

# SplitNet: Divide and Co-training

Shuai Zhao<sup>2,3</sup>, Liguang Zhou<sup>1,2</sup>, Wenxiao Wang<sup>3</sup>, Deng Cai<sup>3</sup>, Tin Lun Lam<sup>1,2\*</sup>, Yangsheng Xu<sup>1,2</sup>

<sup>1</sup>Robotics and Artificial Intelligence Laboratory, The Chinese University of Hong Kong, Shenzhen

<sup>2</sup>Shenzhen Institute of Artificial Intelligence and Robotics for Society

<sup>3</sup>State Key Lab of CAD&CG, College of Computer Science, Zhejiang University

zhaoshuaimcc@gmail.com, {dcai, wenxiaowang}@zju.edu.cn, {liguangzhou@link., tllam@, ysxu@}cuhk.edu.cn

## Abstract

The width of a neural network matters since increasing the width will necessarily increase the model capacity. However, the performance of a network does not improve linearly with the width and soon gets saturated. To tackle this problem, we propose to increase the number of networks rather than purely scaling up the width. To prove it, one large network is divided into several small ones, and each of these small networks has a fraction of the original one's parameters. We then train these small networks together and make them see various views of the same data to learn different and complementary knowledge. During this co-training process, networks can also learn from each other. As a result, small networks can achieve better ensemble performance than the large one with few or no extra parameters or FLOPs. This reveals that the number of networks is a new dimension of effective model scaling, besides depth/width/resolution. Small networks can also achieve faster inference speed than the large one by concurrent running on different devices. We validate the idea — increasing the number of networks is a new dimension of effective model scaling — with different network architectures on common benchmarks through extensive experiments. The code is available at <https://github.com/mzhaoshuai/SplitNet-Divide-and-Co-training>.

## 1. Introduction

Width matters. Increasing the neural network's width to pursue better performance is common sense in neural network architecture design [52, 57, 55]. However, the performance of a network does not improve linearly with its width. As shown in Fig. 1, at the initial stage, increasing width can gain promising improvement of accuracy; at the later stage, the improvement becomes relatively slight and no longer matches the increasingly expensive cost. For example,

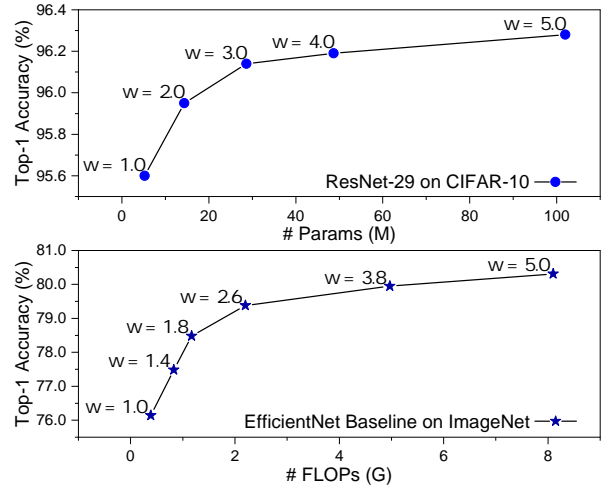


Figure 1: The width saturation of a network — the gain does not match the expensive extra cost when the width is already very large.  $w$  is the width factor. Data is shamelessly copied from ResNeXt [55] and EfficientNet [52].

in Fig. 1, EfficientNet baseline ( $w = 5.0$ , width factor of the network) gains less than +0.5% accuracy improvement compared to EfficientNet baseline ( $w = 3.8$ ) with nearly doubled FLOPs (floating point operations). Increasing the width of ResNet-29 produces a similar phenomenon. We call this the *width saturation of a network*.

Besides the width saturation, in Fig. 1, ResNet-29 ( $w = 3.0$ ) achieve relatively close accuracies as ResNet-29 ( $w = 5.0$ ). The situation is similar for EfficientNet baseline when ( $w = 2.6$ ) and ( $w = 5.0$ ). In this case, an interesting question arises, can two small networks with a half width of a large one achieve or even surpass the performance of the large one? Firstly, ensemble is always a practical technique [34, 13, 64, 1]. Secondly, we can train these two or more small models simultaneously [41, 62, 12], make them learn different and complementary knowledge about the data, and finally, achieve better ensemble performance.

\*Tin Lun Lam is the corresponding author.

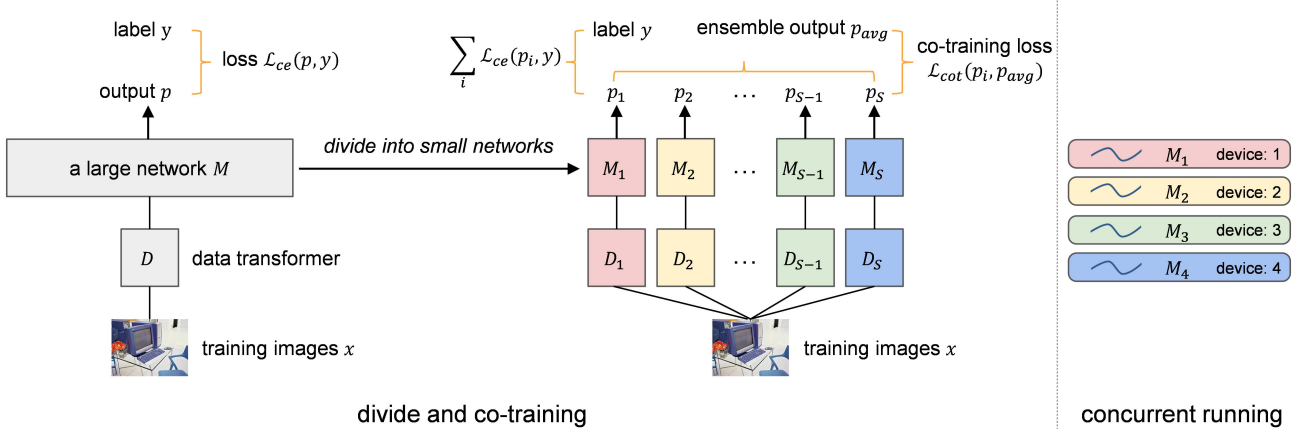


Figure 2: Left: division of one large network and co-training of several small ones. Right: concurrent running of small networks on different devices during inference.

Based on the above observation and conjecture, we argue that the number of networks should be a new dimension of effective model scaling, besides the commonly known depth, width, and resolution [52]. To prove it, we divide one large network into several small networks and show that these small ones can achieve better ensemble performance than the large one.

The overall framework of our method, *i.e.*, SplitNet, is shown in Fig. 2. Given one large network, we first divide the network according to its width, more precisely, the parameters or FLOPs of the network. For instance, if we want to divide a network into two small networks, the number of the small one’s parameters will become half of the original. After dividing, to make small networks learn different and complementary knowledge, they are trained with different views of the same data [41, 12]. This is implemented by applying different data augmentation transformers in practice. Furthermore, following the assumption of co-training [41], small networks are also expected to produce consistent predictions when seeing different views of the same data. Thus we add Jensen-Shannon divergence among all predictions, *i.e.*, co-training loss in Fig. 2. Besides, one network can also learn valuable knowledge about intrinsic object structure information from predicted posterior probability distributions of its peers [20, 62]. Finally, small networks can achieve better ensemble performance than the large one with few or no additional parameters or FLOPs.

Generally, small networks can also achieve faster inference speed than the original network through concurrent running on different devices when resources are sufficient. Due to the separability of small networks, they can achieve higher parallelism than a single large model when computational resources are adequate. Any tricks used to accelerate the large one can also be applied to one small model. Different networks and their average latency are shown in Fig. 3.

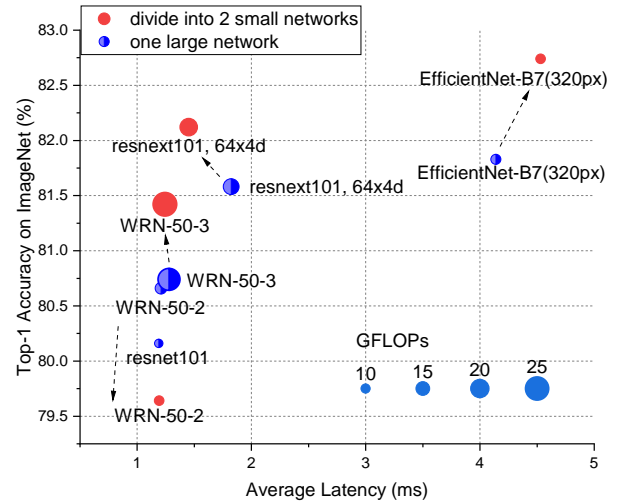


Figure 3: Networks and their inference latency. ResNet-101 is used as a speed reference. All networks are tested on Tesla V100(s) with mixed precision [36] and batch size 100.

