–Supplemental Materials–
CONet: Channel Optimization for Convolutional Neural Networks

https://github.com/mahdihosseini/CONet

1. Appendix A: Experiment setup Details

1.1. General Experiment Setup

For our experiments, we used the CIFAR10, CIFAR100 and ImageNet (ILSVRC2012) [3, 2] datasets. The CIFAR10 and CIFAR100 datasets have 50k training images with 10 and 100 classes with 5000 and 500 images per class respectively. The ImageNet dataset has 1.2M training images with 1000 classes.

We search and evaluate models on CIFAR datasets using one Nvidia V100. For architecture transfer to ImageNet, we evaluate smaller models using one Nvidia RTX 2080Ti and larger models using two Nvidia RTX 2080Tis to accommodate model size.

1.2. Experiment setup for channel search

Before the channel search, we set the channel sizes of the model to 32 for each layer. For the DARTS7 and DARTS+7 experiments on CIFAR10, we used the original channel sizes for our initial model. In order to facilitate optimal rank growth, the initial learning rate η is a valuable hyperparameter, therefore for DARTS and DARTS+, we chose 0.1, and for ResNet34, we chose 0.03.

Within the search phase, we used the Adas[1] scheduler [1] with $\beta = 0.8$ for channel search. Adas is an adaptive learning rate scheduler that also uses local indicators to individually adjust the learning rate of each convolutional layer. β is the learning momentum factor used by Adas. We found that setting β to 0.8 led to quick stabilization of rank and mapping condition while maintaining the underlying structure for each convolutional layer.

1.3. Experiment setup: CIFAR Evaluation

We use the stochastic gradient descent optimizer (SGD) with momentum of 0.9 and a step decaying learning rate scheduler (StepLR) with a step size of 25 epochs and step decay of 0.5. For all evaluations, we set the initial learning rate, η, to 0.1. For ResNet34 model evaluation, we used a weight decay of 0.0005. For DARTS/DARTS+ model evaluation, we used a weight decay of 0.0003 and all of the additional hyperparameters outlined in [4] (e.g. auxiliary head/weight, cutout). Every model is evaluated for 250 epochs for at least 3 runs and the averaged results are reported.

[1] https://github.com/mahdihosseini/AdaS

1.4. Experiment setup: ImageNet Evaluation

For DARTS-14 searched on CIFAR100 and DARTS-7 searched on CIFAR10/100, we used the same settings as the evaluation of DARTS/DART+ models on CIFAR except the weight decay is set to 0.00003. For DARTS-14 searched on CIFAR10 with δ_1, δ_2, we additionally use step size of 1 and a decay rate of 0.97 for the StepLR scheduler. Each model is evaluated for 200 epochs for 1 run.

2. Appendix B: Further Experiment Result

We have included the training curves for all model evaluations reported in subsection 5.1. For each evaluation, we plot the training loss and test accuracy per epoch. We categorize the results by dataset. For CIFAR experiments, we report the average train loss and test accuracy across multiple runs. Please see Figure 1 for CIFAR10 plots, Figure 2 for CIFAR100 plots, and Figure 3 for ImageNet plots.

Please note that for the ImageNet experiments with the DARTS14 models searched on CIFAR10, we used a different step size and decay rate for the StepLR scheduler in contrast to our other experiments. Following the training procedure outline in [4], we used a step size of 1 and decay rate of 0.97.

References

[1] Anonymous. Adas: Adaptive scheduling of stochastic gradients. In *Submitted to International Conference on Learning Representations*, 2021. under review. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[3] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[4] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search, 2019. 1

