

## A APPENDIX

### A.1 ADDITIONAL TRAINING DETAILS

All unstructured pruning experiments used data augmentation (RandomResizedCrop and RandomHorizontalFlip from torchvision), batch size of 512, and weight decay of 0.0001. We pruned convolutional layers 20% at each pruning iteration, for a total of 30 iterations.

ResNet-50 on Tiny ImageNet: we trained 110 epochs, decaying LR by gamma = 0.2 at epochs 70, 85, and 100. Experiments with warmup, late rewinding, and learning rate rewinding (Section 3.2) use the same hyperparameters, except for modifications explicitly mentioned in the main text.

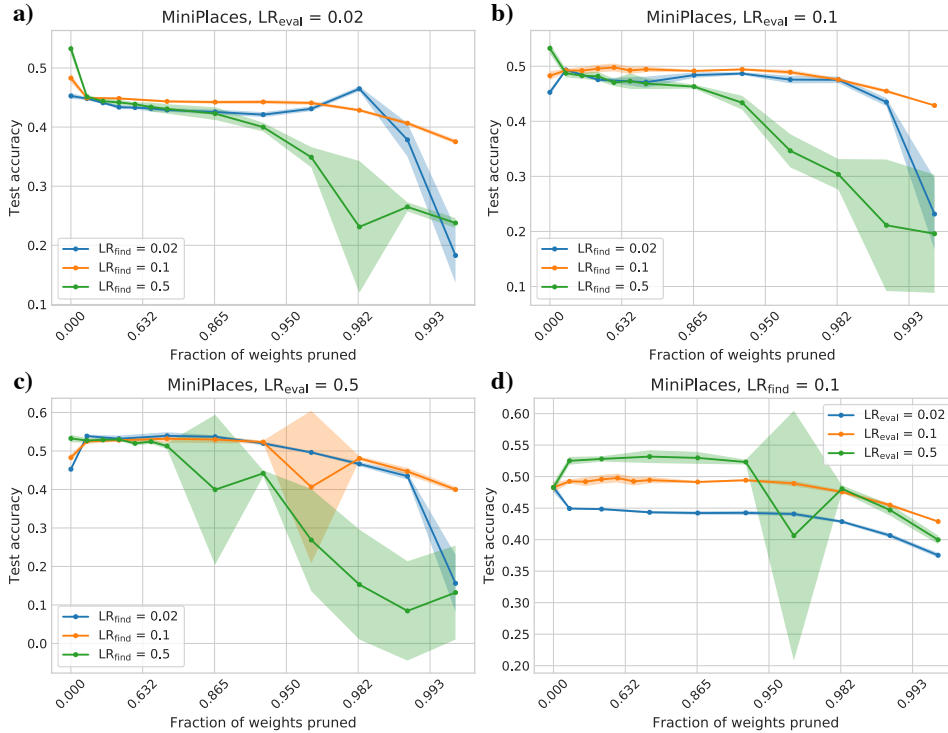
ResNet-50 on MiniPlaces: we trained 120 epochs, decaying LR by gamma = 0.2 at epochs 80, 110, and 115.

ResNet-18 on Tiny ImageNet: we trained 110 epochs, decaying LR by gamma = 0.1 at epochs 80 and 100.

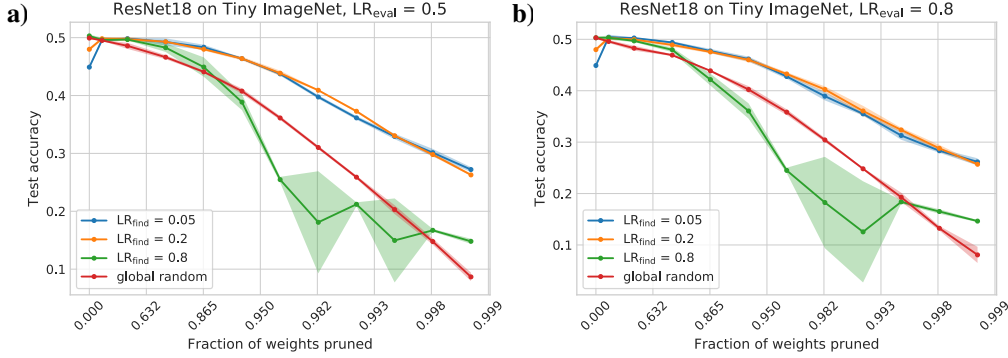
Structured pruning: we use the magnitude of  $\gamma$  of batch normalization layers as the ranking metric to decide which filters to prune. For channels that are connected by residual connections, we sum their measures as they have to be pruned jointly. We train the pruned network using 200 epochs for both pre-training and fine-tuning. We use a batch size of 256, weight decay of  $5e-5$ , cosine learning rate decay, linear learning rate warmup from 0 to the set learning rate within the first 5 epochs, SGD with 0.9 Nesterov momentum, and 0.1 label smoothing.

