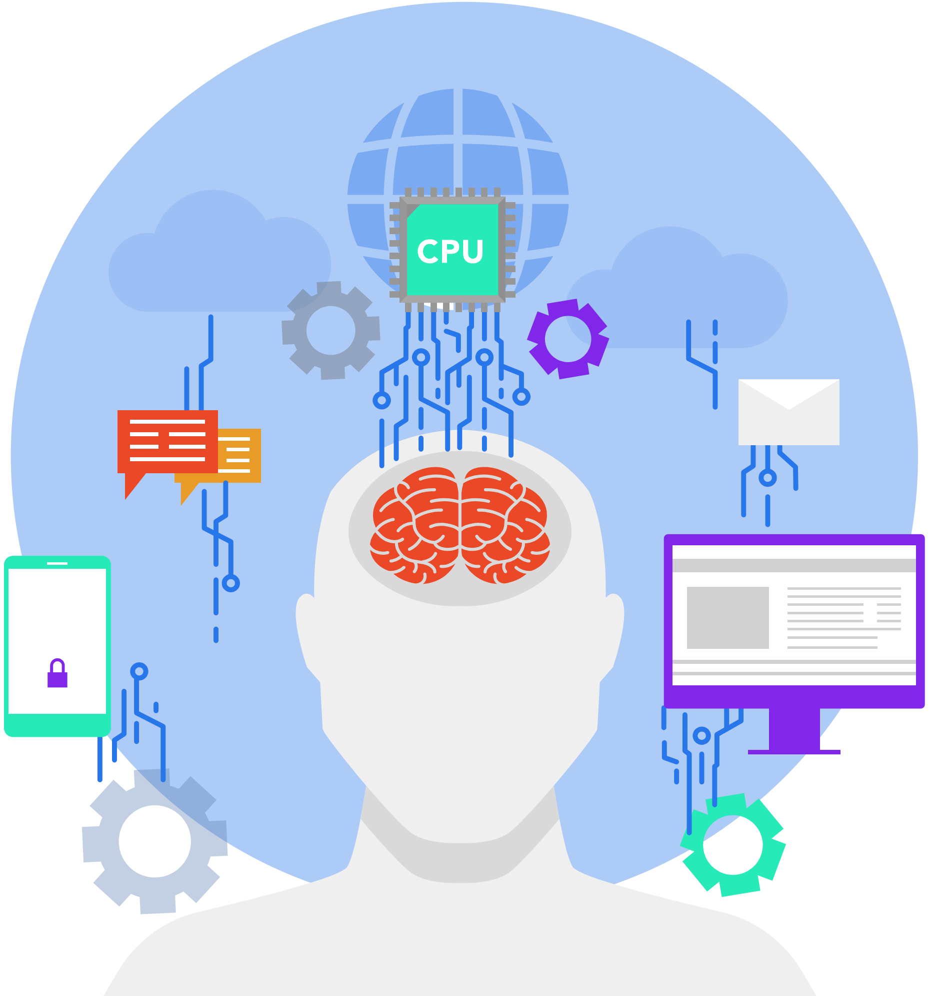
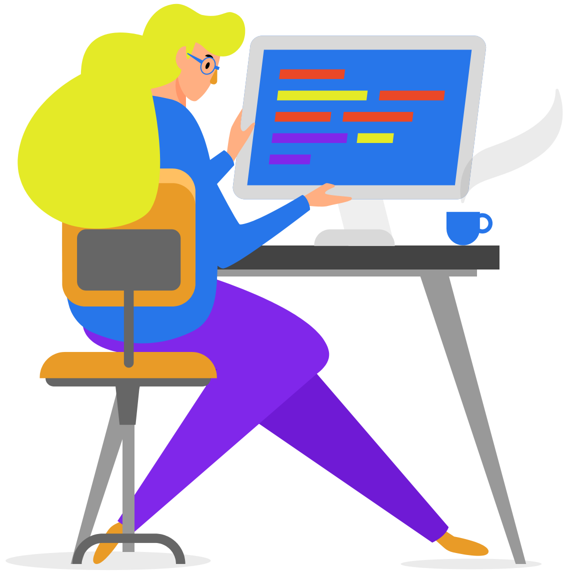


NEURAL NETWORK(s) PRUNING : ONE AND ENSEMBLES

Sumanth Manduru
G01380318

Dileep Kumar Latchireddi
G01389937

AGENDA



01 Motivation & Goals

02 Applications

03 DropNet - Iterative Pruning

04 HQ Ensemble Pruning

05 Experiments & Results

06 Challenges & Future Work

PROBLEM DEFINITION

Motivation

- Reduce the complexity of DNNs and E-DNNs.
- Maintain performance while reducing complexity.

