Pruning Video Super Resolution Models

Ashish Kumar Singh
ashish23ks@cs.ucla.edu

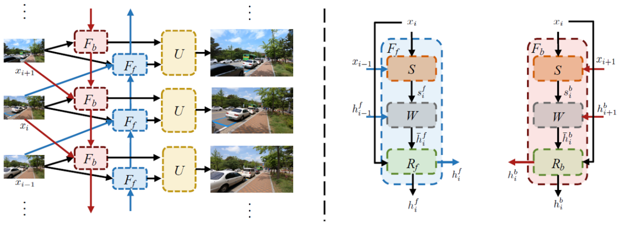
Abstract

Video Super Resolution (VSR) methods are generally very complex and inefficient to deploy on resource constrained platforms. VSR has vast applications in fields like streaming, surveillance, movies, video and astronomy. With these many usecases, there is a need for lightweight general purpose VSR models. This project focuses on using pruning techniques on VSR models and observe the impact on their performance. With unstructured pruning we can reduce model size by 50% with a tradeoff of 0.82 dB in PSNR. With structured pruning, we can speed up the model by 15% with a tradeoff of 3.7 dB in PSNR. A method for pruning aware training is also discussed.

1. Introduction

Single Image Super-Resolution (SISR) [4] aims to restore high-quality images from low quality input images. On the other hand, Video Super Resolution (VSR) [2, 1, 3] aims to restore videos containing multiple frames. We have potentially unlimited data for this task as the inverse mapping (high resolution to low resolution) is trivial. VSR approaches generally have more components than SISR as we have an extra temporal dimension information. VSR poses an extra challenge as it involves aggregation of information from multiple highly-related but misaligned frames. An inherent assumption here is that the videos are natural which put additional constraints to have temporal consistency in the output. For these reason, VSR approaches are generally more complex.

VSR approaches have vast applications in many fields. In gaming, Nvidia uses DLSS (Deep Learning Super Sampling) to upscale rendered low resolution frames to high resolution or increase frame rate [7]. In virtual conference, VSR can provide better quality video conferencing. Streaming sites can use low bandwidth over network to move low resolution video and then use VSR to increase resolution. VSR approaches can be used to restore old low-resolution movies. Most of the surveillance camera are low resolution, VSR methods can be used to improve quality for better recognition. VSR methods can be used to save videos in less amount of memory. In Astronomy, VSR methods can help improve quality of videos of stars and planets.



(a) BasicVSR architecture

(b) Forward and backward propagation branches

Figure 1. BasicVSR[1] architecture

With these many use-cases, there is a need for lightweight general purpose VSR models. This project focuses on using pruning techniques on VSR models and observe the impact on their performance. Specifically, we experimented with unstructured and structured pruning of one of the state-of-the-art VSR model and presented the results. We also discussed about a method for pruning aware training.

2. Related Works

[5] survey talks about various video super resolution techniques and the use of optical flow for aligning and aggregating features from different frames. [2] uses multi-scale deformable alignment module and multiple attention layers for aligning and integrating the features. [3] uses multiple projection modules to sequentially aggregate features. These methods are very complex. Complex designs are effective but inevitably increase the runtime and model complexity. Complex designs poses difficulties in implementing and extending existing approaches and are inefficient to deploy on resource constrained platforms.

To overcome the high complexity of above methods, [1] reconsiders most essential component of VSR guided by propagation, alignment, aggregation and upsampling and proposes BasicVSR model. Propagation refers to the way in which features are propagated temporally. Alignment concerns on the spatial transformation applied to misaligned images/features. Aggregation defines the steps to combine

Model	No. of params	PSNR	PSNR(finetune)	SSIM	SSIM(finetune)
Baseline	7.73M	31.41	-	0.8909	-
Prune 40%	4.74M	28.16	30.84	0.8395	0.8794
Prune 45%	4.35M	27.62	30.72	0.8265	0.8766
Prune 50%	3.87M	22.35	30.59	0.7022	0.8736

Table 1. Unstructured Pruning Comparison

Model	FLOPS	PSNR	PSNR(finetune)	SSIM	SSIM(finetune)
Baseline	101G	31.41	-	0.8909	-
Prune 5%	96.5G	13.69	29.77	0.3556	0.7321
Prune 10%	87.4G	6.53	27.68	0.1038	0.7948
Prune 15%	78.7G	6.95	22.74	0.1946	0.7321

Table 2. Structured Pruning Comparison

aligned features. Upsampling describes the method to transform the aggregated features to the final output image.

