BTP - II (CS47006)

Streamlining Convolutional Neural Networks: Channel-Level Sparsity for Efficient Compression

By - Nikhil Saraswat (20CS10039)

Supervisor - Prof. Pabitra Mitra

Problem

- CNN deployment challenges:
  - model size
  - memory
  - computational complexity

  Hinders CNN deployment in real-world applications

- Solutions:
  - Compression
  - sparsity-inducing techniques

    etc.

Motivation

- Need for further efficiency enhancement by innovative refinement strategies.
- Nonconvex regularization techniques (*Lp* and *TL1* penalties) are promising.
- Aim for improved performance without compromising accuracy.

VGG16 Architecture

Effective image classification

Approx. 20.04 M parameters

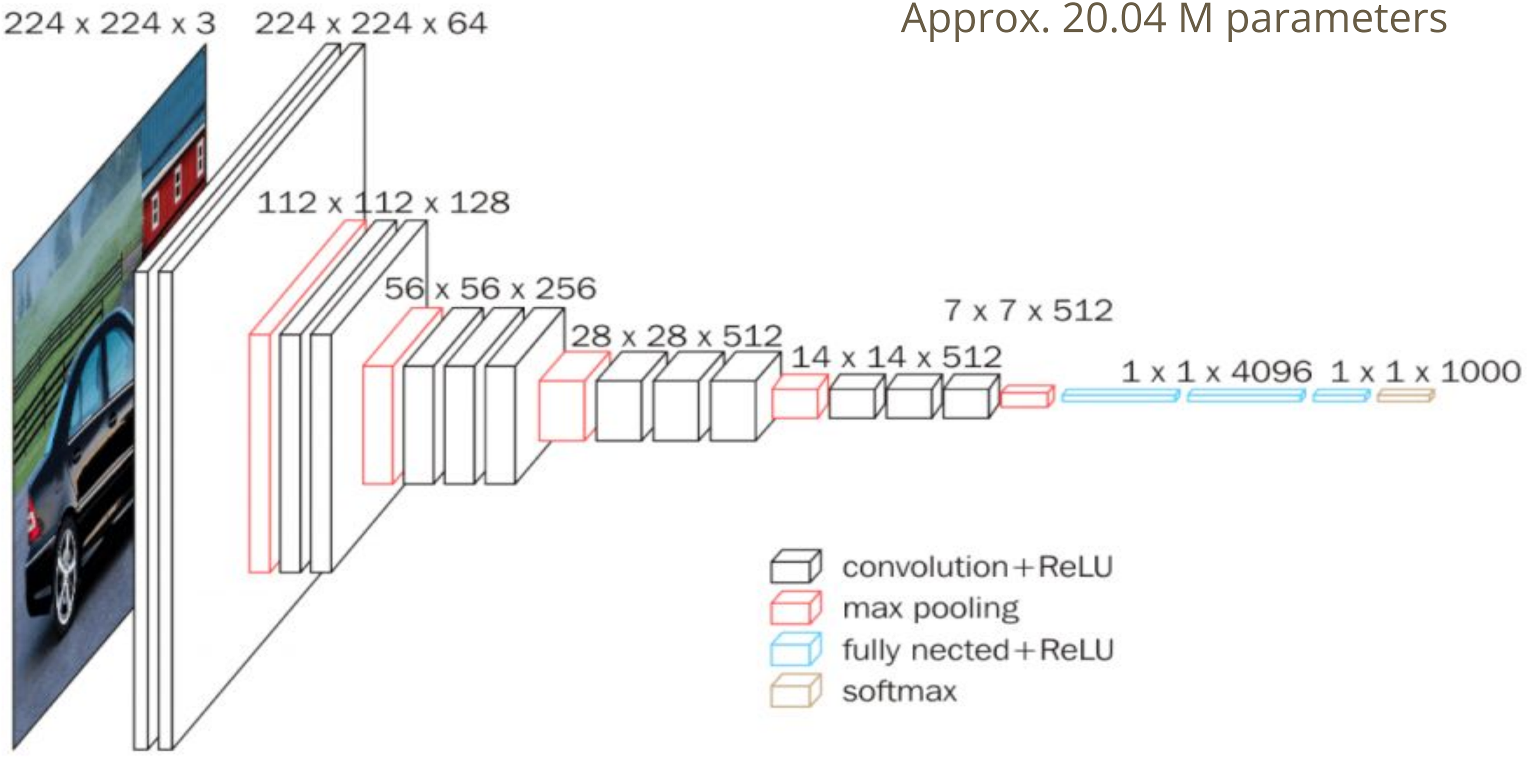


Figure: Visualization of VGG-16 architecture (link)

Non-convex Penalties

- *Lp* and *TL1* boosts sparsity due to their non-convex behavior.
- Higher sparsity ⇒ less bias in parameter selection, aiding in accurate pruning.
- Fosters continuity, ensuring smoother transitions in parameter magnitudes.

