A revisit to Mackay algorithm and its application to deep network compression

Chune LI, Yongyi MAO, Richong ZHANG, Jinpeng HUAI

Frontiers of Computer Science, DOI: 10.1007/s11704-019-8390-z

Problems & Ideas

- Problems of iterative procedure in MacKay's evidence framework for estimating hyperparameter in empirical Bayes and pruning model parameters
  - The iterative estimation procedure has stayed primarily as a heuristic to date
  - Its application to deep neural network has not yet been explored.
- Ideas: Apply MacKay algorithm to deep network compression
  - Justify the iterative procedure in MacKay evidence framework as a well-principled algorithm.
  - Adopt MacKay algorithm to prune the parameters of deep convolution networks.

$$W \sim \prod_{oi} \mathcal{N}(W_{oi}; 0, \frac{1}{\alpha_i}) \quad \text{(linear layer).} \qquad \alpha_i := \frac{O}{\sum_o W^2_{\text{MP}oi}} \quad \text{(for linear layer)}$$

$$W \sim \prod_{oivr} \mathcal{N}(W_{oivr}; 0, \frac{1}{\alpha_o}) \quad \text{(convolutional layer)} \qquad \alpha_o := \frac{IVV}{\sum_{ivr} W^2_{\text{MP}oivr}} \quad \text{(for convolutional layer)}$$

Main Contributions

The MacKay algorithm is a coordinate ascent procedure for optimizing a lower bound of the log-evidence.

OptMacKay

Find $\boldsymbol{\alpha}$ and $\boldsymbol{z}$ that maximize $\mathcal{F}_{\text{MacKay}}(\boldsymbol{z}, \boldsymbol{\alpha})$.

The MacKay algorithm is then defined as the following coordinate ascent procedure for optimizing $\mathcal{F}_{\text{MacKay}}$.

MacKay Algorithm

$\boldsymbol{z}$-Step:

$$\begin{aligned} \boldsymbol{z}^{(t)} &:= \arg\max_{\boldsymbol{z}} \mathcal{F}_{\text{MacKay}}(\boldsymbol{z}, \boldsymbol{\alpha}^{(t)}) \\ &= \arg\max_{\boldsymbol{z}} \log p(\boldsymbol{D}, \boldsymbol{z} | \boldsymbol{\alpha}^{(t)}) \end{aligned}$$

$\boldsymbol{\alpha}$-Step:

$$\boldsymbol{\alpha}^{(t+1)} := \arg\max_{\boldsymbol{\alpha}} \mathcal{F}_{\text{MacKay}}(\boldsymbol{z}^{(t)}, \boldsymbol{\alpha}).$$

The algorithm can compress neural networks to a high level of sparsity with little loss of prediction accuracy, which is comparable with the state-of-the-art.

Table 6 Comparison of the compression of ResNet networks on CIFAR-10.

network	method	error(%)	edge(%)	FLOP(%)
ResNet-56	SFP-10 [40]	**6.11**		85.3
	SFP-40y [40]	6.65		47.4
	ARD@2	6.70	47.0	53.1
	PF-A [25]	6.90	90.6	89.6
	PF-B [25]	6.94	86.3	72.4
	CP [39]	8.20		50.0
	ARD@ 3	10.49	**12.7**	**21.9**
ResNet-110	SFP-20 [40]	**6.07**		71.8
	SFP-30y [40]	6.14		59.2
	ARD@2	6.40	24.5	31.1
	PF-A [25]	6.45	97.7	84.1
	LCCN [38]	6.56		65.8
	PF-B [25]	6.70	67.6	61.4
	ARD@3	10.32	**6.2**	**11.0**