
Accelerating CNN Training by Sparsifying Activation Gradients

Xucheng Ye

School of Computer Science and Engineering
Beihang University

Jianlei Yang

School of Computer Science and Engineering
Beihang University
jianlei@buaa.edu.cn

Pengcheng Dai

School of Microelectronics
Beihang University

Yiran Chen

Department of Electrical and Computer Engineering
Duke University

Weisheng Zhao

School of Microelectronics
Beihang University
weisheng.zhao@buaa.edu.cn

Abstract

Gradients to activations get involved in most of the calculations during back propagation procedure of Convolution Neural Networks (CNNs) training. However, an important known observation is that the majority of these gradients are close to zero, imposing little impact on weights update. These gradients can be then pruned to achieve high gradient sparsity during CNNs' training and reduce the computational cost. In particular, we randomly change a gradient to zero or a threshold value if the gradient is below the threshold which is determined by the statistical distribution of activation gradients. We also theoretically proved that the training convergence of the CNN model can be guaranteed when the above activation gradient sparsification method is applied. We evaluated our method on AlexNet, MobileNet, ResNet-{18, 34, 50, 101, 152} with CIFAR-{10, 100} and ImageNet datasets. Experimental results show that our method can substantially reduce the computational cost with negligible accuracy loss or even accuracy improvement. Finally, we analyzed the benefits that the sparsity of activation gradients introduced in detail.

1 Introduction

Convolutional Neural Networks (CNNs) have been widely applied on many applications and various devices in recent years. However, the network structures are becoming more and more complex so that training them on large scale datasets are very time consuming especially with limited hardware resources. Several research works have reported that training them could be finished within minutes on high performance computation platforms [1][2][3], but almost thousands of GPUs are utilized which are not feasible for many researchers. Even though there are many research works on network compression, most of them are focused on inference [4]. Our work aims to reduce the training workloads efficiently so that large scale training could be performed on budgeted computation platforms.

The kernel optimization step of CNN training is to perform Stochastic Gradient Descent (SGD) algorithm in backward propagation procedure. There are several data types among the training dataflow:

weights, gradients to weight, activations, gradients to activation. Backward propagation starts from computing the weight gradients with the activations and then performs weights update [5]. Among them, activation gradients back-propagation and weight gradients computation require intensive convolution operations so that they dominate the total training cost. It is well known that computation cost can be reduced by skipping over zero-values. Since the above two convolution steps require the activation gradients as input, improving the sparsity of activation gradients should significantly reduce the computation cost and memory footprint during backward propagation procedure.

We assume that activation gradients are distributed normally, then a threshold τ can be calculated based on this hypothesis. Our work applies stochastic pruning to the activation gradients under τ and map them to zero or τ randomly. Note that the gradients passed ReLU layer distribute irregularly so we divide common networks into two categories, one is networks using Conv-ReLU as basic block such as AlexNet [6] and VGGNet [7], another one is that using Conv-BN-ReLU structure such as ResNet [8] and MobileNet [9]. Our experiments show stochastic pruning method works for both Conv-ReLU structure and Conv-BN-ReLU structure in modern networks. A mathematical proof is proposed to demonstrate that stochastic pruning can maintain the convergence of original CNN model. Our approach can be classified as gradient sparsification, and can be combined with gradient quantization method.

2 Related Works

Pruning Acceleration of CNN inference phase by pruning has been researched widely and has achieved outstanding advances. Pruning of inference can be divided into five categories [4]: element-level [10], vector-level [11], kernel-level [12], group-level [13] and filter-level pruning [14][15] [16]. Inference pruning focus on raising parameter sparsity of convolution layers. There are also pruning methods designed for training, e.g. weight gradients sparsification, which can reduce communication cost for weight gradients exchanging in a distributed learning system. Aji [17] pruned 99% weight gradients with smallest absolute value by a heuristic algorithm. According to filters' correlation, Prakash [18] prunes 30% filters temporarily when training, which can improve training efficiency. Different from those works mentioned above, the purpose of our work is to decrease computing cost of backward propagation by pruning activation gradients.

