BiDet: An Efficient Binarized Object Detector

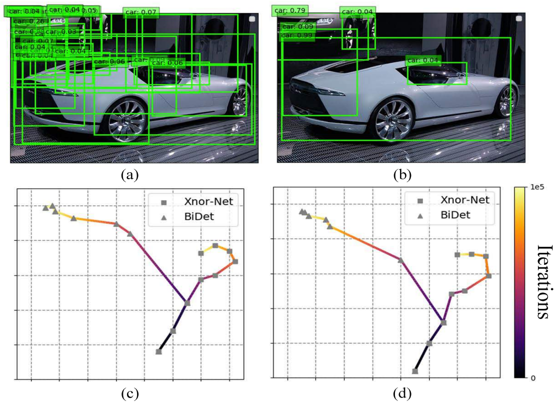
Ziwei Wang[1,2,3], Ziyi Wu[1], Jiwen Lu[1,2,3,*], Jie Zhou[1,2,3,4]
[1] Department of Automation, Tsinghua University, China
[2] State Key Lab of Intelligent Technologies and Systems, China
[3] Beijing National Research Center for Information Science and Technology, China
[4] Tsinghua Shenzhen International Graduate School, Tsinghua University, China
{wang-zw18, wuzy17}@mails.tsinghua.edu.cn; {lujiwen,jzhou}@tsinghua.edu.cn

Abstract

In this paper, we propose a binarized neural network learning method called BiDet for efficient object detection. Conventional network binarization methods directly quantize the weights and activations in one-stage or two-stage detectors with constrained representational capacity, so that the information redundancy in the networks causes numerous false positives and degrades the performance significantly. On the contrary, our BiDet fully utilizes the representational capacity of the binary neural networks for object detection by redundancy removal, through which the detection precision is enhanced with alleviated false positives. Specifically, we generalize the information bottleneck (IB) principle to object detection, where the amount of information in the high-level feature maps is constrained and the mutual information between the feature maps and object detection is maximized. Meanwhile, we learn sparse object priors so that the posteriors are concentrated on informative detection prediction with false positive elimination. Extensive experiments on the PASCAL VOC and COCO datasets show that our method outperforms the state-of-the-art binary neural networks by a sizable margin.[1]

1. Introduction

Convolutional neural network (CNN) based object detectors [7, 10, 22, 24, 32] have achieved state-of-the-art performance due to the strong discriminative power and generalization ability. However, the CNN based detection methods require massive computation and storage resources to achieve ideal performance, which limits their deployment on mobile devices. Therefore, it is desirable to develop detectors with lightweight architectures and few parameters.



(a) (b)

(c) (d)

Figure 1. An example of the predicted objects with the binarized SSD detector on PASCAL VOC. (a) and (b) demonstrate the detection results via Xnor-Net and the proposed BiDet, where the false positives are significantly reduced in our method. (c) and (d) reveal the information plane dynamics for the training set and test set respectively, where the horizontal axis means the mutual information between the high-level feature map and input and the vertical axis represents the mutual information between the object and the feature map. Compared with Xnor-Net, our method removes the redundant information and fully utilizes the network capacity to achieve higher performance. (best viewed in color).

To reduce the complexity of deep neural networks, several model compression methods have been proposed including pruning [12, 27, 45], low-rank decomposition [16, 20, 28], quantization [9, 19, 41], knowledge distillation [3, 40, 42], architecture design [29, 34, 44] and architecture search [37, 43]. Among these methods, network quantization reduces the bitwidth of the network parameters and activations for efficient inference. In the extreme cases, binarizing weights and activations of neural networks decreases the storage and computation cost by 32× and 64× respectively. However, deploying binary neural networks with constrained representational capacity in object detection causes numerous false positives due to the information redundancy in the networks.

[*] Corresponding author
[1] Code: https://github.com/ZiweiWangTHU/BiDet.git

In this paper, we present a BiDet method to learn binarized neural networks including the backbone part and the detection part for efficient object detection. Unlike existing methods which directly binarize the weights and activations in one-stage or two-stage detectors, our method fully utilizes the representational capacity of the binary neural networks for object detection via redundancy removal, so that the detection precision is enhanced with false positive elimination. More specifically, we impose the information bottleneck (IB) principle on binarized object detector learning, where we simultaneously limit the amount of information in high-level feature maps and maximize the mutual information between object detection and the learned feature maps. Meanwhile, the learned sparse object priors are utilized in IB, so that the posteriors are enforced to be concentrated on informative prediction and the uninformative false positives are eliminated. Figure 1 (a) and (b) show an example of predicted positives obtained by Xnor-Net [30] and our BiDet respectively, where the false positives are significantly reduced in the latter. Figure 1 (c) and (d) depict the information plane dynamics for the training and test sets respectively, where our BiDet removes the information redundancy and fully utilizes the representational power of the networks. Extensive experiments on the PASCAL VOC [6] and COCO [23] datasets show that our BiDet outperforms the state-of-the-art binary neural networks in object detection across various architectures. Moreover, BiDet can be integrated with other compact object detectors to acquire faster speedup and less storage. Our contributions include:

- To the best of our knowledge, we propose the first binarized networks containing the backbone and detection parts for efficient object detection.

- We employ the IB principle for redundancy removal to fully utilize the capacity of binary neural networks and learn the sparse object priors to concentrate posteriors on informative detection prediction, so that the detection accuracy is enhanced with false positive elimination.

