

# DNN Feature Map Compression using Learned Representation over GF(2)

Denis Gudovskiy<sup>1</sup>[0000-0002-6829-6667], Alec Hodgkinson<sup>1</sup>[0000-0003-0416-6916],  
and Luca Rigazio<sup>1</sup>

Panasonic Beta Research Lab, Mountain View, CA, 94043, USA  
{denis.gudovskiy,alec.hodgkinson,luca.rigazio}@us.panasonic.com

**Abstract.** In this paper, we introduce a method to compress intermediate feature maps of deep neural networks (DNNs) to decrease memory storage and bandwidth requirements during inference. Unlike previous works, the proposed method is based on converting fixed-point activations into vectors over the smallest GF(2) finite field followed by nonlinear dimensionality reduction (NDR) layers embedded into a DNN. Such an end-to-end learned representation finds more compact feature maps by exploiting quantization redundancies within the fixed-point activations along the channel or spatial dimensions. We apply the proposed network architectures derived from modified SqueezeNet and MobileNetV2 to the tasks of ImageNet classification and PASCAL VOC object detection. Compared to prior approaches, the conducted experiments show a factor of 2 decrease in memory requirements with minor degradation in accuracy while adding only bitwise computations.

**Keywords:** feature map compression · dimensionality reduction · network quantization · memory-efficient inference

## 1 Introduction

Recent achievements of deep neural networks (DNNs) make them an attractive choice in many computer vision applications including image classification [7] and object detection [10]. The memory and computations required for DNNs can be excessive for low-power deployments. In this paper, we explore the task of minimizing the memory footprint of DNN feature maps during inference and, more specifically, finding a network architecture that uses minimal storage without introducing a considerable amount of additional computations or on-the-fly heuristic encoding-decoding schemes. In general, the task of feature map compression is tightly connected to input sparsity. The input sparsity can determine several different usage scenarios. This may lead to substantial decrease in memory requirements and overall inference complexity. First, pen sketches are spatially sparse and can be processed efficiently by recently introduced submanifold sparse CNNs [5]. Second, surveillance cameras with mostly static input contain temporal sparsity that can be addressed by Sigma-Delta networks [16]. A more general scenario presumes a dense input e.g. video frames from a high-resolution

camera mounted on a moving autonomous car. In this work, we address the latter scenario and concentrate on feature map compression in order to minimize memory footprint and bandwidth during DNN inference which might be prohibitive for high-resolution cameras.

We propose a method to convert intermediate fixed-point feature map activations into vectors over the smallest finite field called the Galois field of two elements ( $\text{GF}(2)$ ) or, simply, binary vectors followed by compression convolutional layers using a nonlinear dimensionality reduction (NDR) technique embedded into DNN architecture. The compressed feature maps can then be projected back to a higher cardinality representation over a fixed-point (integer) field using decompression convolutional layers. A layer fusion method allows to keep only the compressed feature maps for inference while adding only computationally inexpensive bitwise operations. Compression and decompression layers over  $\text{GF}(2)$  can be repeated within the proposed network architecture and trained in an end-to-end fashion. In brief, the proposed method resembles autoencoder-type [8] structures embedded into a base network that work over  $\text{GF}(2)$ . Binary conversion and compression-decompression layers are implemented in the Caffe [13] framework and are publicly available<sup>1</sup>.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 gives notation for convolutional layers, describes conventional fusion and NDR methods, and explains the proposed method including details about network training and the derived architectures from SqueezeNet [12] and MobileNetV2 [20]. Section 4 presents experimental results on ImageNet classification and PASCAL VOC object detection using SSD [14], memory requirements, and obtained compression rates.

## 2 Related Work

**Feature Map Compression Using Quantization** Unlike weight compression, surprisingly few papers consider feature map compression. This can most likely be explained by the fact that feature maps have to be compressed for every network input as opposed to offline weight compression. Previous feature map compression methods are primarily developed around the idea of representation *approximation* using a certain quantization scheme: fixed-point quantization [3,6], binary quantization [11,18,22,21], and power-of-two quantization [15]. The base floating-point network is converted to the approximate quantized representation and, then, the quantized network is retrained to restore accuracy. Such methods are inherently limited in finding more compact representations since the base architecture remains unchanged. For example, the dynamic fixed-point scheme typically requires around 8-bits of resolution to achieve baseline accuracy for state-of-the-art network architectures. At the same time, binary networks experience significant accuracy drops for large-scale datasets or compact (not over-parametrized) network architectures. Instead, our method can be

<sup>1</sup> <https://github.com/gudovskiy/fmap.compression>

considered in a narrow sense as a learned quantization using binary representation.

**Embedded NDR and Linear Layers** Another interesting approach is implicitly proposed by Iandola et al. [12]. Although the authors emphasized weight compression rather than feature map compression, they introduced NDR-type layers into network architecture that allowed to decrease not only the number of weights but also feature map sizes by a factor of 8, if one keeps only the outputs of so-called *squeeze* layers. The latter is possible because such network architecture does not introduce any additional convolution *recomputations* since *squeeze* layer computations with a  $1 \times 1$  kernel can be fused with the preceding *expand* layers.

