Mostafa Sedky & Tommy Hosmer
Department of Mechanical Engineering
University of California, Berkeley

Low-Rank Approximations for Convolutional Layers in Super Resolution Applications

Abstract

Convolutional Neural Networks (CNNs) rely on convolution 4D tensor operations to discover features in image data. In this work, we propose using singular value decomposition (SVD) and canonical polyadic decomposition to convolution layer's 4D kernels to improve the inference speed and reduce the model size of the Enhanced Super-Resolution Generative Adversarial Network (ESGRAN) [1] with minimal loss in accuracy. ESRGAN enhances resolution of images, so we benchmark performance based on runtime of the network's inference, peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM). We also test the "image to column" (im2col) transformation, which converts the convolution to a matrix-multiplication and allows the use of BLAS3 GEMM routines.

Introduction

First, we compare FLOP counts to attain theoretical speed up. For direct convolution [2]:

$$\mathcal{O}(d^2NCXY)$$

where d is the dimensions of the kernel matrix, N is the number of output channels, C is the number of input channels, and X and Y are the dimensions of the data matrix. When we decompose the original 4D kernel into 2 lower rank kernels with rank K and complexities:

$$\mathcal{O}(dK(C+N)XY)$$

Acceleration is possible with K:

$$K < \frac{dNC}{N+C}$$

The low-rank approximation of a 4D kernel tensor via CP-decomposition is given by [3]:

$$A(i,j,N,C) = \sum_{k=1}^{K} A^x(i-x+\delta,k)A^y(j-y+\delta,k)A^n(n,k)A^c(c,k)$$

Each decomposed A in the sum is a component representing a matrix with sizes dxK, dxK, NxK, and CxK, respectively. For our approach, the complexity is:

$$\mathcal{O}(K(N+2d+C)XY)$$

We expect a theoretical speed up of ~d^2 when

$$R \approx \frac{NC}{N+C}$$

For the im2col, the FLOPs for the actual matrix-multiplication is

$$\mathcal{O}(2mnp)$$

for mxn and nxp matrices [4].

Methods

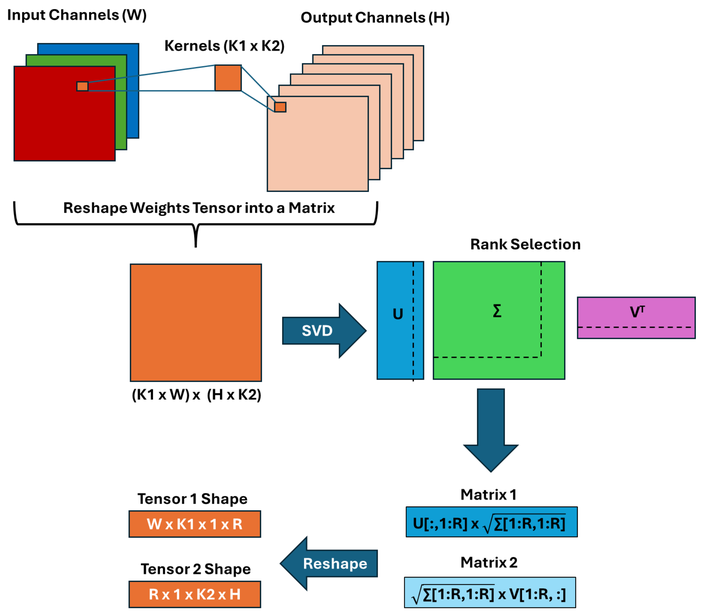

Figure 1. Using the SVD for convolutional layer compression to obtain two tensors to plug back into the ESRGAN architecture.

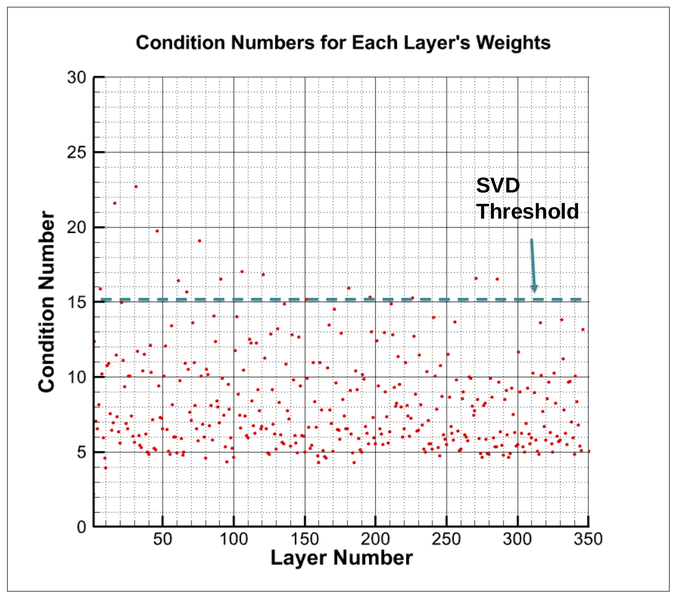

Figure 2. A plot of the condition numbers of each layer showing relatively small values for all layers and the threshold picked for marking a layer for decomposition.

Results

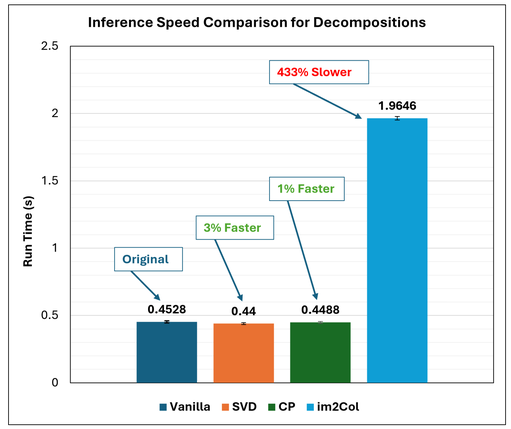

Figure 3. Inference speed comparison for each technique.

Figure 4. Resulting images for each technique.

Table 1. Image Quality Comparison

	Original (or im2col)	SVD	CP
PSNR	23.27	23.31	22.99
SSIM	0.64	0.68	0.63

Conclusions

The low rank approximations via the SVD and the CP-decomposition both slightly outperformed the direct convolution in the inference stage of the ESRGAN. The SVD decomposition even improved PSNR and SSIM values by removing parts of overdetermined layers. Pushing the SVD to use a lower threshold also slows down the inference speed. The im2col method is too memory intensive and is slower than the original implementation.

References

[1] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y. and Change Loy, C., 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops* (pp. 0-0).

[2] Tai, C., Xiao, T., Zhang, Y. and Wang, X., 2015. Convolutional neural networks with low-rank regularization. *arXiv preprint arXiv:1511.06067*.

[3] Lebedev, V., Ganin, Y., Rakhuba, M., Oseledets, I. and Lempitsky, V., 2014. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *arXiv preprint arXiv:1412.6553*.

[4] Demmel, J.W., 1997. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics.