22}$$

$$\left(\frac{\partial x_i^{(m)}}{\partial c_{ji}^{(m)}} \right) = \frac{\partial}{\partial c_{ji}^{(m)}} \left(x_i^{(m)} = \sum_j w_{ji}^{(m)} o_j^{(m+1)} \right) = \frac{\partial}{\partial c_{ji}^{(m)}} \left(c_{ji}^{(m)} + k \right) = 1 \tag{23}$$

$$\frac{\partial E}{\partial c_{ji}^{(0)}} = \left(o_i^{(0)} - t_i \right) \left(o_i^{(0)} \left(1 - o_i^{(0)} \right) \right) \tag{24}$$

Second derivative of the error with respect to a single input contribution $c_{ji}^{(0)}$:

$$\frac{\partial^2 E}{\partial {c_{ji}^{(0)}}^2} = \frac{\partial^2 E}{\partial c_{ji}^{(0)} o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \underbrace{\frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}}}_{(23)} + \underbrace{\frac{\partial E}{\partial o_i^{(0)}}}_{(5)} \frac{\partial^2 o_i^{(0)}}{\partial c_{ji}^{(0)} x_i^{(0)}} \underbrace{\frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}}}_{(23)} + \underbrace{\frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \frac{\partial^2 x_i^{(0)}}{\partial {c_{ji}^{(0)}}^2}}_{\frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}} = 1, \text{ so } \frac{\partial^2 x_i^{(0)}}{\partial {c_{ji}^{(0)}}^2} = 0} \tag{25}$$

$$= \left(o_i^{(0)} - t_i \right) \tag{26}$$

2.2 RETRIEVAL OF STYLE FILES

The style files for ICLR and other conference information are available on the World Wide Web at

`http://www.iclr.cc/`

The file `iclr2016_conference.pdf` contains these instructions and illustrates the various formatting requirements your ICLR paper must satisfy. Submissions must be made using LaTeX and the style files `iclr2016_conference.sty` and `iclr2016_conference.bst` (to be used with LaTeX2e). The file `iclr2016_conference.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in sections **??**, **??**, and **??** below.