Recently, a similar method was proposed for MobileNet architecture [20] with embedded *bottleneck* compression layers, which are, unlike SqueezeNet, linear. Authors view the task of compression aside from the rest of the network and argue that linear layers are more suitable for compression because no information is lost. While a small accuracy gain achieved for such layers compared to NDR layers in floating-point according to their experiments, we believe that it is due to a larger set of numbers ( $\mathbb{R}$  vs.  $\mathbb{R}_{\geq 0}$  for NDR with rectified linear unit (ReLU)). This is justified by our experiments using quantized models with limited set of available values. We consider the linear compression approach as a subset of nonlinear. Our work goes beyond previous approaches [12,20] by extending compression layers to work over GF(2) to find a more compact feature map representation.

**Hardware Accelerator Architectures** Horowitz [9] estimated that off-chip DRAM access requires approximately  $100 \times$  more power than local on-chip cache access. Therefore, currently proposed DNN accelerator architectures propose various schemes to decrease memory footprint and bandwidth. One obvious solution is to keep only a subset of intermediate feature maps at the expense of recomputing convolutions [1]. The presented fusion approach seems to be oversimplified but effective due to high memory access cost. Our approach is complementary to this work but proposes to keep only compressed feature maps with minimum additional computations.

Another recent work [17] exploits weight and feature map sparsity using a more efficient encoding for zeros. While this approach targets similar goals, it requires having high sparsity, which is often unavailable in the first and the largest feature maps. In addition, a special control and encoding-decoding logic decrease the benefits of this approach. In our work, compressed feature maps are stored in a dense form without the need of special control and encoding-decoding logic.

### 3 Feature Map Compression Methods

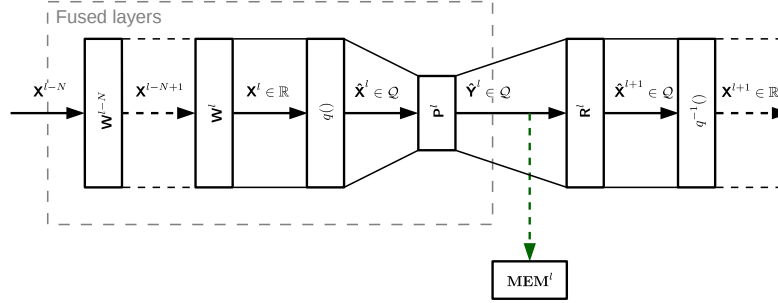
#### 3.1 Model and Notation

The input feature map of  $l$ th convolutional layer in commonly used DNNs can be represented by a tensor  $\mathbf{X}^{l-1} \in \mathbb{R}^{\acute{C} \times H \times W}$ , where  $\acute{C}$ ,  $H$  and  $W$  are the number of input channels, the height and the width, respectively. The input  $\mathbf{X}^{l-1}$  is convolved with a weight tensor  $\mathbf{W}^l \in \mathbb{R}^{C \times \acute{C} \times H_f \times W_f}$ , where  $C$  is the number of output channels,  $H_f$  and  $W_f$  are the height and the width of filter kernel, respectively. A bias vector  $\mathbf{b} \in \mathbb{R}^C$  is added to the result of convolution operation. Once all  $C$  channels are computed, an element-wise nonlinear function is applied to the result of the convolution operations. Then, the  $c$ th channel of the output tensor  $\mathbf{X}^l \in \mathbb{R}^{C \times H \times W}$  can be computed as

$$\mathbf{x}_c^l = g(\mathbf{W}_c^l * \mathbf{X}^{l-1} + \mathbf{b}_c), \quad (1)$$

where  $*$  denotes convolution and  $g()$  is some nonlinear function. In this paper, we assume  $g()$  is the most commonly used ReLU defined as  $g(x) = \max(0, x)$  such that all activations are non-negative.

#### 3.2 Conventional Methods



**Fig. 1.** The unified model of conventional methods: fusion allows to keep only *bottleneck* feature maps and quantization compresses each activation.

We formally describe previously proposed methods briefly reviewed in Section 2 using the unified model illustrated in Figure 1. To simplify notation, biases are not shown. Consider a network built using multiple convolutional layers and processed according to (1). Similar to Alwani et al. [1], calculation of  $N$  sequential layers can be fused together without storing intermediate feature maps  $\mathbf{X}^{l-N+1}, \dots, \mathbf{X}^{l-1}$ . For example, fusion can be done in a channel-wise fashion using memory buffers which are much smaller than the whole feature map. Then,

feature map  $\mathbf{X}^l \in \mathbb{R}$  can be quantized into  $\hat{\mathbf{X}}^l \in \mathcal{Q}$  using a nonlinear quantization function  $q(\cdot)$  where  $\mathcal{Q}$  is a finite field over integers. The quantization step may introduce a drop in accuracy due to imperfect approximation. The network can be further finetuned to restore some of the original accuracy [3,6]. The network architecture is not changed after quantization and feature maps can be compressed only up to a certain suboptimal bitwidth resolution.

