Qualcomm INTERNSHIP

DAC: Data-free Automatic Acceleration of Convolutional Networks

Xin Li*, Shuai Zhang*, Bolan Jiang, Yingyong Qi, Mooi Choo Chuah, and Ning Bi
Qualcomm AI Research
Department of Computer Science and Engineering, Lehigh University
LEHIGH UNIVERSITY

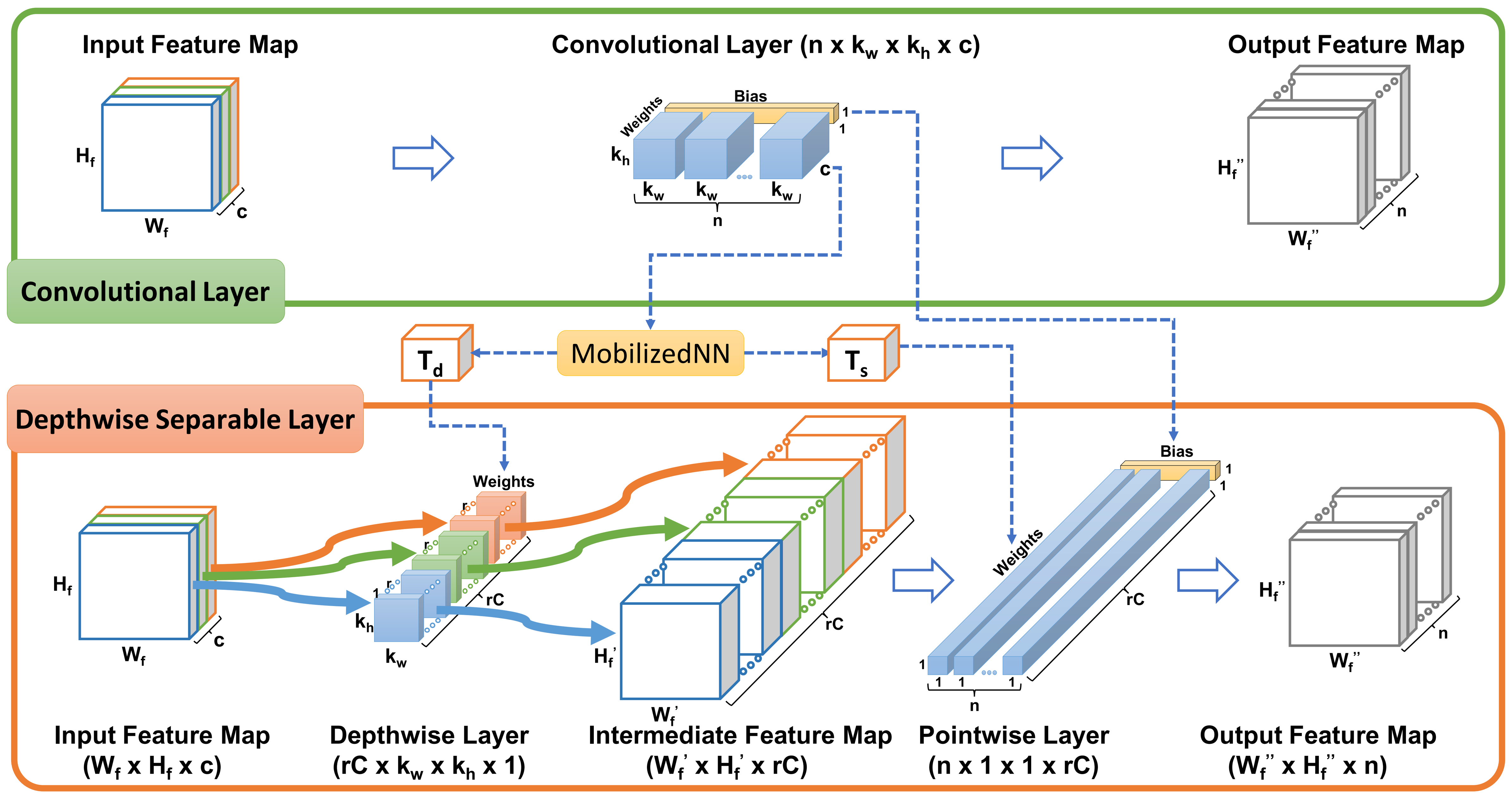
Introduction

We propose a novel Convolutional Networks Decomposition method that is capable of factorizing an ordinary convolutional layer into two layers with much fewer parameters. No training or any data is needed.

