Diversifying Sample Generation for Accurate Data-Free Quantization

Xiangguo Zhang*[1], Haotong Qin*[1], Yifu Ding[1], Ruihao Gong[3,4],
Qinghua Yan[1], Renshuai Tao[1], Yuhang Li[2], Fengwei Yu[3,4], Xianglong Liu[1†]
[1]Beihang University [2]Yale University [3]SenseTime Research [4]Shanghai AI Laboratory
{xiangguozhang, zjdyf, yanqh, rstao}@buaa.edu.cn, yuhang.li@yale.edu,
{qinhaotong, xlliu}@nlsde.buaa.edu.cn, {gongruihao, yufengwei}@sensetime.com

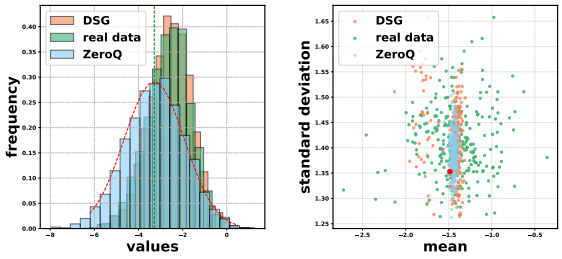
Abstract

*Quantization has emerged as one of the most prevalent approaches to compress and accelerate neural networks. Recently, data-free quantization has been widely studied as a practical and promising solution. It synthesizes data for calibrating the quantized model according to the batch normalization (BN) statistics of FP32 ones and significantly relieves the heavy dependency on real training data in traditional quantization methods. Unfortunately, we find that in practice, the synthetic data identically constrained by BN statistics suffers serious homogenization at both distribution level and sample level and further causes a significant performance drop of the quantized model. We propose **D**iverse **S**ample **G**eneration (DSG) scheme to mitigate the adverse effects caused by homogenization. Specifically, we slack the alignment of feature statistics in the BN layer to relax the constraint at the distribution level and design a layerwise enhancement to reinforce specific layers for different data samples. Our DSG scheme is versatile and even able to be applied to the state-of-the-art post-training quantization method like AdaRound. We evaluate the DSG scheme on the large-scale image classification task and consistently obtain significant improvements over various network architectures and quantization methods, especially when quantized to lower bits (*e.g.*, up to 22% improvement on W4A4). Moreover, benefiting from the enhanced diversity, models calibrated with synthetic data perform close to those calibrated with real data and even outperform them on W4A4.*

1. Introduction

Recently, Deep Neural Networks (DNNs), especially Convolutional Neural Networks (CNNs) achieve great success in a variety of domains, such as image classification [16, 29, 30, 33, 34, 37, 36, 32], object detection [9, 8, 21, 27, 17, 35], semantic segmentation [6, 42], *etc*. Nevertheless, deploying state-of-the-art models on resource-constrained devices is still challenging due to massive parameters and high computational complexity. With more and more hardware support low-precision computations [41, 26, 25], quantization has emerged as one of the most promising approaches to obtain efficient neural networks. Since the whole training stage is required, quantization-aware training methods are considered to be time-consuming and computationally intensive [10, 14, 24]. Therefore, post-training quantization methods are proposed, which directly quantize the FP32 models without retraining or fine-tuning [1, 4, 40, 19, 18]. Nevertheless, they still require real training data to calibrate quantized models that is not often ready-to-use for privacy or security concerns, such as medical data and user data.

Fortunately, recent work have proposed data-free quantization to quantize models without any access to real data. Existing data-free quantization methods [20, 2, 11, 3], such as ZeroQ [2], generate "optimal synthetic data", which learns



(a) (b)

Figure 1: Comparison between real data and synthetic data (generated by DSG and ZeroQ [2]) with 256 samples of each. (a) shows the activation distribution of one channel in ResNet-18 [13]. ZeroQ data mostly fits the normal distribution of BN statistics (the red line), while real data and DSG data have an offset. (b) is a chart of the mean and standard deviation of one channel. ZeroQ data gathers near BN statistics (the red dot) but real data and DSG data are more scattered.

*Equal contribution.
†Corresponding author

an input data distribution to best match the Batch Normalization statistics of the FP32 model. However, models calibrated with real data perform better than those calibrated with synthetic data, though the synthetic data matches BN statistics better. Our study reveals that the data generation process in typical data-free quantization methods has significant homogenization issues at both distribution and sample levels, which prevent models from higher accuracy. **First**, since the synthetic data is constrained to match the BN statistics, the feature distribution in each layer might overfit the BN statistics when data is fed forward in neural networks. As shown in Figure 1(a), the distribution of the synthetic samples generated by ZeroQ almost fits the normal distribution obeying the corresponding BN statistics, while those of real data have an obvious offset leading to more diverse distribution. We call the phenomenon of the synthetic samples as the distribution level homogenization. **Second**, all samples of synthetic data are optimized by the same objective function in existing generative data-free quantization methods. For instance, ZeroQ and GDFQ [28] apply the same constraint to all data samples, which directly sums the loss objective (KL loss or statistic loss) of each layer. Therefore, as shown in Figure 1(b), the feature distribution statistics of these samples are similar. Specifically, the distribution statistics of synthetic data generated by existing methods are centralized while those of real data are dispersed, as so-called sample level homogenization.

To mitigate the adverse effects caused by these issues, we propose a novel **D**iverse **S**ample **G**eneration (DSG) scheme, a simple yet effective data generation method for data-free quantization to enhance the diversity of data. Our DSG scheme consists of two technical contributions: (1) *Slack Distribution Alignment* (SDA): slack the alignment of the feature statistics in BN layers to relax the constraint of distribution; (2) *Layerwise Sample Enhancement* (LSE): apply the layerwise enhancement to reinforce specific layers for different data samples.

