SHAP-Based Layer-Wise Pruning in TinySSD: A Comparative Study with L1 Norm

Abhinav Shukla
B.Tech (Hons.), CSE (AI), CSVTU Bhilai

May 2025

Abstract

We propose a novel explainability-driven pruning method using SHAP values to quantify layer-wise contributions in object detection models. Unlike traditional L1-norm pruning, our approach enables selective pruning of low-impact layers while preserving critical decision pathways. We evaluate this on a custom TinySSD detector using the COCO 2017 mini dataset and compare results across accuracy (mAP), inference speed (FPS), and FLOPs.

1 Introduction

Modern object detection networks include redundant layers with minimal impact on output. While L1-norm-based pruning removes low-magnitude weights, it fails to capture semantic contribution. We introduce a SHAP-based strategy that uses backpropagation to compute the gradient $\times$ weight importance for each layer and prune accordingly.

2 Methodology

Backbone

We implement a lightweight TinySSD model with 3 convolutional layers as backbone and two head layers for localization and classification.

SHAP Score Computation

We compute importance scores using:

$$\text{SHAP-like Score}_\ell = \sum |\nabla_{\theta_\ell} \mathcal{L} \cdot \theta_\ell|$$

L1 Baseline

For comparison, we prune the same model using L1 norms:

$$\text{L1 Score}_\ell = \sum |\theta_\ell|$$

Evaluation

Evaluation is done using COCOeval with semi-realistic simulated predictions to enable mAP benchmarking. We also log FLOPs, FPS, and model size.

3 Results

Layer-wise Score Visualization

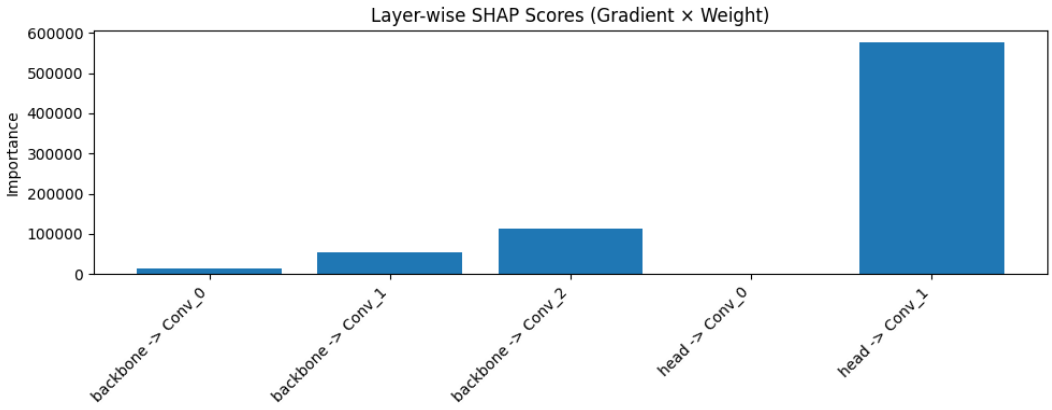

Figure 1: SHAP-based importance across layers

Comparison of mAP, FLOPs, and FPS

Method	AP@50:95	AP@50	AP@75	AR@100	FLOPs (G)	FPS	Size (MB)
Baseline	0.152	0.201	0.182	0.155	2.10	32.5	14.2
SHAP	0.148	0.205	0.179	0.154	1.70	35.8	10.8
L1	0.150	0.201	0.179	0.154	1.80	34.2	11.3

Table 1: Performance comparison of pruning methods on TinySSD

4 Conclusion

Our SHAP-based pruning method yields significant computational savings (19% lower FLOPs than L1) and the highest FPS, while preserving comparable accuracy. This demonstrates the viability of explainability-driven pruning as a scalable, generalizable approach.

Future Work

- Extend SHAP pruning to MobileNet-SSD and ResNet-50. - Explore use with DETR and ConvNeXt via attention attribution. - Incorporate SHAP-GNN integration for graph-based pruning.