Quantization Quantization is another common way to release the computational complexity and memory consumption of training. Gupta's work [19] maintains the accuracy by training the model in 16-bit fixed-point number with stochastic rounding. DoReFaNet [20] derived from AlexNet [6] uses 1-bit, 2-bit and 6-bit fixed-point number to store weights, activations and gradients, respectively, but brings visible accuracy drop. Terngrad [21] is designed for distributed learning. Their approach only requires three numerical levels for weight gradient. They demonstrate that model can still converge using ternary weight gradients.

Mixed precision is a new research direction of quantization. Park [22] proposed a value-aware quantization method that use low-precision for small values, which can significantly reduce memory consumption when training ResNet-152 [8] and Inception-V3 [23] with 98% activations quantified to 3-bit. Micikevicius [5] keeps a FP32 copy for weights update and uses FP16 for computation. This method has been proved to work in various fields of deep learning and speedup the training process.

3 Methodologies

3.1 Original Dataflow

Fig. 1 represents one backward propagation iteration for a convolution layer. *Activation gradients back-propagation* (AGB) perform convolution between gradients g_{A_O} and layer's weight w and get gradient of A_I , which will be propagated to the next layer. *Weight gradients computation* (WGC) takes A_I and g_{A_O} as input, and the result of convolution is weight gradients g_w . Note that without explicit mention, *next layer* indicate the next layer in backward procedure in this paper.

We found that activation gradients are full of *small values*, which are values extremely near to zero. Obviously, the closer one activation gradient is to 0, the less it will affect the update of weights. So

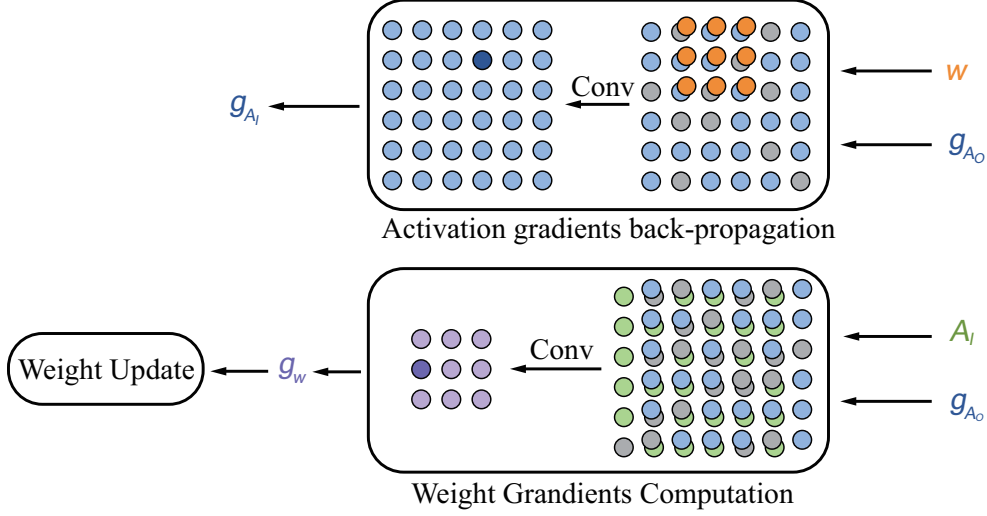


Figure 1: Backward propagation of convolutional layer.

it's reasonable to assume that pruning those small values will not harm CNN model's performance after training.

To achieve this goal, the easiest way is to sort the absolute values of gradients for a threshold, then dropout those small values whose absolute values are under this threshold directly. But here comes two problems. Firstly, sorting requires serious computational and memory overhead. Secondly, experiment results show that direct drop all these small values will significantly influence convergence and accuracy for there are tons of them.

To address these problems, we propose two algorithms: *Distribution Based Threshold Determination* and *Stochastic Pruning*.