The key contribution of our work is that we propose and conduct extensive experiments to validate the idea — the number of networks is a new dimension of effective model scaling. From the perspective of neural network architecture design, to our best knowledge, we are the first one who raises such a discussion. From the big picture, increasing the number of networks is a kind of ensemble learning. However, introducing ensemble learning into the process of network architecture design is new, which is also demonstrated to be more effective than purely scaling up width/depth/resolution.

Compared to the previous multi-head model [48] or layer splitting methods [50, 55, 61], we are discussing how to increase the performance of the model from a higher level, *i.e.*, model level *vs.* layer level or branch level. More similar works are deep mutual learning [62] and some other deep

collaborative learning or co-training methods [41, 12], which try to train two or more networks simultaneously to get better performance. The co-training process in our paper is highly inspired by them. We much appreciate their work, but apparently, they do not formally consider the possibility and discuss the consequent meaning of increasing the number of networks in the field of effective model scaling.

## 2. Related works

**Neural network architecture design** Since the success of AlexNet [25], deep learning methods dominate the field of computer vision and network engineering becomes a core topic. Many excellent architectures emerged since AlexNet, *e.g.*, NIN [30], VGG-Net [45], Inception [50], ResNet [17], Xception [7], *etc.* They explored different ways to design an effective and efficient model, *e.g.*,  $1 \times 1$  convolution kernels, stacked convolution layers with small kernels, combination of different convolution and pooling operations, residual connection, depthwise separable convolution, and so on.

In recent years, neural architecture search (NAS) has become more and more popular. People hope to automatically learn or search for the best neural architectures for certain tasks with machine learning methods. We just name a few here, reinforcement learning based NAS [65], progressive neural architecture search (PNASNet [31]), differentiable architecture search (DARTS [33]), *etc.*

**Collaborative learning** Originally, collaborative learning is an umbrella term in education, which refers to a variety of educational approaches involving joint intellectual effort by students, or students and teachers together [46]. It was formally introduced in deep learning by [48], which was used to describe the simultaneous training of multiple classifier heads of the same network. However, according to its original definition, many works involving two or more models learning together can also be called collaborative learning, *e.g.*, deep mutual learning (DML) [62], co-training [4, 41], mutual mean-teaching [12], cooperative learning [2], knowledge distillation [20], *etc.* Although they have different names, their core idea is similar, *i.e.*, enhancing the performance of one or all models by training them with some peers or teachers. They inspire the co-training part of our work.

## 3. Method

First of all, we simply recall the common framework of deep image classification. Given a neural network model  $M$  and  $N$  training samples  $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^N$  from  $C$  classes, the most commonly used objective function to train  $M$  is the cross entropy loss function:

$$\mathcal{L}_{ce}(p, y) = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C y_{n,c} \log(p_{n,c}), \quad (1)$$

where  $y \in \{0, 1\}$  is the ground truth label and  $p \in [0, 1]$  is the estimated probability, which is usually given by the last softmax layer of  $M$ .

### 3.1. Division

In Fig. 2, we divide one large network  $M$  into  $S$  small networks  $\{M_1, M_2, \dots, M_S\}$  according to its width. When talking about dividing according to width, we are de facto referring to divide according to the number of parameters or FLOPs. For example, if we want to divide  $M$  into two networks  $\{M_1, M_2\}$ , the number of parameters of  $M_1$  or  $M_2$  should be roughly half of  $M$ . Therefore, we first need to know how to calculate the number of parameters and FLOPs of a neural network.

$M$  is usually a stack of convolutional (conv.) layers, so we only discuss about how to calculate the number of parameters and FLOPs of a conv. layer. Following the definition of a conv. layer in PyTorch [40], its kernel size is  $K \times K$ , numbers of channels of input and output feature maps are  $C_{in}$  and  $C_{out}$ , and the number of groups is  $d$ , which means every  $\frac{C_{in}}{d}$  input channels will be convolved with its own sets of filters, of size  $\frac{C_{out}}{d}$ . In this case, the number of parameters and FLOPs of this conv. layer is

$$\# \text{ Params} : K^2 \times \frac{C_{in}}{d} \times \frac{C_{out}}{d} \times d = K^2 \times \frac{C_{in}}{d} \times C_{out}, \quad (2)$$

$$\text{FLOPs} : (2 \times K^2 \times \frac{C_{in}}{d} - 1) \times H \times W \times C_{out}, \quad (3)$$

where  $H \times W$  is the size of the output feature map and  $-1$  occurs because the addition of  $\frac{C_{in}}{d} \times K^2$  numbers only needs  $(\frac{C_{in}}{d} \times K^2 - 1)$  times operations. The bias is omitted for the sake of brevity. For depthwise convolution [44],  $d = C_{in}$ .

Generally,  $C_{out} = t_1 \times C_{in}$ , where  $t_1$  is a constant. Therefore, if we want to divide a conv. layer by a factor  $S$ , we just need to divide  $C_{in}$  by  $\sqrt{S}$ :

$$\frac{K^2 \times C_{in} \times C_{out}}{S} \times \frac{1}{d} = K^2 \times t_1 \times (\frac{C_{in}}{\sqrt{S}})^2 \times \frac{1}{d}, \quad (4)$$

*e.g.*, if we divide a bottleneck in ResNet [17]  $\begin{bmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 256 \end{bmatrix}$

by 4, it becomes 4 small  $\begin{bmatrix} 1 \times 1, & 32 \\ 3 \times 3, & 32 \\ 1 \times 1, & 128 \end{bmatrix}$  blocks, and each small block only has a quarter of the parameters and FLOPs of the original block.

In practice, the numbers of output channels of feature maps in a network have a greatest common divisor (GCD). The GCD of most ResNet variants [17, 57, 55, 21] is the  $C_{out}$  of the first convolutional layer. For other networks, like EfficientNet [52], their GCD is a multiple of 8 or some other numbers. In general, when we want to divide a network by  $S$ , we just need to find its GCD, and replace it with  $\text{GCD}/\sqrt{S}$ , then it is done. For networks like ResNeXt [55], when  $\frac{C_{in}}{d}$

is fixed, i.e.,  $C_{in} = t_2 \times d$ , where  $t_2$  is a constant, Eq. (4) turns to be

$$K^2 \times t_1 \times t_2^2 \times \frac{d}{S}. \quad (5)$$

It means we only need to divide the number of groups  $d$  by  $S$ . For networks like WRN [57], which have a large widen factor, we simply divide the widen factor by  $\sqrt{S}$ . Similarly, division of networks which have the concept of growth rate (DenseNet [22]) or additional rate (PyramidNet [15]) is achieved by dividing these rates by  $\sqrt{S}$ . If there are dropout [49] layers or other drop modules (e.g., ShakeDrop [56]) in the network, their drop ratio will be divided by  $\sqrt{S}$ . More details can be found in Appendix A.

As for weight decay, it is a little complex as its intrinsic mechanism is still vague. Two dividing strategies are used in this work: no dividing and exponential dividing:

$$wd^* = wd \times \exp\left(\frac{1}{S} - 1.0\right), \quad (6)$$

where  $wd$  is the original weight decay value and  $wd^*$  is the new value after dividing. No dividing means the weight decay value keeps unchanged. As mentioned above, the underlying mechanism of weight decay is unclear, so it is hard to find the best and universal solution. The above two dividing strategies are just empirical criteria. In practice, the best way now is trial and error. A detailed discussion about dividing weight decay can be found in Appendix B.2.

The above dividing mechanisms can be easily extended to any existed networks. However, the division of  $M$  is usually not perfect, i.e., the number of parameters of  $M_i$  may not be exactly  $1/S$  of  $M$ . Firstly,  $\sqrt{S}$  may not be an integer. In this case, we round  $C_{in}/\sqrt{S}$  in Eq. (4) to a nearby even number. Secondly, the division of the first layer and last fully-connected layer is not perfect because the input channel (color channels) and output channel (number of object classes) are fixed.

**Concurrent running of small networks** Despite better ensemble performance, small networks can also achieve faster inference speed than the large network in most cases. This is achieved by deploying different small models on different devices to make them run concurrently, as shown in Fig. 2. Typically, a device is an NVIDIA GPU. Theoretically, if one GPU has enough processing units, e.g., streaming multiprocessor (SM), CUDA cores, etc, small networks can also run concurrently within one GPU [39]. However, we sadly find this is impractical now, at least on a Tesla V100 GPU. One small network is already able to occupy most of the computational resources, and different networks can only run in sequence. Therefore, we only discuss the multiple devices fashion in this paper.

The success of concurrent inference of small networks also indicates the possibility of concurrency in training. Currently, small networks run in sequence during training, which leads to longer training time than the large one’s training time. However, it is pretty hard to design a flexible and scalable framework, which can support the asynchronous training of multiple models on multiple devices with possible communication during forward and backward processes. This is challenging and left as a future topic.

