

Dynamic Normalization

Chuan Liu¹ Yi Gao¹ Jiancheng Lv¹

Abstract

Batch Normalization has become one of the essential components in CNN. It allows the network to use a higher learning rate and speed up training. And the network doesn't need to be initialized carefully. However, in our work, we find that a simple extension of BN can increase the performance of the network. First, we extend BN to adaptively generate scale and shift parameters for each mini-batch data, called DN-C (Batch-shared and Channel-wise). We use the statistical characteristics of mini-batch data ($E[X], Std[X] \in \mathbb{R}^c$) as the input of SC module. Then we extend BN to adaptively generate scale and shift parameters for each channel of each sample, called DN-B (Batch and Channel-wise). Our experiments show that DN-C model can't train normally, but DN-B model has very good robustness. In classification task, DN-B can improve the accuracy of the MobileNetV2 on ImageNet-100 more than 2% with only 0.6% additional Mult-Adds. In detection task, DN-B can improve the accuracy of the SSDLite on MS-COCO nearly 4% mAP with the same settings. Compared with BN, DN-B has stable performance when using higher learning rate or smaller batch size.

1. Introduction

Deep learning has achieved great success in many fields, such as computer vision and natural language processing. In the field of computer vision, there are many famous network architectures. LeNet (LeCun et al., 1998) network has laid the foundation for the development of CNN (convolutional neural network). AlexNet (Krizhevsky et al., 2017) is the winner of the 2012 ImageNet competition. AlexNet uses ReLU (Nair & Hinton, 2010; Jarrett et al., 2009), Dropout (Hinton et al., 2012; Srivastava et al., 2014) and other effective tricks, which affects the design of the following network

¹Sichuan University, Chengdu 610065, P. R. China. Correspondence to: Chuan Liu <liu.ca@qq.com>, Jiancheng Lv <lvjiancheng@scu.edu.cn>.

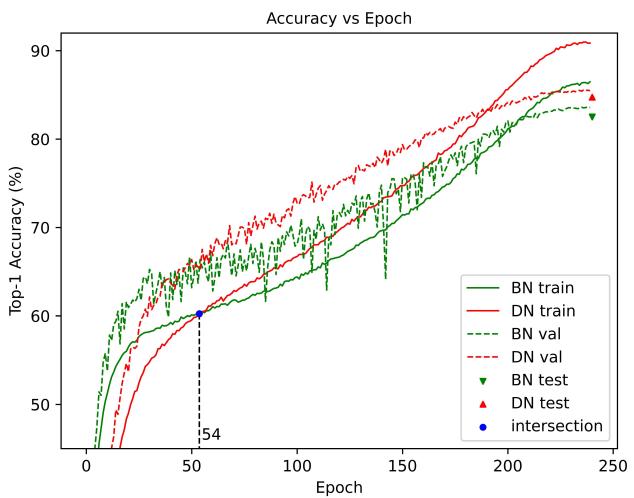


Figure 1. Accuracy vs Epoch on ImageNet-100, using MobileNetV2 (best viewed in color). Compared with BN, DN-B has higher test accuracy and more stable validation accuracy during training. r is 16, g is oup in DN-B.

architectures. ResNet (He et al., 2016) has achieved great success and it proposes an effective strategy to train deeper networks. VGG (Simonyan & Zisserman, 2014), GoogleNet (Szegedy et al., 2015; Ioffe & Szegedy, 2015; Szegedy et al., 2016) and other network architectures (Huang et al., 2017; Xie et al., 2017) also promote the development of CNN. With the rapid development of CNN model and mobile demand, there are some very powerful lightweight models. MobileNetV1 (Howard et al., 2017) achieves a good balance between accuracy and latency due to depthwise separable convolutions. MobileNetV2 (Sandler et al., 2018) utilizes the inverted residual with linear bottleneck. MobileNetV3 (Howard et al., 2019) combined with NAS (network architecture search) has achieved very good results. ShuffleNetV1 (Zhang et al., 2018) uses channel shuffle and group convolution to design the network architecture. ShuffleNetV2 (Ma et al., 2018) provides some guidelines for designing lightweight network structures.

The success of deep learning in computer vision is closely related to the popular components, such as normalization (Ioffe & Szegedy, 2015; Ba et al., 2016; Wu & He, 2018; Ulyanov et al., 2016). (Ioffe & Szegedy, 2015) find the phe-

nomenon of internal covariate shift and proposes the strategy of normalizing the data of mini-batch. BN (Ioffe & Szegedy, 2015) makes it easier to train neural networks, such as using a larger learning rate and it reduces the network initialization requirement. However, BN relies on the statistical characteristics of mini-batch data, the performance of network is not good when the batch size is small. GN (Group Normalization) (Wu & He, 2018) keeps its normalization separate from the batch dimension by grouping channels. With such a simple strategy, GN can keep the network performance even when the batch size is small. But the learnable weight parameters in BN and GN are the dimension of channel. In other words, the data in mini-batch share scale and shift parameters. The parameters are static and don't change with the input when the network finishes training.