The next step implicitly introduced by SqueezeNet [12] is to perform NDR using an additional convolutional layer. A mapping  $\hat{\mathbf{X}}^l \in \mathcal{Q}^{C \times H \times W} \rightarrow \hat{\mathbf{Y}}^l \in \mathcal{Q}^{\tilde{C} \times H \times W}$  can be performed using projection weights  $\mathbf{P}^l \in \mathbb{R}^{\tilde{C} \times C \times H_f \times W_f}$ , where the output channel dimension  $\tilde{C} < C$ . Then, only the compressed *bottleneck* feature map  $\hat{\mathbf{Y}}^l$  needs to be stored in the memory buffer. During the inverse steps, the compressed feature map can be projected back onto the higher-dimensional tensor  $\hat{\mathbf{X}}^{l+1} \in \mathcal{Q}$  using weights  $\mathbf{R}^l \in \mathbb{R}^{C \times \tilde{C} \times H_f \times W_f}$  and, lastly, converted back to  $\mathbf{X}^{l+1} \in \mathbb{R}$  using an inverse quantization function  $q^{-1}(\cdot)$ . In the case of a fully quantized network, the inverse quantization can be omitted.

In practice, the number of bits for the feature map quantization step depends on the dataset, network architecture and desired accuracy. For example, over-parameterized architectures like AlexNet may require only 1 or 2 bits for small-scale datasets (CIFAR-10, MNIST, SVHN), but experience significant accuracy drops for large-scale datasets like ImageNet. In particular, the modified AlexNet (with the first and last layers kept in full-precision) top-1 accuracy is degraded by 12.4% and 6.8% for 1-bit XNOR-Net [18] and 2-bit DoReFa-Net [22], respectively. At the same time, efficient network architectures e.g. [12] using NDR layers require 6-8 bits for the fixed-point quantization scheme on ImageNet and fail to work with lower precision activations. In this paper, we follow the path to select an efficient base network architecture and then introduce additional compression layers to obtain smaller feature maps as opposed to initially selecting an over-parametrized network architecture for quantization.

### 3.3 Proposed Method

**Representation over GF(2)** Consider a scalar  $x$  from  $\mathbf{X}^l \in \mathbb{R}$ . A conventional feature map quantization step can be represented as a scalar-to-scalar mapping or a nonlinear function  $\hat{x} = q(x)$  such that

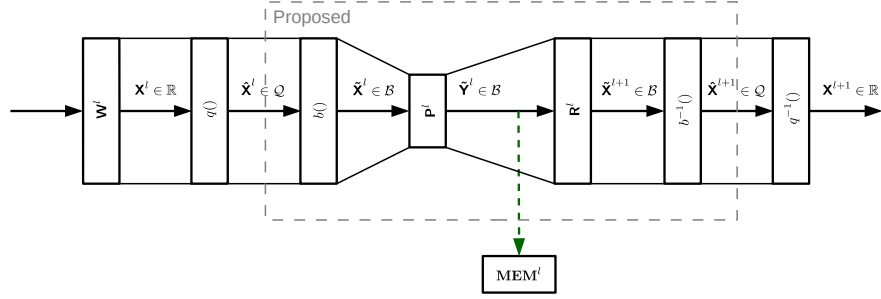
$$x \in \mathbb{R}^{1 \times 1} \xrightarrow{q(\cdot)} \hat{x} \in \mathcal{Q}^{1 \times 1} : \min \|x - \hat{x}\|_2, \quad (2)$$

where  $\hat{x}$  is the quantized scalar,  $\mathcal{Q}$  is the GF( $2^B$ ) finite field for fixed-point representation and  $B$  is the number of bits.

We can introduce a new  $\hat{x}$  representation by a linear binarization function  $b(\cdot)$  defined by

$$\hat{x} \in \mathcal{Q}^{1 \times 1} \xrightarrow{b(\cdot)} \tilde{\mathbf{x}} \in \mathcal{B}^{B \times 1} : \tilde{\mathbf{x}} = \mathbf{b} \otimes \hat{x}, \quad (3)$$

where  $\otimes$  is a bitwise AND operation, vector  $\mathbf{b} = [2^0, 2^1, \dots, 2^{B-1}]^T$  and  $\mathcal{B}$  is GF(2) finite field.



**Fig. 2.** Scheme of the proposed method: binarization is added and compression happens in GF(2) followed by inverse operations.

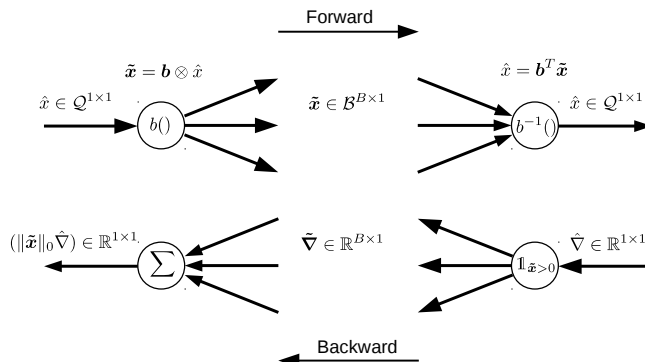
An inverse linear function  $b^{-1}()$  can be written as

$$\tilde{\mathbf{x}} \in \mathcal{B}^{B \times 1} \xrightarrow{b^{-1}()} \hat{\mathbf{x}} \in \mathcal{Q}^{1 \times 1} : \hat{\mathbf{x}} = \mathbf{b}^T \tilde{\mathbf{x}} = \mathbf{b}^T \mathbf{b} \otimes \hat{\mathbf{x}} = (2^B - 1) \otimes \hat{\mathbf{x}}. \quad (4)$$

