Bayesian Neural Networks (BNNs)

Represent weights by probability distributions over possible values, rather than a single fixed value.

Variational approach: Approximate the posterior $P(w|D)$ with the variational distribution $q(w; \theta)$ minimizing the Kullback-Leibler (KL) divergence

$$ \mathcal{L}(D, \theta) = \sum_{i=1}^{n} \left[ kl(q(w_i; \theta)||P(w_i)) - log P(D|w_i) \right]$$

where $P(D, \theta)$ is called variational free energy

$$F(D, \theta) = KL(q(w_i; \theta)||P(w)) - \sum_{i=1}^{n} log P(D|w_i)$$

Advantages
- Uncertainty estimation
- Regularization

Disadvantages
- Long training time
- Intractable posteriors

Bayes By Backprop (BBB)

Approximate $F(D, \theta)$ using Monte Carlo:

$$F(D, \theta) \approx \sum_{i=1}^{n} \left[ log q(w_i; \theta) - log P(w_i) - log P(D|w_i) \right]$$

where $w_i$ is the $i$th MC sample drawn from the variational posterior $q(w_i; \theta)$

Advantages
- Accurate predictions from cheap model averaging

Disadvantages
- Requires MC variance control
- Requires careful prior elicitation

Deterministic Variational Inference (DVI)

Likelihood cost:
- Deterministic form to approximate the final layer activation distribution $q(w_i)$

Advantages
- Remove MC stochasticity
- Automatic prior selection

Disadvantages
- Closed-form limits design
- High compute cost on wide nets

Weight Uncertainty in Neural Networks

John Boom 1  Emma Prévot 2  Ilaria Sartori 1

1University of Cambridge, Department of Engineering

Figure 1. Bayesian Neural Networks.

Figure 4. Model diagnosis confidence on dermatoscope pictures from DermaMNIST.

Table 1. MNIST Classification Accuracy. SGD and BBB methods were trained for 300 epochs, with 400 hidden units (480k) and 1200 hidden units (2.4m). DVI trained only for 30 epochs for computational complexity.

Table 2. BBB DermaMNIST Classification accuracy against state-of-the-art.

Table 3. BBB DermaMNIST Classification accuracy against state-of-the-art.

Figure 5. Model diagnosis confidence on dermatoscope pictures from DermaMNIST.

Figure 6. Comparison of BBB, DVI, and a Standard Neural Net on toy datasets with homoskedastic noise (row 1), heteroskedastic noise (row 2), and discontinuous data (row 3). The scattered data points are the training data. Grey is the true function; Red is the mean prediction; Blue is 2 standard deviations.

Figure 3. Histogram of the trained weights.

Table 2. BBB classification accuracy after weight pruning of the 400 hidden units. Mixture BBB model.

Table 1. MNIST Classification Accuracy. SGD and BBB methods were trained for 300 epochs, with 400 hidden units (480k) and 1200 hidden units (2.4m). DVI trained only for 30 epochs for computational complexity.

Table 3. BBB DermaMNIST Classification accuracy against state-of-the-art.

Figure 6. Comparison of BBB, DVI, and a Standard Neural Net on toy datasets with homoskedastic noise (row 1), heteroskedastic noise (row 2), and discontinuous data (row 3). The scattered data points are the training data. Grey is the true function; Red is the mean prediction; Blue is 2 standard deviations.

Table 2. BBB classification accuracy after weight pruning of the 400 hidden units. Mixture BBB model.

Table 2. BBB classification accuracy after weight pruning of the 400 hidden units. Mixture BBB model.

Table 3. BBB DermaMNIST Classification accuracy against state-of-the-art.

Figure 5. Model diagnosis confidence on dermatoscope pictures from DermaMNIST.

Figure 6. Comparison of BBB, DVI, and a Standard Neural Net on toy datasets with homoskedastic noise (row 1), heteroskedastic noise (row 2), and discontinuous data (row 3). The scattered data points are the training data. Grey is the true function; Red is the mean prediction; Blue is 2 standard deviations.

Table 2. BBB classification accuracy after weight pruning of the 400 hidden units. Mixture BBB model.

Figure 5. Model diagnosis confidence on dermatoscope pictures from DermaMNIST.