### 3.2. Co-training

After dividing, one large network  $M$  becomes  $S$  small networks  $\{M_1, M_2, \dots, M_S\}$ . The problem now is how to make these small networks learn different and complementary knowledge from the data. Before introducing the co-training part, we emphasize that dividing and co-training are just designed to demonstrate the core idea — *increasing the number of networks is a new dimension of effective model scaling*. The idea is obviously more vital than the tools used to demonstrate it. Besides, we state again that the co-training part is inspired by deep mutual learning (DML) [62], co-training [41], and mutual mean-teaching (MMT) [12].

**Different initialization and data views** A basic understanding is that learning some same networks is meaningless. In contrast, small networks are expected to learn different and complementary knowledge about the data to get a comprehensive picture. To this end, small networks are firstly initialized with different weights. Then, when feeding the training data, we apply different data transformers  $D_i$  on the same data for different networks  $M_i$ , as shown in Fig. 2. In this way,  $\{M_1, M_2, \dots, M_S\}$  are trained on different views  $\{D_1(\mathbf{x}), D_2(\mathbf{x}), \dots, D_S(\mathbf{x})\}$  of  $\mathbf{x}$ .

In practice, the different data views are generated by the randomness in the procedure of data augmentation. Besides the commonly used random resize/crop/flip, we further introduce the random erasing [63] and AutoAugment [8] policies. The AutoAugment has 14 image transformation operations, e.g., shear, translate, rotate, auto contrast, etc. It searched tens of different policies which are consisted of two transformation operations for different datasets and randomly choose one policy during the data augmentation process. Through applying these random data augmentation operations, we can guarantee that  $\{D_1(\mathbf{x}), D_2(\mathbf{x}), \dots, D_S(\mathbf{x})\}$  produce different views of  $\mathbf{x}$  in most cases.

**Co-training loss** Following the co-training assumption [41] in semi-supervised learning, small networks are expected to have consistent predictions on  $\mathbf{x}$  although they see different views of  $\mathbf{x}$ :

$$M_1(D_1(\mathbf{x})) = M_2(D_2(\mathbf{x})) = \dots = M_S(D_S(\mathbf{x})), \quad (\text{co-training assumption}). \quad (7)$$



Therefore, Jensen-Shannon (JS) divergence among predicted probability distributions is added into the objective function, *i.e.*, the co-training loss:

$$\mathcal{L}_{cot}(p_1, p_2, \dots, p_S) = H\left(\frac{1}{S} \sum_{i=1}^S p_i\right) - \frac{1}{S} \sum_{i=1}^S H(p_i), \quad (8)$$

where  $p_i$  is the estimated probability of one small network  $M_i$ , and  $H(p) = \mathbb{E}[-\log(p)]$  is the Shannon entropy of distribution  $p$ . The co-training loss is also used in DML [62], but its concrete form is different, *i.e.*, DML uses Kullback-Leibler (KL) divergence between two predictions. Through this co-training manner, one network can also learn something valuable from its peers because the predicted probability contains meaningful information about objects. For example, a model classifies an object as *Chihuahua* may also give high confidence about *Japanese spaniel* since they are both dogs. This is valuable information that defines a rich similarity structure over objects [20].

The overall objection function is:

$$\mathcal{L}_{all}(p_1, p_2, \dots, p_S, y) = \sum_{i=1}^S \mathcal{L}_{ce}(p_i, y) + \lambda_{cot} \mathcal{L}_{cot}(p_1, p_2, \dots, p_S), \quad (9)$$

where  $\lambda_{cot} = 0.5$  is the weight factor of  $\mathcal{L}_{cot}(\cdot)$  and it is chosen by cross-validation. Besides, the output of the network are full of randomness during the early stage of training, so we adopt the warm-up scheme for  $\lambda_{cot}$  [27, 41]. Specifically, we use a linear scaling up strategy when the current training epoch is less than a certain number, *i.e.*, 40/60 on CIFAR [24]/ImageNet [43].  $\lambda_{cot} = 0.5$  is also an equilibrium point between learning different networks and producing consistent predictions.

During inference, we simply use the averaged output as the ensemble output:

$$p_{avg} = \frac{1}{S} \sum_{i=1}^S p_i. \quad (10)$$

It is worth noting that  $p_i$  in Eq. (10) is the output of a model before the softmax operation. This is slightly different with what we discussed above, *i.e.*, softmax normalized probability. The same symbol  $p_i$  is used here just for brevity.

## 4. Experiment

### 4.1. Experimental setup

**Datasets** We adopt CIFAR-10, CIFAR-100 [24], and ImageNet 2012 [43] datasets. CIFAR-10 and CIFAR-100 datasets contain 50K training and 10K test RGB images of size  $32 \times 32$ , labeled with 10 and 100 classes, respectively.

ImageNet 2012 dataset contains 1.28 million training images and 50K validation images from 1000 classes.

**Learning rate and training epochs** We apply warm up and cosine learning rate decay policy [14, 19]. Specifically, if the initial learning rate is  $lr$  and current epoch is  $epoch$ , for the first  $slow\_epoch$  steps, the learning rate is  $lr \times \frac{epoch}{slow\_epoch}$ ; for the rest epochs, the learning rate is  $0.5 \times lr \times (1 + \cos(\pi \times \frac{epoch - slow\_epochs}{max\_epoch - slow\_epoch}))$ . Generally,  $lr$  is 0.1;  $\{max\_epoch, slow\_epoch\}$  is  $\{300, 20\}/\{120, 5\}$  for CIFAR/ImageNet, respectively.

**Batch size and crop size** For CIFAR/ImageNet, crop size is  $32/224$  and batch size is  $128/256$ , respectively. The output stride — the ratio of input image spatial resolution to final output resolution [5, 6], is  $4/32$  for CIFAR/ImageNet, respectively. This means the size of the output feature map before the last fully-connected (fc.) layer is  $8 \times 8$  or  $7 \times 7$ .

**Data augmentation** Random crop and resize (bicubic interpolation), random left-right flipping, AutoAugment [8], normalization, random erasing [63], and mixup [59] are used during training. Label smoothing [51] is only applied to models on ImageNet.

**Weight decay** Generally,  $wd = 1e-4$ . For {EfficientNet-B3, ResNeXt-29 ( $8 \times 64d$ ), WRN-28-10, WRN-40-10} on CIFAR,  $wd = 5e-4$ . It is only applied to weights of conv. and fc. layers [19, 14]. Bias and parameters of batch normalization [23] are left undecayed.

Besides, we use kaiming weight initialization [16]. The optimizer is nesterov [37] accelerated SGD with momentum 0.9. ResNet variants adopt the modifications introduced in [19], *i.e.*, replacing the first  $7 \times 7$  conv. with three consecutive  $3 \times 3$  conv. and put an  $2 \times 2$  average pooling layer before  $1 \times 1$  conv. when there is a need to downsample. The code is implemented in PyTorch [40]. Influence of some settings is shown in Tab. 1. The baseline is very solid.

Table 1: Influence of various settings of ResNet-110 on CIFAR-100. step-lr means step learning rate decay policy as described in ResNet [17]. When the erasing probability of random erasing  $p_e = 1.0$ , it acts like cutout [10].

Method	step-lr	cos-lr	random erasing	mixup	AutoAug	Top-1 err. (%)
	✓					24.71 ± 0.22
ResNet-110 [18]		✓				24.15 ± 0.07
original: 26.88%		✓	✓, $p_e = 1.0$			23.43 ± 0.01
		✓	✓, $p_e = 0.5$			23.11 ± 0.29
		✓	✓, $p_e = 0.5$	✓, $\lambda = 0.2$		21.22 ± 0.28
		✓	✓, $p_e = 0.5$	✓, $\lambda = 0.2$	✓	<b>19.19 ± 0.23</b>

### 4.2. Results on CIFAR dataset

#### Results on CIFAR-100 dataset

Results on CIFAR-100 are shown in Tab. 2. Division and co-training achieve consistent improvements with few extra or even fewer parameters or FLOPs. Additional cost occurs since the division of a network is not perfect, as mentioned in Sec. 3.1. Some conclusions can be drawn from the data.

Table 2: Results on CIFAR-100. The last three rows are trained for 1800 epochs.  $S$  is the number of small networks after dividing. Train from scratch, no extra data. Weight decay value keeps unchanged except {WRN-40-10, 1800 epochs}, which applies Eq. (6). DenseNet and PyramidNet are trained with mixed precision [36]. The smaller, the better.

