Sparse Double Descent: Where Network Pruning Aggravates Overfitting

Zheng He[1], Zeke Xie[2,3], Quanzhi Zhu[1], Zengchang Qin[1]

[1]Beihang University, [2]The University of Tokyo, [3]RIKEN Center for AIP

Overview

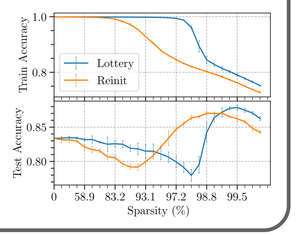
Context. Double descent phenomenon may occur through model sparsification (the *sparse double descent* phenomenon).

Main findings. The *sparse double descent* phenomenon:

- exists consistently across various experimental settings;
- correlates with learning distance rather than minima flatness;
- shows random initialization might surpass the winning tickets.

Random Initialization Might Win

Contrary to *lottery ticket hypothesis*[1], random reinitialized models sometimes could largely surpass the wining ticket models with the same sparsity mask.



Reference

[1] Frankle, J. & Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. ICLR, 2019.

[2] Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. Deep double descent: Where bigger models and more data hurt. ICLR 2020.

[3] Bartoldson, B., Morcos, A. S., Barbu, A., & Erlebacher, G. The generalization-stability tradeoff in neural network pruning. NIPS, 2020.

[4] Nagarajan, V., & Kolter, J. Z. Generalization in deep networks: The role of distance from initialization. arXiv preprint arXiv:1901.01672, 2019.

Contact

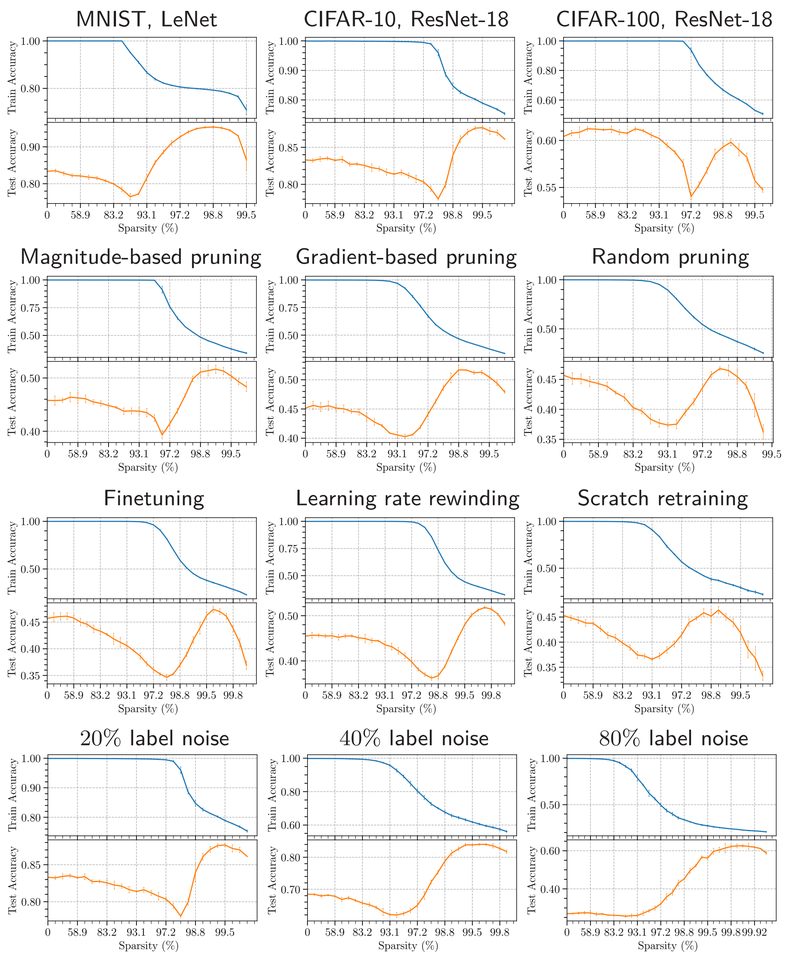
Zheng is actively looking for **PhD opportunities** for 2023. Scan here to see her personal website!

email: zhenghe@buaa.edu.cn

web: https://hezheug.github.io

Sparse Double Descent

The sparse double descent phenomenon exists widely across different datasets, models, pruning strategies, retraining methods and label noise settings.



Consistent with *deep double descent*[2], increasing data complexity shifts the interpolation threshold towards larger capacity, i.e., lower sparsities.

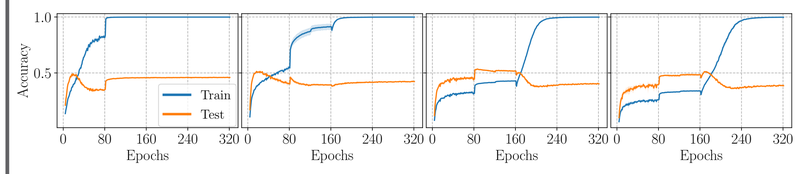
To combat the side effects brought by heavier labels noise, more parameters in the network need to be pruned.

Why Does It Occur?

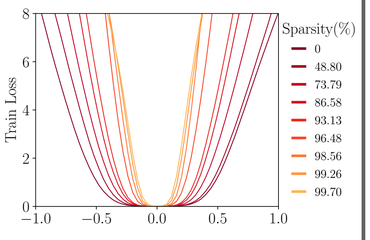
Hypothesis 1: Minima Flatness

Motivation. Previous works[3] hypothesized that pruning could encourage the optimizer to move towards flatter minima that benefit generalization. May such minima flatness hypothesis explain sparse double descent?

Re-dense training. We use re-dense training results as an indirect evidence to estimate minima flatness in the same dimensions. Pruned weights are recovered after 160 epochs.



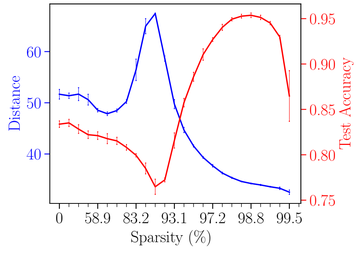
Results. The final solutions of re-dense training do not generalize well. Minima sharpness is compared using 1D loss visualization. Sharpness doesn't coincide with sparse double descent.



Hypothesis 2: Learning Distance

Motivation. Learning distance has been observed to be very related to generalization[4]. We suspect that sparsity may affect l_2 distance from initialization thus affect model capacity.

Results. The changing curve of learning distance correlates with test accuracy. Staying closer to initialization coincides with better robustness, while staying farther from initialization presents an inferior performance.



Conclusion. The l_2 learning distance of models may correlate with the double descent curve and reflects generalization better than minima flatness for sparse models.