3.2 Sparsification Algorithms

Distribution Based Threshold Determination (DBTD) As mentioned above, it is unfeasible to find the threshold by sorting, so a new method that is easier to implement and has less overhead is needed. First, we analyzed the data distribution of activation gradients in the training process. According to the structure of the modern CNN models, it needs to be divided into two cases for discussion.

For those networks using *Conv-ReLU* structure, which means that the convolutional layer is followed by ReLU layer, as a basic block, g_{A_O} is sparse but has anomalous distribution. In addition, the gradients going to propagate to previous layer g_{A_I} , is full of non-zero values. At the same time, statistical data shows that the proportion of values is symmetric with 0 and decreases with the increment of absolute value. For those network taking *Conv-BN-ReLU* structure as a basic block, g_{A_O} has the same properties with g_{A_I} .

In the first case, next block can inherit the sparsity of g_{A_I} because ReLU layer won't map zeros to non-zeros, So the g_{A_I} from Conv-ReLU structure and the g_{A_O} from Conv-BN-ReLU is considered as pruning target, we use g to denote both of them. According to the attribute of g , we made a hypothesis that g is a set of simple random samples of a normal distribution with mean 0 and variance σ^2 , and σ is unknown. Suppose the length of g is n , let

$$A \triangleq \sum_{i=1}^n |g_i|, \quad g_i \in g$$

The expectation of A is:

$$\begin{aligned} E(A) &= E\left(\sum_{i=1}^n |g_i|\right) = nE(|g_1 - 0|) \\ &= \frac{n}{\sqrt{2\pi\sigma^2}} \int |x| \exp\left\{-\frac{x^2}{2\sigma^2}\right\} dx = \sqrt{\frac{2}{\pi}} n\sigma \end{aligned}$$

Let

$$\hat{\sigma} = \frac{1}{n} \sqrt{\frac{2}{\pi}} \sum_{i=1}^n |g_i|$$

then

$$E(\hat{\sigma}) = E\left(\frac{1}{n} \sqrt{\frac{2}{\pi}} \sum_{i=1}^n |g_i|\right) = \sigma$$

Clearly, $\hat{\sigma}$ is an unbiased estimator of parameter σ .

Here we use the mean of the absolute values because the computational overhead is acceptable. Now, we got the approximated distribution of gradient g . Based on the distribution, how to find the threshold can be solved by computing the percentile of g . For instance, p -th percentile of g ($50 \leq p \leq 100$) is greater than $(2p - 100)\%$ gradients' absolute value. So, threshold $\tau = \frac{2p-100}{100}$, and $p = (\tau + 1) \times 50$.

In summary, this algorithm has two advantages,

1. Less cost: its arithmetic complexity is $\mathcal{O}(n)$, less than sorting which is at least $\mathcal{O}(n \log n)$, it requires almost no extra storage at the same time.
2. No extra hyper-parameters, only p need to be set.

Stochastic Pruning During the experiment we found that there are too many values which are near zero. Single small value has little impact on weights update, however, once all these values are set to 0, the distribution of g will have a huge change, which will influence the update of weights and cause accuracy loss. Inspired by *Stochastic Rounding* in [19], we adopt stochastic pruning to solve this problem.

The algorithm treat g as a one-dimensional vector with length of n , threshold τ is determined by DBTD with percentile p .

Algorithm 1: Stochastic Pruning

Input: original activation gradient g , threshold τ

Output: sparse activation gradient \hat{g}

```

for  $i = 1; i \leq n; i = i + 1$  do
  if  $|g_i| < \tau$  then
    Generate a random number  $r \in [0, 1]$ ;
    if  $|g_i| > \tau r$  then
       $\hat{g}_i = (g_i > 0) ? \tau : (-\tau)$ ;
    else
       $\hat{g}_i = 0$ ;
    end
  end
end

```

Fig. 2 shows the effect of stochastic pruning and where it works in different situations mentioned above. The mathematical proof in section 3.3 shows that applying this gradient sparsification method to a convolutional neural network during training does not affect its convergence.