- We evaluate the proposed BiDet on the PASCAL VOC and the large scale COCO datasets for comprehensive comparison with state-of-the-art binary neural networks in object detection.

2. Related Work

Network Quantization: Network quantization has been widely studied in recent years due to its efficiency in storage and computation. Existing methods can be divided into two categories: neural networks with weights and activations in one bit or multiple bits. Binary neural networks reduce the model complexity significantly due to the extremely high compression ratio. Hubara *et al.* [14] and Rastegari *et al.* [30] binarized both weights and activations in neural networks and replaced the multiply-accumulation with xnor and bitcount operations, where straight-through estimators were applied to relax the non-differentiable sign function for back-propagation. Liu *et al.* [25] added extra shortcut between consecutive convolutional blocks to strengthen the representational capacity of the network. They also used custom gradients to optimize the non-differentiable networks. Binary neural networks perform poorly on difficult tasks such as object detection due to the low representational capacity, multi-bit quantization strategies have been proposed with wider bitwidth. Jacob *et al.* [15] presented an 8-bit quantized model for inference in object detection and their method can be integrated with efficient architectures. Wei *et al.* [42] applied the knowledge distillation to learn 8-bit neural networks in small size from large full-precision models. Li *et al.* [19] proposed fully quantized neural networks in four bits with hardware-friendly implementation. Meanwhile, the instabilities during training were overcome by the presented techniques. Nevertheless, multi-bit neural networks still suffer from heavy storage and computation cost. Directly applying binary neural networks with constrained representational power in object detection leads to numerous false positives and significantly degrades the performance due to the information redundancy in the networks.

Object Detection: Object detection has aroused comprehensive interest in computer vision due to its wide application. Modern CNN based detectors are categorized into two-stage and one-stage detectors. In the former, R-CNN [8] was among the earliest CNN-based detectors with the pipeline of bounding box regression and classification. Progressive improvements were proposed for better efficiency and effectiveness. Fast R-CNN [7] presented the ROIpooling in the detection framework to achieve better accuracy and faster inference. Faster R-CNN [32] proposed the Region Proposal Networks to effectively generate region proposals instead of hand-crafted ones. FPN [21] introduced top-down architectures with lateral connections and the multi-scale features to integrate low-level and high-level features. In the latter regard, SSD [24] and YOLO [31] directly predicted the bounding box and the class without region proposal generation, so that real-time inference was achieved on GPUs with competitive accuracy. RetinaNet [22] proposed the focal loss to solve the problem of foreground-background class imbalance. However, CNN based detectors suffer from heavy storage and computational cost so that their deployment is limited.

Information Bottleneck: The information bottleneck (IB) principle was first proposed by [38] with the goal of extracting relevant information of the input with respect to the task, so that the IB principle are widely applied in compression. The IB principle enforces the mutual informa-

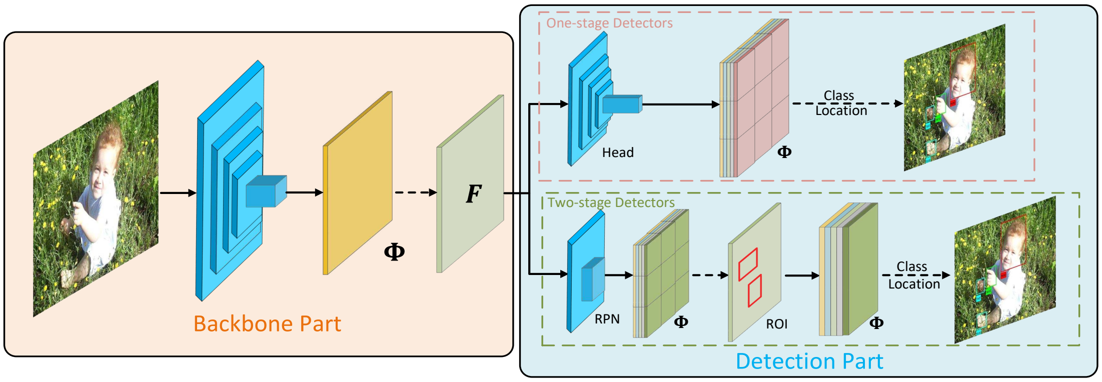


Figure 2. The pipeline of the information bottleneck based detectors, which consist of the backbone part and the detection part. The solid line represents the forward propagation in the network, while the dashed line means sampling from a parameterized distribution Φ. The high-level feature map F is sampled from the distribution parameterized by the backbone network. The one-stage and two-stage detector framework can be both employed in the detection part of our BiDet. For the one-stage detectors, the head network parameterizes the distribution of object classes and location. For two-stage detectors, Region Proposal Networks (RPN) parameterize the prior distribution of location and the posteriors are parameterized by the refining networks. (best viewed in color).

tion between the input and learned features to be minimized while simultaneously maximizing the mutual information between the features and groundtruth of the tasks. Louizos *et al*. [26] and Ullrich *et al*. [39] utilized the Minimal Description Length (MDL) principle that is equivalent to IB to stochastically quantize deep neural networks. Moreover, they used the sparse horseshoe and Gaussian mixture priors for weight learning in order to reduce the quantization errors. Dai *et al*. [5] pruned individual neurons via variational IB so that redundancy between adjacent layers was minimized by aggregating useful information in a subset of neurons. Despite the network compression, IB is also utilized in compact feature learning. Amjad *et al*. [1] proposed stochastic deep neural networks where IB could be utilized to learn efficient representations for classification. Shen *et al*. [35] imposed IB on existing hash models to generate effective binary representations so that the data semantics were fully utilized. In this paper, we extend the IB principle to squeeze the redundancy in binary detection networks, so that the false positives are alleviated and the detection precision is significantly enhanced.