Equations (3)-(4) show that a scalar over a higher cardinality finite field can be linearly converted to and from a vector over a finite field with two elements. Based on these derivations, we propose a feature map compression method shown in Figure 2. Similar to [3], we quantize activations to obtain  $\tilde{\mathbf{X}}^l$  and, then, apply transformation (3). The resulting feature map can be represented as  $\tilde{\mathbf{X}}^l \in \mathcal{B}^{B \times C \times H \times W}$ . For implementation convenience, a new bit dimension can be concatenated along channel dimension resulting in the feature map  $\tilde{\mathbf{X}}^l \in \mathcal{B}^{BC \times H \times W}$ . Next, a single convolutional layer using weights  $\mathbf{P}^l$  or a sequence of layers with  $\mathbf{P}_i^l$  weights can be applied to obtain a compressed representation over GF(2). Using the fusion technique, only the compressed feature maps  $\tilde{\mathbf{Y}}^l \in \mathcal{B}$  need to be stored in memory during inference. Non-compressed feature maps can be processed using small buffers e.g. in a sequential channel-wise fashion. Lastly, the inverse function  $b^{-1}()$  from (4) using convolutional layers  $\mathbf{R}_i^l$  and inverse of quantization  $q^{-1}()$  undo the compression and quantization steps.

**Learning over GF(2)** The graph model shown in Figure 3 explains details about the inference (forward pass) and backpropagation (backward pass) phases of the newly introduced functions. The inference pass represents (3)-(4) as explained above. The backpropagation pass may seem non-obvious at first glance. One difficulty related to the quantized network is that the quantization function itself is not differentiable. Many studies e.g. [3] show that a mini-batch-averaged floating-point gradient practically works well assuming a quantized forward pass. The new functions  $b()$  and  $b^{-1}()$  can be represented as gates that make hard decisions similar to ReLU. The gradient of  $b^{-1}()$  can then be calculated using results of Bengio et al. [2] as

$$\hat{\nabla} \in \mathbb{R}^{1 \times 1} \xrightarrow{b^{-1}()} \tilde{\nabla} \in \mathbb{R}^{B \times 1} : \tilde{\nabla} = \mathbb{1}_{\hat{\mathbf{x}} > 0} \nabla. \quad (5)$$



**Fig. 3.** Forward and backward passes during inference and backpropagation.

Lastly, the gradient of  $b(\cdot)$  is just a scaled sum of the gradient vector calculated by

$$\tilde{\nabla} \in \mathbb{R}^{B \times 1} \xrightarrow{b(\cdot)} \hat{\nabla} \in \mathbb{R}^{1 \times 1} : \hat{\nabla} = \mathbb{1}^T \tilde{\nabla} = \mathbb{1}^T \mathbb{1}_{\tilde{x} > 0} \tilde{\nabla} = \|\tilde{x}\|_0 \tilde{\nabla}, \quad (6)$$

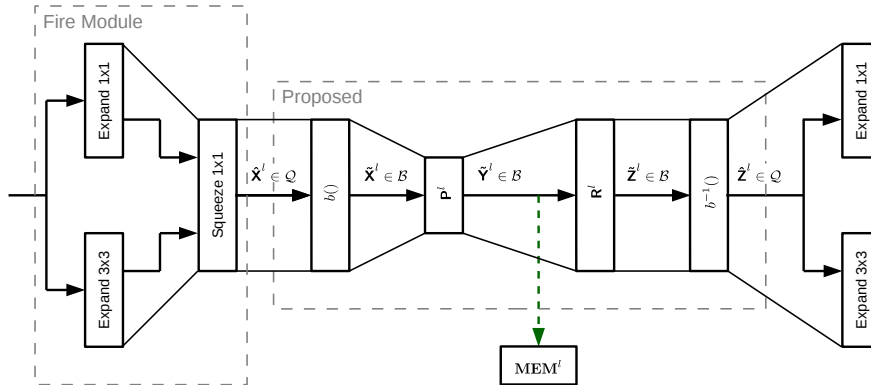
where  $\|\tilde{x}\|_0$  is a gradient scaling factor that represents the number of nonzero elements in  $\tilde{x}$ . Practically, the scaling factor can be calculated based on statistical information only once and used as a static hyperparameter for gradient normalization.

Since the purpose of the network is to learn and keep only the smallest  $\tilde{\mathbf{Y}}^l$ , the choice of  $\mathbf{P}^l$  and  $\mathbf{R}^l$  initialization is important. Therefore, we can initialize these weight tensors by an identity function that maps the non-compressed feature map to a truncated compressed feature map and vice versa. That provides a good starting point for training. At the same time, other initializations are possible e.g. noise sampled from some distribution studied by [2] can be added as well.

**Network Architecture** A base network architecture can be selected among existing networks with the embedded *bottleneck* layers e.g. SqueezeNet [12] or MobileNetV2 [20]. We explain how a base network architecture can be modified according to Section 3.3 using SqueezeNet example.

