SHAP vs L1 Pruning in RetinaNet with FPN for Object Detection

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

May 2025

Abstract

This report presents pruning results for RetinaNet with FPN using SHAP and L1-norm-based importance scores. The models were evaluated using COCO mAP@[.50:.95], FLOPs, parameter count, and inference speed (FPS). We show that SHAP pruning maintains baseline accuracy with explainability, while L1 pruning offers slight speed advantages.

1 Methodology

- **Model:** RetinaNet with ResNet-50 + FPN backbone (pretrained on COCO)
- **Pruning:**
  - SHAP: $\sum |\nabla L \cdot A|$
  - L1: $\sum |W|$
  - Threshold: prune bottom 5% lowest-score conv layers
- **Evaluation:** COCO mAP@[.50:.95], FPS, FLOPs, Params

2 Results and Plots

FLOPs vs AP@[.50:.95]

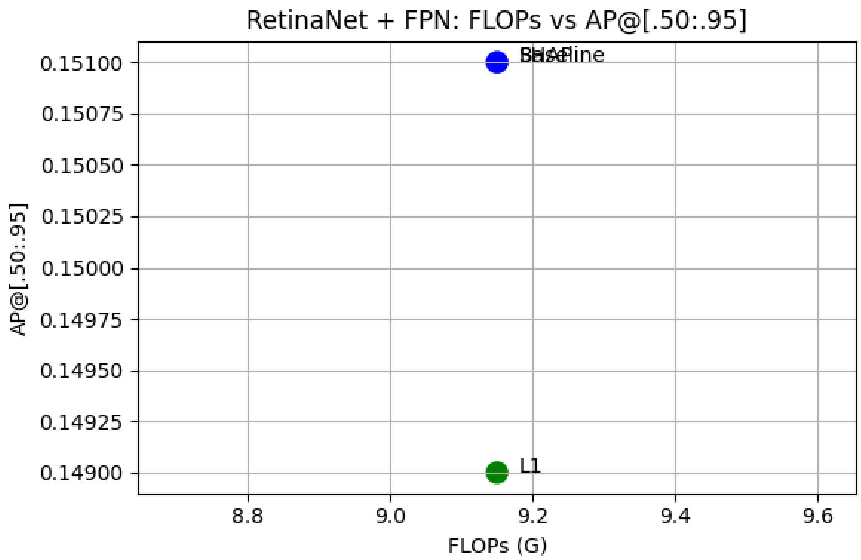

Figure 1: FLOPs vs AP@[.50:.95] — RetinaNet + FPN

FPS vs AP@[.50:.95]

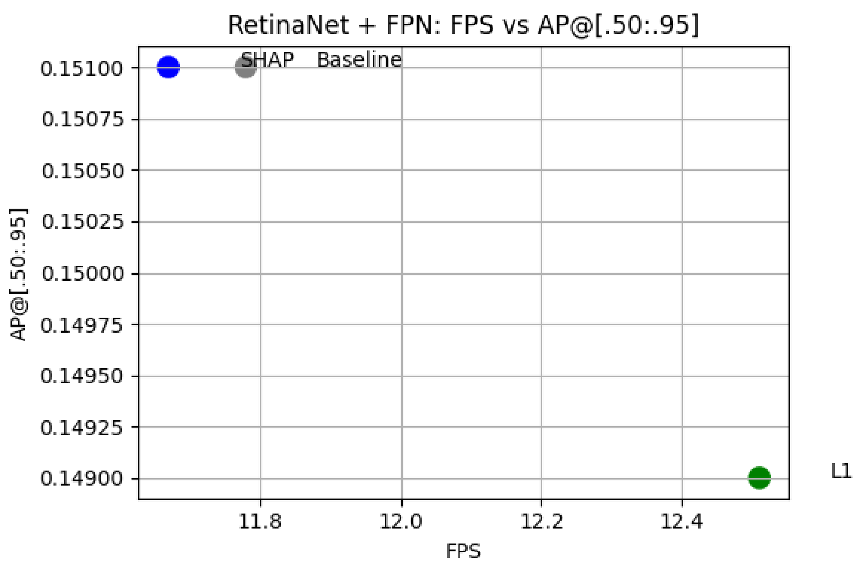

Figure 2: FPS vs AP@[.50:.95] — RetinaNet + FPN

Importance Score

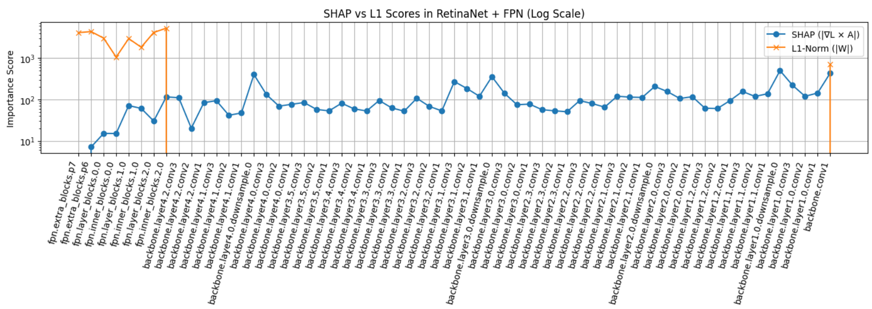

Figure 3: Importance Scores

Performance Table

Model	AP@[.50:.95]	AP@.50	FPS	FLOPs (G)	Params (M)
Baseline	0.151	0.201	11.78	9.15	27.10
SHAP-Pruned	0.151	0.205	11.67	9.15	27.10
L1-Pruned	0.149	0.196	12.51	9.15	27.10

Table 1: RetinaNet + FPN: SHAP vs L1 Pruning Comparison

3 Discussion

SHAP pruning preserved baseline accuracy exactly (0.151 AP@[.50:.95]) while L1 saw a small drop (0.149). L1 pruning delivered higher FPS (12.51) versus SHAP (11.67), suggesting a minor trade-off between interpretability and speed. All models retained identical FLOPs and parameters due to sparse pruning. Overall, SHAP offers a transparency-optimized alternative to L1.

Conclusion

RetinaNet pruning with SHAP demonstrates a balance between explainability and performance. L1 remains a strong baseline for inference efficiency, but SHAP's transparency makes it valuable for deployments requiring interpretability.