3. Approach

In this section, we first extend the IB principle that removes the information redundancy to object detection. Then we present the details of learning the sparse object priors for object detection, which concentrate posteriors on informative prediction with false positive elimination. Finally, we propose the efficient binarized object detectors.

3.1. Information Bottleneck for Object Detection

The information bottleneck (IB) principle directly relates to compression with the best hypothesis that the data misfit and the model complexity should simultaneously be minimized, so that the redundant information irrelevant to the task is exclusive in the compressed model and the capacity of the lightweight model is fully utilized. The task of object detection can be regarded as a Markov process with the following Markov chain:

$$X \rightarrow F \rightarrow L, C \tag{1}$$

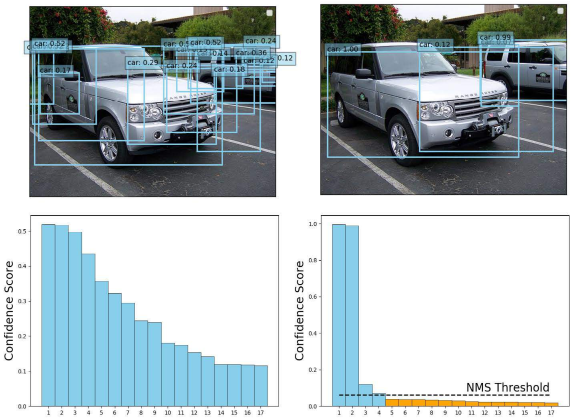
where X means the input images and F stands for the high-level feature maps output by the backbone part. C and L represent the predicted classes and location of the objects respectively. According to the Markov chain, the objective of the IB principle is written as follows:

$$\min_{\phi_b, \phi_d} \quad I(X; F) - \beta I(F; C, L) \tag{2}$$

where ϕ_b and ϕ_d are the parameters of the backbone and the detection part respectively. $I(X;Y)$ means the mutual information between two random variables X and Y. Minimizing the mutual information between the images and the high-level feature maps constrains the amount of information that the detector extracts, and maximizing the mutual information between the high-level feature maps and object detection enforces the detector to preserve more information related to the task. As a result, the redundant information irrelevant to object detection is removed. Figure 2 shows the pipeline for information bottleneck based detectors, the IB principle can be imposed on the conventional one-stage and two-stage detectors. We rewrite the first term of (2) according to the definition of mutual information:

$$I(X; F) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x})} \mathbb{E}_{\boldsymbol{f} \sim p(\boldsymbol{f}|\boldsymbol{x})} \log \frac{p(\boldsymbol{f}|\boldsymbol{x})}{p(\boldsymbol{f})} \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{f}$ are the specific input images and the corresponding high-level feature maps. $p(\boldsymbol{x})$ and $p(\boldsymbol{f})$ are the



(a) (b)

Figure 3. The detected objects and the corresponding confidence score (a) before and (b) after optimizing (6). The contrast of confidence score among different detected objects is significantly enlarged by minimizing alternate objective. As the NMS eliminates the positives with confidence score lower than the threshold, the sparse object priors are acquired and the posteriors are enforced to be concentrated on informative prediction. (best viewed in color).

prior distribution of $\boldsymbol{x}$ and $\boldsymbol{f}$ respectively, and $\mathbb{E}$ represents the expectation. $p(\boldsymbol{f}|\boldsymbol{x})$ is the posterior distribution of the high-level feature map conditioned on the input. We parameterize $p(\boldsymbol{f}|\boldsymbol{x})$ by the backbone due to its intractability, where evidence-lower-bound (ELBO) minimization is applied for relaxation. To estimate $I(X;F)$, we sample the training set to obtain the image $\boldsymbol{x}$ and sample the distribution parameterized by the backbone to acquire the corresponding high-level feature map $\boldsymbol{f}$.

The location and classification of objects based on the high-level feature map are independent, as the bounding box location and the classification probability are obtained via different network branches in the detection part. The mutual information in the second term of (2) is factorized:

$$I(F;C,L) = I(F;C) + I(F;L) \tag{4}$$

Similar to (3), we rewrite the mutual information between the high-level feature maps and the classes as follows:

$$I(F;C) = \mathbb{E}_{\boldsymbol{f}\sim p(\boldsymbol{f}|\boldsymbol{x})}\mathbb{E}_{\boldsymbol{c}\sim p(\boldsymbol{c}|\boldsymbol{f})}\log\frac{p(\boldsymbol{c}|\boldsymbol{f})}{p(\boldsymbol{c})} \tag{5}$$

where $\boldsymbol{c}$ is the object class labels including the background class. $p(\boldsymbol{c})$ and $p(\boldsymbol{c}|\boldsymbol{f})$ are the prior class distribution and posterior class distribution when given the feature maps respectively. Same as the calculation of (3), we employ the classification branch networks in the detection part to parameterize the distribution. Meanwhile, we divide the images to blocks for multiple object detection. For one-stage detectors such as SSD [24], we project the high-level feature map cells to the raw image to obtain the block partition. For two-stage detectors such as Faster R-CNN [32], we scale the ROI to the original image for block split. $\boldsymbol{c}\in\mathbb{Z}^{1\times b}$ represents the object class in b blocks of the image. We define c_i as the i_{th} element of $\boldsymbol{c}$, which demonstrates the class of the object whose center is in the i_{th} block of the image. The class of a block is assigned to background if the block does not contain the center of any groundtruth objects.