The latter network architecture consists of a sequence of *fire* modules where each module contains two concatenated *expand* layers and a *squeeze* layer illustrated in Figure 4. The *squeeze* layers perform NDR over the field of real or, in case of the quantized model, integer numbers. Specifically, the size of the concatenated *expand*  $1 \times 1$  and *expand*  $3 \times 3$  layers is compressed by a factor of 8 along channel dimension by *squeeze*  $1 \times 1$  layer. Activations of only the latter one can be stored during inference using the fusion method. According to the analysis presented in Gysel et al. [6], activations quantized to 8-bit integers do not experience significant accuracy drop.

The quantized *squeeze* layer feature map can be converted to its binary representation following Figure 2. Then, the additional compression rate is defined



**Fig. 4.** SqueezeNet architecture example: *fire* module is extended by the proposed method.

by selecting parameters of  $\mathbf{P}_i^l$ . In the simplest case, only a single NDR layer needs to be introduced with the weights  $\mathbf{P}^l$ . In general, a number of NDR layers can be added with  $1 \times 1$ ,  $3 \times 3$  and other kernels with or without pooling at the expense of increased computational cost. For example,  $1 \times 1$  kernels allow to learn optimal quantization and to compensate redundancies along channel dimension only. But  $3 \times 3$  kernels can address spatial redundancies and, while being implemented with stride 2 using convolutional-deconvolutional layers, decrease feature map size along spatial dimensions.

MobileNetV2 architecture can be modified using the proposed method with few remarks. First, its *bottleneck* layers compress feature maps by a  $1 \times 1$  kernel with variable compression factor from 2 to 6 unlike fixed factor in SqueezeNet. Second, linear compression layers either have to be turned into NDR layers by adding ReLUs or implementation of compression-decompression layers needs to support negative integers. In practice, the former approach might be less cumbersome.

## 4 Experiments

### 4.1 ImageNet Classification

We implemented the new binarization layers from Section 3 as well as quantization layers using modified [6] code in the Caffe [13] framework. The latter code is modified to accurately support binary quantization during inference and training. SqueezeNetV1.1 and MobilenetV2 are selected as a base floating-point network architectures, and their pretrained weights were downloaded from the publicly available sources<sup>2,3</sup>.

<sup>2</sup> <https://github.com/DeepScale/SqueezeNet>

<sup>3</sup> <https://github.com/shicai/MobileNet-Caffe>



**Table 1.** SqueezeNet ImageNet accuracy: A - *fire2,3/squeeze* feature maps and W - weights.

Model	W size, MB	A size, KB	Top-1 Accuracy, %	Top-5 Accuracy, %
fp32	4.7	392.0	58.4	81.0
Quantized				
uint8	4.7	98.0	58.6(58.3)	81.1(81.0)
uint6	4.7	73.5	57.8(55.5)	80.7(78.7)
uint4	4.7	49.0	54.9(18.0)	78.3(34.2)
Proposed: $b() \rightarrow 1 \times 1 \rightarrow 1 \times 1 \rightarrow b^{-1}()$				
uint6	5.0	73.5	58.8	81.3
uint4	4.9	49.0	57.3	80.0
Proposed: $b() \rightarrow 3 \times 3 / 2 \rightarrow 3 \times 3 * 2 \rightarrow b^{-1}()$				
uint8	7.6	24.5	54.1	77.4
uint6	6.9	18.4	53.8	77.2

**SqueezeNet Architecture** We compress the *fire2/squeeze* and *fire3/squeeze* layers which consume 80% of total network memory footprint when fusion is applied due to high spatial dimensions. The input to the network has a resolution of  $227 \times 227$ , and the weights are all floating-point.

The quantized and compressed models are retrained for 100,000 iterations with a mini-batch size of 1024 on the ImageNet [19] (ILSVRC2012) training dataset, and SGD solver with a step-policy learning rate starting from  $1e-3$  divided by 10 every 20,000 iterations. Although this large mini-batch size was used by the original model, it helps the quantized and compressed models to estimate gradients as well. The compressed models were derived and retrained iteratively from the 8-bit quantized model. Table 1 reports top-1 and top-5 inference accuracies of 50,000 images from ImageNet validation dataset.

According to Table 1, the retrained quantized models experience -0.2%, 0.6% and 3.5% top-1 accuracy drops for 8-bit, 6-bit and 4-bit quantization, respectively. For comparison, the quantized models without retraining are shown in parentheses. The proposed compression method using  $1 \times 1$  kernels allows us to restore corresponding top-1 accuracy by 1.0% and 2.4% for 6-bit and 4-bit versions at the expense of a small increase in the number of weights and bit-wise convolutions. Moreover, we evaluated a model with a convolutional layer followed by a deconvolutional layer both with a  $3 \times 3$  stride 2 kernel at the expense of a 47% increase in weight size for 6-bit activations. That allowed us to decrease feature maps in spatial dimensions by exploiting local spatial quantization redundancies. Then, the feature map size is further reduced by a factor of 4, while top-1 accuracy dropped by 4.3% and 4.6% for 8-bit and 6-bit activations, respectively. A comprehensive comparison for fully quantized models with the state-of-the-art binary and ternary networks is given below.

