Special Paper
방송공학회논문지 제26권 제7호, 2021년 12월 (JBE Vol. 26, No. 7, December 2021)
https://doi.org/10.5909/JBE.2020.26.7.862
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Pruning for Robustness by Suppressing High Magnitude and Increasing Sparsity of Weights

Incheon Cho[a)], Muhammad Salman Ali[a)], and Sung-Ho Bae[a)‡]

Abstract

Although Deep Neural Networks (DNNs) have shown remarkable performance in various artificial intelligence fields, it is well known that DNNs are vulnerable to adversarial attacks. Since adversarial attacks are implemented by adding perturbations onto benign examples, increasing the sparsity of DNNs minimizes the propagation of errors to high-level layers. In this paper, unlike the traditional pruning scheme removing low magnitude weights, we eliminate high magnitude weights that are usually considered high absolute values, named 'reverse pruning' to ensure robustness. By conducting both theoretical and experimental analyses, we observe that reverse pruning ensures the robustness of DNNs. Experimental results show that our reverse pruning outperforms previous work with 29.01% in Top-1 accuracy on perturbed CIFAR-10. However, reverse pruning does not guarantee benign samples. To relax this problem, we further conducted experiments by adding a regularization term for the high magnitude weights. With adding the regularization term, we also applied conventional pruning to ensure the robustness of DNNs.

Keyword: convolutional neural networks, image classification, adversarial attack, pruning

Ⅰ. Introduction

Deep Neural Networks (DNNs) have shown state-of-the-art performance in various fields of artificial intelligence[1]-[7]. However, their characteristics, such as excessive memory usage and high computational complexity, hindered their applications in the real world. To solve this problem and make DNNs less resource-intensive, various model compression methods[8]-[13] were proposed. Pruning is one of the model compression methodologies, and previous pruning studies[14]-[16] have conducted research results that show no degradation in performance even after removing more than 90% of the weights.

On the other hand, it is well known that DNN models are vulnerable to adversarial attacks[17]. Adversarial attacks usually generate images that are perceptually similar to benign samples, but it leads to misclassification by the neural

a) Department of Computer Science and Engineering, Kyung Hee Univ.
‡ Corresponding Author : Sung-Ho Bae
E-mail: shbae@khu.ac.kr
Tel: +82-31-201-2593
ORCID: https://orcid.org/0000-0002-3389-1159

※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2018R1C1B 3008159). Also, this research was a result of a study on the "HPC Support" Project, supported by the 'Ministry of Science and ICT' and NIPA.
· Manuscript received October 25, 2021; Revised November 30, 2021; Accepted November 30, 2021.

Copyright ⓒ 2021 Korean Institute of Broadcast and Media Engineers. All rights reserved.
"This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered."

network. Furthermore, generic regularizations (e.g., dropout and weight decay) are ineffective against adversarial attacks[18]. To ensure the robustness of DNNs, numerous attempts[18]-[20] have been made to develop robust architectures.

Therefore, it is crucial to study the relationship between the efficiency and robustness of neural networks. In this paper, we thoroughly analyze this relationship and provide analyses from both theoretical and empirical perspectives. We found that increasing sparsity and limiting the weight's magnitude contributes to the robustness of DNNs. In addition to conventional pruning, which removes the weights under a certain threshold, we proposed a method for generating higher magnitude weights. Our results show that our proposed method performs well against adversarial attacks.

Ⅱ. Main Subject

1. Related Works

Previous studies[21][22] showed a relationship between robustness and sparsity. Authors in [21] proved the relationship between robustness and sparsity both theoretically and practically. In the theorem presented by the previous study[21], it was shown that the ℓ_1 norm of the weights is inversely proportional to the robustness. Based on this evidence, it was experimentally proved that the more sparsity is ensured, the more robust the model is. The weights having small magnitude were removed through pruning to ensure high sparsity in the previous study. Removal of small weights using pruning helps to ensure sparsity, but it does not minimize distortion caused by adversarial attacks. In this paper, we conducted research focusing on the minimization of output distortion caused by adversarial attacks. Our study theoretically and experimentally proved that not only an increase in sparsity but also a decrease in the magnitude of the weights are related to increasing robustness.

