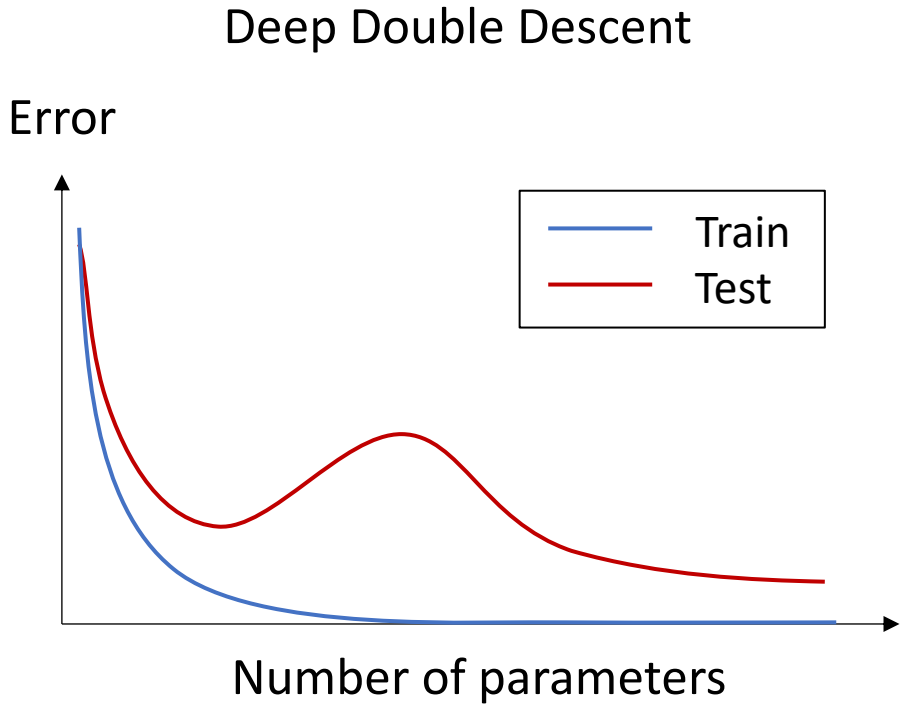
Sparse Double Descent: Where Network Pruning Aggravates Overfitting

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

Motivation

- Deep neural networks are overparameterized
- As the model capacity increases, the double descent phenomenon occurs
- Network pruning could also affect model capacity

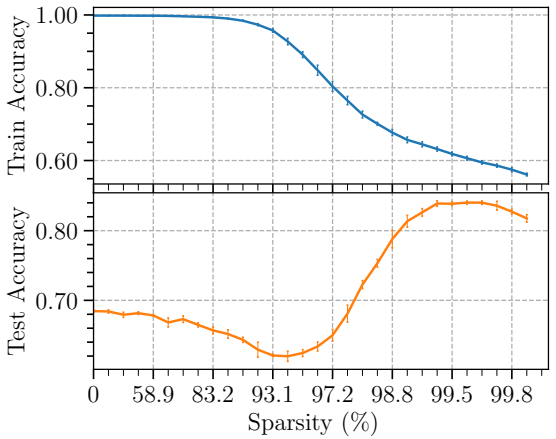


Q: May the sparsification of DNNs also cause double descent?

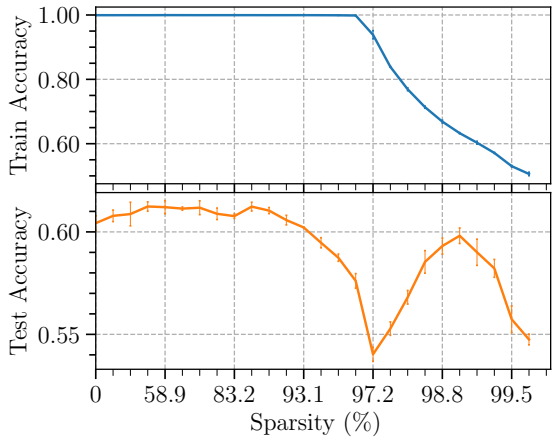
Sparse Double Descent

- Sparse double descent exists consistently across different experimental settings under label noise