As the localization contains shift parameters and scale parameters for anchors, we rewrite the mutual information between the object location and high-level feature maps:

$$I(F;L) = \mathbb{E}_{\boldsymbol{f}\sim p(\boldsymbol{f}|\boldsymbol{x})}\mathbb{E}_{\boldsymbol{l}_1\sim p(\boldsymbol{l}_1|\boldsymbol{f})}\mathbb{E}_{\boldsymbol{l}_2\sim p(\boldsymbol{l}_2|\boldsymbol{f})}\log\frac{p(\boldsymbol{l}_1|\boldsymbol{f})p(\boldsymbol{l}_2|\boldsymbol{f})}{p(\boldsymbol{l}_1)p(\boldsymbol{l}_2)}$$

where $\boldsymbol{l}_1\in\mathbb{R}^{2\times b}$ represents the horizontal and vertical shift offset of the anchors in b blocks of the image, and $\boldsymbol{l}_2\in\mathbb{R}^{2\times b}$ means the height and width scale offset of the anchors. For the anchor whose center (x,y) is in the j_{th} block with height h and width w, the offset changes the bounding box in the following way: $(x,y)\rightarrow(x,y)+\boldsymbol{l}_{1,j}$ and $(h,w)\rightarrow(h,w)\cdot exp(\boldsymbol{l}_{2,j})$, where $\boldsymbol{l}_{1,j}$ and $\boldsymbol{l}_{2,j}$ represent the j_{th} column of $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$. The priors and the posteriors of shift offset conditioned on the feature maps are denoted as $p(\boldsymbol{l}_1)$ and $p(\boldsymbol{l}_1|\boldsymbol{f})$ respectively. Similarly, the scaling offset has the prior and the posteriors given feature maps $p(\boldsymbol{l}_2)$ and $p(\boldsymbol{l}_2|\boldsymbol{f})$. We leverage the localization branch networks in the detection part for distribution parameterization.

3.2. Learning Sparse Object Priors

Since the feature maps are binarized in BiDet, we utilize the binomial distribution with equal probability as the priors for each element of the high-level feature map $\boldsymbol{f}$. We assign the priors for object localization in the following form: $p(\boldsymbol{l}_{1,j}) = N(\boldsymbol{\mu}^0_{1,j},\boldsymbol{\Sigma}^0_{1,j})$ and $p(\boldsymbol{l}_{2,j}) = N(\boldsymbol{\mu}^0_{2,j},\boldsymbol{\Sigma}^0_{2,j})$, where $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ means the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For one-stage detectors, the object localization priors $p(\boldsymbol{l}_{1,j})$ and $p(\boldsymbol{l}_{2,j})$ are hypothesized to be the two-dimensional standard normal distribution. For two-stage detectors, Region Proposal Networks (RPN) output the parameters of the Gaussian priors.

As numerous false positives emerge in the binary detection networks, learning sparse object priors for detection part enforces the posteriors to be concentrated on informative detection prediction with false positive elimination. The priors for object classification is defined as follows:

$$p(c_i) = \mathbb{I}_{M_i}\cdot cat(\frac{1}{n+1}\cdot\mathbf{1}^{n+1}) + (1-\mathbb{I}_{M_i})\cdot cat([1,\mathbf{0}^n])$$

where $\mathbb{I}_x$ is the indicator function with $\mathbb{I}_1 = 1$ and $\mathbb{I}_0 = 0$, and M_i is the i_{th} element of the block mask $\boldsymbol{M}\in\{0,1\}^{1\times b}$. $cat(\boldsymbol{K})$ means the categorical distribution with the parameter $\boldsymbol{K}$. $\mathbf{1}^n$ and $\mathbf{0}^n$ are the all-one and zero vectors in n dimensions respectively, where n is the number of class. The multinomial distribution with equal probability is utilized for the class prior in the i_{th} block if M_i equals to one. Otherwise, the categorical distribution with the probability 1 for background and zero probability for other classes is leveraged for the prior class distribution. When M_i equals to zero, the detection part definitely predicts the background for object classification in the i_{th} block according to (5). In order to obtain sparse priors for object classification with fewer predicted positives, we minimize the L_1 norm of the block mask $\boldsymbol{M}$. We propose an alternative way to optimize $\boldsymbol{M}$ due to the non-differentiability, where the objective is written as follows:

$$\min_{s_i} -\frac{1}{m}\sum_{i=1}^{m} s_i \log s_i \qquad (6)$$

where $m = ||\boldsymbol{M}||_1$ represents the number of detected foreground objects in the image, and s_i is the normalized confidence score for the i_{th} predicted foreground object with $\sum_{i=1}^{m} s_i = 1$. As shown in Figure 3, minimizing (6) increases the contrast of confidence score among different predicted objects, and predicted objects with low confidence score are assigned to be negative by the non-maximum suppression (NMS) algorithms. Therefore, the block mask becomes sparser with fewer predicted objects, and the posteriors are concentrated on informative prediction with uninformative false positive elimination.