3.3 Convergence Analysis

In this section, we will prove that the model training with pruning algorithm has the same convergence ability with original training schedule under the GOGA (General Online Gradient Algorithm) framework [24].

In [24], L. Bottou considered a learning problem as follows: suppose that there is an unknown distribution $P(z)$ and can only get a batch of samples z_t each iteration, which t denotes iteration times. The goal of training is to find the optimal parameters w which minimize the loss function $Q(z, w)$. For convenience, we define the cost function as

$$C(w) \triangleq \mathbf{E}_z Q(z, w) \triangleq \int Q(z, w) dP(z) \quad (1)$$

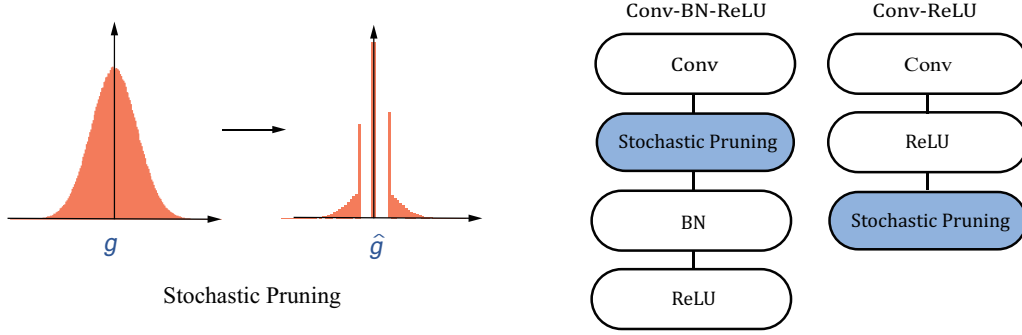


Figure 2: Effect of stochastic pruning and its placement in different structures.

Under this framework, L. Bottou proved that an online learning system with update rule as

$$w_{t+1} = w_t - \gamma_t H(z_t, w_t)$$

(where γ_t is the learning rate) will finally converge as long as the assumptions below are satisfied.

Assumption 1. The cost function $C(w_t)$ has a single global minimum w^* and satisfies the condition that

$$\forall \varepsilon, \inf_{(w-w^*)^2 > \varepsilon} (w - w^*) \nabla_w C(w) > 0 \quad (2)$$

Assumption 2. Learning rate γ_t fulfills that

$$\sum_{t=1}^{\infty} \gamma_t = \infty, \sum_{t=1}^{\infty} \gamma_t^2 < \infty \quad (3)$$

Assumption 3. The update function $H(z_t, w_t)$ meets that

$$\mathbf{E}_z H(z, w_t) = -\nabla_w C(w_t) \quad (4)$$

and

$$\mathbf{E}_z [H(z, w_t)^2] \leq A + B(w - w^*)^2 \quad (5)$$

The only difference between proposed algorithm and original algorithm is the update function $H(z_t, w_t)$, where in original algorithm

$$H(z_t, w_t) = -\nabla_w C(w_t) \quad (6)$$

while in proposed algorithm

$$\hat{H}(z_t, w_t) \neq -\nabla_w C(w_t)$$

In this case, if we assume original algorithm meets all the assumptions, the proposed algorithm also meets *Assumption 1* and *2*.

To prove *Assumption 3*, we first give the following lemma:

Lemma 1. For a stochastic variable x , we get another stochastic variable y by applying *Algorithm 1* to x with threshold τ , which means

$$y = \text{Prune}(x) = \begin{cases} x & \text{i.f.f. } |x| \geq \tau \\ 0 & \text{with probability } p = \frac{\tau - x}{\tau} \text{ i.f.f. } |x| < \tau \\ \tau & \text{with probability } p = \frac{x}{\tau} \text{ i.f.f. } |x| < \tau \end{cases}$$

Then y satisfies

$$\mathbf{E}[y] = \mathbf{E}[\text{Prune}(x)] = \mathbf{E}[x] \quad (7)$$

$$\mathbf{E}[y^2] = \mathbf{E}[\text{Prune}(x)^2] \leq \tau^2 + \mathbf{E}[x^2] \quad (8)$$

Then we can prove eq. (4) using *lemma 1*.