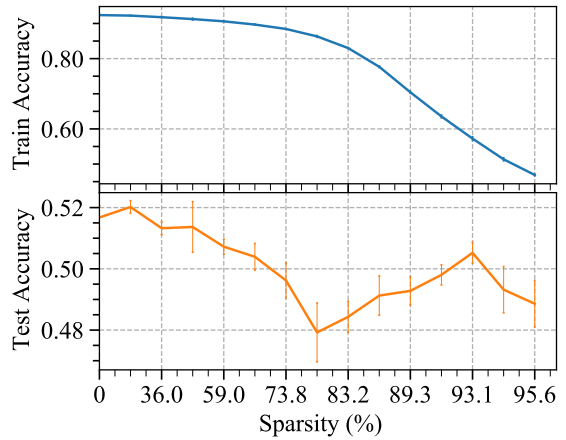
ResNet-18, CIFAR-10



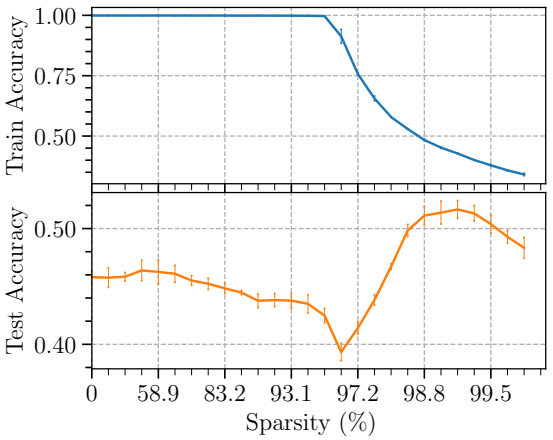
ResNet-18, CIFAR-100



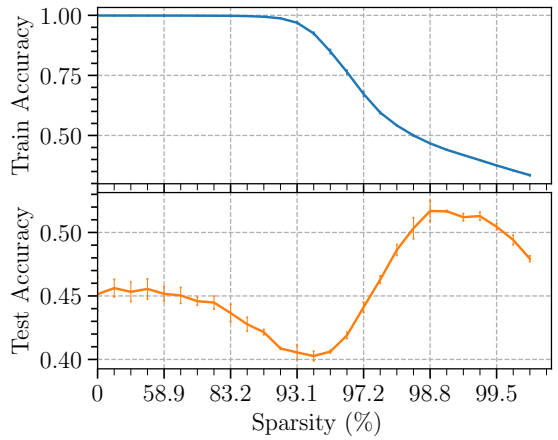
ResNet-101, Tiny ImageNet



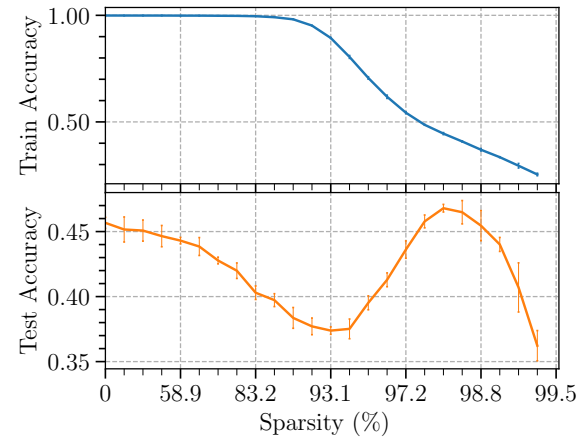
Magnitude-based pruning



Gradient-based pruning



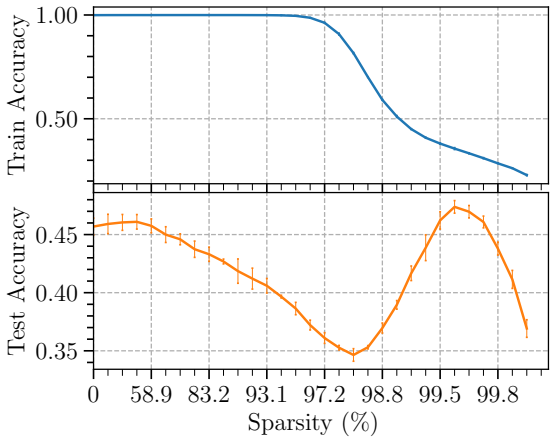
Random pruning



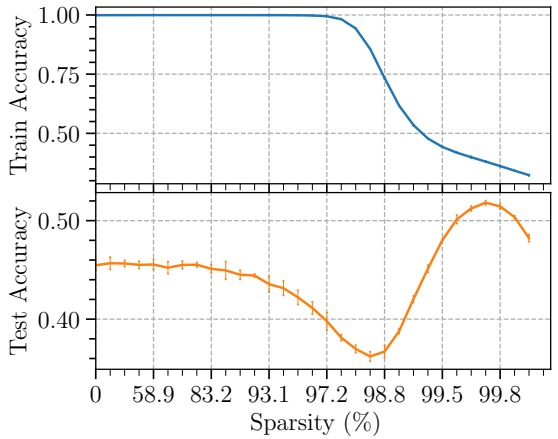
Sparse Double Descent

- Sparse double descent exits consistently across different experimental settings under label noise