Sparse Regularization

- ***L1*** **Regularization:**

$$L_1 \text{ regularization term} = \lambda \sum_{i=1}^{n} |w_i|$$

- ***Lp*** **Regularization:**

$$L_p \text{ regularization term} = \lambda \left(\sum_{i=1}^{n} |w_i|^p \right)^{\frac{1}{p}}$$

- ***TL1*** **Regularization:**

$$TL_1 \text{ regularization term} = \lambda \sum_{i=1}^{n} \frac{(a+1)|w_i|}{a+|w_i|}$$

Channel Pruning

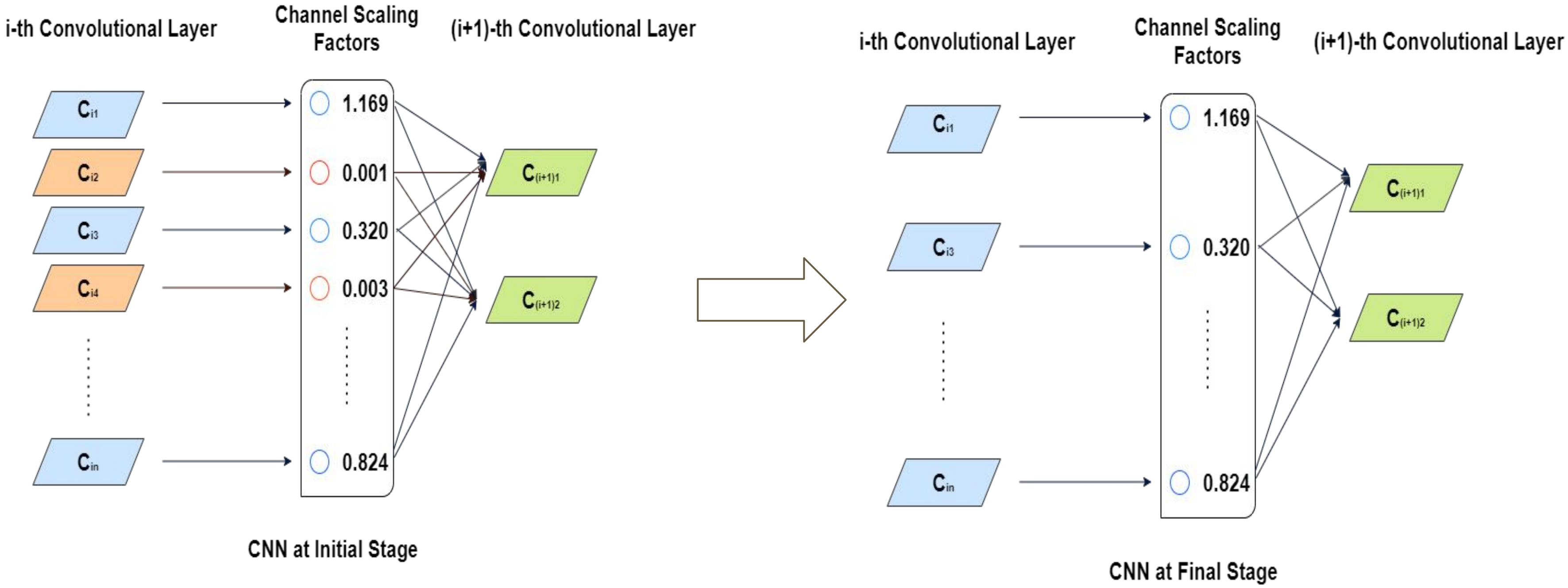
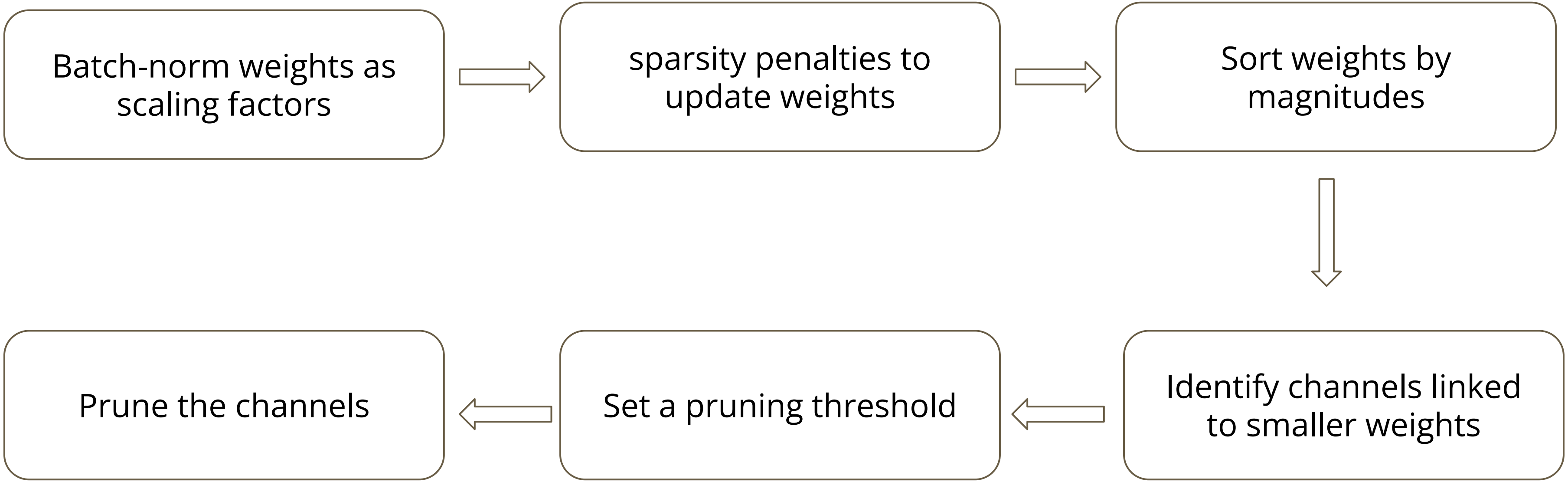