Our DSG scheme presents a novel perspective of data diversity for data-free quantization. We evaluate the DSG scheme on the large-scale image classification task, and the results show that DSG performs remarkably well across various network architectures such as ResNet-18 [13], ResNet-50, SqueezeNext [7], ShuffleNet [39], and InceptionV3 [31], and surpasses previous methods by a wide margin, even outperforms models calibrated with real data. Moreover, we show that the synthetic data generated by DSG scheme can be applied to the most advanced post-training quantization methods, such as AdaRound [19].

We summarize our main contributions as:

1. We revisit the data generation process of data-free quantization methods from the diversity perspective. Our study reveals the homogenization problems of synthetic data existing at two levels that harm the performance of the quantized models.
2. We propose Diverse Sample Generation (DSG) scheme, a novel sample generation method for accurate data-free quantization that enhances the diversity of data by combining two practical techniques: *Slack Distribution Alignment* and *Layerwise Sample Enhancement* to solve the homogenization at distribution and sample levels.
3. We evaluate the proposed method on the large-scale image classification task and consistently obtain significant improvements over various base models and state-of-the-art (SOTA) post-training quantization methods.

2. Related Work

Data-Driven Quantization. Quantization is a potent approach to accelerate the inference phase due to its low-bit representations of weights and activations. However, models often suffer an accuracy degeneration after quantization, especially when quantized to ultra-low bit-width. Quantization-aware training is proposed to retrain or fine-tune the quantized model with training/validation data to improve the accuracy, as described in earlier work such as [10, 14]. They often give satisfactory results, but the training process is computationally expensive and time-consuming. More crucially, the original training/validation data are not always accessible, especially on private and secure occasions.

Post-training quantization focuses on obtaining accurate quantized models with small computation and time cost, which has achieved relatively good performance without any fine-tuning or training process. Particularly, [1] approximates the optimal clipping value analytically and introduces a bit allocation policy and bias-correction to quantize both activations and weights to 4-bit. [4] formalizes the linear quantization task as a minimum mean squared error problem for both weights and activations. [40] exploits channel splitting to avoid containing outliers. [19] proposes AdaRound, a better weight-rounding mechanism for post-training quantization that adapts to the data and the task loss. However, these aforementioned methods also require access to limited data for recovering the performance.

Data-Free Quantization. Recent work [20, 2, 11, 3, 28] go further to data-free quantization, which requires neither training nor validation data for quantization. [20] uses weight equalization and bias correction to achieve competitive accuracy on layerwise quantization compared with channel-wise quantization. However, it suffers a non-negligible performance drop when quantized to 6 or lower bit-width. While [2] utilizes mixed-precision quantization with synthetic data to support ultra-low precision quantization. [11] proposes inception scheme and BN statistics scheme to generate data for calibration and time-consuming knowledge distillation finetuning. Furthermore, [3] proposes a data-free adversarial knowledge distillation method, which minimizes the maximum distance between the outputs of the

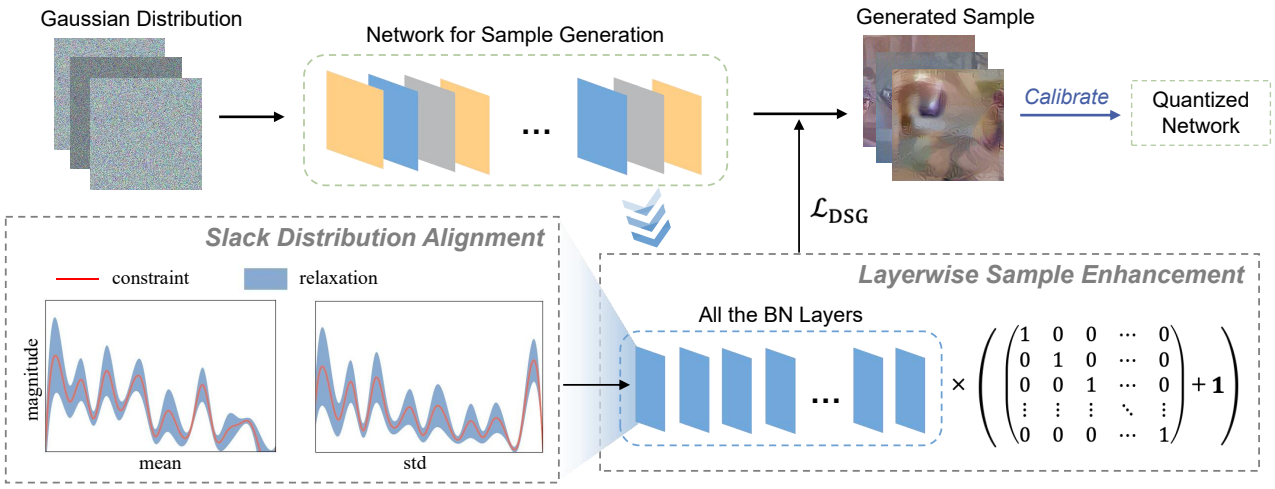


Figure 2: The framework of Diverse Sample Generation (DSG) scheme, which consists of Slack Distribution Alignment (SDA) and Layerwise Sample Enhancement (LSE). SDA relaxes BN statistics constraint in each layer, and LSE provides specific loss term for each sample. First, we initialize synthetic data from Gaussian. Then we compute the loss using our SDA and LSE and update synthetic data with this loss. Finally, we calibrate the quantized model with synthetic data.