In addition, some work (Hu et al., 2018; Yang et al., 2019; Ma et al., 2020; Zhao et al., 2020; Chen et al., 2020) shows that the performance of the dynamic module is better than the static module in network. SENet (Hu et al., 2018) utilizes SE (Squeeze-and-Excitation block) to adaptively learn and calibrate the relationship between channels. SENet has greatly improved the network performance at a small cost and the application scope is very wide. CondConv (Yang et al., 2019) extends the boundary of convolution kernel, so that convolution kernel can adaptively learn the corresponding convolution kernel weight for each sample. This strategy can improve the performance of the model under the condition of relatively balanced performance and cost. WeightNet (Ma et al., 2020) further unifies SENet and CondConv, reducing costs and improving the performance of the model. Dynamic ReLU (Chen et al., 2020) and APReLU (Zhao et al., 2020) believe that unified ReLU may limit the learning ability of the model. So Dynamic ReLU and APReLU adaptively learn the corresponding activation function for the sample, which greatly improves the performance of the model.

In this paper, We first let SC-Module adaptively learn the scale and shift parameters of BN (Ioffe & Szegedy, 2015). Then, we extend SC-Module to learn the scale and shift parameters for each sample. Because the scale coefficient in BN can also be considered as a measure of the importance of the channel, we think that Dynamic Normalization also has the channel recalibration ability of SENet (Hu et al., 2018) to some extent. And compared with using SENet and BN together, Dynamic Normalization can reduce the number of parameters and maintain the accuracy of the model. In addition, we set two adjustable coefficients, coefficient r (reduction ratio of channel) and coefficient g (number of channels per group), to balance the accuracy and cost of the model. Our experiments show that DN-B model has very good robustness. And DN-B can improve the accuracy of the MobileNetV2 (Sandler et al., 2018) on ImageNet-100 (Deng et al., 2009) more than 2% with only 0.6% additional

Mult>Adds. Compared with BN, DN-B has stable performance when using higher learning rate or smaller batch size.

2. Background and Related Work

Dynamic Network. Dynamic network has achieved good performance at present. The success of the dynamic network mainly comes from the adaptive adjustment of some components in the network. SENet (Hu et al., 2018) improves the performance of the model by adaptively recalibrating the channel for each sample. CondConv (Yang et al., 2019) enables the convolution kernel weights to be generated adaptively according to each sample. WeightNet (Ma et al., 2020) has unified SENet and CondConv in the weight space, reducing the number of parameter and improving the performance of the model. APReLU (Zhao et al., 2020) improves the learning ability of the model and the performance of the model by adaptively learning the activation function of the sample. Dynamic ReLU (Chen et al., 2020) proposes three different ways of dynamically learning activation functions: DY-ReLU-A (Spatial and Channel-shared), DY-ReLU-B (Spatial-shared and Channel-wise), DY-ReLU-C (Spatial and Channel-wise).

Attention. Attention has been widely used in CNN model, such as SENet (Hu et al., 2018) and CBAM (Woo et al., 2018) in computer vision. SENet is the winner of the 2017 ImageNet competition. SENet utilizes Squeeze-and-Excitation block to generate a weight coefficient for each channel of each sample, and then uses these weight coefficients to recalibrate the channels. CBAM uses attention in both channel and spatial dimensions, which achieves better performance than SENet. CBAM (Woo et al., 2018) adds max-pooling operation, which makes the output of max-pooling and ave-pooling operations share multi-layer perceptron (MLP) with one hidden layer in Channel Attention Module. Then CBAM concats the output of max-pooling and ave-pooling operations, and obtains the spatial attention coefficients through a convolution layer and activation function. And CBAM first recalibrates the channel dimension with Channel Attention Module, and then recalibrates the spatial dimension with Spatial Attention Module.