Shaded error bars in all figures represent standard deviations over 5 runs with different random seeds.

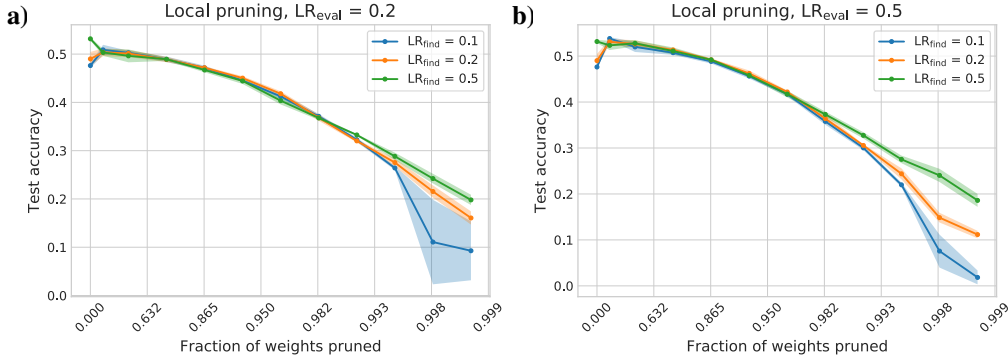
### A.2 ADDITIONAL FIGURES REFERENCED FROM SECTION 3



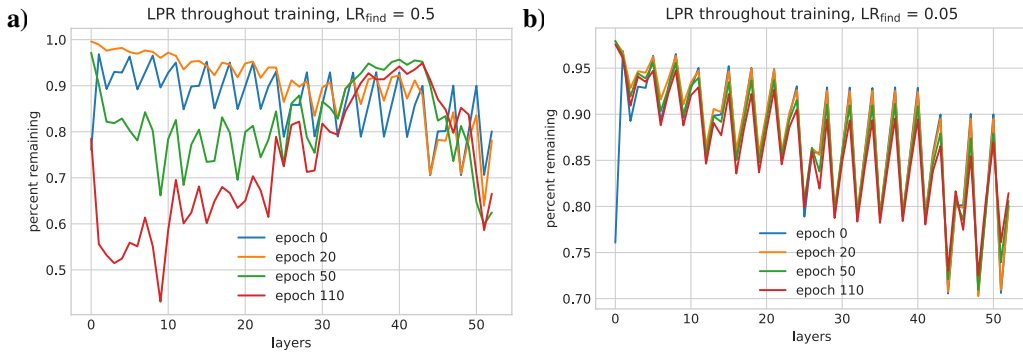
**Figure A1:** ResNet-50 on MiniPlaces with  $LR_{find}$  and  $LR_{eval}$  from  $\{0.02, 0.1, 0.5\}$ . In (a), (b), and (c), we use a constant  $LR_{eval}$  of 0.02, 0.1, and 0.5 respectively, with each plot showing all three values of  $LR_{find}$ . They show that  $LR_{eval} = 0.1$  is the best. Thus, in (d) we compare the different values of  $LR_{eval}$ , all with  $LR_{find} = 0.1$ , to demonstrate that the best  $LR_{eval}$  is 0.5, until very high sparsity levels where 0.1 becomes better. In line with other results, the best LR for standard training (0.5) is not the best  $LR_{find}$  for pruning.



