The State of Sparsity in Deep Neural Networks[1]

May 6, 2020

[1] "The State of Sparsity in Deep Neural Networks" by Gale, Elsen, and Hooker

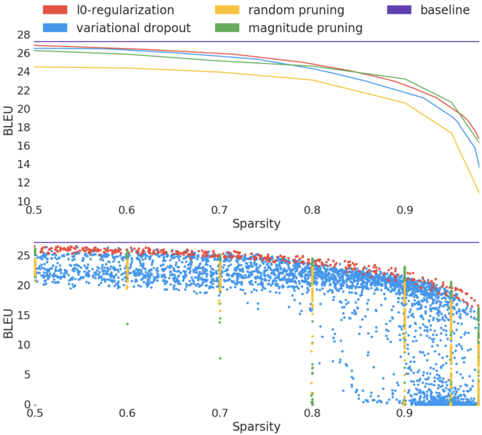
Overview

- Authors benchmark different pruning techniques on different large-scale tasks
- Their benchmarks demonstrate that “state-of-the-art” techniques often lose to simple baselines
- They failed to reproduce lottery ticket hypothesis in a large-scale scenario

Sparsity techniques

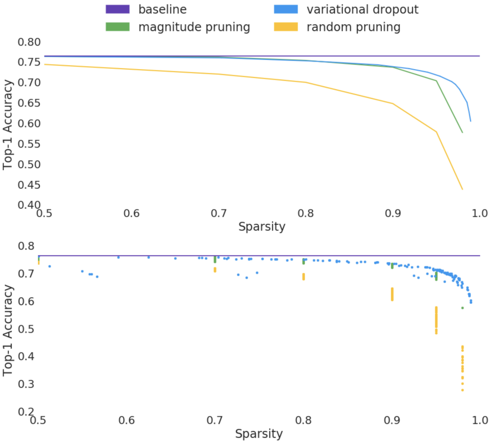
- *Magnitude pruning*: gradually mask out more and more weights during training based on their magnitudes.
- *Variational dropout*: train a bayesian model with gaussian posterior and remove weights with large variance.
- l_0 *regularization*: train an importance weight for each weight and push the importances towards zero.
- *Random pruning baseline*: gradually prune more and more weights during training on random

BLEU/sparsity tradeoff for Transformer on WMT'14 en-de



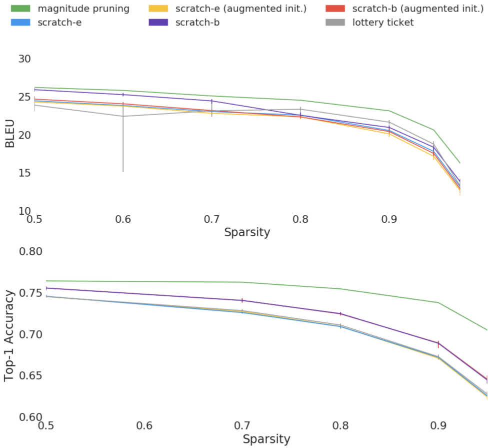
- Magnitude pruning prunes layers uniformly regardless of their type, vardrop and l0-reg prunes some layer types less aggressively
- Magnitude pruning outperforms vardrop and l0-reg in high-sparsity scenarios

Accuracy/sparsity tradeoff for ResNet-50 on ImageNet



- Authors couldn't manage to make l0-reg work for this setup
- Vardrop worked really well (but consumes much more memory)
- Authors tried not to prune the first layer for magnitude pruning and it outperformed vardrop everywhere except for extreme sparsification values

Testing LTH



- scratch-e: apply the learned mask, reinit the subnetwork and retrain
- scratch-b: increase the number of training steps up to 2x
- lottery ticket: like scratch-e, but use original init
- augmented init: scale the variance at init by the number of non-zeros

Conclusion

- Modern SotA pruning techniques do not work that well on large-scale tasks
- LTH outperformed random init only for high-sparsity and only for Transformer
- Different pruning techniques prune layers non-uniformly, i.e. this implies and some layers are more important (for example, the initial layer and the last one)