Our Contributions:

1. Design an efficient way to decompose a pre-trained model into a model with much **fewer parameters**.
2. The new model maintains high accuracy **without any data and training process.**
3. The proposed method **can be applied to any model** with convolutional layers.
4. It works pretty well in the task of **Image Classification, Object Detection, and Multi-person Pose Estimation**, etc..

Proposed Solution



Decomposition Algorithm

Input : Weights of a convolutional layer: $T \in \mathbb{R}^{n \times k_w \times k_h \times c}$;
Decomposition Rank: r.
Output: Weights of the depthwise layer: $Td \in \mathbb{R}^{rC \times k_w \times k_h \times 1}$;
Weights of the pointwise layer: $Ts \in \mathbb{R}^{n \times 1 \times 1 \times rC}$

begin
- $list\_d \in \mathbb{R}^{c \times r \times k_w \times k_h \times 1} \leftarrow \emptyset$
- $list\_s \in \mathbb{R}^{n \times 1 \times 1 \times r \times c} \leftarrow \emptyset$
- **for** $i \in c$ **do**
  - $T_i \leftarrow T[:, :, :, i] \in \mathbb{R}^{n \times k_w \times k_h}$
  - $M_i \leftarrow Reshape(Ti, (n, k_w \times k_h)) \in \mathbb{R}^{n \times k_w k_h}$
  - $D_i, S_i \leftarrow Decompose(M_i, r)$
  - $list\_d[i, :, :, :, :] \leftarrow D_i \in \mathbb{R}^{r \times k_w \times k_h \times 1}$
  - $list\_s[:, :, :, :, i] \leftarrow S_i \in \mathbb{R}^{n \times 1 \times 1 \times r}$
- $Td \leftarrow Reshape(list\_d, (r \times c, k_w, k_h, 1))$
- $Ts \leftarrow Reshape(list\_s, (n, 1, 1, r \times c))$

function Decompose(M, r)
begin
- $U, Sigma, V \leftarrow SVD(M)$
- $Ur \leftarrow U[:, :r] \in \mathbb{R}^{n \times r}$
- $Vr \leftarrow V[:r, :] \in \mathbb{R}^{r \times k_w k_h}$
- $Sr \leftarrow Sigma[:r, :r] \in \mathbb{R}^{r \times r}$
- $D \leftarrow Reshape(Vr, (r, k_w, k_h, 1))$
- $S \leftarrow Ur\ Sr$
- $S \leftarrow Reshape(S, (n, 1, 1, r))$
- **return** D, S

Experimental Results

Object Detection

SSD300 | Saved 20% Parameters | Saved 40% Parameters

Multi-Person Pose Estimation

Openpose | Saved 30% Parameters | Saved 50% Parameters

Image Classification

	ImageNet Top-5 Acc. (%)
VGG16	88.9

Parameters	Spatial Decom.	Ours
Save 40%	88.6	88.6
Save 50%	86.3	87.5
Save 60%	78.0	84.7

Object Detection

	SSD300	Ours (save 30%)
mAP(%)	76.5	75.3
Car	84.7	83.3
Person	77.5	75.8

Future Work

1. Test the computation time of decomposed models on mobile devices.
2. Evaluate the performance of our proposed method in other application fields, e.g., voice recognition, language translation, etc..
3. Explore the possibility of adapting the proposed method on other kinds of layers, e.g., 3D convolutional layer.
4. Improve decomposition performance using feature approximation.

Qualcomm Proprietary and Confidential
Qualcomm