3. Method

We use pre-trained BasicVSR model to prune using unstructured and structured techniques. All models are tested on REDS video dataset.

Unstructured pruning aka weight pruning make the weight matrix sparse by pruning unimportant weights while maintaining acceptable accuracy. Weight pruning is mostly useful for model compression. No practical speed up as hardware acceleration of sparse matrix multiplication is not feasible. For unstructured pruning we pruned all Conv2D layers of the BasicVSR model with varying sparsity. Finetuning is done for 1K iterations to regain some of the lost accuracy. Quantitative results are shared in next section.

Structured pruning aka filter pruning prunes whole filter resulting in fewer computation. Filter pruning is mainly focused for acceleration over weight pruning. Number of FLOPS (Floating Point Operations) are used as proxy to compare execution time of different model. For structured pruning also we pruned all Conv2D layers of the BasicVSR model with varying sparsity. Finetuning is done for 1K iterations to regain some of the lost accuracy. Quantitative results are shared in next section.

Inspired from [6], we propose to use pruning aware training as future work. The idea is to regularize the weight of filters during training such that the model can be pruned easily later. We can introduce a gate variable to control the throughput of each filter. By regularizing the gate variable, we can know which filters to remove. Weight normalization can be done to decouple learning of filter and its norm.

$$\hat{W}_i = \frac{W_i}{||W_i||_2}, W_i = \gamma_i \hat{W}_i$$

During training, we can sort the filters in a layer by their L1 norm and set those with least norm using predefined sparsity level as unimportant and regularize them with extra penalty term.

$$\mathcal{L}_{SI} = \alpha \sum_{l=1}^{L} \sum_{i \in S^{(l)}} \gamma_i^2$$

After training, we can perform structural pruning with the same predefined sparsity level to get better pruned model.

4. Experiments Results

4.1. Unstructured Pruning

Table 1. compares quantitative performance of unstructured pruning of BasicVSR model. There is considerable drop in PSNR and SSIM accuracy with 50% pruning ratio but we gain most of the lost accuracy by finetuning. Fig 2 shows PSNR comparison for Unstructured pruning with different pruning ratio. There is major drop from 40 % pruning ratio, but finetuned model performs similar to baseline.