**Table 2.** MobileNetV2 ImageNet accuracy: A - *conv2\_1/linear* feature maps and W - weights.

Model	W size, MB	A size, KB	Top-1 Accuracy, %	Top-5 Accuracy, %
fp32	13.5	784.0	71.2	90.2
Quantized				
int9	13.5	220.5	71.5(71.2)	89.9(90.2)
int7	13.5	171.5	71.5(68.5)	89.8(88.4)
int5	13.5	122.5	70.9(7.3)	89.4(17.8)
Modified: ReLU nonlinearity added				
uint8	13.5	196.0	71.6	90.0
Proposed: $b() \rightarrow 1 \times 1 \rightarrow 1 \times 1 \rightarrow b^{-1}()$				
uint6	13.7	147.0	70.9	89.4
uint4	13.6	98.0	69.5	88.5
Proposed: $b() \rightarrow 2 \times 2 / 2 \rightarrow 2 \times 2 * 2 \rightarrow b^{-1}()$				
uint8	14.2	49.0	66.6	86.9
uint6	14.0	36.8	66.7	86.9

**MobileNetV2 Architecture** We compress the *conv2\_1/linear* feature map which is nearly  $3\times$  larger than any other *bottleneck* layer feature map. The same training hyperparameters are used as in previous experiment setup with few differences. The number of iterations is 50,000 with proportional change in learning rate policy. Second, we add a ReLU layer after *conv2\_1/linear* to be compatible with the current implementation of the compression method. Hence, the *conv2\_1/linear* feature map contains signed integers in the original model and unsigned integers in the modified one. Lastly, we found that batch normalization layers cause some instability to training process. Therefore, normalization and scaling parameters are fixed and merged into weights and biases of convolutional layers. Then, the modified model was retrained from the original one.

According to Table 2, the original (without ReLU) quantized models after retraining experience -0.3%, -0.3% and 0.3% top-1 accuracy drops for 9-bit, 7-bit and 5-bit quantization, respectively. For comparison, the quantized models without retraining are shown in parentheses. Surprisingly, quantized MobileNetV2 is resilient to smaller bitwidths with only 0.6% degradation for 5-bit model compared to 9-bit one. The modified (with ReLU nonlinearity) 8-bit model outperforms all the original quantized model by 0.1%, even the one with more bits, unlike results reported by [20] for floating-point models. Hence, the conclusions about advantages of linear compression layers could be reconsidered in finite (integer) field. Accuracies of the proposed models using  $1 \times 1$  kernels are on par with the conventional quantization approaches. Most likely, lack of batch normalization layers does not allow to increase accuracy which should be investigated. The proposed models with a convolutional-deconvolutional layers and  $2 \times 2$  stride 2 kernel compress feature maps by another factor of 2 with around

**Table 3.** ImageNet accuracy: W - weights, A - feature maps, F - fusion, Q - quantization, C - compression.

Model	Base Network	W, bits	W size, MB	A, bits	A size, KB	Top-1 Acc., %	Top-5 Acc., %
AlexNet	-	32	232	32	3053.7	56.6	79.8
AlexNet	-	32	232	6	572.6	55.8	79.2
XNOR-Net	AlexNet	1 <sup>1</sup>	22.6	1	344.4 <sup>2</sup>	44.2	69.2
XNOR-Net	ResNet-18	1 <sup>1</sup>	3.34	1	1033.0 <sup>2</sup>	51.2	73.2
DoReFa-Net	AlexNet	1 <sup>1</sup>	22.6	1	95.4	43.6	-
DoReFa-Net	AlexNet	1 <sup>1</sup>	22.6	2	190.9	49.8	-
DoReFa-Net	AlexNet	1 <sup>1</sup>	22.6	4	381.7	53.0	-
Tang'17	AlexNet	1 <sup>3</sup>	7.43	2	190.9	46.6	71.1
Tang'17	NIN-Net	1 <sup>3</sup>	1.23	2	498.6	51.4	75.6
The proposed models							
SqueezeNet	-	32	4.7	32	12165.4	58.4	81.0
F+Q	SqueezeNet	8	1.2	8	189.9	58.3	80.8
F+Q+C(1×1)	SqueezeNet	8	1.2	6(8) <sup>4</sup>	165.4	58.3(58.8) <sup>5</sup>	81.0(81.3) <sup>5</sup>
F+Q+C(1×1)	SqueezeNet	8	1.2	4(8) <sup>4</sup>	140.9	56.6(57.3) <sup>5</sup>	79.7(80.0) <sup>5</sup>
F+Q+C(3×3s2)	SqueezeNet	8	1.9	8(8) <sup>4</sup>	116.4	53.5(54.1) <sup>5</sup>	76.7(77.4) <sup>5</sup>
F+Q+C(3×3s2)	SqueezeNet	8	1.7	6(8) <sup>4</sup>	110.3	53.0(53.8) <sup>5</sup>	76.8(77.2) <sup>5</sup>

<sup>1</sup> Weights are not binarized for the first and the last layer.

<sup>2</sup> Activation size estimates are based on 8-bit assumption since it is not clear from [18] whether the activations were binarized or not for the first and the last layer.

<sup>3</sup> Weights are not binarized for the first layer.

<sup>4</sup> Number of bits for the compressed *fire2,3/squeeze* layers and, in parentheses, for the rest of layers.

<sup>5</sup> For comparison, accuracy in parentheses represents result for the corresponding model in Table 1.