Method	original	re-implementation				SplitNet ( $S = 2$ )			SplitNet ( $S = 4$ )		
	Top-1 err.	Top-1 err.	# params (M)	GFLOPs		Top-1 err.	# params (M)	GFLOPs	Top-1 err.	# params (M)	GFLOPs
ResNet-110 [18]	26.88	18.96	1.17	0.17		<b>18.32</b> <sub>(0.64 ↓)</sub>	1.33	0.20	19.56 <sub>(0.63 ↓)</sub>	1.21	0.18
ResNet-164 [18]	24.33	18.38	1.73	0.25		<b>17.12</b> <sub>(1.26 ↓)</sub>	1.96	0.29	18.05 <sub>(0.33 ↓)</sub>	1.78	0.26
SE-ResNet-110 [21]	23.85	17.91	1.69	0.17		<b>17.37</b> <sub>(0.54 ↓)</sub>	1.89	0.20	18.33 <sub>(0.42 ↓)</sub>	1.70	0.18
SE-ResNet-164 [21]	21.31	17.37	2.51	0.26		<b>16.31</b> <sub>(1.06 ↓)</sub>	2.81	0.29	17.21 <sub>(0.16 ↓)</sub>	2.53	0.27
EfficientNet-B0 [52] <sup>†</sup>	-	18.50	4.13	0.23		18.20 <sub>(0.30 ↓)</sub>	4.28	0.24	<b>17.85</b> <sub>(0.65 ↓)</sub>	4.52	0.30
EfficientNet-B3 [52]	-	18.10	10.9	0.60		17.00 <sub>(1.10 ↓)</sub>	11.1	0.60	<b>16.62</b> <sub>(1.48 ↓)</sub>	11.7	0.65
WRN-16-8 [57]	20.43	18.69	11.0	1.55		17.37 <sub>(1.32 ↓)</sub>	12.4	1.75	<b>17.07</b> <sub>(1.62 ↓)</sub>	11.1	1.58
ResNeXt-29, 8×64d [55]	17.77	16.43	34.5	5.41		14.99 <sub>(1.44 ↓)</sub>	35.4	5.50	<b>14.88</b> <sub>(1.55 ↓)</sub>	36.9	5.67
WRN-28-10 [57]	19.25	15.50	36.5	5.25		14.48 <sub>(1.02 ↓)</sub>	35.8	5.16	<b>14.20</b> <sub>(1.24 ↓)</sub>	36.7	5.28
WRN-40-10 [57]	18.30	15.56	55.9	8.08		14.28 <sub>(1.28 ↓)</sub>	54.8	7.94	<b>13.96</b> <sub>(1.60 ↓)</sub>	56.0	8.12
WRN-40-10 [57]	18.30	16.02	55.9	8.08		14.09 <sub>(1.93 ↓)</sub>	54.8	7.94	<b>13.10</b> <sub>(2.92 ↓)</sub>	56.0	8.12
DenseNet-BC-190 [22]	17.18	14.10	25.8	9.39		12.64 <sub>(1.46 ↓)</sub>	25.5	9.24	<b>12.56</b> <sub>(1.54 ↓)</sub>	26.3	9.48
PyramidNet-272 [15] + ShakeDrop [56]	14.96	11.02	26.8	4.55		10.75 <sub>(0.27 ↓)</sub>	28.9	5.24	<b>10.54</b> <sub>(0.48 ↓)</sub>	32.8	6.33

<sup>†</sup> When training on CIFAR-100, the stride of EfficientNet at the stage 4&7 is set to be 1. Original EfficientNet on CIFAR-100 is pre-trained while we train it from scratch.

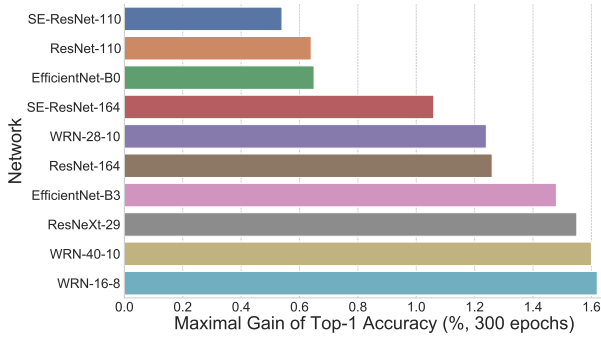


Figure 4: Network architecture and its maximal gain of accuracy with division and co-training on CIFAR-100.

**Conclusion 1** Increasing the number of networks is more effective than purely increasing the width/depth of a network.

For all networks in Tab. 2, division and co-training gain promising improvement. We also notice, ResNet-110 ( $S = 2$ ) > ResNet-164, SE-ResNet-110 ( $S = 2$ ) > SE-ResNet-164, EfficientNet-B0 ( $S = 4$ ) > EfficientNet-B3, and WRN-28-10 ( $S = 4$ ) > WRN-40-10 with fewer parameters, where > means the former has better performance. By contrast, the latter is deeper or wider. This exactly demonstrates the superiority of increasing the number of networks.

The relationship between network architectures and the maximal gain of accuracy with division and co-training is shown in Fig. 4. Generally, with wider or deeper networks, increasing the number of networks can gain more improvement, *e.g.*, ResNet-110 (+0.64) *vs.* ResNet-164 (+1.26), SE-ResNet-110 (+0.54) *vs.* SE-ResNet-164 (+1.06), WRN-28-10 (+1.24) *vs.* WRN-40-10 (+1.60), and EfficientNet-B0 (+0.65) *vs.* EfficientNet-B3 (+1.48). *It is clear that increasing the number, width, and depth of networks together is more efficient and effective than purely increasing the width and depth.*

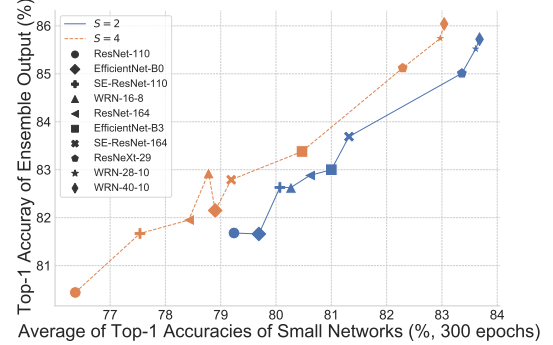


Figure 5: The average of accuracies and ensemble accuracy of small networks on CIFAR-100.

**Conclusion 2** The ensemble performance is closely related to individual performance.

The relationship between the average of accuracies and ensemble accuracy is shown in Fig. 5. When calculating the average accuracy of small networks, we first calculate the accuracy of small networks separately and then average them. The ensemble accuracy is calculated by Eq. (10). From the big picture, the average of accuracies is positively correlated with the ensemble accuracy. The higher the average of accuracies, the better the ensemble performance.

At the same time, we note that there is a big gap between the average of accuracies and ensemble accuracy, namely, the ensemble accuracy is higher than the average of accuracies by a large margin. This verifies our motivation in Sec. 1: several small networks can achieve better performance than only one large network through learning different and complementary knowledge. Only in this way, the small networks with relatively low performance can surpass the performance of the large one.

**Conclusion 3** A necessary width/depth of network matters.

In Tab. 2, ResNet-110 and SE-ResNet-110 with SplitNet ( $S = 4$ ) get a drop in performance, 0.63%↓ and 0.42%↓, respectively. In Fig. 5, their small networks also get relatively low accuracies. If we take one step further to look at the architecture of these small networks, we will find they have input channel numbers [8, 16, 32] at the first layer of their three blocks. These networks are too thin compared to the original one with [16, 32, 64] channels. Consequently, individual networks cannot gain satisfying accuracy and lead to inferior ensemble performance as conclusion 2 said. This demonstrates that a necessary width is essential.

In Tab. 2, ResNet-164 and SE-ResNet-164 with SplitNet ( $S = 4$ ) still get improvements, although their performance is not as good as applying ( $S = 2$ ). In Fig. 5, small networks of ResNet-164 and SE-ResNet-164 with ( $S = 4$ ) also get better performance than ResNet-110 and SE-ResNet-110 with ( $S = 4$ ), respectively. This reveals a necessary depth is also important, and it can help small networks gain necessary accuracies to achieve satisfying ensemble performance. The conclusion also applies to PyramidNet-272, which gets a relatively slight improvement with division and co-training as its base channel is small, *i.e.*, 16.

In a word, *after dividing, the small networks should have a necessary width or depth to guarantee their model capacity.* Only in this way, they can learn enough knowledge from the data and achieve better ensemble performance.

From conclusion 1–3, We can learn some practices about effective model scaling. When the width/depth of networks is small, increasing width/depth can get substantial improvement. However, when width/depth becomes large and increasing them yields little gain (Fig. 1), it is more effective to increase the number of networks. This comes to our core idea — *the number of networks is a new dimension of effective model scaling.* Increasing the number, width, and depth of networks together is more efficient and effective.