3.3. Efficient Binarized Object Detectors

In this section, we first briefly introduce neural networks with binary weights and activations, and then detail the learning objectives of our BiDet. Let $\boldsymbol{W_r^l}$ be the real-valued weights and $\boldsymbol{A_r^l}$ be the full-precision activations of the l_{th} layer in a given L-layer detection model. During the forward propagation, the weights and activations are binarized via the sign function: $\boldsymbol{W_b^l} = sign(\boldsymbol{W_r^l})$ and $\boldsymbol{A_b^l} = sign(\boldsymbol{W_r^l} \odot \boldsymbol{A_b^l})$. $sign$ means the element-wise sign function which maps the number larger than zero to one and otherwise to minus one, and $\odot$ indicates the element-wise binary product consisting of xnor and bitcount operations. Due to the non-differentiability of the sign function, straight-through estimator (STE) is employed to calculate the approximate gradients and update the real-valued weights in the back-propagation stage. The learning objectives for the proposed BiDet is written as follows:

$$\begin{aligned}
&\min J = J_1 + J_2\\
&= (\sum_{t,s}\log\frac{p(f_{st}|\boldsymbol{x})}{p(f_{st})} - \beta\sum_{i=1}^{b}\log\frac{p(c_i|\boldsymbol{f})p(\boldsymbol{l}_{1,i}|\boldsymbol{f})p(\boldsymbol{l}_{2,i}|\boldsymbol{f})}{p(c_i)p(\boldsymbol{l}_{1,i})p(\boldsymbol{l}_{2,i})})\\
&\quad - \gamma\cdot\frac{1}{m}\sum_{i=1}^{m} s_i\log s_i \qquad (7)
\end{aligned}$$

where γ is a hyperparameter that balances the importance of false positive elimination. The posterior distribution $p(c_i|\boldsymbol{f})$ is hypothesized to be the categorical distribution $cat(\boldsymbol{K}_i)$, where $\boldsymbol{K}_i \in \mathbb{R}^{1\times(n+1)}$ is the parameter and n is the number of classes. We assume the posterior of the shift and scale offset follows the Gaussian distribution: $p(\boldsymbol{l}_{1,j}|\boldsymbol{f}) = N(\boldsymbol{\mu}_{1,j}, \boldsymbol{\Sigma}_{1,j})$ and $p(\boldsymbol{l}_{2,j}|\boldsymbol{f}) = N(\boldsymbol{\mu}_{2,j}, \boldsymbol{\Sigma}_{2,j})$. The posteriors of the element in the s_{th} row and t_{th} column of binary high-level feature maps $p(f_{st}|\boldsymbol{x})$ is assigned to binomial distribution $cat([p_{ts}, 1-p_{ts}])$, where p_{ts} is the probability for f_{st} to be one. All the posterior distribution is parameterized by the neural networks. J_1 represents for the information bottleneck employed in object detection, which aims to remove information redundancy and fully utilize the representational power of the binary neural networks. The goal of J_2 is to enforce the object priors to be sparse so that the posteriors are encouraged to be concentrated on informative prediction with false positive elimination.

In the learning objective, $p(f_{st})$ in the binomial distribution is a constant. Meanwhile, the sparse object classification priors are imposed via J_2 so that $p(c_i)$ is also regarded as a constant. For one-stage detectors, constant $p(\boldsymbol{l}_{1,i})$ and $p(\boldsymbol{l}_{2,i})$ follows standard normal distribution. For two-stage detectors, $p(\boldsymbol{l}_{1,i})$ and $p(\boldsymbol{l}_{2,i})$ are parameterized by RPN, which is learned by the objective function. The last layer of the backbone that outputs the parameters of the binary high-level feature maps is real-valued in training for Monte-Carlo sampling and is binarzed with the sign function during inference. Meanwhile, the layers that output the parameters for object class and location distribution remain real-valued for accurate detection. During inference, we drop the network branch of covariance matrix for location offset, and assign all location prediction with the mean value to accelerate computation. Moreover, the prediction of object classes is set to that with the maximum probability to avoid time-consuming stochastic sampling in inference.

4. Experiments

In this section, we conducted comprehensive experiments to evaluate our proposed method on two datasets for object detection: PASCAL VOC [6] and COCO [23]. We first describe the implementation details of our BiDet, and then we validate the effectiveness of IB and sparse object priors for binarized object detectors by ablation study. Finally, we compare our method with state-of-the-art binary neural networks in the task of object detection to demonstrate superiority of the proposed BiDet.

4.1. Datasets and Implementation Details

We first introduce the datasets that we carried out experiments on and data preprocessing techniques:

PASCAL VOC: The PASCAL VOC dataset contains natural images from 20 different classes. We trained our model on the VOC 2007 and VOC 2012 trainval sets which consist of around 16k images, and we evaluated our method on VOC 2007 test set including about 5k images. Following [6], we used the mean average precision (mAP) as the evaluation criterion.

COCO: The COCO dataset consists of images from 80 different categories. We conducted experiments on the 2014 COCO object detection track. We trained our model with the combination of 80k images from the training set and 35k images sampled from validation set (trainval35k [2]) and tested our method on the remaining 5k images in the validation set (minival [2]). Following the standard COCO evaluation metric [23], we report the average precision (AP) for IoU $\in [0.5 : 0.05 : 0.95]$ denoted as mAP@$[.5, .95]$. We also report AP$_{50}$, AP$_{75}$ as well as AP$_s$, AP$_m$ and AP$_l$ to further analyze our method.