FP32 teacher model and the quantized student model for any adversarial samples from a generator. [28] couples the data generation and model finetuning, and uses time-consuming distillation to improve the accuracy of quantized models. However, [11], [28] and [3] have the same limitations as [2] while generating data that we explain in Section 3.2.

3. Diverse Sample Generation

In this section, we revisit the image generation process in data-free quantization and point out the homogenization of the synthetic data in previous work. We present our Diverse Sample Generation (DSG) scheme to diversify the synthetic data for obtaining accurate quantized models.

3.1. Preliminaries

Data-free quantization methods are proposed to quantize the FP32 model, and ZeroQ [2] is a typical representation of these work, which is proposed to learn an input data distribution that best matches the BN statistics, *i.e.*, the mean and standard deviation, by solving the following optimization problem:

$$\min_{\mathbf{x}^s} \mathcal{L}_{\text{BN}} = \frac{1}{N}\sum_{i=0}^{N}\left(\|\tilde{\boldsymbol{\mu}}_i^s - \boldsymbol{\mu}_i\|_2^2 + \|\tilde{\boldsymbol{\sigma}}_i^s - \boldsymbol{\sigma}_i\|_2^2\right) \quad (1)$$

where $\mathcal{L}_{\text{BN}}$ is the BN statistics loss to be minimized, $\mathbf{x}^s$ is the synthetic input data. $\tilde{\boldsymbol{\mu}}_i^s/\tilde{\boldsymbol{\sigma}}_i^s$ are the mean/standard deviation of the feature distribution of synthetic data at the i-th BN layer, $\boldsymbol{\mu}_i/\boldsymbol{\sigma}_i$ are mean/standard deviation parameters stored in i-th BN layer of pre-trained FP32 model.

3.2. Homogenization of Synthetic Data

Although many generative data-free quantization methods have been proposed to resolve the problem of accessing real data, they are considered to suffer a huge drop in performance compared with post-training quantization calibrated with real data. We explore the commonly practiced data synthesizing processes and find that the homogenization issue exists at two levels, which degrades the fidelity and quality of synthesized data and thus behave differently with real images:

1) Distribution level homogenization: BN statistics loss in Eq. (1) strictly constrains the feature distribution of the synthetic data, aiming to generate samples that exhibit similarities to original training data, which is generally regarded as the upper limit of the performance of synthetic data. However, fitting BN statistics obtained from training data is not equivalent to imitating real data. As shown in Figure 1(a), the distribution of synthetic data generated by ZeroQ is almost in the immediate vicinity of the BN statistics. In contrast, the distribution of the real samples deviates from BN statistics. Therefore, this constraint causes the feature distribution of synthetic data overfitting to the BN statistics in each layer, as so-called distribution homogenization.

2) Sample level homogenization: Besides homogenization at the distribution level, the application of $\mathcal{L}_{\text{BN}}$ also leads to the homogenization at the sample level. Existing methods, such as ZeroQ, generate a batch of synthetic data to calibrate or finetune quantized models. However, since all the samples are initialized and optimized by the same objective function, the statistics of each sample are quite similar, while those of real samples are more versatile. As shown in Figure 1(b), the feature distribution statistics of ZeroQ samples are almost overlapping, while the real-world images have larger variance of statistic distribution. That is to say, ZeroQ data suffers the homogenization at the sample level. Thus, it cannot be applied to determine the clip values of activations for quantized models in place of the real data. Otherwise, it might cause a significant performance drop.

As described above, homogenization exists in the data generation process in previous work of synthetic data-free quantization, including distribution level and sample level, which results in the synthetic data lacking diversity, consequently preventing the quantized models from good performance. In this paper, we propose a novel synthetic data-free quantization method, namely **D**iverse **S**ample **G**eneration (DSG) scheme, which aims to address this issue by enhancing the diversity of synthetic data. Calibrating with data generated by our synthesizing scheme, quantized models gain non-negligible improvements in accuracy.

3.3. Slack Distribution Alignment

We propose the Slack Distribution Alignment (SDA) to eliminate the distribution homogenization, which relaxes BN statistics constraint. Intuitively, we introduce margins for mean and standard deviation statistics of activations, respectively, which allow sufficient distribution variation. In more detail, we add the relaxation constants to the original BN statistics loss function to tackle the distribution homogenization issue. The loss term of SDA of i-th BN layer l_{SDA_i} is defined as follow:

$$\begin{aligned} l_{\text{SDA}_i} = &\left\|\max\left(|\tilde{\boldsymbol{\mu}}_i^s - \boldsymbol{\mu}_i| - \delta_i, 0\right)\right\|_2^2 \\ &+ \left\|\max\left(|\tilde{\boldsymbol{\sigma}}_i^s - \boldsymbol{\sigma}_i| - \gamma_i, 0\right)\right\|_2^2 \end{aligned} \tag{2}$$

where δ_i and γ_i denote the relaxation constants for the mean and standard deviation statistics of features at the i-th BN layers, we admit a gap between the statistics of synthetic data and the statistic parameters of BN layers. Within a specific range, the statistics of synthetic data can fluctuate with relaxed constraints. Thus the feature distribution of synthetic data becomes more diverse, as shown in Figure 1(a).