&

Goals

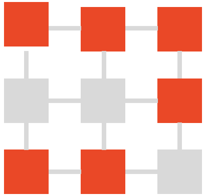
- Find a subnetwork that has fewer parameters with a similar accuracy.
- Find a smaller high quality deep ensembles of size S ($\ll$ M) with higher ensemble accuracy than the entire deep ensemble of all M models.

DATASETS

01

MNIST
70000 Images
10 Handwritten Digits

02



CIFAR10
60,000 Images
10 Classes

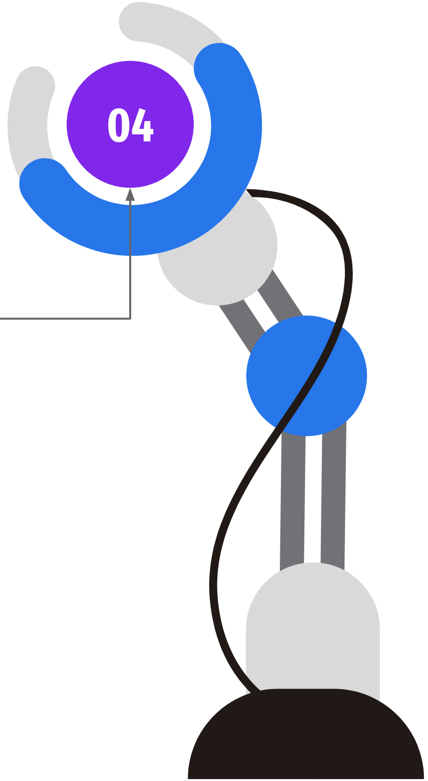
NETWORK PRUNING APPLICATIONS

01 Resource Constrained Devices

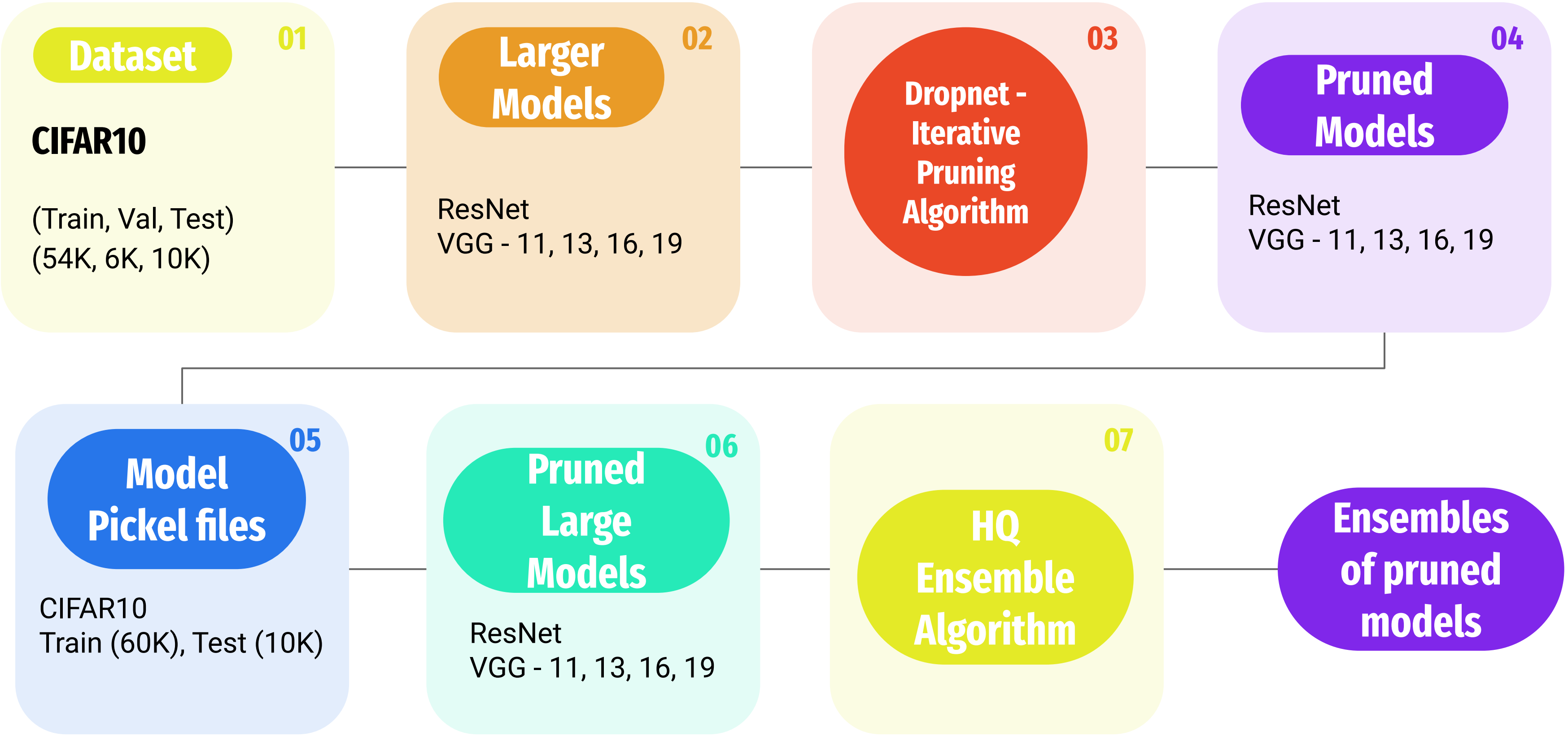
02 Model Compression

03 Interpretation of Complex Models

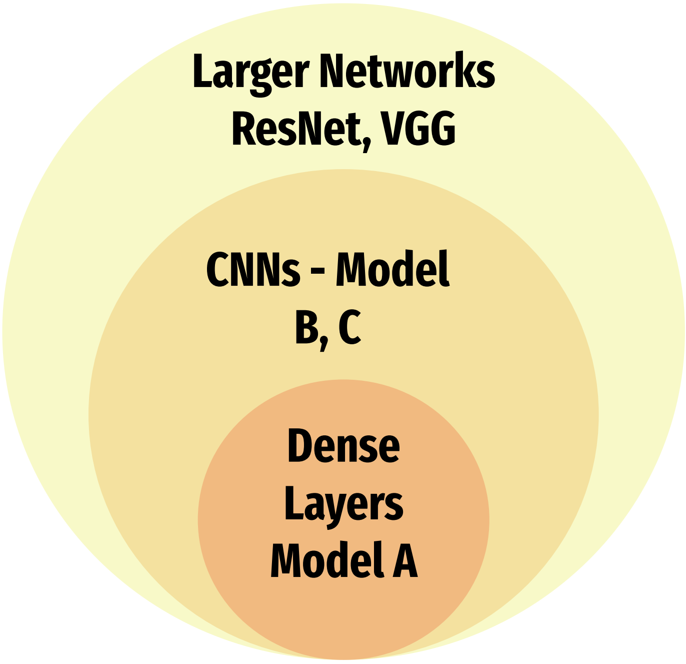
04 Energy Efficiency