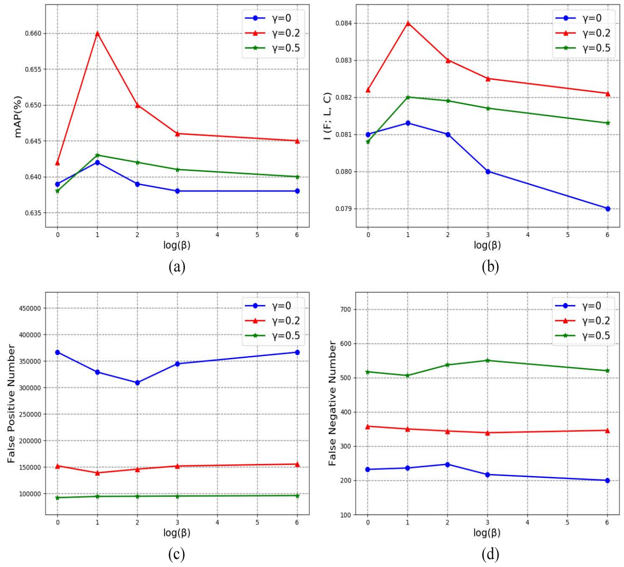
We trained our BiDet with the SSD300 [24] and Faster R-CNN [32] detection framework whose backbone were VGG16 [36] and ResNet-18 [11] respectively. Following the implementation of binary neural networks in [14], we remained the first and last layer in the detection networks real-valued. We used the data augmentation techniques in [24] and [32] when training our BiDet with SSD300 and Faster R-CNN detection frameworks respectively.

In most cases, the backbone network was pre-trained on ImageNet [33] in the task of image classification. Then we jointly finetuned the backbone part and trained the detection part for the object detection task. The batchsize was assigned to be 32, and the Adam optimizer [17] was applied. The learning rate started from 0.001 and decayed twice by multiplying 0.1 at the 6_{th} and 10_{th} epoch out of 12 epochs. Hyperparamters β and γ were set as 10 and 0.2 respectively.

4.2. Ablation Study

Since the IB principle removes the redundant information in binarized object detectors and the learned sparse object priors concentrate the posteriors on informative prediction with false positive alleviation, the detection accuracy is enhanced significantly. To verify the effectiveness of the IB principle and the learned sparse priors, we conducted the ablation study to evaluate our BiDet w.r.t. the hyperparameter β and γ in the objective function. We adopted the SSD detection framework with VGG16 backbone for our BiDet on the PASCAL VOC dataset. We report the mAP, the mutual information between high-level feature maps and the object detection $I(F; L, C)$, the number of false positives and the number of false negatives with respect to β and γ in Figure 4 (a), (b), (c) and (d) respectively. Based on the results, we observe the influence of the IB principle and the learned sparse object priors as follows.

By observing Figure 4 (a) and (b), we conclude that mAP and $I(F; L, C)$ are positively correlated as they demonstrate the detection performance and the amount of related information respectively. Medium β provides the optimal trade-off between the amount of extracted information and the related information so that the representational capacity of the binary object detectors is fully utilized with redundancy removal. Small β fails to leverage the representational power of the networks as the amount of extracted information is limited by regularizing the high-level feature maps, while large β enforces the networks to learn redundant information which leads to significant over-fitting. Meanwhile, medium γ offers optimal sparse object priors that enforces the posteriors to concentrate on most informative prediction. Small γ is not capable of sparsifying the predicted objects, and large γ disables the posteriors to represent informative objects with excessive sparsity.



(a)

(b)

(c)

(d)

Figure 4. Ablation study w.r.t. hyperparameters β and γ, where the variety of (a) mAP, (b) the mutual information between high-level feature maps and the object detection $I(F; L, C)$, (c) the number of false positives and (d) the number of false negatives are demonstrated. (best viewed in color).

By comparing the variety of false positives and false negatives w.r.t. β and γ, we know that medium β decreases false positives most significantly and changing β does not varies the number of false negatives notably, which means that the redundancy removal only alleviates the uninformative false positives while remains the informative true positives unchanged. Meanwhile, small γ fails to constrain the false positives and large γ clearly increases the false negatives, which both degrades the performance significantly.

4.3. Comparison with the State-of-the-art Methods

In this section, we compare the proposed BiDet with the state-of-the-art binary neural networks including BNN [4], Xnor-Net [30] and Bi-Real-Net [25] in the task of object detection on the PASCAL VOC and COCO datasets. For reference, we report the detection performance of the multi-bit quantized networks containing DoReFa-Net [46] and TWN [18] and the lightweight networks MobileNetV1 [13].

Table 1. Comparison of parameter size, FLOPs and mAP (%) with the state-of-the-art binary neural networks in both one-stage and two-stage detection frameworks on PASCAL VOC. The detector with the real-valued and multi-bit backbone is given for reference. BiDet (SC) means the proposed method with extra shortcut for the architectures.

Framework	Input	Backbone	Quantization	W/A (bit)	#Params	MFLOPs	mAP
SSD300	300 × 300	VGG16	–	32/32	100.28MB	31,750	72.4
		MobileNetV1			30.07MB	1,150	68.0
		VGG16	TWN	2/32	24.54MB	8,531	67.8
			DoReFa-Net	4/4	29.58MB	4,661	69.2
			BNN	1/1	22.06MB	1,275	42.0
			Xnor-Net		22.16MB	1,279	50.2
			BiDet		22.06MB	1,275	**52.4**
			Bi-Real-Net	1/1	21.88MB	1,277	63.8
			BiDet (SC)		21.88MB	1,277	**66.0**
		MobileNetV1	Xnor-Net	1/1	22.48MB	836	48.9
			BiDet		22.48MB	836	**51.2**
Faster R-CNN	600 × 1000	ResNet-18	–	32/32	47.35MB	36,013	74.5
			TWN	2/32	3.83MB	9,196	69.9
			DoReFa-Net	4/4	6.73MB	4,694	71.0
			BNN	1/1	2.38MB	779	35.6
			Xnor-Net		2.48MB	783	48.4
			BiDet		2.38MB	779	**50.0**
			Bi-Real-Net	1/1	2.39MB	781	58.2
			BiDet (SC)		2.39MB	781	**59.5**

Table 2. Comparison of mAP@[.5, .95] (%), AP with different IOU threshold and AP for objects in various sizes with state-of-the-art binarized object detectors in SSD300 and Faster R-CNN detection framework on COCO, where the performance of real-valued and multi-bit detectors is reported for reference. BiDet (SC) means the proposed method with extra shortcut for the architectures.