Figure: Channel Pruning (link)

Channel-Level Sparsity: Leveraging Scaling Layers



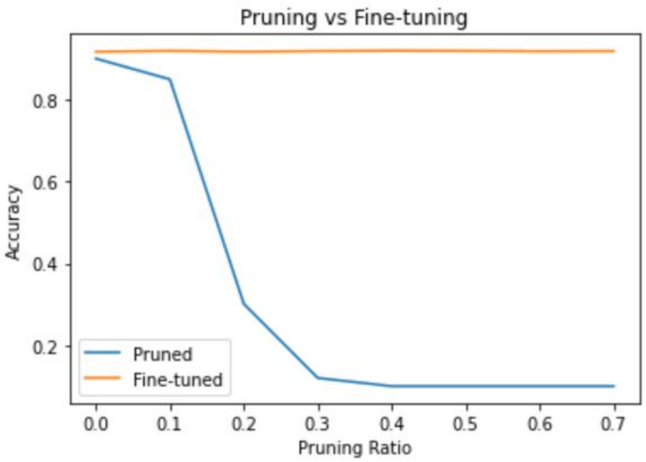
Re-training/Fine-Tuning for Recovery of Accuracy

- The pruned model exhibits reduced accuracy due to the removal of potentially valuable channel information.

- Fine-tuning process adjusts the remaining parameters, enabling the model to regain performance comparable to the original, unpruned model.

Results on Pruning and Fine-tuning

Pruning %	Pruned Acc. %	Fine-tuned Acc. %
No Pruning	91.71	91.75
10	85.01	91.97
20	30.12	91.75
30	12.00	91.91
40	10.00	92.01
50	10.00	91.96
60	10.00	91.87
70	10.00	91.90