4.5% accuracy degradation and 5% increase in weight size. A comparison in the object detection section further compares  $2 \times 2$  and  $3 \times 3$  stride 2 kernels and concludes that the former one is preferable due to accuracy and size.

**Comparison with Binary and Ternary State-of-the-Art** We compare recently reported ImageNet results for low-precision networks as well as several configurations of the proposed approach for which, unlike previous experiments, all weights and activations are quantized. Most of the works use the over-parametrized AlexNet architecture while ours is based on the SqueezeNet architecture in this comparison. Table 3 shows accuracy results for base networks as well as their quantized versions. Binary XNOR-Net [18] estimates based on AlexNet as well as ResNet-18. DoReFa-Net [22] is more flexible and can adjust the number of bits for weights and activations. Since its accuracy is limited by the number of activation bits, we present three cases with 1-bit, 2-bit, and

4-bit activations. The most recent work [21] solves the problem of binarizing the last layer weights, but weights of the first layer are full-precision. Overall, AlexNet-based low-precision networks achieve 43.6%, 49.8%, 53.0% top-1 accuracy for 1-bit, 2-bit and 4-bit activations, respectively. Around 70% of the memory footprint is defined by the first two layers of AlexNet. The fusion technique is difficult in such architectures due to large kernel sizes ( $11\times 11$  and  $5\times 5$  for AlexNet) which can cause extra recomputations. Thus, activations require 95.4KB, 190.0KB and 381.7KB of memory for 1-bit, 2-bit and 4-bit models, respectively. The NIN-based network from [21] with 2-bit activations achieves 51.4% top-1 accuracy, but its activation memory is larger than AlexNet due to late pooling layers.

The SqueezeNet-based models in Table 3 are finetuned from the corresponding models in Table 1 for 40,000 iterations with a mini-batch size of 1024, and SGD solver with a step-policy learning rate starting from  $1e-4$  divided by 10 every 10,000 iterations. The model with fusion and 8-bit quantized weights and activations, while having an accuracy similar to floating-point model, outperforms the state-of-the-art networks in terms of weight and activation memory. The proposed four models from Table 1 further decrease activation memory by adding compression-decompression layers to *fire2,3* modules. This step allowed us to shrink memory from 189.9KB to 165.4KB, 140.9KB, 116.4KB and 110.3KB depending on the compression configuration. More compression is possible, if the proposed approach is applied to other *squeeze* layers.

## 4.2 PASCAL VOC Object Detection using SSD

**Accuracy Experiments** We evaluate object detection using the Pascal VOC [4] dataset which is a more realistic application for autonomous cars where the high-resolution cameras emphasize feature map compression benefits. The VOC2007 test dataset contains 4,952 images and a training dataset of 16,551 images is a union of VOC2007 and VOC2012. We adopted the SSD512 model [14] for the proposed architecture. SqueezeNet pretrained on ImageNet is used as a feature extractor instead of the original VGG-16 network. This reduces number of parameters and overall inference time by a factor of 4 and 3, respectively. The original VOC images are rescaled to  $512\times 512$  resolution. As with ImageNet experiments, we generated several models for comparisons: a base floating-point model, quantized models, and compressed models. We apply quantization and compression to the *fire2/squeeze* and *fire3/squeeze* layers which represent, if the fusion technique is applied, more than 80% of total feature map memory due to their large spatial dimensions. Typically, spatial dimensions decrease quadratically because of max pooling layers compared to linear growth in the depth dimension. The compressed models are derived from the 8-bit quantized model, and both are retrained for 10,000 iterations with a mini-batch size of 256 using SGD solver with a step-policy learning rate starting from  $1e-3$  divided by 10 every 2,500 iterations.

Table 4 presents mean average precision (mAP) results for SqueezeNet-based models as well as size of the weights and feature maps to compress. The 8-bit quantized model with retraining drops accuracy by less than 0.04%, while 6-bit,

**Table 4.** VOC2007 SSD512 accuracy: A - *fire2,3/squeeze* feature maps and W - weights.

Model	W size, MB	A size, KB	mAP, %
fp32	23.7	2048	68.12
Quantized			
uint8	23.7	512	68.08(68.04)
uint6	23.7	384	67.66(67.14)
uint4	23.7	256	65.92(44.13)
uint2	23.7	128	55.86(0.0)
Proposed $b() \rightarrow 1 \times 1 \rightarrow 1 \times 1 \rightarrow b^{-1}()$			
uint6	23.9	384	68.17
uint4	23.8	256	65.42
Proposed: $b() \rightarrow 3 \times 3 / 2 \rightarrow 3 \times 3 * 2 \rightarrow b^{-1}()$			
uint8	26.5	128	63.53
uint6	25.9	96	62.22
Proposed: $b() \rightarrow 2 \times 2 / 2 \rightarrow 2 \times 2 * 2 \rightarrow b^{-1}()$			
uint8	24.9	128	64.39
uint6	24.6	96	62.09

4-bit and 2-bit models decrease accuracy by 0.5%, 2.2% and 12.3%, respectively. For reference, mAPs for the quantized models without retraining are shown in parentheses. Using the proposed compression-decompression layers with a  $1 \times 1$  kernel, mAP for the 6-bit model is increased by 0.5% and mAP for the 4-bit is decreased by 0.5%. We conclude that compression along channel dimension is not beneficial for SSD unlike ImageNet classification either due to low quantization redundancy in that dimension or the choice of hyperparameters e.g. mini-batch size. Then, we evaluate the models with spatial-dimension compression which is intuitively appealing for high-resolution images. Empirically, we found that a  $2 \times 2$  kernel with stride 2 performs better than a corresponding  $3 \times 3$  kernel while requiring less parameters and computations. According to Table 4, an 8-bit model with  $2 \times 2$  kernel and downsampling-upsampling layers achieves 1% higher mAP than a model with  $3 \times 3$  kernel and only 3.7% lower than the base floating-point model.

