Practical Variational Inference for Neural Networks

Alex Graves

CIFAR Junior Fellow
University of Toronto
Canada

Method

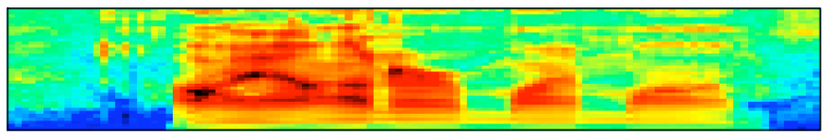
- Instead of learning neural network weights, we learn the mean and variance of a separate Gaussian for each weight: adaptive weight noise
- The loss is the number of bits to transmit the errors plus the number of bits to transmit the weights: optimisation = compression
- The more information the weights store about the training data, the more they cost to send: no overfitting
- Can interpret as MDL or stochastic variational inference

Advantages

- Applies to any differentiable log-loss model (previous variational methods for neural networks were limited to very simple architectures)
- No validation set required (as long as the training data is compressed)
- The weight costs tell you how important each weight is to the network
- Can prune the network by removing weights with high probability at zero

Results

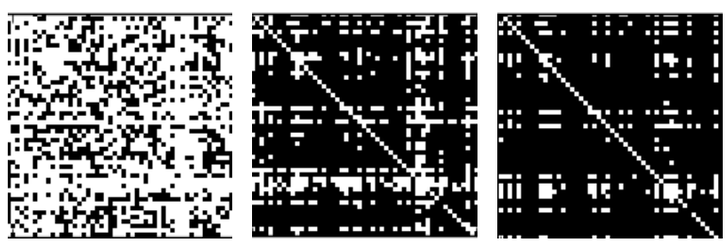
- Outperformed other regularisers for phoneme recognition on TIMIT with a complex neural network



ay aa nx er m ay m aa m

Regulariser	Error Rate
L2	27.4%
L1	26.0%
Weight noise	25.4%
Adaptive weight noise	23.8%

- Allowed many weights to be pruned with little impact (even improvement!) on performance



Weight matrix at different pruning thresholds: black=prune, white=keep

Weights Pruned	Error Rate
22.6%	24.0%
54.8%	23.5%
69.1%	23.7%
88.5%	24.5%