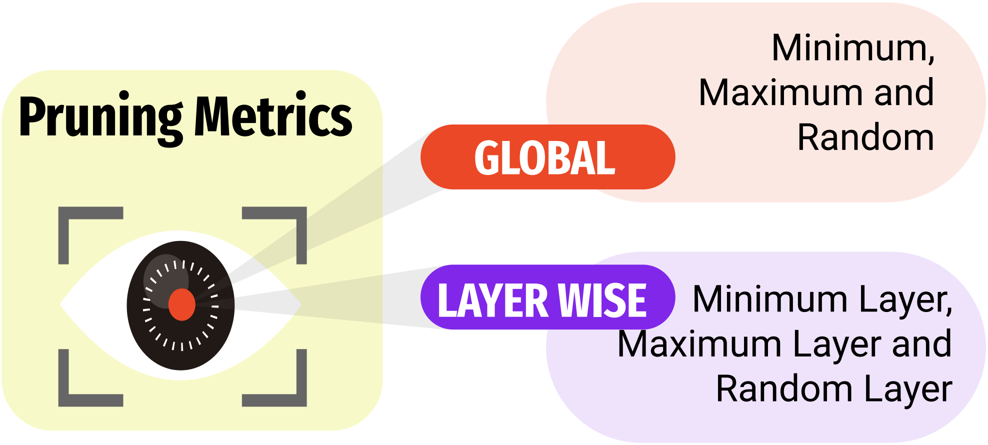


NETWORK PRUNING PIPELINE



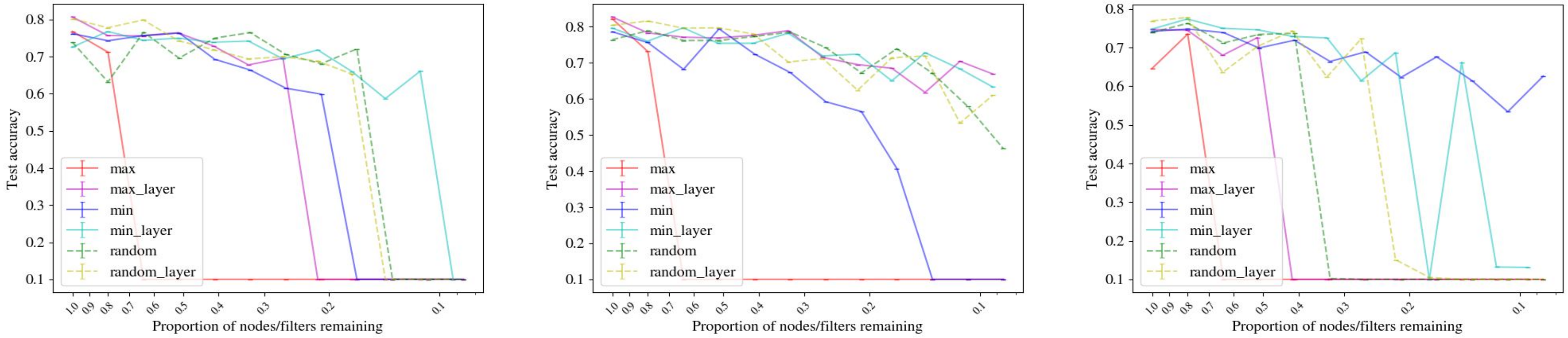
DROPNET - ITERATIVE PRUNING



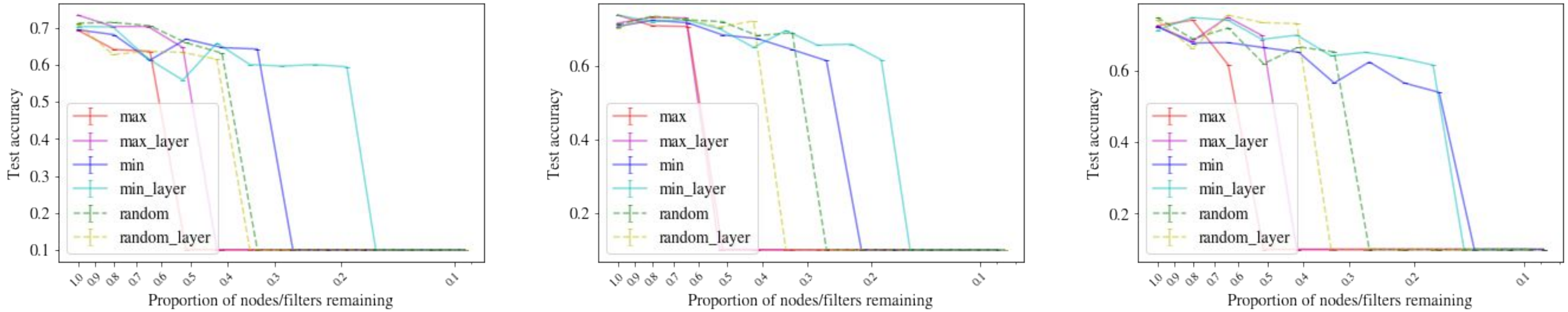


Pruning Metrics	Importance Score
Minimum and Min Layer	E[a_i] or E[f_i]
Maximum and Max Layer	-ve E[a_i] or -ve E[f_i]
Random and Random Layer	0

