On Device Single Image Super Resolution on iPhone

Team name: SuperRez

Myles, Bharathi, Vedant, Nick

Introduction

Current Issue: Mobile cameras are widely used but still face limitations, such as unwanted artifacts, lighting issues, or complex settings, which can result in undesirable photos.

Task: Take images as user inputs and output auto-enhanced / augmented images back to the user with an on device model

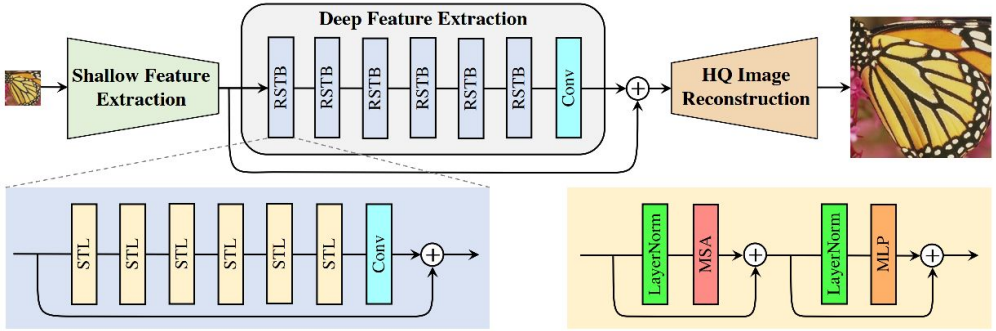
Why: Allow enhancement/augmentation of photos on device ensuring user privacy and creative control without relying on cloud processing

Model Structure

Model: SwinIR Transformer architecture

Inputs: Images - low resolution

Output: Larger (2x) enhanced Images - higher resolution



Evaluation - Data

Training set: DIV2k (800 imgs) + Flickr2K (2650 imgs)
Sometimes: OST (10324 imgs) + WED (4744 imgs) + FFQH (200 imgs)

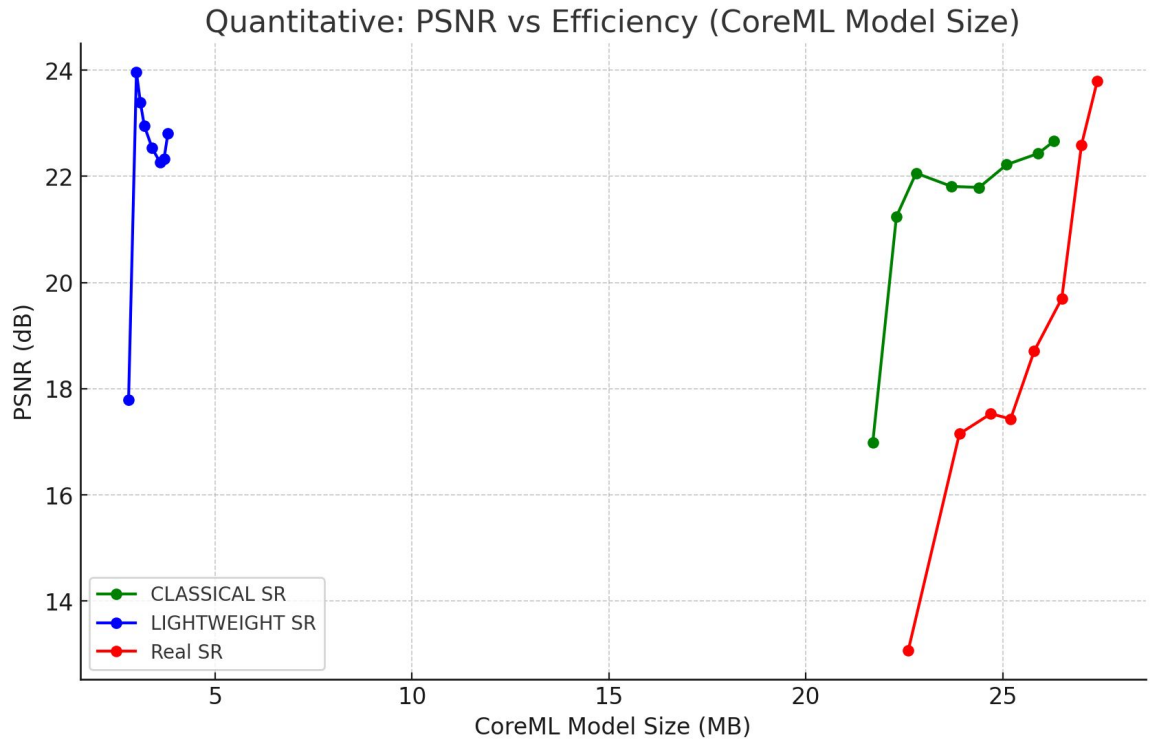
Testing set: Given image testset from SwinIR repo (examples below)