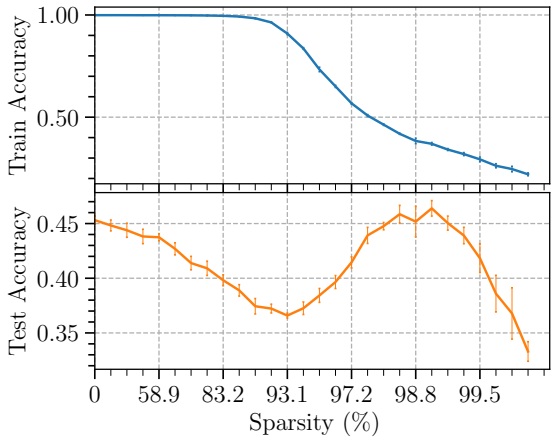
Finetuning



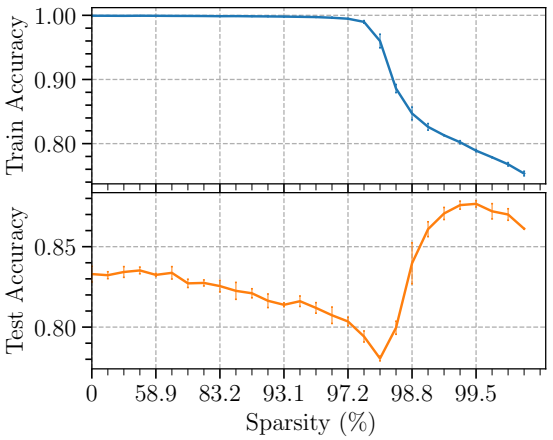
Learning rate rewinding



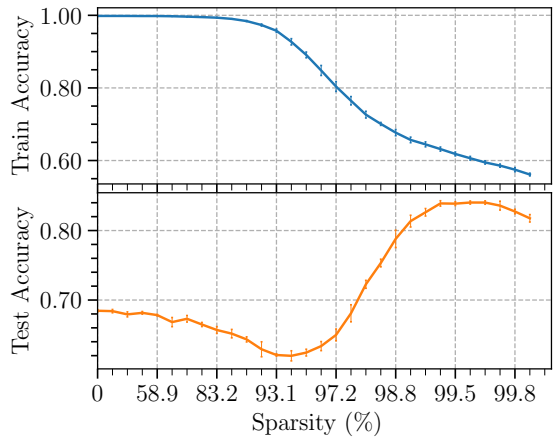
Scratch retraining



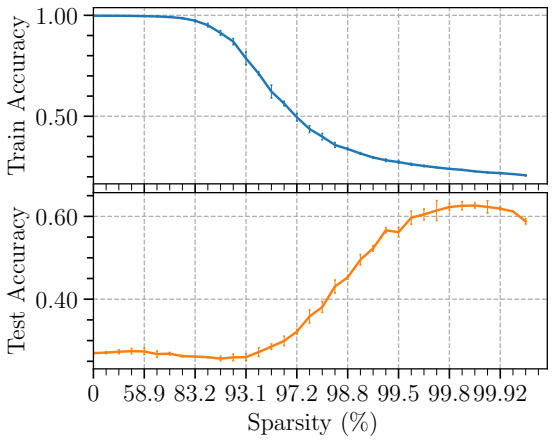
20% symmetric label noise



40% symmetric label noise



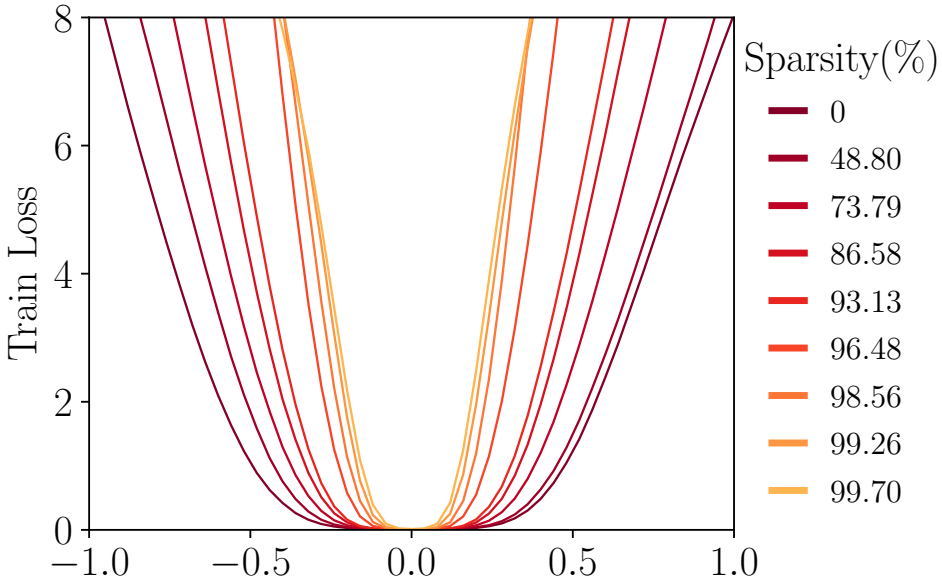
80% symmetric label noise



Why Sparse Double Descent Occurs?

Minima flatness hypothesis