Normalization. Normalization (Ioffe & Szegedy, 2015; Ba et al., 2016; Wu & He, 2018; Ulyanov et al., 2016) is one of the essential components in deep learning, which promotes the development of deep learning. As one of the most commonly used normalization techniques in computer vision, BN (Ioffe & Szegedy, 2015) can speed up the training of neural network and avoid careful initialization to some extent. However, BN uses the statistical characteristics of mini-batch data. When the batch size is small, the network performance using BN is not good. So GN (Wu & He, 2018) groups channels so that it is independent on the batch dimen-

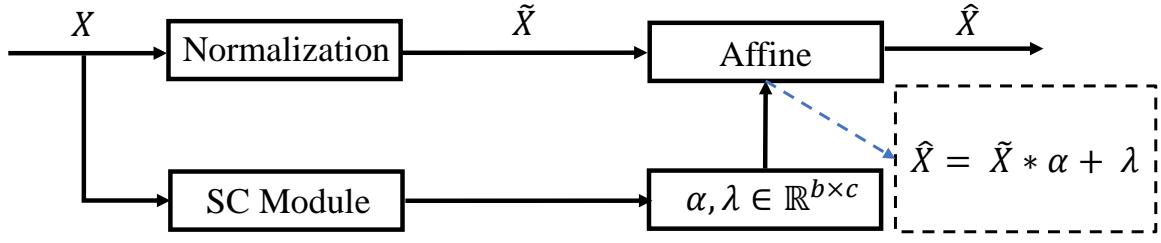


Figure 2. The DN-B structure. The scale and shift parameters are generated by SC-Module. See equation 6 for the details of Normalization module.

However, the scale and shift parameters of BN and GN are both channel dimensions and are shared in batch dimension. In other words, BN and GN are batch-shared and channel-wise. In this work, We first extend normalization to generate parameters adaptively for mini-batch data (DN-C, Batch-shared and Channel-wise). Then, we extend DN-B to generate parameters adaptively for each channel of each sample (DN-B, Batch and Channel-wise).

3. Dynamic Normalization

3.1. Batch Normalization

In this part, We will give a brief introduction to BN.

$$X = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} \quad (1)$$

where x denotes the input features. $E[x]$ denotes the mean of x , and $Var[x]$ denotes the variance of x , $E[x], Var[x] \in \mathbb{R}^c$. ϵ is the minimum to prevent division by zero.

Affine.

$$y = X * \gamma + \beta \quad (2)$$

where γ and β are learnable parametes, $\gamma, \beta \in \mathbb{R}^c$, and c denotes the channel number of the input feature.

3.2. SENet and WeightNet

Squeeze-and-Excitation (SE) block (Hu et al., 2018) is powerful and plug and play. SE first uses pooling operation to compress the input features in spatial dimension. And uses a fully connected layer to compress the features in channel dimension. Then a fully connected layer is used to restore the channel dimension. Finally, the final channel coefficients are obtained through Sigmoid function, and the channels are recalibrated with the coefficients.

WeightNet (Ma et al., 2020) improves the performance of the network by unifying SE (Hu et al., 2018) and CondConv

(Yang et al., 2019) in the weight space. WeightNet replaces the two fully connected layers in SE with $1 * 1$ convolution operations (FC and Grouped FC). By using convolution, the grouped fully connected operation can be realized. At the same time, it also reduces the amount of parameters. Through these strategies, WeightNet can adaptively generate weight parameters for each sample. It enhances the learning ability of the network, and can well balance the cost and accuracy of the model. In our implementation of SC-Module, we also use FC and Grouped FC to generate our scale and shift parameters.

3.3. Dynamic Normalization Structure

In this section, we will introduce the Dynamic Normalization with batch dimension (DN-B), without batch dimension (DN-C). If we use SE block before BN, we can get equation 3, 4 and 5.

$$x = \alpha * X \quad (3)$$

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} * \gamma + \beta \quad (4)$$

$$= \frac{x - E[\alpha * X]}{\sqrt{Var[\alpha * X] + \epsilon}} * \gamma + \beta \quad (5)$$

where X is the output feature through the convolution kernels, α denotes the the importance of each channel derived from the SE-Module. $\gamma, \beta \in \mathbb{R}^c$, and c denotes the channel number of the input feature.

$$\tilde{X} = \frac{X - E[X]}{\sqrt{Var[X] + \epsilon}} \quad (6)$$

$$\hat{X} = \tilde{X} * \alpha + \lambda \quad (7)$$

where X is the output feature through the convolution kernels, α denotes the the scale parameter of each channel derived from the SC-Module. In DN-B, $\alpha, \lambda \in \mathbb{R}^{n, c}$, n denotes the batch size, and c denotes the channel number of the input feature. And in DN-C, $\alpha, \lambda \in \mathbb{R}^c$. See Figure 2 for details of DN.

DN-B and DN-C. First, we extend BN to adaptively generate scale and shift parameters for each mini-batch data, called DN-C (Batch-shared and Channel-wise). We use the statistical characteristics of mini-batch data ($E[X], Std[X] \in \mathbb{R}^c$) as the input of SC-Module. We can adaptively generate scale and shift parameters through SC-Module. Finally, through the Affine operation, we get final output.