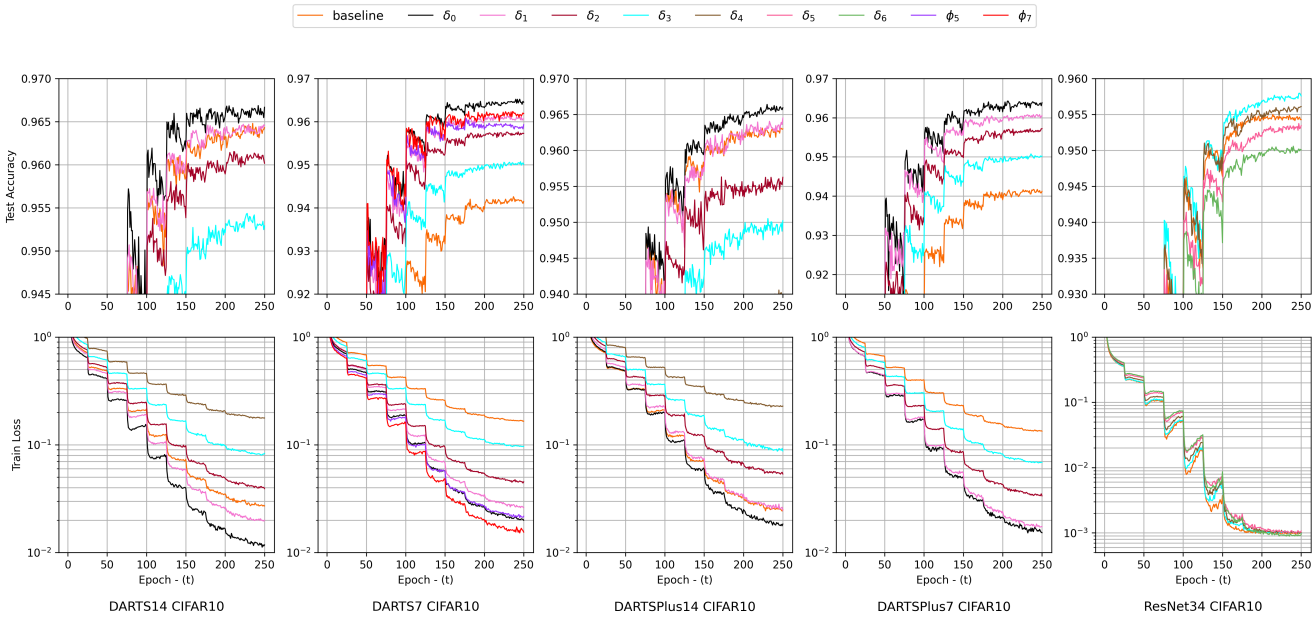

Figure 1. Full training results on CIFAR10 for DARTS-7, DARTS-14, DARTS+-7, DARTS+-14 and ResNet searched on CIFAR10 with different delta thresholds: $\delta_0 = 0.0025$, $\delta_1 = 0.005$, $\delta_2 = 0.0075$, $\delta_3 = 0.01$, $\delta_4 = 0.015$, $\delta_5 = 0.02$, $\delta_6 = 0.025$. Compound Scaling for DARTS7 with different Phi values: $\phi_5 = \sqrt{2}^5$, $\phi_7 = \sqrt{2}^7$.

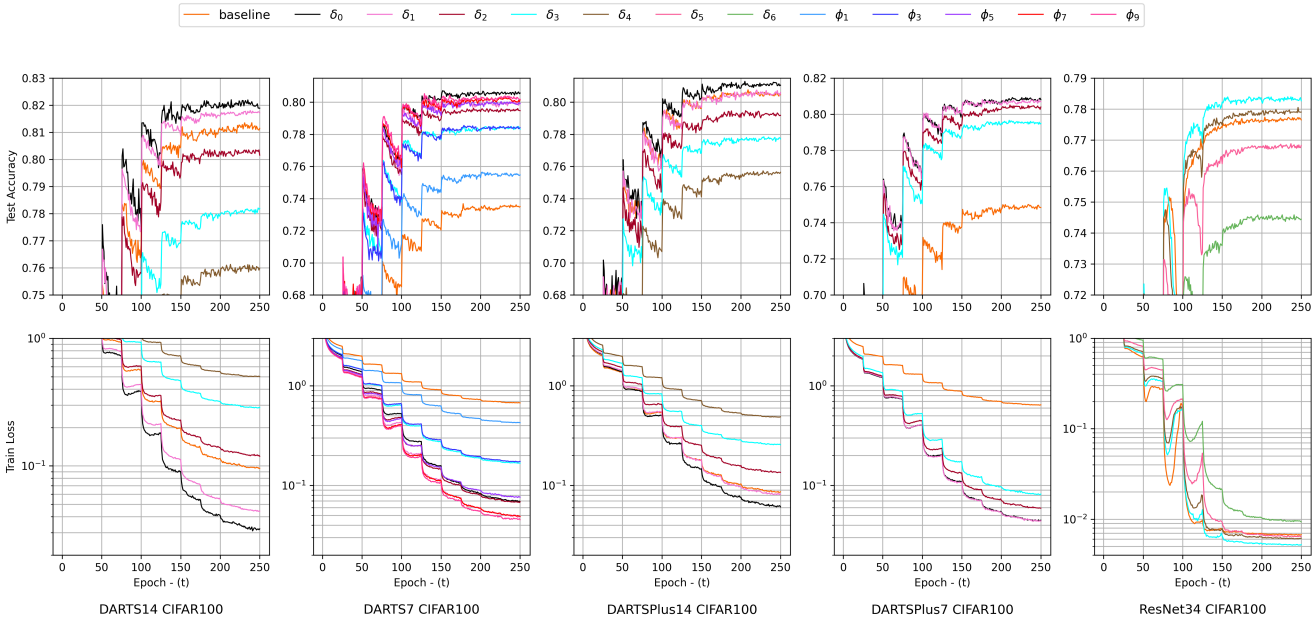

Figure 2. Full training results on CIFAR100 for DARTS-7, DARTS-14, DARTS+-7, DARTS+-14 and ResNet searched on CIFAR100 with different delta thresholds: $\delta_0 = 0.0025$, $\delta_1 = 0.005$, $\delta_2 = 0.0075$, $\delta_3 = 0.01$, $\delta_4 = 0.015$, $\delta_5 = 0.02$, $\delta_6 = 0.025$. Compound Scaling for DARTS7 with different Phi values: $\phi_1 = \sqrt{2}$, $\phi_3 = \sqrt{2}^3$, $\phi_5 = \sqrt{2}^5$, $\phi_7 = \sqrt{2}^7$, and $\phi_9 = \sqrt{2}^9$.