- Previous works hypothesized that pruning encourages the optimizer to move towards flatter minima [Bartoldson et al., 2020]
- Minima flatness is usually correlated with good generalization
- We observed optimizer may not converge to flat regions as sparsity increases

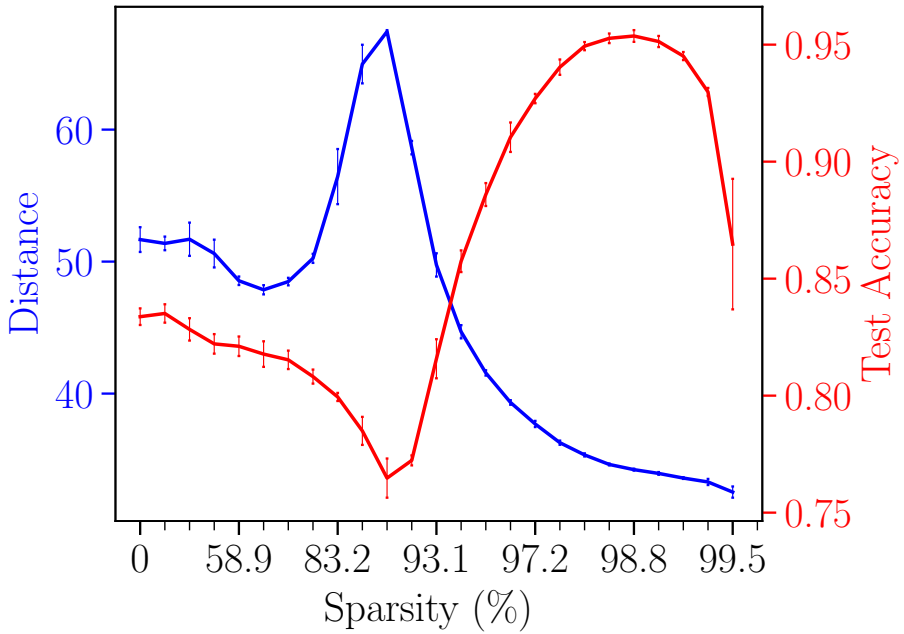


Bartoldson, B., Morcos, A. S., Barbu, A., and Erlebacher, G. The generalization-stability tradeoff in neural network pruning. NIPS, 2020.