Results on *L1* Sparsity Penalty

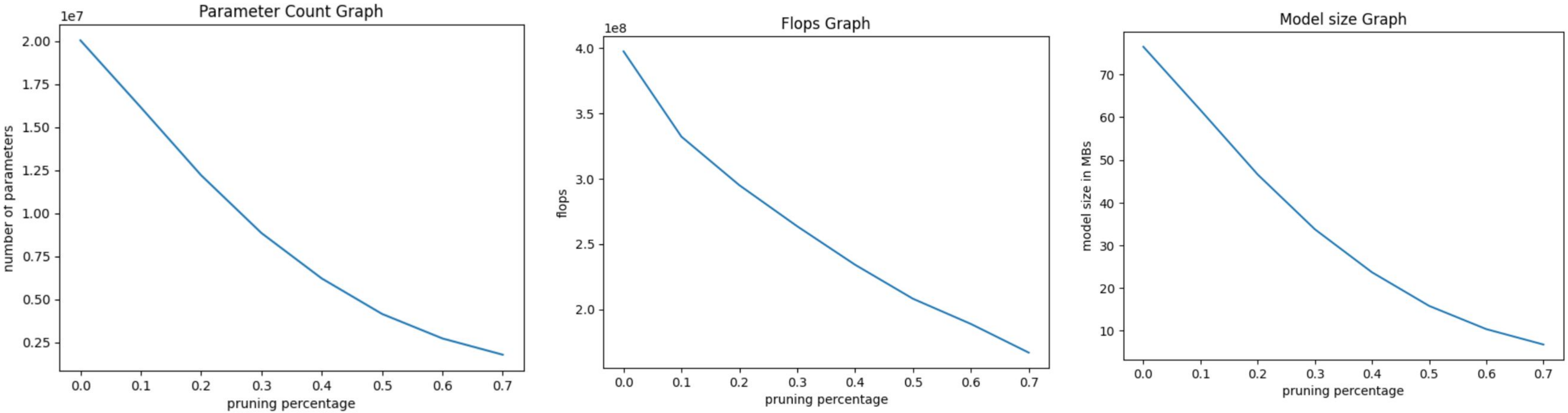
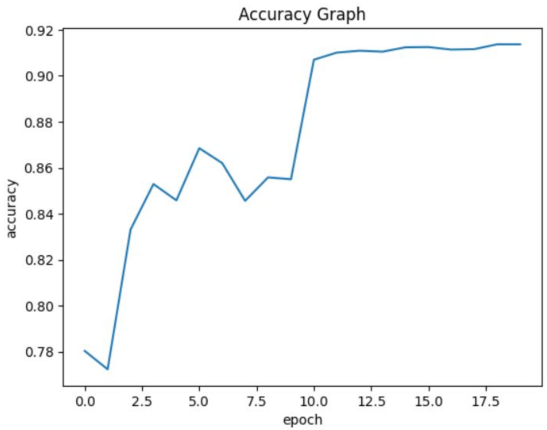
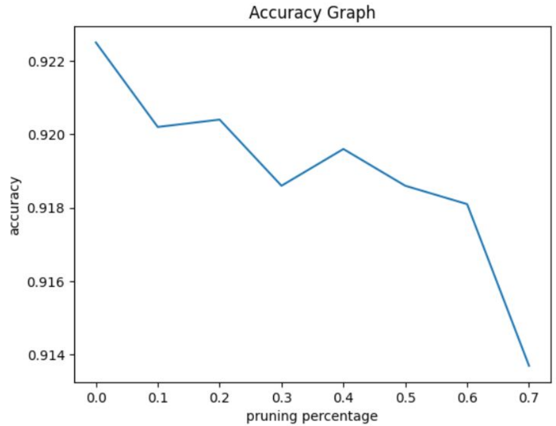


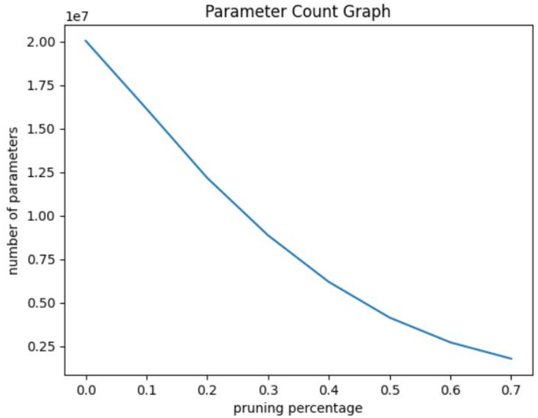
Figure: Plots of No. of Parameters, Flops and ModelSize w.r.t. Pruning ratio