2. Notations

Assume a DNN $\hat{y}=f(x;W^{1:d})$ has a set of d convolution layers where $W^{1:d}=\{W^1, W^2, ..., W^d\}$. x and y are the input and true label as $(x,y)\in D$ where D is data distribution, and $\hat{y}$ is the predicted label of x. The output of DNN which consist of depth-d fully connected neural net, can be formulated as

$$f(x;W^{1:d}) = W^{(d)}\sigma(W^{d-1}\sigma(\cdots W^2\sigma(W^1x)\cdots)) \quad (1)$$

where σ denotes the ReLU activation. $\tilde{x}=x+\delta$ be the adversarial examples generated by adding perturbation term δ.

3. Minimizing Output Distortion

There is a difference between the outputs of the DNN receiving the adversarial attacked inputs and the outputs of DNN receiving the benign inputs. This difference can be referred to as distortion, and creating a model that minimizes this distortion will increase robustness. The distortion can be calculated as

$$\| f(x;W^{1:d}) - f(\tilde{x};W^{1:d}) \|_1. \quad (2)$$

Then, we can use the following upper bound of the Eq. (2): the inequality equation can be derived as

$$\| f(x;W^{1:d}) - f(\tilde{x};W^{1:d}) \|_1 \leq \Pi_{j=1}^{d} \| W^j \|_1 \cdot \| \delta \|_1. \quad (3)$$

In the Appendix section, we describe the derivation procedure in detail and extend Eq. (3) for convolution operation. In Eq. (3), ℓ_1 the norm of W^j is related to sparsity and magnitude of the elements in W^j. The upper bound of distortion may decrease when 1) W^j is sparse and 2) the maximum value of ℓ_1 norm of W^j is decreased.

4. Reverse Pruning

To reduce the magnitude of distortion, we propose a new

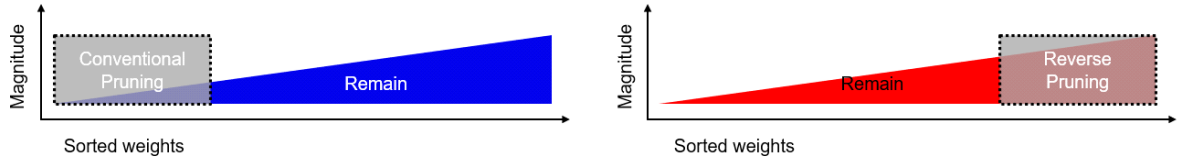

Fig. 1. Overview of conventional pruning and reverse pruning. Reverse pruning method eliminates the weights which have high magnitude

pruning method to guarantee robustness named reverse pruning (RP). Unlike the conventional pruning (CP) method, which removes the elements of the weights lower than the threshold, reverse pruning removes the elements of the weights higher than the threshold. Figure. 1 illustrates an overview of conventional pruning and reverse pruning. After performing reverse pruning, DNN consists of weights that have low magnitude.

To compare conventional pruning and reverse pruning, we conducted experiments on CIFAR-10 with ResNet56[2]. In Figure. 2, reverse pruning gives the best performance for the perturbed data samples generated using adversarial attacks. However, the performance of reverse pruning for benign data samples is lower than that of conventional pruning. Furthermore, a high pruning ratio at reverse pruning impedes training by making it hard to converge to an optimal value. By ensuring sparsity and low magnitude values, we solve this problem. For this purpose, we combined conventional pruning and regularization for high magnitude weights.