Then we extend BN to adaptively generate scale and shift parameters for each channel of each sample, called DN-B (Batch and Channel-wise). We use the original input features as the input of SC-Module to generate the scale and shift parameters ($\alpha, \lambda \in \mathbb{R}^{n,c}$). Finally, through the Affine operation, we get final output. See Figure 2 for details of DN-B.

4. Experiments

In this section, we evaluate DN operation on ImageNet-100 dataset¹ (Deng et al., 2009) and MS-COCO dataset (Lin et al., 2014). On the ImageNet-100 dataset, we chose the lightweight MobileNetV2 (Sandler et al., 2018) and ResNet-18/34 as the baseline network. On the MS-COCO dataset, we chose the lightweight detection system, SSDLite (Liu et al., 2016). And we use MobileNetV2 as the backbone network of SSDLite.

Table 1. The accuracy of the MobileNetV2 with different modules on ImageNet-100.

batch size=64, epoch=240, Lr=0.2								
Module	r	g	#p	Mult-Add	Train	Val	Test	
BN	X	X	2.35M	299.62M	86.48	83.61	82.50	
DN-C ¹	16	X	3.03M	301.02M	NAN	NAN	X	
DN-C ²	16	X	3.03M	301.05M	NAN	NAN	X	
DN-B	16	1	3.03M	300.34M	91.62	85.63	84.08	
DN-B ³	16	oup	4.36M	301.66M	90.86	85.57	84.74	

¹ Add the $E[X]$ branch and the $Std[X]$ branch after the first FC. The learning rate is 0.1.

² Add the $E[X]$ branch and the $Std[X]$ branch after the second FC. The learning rate is 0.1.

³ The SC-Module in this experiment is implemented by two full connection operations.

4.1. The performance of MobileNetV2 on ImageNet-100

Our implementation code is based on DenseNAS (Fang et al., 2020b). In order to prove the effectiveness of our method, we directly replace BN operation in MobileNetV2

¹We randomly select 100 classes in the ImageNet dataset (Deng et al., 2009) called ImageNet-100, where twenty percent of the selected data is used as the validation set. And the original validation set is used as the test set.

Table 2. The accuracy of the MobileNetV2 with different modules on ImageNet-100.

batch size=64, epoch=240, Lr=0.2								
Module	r	g	Lr	#p	Mult-Add	Train	Val	Test
BN	X	X	0.2	2.35M	299.62M	86.48	83.61	82.50
SE	16	X	0.2	3.73M	300.98M	89.96	83.90	82.00
DN-B	16	1	0.2	3.03M	300.34M	91.62	85.63	84.08
DN-B ¹	16	oup	0.2	4.36M	301.66M	90.86	85.57	84.74

¹ The SC-Module in this experiment is implemented by two full connection operations.

Table 3. The accuracy of the ResNet with different modules on ImageNet-100.

ResNet18/ResNet-34, Lr=learning rate, b=batch size								
res-18	r	g	Lr	#p	Mult-Add	Train	Val	Test
BN	X	X	0.2	11.22M	1730.12M	94.57	85.89	85.00
BN	X	X	0.5	11.22M	1730.12M	93.35	86.07	84.64
BN	X	X	0.8	11.22M	1730.12M	88.61	84.75	82.88
DN-B	4	oup	0.2	12.52M	1731.43M	94.19	85.58	84.36
DN-B	4	oup	0.5	12.52M	1731.43M	93.05	85.75	85.34
DN-B	16	oup	0.8	11.54M	1730.45M	92.56	85.85	84.78
res-34	r	g	b	#p	Mult-Add	Train	Val	Test
² BN	X	X	64	21.33M	3579.81M	94.98	86.89	85.84
³ BN	X	X	8	21.33M	3579.81M	83.76	83.51	82.32
⁴ BN	X	X	4	21.33M	3579.81M	47.79	54.10	53.20
⁴ DN-B	16	oup	4	21.85M	3580.35M	75.14	78.41	77.10
³ DN-B	16	oup	8	21.85M	3580.35M	84.86	84.65	83.50
² DN-B	16	oup	64	21.85M	3580.35M	95.02	86.97	85.42
² DN-B	32	oup	64	21.58M	3580.08M	95.17	86.72	85.96

¹ The SC-Module in this experiment is implemented by two full connection operations.

² The models take 240 epochs.

³ The models only take 120 epochs.