Proof.

$$\hat{\mathbf{H}}(z_t, w_t) = (g_w^1 \quad g_w^2 \quad \cdots \quad g_w^N)^T \quad (9)$$

where N is the total layer number of neural networks. The weight gradient of n -th layer

$$\mathbf{E}[\hat{g}_w^n] = \mathbf{E} \left[(F^n)'_w (\hat{g}_{A_O^n}) \Big|_{A_I^n} \right] \quad (10)$$

$$= (F^n)'_w (\mathbf{E} [\hat{g}_{A_O^n}]) \Big|_{A_I^n} \quad (11)$$

$$= (F^n)'_w (\mathbf{E} [\text{Prune}(\hat{g}_{A_I^{n+1}})]) \Big|_{A_I^n} \quad (12)$$

$$= (F^n)'_w \left(\mathbf{E} \left[\text{Prune} \left((F^{n+1})'_A (\hat{g}_{A_O^{n+1}}) \Big|_{A_I^{n+1}} \right) \right] \right) \Big|_{A_I^n} \quad (13)$$

$$= (F^n)'_w \left(\mathbf{E} \left[(F^{n+1})'_A (\hat{g}_{A_O^{n+1}}) \Big|_{A_I^{n+1}} \right] \right) \Big|_{A_I^n} \quad (14)$$

$$= (F^n)'_w \left((F^{n+1})'_A (\mathbf{E} [\hat{g}_{A_O^{n+1}}]) \Big|_{A_I^{n+1}} \right) \Big|_{A_I^n} \quad (15)$$

$$= (F^n)'_w \left(\cdots (F^N)'_A (\mathbf{E} [\hat{g}_{A_O^N}]) \Big|_{A_I^N} \cdots \right) \Big|_{A_I^n} \quad (16)$$

where F^n means the operation of n -th layer and Prune denotes our pruning algorithm. Equation (11) is right because convolution and its derivatives are linear.

Since $\hat{g}_{A_O}^N = g_{A_O}^N = \mathbf{Q}(z_t, w_t)$, we got

$$\forall n, \quad \mathbf{E}_{\mathbf{z}} [\hat{g}_w^n] = \mathbf{E}_{\mathbf{z}} [g_w^n]$$

which is combined with eq. (9) and finally gives that

$$\mathbf{E}_{\mathbf{z}} \hat{\mathbf{H}}(z, w_t) = \mathbf{E}_{\mathbf{z}} \mathbf{H}(z, w_t) = -\nabla_w C(w_t)$$

□

Similarly, it's easy to prove that eq. (5) is satisfied if original training method meets this condition with the same strategies. Thus, we can say that the proposed pruning algorithm has the same convergence with original algorithm under the GOGA framework.

4 Experimental Results

In this section, several experiments are conducted to demonstrate that the proposed approach could reduce the training complexity significantly with a negligible model accuracy loss. PyTorch [25] framework is adopted for all the evaluations.

4.1 Datasets and Models

Three datasets are utilized including CIFAR-10, CIFAR-100 [26] and ImageNet [27]. CIFAR-10 and CIFAR-100 datasets contains 32×32 pixels of RGB-colored images with 10 classes and 100 classes, respectively. Each CIFAR dataset contains 50000 pictures for training and 10000 pictures for testing, which is distributed uniformly for each class. ImageNet dataset contains 227×227 pixels of RGB-color images with 1000 classes, in which 1 million images for training and 50000 images for testing. AlexNet [6], ResNet [8] and MobileNet [9] are evaluated while ResNet include Res-{18, 34, 50, 101, 152} models.