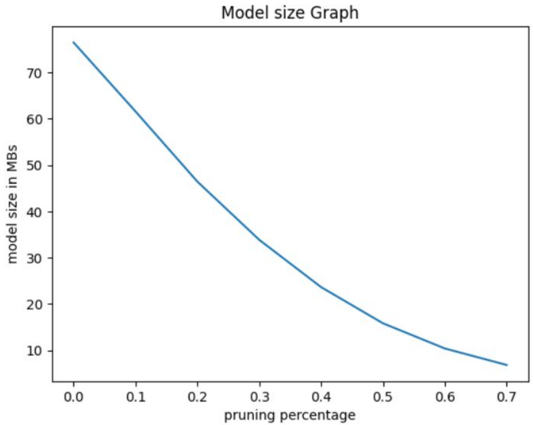
Pruning with *Lp* regularization

For p = 0.50



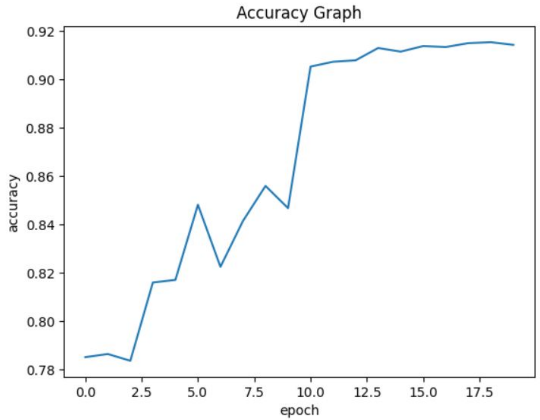


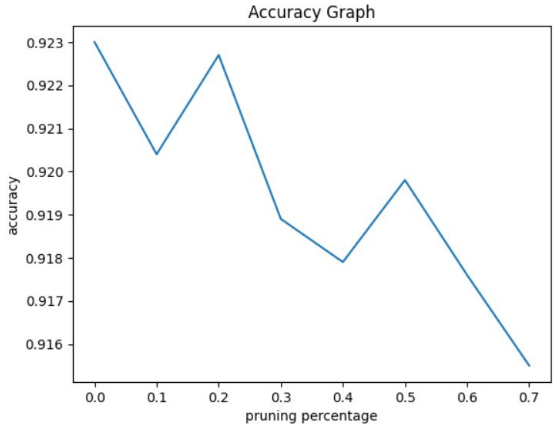




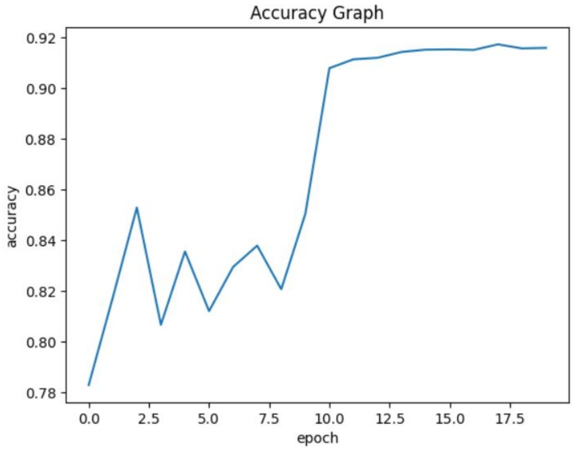
Pruning with *TL1* regularization

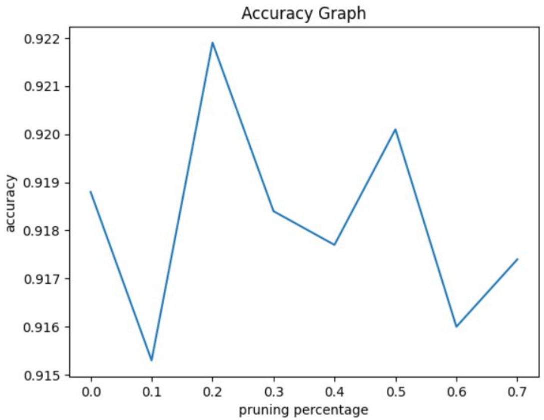
For a = 1.0





For a = 0.5





Comparison of various Regularization penalties

Regularization Penalty	Acc % (30%)	Acc % (70%)	Params (30%)	Params (70%)
L1	91.71	91.61	8838650	1779829
L1/4	91.24	91.12	8835852	1775829
L1/2	**91.86**	91.37	8859907	1772747
L3/4	91.59	91.65	8865653	1765178
TL1 (a = 1.0)	**91.89**	91.55	**8832892**	1792061
TL1 (a = 0.5)	91.84	**91.74**	8859391	**1754922**

Observations

- Channel pruning achieves up to 10x parameter reduction, leading to significant memory savings.
- Floating-point operation reductions reach around 50%, indicating substantial computational overhead decrease.
- *TL1* and *L1/2* nonconvex regularization techniques can maintain or improve mean test accuracy compared to L1.
- *TL1 (a = 1.0)* and *L1/2* show improved mean test accuracy over *L1*, highlighting their effectiveness.
- *L1/4* exhibits decreased test accuracy due to extensive channel pruning, emphasizing pruning percentage impact on model performance with nonconvex regularization.

Conclusion

- Introduces sparsity-induced regularization for automatic channel pruning without accuracy loss.
- Demonstrates up to 10x reduction in computational costs, decreased model size, and memory requirements.
- No significant training overhead; doesn't require specialized libraries or hardware for efficient inference.
- *TL1*, and *L1/2* nonconvex regularizers outperform traditional *L1*, with *TL1* preserving accuracy post-retraining and achieving superior compression.