⁴ The models only take 60 epochs.

with DN operation. First, we want BN to be dynamic, so we use the statistical characteristics of data in mini-batch as the input of SC-Module. But from table 1, When we use DN-C directly, we find that the model can't train normally. Therefore, we further extend DN-C to generate scale and shift parameters for each sample. From table 2, we can see that DN-B improves the accuracy of the model with less parameter cost. At the same time, we use SE block directly in front of BN, which reduces the accuracy of the model, thus proving that our method is not just an extension of SE block. In addition, we also study the influence of other parameters on MobileNetV2 with DN-B, such as r, g, learning rate and batch size. We find that compared with BN, DN-B has better robustness, and the model can still maintain better accuracy even when using the higher learning rate or smaller batch size. For details of the experiment, see 4.4.

Table 4. The influence of batch size.

MobileNetV2, ImageNet-100, Lr=0.2, DN-B						
BN/DN-B	r and g	#p	Mult-Adds	Train acc(last)	Val acc(best)	Test acc
BN-8 ¹	X	2.35M	299.62M	1.00/4.95	2.50/12.04	X
DN-8 ¹	16, 1	3.03M	300.34M	73.81/90.81	79.60/94.01	78.42/92.92
BN-64 ²	X	2.35M	299.62M	86.48/96.45	83.61/95.54	82.50/94.80
DN-64 ²	16, 1	3.03M	300.34M	91.62/97.74	85.63/95.92	84.08/94.78
MobileNetV2, ImageNet-100, Lr=0.2, DN-B						
BN/DN-B	r and g	#p	Mult-Adds	Train acc(last)	Val acc(best)	Test acc
BN-64-1 ³	X	2.35M	299.62M	86.48/96.45	83.61/95.54	82.50/94.80
DN-64-1 ³	16, 1	3.03M	300.34M	73.81/90.81	79.60/94.01	83.92/94.94
BN-64-2	X	2.35M	299.62M	86.48/96.45	83.61/95.54	82.50/94.80
DN-64-2	16, 1	3.03M	300.34M	91.62/97.74	85.63/95.92	84.02/94.90
BN-64-32	X	2.35M	299.62M	86.48/96.45	83.61/95.54	82.50/94.80
DN-64-32	16, 1	3.03M	300.34M	91.62/97.74	85.63/95.92	84.08/94.78

¹ BN-8 and DN-8 denote the training batch size is 8. The model takes 120 epochs.

² BN-64 and DN-64 denote the training batch size is 64. The model takes 240 epochs.

³ BN-64-1 and DN-64-1 denote the training batch size is 64, test batch size is 1.

4.2. The performance of ResNet on ImageNet-100

Our implementation code is based on DenseNAS (Fang et al., 2020b). Using the higher learning rate on ResNet-18, we find that the model with DN can maintain better accuracy. In addition, we use smaller batch size on ResNet-34 and find that the model using DN has higher accuracy. These experiments are consistent with the performance of DN in MobileNetV2, which proves the robustness of DN.

Table 5. Object detection on MS-COCO.

batch size=64, GPU=2						
Module	r	g	epoch	#p	MAdds	mAP(%)
BN	X	X	30	4.32M	0.81GB	15.3
DN-B ¹	16	oup	30	6.33M	0.81GB	19.2

¹ The SC-Module in this experiment is implemented by two full connection operations.

4.3. Object Detection

We verify the effectiveness of our method on the MS-COCO dataset (Lin et al., 2014). We use a lightweight detection system, SSDLite (Liu et al., 2016; Sandler et al., 2018). For convenience, our code is based on MMDetection (Chen et al., 2019) and FNA (Fang et al., 2020a). First, we train the MobileNetV2-BN and the MobileNetV2-DN-B with the same settings on ImageNet-100. Then, we use these two models as the backbone network of SSDLite. To make a fair comparison, instead of SyncBN, we use BN. We only train 30 epochs, the initial learning rate is 0.05 and decays at 18, 25, 28 epochs. The batch size of each GPU is 64. From the tabel 5 and figure 3, we can see using DN can get better accuracy. Note that all settings are the same except

that the backbone network uses DN-B operation instead of BN operation.

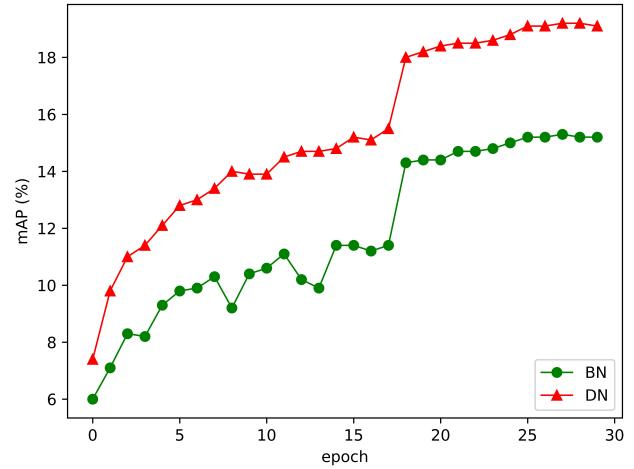


Figure 3. mAP vs Epoch on MS-COCO, using SSDLite (best viewed in color). Compared with BN, DN-B has better accuracy. r is 16, g is oup in DN-B.

