What Happens To Uncertainty In Neural Networks Upon Pruning?

Ksheer Sagar Agrawal, Lipika Rajpal, Kanishk Singhal, Prof. Nipun Batra

IIT Gandhinagar

What is pruning?

Pruning is a technique that aims at reducing the size or complexity of neural networks by reducing the number of parameters. We have explored three techniques of pruning, viz., iterative pruning, one-shot pruning and pruning with and without re-initialization of weights.

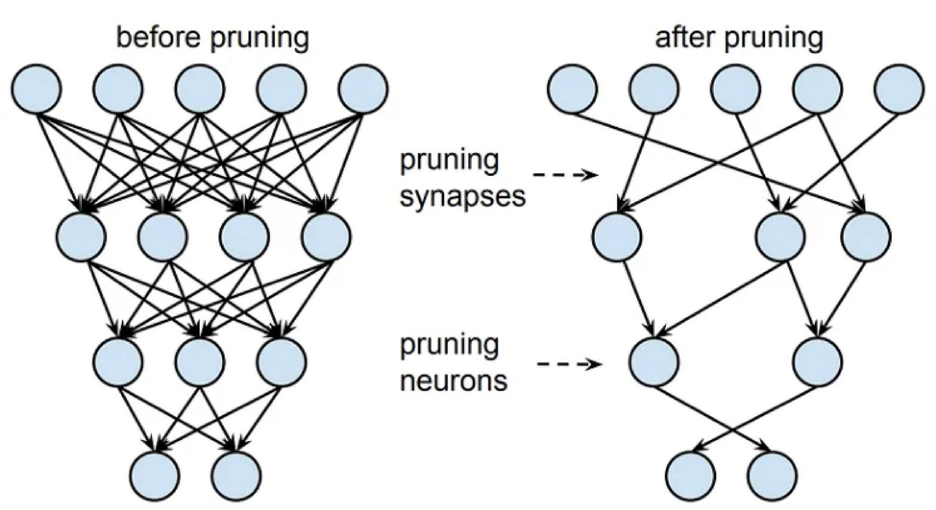


Figure 1. Structured and Unstructured Pruning

Pruning can be classified as structured and unstructured. Structured pruning involves removing entire neurons, channels or other structured group of parameters from a neural network. Unstructured pruning, on the other hand, involves removing individual weights or connections from the neural network without regard for the networks original structure.