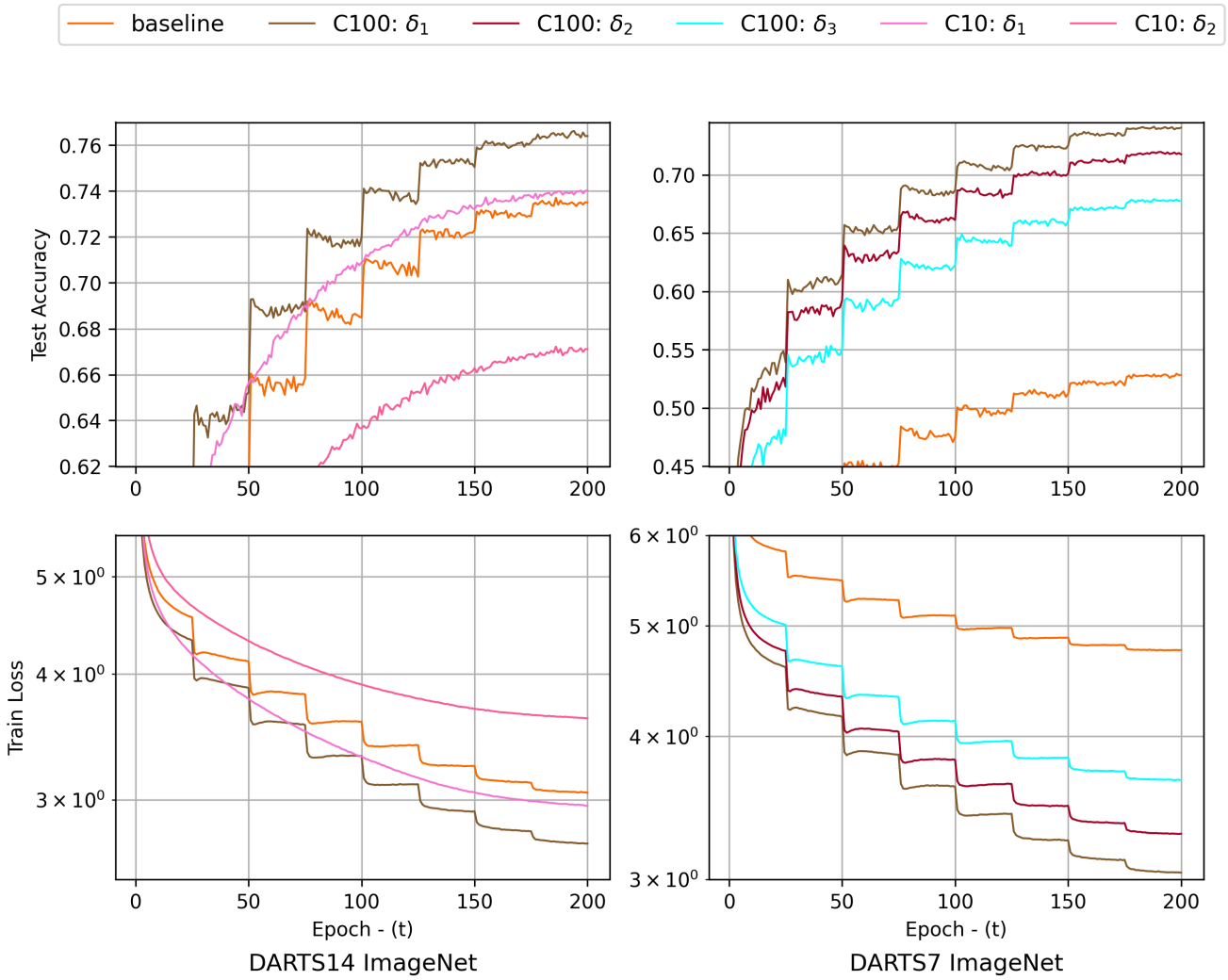

Figure 3. Full training results on ImageNet for DARTS-7 and DARTS-14 searched on CIFAR10/100 with different delta thresholds: $\delta_1 = 0.005$, $\delta_2 = 0.0075$, $\delta_3 = 0.01$, $\delta_4 = 0.015$.