The last layer size of each model is changed in order to adapt them on CIFAR datasets. Additionally for AlexNet, the kernels in first two convolution layers are set as 3×3 with padding = 2 and stride = 1. For FC-1 and FC-2 layers in AlexNet, they are also re-sized to 4096×2048 and 2048×2048 , respectively. For ResNet, kernels in first layer are replaced by 3×3 kernels with padding = 1 and stride = 1. Meanwhile, the pooling layer before FC-1 in ResNet is set to Average-Pooling with the size of 4×4 . For MobileNet, the kernels stride of first layer is changed to 1 and the last pooling layer is changed to Average-Pooling with size of 2×2 .

4.2 Training Settings

All the 7 models mentioned above are trained for 300 epochs on CIFAR-10, 100 datasets. While for ImageNet, only AlexNet, ResNet-18, 34, 50 and MobileNet are trained for 180 epochs due to our limited computing resources.

The Momentum SGD is used for all training with momentum = 0.9 and weight decay = 5×10^{-4} . Learning rate 1r is set to 0.05 for AlexNet and 0.1 for the others. 1r-decay is set to 0.1/100 for CIFAR-10, 100 and 0.1/45 for ImageNet.

Table 1: Evaluation results on CIFAR-10, where acc% means the training accuracy and ρ_{nnz} means the density of non-zeros.

Model	Baseline		$p = 85\%$		$p = 90\%$		$p = 95\%$		$p = 99.5\%$	
	acc%	ρ_{nnz}	acc%	ρ_{nnz}	acc%	ρ_{nnz}	acc%	ρ_{nnz}	acc%	ρ_{nnz}
AlexNet	90.50	0.09	90.34	0.01	90.55	0.01	90.31	0.01	89.66	0.01
ResNet-18	95.04	1	95.23	0.36	95.04	0.35	94.91	0.34	94.86	0.31
ResNet-34	94.90	1	95.13	0.34	95.09	0.32	95.16	0.31	95.02	0.28
ResNet-50	94.94	1	95.36	0.22	95.13	0.20	95.01	0.17	95.28	0.14
ResNet-101	95.60	1	95.61	0.24	95.48	0.22	95.60	0.19	94.77	0.12
ResNet-152	95.70	1	95.13	0.18	95.58	0.18	95.45	0.16	93.84	0.08
MobileNet	92.28	1	92.12	0.26	92.10	0.23	28.71	0.04	67.95	0.13

Table 2: Evaluation results on CIFAR-100, where acc% means the training accuracy and ρ_{nnz} means the density of non-zeros.

Model	Baseline		$p = 85\%$		$p = 90\%$		$p = 95\%$		$p = 99.5\%$	
	acc%	ρ_{nnz}	acc%	ρ_{nnz}	acc%	ρ_{nnz}	acc%	ρ_{nnz}	acc%	ρ_{nnz}
AlexNet	67.61	0.10	67.49	0.03	68.13	0.03	67.99	0.03	67.93	0.02
ResNet-18	76.47	1	76.89	0.40	77.16	0.39	76.44	0.37	76.66	0.34
ResNet-34	77.51	1	77.72	0.36	78.04	0.35	77.84	0.33	77.40	0.31
ResNet-50	77.74	1	78.83	0.25	78.27	0.22	78.92	0.20	78.52	0.16
ResNet-101	79.70	1	78.22	0.23	79.10	0.21	79.08	0.19	77.13	0.13
ResNet-152	79.25	1	80.51	0.22	79.42	0.19	79.76	0.18	76.40	0.10
MobileNet	68.21	1	8.68	0.02	68.55	0.25	53.45	0.16	9.82	0.03

Table 3: Evaluation results on ImageNet, where acc% means the training accuracy and ρ_{nnz} means the density of non-zeros.

