Weight Uncertainty in Neural Networks

ICML 2015

Summary

In this paper, a new algorithm for training probabilistic neural networks, referred to as *Bayes by Backprop*, is proposed. The algorithm improves upon the work proposed by Graves in 2011 (see paper "Practical Variational Inference in Neural Networks) by computing unbiased gradients and allowing for non-gaussian priors. Training a network using Bayes by Backprop corresponds to an ensemble of networks with weights drawn from a shared posterior distribution. Evaluated on MNIST, the algorithm performs on-par with networks regularized using dropout.

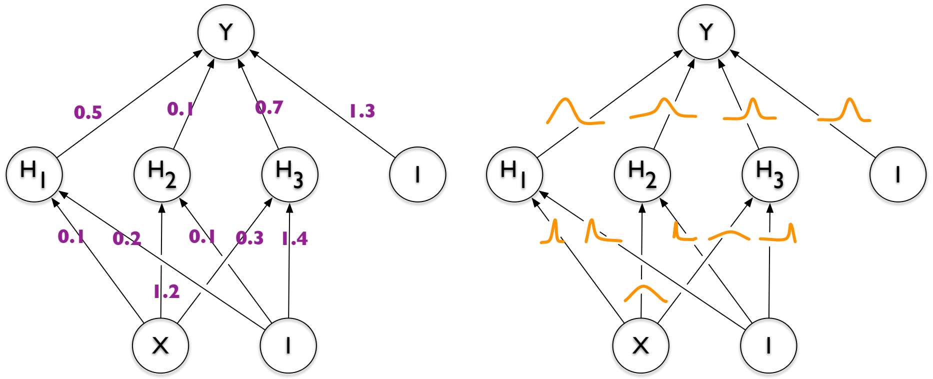

Figure 1: Comparison of network trained using conventional backpropagation and a network trained using Bayes by Backprop. Instead of learning a point estimate of each weight, the full posterior distribution of each weight given the training data is learned.

Main Contributions

- **Unbiased Monte Carlo Gradients** The authors put forward a simple formula for the gradients of the variational free energy w.r.t the parameters of the variational distribution that reuses the gradients needed for conventional backpropagation.
- **Scale Mixture Prior** Contrary to other works that tried to optimize the parameters of the prior distribution of the weights, a mixture of two gaussian is proposed that alleviates the need for optimization of more parameters.
- **Mini-Batch Re-Weighting** Based on the assumption that uniform sampling of mini-batches can result in a non-uniform distribution of the cost function, a weighting scheme is proposed that puts more emphasis on the prior in the beginning of the training. The more data is observed, the more influence is then assigned to the likelihood of the data.
- **Model Pruning** It is shown that the weights obtained using Bayes by Backprop display a significantly higher level of scarcity compared to conventional SGD. Consequently, a large number of weights can be removed from the network after training without impairing performance.

Implementation Details

- **Approximation of the Variational Free Energy** Following the popular approach of approximating the intractable posterior over the network's weights $P(\mathbf{w}|\mathcal{D})$ by a simpler variational distribution $q(\mathbf{w}|\theta)$, the cost function to be minimized corresponds to the variational free energy:

$$\begin{aligned}\mathcal{F}(\mathbf{w}|\theta) = &\quad \mathbf{KL}\left[q(\mathbf{w}|\theta)\,||\,P(\mathbf{w})\right] - \mathbb{E}_{q(\mathbf{w}|\theta)}\left[\log P(\mathcal{D}|\mathbf{w})\right] \\ \approx &\sum_{i=1}^{n} \log q(\mathbf{w}^{(i)}|\theta) - \log P(\mathbf{w}^{(i)}) - \log P(\mathcal{D}|\mathbf{w}^{(i)})\end{aligned}$$

  This approximation ensures that all terms in the cost function depend on the particular Monte Carlo sample to the weights $\mathbf{w}^{(i)}$.

- **Parameter Update** Given a variational gaussian posterior distribution, a sample of the weights is given by $\mathbf{w} = \mu + \log(1 + \exp(\rho)) \cdot \epsilon$ for $\epsilon \sim \mathcal{N}(0, I)$. The gradients w.r.t. the variational parameters $\theta = (\mu, \rho)$ are given by:

$$\begin{aligned}\Delta_\mu &= \frac{\partial f(\mathbf{w},\theta)}{\partial \mathbf{w}} + \frac{\partial f(\mathbf{w},\theta)}{\partial \mu} \\ \Delta_\rho &= \frac{\partial f(\mathbf{w},\theta)}{\partial \mathbf{w}} \frac{\epsilon}{1 + \exp(-\rho)} + \frac{\partial f(\mathbf{w},\theta)}{\partial \rho}\end{aligned}$$

Evaluation

- **MNIST Classification** Bayes by Backprop is used to train a 2-Layer ReLU CNN with 1200 units per layer and the performance is compared to the same architecture trained using conventional SGD and dropout SGD. The Bayes by Backprop variant with a gaussian mixture prior achieves a test error of 1.32% on-par with dropout SGD (1.36%).

- **Weight Scarcity & Model Pruning** It can be observed that the weight values obtained through Bayes by Backprop display a much wider spread compared to SGD and dropout. This phenomenon is illustrated in Figure 2. Consequently, a large fraction of the weights (up to 75%) can be removed based on the Signal-to-Noise ratio ($|\mu_i|/\sigma_i$) without negatively affecting the test error.

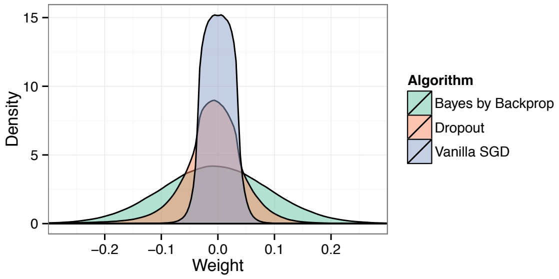

Figure 2: Histogram of the trained networks' weights.

References

This summary is solely based on my understanding of the original paper. All images used here are taken from the original paper as well. The paper can be found under the following link:
`https://arxiv.org/pdf/1505.05424.pdf`