Why Sparse Double Descent Occurs?

Learning Distance hypothesis

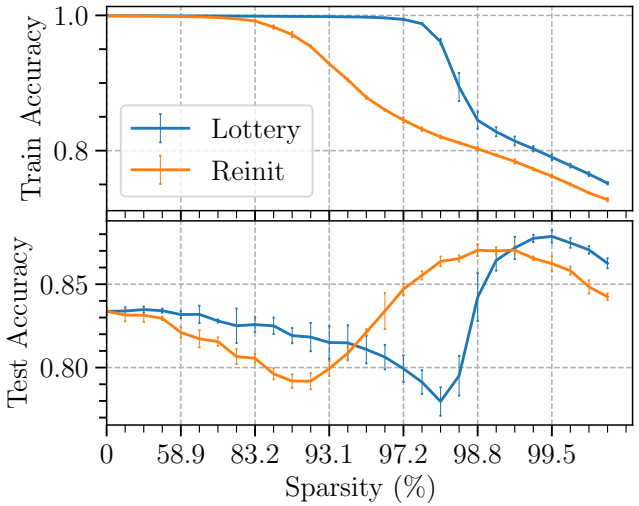
- model capacity could be restricted by the l2 learning distance from initialization [Nagarajan & Kolter, 2019]
- Pruning may affect the learning distance
- We observed the curve of learning distance correlates with test accuracy

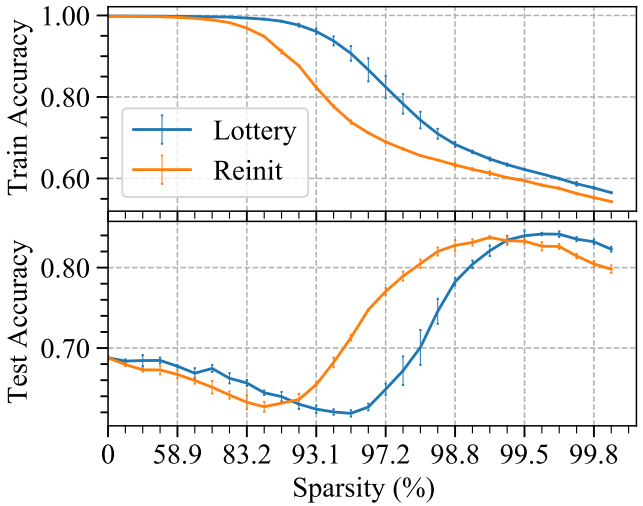


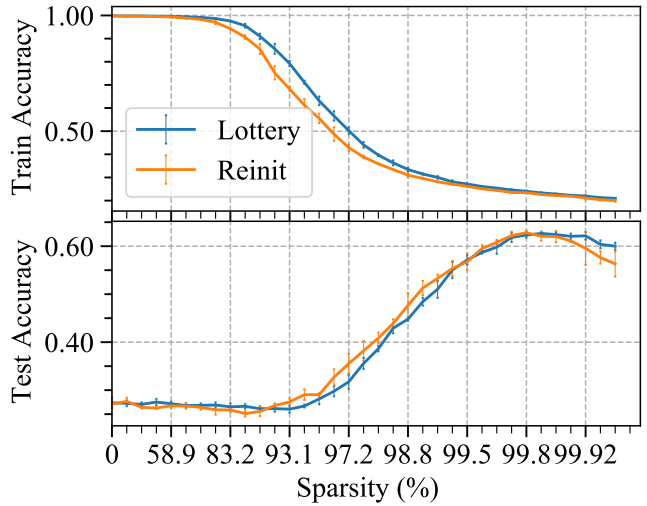
Nagarajan, V. and Kolter, J. Z. Generalization in deep networks: The role of distance from initialization. *arXiv preprint arXiv:1901.01672*, 2019.

Winning tickets may not always win

- Random reinitializations sometimes surpass the wining ticket initializations in the Lottery Ticket Hypothesis [Frankle & Carbin, 2019]







Thanks for your attention!

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. ICLR, 2019.