Future-work

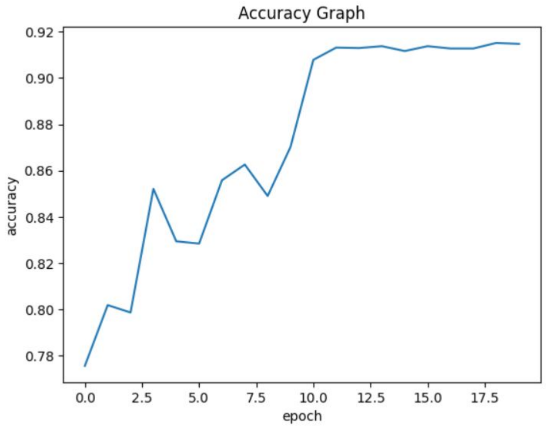
- **Experimentation with Larger Models:** Extend method to ResNet and DenseNet to evaluate effectiveness across diverse architectures and datasets.
- **Exploration with SVHN & CIFAR-100:** Assess scalability and performance on CIFAR-100 and SVHN datasets for insights into handling more complex data.
- **Assessment of Additional Nonconvex Regularizers:** SCAD (Smoothly Clipped Absolute Deviation) and MCP (Minimax Concave Penalty)
- **Optimization for Real-World Deployment:** Optimize method for scalability, efficiency, and practical implementation, including enhancements for inference speed and memory usage.