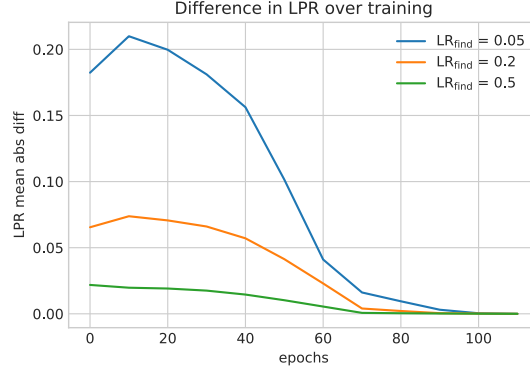
**Figure A2:** ResNet-18 on Tiny ImageNet with  $LR_{find} \in \{0.05, 0.2, 0.8\}$ , compared using a constant  $LR_{eval}$  of 0.5 in (a) and 0.8 in (b). For both values of  $LR_{eval}$ ,  $LR_{find}$  of 0.05 and 0.2 are very similar and both perform significantly better than  $LR_{find} = 0.8$ , despite 0.8 performing the best on the unpruned model.  $LR_{find}$  of 0.8 also does worse than the random baseline (red).



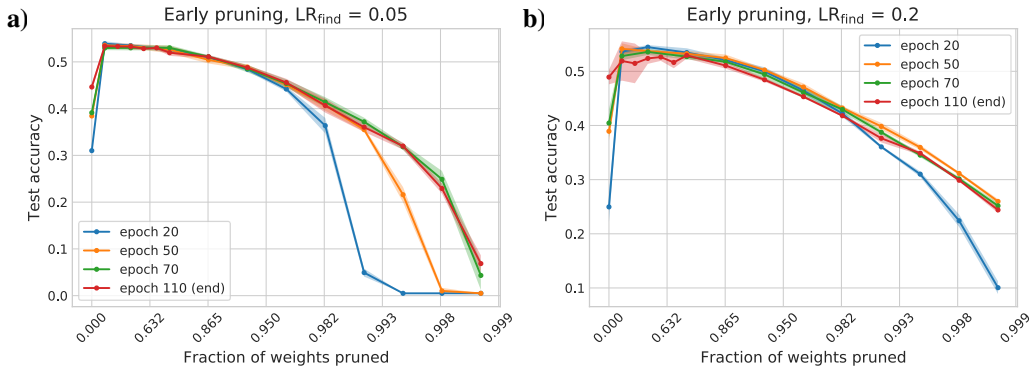
**Figure A3:** Local pruning: pruning all layers the same amount. We compare  $LR_{find} \in \{0.1, 0.2, 0.5\}$  using  $LR_{eval}$  of 0.2 (a) and 0.5 (b). For both values of  $LR_{eval}$ , we see that the larger the  $LR_{find}$ , the better the pruned models perform.



**Figure A4:** Layerwise pruning ratios if you were to prune at different epochs, for  $LR_{find} = 0.5$  (a) and 0.05 (b), at the first pruning iteration (starting from an unpruned model). In (a), the LPR change significantly throughout training, whereas for (b), LPR changes very slightly except for the first layer.



**Figure A5:** Distance between LPR if pruned at epoch  $n$  vs. actual LPR when pruned at the end, for different  $LR_{find}$ . Distance is calculated as the mean absolute difference in pruning ratio for each layer. This is done at the first pruning iteration (starting from an unpruned model). Smaller  $LR_{find}$  values start with much smaller distances; all  $LR_{find}$  values start to plateau around epoch 70, which is when the learning rate decays.



**Figure A6:** Pruning early for  $LR_{find} = 0.05$  and  $0.2$ . Pruning at epoch 20 (blue) is bad, but pruning at epochs 50 (orange) or 70 (green) is actually better than pruning at convergence (110 epochs, red) for  $LR_{find} = 0.2$ . Pruning at 70 for  $LR_{find} = 0.05$  also performs well. All masks are evaluated with  $LR_{eval} = 0.5$ , except for the points at 0% sparsity, which represent the performance of an unpruned model trained only for the corresponding number of epochs. These unpruned accuracies shows how the early pruning epochs are clearly far from convergence, yet some of them find masks that are quite good.