A significant challenge is determining the values of δ_i and γ_i without any real data access. Since real data performs well in calibrating the quantized model, the gap between the feature statistics of real data and BN statistics is considered as a reasonable reference, even an optimal degree of relaxation. Since the neural input is a sum of many inputs, the Gaussian assumption can be seen as a common approximation. From the central limit theorem, we expect it to be approximately Gaussian distribution [23]. Therefore, we propose to use synthetic data randomly sampled from Gaussian distribution to determine δ_i and γ_i. **First**, we initialize 1024 synthetic samples by the Gaussian distribution with $\mu = 0$ and $\sigma = 1$. We input the synthetic samples into models and save the feature statistics at all BN layers, specifically the mean and standard deviation of the feature distribution. **Second**, we calculate the gap between the saved statistics and the corresponding BN statistics. We take the percentile of the absolute values of the gaps as δ_i and γ_i. Formally, δ_i and γ_i are defined as follows:

$$\delta_i = \left|\tilde{\boldsymbol{\mu}}_i^0 - \boldsymbol{\mu}_i\right|_\epsilon \quad \gamma_i = \left|\tilde{\boldsymbol{\sigma}}_i^0 - \boldsymbol{\sigma}_i\right|_\epsilon \tag{3}$$

where $\tilde{\boldsymbol{\mu}}_i^0/\tilde{\boldsymbol{\sigma}}_i^0$ are mean/standard deviation of the feature of Gaussian initialized data $\mathbf{x}^0$ at the i-th BN layer. $\left|\tilde{\boldsymbol{\mu}}_i^0 - \boldsymbol{\mu}_i\right|_\epsilon$ and $\left|\tilde{\boldsymbol{\sigma}}_i^0 - \boldsymbol{\sigma}_i\right|_\epsilon$ denote the ϵ percentile of $\left|\tilde{\boldsymbol{\mu}}_i^0 - \boldsymbol{\mu}_i\right|$ and $\left|\tilde{\boldsymbol{\sigma}}_i^0 - \boldsymbol{\sigma}_i\right|$, respectively. The value of ϵ in range $(0, 1]$ determines the values of δ_i and γ_i, further settling the degree of relaxation. When the value of ϵ becomes larger, the constraints in Eq. (2) are more relaxing. The default value of ϵ is set as 0.9 to prevent the influence of outliers.

3.4. Layerwise Sample Enhancement

Existing generative data-free quantization methods always use the same objective function to optimize all samples of data. Specifically, as shown in Eq. (1), the loss terms of all layers are equivalently summed. Meanwhile, the optimizing strategy is invariant among samples. In other words, all the samples have the same attention to each layer, which leads to homogeneity at the sample level.

In order to enhance the diversity at the sample level, we propose **L**ayerwise **S**ample **E**nhancement (LSE), which reinforces the loss of a specific layer for each sample. Specifically, the loss function of every synthetic image in a batch may be different. In fact, for a network with N BN layers, LSE can provide N loss terms and apply each of them to the specific data sample. Here, we suppose to generate a batch of images, and the batch size is set as N, equaling to the number of BN layers of the model. Therefore, the loss term $\mathcal{L}$ of LSE for this batch is defined as:

$$\mathcal{L} = \frac{1}{N} \cdot \mathbf{1}^T \left(\left(\mathbf{I} + \mathbf{1}\mathbf{1}^T\right)\mathbf{L}\right) \tag{4}$$

where $\mathbf{I}$ is an N-dimensional identity matrix, $\mathbf{1}$ is an N dimension column vector of all ones, $\mathbf{L} = \{l_0, l_1, ..., l_N\}^T$ represents the vector containing loss terms of all BN layers. We define $\mathbf{X}_{\text{LSE}} = \mathbf{I} + \mathbf{1}\mathbf{1}^T$ as the enhancement matrix, and the Eq. (4) can be simplified as:

$$\mathcal{L} = \frac{1}{N} \cdot \mathbf{1}^T \left(\mathbf{X}_{\text{LSE}}\mathbf{L}\right) \tag{5}$$

where $\mathbf{X}_{\text{LSE}}\mathbf{L}$ can be seen as a N-dimensional column vector, the i-th element of which represents the loss function of the i-th image in this batch. Thus, we impose a unique constraint on each sample of this batch and jointly optimize the whole batch.

For a network with N BN layers, LSE can simultaneously generate various samples in a batch, and each kind of sample has a enhancement on a particular layer. Specifically, the SDA is applied to obtain the loss term of layers $\mathbf{L}_{\text{SDA}} = \{l_{\text{SDA}_0}, l_{\text{SDA}_1}, \ldots, l_{\text{SDA}_N}\}^T$, and our DSG loss $\mathcal{L}_{\text{DSG}}$ is defined as:

$$\mathcal{L}_{\text{DSG}} = \frac{1}{N} \cdot \mathbf{1}^T \left(\mathbf{X}_{\text{LSE}}\mathbf{L}_{\text{SDA}}\right) \tag{6}$$

Compared with the synthetic data generated with BN statistics loss, images generated by the DSG scheme may have

Algorithm 1: The generation process of our DSG scheme.

Input: pretrained model M with N BN layers, training iterations T.
Output: synthetic data: $\mathbf{x}^s$
Initialize $\mathbf{x}^s$ from Gaussian distribution $\mathcal{N}(0,1)$;
Initialize $\mathbf{x}^0$ from Gaussian distribution $\mathcal{N}(0,1)$;
Get $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ from BN layers of M, $i = 1, 2, \ldots, N$;
Forward propagate $\mathrm{M}(\mathbf{x}^0)$ and gather activations;
Compute δ_i and γ_i based on Eq. (3);
for *all* $t = 1, 2, \ldots, T$ **do**
  Forward propagate $\mathrm{M}(\mathbf{x}^s)$ and gather activations;
  Get $\tilde{\boldsymbol{\mu}}_i^s$ and $\tilde{\boldsymbol{\sigma}}_i^s$ from activations;
  Compute all $l_{\mathrm{SDA}\,i}$ based on Eq. (2);
  Compute $\mathcal{L}_{\mathrm{DSG}}$ based on Eq. (5);
  Descend $\mathcal{L}_{\mathrm{DSG}}$ and update $\mathbf{x}^s$;
Get synthesized data $\mathbf{x}^s$;

various distributions in one layer. This significantly improves the diversity of synthetic data, which is important to calibrate the quantized models. As shown in Figure 3, significant fluctuation exists in the distribution statistics of synthetic samples generated by the DSG scheme, which is closer to the behavior of real data and does not strictly fit the BN statistics.

