Learning Sparse Neural Networks via Sensitivity-Driven Regularization

Enzo Tartaglione[1], Skjalg Lepsøy[2], Attilio Fiandrotti[1,3] and Gianluca Francini[4]

[1]Politecnico di Torino, [2]Nuance Communications, [3]Télécom ParisTech, [4]Telecom Italia

Complexity in neural networks

- More and more complex problems to be solved.
- Larger and larger ANNs used to solve these problems.
- More parameters, more operations → higher complexity → difficult execution on mobile/embedded devices.

Architecture	# params
VGG-16	144M
ResNet-101	44M
Inception	6.8M
Se-ResNet51	28M

State-of-the-art solutions

- Accept lower test performance and use smaller ANNs.
- Modify the connectivity → more complex operations.
- Use shared parameters → still high number of operations.
- Remove all the parameters considered not useful:
  - Custom filter pruning for CNN.
  - Use of the ℓ_1 regularizer.
  - Design of complex ℓ_0-like regularizers.

What our method provides

- Great test set performance.
- Pruning parameters not contributing in the generation of the correct classification.
- Minimum computational overhead.
- Generality:
  - **Any ANN architecture** for classification tasks.
  - **Any activation** function.
  - Trained with **any loss** function.

Sensitivity of the output to a parameter

$$\mathrm{S}(\boldsymbol{y}, w_i) = \sum_k \alpha_k \left|\frac{\partial y_k}{\partial w_i}\right|$$

- Positive quantity.
- Estimates the importance of a parameter for the generation of a result.
- Uses the Jacobians from backpropagation.
- Low S parameters can be “pushed” to zero.

	Value	Importance
S	↗	↗
	↘	↘
|w|	↗	?
	↘	?

Learning rule

$$w_i^t := w_i^{t-1} - \eta \frac{\partial \mathrm{L}}{\partial w_i^{t-1}} - \lambda w_i^{t-1} \max\left[0, 1 - \mathrm{S}(\boldsymbol{y}, w_i^{t-1})\right]$$

- Update term pushes parameters with low S towards zero.
- If we want the R function to be minimized, we integrate over w.
- Assuming ReLU activations, its form is very simple, but in general it is more complex.

$$\mathrm{R}_{\mathrm{ReLU}}(w_i) = \frac{w_i^2}{2} \max[0, 1 - \mathrm{S}(\boldsymbol{y}, w_i)]$$

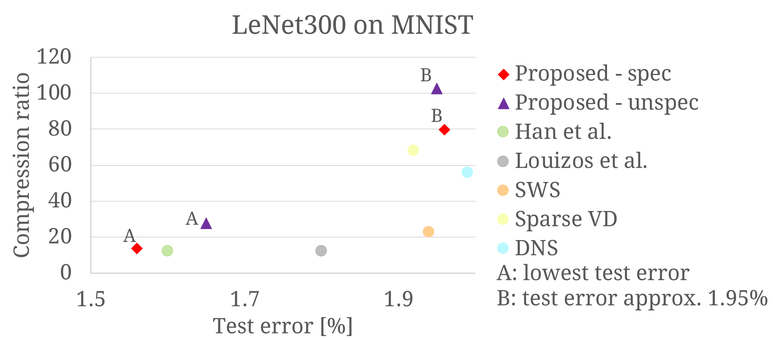
$$\mathrm{R}(w_i) = \mathrm{H}[1 - \mathrm{S}(\boldsymbol{y}, w_i)] \frac{w_i^2}{2} \left[1 - \sum_k \alpha_k \mathrm{sign}\left(\frac{\partial y_k}{\partial w_i}\right) \sum_{m=1}^{\infty} (-1)^{m+1} \frac{w_i^{m-1}}{(m+1)!} \frac{\partial^m y_k}{\partial w_i^m}\right]$$

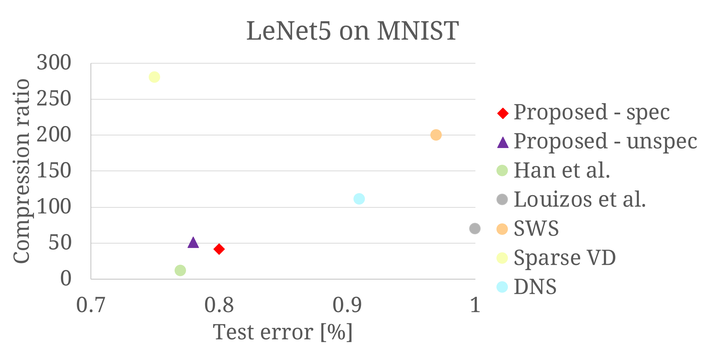
Unspecific vs specific sensitivity

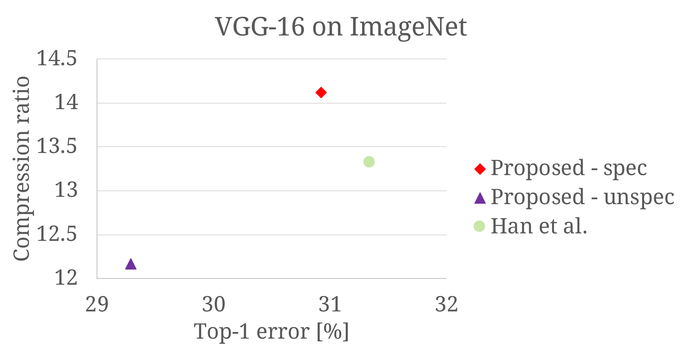
$$\mathrm{S}^{\mathrm{unspec}}(\boldsymbol{y}, w_i) = \frac{1}{C} \sum_{k=1}^{C} \left|\frac{\partial y_k}{\partial w_i}\right|$$

$$\mathrm{S}^{\mathrm{spec}}(\boldsymbol{y}, \boldsymbol{y}^*, w_i) = \sum_{k=1}^{C} y_k^* \left|\frac{\partial y_k}{\partial w_i}\right|$$

Results







POLITECNICO DI TORINO

NUANCE

TELECOM ParisTech

TIM