4.4. Ablation study and Analysis

We set two coefficients in SC-Module to balance accuracy and cost of the module. In this part, we give the results of models when using different values of r and g. In addition, we also give the results of larger learning rate and different batch size.

The influence of r. To balance the accuracy and cost of the model, we add a parameter r in the first FC. This is the same as SENet, but we use convolution operation to implement

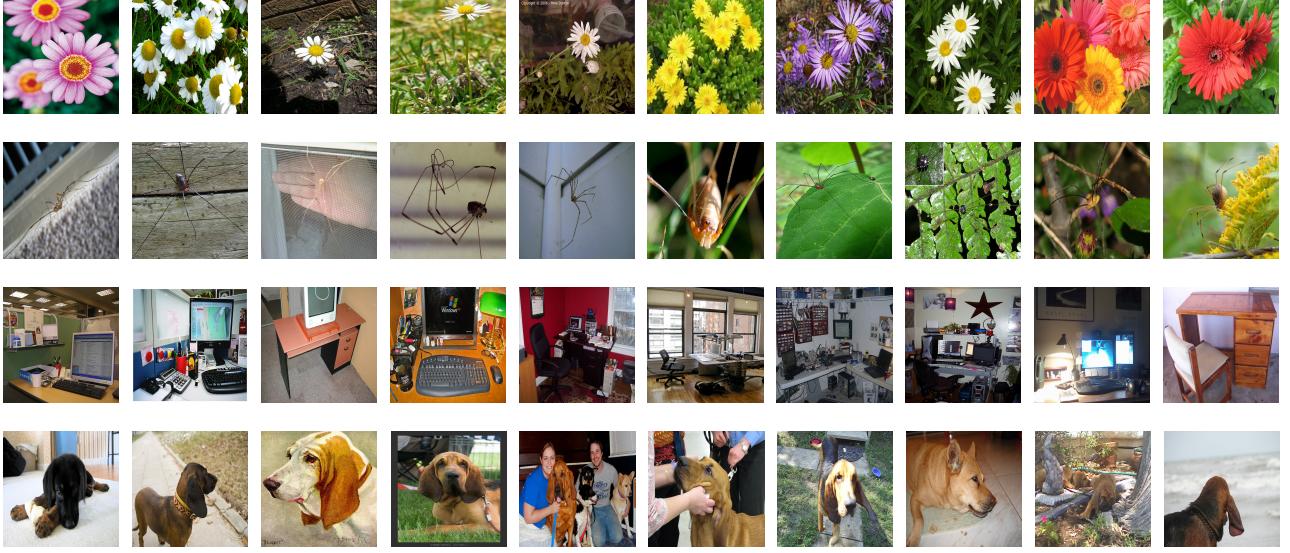


Figure 4. The 4 classes randomly selected from ImageNet-100 validation set. For each class, we randomly show 10 figures. The top-down class numbers are n11939491, n01770081, n03179701, n02088466.

FC. From the table 6, We can see that when r takes different values, it can balance accuracy and cost. Normally, when r is smaller, the parameters of the model increase and the accuracy of the model improves.

Table 6. The influence of r. The accuracy of the MobileNetV2 with different r on ImageNet-100.

DN-B, batch size=64, lr=0.2, epoch=240						
r	g	#p	Mult-Adds	Train	Val	Test
8	1	3.72M	301.02M	91.20	85.30	84.26
16	1	3.03M	300.34M	91.62	85.63	84.08
32	1	2.69M	300.00M	91.28	85.34	83.36

Table 7. The influence of g. The accuracy of the MobileNetV2 with different g on ImageNet-100.

DN-B, batch size=64, lr=0.2, epoch=240						
r	g	#p	Mult-Adds	Train	Val	Test
16	1	3.03M	300.34M	91.62	85.63	84.08
16	2	3.07M	300.37M	90.81	84.08	82.42
16	4	3.14M	300.44M	90.94	83.95	82.32
16 ¹	oup	4.36M	301.66M	90.86	85.57	84.74

¹ The SC-Module in this experiment is implemented by two full connection operations.