Our DSG scheme applies both SDA and LSE to tackle the homogenization issues at both distribution and sample levels and generate diverse samples for accurate data-free quantization. Figure 2 shows the whole process of the DSG scheme, and it is summarized in Algorithm 1. Instead of imposing a strict constraint on all samples, we relax this constraint using SDA and introduce LSE to reinforce the loss of a specific layer for one sample to mitigate the homogenization at two levels. Therefore, synthetic data generated by the DSG scheme performs better in quantization than those generated by existing generative data-free quantization methods, even can take the place of real data in SOTA post-training quantization methods when real data is not available.

3.5. Analysis and Discussion

We further present discussion with visualization results for our DSG scheme. The distribution of statistics is shown in Figure 3, including that of the real data, ZeroQ data, and DSG data. The figure illustrates the homogenization at the aforementioned two levels.

As shown in Figure 3, there is a significant offset between feature statistics of DSG samples and the BN statistics, which is similar to the behavior of real data samples. While the feature statistics of ZeroQ synthetic samples almost obeys the BN statistics. This phenomenon proves that our DSG scheme diversifies the synthetic data at the distribution level. Specifically, the SDA slacks the constraint of statistics during the generation process to make the distribution diverse.

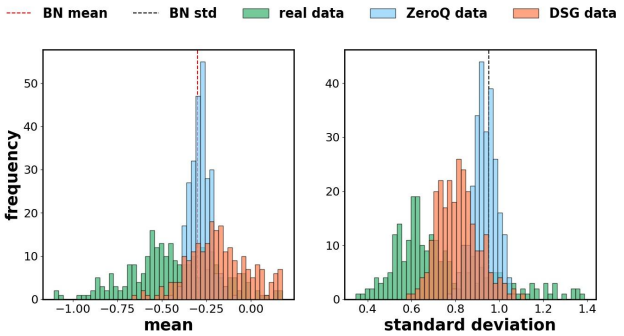


Figure 3: Mean and standard deviation of the activations in one channel of ResNet-18 when feeding different types of data (with 256 samples), including real data and synthetic data (generated by ZeroQ and DSG). Each sample generated by ZeroQ behaves similarly overfitting BN statistics compared with real data, which shows the homogenization at both distribution and sample levels. Our DSG data shows the diversity close to real data, which is the key to obtaining the accurate quantized model.

Moreover, with a larger variance of the statistic distribution on both mean and standard deviation, the statistics of DSG synthetic data are more disperse, which behave almost in the same as real data. On the contrary, statistics of ZeroQ data seem to be centralized. This phenomenon results from both SDA and LSE, which jointly enhance the data diversity at the sample level. The results imply that since the statistics are widely dispersed, the synthetic data might be more plentiful in feature and have more comprehensive content.

4. Experiments

In this section, we conduct experiments on the benchmark ImageNet dataset (ILSVRC12) [5] large-scale image classification tasks to verify the effectiveness of the proposed DSG scheme and compare it with other state-of-the-art (SOTA) data-free quantization methods.

DSG scheme: Our DSG scheme is implemented by PyTorch for its high flexibility and powerful automatic differentiation mechanism. The proposed DSG scheme is used for generating synthetic data for calibration, and the effectiveness is evaluated by measuring the accuracy of quantized models with various quantization methods, such as Percentile, EMA, and MSE. Also, we further apply the synthetic data generated by the DSG scheme to the state-of-the-art post-training quantization method: AdaRound [19].

Network Architectures: To prove the versatility of our DSG scheme, we employ various widely-used network architectures, including ResNet-18, ResNet-50, SqueezeNext, InceptionV3, and ShuffleNet. We also evaluate our DSG scheme with various bit-widths, including W4A4 (means 4-bit Weight and 4-bit Activation), W6A6, W8A8, *etc*.

Initialization: For fair comparison, when evaluating our

DSG scheme on various network architectures, we mostly follow the hyper-parameter settings (*e.g.*, the number of iterations to generate synthetic data) of their original papers or released official implementations. The data generated by our DSG scheme is initialized by Gaussian random initialization. We adopt Adam [15] as optimization algorithm in our experiments.

Method	W-bit	A-bit	Top-1
Baseline	32	32	71.47
Vanilla (ZeroQ)	4	4	26.04
Layerwise Sample Enhancement	4	4	27.12
Slack Distribution Alignment	4	4	33.39
DSG (Ours)	4	4	**34.53**

Table 1: Ablation study for DSG scheme on ResNet-18. We abbreviate quantization bits used for weights as "W-bit" (for activations as "A-bit"), top-1 test accuracy as "Top-1."

4.1. Ablation Study

We perform ablation studies to investigate the effect of components of the proposed DSG scheme, with the ResNet-18 model on the ImageNet dataset. We evaluate our method on W4A4 bit-width, which can reveal the effect of each part most obviously.

4.1.1 Effect of SDA

We first evaluate the effectiveness of SDA. As described in Section 3.3, the degree of the slack is adaptively determined by the value of ϵ in SDA. Thus, in Figure 4, we explore the impact of different values of ϵ in Eq. (3). Since ϵ is in the range of $(0, 1]$, we adopt a moderate interval, which is 0.1. Furthermore, we add $\epsilon = 0$ to complement the vanilla case that the constraints are not slacked.

