Flexible Learning of Sparse Neural Networks via Constrained L_0 Regularization

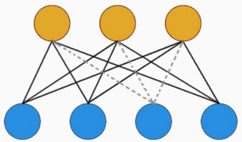
Jose Gallego-Posada, Juan Ramirez and Akram Erraqabi

Mila · Université de Montréal · Universidad EAFIT

Sparsity

Training, storing and deploying NNs can be very expensive. Fortunately, their performance is **robust** to parameter pruning.

A method for obtaining efficient neural networks is by training them to encourage **sparsity** during training.



$$\|\boldsymbol{\theta}\|_0 = \sum_{j=1}^{|\boldsymbol{\theta}|} \mathbb{I}\{\theta_j \neq 0\}$$

Regularization via L_0 penalties

(Louizos et al., 2018)

Penalize the number of active parameters.

$$\mathcal{R}(\boldsymbol{\theta}) = \frac{1}{N}\left(\sum_{i=1}^{N} \ell\left(h(x_i;\, \boldsymbol{\theta}),\, y_i\right)\right) + \lambda\|\boldsymbol{\theta}\|_0$$

Re-parametrize with **differentiable** stochastic gates based on concrete distributions.

$$\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \odot \boldsymbol{z}$$

free parameter magnitudes · gates

$$\mathcal{R}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}) \triangleq \mathbb{E}_{\boldsymbol{z}\,|\,\boldsymbol{\phi}}\left[\mathcal{R}(\tilde{\boldsymbol{\theta}} \odot \boldsymbol{z})\right]$$

$$= \mathbb{E}_{\boldsymbol{z}\,|\,\boldsymbol{\phi}}\left[\frac{1}{N}\sum_{i=1}^{N} \ell\left(h(x_i;\, \tilde{\boldsymbol{\theta}} \odot \boldsymbol{z}),\, y_i\right)\right] + \lambda\, \mathbb{E}_{\boldsymbol{z}\,|\,\boldsymbol{\phi}}\left[\|\boldsymbol{z}\|_0\right]$$

Regularization via L_0 constraints

$$\min_{\tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}} \quad \mathfrak{f}_{\text{obj}}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}) \triangleq \mathbb{E}_{\boldsymbol{z}\,|\,\boldsymbol{\phi}}\left[\frac{1}{N}\sum_{i=1}^{N} \ell\left(h(x_i;\, \tilde{\boldsymbol{\theta}} \odot \boldsymbol{z}),\, y_i\right)\right]$$

$$\text{subject to} \quad \mathfrak{g}_{\text{const}}(\boldsymbol{\phi}) \triangleq \mathbb{E}_{\boldsymbol{z}\,|\,\boldsymbol{\phi}}\left[\, \|\boldsymbol{z}\|_0 \right] \leq \epsilon \cdot |\boldsymbol{\theta}|$$

expected # of active params · In [0, 1] · # of network params

Consider the associated Lagrangian and min-max game:

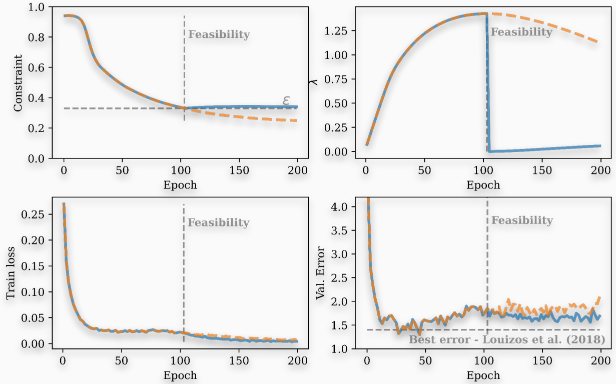
$$\mathcal{L}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}, \lambda) \triangleq \mathfrak{f}_{\text{obj}}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}) + \lambda\left(\frac{\mathfrak{g}_{\text{const}}(\boldsymbol{\phi})}{|\boldsymbol{\theta}|} - \epsilon\right)$$

$$\tilde{\boldsymbol{\theta}}^*, \boldsymbol{\phi}^*, \lambda^* = \operatorname*{argmin}_{\tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}} \operatorname*{argmax}_{\lambda \geq 0} \mathcal{L}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}, \lambda)$$

Constraints can be liberating

- ϵ has **straightforward semantics**: the maximum proportion of active gates. Application specific requirements on sparsity can be incorporated into it.
- Tuning the penalization λ to get a satisfactory model may require running several experiments. Even harder when introducing various sources of regularization.
- It is **transparent** whether a model is respecting the sparsity constraints.

Training dynamics



Predictive performance

Architecture	Approach	Pruned	Best error (%)
MLP 784-300-100	**Penalized**†: $\lambda = 0.1/N$	219-214-100	1.4
	Constrained: $\epsilon = 33\%$	198-233-100	1.4
LeNet 20-50-800-500	**Penalized**†: $\lambda = 0.1/N$	20-25-45-462	0.9
	Constrained: $\epsilon = 10\%$	20-21-34-407	0.53

†C. Louizos, M. Welling, and D. P. Kingma. Learning Sparse Neural Networks through L0 Regularization. *ICLR*, 2018.

The constrained approach is ***not universally*** *better than the penalized approach!*

It **can** provide more flexibility and interpretability without compromising predictive performance.