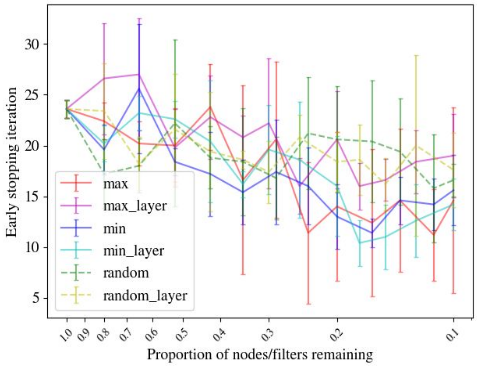
VGG19, VGG16 and ResNet

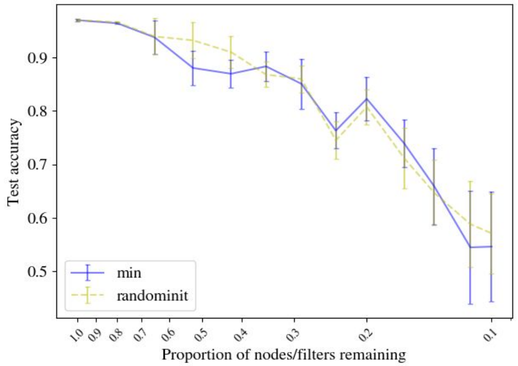


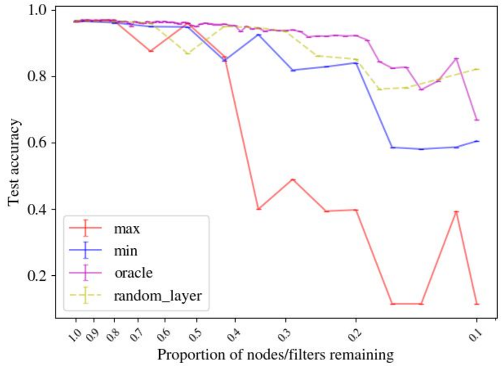
Model C - 64x2- 128x2, 128x2 - 128x2, 128x2 - 256x2



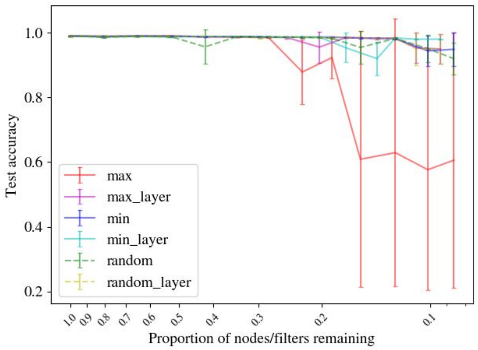
EARLY STOPPING, COMPARE RANDOM AND ORACLE

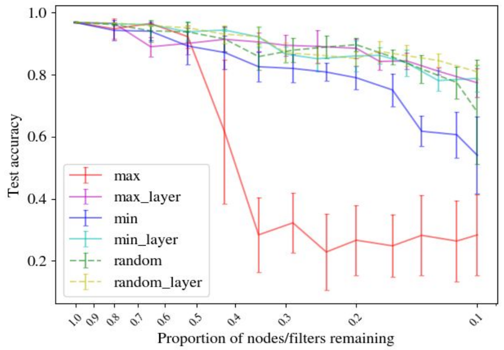


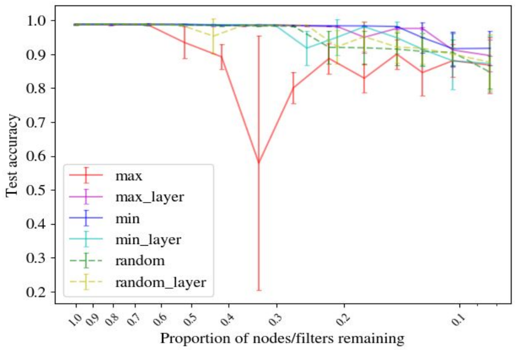




Model B (Conv64-64), Model A (FC40 -FC40), Model B (Conv64-32)







