Dynamic Pruning of a Neural Network via Gradient Signal-to-Noise Ratio

Julien Siems Aaron Klein Cedric Archambeau Maren Mahsereci

Contributions:

1. We extend prior work (Lin et al. (2020)) on dynamic pruning during training and lift its limitation of a hand-designed sparsity schedule.
2. We only consider a weight for pruning if it's no longer receiving reliable gradients. To measure this point, we propose to use the Mini-Batch Gradient Signal-To-Noise Ratio.
3. We benchmark our approach on different datasets and archs.

Background:

1. Mini-Batch Gradient Signal-to-Noise Ratio (Mahsereci et al. (2017))

- Measures reliability of a mini-batch gradient with respect to noise in mini-batch gradient.

$$snr(\theta_k) := \frac{\nabla L^2_{B,k}(\theta)}{\Sigma_{B,kk}(\theta)} = \begin{cases} > 1 & \text{Reliable gradient} \\ \leq 1 & \text{Noisy gradient} \end{cases}$$

2. Dynamic Pruning with Feedback (Lin et al. (2020)):

- Increasingly prunes weights based on magnitude during training.
- Gradients computed also for pruned weights:
    - Pruned weights can become unpruned.

Algorithm:

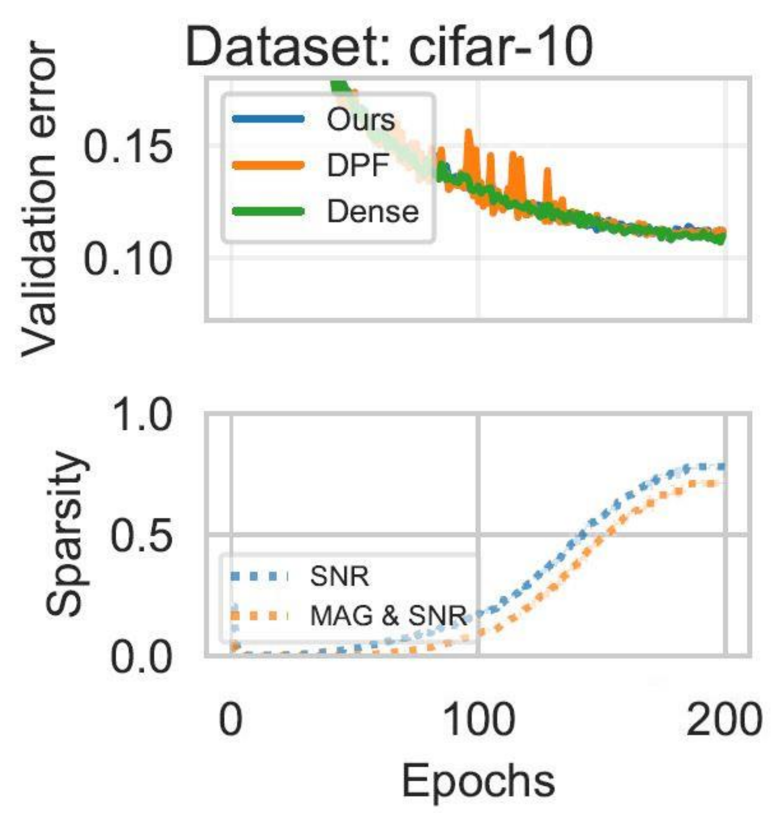
Algorithm 1 *The detailed training procedure of pruning guided by gradient SNR.*

Require: uncompressed model weights $\boldsymbol{\theta} \in \mathbb{R}^d$, pruned weights: $\hat{\boldsymbol{\theta}}$, mask: $\mathbf{m}_{prune} \in \{0,1\}^d$; SNR exp. avg.: γ, burn-in steps: bin_{steps}, mask: $\mathbf{m} \in \{0,1\}^d$; training iterations: T.

for $t = 0, \dots, T$ **do**
  if $t > bin_{steps}$ **then** ▷ trigger mask update, by default after $bin_{steps} = 1$ epoch
    compute SNR mask $\mathbf{m}_{snr} \leftarrow \{\mathrm{snr}(\theta^i_t) > 1 \mid \mathrm{i\ in}\ |\boldsymbol{\theta}_t|\}$ ▷ let sp_{snr} be the sparsity of the resulting mask
    compute MAG mask $\mathbf{m}_{mag} \leftarrow \{|\theta^i_t| > \gamma \mid \mathrm{i\ in}\ |\boldsymbol{\theta}_t|\}$ ▷ γ cut-off weight mag. acc to sp_{snr}
    compute PRUNE mask $\mathbf{m}_{prune} \leftarrow \mathbf{m}_{mag} \wedge \mathbf{m}_{snr}$ ▷ only prune, if $\mathbf{m}_{mag}$ and $\mathbf{m}_{snr}$ agree to prune
  end if
  $\hat{\boldsymbol{\theta}}_t \leftarrow \mathbf{m}_{prune} \odot \boldsymbol{\theta}_t$ ▷ apply resulting mask
  compute (mini-batch) gradient $\nabla L_{\mathcal{B}}(\hat{\boldsymbol{\theta}})$ ▷ forward/backward pass with pruned weights $\hat{\boldsymbol{\theta}}_t$
  update $\mathbf{m}_{snr,t}$ ▷ update SNR exp mov. avg. per weight
  $\boldsymbol{\theta}_{t+1} \leftarrow$ gradient update $\nabla L_{\mathcal{B}}(\hat{\boldsymbol{\theta}})$ to $\boldsymbol{\theta}_t$ ▷ via arbitrary optimizer (e.g. SGD with momentum)
end for

Ensure: $\boldsymbol{\theta}_T$ and $\hat{\boldsymbol{\theta}}_T$

Experiments:



- Sparsity depends on dataset and NN.
- Performs on par with DPF and normal dense training, but with less hyperparameters than DPF.
- Pruning hardly influences validation error (unnecessary weights removed)

References:

Mahsereci, Maren, et al. "Early stopping without a validation set." arXiv preprint arXiv:1703.09580 (2017).
Tao Lin, et al. "Dynamic Model Pruning with Feedback." International Conference on Learning Representations.