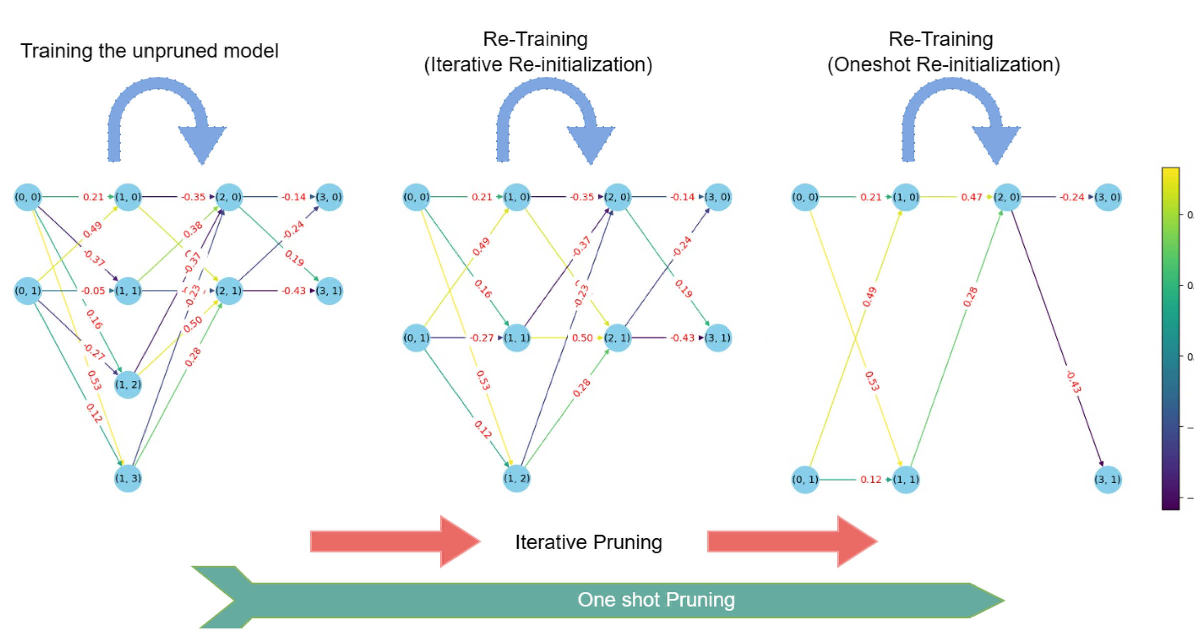


Figure 2. An illustration of Structured Pruning Techniques

The Lottery Ticket Hypothesis

A randomly-initialized, dense neural network contains a subnetwork that is initialized such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations.

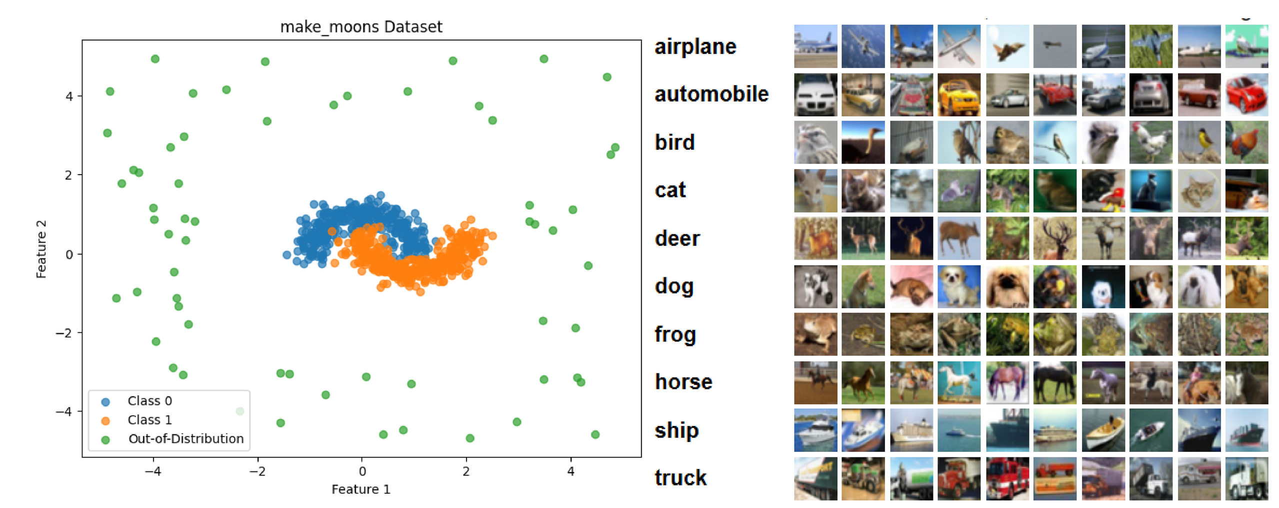
Pruning Algorithm

Input: X, the dataset on which to train the model
Pm, ratio of nodes to be pruned in each layer
N, the number of layers in Neural Network

W ← initialise()
W ← train To Convergence(f(X; W))
for i in 1 to N-1 do
    pruned layers ← prune Pm features from i^{th} layer
    $(i+1)^{th}$ layer ← remove out edges from pruned layer
end for
return pruned model

Datasets used

We have used two dataset for our study, **Make Moons** is used to measure the affect of pruning on small Datasets. **CIFAR10** is a dataset with 60,000 32x32 images divided into 10 classes used to find pruning results of accuracy and calibration on large CNNs.



(a) Make Moons Dataset (b) CIFAR10 Dataset

Figure 3. Datasets Used for study

Accuracy with increasing pruning ratio

We measured the accuracy of different pruning techniques by varying the pruning ratios on the make moons dataset, and we obtained Figure 4.