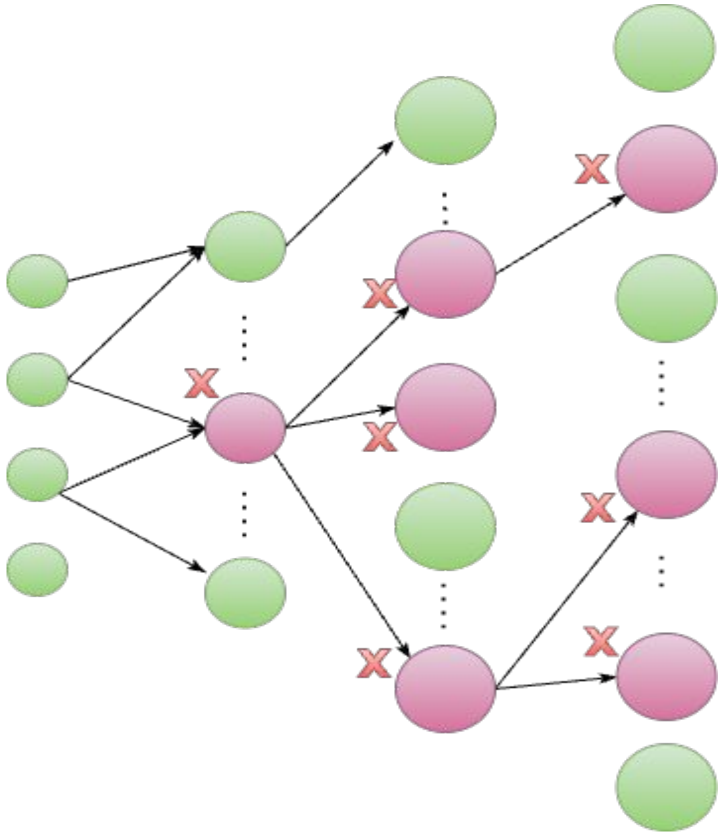
HQ ENSEMBLE PRUNING

Focal Diversity Metrics

- Cohen's Kappa (CK)
- Binary Disagreement(BD)
- Kohavi-Wolpert Variance (KW)
- Generalized Diversity (GD)

β Hyperparameter :

β controls the percentage of ensemble to be pruned



Results

Global Metric	Ensemble	Pruning Percentage	Cost	Accuracy
Minimum	0234 023 02	50% 0% 20%	80% 60% 40%	+2.04 +0.56 - 0.96
Maximum	01 01 124	20% 0% 50%	40% 40% 60%	+2.04 +0.14 -0.4
Random	12 0123 0123	50% 0% 20%	40% 80% 80%	+2.74 +0.7 +0.35

Results

Layerwise Pruning	Ensemble	Pruning Percentage	Cost	Accuracy
Minimum	01 0123 12	0% 20% 50%	40% 80% 40%	+2.39 +0.22 +0.5
Maximum	0123 02 04	0% 20% 50%	80% 40% 40%	+0.14 +2.04 -0.4
Random	0123 012 012	0% 20% 50%	80% 60% 60%	+1.6 +4.23 +1.47

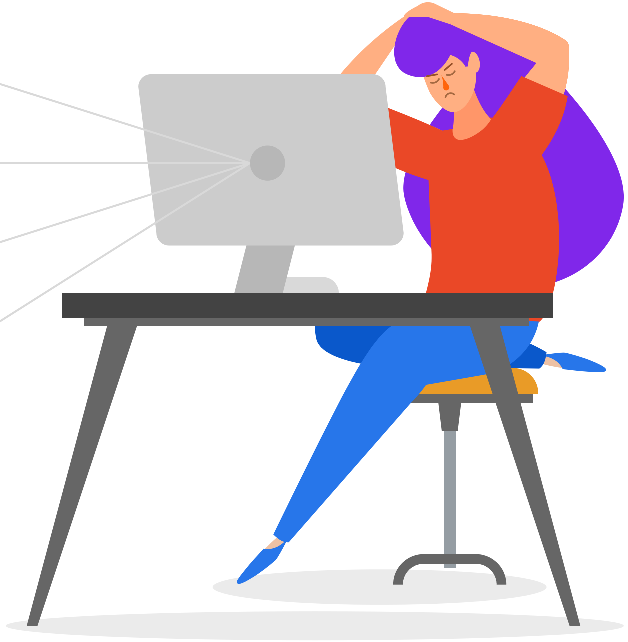
Network Pruning Challenges

Models Training Time 01

Computational Cost 02

Optimal Pruning rate 03

Model Performance, Underfitting 04



Future Work

01

Activation Functions
Similar to ReLU - SoftPlus

02

RNNs and RL Agents

03

Pruning Percentiles

04

Interpretation of Pruned Models

05

Larger pools of Pruned Models

References

[1] Chong Min John Tan and Mehul Motani. 2020. DropNet: Reducing Neural Network Complexity via Iterative Pruning. InProceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119), Hal Daumé III and Aarti Singh (Eds.). PMLR, 9356–9366. https://proceedings.mlr.press/v119/tan20a.html

[2] Yanzhao Wu and Ling Liu. 2021. Boosting Deep Ensemble Performance with Hierarchical Pruning. (Dec. 2021), 1433–1438.