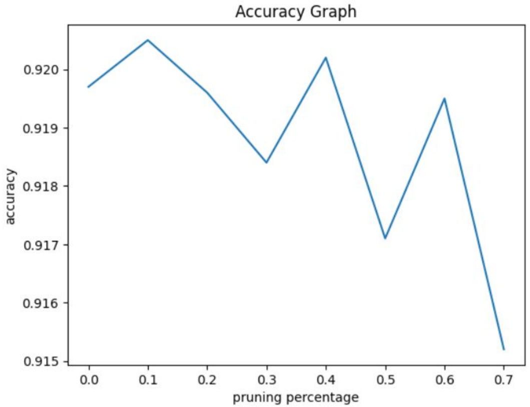
Thank You

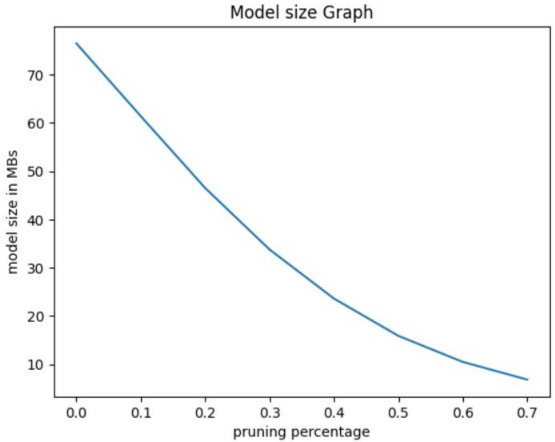
Appendix

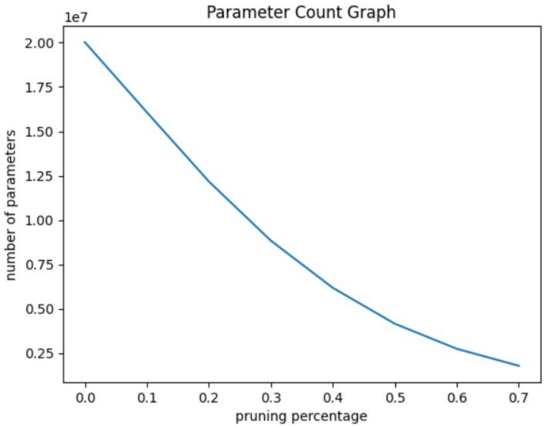
Pruning with *Lp* regularization

For p = 0.25



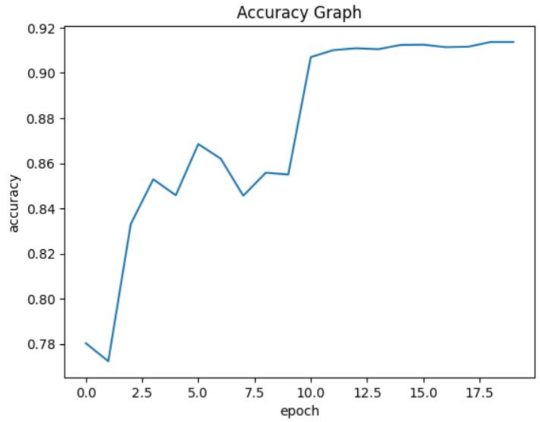


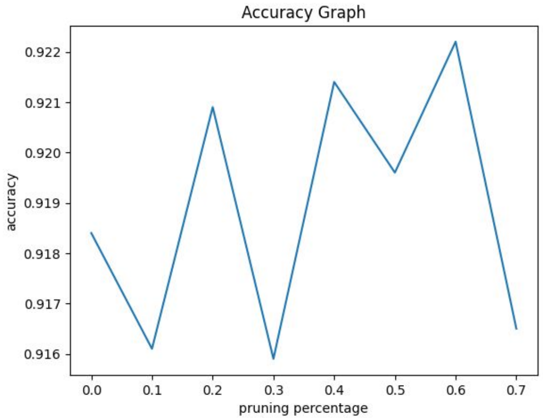


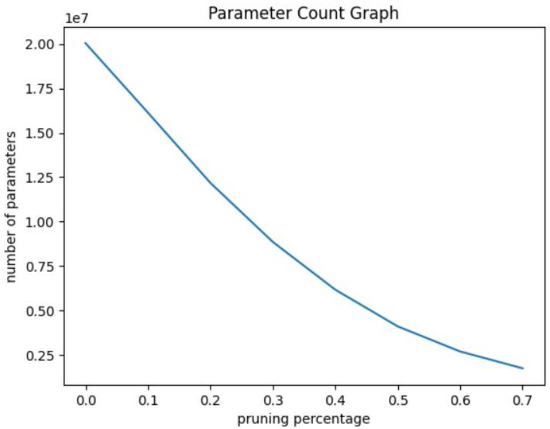


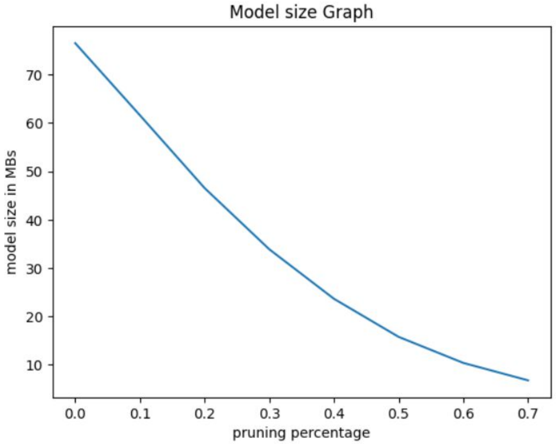
Pruning with *Lp* regularization

For p = 0.75



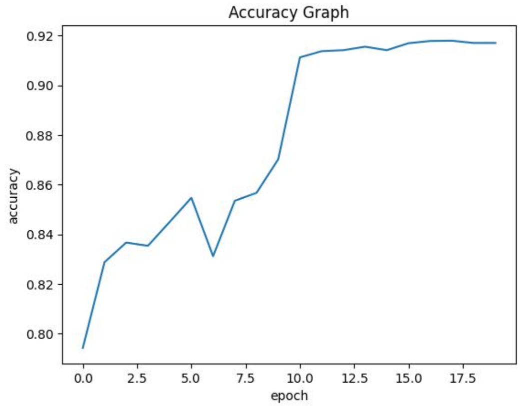