## Results on CIFAR-10 dataset

Results on the CIFAR-10 are shown in Tab. 3. Although the Top-1 accuracy is very high, division and co-training can also achieve significant improvement. It is worth noting {WRN-28-10, epoch 1800} get worse performance than {WRN-28-10, epoch 300}, namely, WRN-28-10 may get over-fitting on CIFAR-10 when trained for more epochs. In contrast, division and co-training can help WRN-28-10 get consistent improvement with the increasing of training epochs. This is because different models can learn different knowledge about the data. In this way, even if one model gets over-fitting, the ensemble of different models also can make a comprehensive and correct prediction. The conclusion also applies to {WRN-40-10, 1800 epochs} on CIFAR-100. {WRN-40-10, 1800 epochs} is worse than {WRN-40-10, 300 epochs} while WRN-40-10 with division

Table 3: Results on CIFAR-10.  $S$  is the number of small networks after dividing. Divide weight decay as Eq. (6). No extra data and train from scratch. PyramidNet is trained with mixed precision [36]. Methods compared here are AutoAugment [8], RandAugment [9], Fixup-init [60], Mixup [59], Cutout [10], ShakeDrop [56], and Fast AutoAugment [29].

Method	Top-1 err. (%)	# params (M)	GFLOPs	
SE-ResNet-164 [21]	4.39	2.51	0.26	
WRN-28-10 [57]	4.00	36.5	5.25	
ResNeXt-29, 8×64d [55]	3.65	34.5	5.41	
WRN-28-10 [57, 8]	2.68	36.5	5.25	
WRN-40-10 [57, 60, 59, 10]	2.30	55.9	8.08	
Shake-Shake 26 2×96d [11, 9]	2.00	26.2	3.78	
PyramidNet-272 [15, 56, 29]	1.70	26.2	4.55	
re-implementation	epochs 300	epochs 1800	# params (M)	GFLOPs
SE-ResNet-164	2.98	2.19	2.49	0.26
ResNeXt-29, 8×64d	2.69	2.23	34.4	5.40
WRN-28-10	2.28	2.41	36.5	5.25
WRN-40-10	2.33	2.19	55.8	8.08
Shake-Shake 26 2×96d		2.00	26.2	3.78
PyramidNet-272		1.33	26.2	4.55
SE-ResNet-164, $S = 2$	<b>2.84</b>	2.20	2.81	0.29
ResNeXt-29, 8×64d, $S = 2$	<b>2.12</b>	<b>1.94</b>	35.1	5.49
WRN-28-10, $S = 2$	<b>2.06</b>	<b>1.81</b>	35.8	5.16
WRN-28-10, $S = 4$	<b>2.01</b>	<b>1.68</b>	36.5	5.28
WRN-40-10, $S = 4$	<b>2.01</b>	<b>1.62</b>	55.9	8.12
Shake-Shake 26 2×96d, $S = 2$		<b>1.75</b>	23.3	3.38
PyramidNet-272, $S = 4$		<b>1.29</b>	32.6	6.33

and co-training achieves expected improvement with more training epochs.

## 4.3. Results on ImageNet dataset

Results on ImageNet are shown in Tab. 4. The inference speed difference between sequential and concurrent running is also shown in Fig. 6. All experiments on ImageNet are conducted with mixed precision training [36]. WRN-50-2 and WRN-50-3 are 2× and 3× wide as ResNet-50.

The results on ImageNet validates our idea again — *the number of networks is a new dimension of effective model scaling.* Specifically, EfficientNet-B7 (84.4%, 66M, crop 600, wider, deeper) *vs.* EfficientNet-B6 (84.2%, 43M, crop 528), WRN-50-3 (80.74%, 135.0M) *vs.* WRN-50-2 (80.66%, 68.9M), the former only produces 0.2%↑ and 0.08%↑ gain of accuracy, respectively. This shows that increasing width/depth can only yield little gain when the model is already very large, while it brings out an expensive extra cost. In contrast, increasing the number of networks rewards with much more satisfying improvement.

## 4.4. Ablation study

In Tab. 5, we study the influence of dividing and ensemble, different data views, and various value of weight factor  $\lambda_{cot}$  of co-training loss in Eq. (9). The contribution of dividing and ensemble is the most significant, *i.e.*, 1.28%↑ at best. Using different data transformers (0.14%↑) and co-training loss (0.24%↑) can also help the model improve performance during several runs. Besides, the effect of co-training is

Table 4: Single crop results on ImageNet 2012 validation dataset. No extra data and train from scratch.  $S$  is the number of small networks after dividing. Acc. of  $M_i$  is the accuracy of small networks. Only WRN applies Eq. (6); others keep weight decay unchanged. Train with mixed precision.

Method	Top-1 / Top-5 Acc.	MParams / GFLOPs	Crop / Batch / Epochs	Acc. of $M_i$
WRN-50-2 [57]	78.10 / 93.97	68.9 / 12.8	224 / 256 / 120	
ResNeXt-101, 64×4d [55]	79.60 / 94.70	83.6 / 16.9	224 / 256 / 120	
ResNet-200 [17] + AutoAugment [8]	80.00 / 95.00	64.8 / 16.4	224 / 4096 / 270	
SENet-154 [21]	82.72 / 96.21	115.0 / 42.3	320 / 1024 / 100	
SENet-154 [21] + MultiGrain [3]	83.10 / 96.50	115.0 / ~83.1	450 / 512 / 120	—
PNASNet-5 ( $N = 4, F = 216$ ) [32] <sup>†</sup>	82.90 / 96.20	86.1 / 25.0	331 / 1600 / 312	
AmoebaNet-C ( $N = 6, F = 228$ ) [42] <sup>†</sup>	83.10 / 96.30	155.3 / 41.1	331 / 3200 / 100	
EfficientNet-B6 [52]	<b>84.20</b> / <b>96.80</b>	43.0 / 19.0	528 / 4096 / 350	
EfficientNet-B7 [52]	<b>84.40</b> / <b>97.10</b>	66.7 / 37.0	600 / 4096 / 350	
WRN-50-2 (re_impl.)	80.66 / 95.16	68.9 / 12.8	224 / 256 / 120	
WRN-50-3 (re_impl.)	80.74 / 95.40	135.0 / 23.8	224 / 256 / 120	
ResNeXt-101, 64×4d (re_impl.)	81.57 / 95.73	83.6 / 16.9	224 / 256 / 120	—
EfficientNet-B7 (re_impl.) <sup>‡</sup>	81.83 / 95.78	66.7 / 10.6	320 / 256 / 120	
WRN-50-2, $S = 2$	79.64 / 94.82	51.4 / 10.9	224 / 256 / 120	78.68, 78.66
WRN-50-3, $S = 2$	<b>81.42</b> / <b>95.62</b>	138.0 / 25.6	224 / 256 / 120	80.32, 80.24
ResNeXt-101, 64×4d, $S = 2$	<b>82.13</b> / <b>95.98</b>	88.6 / 18.8	224 / 256 / 120	81.09, 81.02
EfficientNet-B7, $S = 2^{\#}$	<b>82.74</b> / <b>96.30</b>	68.2 / 10.5	320 / 256 / 120	81.39, 81.83
SE-ResNeXt-101, 64×4d, $S = 2^{\#}$	<b>83.34</b> / <b>96.61</b>	98.0 / 61.1	416 / 256 / 120	82.33, 82.62

<sup>†</sup> Refer to [GitHub: tensorflow/models/pnasnet](https://github.com/tensorflow/models/pnasnet) and [GitHub: tensorflow/tpu/amoeba\\_net](https://github.com/tensorflow/tpu/amoeba_net). They use 100 P100 workers. On each worker, batch size is 16 or 32.

<sup>‡</sup> Batch size 128 on 4 Tesla V100 32GB GPUs with gradient accumulation step 2, i.e., 1 backpropagation per 2 forward processes. Re-implementation of EfficientNet in PyTorch refers to Appendix D.

Table 5: Influence of various settings on CIFAR-100.

Method	SplitNet ( $S = 2$ )			SplitNet ( $S = 4$ )		
	diff. views	$\lambda_{cot}$	Top-1 err. (%)	diff. views	$\lambda_{cot}$	Top-1 err. (%)
WRN-16-8 [57]	✗		17.64 ± 0.18	✗		17.41 ± 0.18
original: 20.43%	✓		17.50 ± 0.05	✓		17.36 ± 0.12
re-impl.: 18.69%	✓	0.1	17.56 ± 0.08	✓	0.1	17.37 ± 0.04
	✓	0.5	17.44 ± 0.08	✓	0.5	17.12 ± 0.05
	✓	1.0	17.51 ± 0.04	✓	1.0	17.15 ± 0.38

more significant when there are more networks, i.e., 0.06%↑ ( $S = 2$ ) vs. 0.24%↑ ( $S = 4$ ). To better understand the effect of training on different data views, we project the weights of the last fc. layers of small networks ( $[\theta_1; \theta_2; \dots; \theta_S]$ ) to a 2D space using PCA [28]. And the standard deviation of these coordinates are  $\{S = 2 : 13.66 \text{ (w) vs. } 13.51 \text{ (w/o)}\}$  and  $\{S = 4 : 29.49 \text{ (w) vs. } 29.19 \text{ (w/o)}\}$ . The projection points of small networks trained with different data views are more discrete, which means the distribution of corresponding networks is more discrete in hypothesis space. Namely, they are more likely to learn different and complementary knowledge about the data. However, the difference is relatively small. This somehow explains the slight gain of using various data views. Nevertheless, it does make changes in several runs. Considering the baseline is very strong, their improvement also matters, especially in practical applications.

