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

1. 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 Dynamic Depth Pruning, inspired by recent work [1], and Feature-based Knowledge Distillation [2, 3] 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.

2. Motivation

As mobile device usage increases and privacy concerns become more and 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, present in previous clustering models, into the training phase.

However, as demonstrated by CATANet-L achieving 86ms on a SOTA GPU (RTX 4090)[4], real-time video processing (24 fps or higher) remains a significant challenge. We identified room for improvement in the fact that most existing architectures are manually and statically designed. (For example, CATANet-S (230K) and CATANet-L (477K) are predefined simply by manually adjusting hyperparameters like `dim` or `blocknum` [4]).

We think manual design introduces two critical problems. First, Lack of Parameter Optimization. There is no guarantee that the 230K parameter structure of CATANet-S is the optimal configuration for performance within that parameter budget. We think that an automatically generated architecture, even with fewer than 230K, could outperform the manually designed model if it finds a more efficient architectural combination. Second, can not use a bigger model. A model like CATANet-S must be trained from scratch with its limited 230K capacity. Consequently, the knowledge from the larger model (CATANet-L) was not recycled even if the larger one has better features and knowledge than small one.

This research is motivated by solving these two problems. We propose a method to automatically generate a smaller, more intelligently trained student model using a pipeline that combines Dynamic Pruning [1] and Feature-based Knowledge Distillation [2, 5].

3. Related Work

3.1. Lightweight SR Transformers

Recently, Transformer architectures have achieved significant success in the field of super resolution. However, the computational complexity of traditional Transformers, or their content-agnostic local window approaches[4], failed to fully use the architecture's potential.

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. Despite this advancement, the resulting CATANet-L/M/S models are limited to being static structures where hyperparameters, such as `dim` and `block num`, are manually-designed.

3.2. Pruning

Model pruning is the most direct method to shrink a model by removing parameters. The TinySR [1] proposed an algorithm named Dynamic Depth Pruning to compress Diffusion Models. This approach does not simply remove layers arbitrarily, but instead, it enables the model to self-learn and explore the combination of layers that retains the most information throughout the training process.

3.3. Knowledge Distillation

Since model pruning inevitably induces performance degradation, a training technique to recover this loss is essential. A recent study in the NLP domain [5] experimentally demonstrated that a pipeline that combines Pruning and Knowledge Distillation is the good way to improve the performance of transformer models, minimizing performance drop while maximizing the inference speed (up to 2.56x)[5].

However, rather than using the logit-based distillation proposed in that NLP study[5], we think that feature-based knowledge distillation — which forces the student to mimic the intermediate features of the teacher — is more effective for super resolution tasks [2].

We aim to combine these three research areas. We will train a Student model (CATANet-XS) by applying this Pruning and Distillation pipeline to the lightweight SR Transformer, CATANet. Our goal is to demonstrate that this combined approach yields significant results even for lightweight models and outperforms manually tuned models.

4. Proposed Approach

This research will use a pruning and distillation pipeline to overcome the limitations of manually designed architectures like CATANet-S [4] and to achieve higher performance with fewer parameters.

4.1. Dynamic Pruning

Instead of manually designing block numbers like CATANet-L/M/S [4], we will discover the optimal parameter configuration from the Teacher model, CATANet-L.

To achieve this, we will apply Dynamic Depth Pruning[1] to the CATANet architecture. We expect that this algorithm will help construct a learnable binary mask during the training process, which selectively identifies the most critical Residual Group blocks to keep.

4.2. Feature-Based Knowledge Distillation

A model generated via dynamic pruning will inevitably suffer from performance degradation. To recover this performance, we will add distillation pipeline, which was proven effective in the previous study [5].

However, mimicking feature representations from a teacher model is widely considered more effective for Super-Resolution tasks than logit-based distillation [3]. Therefore, we will employ Feature-based Knowledge Distillation [2].

5. Evaluation

All experiments will be conducted on the standard benchmark datasets used in the CATANet paper [4] (Set5, Set14, Urban100, and Manga109). To determine the success of our goal, we use PSNR and SSIM as the primary performance metrics, consistent with the original evaluation [4].

Furthermore, inference time will be measured to verify our goal of real-time processing (24 fps or more). Our primary goal is to demonstrate that CATANet-XS can achieve superior (or comparable) performance metrics and faster inference times than the baseline CATANet-S, all while utilizing fewer parameters.

Finally, to confirm its practicality in resource-constrained settings, we plan to develop a demonstration mobile application. This will allow us to assess whether our model achieves the expected results on actual on-device hardware.

References

[1] Linwei Dong, Qingnan Fan, Yuhang Yu, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TinySR: Pruning Diffusion for Real-World Image Super-Resolution. *arXiv preprint arXiv:2508.17434*, 2025. 1, 2

[2] 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

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

[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, 2

[5] Aishwarya Mirashi, Purva Lingayat, Srushti Sonavane, Tejas Padhiyar, Raviraj Joshi, and Geetanjali Kale. On importance of pruning and distillation for efficient low resource nlp. 2024. 1, 2