Model	Baseline		$p = 85\%$		$p = 90\%$		$p = 95\%$		$p = 99.5\%$	
	acc%	ρ_{nnz}	acc%	ρ_{nnz}	acc%	ρ_{nnz}	acc%	ρ_{nnz}	acc%	ρ_{nnz}
AlexNet	56.38	0.07	57.10	0.05	56.84	0.04	55.38	0.04	39.58	0.02
ResNet-18	68.73	1	69.02	0.41	68.85	0.40	68.66	0.38	68.74	0.36
ResNet-34	72.93	1	72.92	0.39	72.86	0.38	72.74	0.37	72.42	0.34
MobileNet	70.76	1	70.94	0.32	70.09	0.28	70.23	0.27	0.84	0.01

4.3 Results and Discussion

As discussed previously, sparsity is different for Conv-ReLU and Conv-BN-ReLU structures but are covered by our evaluated three type of models. The percentage p in the proposed method varies from 85%, 90%, 95% to 99.5% for comparison with the baseline. All the training are run directly without any fine-tuning. The evaluation results are shown in Table 1, Table 2 and Table 3, where the non-zero gradients density means the percentage of non-zero gradients over all gradients.

Accuracy Analysis. From Table 1, Table 2 and Table 3, there is no accuracy lost for most situations. And even for ResNet-50 on CIFAR-100, there is 1% accuracy improvement. But for MobileNet and AlexNet on ImageNet, there is a significant accuracy loss when using very aggressive pruning policy like $p = 99.5\%$. On the other hand, it even does not converge when setting

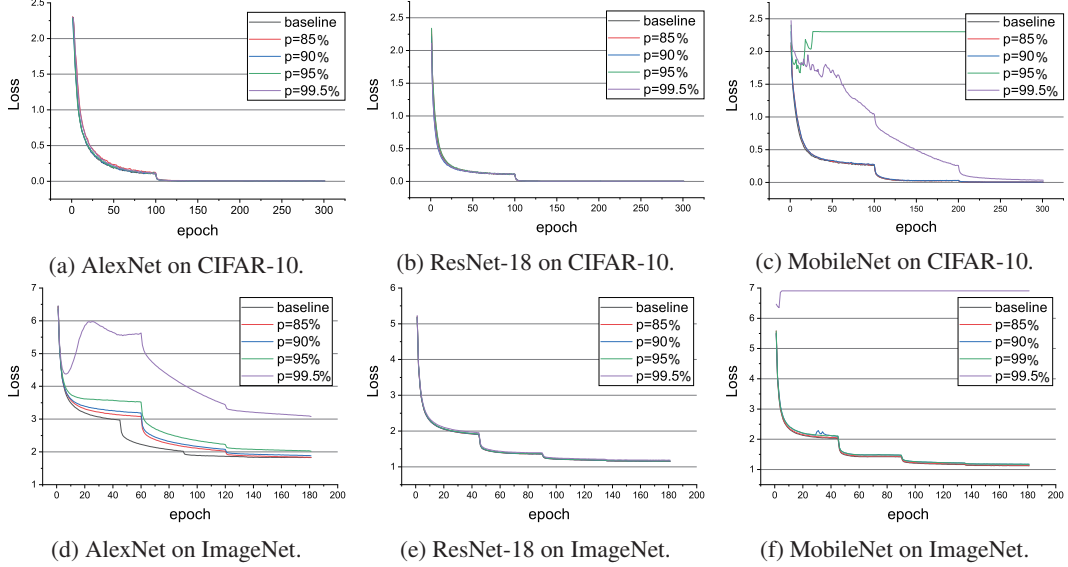


Figure 3: Training loss of AlexNet/ResNet/MobileNet on CIFAR-10/100 and ImageNet.

$p = 99.5\%$ for training MobileNet on ImageNet or CIFAR-100. The possible reason is that MobileNet is a compact neural network and there are not much redundancies on gradients, which makes it sensitive to gradients pruning. In summary, the accuracy loss is almost negligible when a non-aggressive policy is adopted for gradients pruning.

Gradients Sparsity. The gradients density illustrated in Table 1, Table 2, Table 3 has shown the ratio of non-zero gradients over all gradients, which could be a good measurement for sparsity. Although the basic block of AlexNet is Conv-ReLU whose natural density is relatively low, our method could still reduce the gradients density for more than $10\times$ on CIFAR-10 and $3\times \sim 5\times$ on CIFAR-100. While it comes to ResNet, whose basic block is Conv-BN-ReLU and activation gradients are naturally fully dense, our method could reduce the gradients density as $10\% \sim 30\%$. In addition, the deeper networks could obtain a relative lower gradients density with our sparsification, which means that it works better for complex networks. The potential benefits brought by gradients sparsification include reducing computation cost up to $10\times$ and the memory footprint up to $12\times$ on AlexNet.