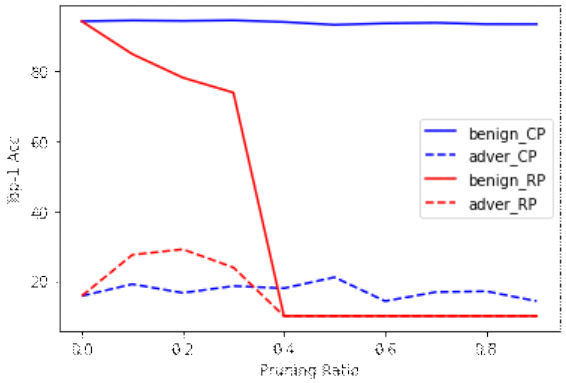

Fig. 2. Comparison of reverse pruning (RP) with conventional pruning (CP)

5. Combining Conventional Pruning and Regularization

As mentioned in the previous section, we need to increase sparsity and reduce the magnitude of the weights for reducing the weights norm. To satisfy these conditions, we combined the characteristics of both conventional pruning and reverse pruning.

First, to prevent generating high magnitude weights, we added a regularization term to the objective function. Then, the final objective function is calculated as

$$Loss = CE(\hat{y}, y) + \alpha \cdot ReLU(\| W^{1:d} \|_2^2 - \tau) \quad (4)$$

where α is the scale value to reflect the confidence of the regularization term and τ is threshold value. With this regularization term, we are able only to restrain generating high magnitude weights.

Second, to generate sparse weights, we applied conventional pruning with Eq. (4). By training DNN in these constraints, we can minimize the distortion of samples on which adversarial attacks were applied. After training with these constraints, we check representative values such as the quantiles of the weights, mean and ℓ_1-Norm of the weights. In Table. 1, we can compare our proposed method with the previous method for both ranges of the weights and ℓ_1-Norm of the weights. Our proposed method has a low range of the weights and ℓ_1-Norm of the weights. Since pruning converts the weights to zero, a significant number of weights in the distribution of quartiles is measured as zero.

Table. 1. Representative values for each method. Our proposed method reduces both range of the weights and ℓ_1-Norm of the weights

	Quantile-0%	Quantile-25%	Quantile-50%	Quantile-75%	Quanitle-100%	ℓ_1-Norm
Baseline	-1.337	-0.029	-0.004	0.021	1.811	28082.62
CP	-1.176	-0.023	0.000	0.000	1.613	20576.11
RP	-0.278	-0.039	0.000	0.030	0.322	38442.54
Ours	-0.223	-0.022	0.000	0.000	0.223	22022.85

6. Experimental Results

To validate our proposed method, we conducted experiments for the image classification task on the CIFAR-10 dataset with ResNet56. Our proposed method can be applied to various adversarial attacks which perturb data samples by adding noise. In this paper, we focus on PGD[23] attack to generate perturbed samples. Except for reverse pruning, all experiments are conducted on 50% of sparsity and we observed performance change with or without the regularization of the weights. For reverse pruning, we set the pruning ratio to 10% to ensure the adequate performance of both benign data and perturbed data. The reason why the pruning ratio is set differently for each method is because the ratio when robustness is maximized was used. In this paper, we set the hyperparameters of Eq. (4) with $\alpha = 0.1, \tau = 0.05$. τ was established based on the distribution of weights generated by reverse pruning (details in Table 1). Table 2 illustrates the performance of robustness for various sparse methods. We conducted experiments to observe performance changes with the strength of the noise. Compared to the previous method[21], which utilizes conventional pruning to ensure sparsity, our proposed method performs well against adversarial attack. Reverse pruning performs poorly when the noise intensity is low. Reverse pruning seems to cause some underfitting because it removes the high magnitude weights. Therefore, it is assumed that poor performance occurs due to underfitting for weakly disturbed data. With these results, we showed that our experimental results prove our theoretical analysis.

