Reweighted ℓ_1 Minimization for pruning

Northeastern University
MIT-IBM Watson AI Lab

Consider regularized weight pruning problem

$$\underset{\mathbf{w}}{\text{minimize}} \quad f(\mathbf{w}) + \gamma \textstyle\sum_i \alpha_i |w_i|, \tag{1}$$

where $\alpha_i > 0$ is a positive weight that is determined by reweighted ℓ_1 algorithm, and $\boldsymbol{w}_i$ is the weight parameter of the i-th layer. $f(\boldsymbol{W})$ is the loss function and $|\cdot|$ is the L1 norm.

The algorithm sketch is provided as below. Given initial weights $\boldsymbol{\alpha}^{(0)} = \mathbf{1}$, namely, $\alpha_i = 1$ for any i, the proposed algorithm performs for $t = 1, 2, \ldots, T$

1. Set $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(t-1)}$, and solve problem (1) via gradient descent method (SGD or Adam) to obtain a solution $\mathbf{w}^{(t)}$.

2. Update weights $\alpha_i^{(t)} = \frac{1}{|w_i^{(t)}| + \epsilon}$ for any i, where ϵ is a given small number, e.g., $\epsilon = 0.01$.

3. Iteration terminates when $t = T$ (it is common to choose $T < 10$), or $\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\|_2 \leq \epsilon$, where ϵ is a given stopping tolerance, e.g., 0.01.