Evaluation - Metrics

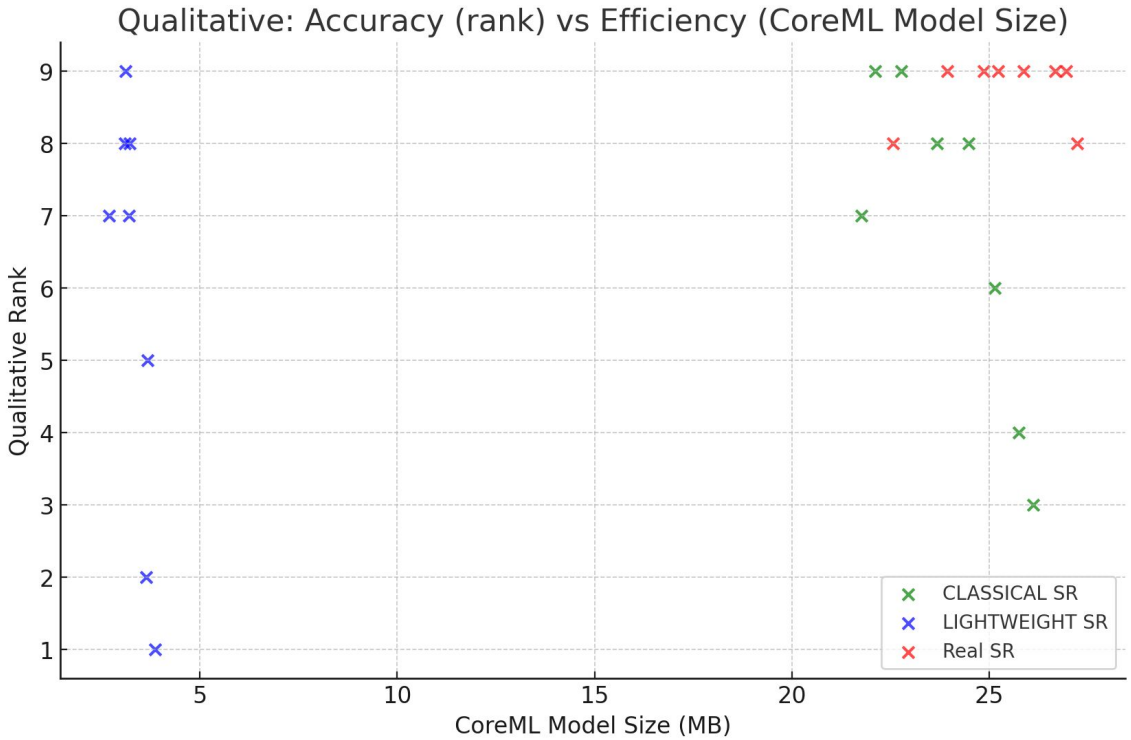
Metric	Purpose of Use
# Params	Relates to the size of the model with how large it is
PSNR	Measures quality of restructured image
Memory	Relates to the size of the model with how much storage it requires
Latency	Captures responsiveness and efficiency of model
dtype	Determines the precision and memory efficiency of computations
Sparsity	Measures how much of the model has been removed

Data:
https://docs.google.com/spreadsheets/d/11k01WrqfHIZPwY8tq7xB6xZ5dlMk2BbCbH1Uu0fDH6o/edit?usp=sharing

Efficiency-accuracy trade-off curve



Efficiency-accuracy trade-off curve



Results - Curve Explanations

PSNR vs Efficiency (Model Size):

- Lightweight SR can obtain similarly high PSNR values compared to classical and real SR with a much smaller model size
- As model size drops for classical and real SR, so does the PSNR with a steep drop for the lowest model size which used unstructured pruning
- For lightweight SR we see an initial increase of PSNR as the model size decreases with a steep drop in unstructured quality for the unstructured pruning method trial

Results - Curve Explanations

Accuracy vs Efficiency (Model Size):

- Lightweight SR can obtain similarly high accuracy compared to classical and real SR with a much smaller model size
- Accuracy remains high for real SR as model decrease with a small decrease in accuracy for the unstructured trial
- Accuracy initially increases as model size shrinks for classical and lightweight SR but decreases for the unstructured trials
- The unstructured trials' accuracy still outperforms the initial accuracies for classical and lightweight

Results - Methods

Attention Head Pruning:

Iterative Magnitude Pruning:

Key insights

1. **Quantization vs. Mixed Precision**: Static and dynamic quantization underperform, while mixed precision compute and FP16 storage work well with Apple's GPU and Neural Engine.
2. **Pruning** heads results in better image quality than pruning neurons per layer or unstructured pruning.
3. **Model Complexity & Quantization Sensitivity:** Larger models seem more sensitive to quantization potentially because SISR and image reconstruction tasks rely on precise, high-fidelity computations and diverse weight distributions to accurately capture fine-grained details and textures.

Challenges

1. **Model Conversion Issues:** Converting models to CoreML somtimes introduces inconsistencies in results.
2. **Performance Tuning Challenges**: Optimizing for real-time inference on varying iOS devices is difficult and does not always produce consistent results.

Future Work

- **Tinker with Model Architecture:** Experiment with architecture and patch size to better utilize Apple's GPU and ANE (Apple Neural Engine)
- **Pruning:** Experiment with alternative pruning techniques such as gradient pruning or sensitivity based pruning
- **Video enhancement**