# Exploiting Kernel Sparsity and Entropy for Interpretable CNN Compression

Yuchao Li[1#], Shaohui Lin[1#], Baochang Zhang[2], Jianzhuang Liu[3],
David Doermann[4], Yongjian Wu[5], Feiyue Huang[5], Rongrong Ji[1*]

[1]School of Information Science and Engineering, Xiamen University, China
[2]School of Automation Science and Electrical Engineering, Beihang University, China
[3]Huawei Noahs Ark Lab, China
[4]University at Buffalo, New York, USA
[5]BestImage, Tencent Technology (Shanghai) Co.,Ltd, China

xiamenlyc@gmail.com, shaohuilin007@gmail.com, bczhang@buaa.edu.cn, liu.jianzhuang@huawei.com,
doermann@buffalo.edu, littlekenwu@tencent.com, garyhuang@tencent.com, rrji@xmu.edu.cn

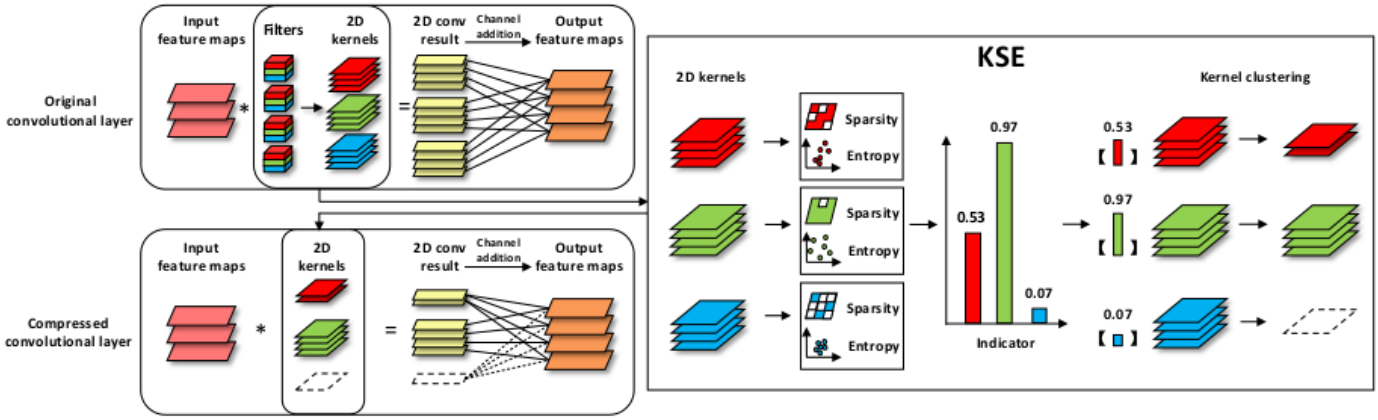


Figure 2. The framework of our method. The convolution operation is split into two parts, 2D convolution and channel fusion (addition). The 2D convolution is used to extract features from each input feature map, and the channel addition is used to obtain an output feature map by summing the intermediate results of the 2D convolution across all the input feature maps. In our KSE method, we first obtain the 2D kernels corresponding to an input feature map and calculate their sparsity and entropy as an indicator, which is further used to reduce the number of the 2D kernels by kernel clustering and generate a compact network.

为了分析神经网络特征图的冗余性问题，发现特征图的重要性取决于它的稀疏性和信息丰富度。但直接计算特征图的稀疏性和信息丰富度，需要巨大计算开销。为此，建立了特征图和其对应的二维卷积核之间的关系，通过计算卷积核的稀疏性和密度熵来表征对应特征图的重要程度。

特征图和 kernel

$$Y_n = \sum_{c=1}^{C} W_{n,c} * X_c,$$

For an input feature map X c , we call the set {W n,c } $^{N}_{n=1}$ the corresponding 2D kernels of X c .

Kernel Sparsity

$$s_c = \sum_{n=1}^{N} |W_{n,c}|.$$

Kernel Entropy

we define the kernel entropy to measure the complexity of the distribution of the 2D kernels:

$$e_c = -\sum_{i=1}^{N} \frac{dm(W_{i,c})}{d_c} log_2 \frac{dm(W_{i,c})}{d_c},$$

where $d_c = \sum_{i=1}^{N} dm(W_{i,c})$. 核的熵越小，2D kernels 分布越复杂，kernels 越具有多样性。

KSE Indicator

$$v_c = \frac{s_c}{1 + \alpha e_c},$$

where α is a parameter to control the balance between the sparsity and entropy, which is set to 1 in this work.

Kernel Clustering

$q_c$. Thus, the $c$-th input feature map generates $q_c$ centroids (new 2D kernels) $\{B_{i,c} \in \mathbb{R}^{K_h \times K_w}\}_{i=1}^{q_c}$ and an index set $\{I_{n,c} \in \{1, 2, ..., q_c\}\}_{n=1}^{N}$ to replace the original 2D ker-

For example, $I_{1,c} = 2$ denotes that the first original kernel is classified to the second cluster $B_{2,c}$.

When $q_c = 0$, the c-th input feature map is considered as unimportant, and all its corresponding kernels are pruned.  qc=N,相反。

channel addition

2D activation maps:

$$Z_{i,c} = B_{i,c} * X_c.$$

$$Y_n = \sum_{c=1}^{C} Z_{I_{n,c},c}.$$

For ResNet-50, we obtain 4.7× FLOPs reduction and 2.9× compression with only 0.35% Top-5 accuracy drop on ImageNet 2012.