The influence of g. Our second FC in SC-Module is implemented by group convolution operation. This is the same as WeightNet, but our g parameter represents that the number of channels in each group is g. From table 6, we can see when g is the number of output channel, the model has the

highest accuracy. When g is 1, the accuracy of the model with DN-B is still higher than that of the model with BN, and only a few parameters are added.

The influence of batch size. We also verified the robustness of DN-B under different batch sizes. When we reduce the batch size to 8 and train only 120 epochs, we find that the BN model can not train normally. However, the model using DN operation can maintain the accuracy for normal training. In addition, we also study the influence of batch size on the test accuracy of model using DN-B operation. We find that when the small batch size is used for model inference, the accuracy of model using DN-B decreases very little. For details of the results, see tabel 4.

Table 8. The influence of learning rate. The accuracy of the MobileNetV2 with different learning rate on ImageNet-100.

DN-B, batch size=64, epoch=240						
BN/DN-B	Lr	#p	Mult-Adds	Train	Val	Test
BN	0.2	2.35M	299.62M	86.48	83.61	82.50
BN	0.5	2.35M	299.62M	71.43	75.85	74.12
DN(16, 1)	0.2	3.03M	300.34M	91.62	85.63	84.08
DN(16, 1)	0.5	3.03M	300.34M	88.69	84.60	83.24

The influence of learning rate. As one of BN's advantages is that the network can use a relatively high learning rate, we carry out some experiments on a higher learning rate to verify the effectiveness of our method. When we increase the learning rate to 0.5, we find that the training accuracy of the model using BN decreases by 15+%, and the test accuracy decreases by 8+%. However, when the DN model