We do not divide the large network into 8, 16, or more small networks. Firstly, the large the  $S$ , the thinner the small networks. As conclusion 3 said, after dividing, the small networks may be too thin to guarantee a necessary model capacity and satisfying ensemble performance – unless the

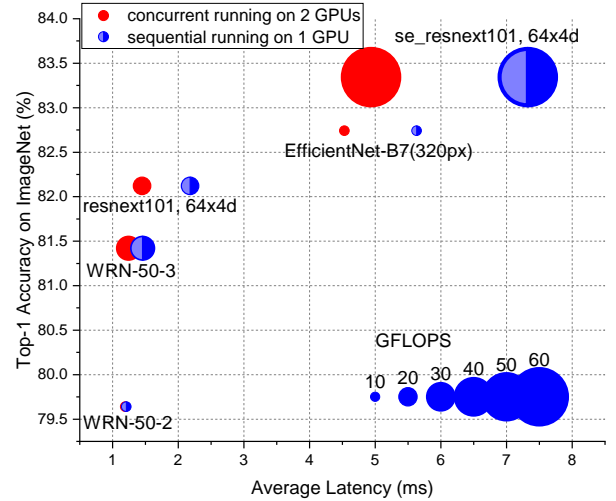


Figure 6: The difference of inference latency between sequential and concurrent running. Test on Tesla V100(s) with mixed precision and batch size 100.

original network is very large. Secondly, as mentioned in 3.1, small networks run in sequence, and training of 8/16 small networks may cost too much time. The implementation of an asynchronous training system needs further hard work. Theoretically, one can abandon the co-training part to achieve an easy one with some sacrifice of performance.



## 5. Conclusion and future work

In this paper, a new idea — *the number of networks should be a new dimension of effective model scaling*, besides commonly known depth/width/resolution is discussed. We verify this idea through extensive experiments and achieve promising results on CIFAR and ImageNet datasets. This new dimension potentially introduces many interesting topics in neural network engineering, *e.g.*, designing a flexible framework for asynchronous training of multiple models, more complex deep ensemble methods, multiple models with different modalities of data, introducing the idea into NAS (given specific parameters/FLOPs constrain, how can we design one or several models to get the best performance on a certain task).

## Acknowledgments

This paper was supported by funding 2019-INT007 from the Shenzhen Institute of Artificial Intelligence and Robotics for Society. This work was also supported in part by The National Nature Science Foundation of China (Grant Nos: 62036009, 61936006). Thanks Yuejin Li for his support in using GPU clusters.

## References

- [1] Monther Alhamdoosh and Dianhui Wang. Fast decorrelated neural network ensembles with random weights. *Inf. Sci.*, 2014.
- [2] Tanmay Batra and Devi Parikh. Cooperative learning with visual attributes. *CoRR*, abs/1705.05512, 2017.
- [3] Maxim Berman, Hervé Jégou, Vedaldi Andrea, Iasonas Kokkinos, and Matthijs Douze. MultiGrain: a unified image embedding for classes and instances. *arXiv e-prints*, 2019.
- [4] Avrim Blum and Tom M. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*. ACM, 1998.
- [5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.
- [6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.
- [7] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.
- [8] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019.
- [9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. *arXiv preprint arXiv:1909.13719*, 2019.
- [10] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017.
- [11] Xavier Gastaldi. Shake-shake regularization. *CoRR*, abs/1705.07485, 2017.
- [12] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020.
- [13] Giorgio Giacinto and Fabio Roli. Design of effective neural network ensembles for image classification purposes. *Image Vis. Comput.*, 2001.
- [14] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017.
- [15] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *CVPR*, 2017.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.
- [19] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. *CoRR*, abs/1812.01187, 2018.
- [20] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.
- [21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.
- [22] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.
- [23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- [24] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.
- [25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.
- [26] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *NeurIPS*, 1991.
- [27] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.
- [28] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *NeurIPS*, 2018.
- [29] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. In *NeurIPS*, 2019.
- [30] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
- [31] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan L. Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *ECCV*, 2018.

- [32] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan L. Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *ECCV*, 2018.
- [33] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. In *ICLR*, 2019.
- [34] Yong Liu and Xin Yao. Simultaneous training of negatively correlated neural networks in an ensemble. *IEEE Trans. Syst. Man Cybern. Part B*, 1999.
- [35] Luke Melas-Kyriazi. Efficientnet in pytorch. <https://github.com/lukemelas/EfficientNet-PyTorch>, 2020.
- [36] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Damos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *ICLR*, 2018.
- [37] Yuri Nesterov. A method for unconstrained convex minimization problem with the rate of convergence  $O(1/k^2)$ . In *Doklady an ussr*, volume 269, pages 543–547, 1983.
- [38] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.
- [39] NVIDIA. Cuda toolkit documentation v11.1.0. <https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#streams>, 2020.
- [40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 2019.
- [41] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan L. Yuille. Deep co-training for semi-supervised image recognition. In *ECCV*, 2018.
- [42] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. In *AAAI*, 2019.
- [43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015.
- [44] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.
- [45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.
- [46] Barbara Leigh Smith and Jean T MacGregor. What is collaborative learning, 1992.
- [47] Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *CoRR*, abs/1803.09820, 2018.
- [48] Guocong Song and Wei Chai. Collaborative learning for deep neural networks. In *NeurIPS*, 2018.
- [49] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014.
- [50] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.
- [51] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*. IEEE Computer Society, 2016.
- [52] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.
- [53] TensorFlow. Efficientnet in tensorflow. <https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet>, 2020.
- [54] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2020.
- [55] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.
- [56] Yoshihiro Yamada, Masakazu Iwamura, Takuya Akiba, and Koichi Kise. Shakedrop regularization for deep residual learning. *IEEE Access*, 2019.
- [57] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.
- [58] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger B. Grosse. Three mechanisms of weight decay regularization. In *ICLR*, 2019.
- [59] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.
- [60] Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. In *ICLR*, 2019.
- [61] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.
- [62] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018.
- [63] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020.
- [64] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. Ensembling neural networks: Many could be better than all. *Artif. Intell.*, 2002.
- [65] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *ICLR*, 2017.

## SplitNet: Divide and Co-training Appendices

### A. Details about dividing a large network

$S$  is the number of small networks after dividing.

**ResNet [17]** For CIFAR [24] and SVHN [38], the numbers of input channels of the three blocks are:

$$\begin{aligned} \text{original} : & [16, 32, 64], \\ S = 2 : & [12, 24, 48], \\ S = 4 : & [8, 16, 32]. \end{aligned}$$

**SE-ResNet [21]** The reduction ratio in the SE module keeps unchanged. Other settings are the same as ResNet.

**EfficientNet [52]** The numbers of output channels of the first conv. layer and blocks in EfficientNet baseline are:

$$\begin{aligned} \text{original} : & [32, 16, 24, 40, 80, 112, 192, 320, 1280], \\ S = 2 : & [24, 12, 16, 24, 56, 80, 136, 224, 920], \\ S = 4 : & [16, 12, 16, 20, 40, 56, 96, 160, 640]. \end{aligned}$$

**WRN [57]** Suppose the widen factor is  $w$ , the new widen factor  $w^*$  after dividing is:

$$w^* = \max(\lfloor \frac{w}{\sqrt{S}} + 0.4 \rfloor, 1.0). \quad (11)$$

**ResNeXt [55]** Suppose original *cardinality* (groups in convolution) is  $d$ , new *cardinality*  $d^*$  is:

$$d^* = \max(\lfloor \frac{d}{S} \rfloor, 1.0). \quad (12)$$

**Shake-Shake [11]** For Shake-Shake 26  $2 \times 96d$ , the numbers of output channels of the first convolutional layer and three blocks are:

$$\begin{aligned} \text{original} : & [16, 96, 96 \times 2, 96 \times 4], \\ S = 2 : & [16, 64, 64 \times 2, 64 \times 4], \\ S = 4 : & [16, 48, 48 \times 2, 48 \times 4]. \end{aligned}$$

**DenseNet [22]** Suppose the grow rate of DenseNet is  $g_{\text{dense}}$ , the new grow rate after dividing is

$$g_{\text{dense}}^* = \frac{1}{2} \times \lfloor 2 \times \frac{g_{\text{dense}}}{\sqrt{S}} \rfloor. \quad (13)$$

**PyramidNet [15] + ShakeDrop [56]** Suppose the additional rate of PyramidNet and final drop probability of ShakeDrop is  $g_{\text{pyramid}}$  and  $p_{\text{shake}}$ , respectively, we divide them as:

$$g_{\text{pyramid}}^* = \frac{g_{\text{pyramid}}}{\sqrt{S}}, \quad (14)$$

$$p_{\text{shake}}^* = \frac{p_{\text{shake}}}{\sqrt{S}}. \quad (15)$$

To pursue better performance, we do not divide the base channel of PyramidNet on CIFAR because it is small, *i.e.*, 16.