**Memory Requirements** Table 5 summarizes memory footprint benefits for the evaluated SSD models. Similar to the previous section, we consider only the largest feature maps that represent more than 80% of total activation memory. Assuming that the input frame is stored separately, the fusion technique allows to compress feature maps by a factor of 19. Note that no additional recomputations are needed. Second, conventional 8-bit and 4-bit fixed-point models decrease the size of feature maps by a factor of 4 and 8, respectively. Third, the proposed model with  $2 \times 2$  stride 2 kernel gains another factor of 2 compression compared

**Table 5.** SSD512 memory requirements: A - feature map, F - fusion, Q - quantization, C - compression ( $2 \times 2s^2$ ).

A size, KB	Base, fp32	F, fp32	F+Q, uint8	F+Q, uint4	F+Q+C, uint8
input (int8)	768	768	768	768	768
conv1	16384	0	0	0	0
mpool1	4096	0	0	0	0
fire2,3/squeeze	2048	2048	512	256	128
fire2,3/expand	16384	0	0	0	0
Total	38912	2048	512	256	128
mAP, %	68.12	68.12	68.08	65.92	64.39
Compression	-	19 $\times$	76 $\times$	152 $\times$	304 $\times$

to 4-bit fixed-point model with only 1.5% degradation in mAP. This result is similar to ImageNet experiments which showed relatively limited compression gain along channel dimension only. At the same time, learned quantization along combined channel and spatial dimensions pushes further compression gain. In total, the memory footprint for this feature extractor is reduced by two orders of magnitude.

## 5 Conclusions

We introduced a method to decrease memory storage and bandwidth requirements for DNNs. Complementary to conventional approaches that use layer fusion and quantization, we presented an end-to-end method for learning feature map representations over  $GF(2)$  within DNNs. Such a binary representation allowed us to compress network feature maps in a higher-dimensional space using autoencoder-inspired layers embedded into a DNN along channel and spatial dimensions. These compression-decompression layers can be implemented using conventional convolutional layers with bitwise operations. To be more precise, the proposed representation traded cardinality of the finite field with the dimensionality of the vector space which makes possible to learn features at the binary level. The evaluated compression strategy for inference can be adopted for GPUs, CPUs or custom accelerators. Alternatively, existing binary networks can be extended to achieve higher accuracy for emerging applications such as object detection and others.

## References

1. Alwani, M., Chen, H., Ferdman, M., Milder, P.A.: Fused-layer CNN accelerators. In: MICRO. pp. 1–12 (Oct 2016)
2. Bengio, Y., Lonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)

3. Courbariaux, M., Bengio, Y., David, J.: Training deep neural networks with low precision multiplications. In: ICLR (May 2015)
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. IJCV **88**(2), 303–338 (Jun 2010)
5. Graham, B., van der Maaten, L.: Submanifold sparse convolutional networks. arXiv preprint arXiv:1706.01307 (2017)
6. Gysel, P., Motamedi, M., Ghiasi, S.: Hardware-oriented approximation of convolutional neural networks. In: ICLR (May 2016)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
8. Hinton, G., Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (Jul 2006)
9. Horowitz, M.: Computing’s energy problem (and what we can do about it). In: ISSCC. pp. 10–14 (Feb 2014)
10. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K.: Speed/accuracy trade-offs for modern convolutional object detectors. In: CVPR (Jul 2017)
11. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks. In: NIPS, pp. 4107–4115 (2016)
12. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv preprint arXiv:1602.07360 (2016)
13. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
14. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: ECCV (Oct 2016)
15. Miyashita, D., Lee, E.H., Murmann, B.: Convolutional neural networks using logarithmic data representation. arXiv preprint arXiv:1603.01025 (2016)
16. O’Connor, P., Welling, M.: Sigma delta quantized networks. In: ICLR (Apr 2017)
17. Parashar, A., Rhu, M., Mulkara, A., Puglielli, A., Venkatesan, R., Khailany, B., Emer, J., Keckler, S.W., Dally, W.J.: SCNN: An accelerator for compressed-sparse convolutional neural networks. In: ISCA. pp. 27–40 (2017)
18. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: XNOR-net: Imagenet classification using binary convolutional neural networks. arXiv preprint arXiv:1603.05279 (2016)
19. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3), 211–252 (2015)
20. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR (Jun 2018)
21. Tang, W., Hua, G., Wang, L.: How to train a compact binary neural network with high accuracy? In: AAAI (2017)
22. Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., Zou, Y.: DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv preprint arXiv:1606.06160 (2016)