Framework	Input	Backbone	Quantization	mAP@[.5, .95]	AP$_{50}$	AP$_{75}$ (%)	AP$_s$	AP$_m$	AP$_l$
SSD300	300 × 300	VGG16	–	23.2	41.2	23.4	5.3	23.2	39.6
			TWN	16.9	33.0	15.8	5.0	16.9	27.2
			DoReFa-Net	19.5	35.0	19.6	5.1	20.5	32.8
			BNN	6.2	15.9	3.8	2.4	10.0	9.9
			Xnor-Net	8.1	19.5	5.6	2.6	8.3	13.3
			BiDet	**9.8**	**22.5**	**7.2**	**3.1**	**10.8**	**16.1**
			Bi-Real-Net	11.2	26.0	8.3	3.1	12.0	18.3
			BiDet (SC)	**13.2**	**28.3**	**10.5**	**5.1**	**14.3**	**20.5**
Faster R-CNN	600 × 1000	ResNet-18	–	26.0	44.8	27.2	10.0	28.9	39.7
			TWN	19.7	35.3	19.7	5.1	20.7	33.3
			DoReFa-Net	22.9	38.6	23.7	8.0	24.9	36.3
			BNN	5.6	14.3	2.6	2.0	8.5	9.3
			Xnor-Net	10.4	21.6	8.8	2.7	11.8	15.9
			BiDet	**12.1**	**24.8**	**10.1**	**4.1**	**13.5**	**17.7**
			Bi-Real-Net	14.4	29.0	13.4	3.7	15.4	24.1
			BiDet (SC)	**15.7**	**31.0**	**14.4**	**4.9**	**16.7**	**25.4**

Results on PASCAL VOC: Table 1 illustrates the comparison of computation complexity, storage cost and the mAP across different quantization methods and detection frameworks. Our BiDet significantly accelerates the computation and saves the storage by 24.90× and 4.55× with the SSD300 detector and 46.23× and 19.81× with the Faster R-CNN detector. The efficiency is enhanced more notably in the Faster R-CNN detector, as there are multiple real-valued output layers of the head networks in SSD300 for multi-scale feature extraction.

Compared with the state-of-the-art binary neural networks, the proposed BiDet improves the mAP of Xnor-Net by 2.2% and 1.6% with SSD300 and Faster R-CNN frameworks respectively with fewer FLOPs and the number of parameters than Xnor-Net. As demonstrated in [25], adding extra shortcut between consecutive convolutional layers can further enhance the representational power of the binary neural networks, we also employ architecture with additional skip connection to evaluate our BiDet in networks with stronger capacity. Due to the information redundancy,

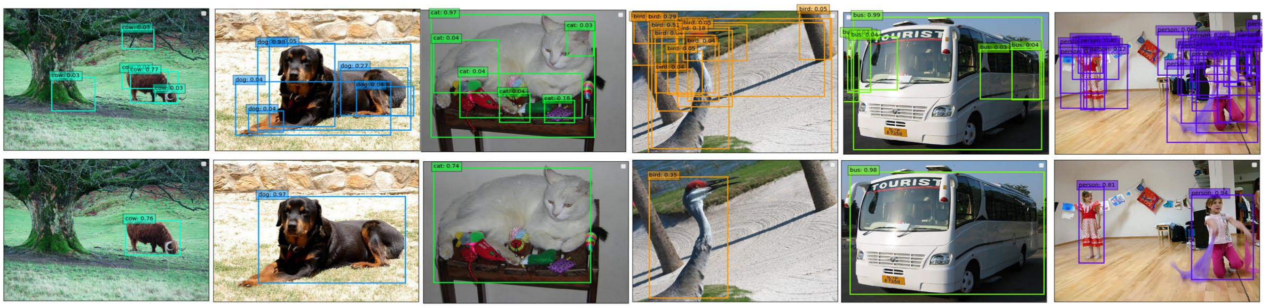


Figure 5. Qualitative results on PASCAL VOC. Images in the top row shows the object predicted by Xnor-Net, while the images with the objects detected by our BiDet are displayed in the bottom row. The proposed BiDet removes the information redundancy to fully utilize the network capacity, and learns the sparse object priors to eliminate false positives (best viewed in color).

the performance of Bi-Real-Net with constrained network capacity is degraded significantly compared with their full-precision counterparts in both one-stage and two-stage detection frameworks. On the contrary, our BiDet imposes the IB principle on learning binary neural networks for object detection and fully utilizes the network capacity with redundancy removal. As a result, the proposed BiDet increases the mAP of Bi-Real-Net by 2.2% and 1.3% in SSD300 and Faster R-CNN detectors respectively without additional computational and storage cost. Figure 5 shows the qualitative results of Xnor-Net and our BiDet in the SSD300 detection framework with VGG16, where the proposed BiDet significantly alleviates the false positives.

