ZACHARY ROSS

SPARSE NETWORKS AND DIFFERENTIAL PRIVACY

Pruning Neural Networks without any data by iteratively conserving synaptic flow

Pruning Algorithms

Pruning algorithms are generally defined by

1. scoring parameters by some metric, and
2. masking parameters according to their scores.

The latter is generally done by removing the parameters (e.g. making the value zero via Hadamard product). This can either be done via global masking or layer-masking, with global-masking performing better but suffering from *layer-collapse*, which is when an entire layer is masked.

Let θ be a collection of network parameters and θ_{prune} be the parameters remaining after pruning. The *compression ratio* ρ of a pruning algorithm is

$$\rho = \frac{|\theta|}{|\theta_{\text{prune}}|} \tag{1}$$

. We define $\rho_{\max}$ as the maximal possible compression ratio for a *network* that doesn't lead to layer collapse and the ρ_{cr} as the maximal compression ratio a given *algorithm* can achieve without inducing layer collapse. Note that the distinction in the two is $\rho_{\max}$ is maximal for a network while ρ_{cr} is maximal for an algorithm. These definitions motivate the following axiom:

Axiom 1 (Maximal Critical Compression)**.** *For any pruning algorithm and any network, we should always have* $\rho_{cr} = \rho_{\max}$.

Synaptic Saliency

Synaptic saliency is a class of score metrics defined by

$$\mathcal{S}(\theta) = \frac{\partial \mathcal{R}}{\partial \theta} \odot \theta \tag{2}$$

where $\mathcal{R}$ is a scalar loss function of the output y of a feed-forward network parameterized by θ. This metric satisfies two conservation laws:

Theorem 1 (Neuron-wise Conservation of Synaptic Saliency)**.** *For a feedforward neural network with continuous, homogeneous activation functions* $\phi(x) = \phi'(x)x$*, let* j *be the index of a hidden neuron in layer* i*. The sum of the synaptic saliency for the incoming parameters to a hidden neuron* $\left(\mathcal{S}_j^{(l)} = \left(\frac{\partial \mathcal{R}}{\partial \theta_j^{(l)}}\right)^{\mathsf{T}} \theta_j^{(l)}\right)$ *is equal to the sum of the synaptic saliency for the outgoing parameters from the hidden neuron* $\left(\mathcal{S}^{(l+1)} = \left(\frac{\partial \mathcal{R}}{\partial \theta_{:,j}^{(l+1)}}\right)^{\mathsf{T}} \theta_{:,j}^{(l+1)}\right)$.

Proof. Let each neuron in layer l be defined by $z_i^{(l)} = \theta_i^{(l)} \phi\left(z^{(l-1)}\right) = \sum_{j=1}^{n} \theta_{i,j}^{(l)} \phi\left(z_j^{(l-1)}\right)$ and note the following

$$\begin{aligned} \frac{\partial R}{\partial \phi\left(z_j^{(l-1)}\right)} &= \sum_{i=1}^{m} \frac{\partial R}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial \phi\left(z_j^{(l-1)}\right)} && \text{(total derivative)} \\ &= \sum_{i=1}^{m} \frac{\partial R}{\partial z_i^{(l)}} \theta_{i,j}^{(l)} && \text{(substituion)} \end{aligned} \tag{3}$$

.

For the incoming parameters, we have that

$$\begin{aligned} \mathcal{S}_i^{(l)} &= \sum_{j=1}^{n} \frac{\partial R}{\partial \theta_{ij}^{(l)}} \theta_{ij}^{(l)} && \text{(definition)} \\ &= \sum_{j=1}^{n} \frac{\partial R}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial \theta_{ij}^{(l)}} \theta_{ij}^{(l)} && \text{(chain rule)} \\ &= \sum_{j=1}^{n} \frac{\partial R}{\partial z_i^{(l)}} \phi\left(z_j^{(l-1)}\right) \theta_{ij}^{(l)} && \text{(substitution)} \\ &= \frac{\partial R}{\partial z_i^{(l)}} z_i^{(l)} && \text{(vectorization)} \end{aligned} \tag{4}$$

and for the outgoing parameters we have that

$$\begin{aligned} \mathcal{S}_j^{(l+1)} &= \sum_{i=1}^{n} \frac{\partial R}{\partial \theta_{ij}^{(l+1)}} \theta_{ij}^{(l+1)} && \text{(definition)} \\ &= \sum_{i=1}^{m} \frac{\partial R}{\partial z_i^{(l+1)}} \frac{\partial z_i^{(l+1)}}{\partial \theta_{ij}^{(l+1)}} \theta_{ij}^{(l+1)} && \text{(chain rule)} \\ &= \sum_{i=1}^{n} \frac{\partial R}{\partial z_i^{(l+1)}} \phi\left(z_j^{(l)}\right) \theta_{ij}^{(l+1)} && \text{(substitution)} \\ &= \left(\sum_{i=1}^{n} \frac{\partial R}{\partial z_i^{(l+1)}} \theta_{ij}^{(l+1)} \right) \phi\left(z_j^{(l)}\right) && \text{(rearrange variables)} \\ &= \frac{\partial R}{\partial \phi\left(z_j^{(l)}\right)} \phi\left(z_j^{(l)}\right) && \text{(substitute equation 3)} \end{aligned} \tag{5}$$

.

This shows that $\mathcal{S}_j^{(l)} = \mathcal{S}_j^{(l+1)}$ so long as ϕ is homogeneous. □

Theorem 2 (Network-wise Conservation of Synaptic Saliency)**.** *The sum of the synaptic saliency across any set of parameters that exactly separates the input neurons x from the output neurons y of a feedforward neural network with homogeneous activation functions equals* $\left(\frac{\partial \mathcal{R}}{\partial x}\right)^{\mathsf{T}} x = \left(\frac{\partial \mathcal{R}}{\partial y}\right)^{\mathsf{T}} y$.