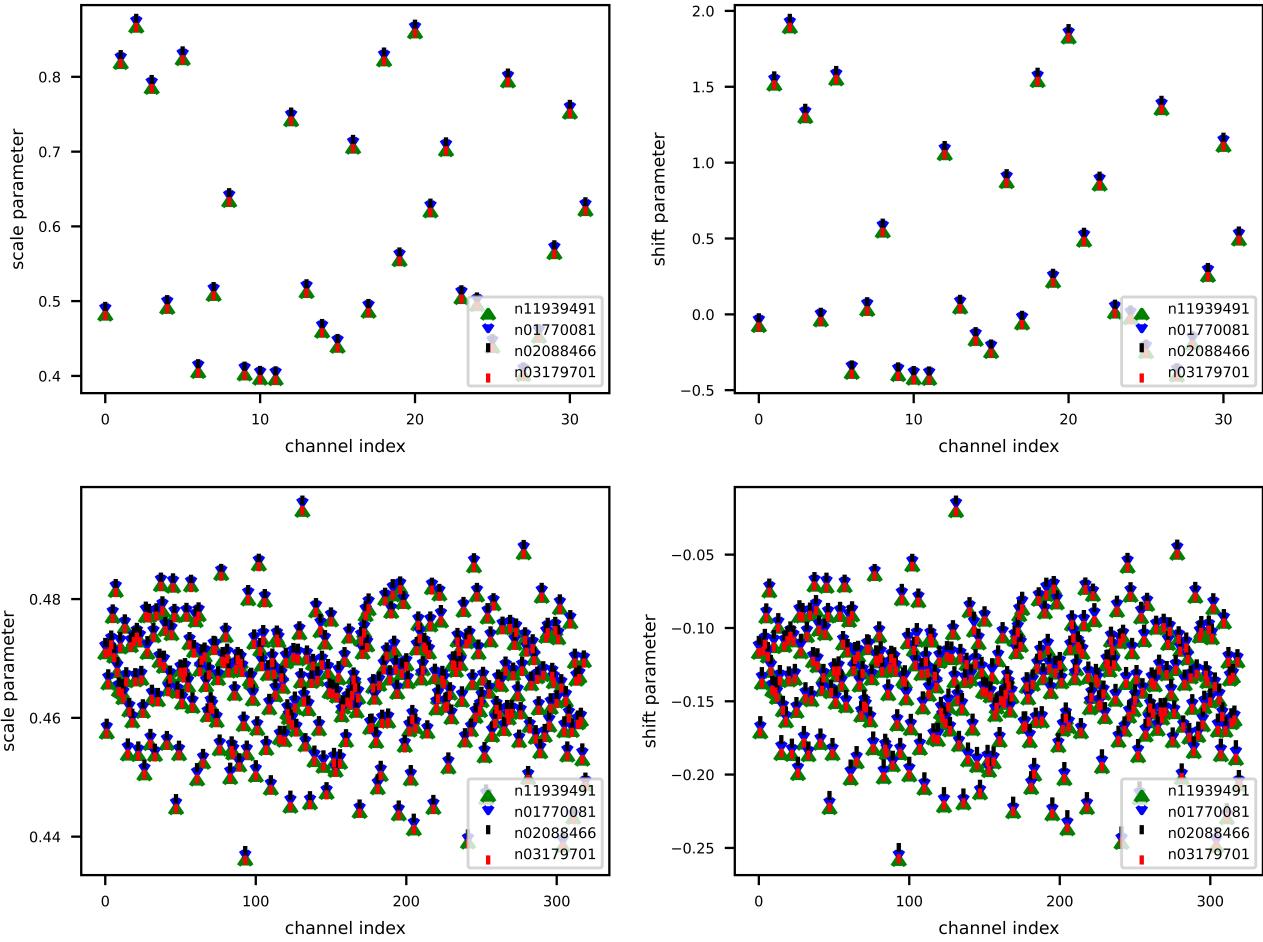


Figure 5. The affine parameters (best viewed in color). The top are the parameters of the DN-start operation in MobileNetV2 near the input. And the down are the parameters of the DN-end in MobileNetV2 near the output.

also increases the learning rate to 0.5, the training accuracy of the model only decreases by 2+%, and the test accuracy only decreases by 0.84%. And the DN-B model with higher learning rate has better accuracy than the BN model with lower learning rate.

Table 9. The test accuracy of the MobileNetV2 with DN-B on the 4 classes in figure 4.

DN-B (16, oup)				
number	module	#p	Mult>Adds	Test
n11939491	DN(16,out)	4.36M	301.66M	96.00
n01770081	DN(16,out)	4.36M	301.66M	90.00
n03179701	DN(16,out)	4.36M	301.66M	74.00
n02088466	DN(16,out)	4.36M	301.66M	68.00

The visualization of Scale and Shift parameters. We randomly select 4 classes in ImageNet-100 and infer them

based on MobileNetV2 with DN-B (16, 1) to get the parameters of affine operation. The 4 classes are shown in Figure 4 and the test accuracy is shown in Table 9. The two DN operations we selected are at the beginning and the end of MobileNetV2, called DN-start and DN-end. From Figure 5, We find that the mean value of affine parameters between classes is almost the same, that is to say, DN can adaptively learn the function of BN in some layer of the network. Note that in DN-end operation, the number of channels is 1280. For visualization, we evenly display 320 channels at 4 intervals.

5. Discussion and Future Work

In this work, we extend the static normalization to the dynamic normalization via SC-Module. We first extend normalization to generate parameters adaptively for mini-batch data (DN-C, Batch-shared and Channel-wise). Then, we extend DN-B to generate parameters adaptively for each

channel of each sample (DN-B, Batch and Channel-wise). Our experiments show that DN-C model can't train normally, but DN-B model has very good robustness. And DN-B adds few parameters and gets better accuracy. In addition, compared with BN, DN-B has stable performance when using higher learning rate or smaller batch size. In this way, DN-B enhances the capability of BN and can be used as an alternative to BN in some cases. In the future, we will combine DN with Layer Norm (Ba et al., 2016), Instance Norm (Ulyanov et al., 2016) and Group Norm (Wu & He, 2018) to explore other possibilities of DN in RNN/LSTM or GAN models.

References

- Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., and Liu, Z. Dynamic relu. *arXiv preprint arXiv:2003.10027*, 2020.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.
- Fang, J., Sun, Y., Peng, K., Zhang, Q., Li, Y., Liu, W., and Wang, X. Fast neural network adaptation via parameter remapping and architecture search. *arXiv preprint arXiv:2001.02525*, 2020a.
- Fang, J., Sun, Y., Zhang, Q., Li, Y., Liu, W., and Wang, X. Densely connected search space for more flexible neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10628–10637, 2020b.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1314–1324, 2019.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th international conference on computer vision*, pp. 2146–2153. IEEE, 2009.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. Ssd: Single shot multibox detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.
- Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.
- Ma, N., Zhang, X., Huang, J., and Sun, J. Weightnet: Revisiting the design space of weight networks. *arXiv preprint arXiv:2007.11823*, 2020.
- Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Woo, S., Park, J., Lee, J.-Y., and So Kweon, I. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Wu, Y. and He, K. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Yang, B., Bender, G., Le, Q. V., and Ngiam, J. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems*, 32:1307–1318, 2019.

Zhang, X., Zhou, X., Lin, M., and Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.

Zhao, M., Zhong, S., Fu, X., Tang, B., Dong, S., and Pecht, M. Deep residual networks with adaptively parametric rectifier linear units for fault diagnosis. *IEEE Transactions on Industrial Electronics*, 2020.