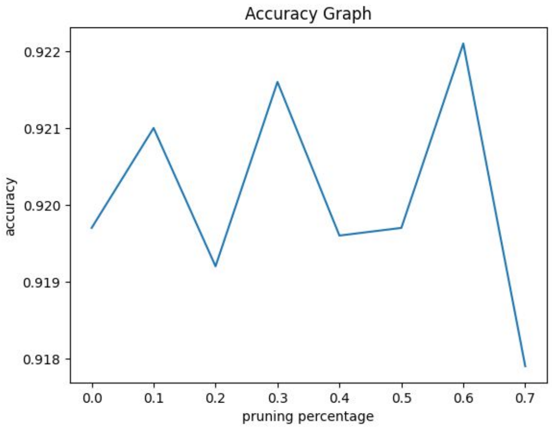


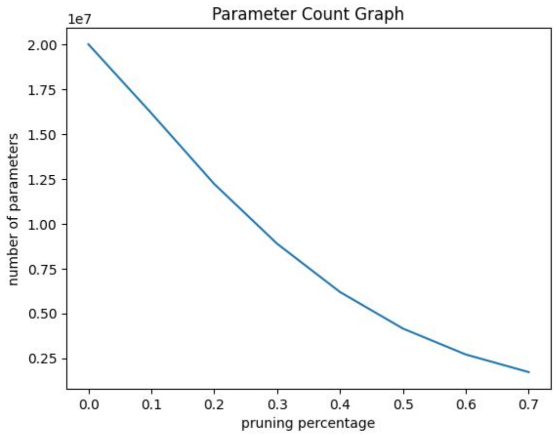


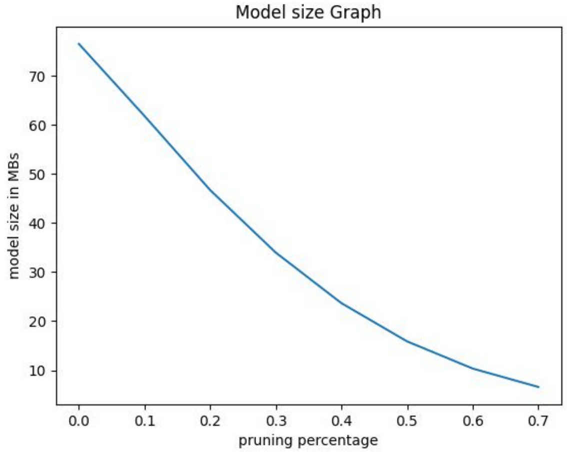
Pruning with *Lp* regularization

For p = 2.0



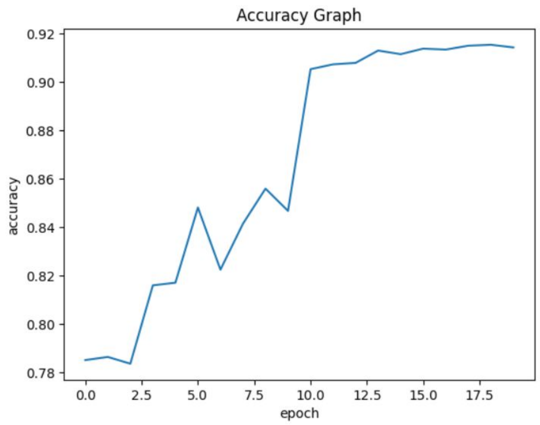


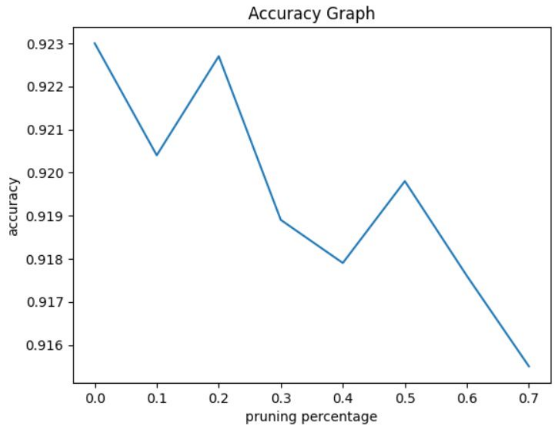


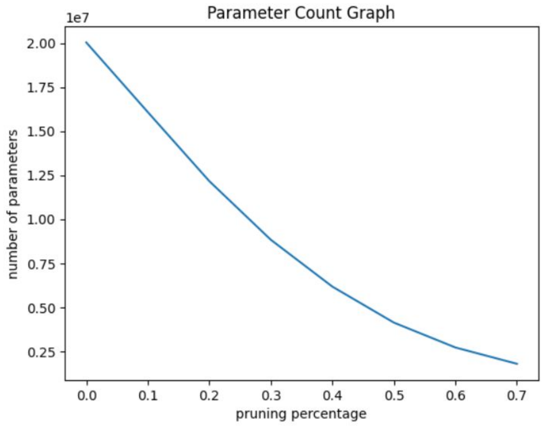


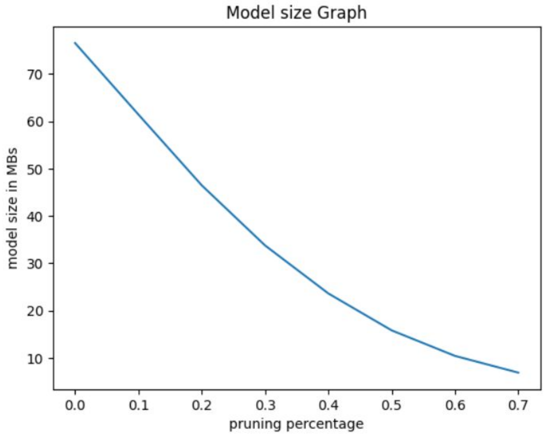
Pruning with *TL1* regularization

For a = 1.0









Pruning with *TL1* regularization

For a = 0.5