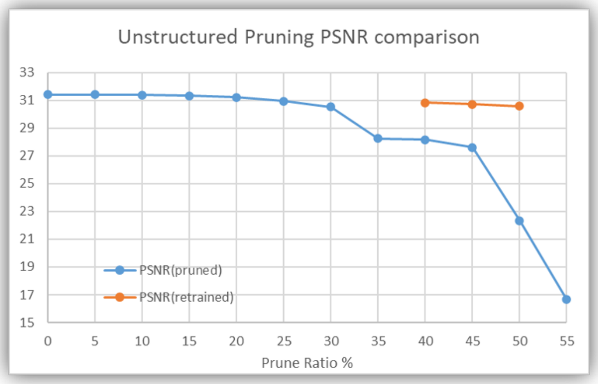


Figure 2. Unstructured Pruning PSNR comparison

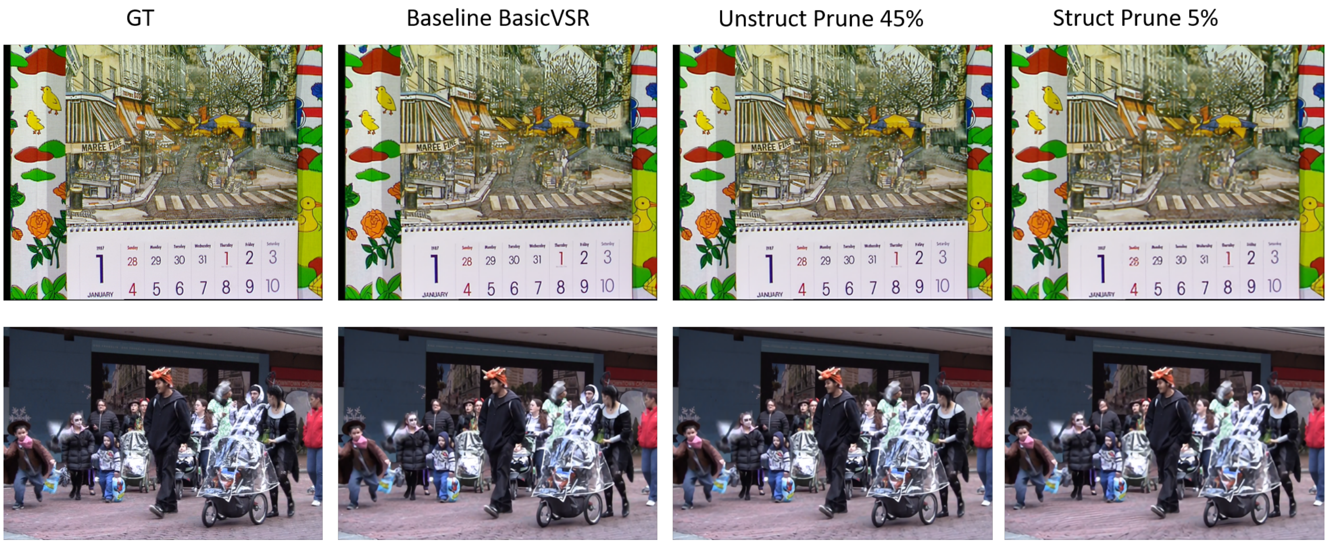


Figure 3. Pruning BasicVSR model Qualitative comparison

4.2. Structured Pruning

Table 2. compares quantitative performance of structured pruning of BasicVSR model. There is considerable drop in PSNR and SSIM accuracy even with only 5% pruning ratio but we gain some of the lost accuracy by finetuning. Though it is still worse that its corresponding value for unstructured pruning. Fig. 4 shows that structured pruning severely affects the performance, though some accuracy is regained after finetuning.

Fig. 3 compares the output of different pruning techniques visually.

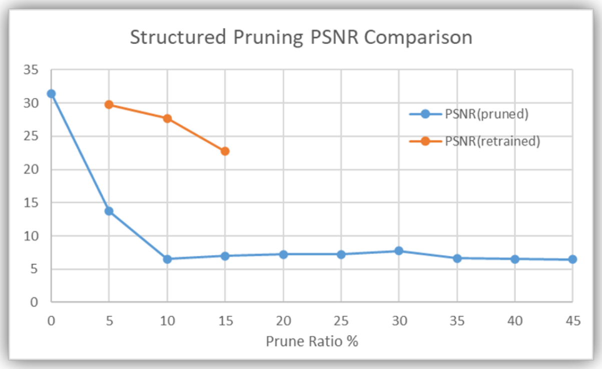


Figure 4. Structured Pruning PSNR comparison

5. Conclusion

Experiments on unstructured pruning shows that we can compress 50% of BasicVSR model without much effect on the performance. And for structured pruning, even to achieve 15% speedup, there is some degradation in the performance. To better condition model for structured pruning, pruning aware training is proposed as future work.

5.1. References

References

[1] Kelvin et. al., BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond 1

[2] Xintao Wang et. al., EDVR: Video Restoration with Enhanced Deformable Convolutional Networks 1

[3] Muhammad Haris et. al., Recurrent Back-Projection Network for Video Super-Resolution 1

[4] Jingyun Liang et. al., SwinIR: Image Restoration Using Swin Transformer 1

[5] Zhigang Tu et. al., Optical Flow for Video Super-Resolution: A Survey 1

[6] Yulun Zhang et. al., Aligned Structured Sparsity Learning for Efficient Image Super-Resolution 2

[7] https://www.nvidia.com/en-us/geforce/news/nvidia-dlss-2-0-a-big-leap-in-ai-rendering/ 1