Table. 2. Robustness of conventional pruning and our proposed method. Robustness for PGD attack is evaluated against four levels of perturbation (ϵ) values. Our proposed method performs well against PGD attacks and perturbation budgets. In the notation of 'PGD-N', N indicates step size of PGD attack

Adversarial Attack	ϵ	Baseline	Conventional Pruning	Reverse Pruning	Ours
Benign	0	94.21	94.17	84.44	93.84
PGD-4	1	44.16	44.33	33.79	45.41
	2	36.08	36.21	31.50	36.93
	4	23.75	24.00	27.02	24.84
	8	15.73	15.85	23.29	17.83

Ⅲ. Conclusion

In this paper, we checked how the high magnitude weights and the sparsity of DNNs affect the robustness. We proposed a mechanism to minimize output distortion between perturbed and original images. In theoretical analysis, we found that the sparser the weights of the model, and the smaller the magnitude of the weights, the more robust the model can be. Our experiments are in line with our theoretical analysis. Compared to the previous study, our proposed method performed well on perturbed samples with a slight degradation in accuracy on benign samples.

Ⅳ. Appendix

1. Proof of Equation (3)

In this section, we prove the Eq. (3). With Cauchy-Schwarz inequality, Eq. (2) can be reconstructed as:

$$\begin{aligned} &\| f(x;W^{1:d}) - f(\tilde{x};W^{1:d}) \|_1 \\ &= \| W^d(\sigma(f(x;W^{1:d-1})) - \sigma(f(\tilde{x};W^{1:d-1}))) \|_1 \\ &\leq \| W^d \|_1 \cdot \| \sigma(f(x;W^{1:d-1})) - \sigma(f(\tilde{x};W^{1:d-1})) \| \end{aligned} \quad (5)$$

Using 1-Lipschitzness of ReLU activation with respect to ℓ_1 norm at Eq. (5), we can modify Eq. (5) as:

$$\begin{aligned} &\| f(x;W^{1:d}) - f(\tilde{x};W^{1:d}) \| \\ &\leq \| W^d \|_1 \cdot \| f(x;W^{1:d-1}) - f(\tilde{x};W^{1:d-1}) \|_1 \end{aligned} \quad (6)$$

With iteratively conducting procedure of Eq. (5) and Eq. (6), Eq. (2) can be finally rewritten as:

$$\begin{aligned} &\| f(x;W^{1:d}) - f(\tilde{x};W^{1:d}) \|_1 \\ &\leq \prod_{j=2}^{d} \| W^j \|_1 \cdot \| W^1 x - W^1(x+\delta) \|_1 \\ &= \Pi_{j=2}^{d} \| W^j \|_1 \cdot \| W^1 \delta \|_1 \leq \Pi_{j=1}^{d} \| W^j \|_1 \cdot \| \delta \|_1 \end{aligned} \quad (7)$$

We can also extend this procedure to convolution operations. With Young's convolution inequality, we can derive Eq. (7).

References

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,''in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1 – 9.

[2] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770 – 778.

[3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, ''MobileNets: Efficient convolutional neural networks for mobile vision applications,'' 2017, arXiv:1704.04861.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ''Rich feature hierarchies for accurate object detection and semantic segmentation,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580 – 587.

[5] S. Ren, K. He, R. B. Girshick, and J. Sun, ''Faster R-CNN: Towards real time object detection with region proposal networks,'' in Proc. NIPS, 2015, pp. 91 – 99.

[6] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431 – 3440.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional net s, atrous convolution, and fully connected CRFs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834 – 848, Apr. 2017.

[8] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, ''Pruning filters for efficient ConvNets,'' 2016, arXiv:1608.08710.

[9] Y. He, P. Liu, Wang Z. Hu, and Y. Yang, "Filter pruning via geometric median for deep convolutional neural networks acceleration," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4340 – 4349.

[10] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, ''Learned step size quantization,'' 2019, arXiv:1902.08153.