Table 1 shows that if SDA is absent, the performance of the quantized network drops significantly by 7.41% compared with that using two methods as a whole, which shows that SDA is essential and even provides the major contribution to the final performance. In more detail, as shown in Figure 4, when ϵ increases from 0 to 0.9, the final performance of the quantized model increase steadily. However, when ϵ is set to 1, the accuracy suffers a huge drop. The phenomenon forcefully confirms that the distribution slack introduced by our SDA relaxes the constraints and allows the statistics of synthetic data to have a certain degree of offset, which enhances the diversity of features and consequently boosts the performance of the quantized model. Nevertheless, suppose ϵ is set to 1, which means that all the outliers in $\tilde{\boldsymbol{\mu}}_i^0$ and $\tilde{\boldsymbol{\sigma}}_i^0$ are taken into consideration, the degree of slack becomes out of bounds and the feature distribution of synthetic data might be far away from the reasonable range.

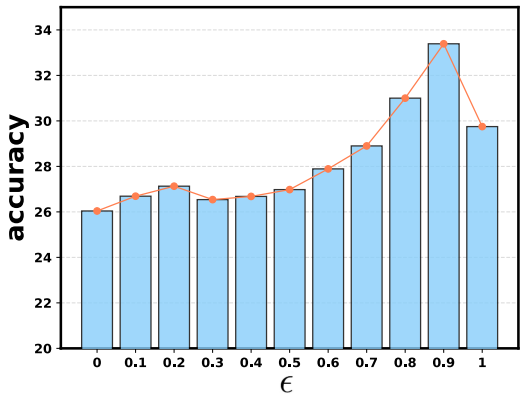


Figure 4: The accuracy comparison of different ϵ values in Eq. (3) on ResNet-18. As ϵ varies from 0 to 0.9, the final accuracy is mainly on the rise. But it suffers a significant drop caused by the outliers when $\epsilon = 1$.

Therefore, the default value of ϵ is set to 0.9 empirically to mitigate the impact of outliers.

4.1.2 Effect of LSE

Table 1 also shows that LSE contributes to the final performance. From the results, LSE achieves up to 1.08% improvement compared with ZeroQ, which is slight but non-negligible.

Experiment results demonstrate that these two methods are compatible and jointly boost the performance of the quantized model. From the results, the improvements brought by these two parts are superimposed because they focus on different root causes, and the processes do not interfere with each other. Consequently, quantized models calibrated with samples generated by our methods surpass vanilla competitors by 8.49%. In short, SDA prevents the synthetic samples from overfitting to BN statistics, and LSE makes the samples have different focuses on the statistics of specific layers.

4.2. Comparison with SOTA methods

We evaluate our DSG scheme on ImageNet dataset for large-scale image classification task, and analysis the performance over various network architectures, including ResNet-18 (Table 2(a)), ResNet-50 (Table 2(b)), SqueezeNext (Table 3(a)), InceptionV3 (Table 3(b)), and ShuffleNet (Table 3(c)). As mentioned above, we adopt SDA for all samples and apply LSE to part of them while generating synthetic data. The bit-width of the quantized model is marked as WwAa standing for w-bit weight and a-bit activation, which is set to W8A8, W6A6, W4A4, *etc*.

Our method outperforms previous state-of-the-art methods in various bit-width settings and is even better than those requiring real data to calibrate quantized models directly. In W8A8 cases, our method surpasses previous post-training quantization and quantization-aware training methods, such as DFC [11], DFQ [20], and RVQuant [22], as shown in the

(a) ResNet-18

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	71.47
Real Data	✗	✓	4	4	31.86
ZeroQ	✓	✓	4	4	26.04
DSG (Ours)	✓	✓	4	4	**34.53**
Real Data	✗	✓	6	6	70.62
Integer-Only	✗	✗	6	6	67.30
DFQ	✓	✓	6	6	66.30
ZeroQ	✓	✓	6	6	69.74
DSG (Ours)	✓	✓	6	6	**70.46**
Real Data	✗	✓	8	8	71.44
RVQuant	✗	✗	8	8	70.01
DFQ	✓	✓	8	8	69.70
DFC	✓	✗	8	8	69.57
ZeroQ	✓	✗	8	8	71.43
DSG (Ours)	✓	✓	8	8	**71.49**

(b) ResNet-50

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	77.72
Real Data	✗	✓	6	6	75.52
OCS	✗	✓	6	6	74.80
ZeroQ	✓	✓	6	6	75.56
DSG (Ours)	✓	✓	6	6	**76.07**
Real Data	✗	✓	8	8	77.70
ZeroQ	✓	✓	8	8	77.67
DSG (Ours)	✓	✓	8	8	**77.68**

Table 2: Quantization results of ResNet-18 and ResNet-50 on ImageNet. Here, "No D" means that none of the data is used to assist quantization, "No FT" stands for no fine-tuning (retraining). "Real Data" represents using real training data and quantization methods in ZeroQ (without any fine-tuning). Our DSG outperforms the SOTA data-free quantization methods and quantization methods requiring real data, such as ZeroQ, DFQ, DFC, Integer-Only [14], and OCS [40].