Convergence Rate. The training loss is also demonstrated in Fig. 3 for AlexNet, ResNet-18 and MobileNet on CIFAR-10 and ImageNet datasets. Fig. 3b and Fig. 3e have shown that ResNet-18 is very robust for gradients pruning. For AlexNet, the gradients pruning could be still robust on CIFAR-10 but will degrade the accuracy on ImageNet. Fig. 3e also confirms that sparsification with a larger p will result in a worse convergence rate. For MobileNet, it will divergence if the pruning policy is too aggressive with $p = 99.5\%$, as shown in Fig. 3c and Fig. 3f.

5 Conclusions

In this paper, a new algorithm is proposed for dynamically gradients pruning for CNNs training. Different from original training, we assume the activation gradients of CNNs satisfy normal distribution and then estimate their variance according to their average value. After that, we calculate the pruning threshold according to the variance and a preset parameter p . The gradients are pruned stochastically if they are less than the threshold, which has been theoretically proved to be convergent. Evaluations on state-of-the-art models have confirmed that the gradients pruning approach could significantly reduce the computation cost of backward-propagation with a negligible accuracy loss.

References

- [1] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [2] Yang You, Zhao Zhang, Cho-Jui Hsieh, James Demmel, and Kurt Keutzer. Imagenet training in minutes. In *Proceedings of the 47th International Conference on Parallel Processing*, page 1. ACM, 2018.
- [3] Xianyan Jia, Shutao Song, Wei He, Yangzihao Wang, Haidong Rong, Feihu Zhou, Liqiang Xie, Zhenyu Guo, Yuanzhou Yang, Liwei Yu, et al. Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes. *arXiv preprint arXiv:1807.11205*, 2018.
- [4] Jian Cheng, Pei-song Wang, Gang Li, Qing-hao Hu, and Han-qing Lu. Recent advances in efficient computation of deep convolutional neural networks. *Frontiers of Information Technology & Electronic Engineering*, 19(1):64–77, 2018.
- [5] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.
- [6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
- [7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [9] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [10] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.
- [11] Huizi Mao, Song Han, Jeff Pool, Wenshuo Li, Xingyu Liu, Yu Wang, and William J Dally. Exploring the regularity of sparse structure in convolutional neural networks. *arXiv preprint arXiv:1705.08922*, 2017.
- [12] Sajid Anwar, Kyuyeon Hwang, and Wonyong Sung. Structured pruning of deep convolutional neural networks. *ACM Journal on Emerging Technologies in Computing Systems*, 13(3):32, 2017.
- [13] Vadim Lebedev and Victor Lempitsky. Fast ConvNets using group-wise brain damage. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2554–2564, 2016.
- [14] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5058–5066, 2017.
- [15] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1389–1397, 2017.
- [16] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2736–2744, 2017.
- [17] Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*, 2017.
- [18] Aaditya Prakash, James Storer, Dinei Florencio, and Cha Zhang. Repr: Improved training of convolutional filters. *arXiv preprint arXiv:1811.07275*, 2018.

- [19] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *Proceedings of the International Conference on Machine Learning*, pages 1737–1746, 2015.
- [20] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.
- [21] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Tern-Grad: Ternary gradients to reduce communication in distributed deep learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1509–1519, 2017.
- [22] Eunhyeok Park, Sungjoo Yoo, and Peter Vajda. Value-aware quantization for training and inference of neural networks. In *Proceedings of the European Conference on Computer Vision*, pages 580–595, 2018.
- [23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [24] Léon Bottou. Online learning and stochastic approximations. *On-Line Learning in Neural Networks*, 17(9):142, 1998.
- [25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Workshop*, 2017.
- [26] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- [27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.