[11] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

[12] Z. Wang, X. Cheng, G. Sapiro, and Q. Qiu, ''ACDC: Weight sharing in atom-coefficient decomposed convolution,'' 2020, arXiv:2009.02386.

[13] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, and V. Lempitsky, ''Speeding-up convolutional neural networks using fine-tuned CP- decomposition,'' in Proc. 3rd Int. Conf. Learn. Represent., (ICLR) Conf. Track, 2015, pp. 2 – 12.

[14] Han. S., Pool. J., Tran. J., and Dally. W. J. "Learning both weights and connections for efficient neural network.", in Proc. Advances in neural information processing systems, Jun. 2015, pp.1135 - 1143.

[15] Y. Guo, A. Yao, and Y. Chen, ''Dynamic network surgery for efficient DNNs,'' in Proc. NIPS, 2016, pp. 1387 – 1395.

[16] Molchanov, D., Ashukha, A., and Vetrov, D. "Variational dropout sparsifies deep neural networks," In International Conference on Machine Learning ,pp. 2498-2507, July 2017.

[17] Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks," In 2017 ieee symposium on security and privacy (sp), pp. 39-57, May 2017.

[18] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[19] Lu, J., Issaranon, T., and Forsyth, D. Safetynet: Detecting and rejecting adversarial examples robustly. In Proceed ings of the IEEE International Conference on Computer Vision, pp. 446 – 454, 2017.

[20] Weng, Tsui-Wei, et al. "Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach," In International Conference on Learning Representations, 2018.

[21] Guo, Y., Zhang, C., Zhang, C., and Chen, Y. "Sparse dnns with improved adversarial robustness," Advances in Neural Information Processing Systems, 31:242 – 251, 2018

[22] Ye, S., Xu, K., Liu, S., Cheng, H., Lambrechts, J.-H., Zhang, H., Zhou, A., Ma, K., Wang, Y., and Lin, X. "Adversarial robustness vs. model compression, or both?," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 111 – 120, 2019.

[23] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. "Towards deep learning models resistant to adversarial attacks," In International Conference on Learning Representations, 2018.

Introduction Authors

Incheon Cho

- May. 2020 ~ : He pursuining the M.S. degree with the Department of Computer Science and Engineering, Kyung Hee Universiity, Yongin, South Korea.
- May. 2014 ~ Feb. 2020 : He received the bachelor's degree from the Department of Computer Science and Engineering, Kyung Hee University, Yongin, South Korea.
- ORCID : http://orcid.org/0000-0001-5641-5596
- Interests : Model Compression, Pruning, Knowledge Distillation, Adversarial Attack

Muhammad Salman Ali

- May. 2018 ~ : He pursuining the M.S. degree leading to Ph. D. degree with the Department of Computer Science and Engineering, Kyung Hee University, South Korea.
- May. 2013 ~ Feb. 2017 : He received the bachelor's degree in computer science fromt the National University of Sciences and Technology (NUST), Islamabad, Pakistan.
- ORCID : http://orcid.org/0000-0002-8548-3827
- Interests : Interpretation of Deep Neural Networks

Sung-Ho Bae

- Sep. 2017 ~ : He has been an Assistant Professor with Department of Computer Science and Engineering, Kyung Hee University, Yongin, South Korea.
- Jul. 2016 ~ Aug. 2017 : He was a Postdoctoral Associate with Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), MA, USA.
- Feb. 2011 ~ Aug. 2016 : He received M.S. and Ph. D. degrees with Department of Electrical and Electronic Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea.
- May. 2004 ~ Feb. 2011 : He received the B.S. degree with he Department of Computer Science and Engineering/Electronics from Kyung Hee University, South Korea.
- ORCID : https://orcid.org/0000-0002-3389-1159
- Interests : Adversarial Attack, Model Compression, Semi-supervised Learning and Data Augmentation, Explainable DNN, Neural Architecture Search for Hardware Implementation, International Standardization