(a) SqueezeNext

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	69.38
Real Data	✗	✓	6	6	62.88
ZeroQ	✓	✓	6	6	39.83
DSG (Ours)	✓	✓	6	6	**60.50**
Real Data	✗	✓	8	8	69.23
ZeroQ	✓	✓	8	8	68.01
DSG (Ours)	✓	✓	8	8	**69.27**

(b) InceptionV3

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	78.80
Real Data	✗	✓	4	4	23.23
ZeroQ	✓	✓	4	4	12.00
DSG (Ours)	✓	✓	4	4	**34.89**
Real Data	✗	✓	6	6	77.96
ZeroQ	✓	✓	6	6	75.14
DSG (Ours)	✓	✓	6	6	**76.52**
Real Data	✗	✓	8	8	78.78
ZeroQ	✓	✓	8	8	78.70
DSG (Ours)	✓	✓	8	8	**78.81**

(c) ShuffleNet

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	65.07
Real Data	✗	✓	6	6	44.75
ZeroQ	✓	✓	6	6	39.92
DSG (Ours)	✓	✓	6	6	**44.88**
Real Data	✗	✓	8	8	64.52
ZeroQ	✓	✓	8	8	64.46
DSG (Ours)	✓	✓	8	8	**64.77**

Table 3: Quantization results of SqueezeNext, InceptionV3 and ShuffleNet on ImageNet.

bottom part in Table 2(a). And meanwhile, we consistently observe significant improvement over various network architectures, as shown in Table 2(b), Table 3(a), Table 3(b), and Table 3(c). For instance, our method outperforms ZeroQ on SqueezeNext by 1.26% on W8A8. However, advancement gets more apparent when it goes further to lower bit-width cases. On W6A6, our method significantly surpasses ZeroQ by more than 20% on SqueezeNext, and even outperforms real data on ShuffleNet by a slight 0.13%.

Moreover, we highlight that our DSG achieves significant improvement with even lower bit-width, *i.e.* W4A4 cases. As shown in Table 2(a), the performance of DSG on ResNet-18 is up to 34.53%, which is 8.49% higher than that of ZeroQ, 2.67% higher than that directly using real data for calibration. As well as on InceptionV3, our method achieves 34.89% accuracy surpassing ZeroQ and real data by 22.89% and 11.66%, respectively, which is quite a wide margin.

In short, sufficient experiments over various network architectures and different bit-widths demonstrate that the synthetic data generated by the proposed DSG scheme can significantly improve the performance of the quantized model. The results also suggest that the diversity of synthetic data is important for improving the quantized model. Especially

Method	No D	W-bit	A-bit	Quant	Top-1
Baseline	–	–	32	32	71.47
Real Data	✗	4	4	Vanilla	31.86
ZeroQ	✓	4	4	Vanilla	26.04
DSG (Ours)	✓	4	4	Vanilla	**34.53**
Real Data	✗	4	4	Percentile	42.83
ZeroQ	✓	4	4	Percentile	32.24
DSG (Ours)	✓	4	4	Percentile	**38.76**
Real Data	✗	4	4	EMA	42.67
ZeroQ	✓	4	4	EMA	32.31
DSG (Ours)	✓	4	4	EMA	**35.18**
Real Data	✗	4	4	MSE	41.45
ZeroQ	✓	4	4	MSE	39.39
DSG (Ours)	✓	4	4	MSE	**40.00**

Table 4: Uniform post-quantization on ImageNet with ResNet-18. We evaluation our DSG scheme on various post-training quantization methods (Percentile, EMA, MSE), and Vanilla means the quantization method adopted by ZeroQ, which simply obtains the quantizer by the maximum and minimum of the weight and activation.

Method	No D	Label	Image Prior	W-bit	A-bit	Top-1
Real Data	✗	✗	✗	3	32	64.16
ZeroQ	✓	✗	✗	3	32	49.86
DSG (Ours)	✓	✗	✗	3	32	**56.09**
DSG (Ours)	✓	✓	✗	3	32	**58.27**
DSG (Ours)	✓	✓	✓	3	32	**61.32**
Real Data	✗	✗	✗	4	32	68.42
ZeroQ	✓	✗	✗	4	32	63.86
DSG (Ours)	✓	✗	✗	4	32	**66.87**
DSG (Ours)	✓	✓	✗	4	32	**67.09**
DSG (Ours)	✓	✓	✓	4	32	**67.78**
Real Data	✗	✗	✗	5	32	69.21
ZeroQ	✓	✗	✗	5	32	68.39
DSG (Ours)	✓	✗	✗	5	32	**68.97**
DSG (Ours)	✓	✓	✗	5	32	**69.02**
DSG (Ours)	✓	✓	✓	5	32	**69.16**

Table 5: AdaRound on ImageNet with ResNet-18. We evaluate the DSG scheme on AdaRound, one of the SOTA methods of post-training quantization, which learns how to quantize weights using several batches of unlabeled samples. We adopt "Label" [12] and "Image Prior" [38] to evaluating our DSG scheme further.

when the model is quantized to lower bit-width, the advantages of diversity becomes even more obvious on the final performance.

We further evaluate the DSG scheme over various quantization methods on W4A4 to verify its versatility. Specifically, we implement Percentile, EMA, and MSE in conjunction with different data generation method. Table 4 shows that the synthetic data generated by the DSG scheme achieves leading accuracy regardless of specific quantization methods and surpasses ZeroQ while using Percentile, EMA, and MSE by 6.52%, 2.87%, and 0.61% respectively. The results also demonstrate that DSG scheme is robust and versatile to various conditions.

4.3. Evaluation with AdaRound

Besides the above-mentioned post-training quantization methods aiming to obtain optimal clipping values for activations, we further evaluate our DSG scheme with AdaRound [19], the novel post-training quantization method which proposes a rounding procedure for quantizing weights. We also conduct experiments in conjunction with other data generation methods. Generating data with labels [12] can extract the class information from the parameters of networks. And image prior [38] helps to steer synthetic data away from unrealistic images with no discernible visual information. More specifically, since AdaRound only aims to quantize weights, we preserve full precision for activations in the experiments. And we generate 1024 samples for AdaRound to learn the rounding scheme. The results are shown in Table 5, and our DSG scheme surpasses ZeroQ under all settings, especially when we quantize the weights to ultra-low bit-width (*i.e.* 3-bit). Combined with labels and image prior approach, our DSG scheme still maintains high accuracy.