### B. Weight decay matters

#### B.1. The value of weight decay matters

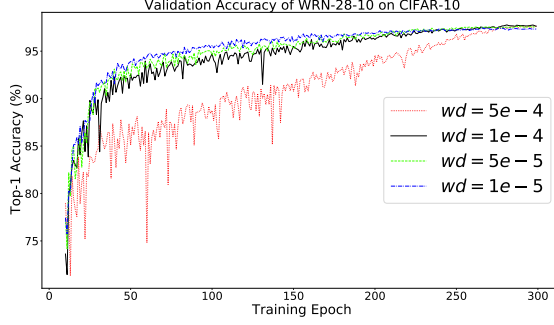
Weight decay is a simple technique that can prevent weights from growing too large and improve the generalization performance of a model [26]. When using SGD in PyTorch, it is equivalent to  $\ell_2$  regularization.

It is commonly known that the generalization of a model depends on the balance between the model capacity and data complexity. In our experiments, we use many data augmentation methods, *e.g.*, AutoAugment [8], mixup [59], random erasing [63]. These regularization methods have relatively fixed hyper-parameters. In this way, to obtain good performance, it is necessary to find an appropriate value of weight decay to balance various forms of regularizations [47]. Some weight decay values are tested on CIFAR datasets with WRN-16-8 and WRN-28-10. The results and corresponding training curves are shown in Tab. 6 and Fig. 7, respectively.

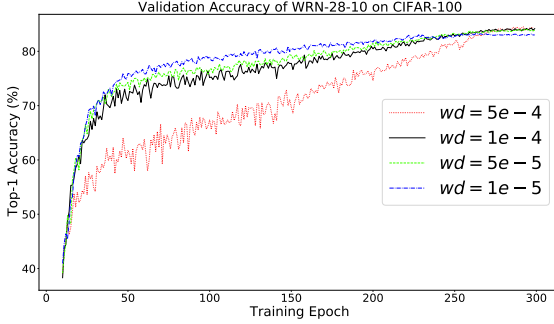
For WRN-16-8 and WRN-28-10, a small weight decay factor like  $1e-5$  and  $5e-5$  is not suitable with initial learning rate 0.1. In contrast,  $1e-4$  and  $5e-4$  are good choices for these two networks on CIFAR datasets. From Tab. 6, we also find

Table 6: Influence of different values of weight decay on CIFAR dataset.  $\{lr = 0.1, max\_epoch = 300, slow\_epoch = 20\}$  with cosine annealing learning rate strategy as shown in Fig. 8a. Generally, weight decay is only applied to weights of conv. and fc. layers. *wd-all* means applying weight decay to all learnable parameters in the model.

Method	$wd$	Top-1 err. (%)	
		CIFAR-10	CIFAR-100
WRN-28-10 [57]	$1e-5$	2.54	16.71
	$5e-5$	2.36	15.91
	$1e-4$	<b>2.23</b>	15.71
	$5e-4$	2.28	<b>15.50</b>
Method	$weight\ decay$	Top-1 err. (%) CIFAR-100	Top-1 err. (%) ( <i>wd-all</i> ) CIFAR-100
WRN-16-8 [57]	$1e-5$	20.55	20.98
	$5e-5$	19.78	19.67
	$1e-4$	<b>18.69</b>	19.30
	$5e-4$	<b>18.69</b>	<b>17.75</b>



(a) Accuracy of WRN-28-10 on CIFAR-10



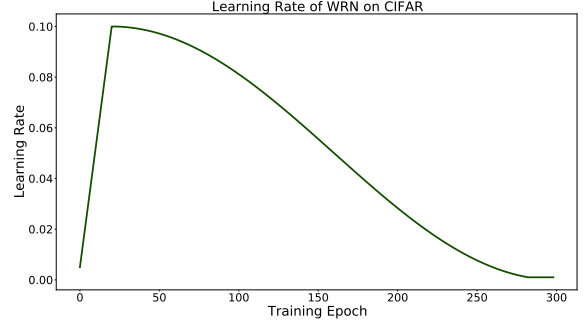
(b) Accuracy of WRN-28-10 on CIFAR-100

Figure 7: Training curves of WRN-28-10 on CIFAR-10 and CIFAR-100. **Best viewed in color with 800% zoom.**

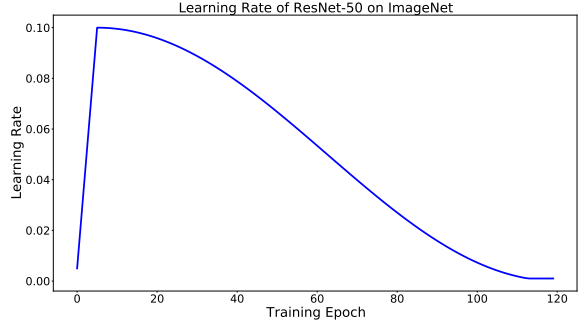
the small model with fewer parameters, *i.e.*, WRN-16-8, is more sensitive to the value of weight decay. Specifically, compared to their best results, WRN-16-8 gets  $\sim 2\%$  drop of performance while WRN-28-10 only gets  $\sim 1\%$  drop when using *weight decay*= $1e-5$  on CIFAR-100.

A small weight decay values ( $5e-5$  and  $1e-5$ ) can help the model gain better performance than large values at the early stage of the training, however, the models with large weight decay values ( $5e-4$  and  $1e-4$ ) finally achieve better generalization performance. This is partly because of the cosine annealing learning rate strategy we use. The decayed weight in an iteration is  $lr \times wd \times \theta$ , where  $lr$  is the learning rate,  $wd$  denotes the value of weight decay, and  $\theta$  is the learnable weight. At the early stage of training, the decayed value ( $lr \times wd$ ,  $wd = 5e-4 / 1e-4$ ) causes large disturbance to the model, and the model cannot fit the data well. In contrast, at the late stage, the decayed value ( $lr \times wd$ ,  $wd = 5e-5 / 1e-5$ ) maybe too small for the model and produce little positive regularization effect. In the latter case, the model may be trapped at some local minima with inferior performance compared to the model with a large weight decay value.

It is worth noting that weight decay is only applied to the weight parameters of the conv. and fc. layers. The parameters of the batch normalization layers and bias are left undecayed. According to the observation of [58], it



(a) Learning rate strategy with warm up of WRN on CIFAR.



(b) Learning rate strategy with warm up of ResNet-50 on ImageNet.

Figure 8: Cosine learning rate strategy.

Table 7: Influence of different values of weight decay on ImageNet validation dataset. We use  $\{lr = 0.1, max\_epoch = 120, slow\_epoch = 5, batch\ size\ 256\}$  with cosine annealing learning rate strategy. Generally, weight decay is only applied to weights of conv. and fc. layers. *wd-all* means applying weight decay to all learnable parameters in the model.

Method	$wd$	Top-1 / Top-5 Acc. (%)	Top-1 / Top-5 Acc. (%) ( <i>wd-all</i> )
ResNet-50 [17]	$1e-5$	78.21 / 94.01	78.37 / 94.25
	$5e-5$	<b>78.84 / 94.47</b>	<b>78.81 / 94.47</b>
	$1e-4$	78.32 / 94.17	78.53 / 94.34
	$5e-4$	72.98 / 91.44	74.89 / 92.49

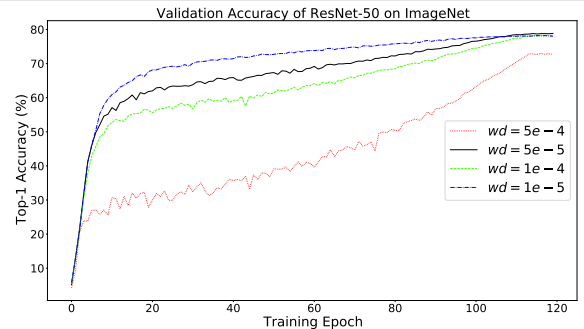


Figure 9: Training curve of ResNet-50 on ImageNet. The training curve of ResNet-50 (*wd-all*) is similar.



