Exploiting Elasticity in Tensor Ranks for Compressing Neural Networks

Jie Ran, Rui Lin, Hayden K.H. So, Graziano Chesi, Ngai Wong

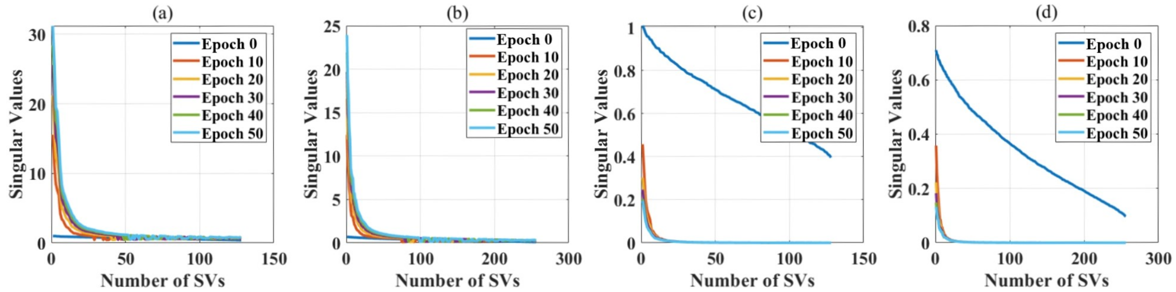
The University of Hong Kong

Motivation:

Elasticities in depth, width, kernel size and resolution have been explored in compressing deep neural networks (DNNs). Recognizing that the kernels in a convolutional neural network (CNN) are 4-way tensors, we further exploit a new elasticity dimension along the input-output channels, dynamically and globally searching for the reduced tensor ranks during training.

Experiments:

- Effect of regularizer on singular values of the parameters



PERFORMANCE COMPARISON ON IMAGENET

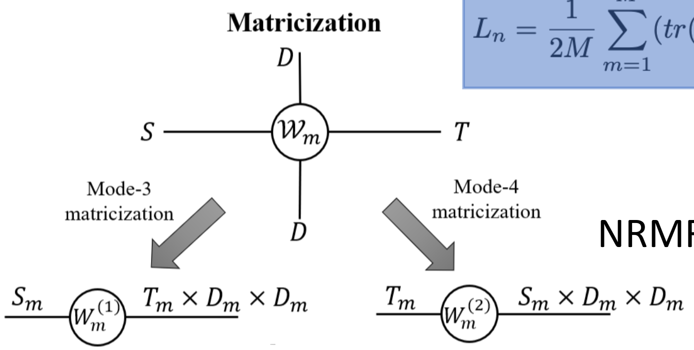
Model	Rank Selection	Top-1 Acc. (%)	Top-5 Acc. (%)	#Parameters
ResNet18	Base	69.76	89.08	$11.69M$
	VBMF	67.20	87.88	$7.50M$
	NRMF	67.27	87.7	$6.81M$

- Performances of ResNet18 on ImageNet

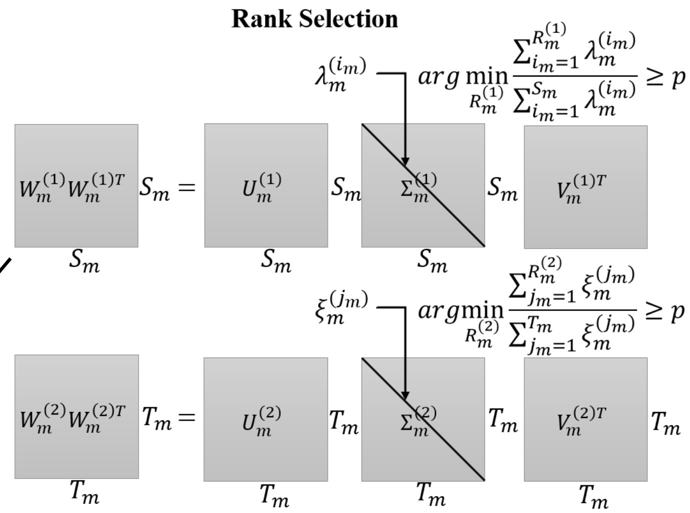
Our Method:

we introduce a nuclear-norm-based regularizer, and demonstrate how it can dynamically locate the ranks during training.

$$L_n = \frac{1}{2M}\sum_{m=1}^{M}\left(tr(\boldsymbol{W}_m^{(1)}\boldsymbol{W}_m^{(1)T}) + tr(\boldsymbol{W}_m^{(2)}\boldsymbol{W}_m^{(2)T})\right)$$



NRMF rank selection strategy



Reference:

S. Nakajima, M. Sugiyama, S. D. Babacan, and R. Tomioka, "Global analytic solution of fully-observed variational bayesian matrix factor- ization," *Journal of Machine Learning Research*, vol. 14, no. Jan, pp. 1–37, 2013.