5. Conclusion

We first revisit the data generation process in data-free quantization and demonstrate that homogenization exists at both distribution and sample level in the data generated by the previous data-free quantization method, which harms the accuracy of quantized models. Toward this end, we have represented a novel sample generation method for accurate data-free quantization, dubbed as Diverse Sample Generation (DSG) scheme, to mitigate the homogenization issue. The proposed DSG scheme consists of Slack Distribution Alignment (SDA) and Layerwise Sample Enhancement (LSE), which are tailored to the aforementioned two levels of homogenization respectively and jointly enhance the diversity of generated data. Extensive experiments on multiple network architectures and post-training quantization methods demonstrate the leading accuracy and versatility of the DSG scheme. Especially on ultra-low bit-width, such as W4A4, our method achieves even up to 22% improvement.

Acknowledgement This work was supported by National Natural Science Foundation of China (62022009, 61872021), Beijing Nova Program of Science and Technology (Z191100001119050), State Key Lab of Software Development Environment (SKLSDE-2020ZX-06), and SenseTime Research Fund for Young Scholars.

References

[1] Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. Post-training 4-bit quantization of convolution networks for rapid-deployment, 2019. 1, 2

[2] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13169–13178, 2020. 1, 2, 3

[3] Yoojin Choi, Jihwan Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation, 2020. 1, 2, 3

[4] Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3009–3018. IEEE, 2019. 1, 2

[5] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, and Fei Fei Li. Imagenet: a large-scale hierarchical image database. In *IEEE CVPR*, 2009. 5

[6] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge. *IJCV*, 2010. 1

[7] Amir Gholami, Kiseok Kwon, Bichen Wu, Zizheng Tai, Xiangyu Yue, Peter Jin, Sicheng Zhao, and Kurt Keutzer. Squeezenext: Hardware-aware neural network design, 2018. 2

[8] Ross Girshick. Fast r-cnn. In *IEEE ICCV*, 2015. 1

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR*, 2014. 1

[10] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning*, pages 1737–1746, 2015. 1, 2

[11] Matan Haroush, Itay Hubara, Elad Hoffer, and Daniel Soudry. The knowledge within: Methods for data-free model compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2020. 1, 2, 3, 6

[12] M. Haroush, I. Hubara, E. Hoffer, and D. Soudry. The knowledge within: Methods for data-free model compression. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8491–8499, 2020. 8

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, 2016. 1, 2

[14] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018. 1, 2, 7

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[16] Krizhevsky, Alex, Sutskever, Ilya, Hinton, and E. Geoffrey. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 2017. 1

[17] Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. Fully quantized network for object detection. In *IEEE CVPR*, 2019. 1

[18] Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021. 1

[19] Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization, 2020. 1, 2, 5, 8

[20] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *IEEE ICCV*, 2019. 1, 2, 6

[21] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *IEEE CVPR*, 2019. 1

[22] Eunhyeok Park, Sungjoo Yoo, and Peter Vajda. Value-aware quantization for training and inference of neural networks, 2018. 6

[23] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *CoRR*, abs/1802.05668, 2018. 4

[24] Haotong Qin, Zhongang Cai, Mingyuan Zhang, Yifu Ding, Haiyu Zhao, Shuai Yi, Xianglong Liu, and Hao Su. Bipointnet: Binary neural network for point clouds. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[25] Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, and Nicu Sebe. Binary neural networks: A survey. *Pattern Recognition*, 2020. 1

[26] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song.

Forward and backward information retention for accurate binary neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 1

[28] Xu Shoukai, Li Haokun, Zhuang Bohan, Liu Jing, Cao Jiezhang, Liang Chuangrun, and Tan Mingkui. Generative low-bitwidth data free quantization. In *The European Conference on Computer Vision*, 2020. 2, 3

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE CVPR*, 2015. 1

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2

[32] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world, 2021. 1

[33] Yiru Wang, Weihao Gan, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *IEEE ICCV*, 2019. 1

[34] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *ICCV*, October 2019. 1

[35] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 138–146, 2020. 1

[36] Yudong Wu, Yichao Wu, Ruihao Gong, Yuanhao Lv, Ken Chen, Ding Liang, Xiaolin Hu, Xianglong Liu, and Junjie Yan. Rotation consistent margin loss for efficient low-bit face recognition. In *CVPR*, June 2020. 1

[37] Jie Yang, Jiarou Fan, Yiru Wang, Yige Wang, Weihao Gan, Lin Liu, and Wei Wu. Hierarchical feature embedding for attribute recognition. In *CVPR*, June 2020. 1

[38] Hongxu Yin, Pavlo Molchanov, Zhizhong Li, Jose M. Alvarez, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion, 2020. 8

[39] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices, 2017. 2

[40] Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Christopher De Sa, and Zhiru Zhang. Improving neural network quantization without retraining using outlier channel splitting. *arXiv preprint arXiv:1901.09504*, 2019. 1, 2, 7

[41] Feng Zhu, Ruihao Gong, Fengwei Yu, Xianglong Liu, Yanfei Wang, Zhelong Li, Xiuqi Yang, and Junjie Yan. Towards unified int8 training for convolutional neural network. In *CVPR*, June 2020. 1

[42] Bohan Zhuang, Chunhua Shen, Mingkui Tan, Lingqiao Liu, and Ian Reid. Structured binary neural networks for accurate image classification and semantic segmentation. In *IEEE CVPR*, 2019. 1