Due to the different pipelines in one-stage and two-stage detectors, the mAP gained from the proposed BiDet with Faster R-CNN is less than SSD300. As analyzed in [22], one-stage detectors face the severe positive-negative class imbalance problem which two-stage detectors are free of, so that one-stage detectors are usually more vulnerable to false positives. Therefore, one-stage object detection framework obtains more benefits from the proposed BiDet, which learns the sparse object priors to concentrate the posteriors on informative prediction with false positive elimination.

Moreover, our BiDet can be integrated with other efficient networks in object detection for further computation speedup and storage saving. We employ our BiDet as a plug-and-play module in SSD detector with the MobileNetV1 network and saves the computational and storage cost by $1.47\times$ and $1.38\times$ respectively. Compared with the detectors that directly binarize weights and activations in MobileNetV1 with Xnor-Net, BiDet improves the mAP by a sizable margin, which depicts the effectiveness of redundancy removal for networks with extremely low capacity.

Results on COCO: The COCO dataset is much more challenging for object detection than PASCAL VOC due to the high diversity and large scale. Table 2 demonstrates mAP, AP with different IOU threshold and AP of objects in various sizes. Compared with the state-of-the-art binary neural networks Xnor-Net, our BiDet improves the mAP by 1.7% and 1.7% in SSD300 and Faster R-CNN detection framework respectively due to the information redundancy removal. Moreover, the proposed BiDet also enhances the binary one-stage and two-stage detectors with extra shortcut by 2.0% and 1.3% on mAP. Comparing with the baseline methods of network quantization, our method achieves better performance in the AP with different IOU threshold and AP for objects in different sizes, which demonstrates the universality in different application settings.

5. Conclusion

In this paper, we have proposed a binarized neural network learning method called BiDet for efficient object detection. The presented BiDet removes the redundant information via information bottleneck principle to fully utilize the representational capacity of the networks and enforces the posteriors to be concentrated on informative prediction for false positive elimination, through which the detection precision is significantly enhanced. Extensive experiments depict the superiority of BiDet in object detection compared with the state-of-the-art binary neural networks.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, Grant U1713214, and Grant 61672306, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564, and in part by Tsinghua University Initiative Scientific Research Program.

References

[1] Rana Ali Amjad and Bernhard Claus Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *TPAMI*, 2019.

[2] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, pages 2874–2883, 2016.

[3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *NIPS*, pages 742–751, 2017.

[4] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *NIPS*, pages 3123–3131, 2015.

[5] Bin Dai, Chen Zhu, and David Wipf. Compressing neural networks using the variational information bottleneck. *arXiv preprint arXiv:1802.10399*, 2018.

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[7] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[9] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. *arXiv preprint arXiv:1908.05033*, 2019.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[12] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *ICCV*, pages 1389–1397, 2017.

[13] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[14] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *NIPS*, pages 4107–4115, 2016.

[15] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *CVPR*, pages 2704–2713, 2018.

[16] Hyeji Kim, Muhammad Umar Karim Khan, and Chong-Min Kyung. Efficient neural network compression. In *CVPR*, pages 12569–12577, 2019.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

[19] Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. Fully quantized network for object detection. In *CVPR*, pages 2810–2819, 2019.

[20] Shaohui Lin, Rongrong Ji, Chao Chen, Dacheng Tao, and Jiebo Luo. Holistic cnn compression via low-rank decomposition with knowledge transfer. *TPAMI*, 2018.

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.

[25] Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *ECCV*, pages 722–737, 2018.

[26] Christos Louizos, Karen Ullrich, and Max Welling. Bayesian compression for deep learning. In *NIPS*, pages 3288–3298, 2017.

[27] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *CVPR*, pages 11264–11272, 2019.

[28] Bo Peng, Wenming Tan, Zheyang Li, Shun Zhang, Di Xie, and Shiliang Pu. Extreme network compression via filter group approximation. In *ECCV*, pages 300–316, 2018.

[29] Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang Yu, Yuxing Peng, and Jian Sun. Thundernet: Towards real-time generic object detection. *arXiv preprint arXiv:1903.11752*, 2019.

[30] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*, pages 525–542, 2016.

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.

[35] Yuming Shen, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, and

Ziyi Shen. Embarrassingly simple binary representation learning. In *ICCVW*, pages 0–0, 2019.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *CVPR*, pages 2820–2828, 2019.

[38] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[39] Karen Ullrich, Edward Meeds, and Max Welling. Soft weight-sharing for neural network compression. *arXiv preprint arXiv:1702.04008*, 2017.

[40] Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and S Yu Philip. Private model compression via knowledge distillation. In *AAAI*, volume 33, pages 1190–1197, 2019.

[41] Ziwei Wang, Jiwen Lu, Chenxin Tao, Jie Zhou, and Qi Tian. Learning channel-wise interactions for binary convolutional neural networks. In *CVPR*, pages 568–577, 2019.

[42] Yi Wei, Xinyu Pan, Hongwei Qin, Wanli Ouyang, and Junjie Yan. Quantization mimic: Towards very tiny cnn for object detection. In *ECCV*, pages 267–283, 2018.

[43] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *CVPR*, pages 10734–10742, 2019.

[44] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856, 2018.

[45] Chenglong Zhao, Bingbing Ni, Jian Zhang, Qiwei Zhao, Wenjun Zhang, and Qi Tian. Variational convolutional neural network pruning. In *CVPR*, pages 2780–2789, 2019.

[46] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.