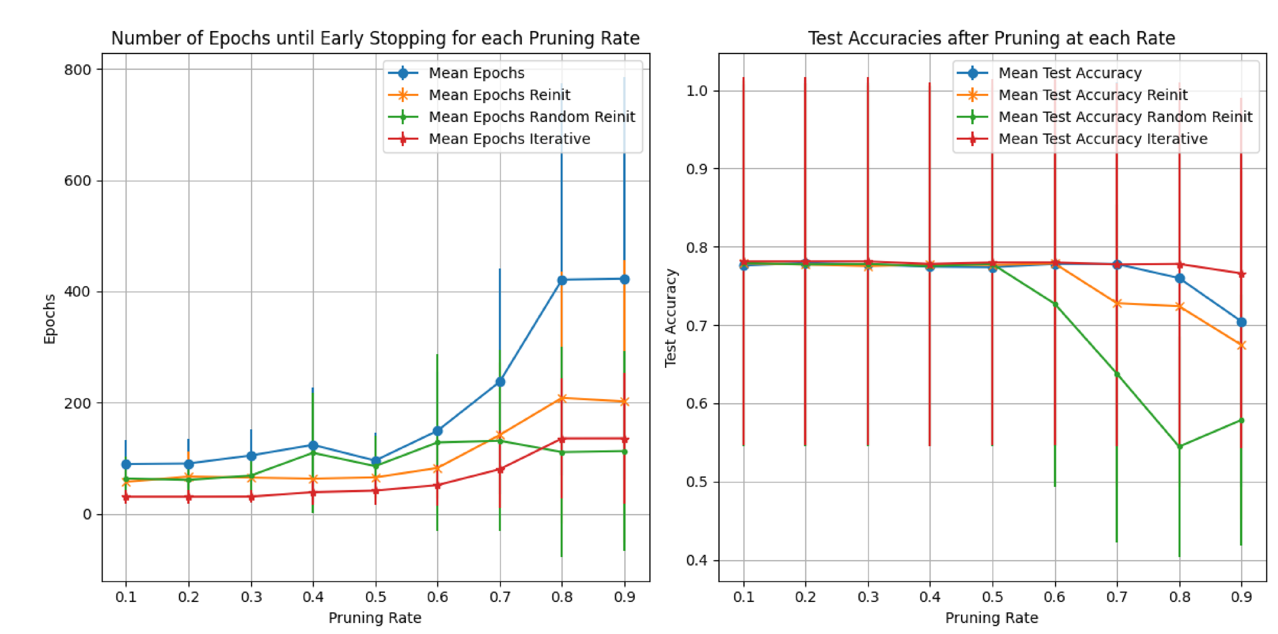


Figure 4. Accuracy vs Pruning Ratio for different techniques

On performing Iterative pruning, on MLP part of a CNN, we obtain a lottery ticket, which is a smaller model and trains to the minimum test validation loss in less iteration, giving higher test accuracy.

