CATANet-XS: Real-Time Lightweight Super-Resolution via Pruning and Distillation

Haejun Lee, Junwoo Park, Jiyun Lee, Chaeeun Lee
{2020440109, 2021440054, 2023440099, 2023440102}
School of Electrical and Computer Engineering, University of Seoul
{cobra1318, junwoo5914, myannie4869, dlcodms6084} @uos.ac.kr

Abstract

Lightweighting is essential for ai models in on-device environments. While existing lightweight models like CATANet [4] offer impressive performance, their architectures are static and manually designed for specific sizes (e.g., S/M/L). This leads us to wonder whether an automatically generated architecture could be more efficient than its manually designed one.

In this project, we create CATANet-XS (student) by applying the OPTIN framework [3] and Feature-based Knowledge Distillation [1, 2] to a CATANet-L (teacher) model. We seek to validate whether CATANet-XS can demonstrate significant results in terms of inference time and performance compared to the baseline CATANet-S.

Our goal is to achieve better performance than CATANet-S, while targeting real-time inference speeds (24 fps or more). Furthermore, we will examine the potential for this pipeline to be generically applied to other Transformer-based Super-Resolution models and lead to significant results.

1. Introduction

As mobile device usage increases and privacy concerns become more pronounced, the importance of On-device AI is increasingly emphasized. In the field of Super-Resolution, CATANet [4] significantly improved inference speed by separating the Token Center Update problem into the training phase. However, real-time video processing (24 fps or higher) remains challenging, as CATANet-L achieves 86ms on RTX 4090 [4].

Most existing architectures are manually and statically designed. CATANet-S (230K) and CATANet-L (477K) are predefined by manually adjusting hyperparameters like `dim` or `blocknum` [4]. This manual design introduces two problems: (1) no guarantee of optimal parameter allocation within a given budget, and (2) inability to leverage knowledge from larger pre-trained models.

This paper proposes an automated pipeline combining OPTIN [3] and Feature-Affinity Knowledge Distillation (FaKD) [1] to address these issues. Table 1 presents preliminary results on Set5. After pruning, performance drops from 38.26 dB to 35.63 dB but recovers to 37.90 dB (output distillation) and 37.78 dB (feature distillation), with the latter still improving. Parameters remain at 0.477M as weights are masked rather than physically removed. FaKD could not be evaluated due to CUDA OoM issues. Model rebuilding and memory optimization are expected to achieve near-teacher performance with actual parameter reduction.

Table 1. Preliminary Results on Set5 (10 iterations)

Model	PSNR (dB)	SSIM	Params (M)
CATANet-L (Teacher)	38.26	0.9616	0.477
Pruned (Weight Masked)	35.63	0.9457	0.477*
+ Output Distillation	37.90	0.9602	0.477*
+ Feature Distillation	37.78	0.9596	0.477*

*Weights masked but not physically removed

2. Related Work

2.1. Lightweight SR Transformers

Traditional Transformers have content-agnostic local window approaches issues. CATANet [4] addresses this by proposing an innovative mechanism called Content-Aware Token Aggregation called CATA. CATA improves the inference speed degradation seen in previous clustering models by updating token centers only during the training phase.

2.2. Pruning

Model pruning is a direct method to compress models by removing parameters. Unlike traditional pruning framework that rely on magnitude-based, OPTIN [3] introduces one-shot pruning framework for Transformers. This allows for identifying the optimal sub-network that minimizes infor-

mation loss without the need for extensive re-training from scratch.

2.3. Knowledge Distillation

Since model pruning inevitably induces performance degradation, a training technique to recover this loss is essential. In Super-Resolution tasks, feature-based distillation is demonstrated to be more effective than logit-based approaches. We specifically adopt Feature-Affinity Knowledge Distillation (FaKD) [1], an advanced method that transfers not only feature values but also their affinity.

3. Proposed Method

We train a student model (CATANet-XS) by applying a pruning and distillation pipeline to CATANet.

3.1. Trajectory-Based Pruning

To identify redundant components within CATANet, we leverage the trajectory-based importance metric (TBI) from OPTIN. Since OPTIN cannot be directly applied to CATANet's hybrid CNN-Transformer structure, we implement custom layers to generate pruning masks.

3.2. Feature-based Distillation Strategy

After pruning, we employ knowledge distillation to minimize performance degradation. We evaluate multiple approaches including output distillation, feature distillation, and Feature-Affinity Knowledge Distillation (FaKD) [1], which transfers both feature values and their affinity relationships.

4. Current Progress and Team Contributions

4.1. Model Implementation

Responsible: Haejun Lee

- **Completed**: CATANet-L baseline implementation, and training pipeline setup
- **In Progress**: CATANet training, Flutter integration for mobile deployment

4.2. OPTIN-based Pruning

Responsible: Junwoo Park

- **Completed**: Adapted OPTIN for CATANet's hybrid architecture, implemented TBI algorithm
- **In Progress**: Physical parameter reduction by rebuilding model architecture

4.3. Knowledge Distillation

Responsible: Junwoo Park

- **Completed**: Developed flexible KD framework with multiple strategies (Output, Feature, and FaKD) and automated analysis pipeline
- **In Progress**: Resolve CUDA OoM issues for FaKD evaluation through memory optimization

4.4. Mobile Application Development

Responsible: Jiyun Lee (UI/UX), Chaeeun Lee (Integration)

- **Completed**: Flutter UI with image/video tabs and native channel setup
- **In Progress**: UI/UX design improvements and interface optimization

5. Plan for Completing the Project

5.1. Project Timeline

Week 1-2 (Nov 28 - Dec 4): Complete CATANet-L training, implement model rebuilding for actual parameter reduction, resolve FaKD CUDA OoM issues, and begin hyperparameter tuning.

Week 3-4 (Dec 5 - Dec 11): Train with FaKD, conduct ablation studies, integrate into mobile app

5.2. Success Criteria

1. **Performance**: CATANet-XS achieves comparable or superior PSNR/SSIM to CATANet-S with fewer parameters (target: $<$230K params)
2. **Efficiency**: Inference speed $\geq$ 24 fps on GPU for real-time processing
3. **Validation**: Ablation study demonstrates effectiveness of Pruning+KD pipeline
4. **Deployment**: Mobile application successfully runs SR inference on actual devices
5. **Generalization**: Discussion on applying this pipeline to other Transformer-based SR models

References

[1] Zibin He, Tao Dai, Jian Lu, Yong Jiang, and Shu-Tao Xia. Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 518–522, 2020. 1, 2

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1

[3] Samir Khaki and Konstantinos N Plataniotis. The need for speed: Pruning transformers with one recipe. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[4] Xin Liu, Jie Liu, Jie Tang, and Gangshan Wu. CATANet: Efficient Content-Aware Token Aggregation for Lightweight Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1