MACHINE INTELLIGENCE COMMUNITY

meProp: Sparsified Back Propagation ...

Sun et al. https://arxiv.org/pdf/1706.06197.pdf

Summary by Brian Fakhoury

---

1) Training neural networks can be a slow process. Deeper networks can have hundreds or thousands of weights, of which all are modified when the model is updated.
    a) Every time the backward pass is calculated, calculating all the local gradients can create a long computation time
2) Furthermore, many of the weights are barely changed since they may have very little contribution to the specific loss scenario.
3) According to the team, only 1-4% of the weights in the parameter matrix need to be updated per backward pass.
    a) The total cost for calculating the backpropagation is directly related to the dimension of the full gradient vector.
4) The proposed method would operate normally on the forward pass, but, instead of calculating the entire gradient vector, only the vectors with greatest magnitudes will be considered, hence the name minimal effort.

$$\frac{\partial L}{\partial W} \leftarrow top_k(\frac{\partial L}{\partial y}) \cdot \frac{\partial y}{\partial W} \qquad (9)$$

$$\frac{\partial L}{\partial x} \leftarrow \frac{\partial y}{\partial x} \cdot top_k(\frac{\partial L}{\partial y}) \qquad (10)$$

    a) The top k vectors will reflect on the local gradients that are largest, and therefore have the most effect on the model output.

5) The team's results were positive when this method was applied to popular problems, like recognizing the MNIST dataset (http://yann.lecun.com/exdb/mnist/)
    a) By only calculating the gradient vector in batches with the Adam optimizer, they only updated a few percent of the weights in their MLP and LSTM models.
6) The team also observed reduced overfitting when using only the top k gradients, and that this method scaled when adding more layers.
    a) This is due to meProp leaving weights that have little contribution untouched.
        i) Therefore, a more general model is produced as connections reflect more accurately on data they need to single out: better noise reduction.
    b) Adding more layers only scaled the weight matrix by one dimension, and picking the top k gradient vectors by magnitude will scale by the increase in linear factor.