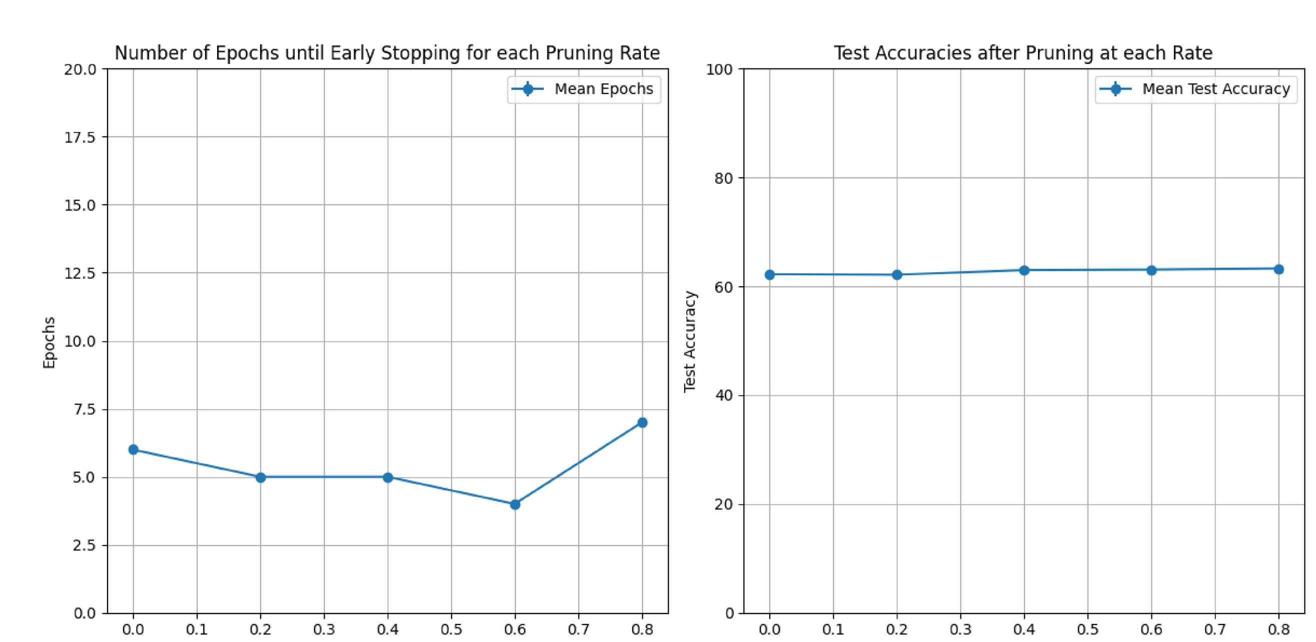


Figure 5. Early-Stop Iteration (Val.) and Test Accuracy vs. Pruning Ratio for Cifar10 Dataset

Expected Calibration Error

$$\text{ECE} = \sum_{i=1}^{K} \frac{|B_i|}{n} |\text{acc}(B_i) - \text{conf}(B_i)|$$

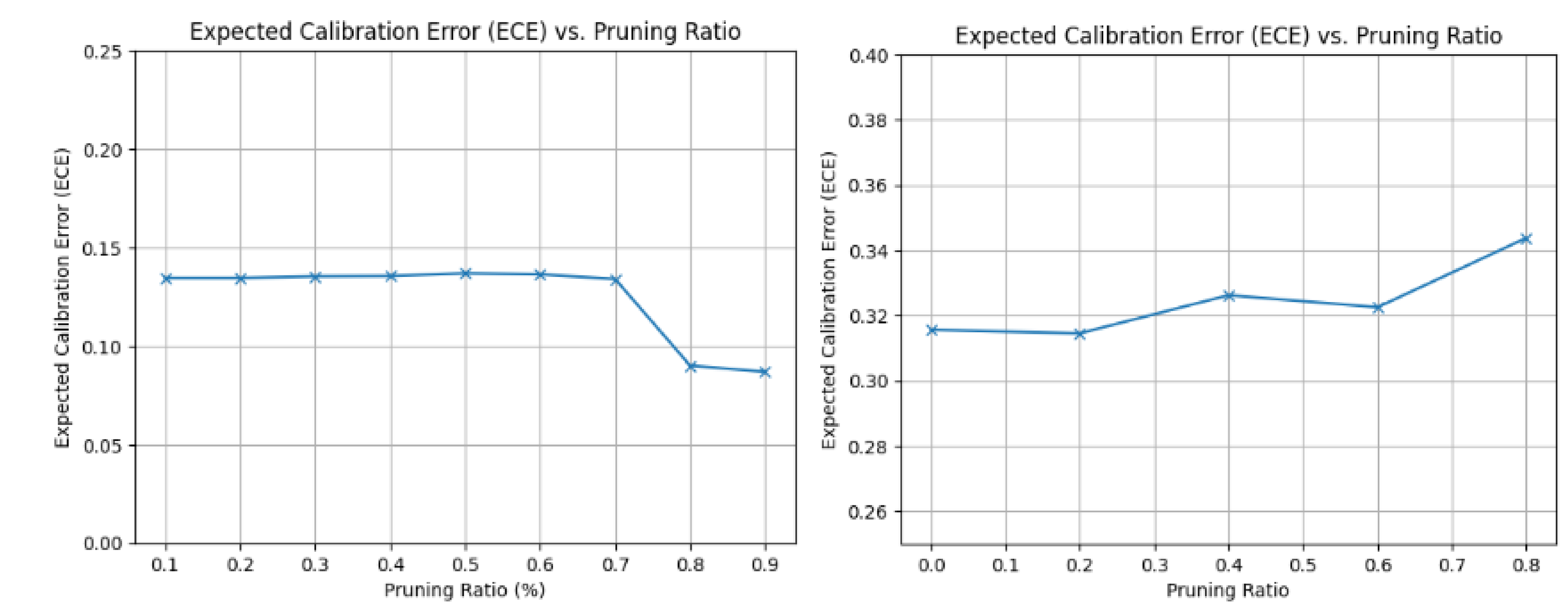


Figure 6. Expected Caliberation Error (ECE) vs Pruning Ratio for Make Moons & CIFAR10 Dataset

Out of Distribution (OOD) Detection

For a better understanding of calibration and the value of ECE, we performed an Out-of-Distribution (OOD) Detection on **CIFAR100** dataset.