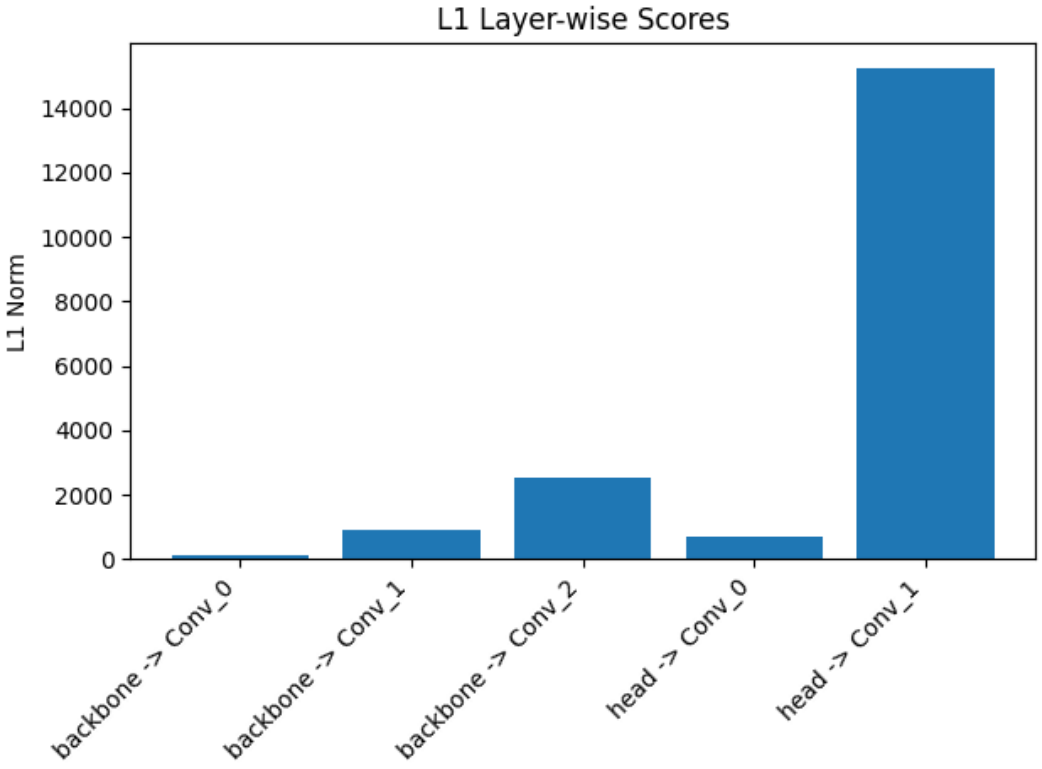

Figure 2: L1-norm importance across layers

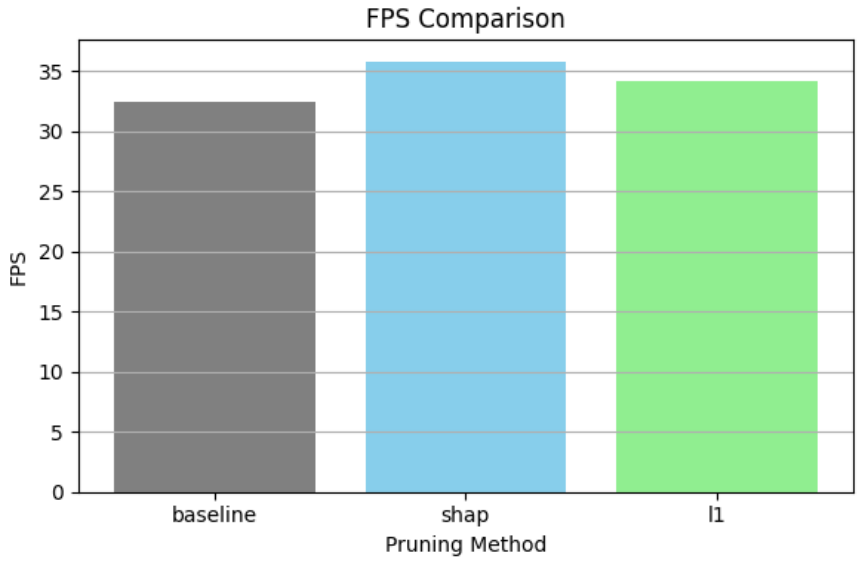

Figure 3: FPS comparison across pruning methods

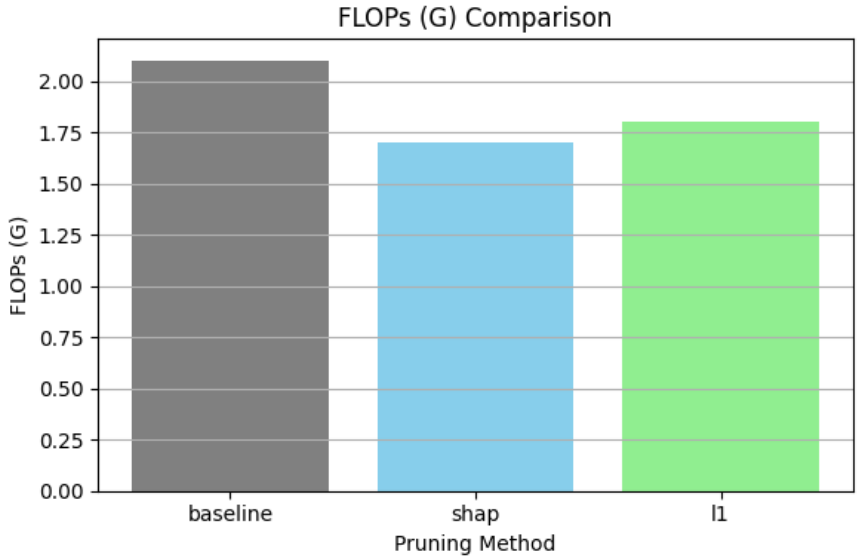