Table 8: Influence of weight decay’s dividing manners. Weight decay is only applied to weights of conv. and fc. layers.  $lr = 0.1/0.2$  for WRNs / Shake-Shake. Three manners,  $wd^* = wd$ ,  $wd^* = \frac{wd}{S}$  (Eq. (16)),  $wd^* = wd \times \exp(\frac{1}{S} - 1.0)$  (Eq. (6)).

(a) Training epochs 300

Method	$wd$	Top-1 err. (%)		Method	$wd$	Top-1 err. (%)
		CIFAR-10	CIFAR-100			CIFAR-100
WRN-28-10	$5e-4$	2.28	15.50	WRN-16-8	$1e-4$	18.69
WRN-28-10, $S = 2$	$5e-4$	2.15	14.48	WRN-16-8, $S = 2$	$1e-4$	<b>17.37</b>
WRN-28-10, $S = 2$	Eq. (16)	2.15	14.43	WRN-16-8, $S = 2$	Eq. (16)	18.11
WRN-28-10, $S = 2$	Eq. (6)	<b>2.06</b>	<b>14.16</b>	WRN-16-8, $S = 2$	Eq. (6)	17.77
WRN-28-10, $S = 4$	$5e-4$	2.36	14.26	WRN-16-8, $S = 4$	$1e-4$	<b>17.07</b>
WRN-28-10, $S = 4$	Eq. (16)	<b>2.00</b>	14.79	WRN-16-8, $S = 4$	Eq. (16)	18.35
WRN-28-10, $S = 4$	Eq. (6)	2.01	<b>14.04</b>	WRN-16-8, $S = 4$	Eq. (6)	17.92

(b) Training epochs 1800

Method	$wd$	Top-1 err. (%)	Method	$wd$	Top-1 err. (%)
		CIFAR-10			CIFAR-10
WRN-28-10	$5e-4$	2.41	Shake-Shake 26 2×96d	$1e-4$	2.00
WRN-28-10, $S = 2$	$5e-4$	1.95	Shake-Shake 26 2×96d, $S = 2$	$1e-4$	1.94
WRN-28-10, $S = 2$	Eq. (16)	<b>1.81</b>	Shake-Shake 26 2×96d, $S = 2$	Eq. (16)	<b>1.74</b>
WRN-28-10, $S = 2$	Eq. (6)	<b>1.81</b>	Shake-Shake 26 2×96d, $S = 2$	Eq. (6)	1.75
WRN-28-10, $S = 4$	Eq. (16)	<b>1.66</b>	Shake-Shake 26 2×96d, $S = 4$	Eq. (16)	<b>1.69</b>
WRN-28-10, $S = 4$	Eq. (6)	1.68	Shake-Shake 26 2×96d, $S = 4$	Eq. (6)	1.84

makes a trivial difference with applying weight decay to the whole model. However, WRN-16-8 get better performance (**17.75%** vs. 18.69%) on CIFAR-100 when applying a large weight decay ( $5e-4$ ) to all parameters. This is interesting and needs further investigation.

We replicate some of the above experiments on ImageNet [43] with ResNet-50. Results are shown in Tab. 7 and Fig. 9. A large weight decay value ( $5e-4$  &  $lr = 0.1$ ) leads to a significant drop in performance ( $\sim 6\%$ ) on ImageNet while WRNs with a large weight decay value work well on CIFAR. This shows that the weight decay value should not be too large on a large and complex dataset with a relatively small model.

## B.2. The manner of dividing weight decay matters

As above said, the value of weight decay matters for the generalization performance of a neural network, consequently, the manner of dividing weight decay matters. Besides the exponential strategy in Eq. (6) and no dividing strategy, a linear dividing strategy is further introduced:

$$wd^* = \frac{wd}{S}. \quad (16)$$

Linear transformations can also be applied to Eq. (6) and Eq. (16) to get new decay strategies. However, this is a large topic, *i.e.*, finding out the underlying relationship between

model capacity and appropriate weight decay values. Therefore, in this work, discussions are restricted within these three manners.

The experimental results are shown in Tab. 8. It is clear that there is not a best and universal solution for different models on different datasets. In most cases, dividing weight decay can gain performance improvement. This is also reasonable because a general assumption is that a small model needs less regularization. However, diving weight decay sometimes lead to worse performance, *i.e.*, WRN-16-8 with Eq. (6) and Eq. (16) performs worse than WRN-16-8 with  $wd^* = wd$ . It is somehow counterintuitive. WRN-28-10 ( $S = 2$ ) with larger capacity can gain improvement with dividing weight decay. In this case, WRN-16-8 ( $S = 2$ ) is expected to work well with dividing weight decay because WRN-16-8 ( $S = 2$ ) is a relatively small model and should need less regularization. One possible reason is  $1e-4$  maybe already a small weight decay value for WRN-16-8 on CIFAR-100, and it should not be further minimized.

### C. Influence of different ensemble manners

Besides the simple averaging ensemble manner (Eq. (10)) we used in experiments, we also test max ensemble, *i.e.*, use the most confident prediction of small networks as the final prediction. Results are shown in Tab. 9. Simple averaging is the most effective way among test methods in most cases. This is consistent with previous work [64]. More complex ensemble manners are left as future topics.

Table 9: Influence of different ensemble manners. Softmax (✓) means we ensemble the results after softmax operation. Generally, we do ensemble operations without softmax.

Method	<i>softmax</i>	<i>ensemble</i>	Top-1 err. (%)	
			CIFAR-10	CIFAR-100
WRN-28-10, $S = 4$	✓	max	1.85	15.04
	✓	avg.	1.77	14.45
	✗	max	1.90	15.27
	✗	avg.	<b>1.68</b>	<b>14.26</b>
ResNeXt-29, $8 \times 64d$ , $S = 2$	✓	max	1.94	15.58
	✓	avg.	<b>1.93</b>	15.08
	✗	max	2.01	15.78
	✗	avg.	1.94	<b>14.99</b>
Method	<i>softmax</i>	<i>ensemble</i>	ImageNet Acc. (%)	
			Top-1	Top-5
WRN-50-3, $S = 2$	✓	max	81.11	95.42
	✓	avg.	81.31	95.52
	✗	max	80.84	95.42
	✗	avg.	<b>81.42</b>	<b>95.62</b>
ResNeXt-101, $64 \times 4d$ , $S = 2$	✓	max	81.92	95.82
	✓	avg.	82.03	95.91
	✗	max	81.78	95.80
	✗	avg.	<b>82.13</b>	<b>95.98</b>

### D. EfficientNet in PyTorch

We re-implement EfficientNet in PyTorch according to [54, 35] and the official implementation in TensorFlow [53]. Some attempts are shown in Tab. 10. We cannot reproduce the original result due to some differences between two frameworks, limited epochs, or smaller batch size, *i.e.*, 4096 vs. 256.

Original EfficientNet uses an exponential learning rate strategy, *i.e.*, decays the learning rate by 0.97 every 2.4 epochs. To match the decay steps (the combination of a forward process and a backward propagation is called a *step*), when using exponential learning rate strategy with batch size 256 and epochs 120, we use decay epochs 0.8. A decay epoch 2.4 in our training settings will lead to much inferior performance. A surprising finding is that mixup will lead to worse performance with EfficientNet.

Wightman [54] reproduces a similar result with original EfficientNet-B2 using PyTorch, *i.e.*, 80.4% Top-1 accuracy. The original result of EfficientNet-B2 is 80.3%. However, they train the model for 450 epochs with batch size 128. This

Table 10: Some attempts of re-implementing EfficientNet-B1. Our re-implementation adopts RandAugment and mixed precision training. Weight decay is  $1e-5$ .

EfficientNet-B1	optim.	lr	mixup	Crop / Batch / Epochs	Top-1 / Top-5 Acc.
Original	RMSProp	exp., 0.256	✗	240 / 4096 / 350	<b>79.20 / 94.50</b>
re-impl.	RMSProp	exp., 0.016	✗	240 / 256 / 120	77.56 / 93.61
	SGD	exp., 0.256	✗	240 / 256 / 120	77.61 / 93.80
	SGD	exp., 0.256	✓	240 / 256 / 120	77.36 / 93.61
	SGD	cos., 0.256	✗	240 / 256 / 120	<b>78.31 / 94.10</b>

means much more training steps and leads to formidable training time. They also use RandAugment during training.

#### Latency of concurrent running EfficientNet-B7 ( $S = 2$ )

As shown in Fig. 3, concurrent running EfficientNet-B7 ( $S = 2$ ) is slower than EfficientNet-B7. This is because our running time includes data loading, pre-processing (still runs in sequence), and transferring (CPU  $\leftrightarrow$  GPU, GPU  $\leftrightarrow$  GPU) time. For EfficientNet-B7, the speedup of concurrent running is not able to cover the extra time brought by additional data manipulating time. This is possible the price for better performance.