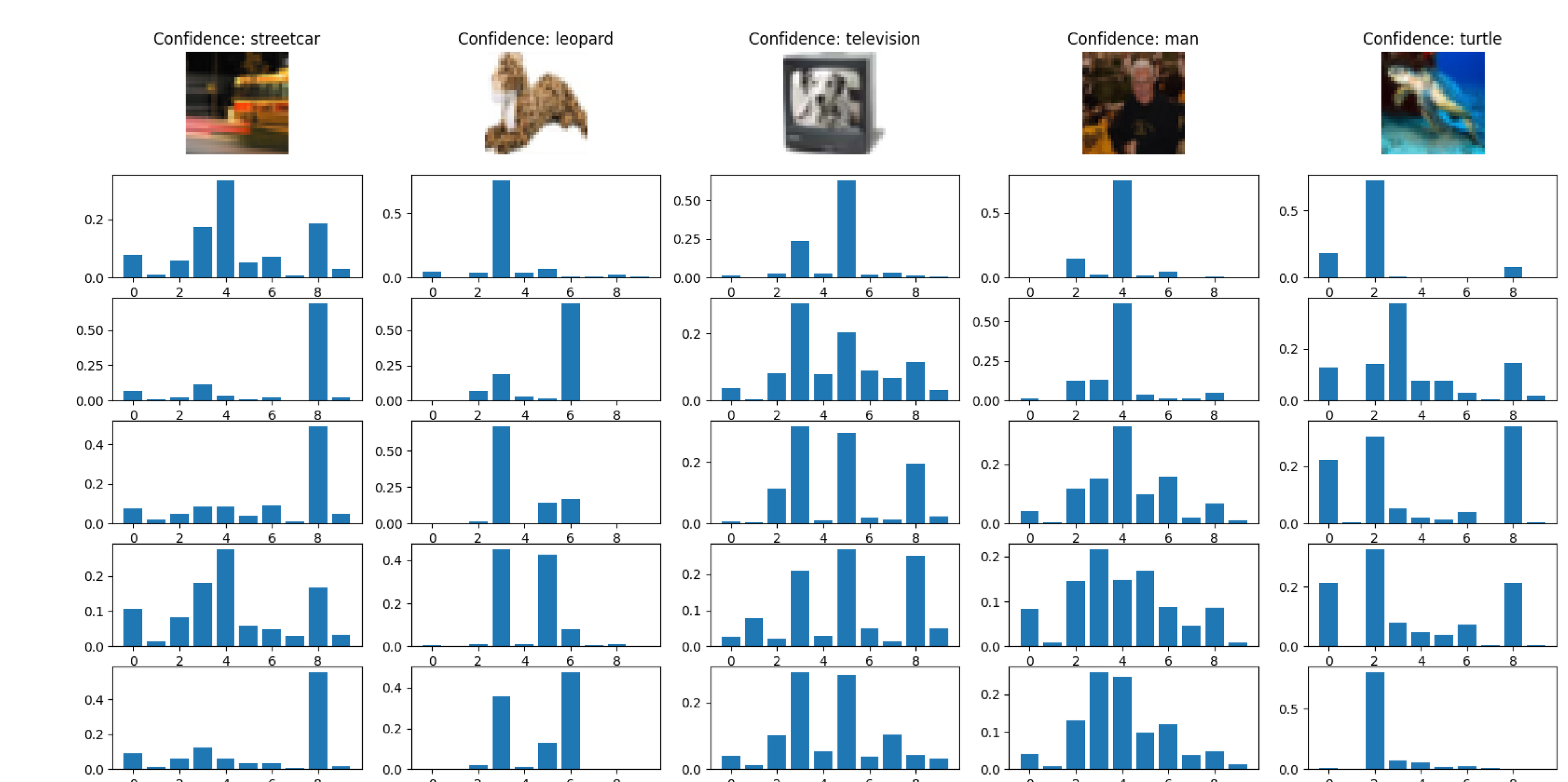


Figure 7. Confidence Histogram vs Pruning Ratio for Cifar100 Dataset

Conclusion

Our results show that pruning may considerably improve the model's calibration without being specifically designed for this purpose. Pruning may thus have a positive effect on **reliability** and **robustness**.

References

[1] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? 2020.

[2] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. 2019.