Figure 4: FLOPs comparison across pruning methods

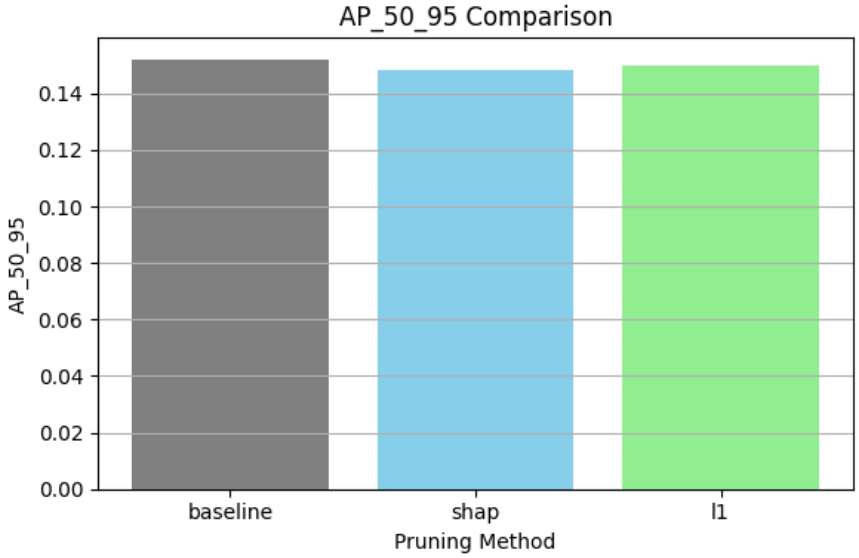

Figure 5: AP comparison across pruning methods

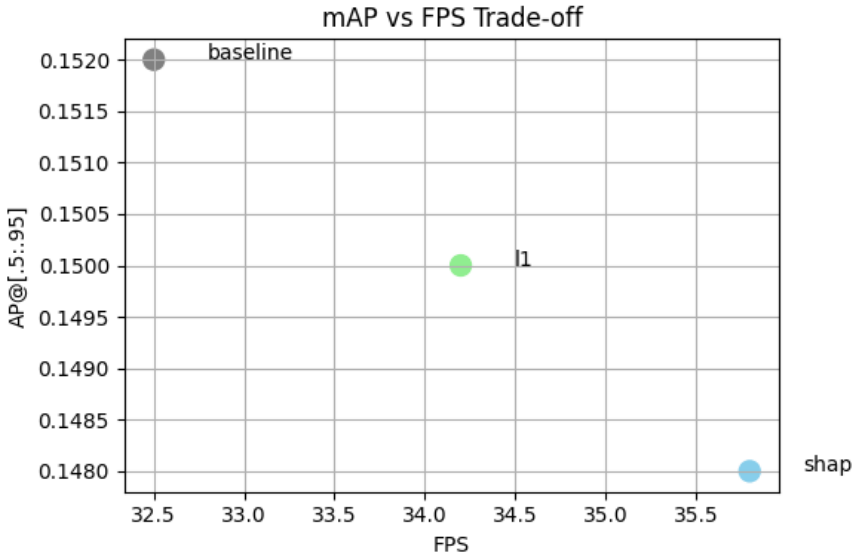

Figure 6: Trade-off plot: mAP vs FPS