Proceedings

and Quantization

Authors: Zhenshan Bao, Jiayang Liu, Wenbo Zhang Authors Info & Affiliations

Publication:
MLMI 2019: Proceedings of the 2019 2nd International Conference on Machine Learning and Machine Intelligence • September 2019 • Pages 3–6 • https://doi.org/10.1145/3366750.3366751

0 10

Get Access

19

ABSTRACT

As the complexity of processing issues increases, deep neural networks require more computing and storage resources. At the same time, the researchers found that the deep neural network contains a lot of redundancy, causing unnecessary waste, and the network model needs to be further optimized. Based on the above ideas, researchers have turned their attention to building more compact and efficient models in recent years, so that deep neural networks can be better deployed on nodes with limited resources to enhance their intelligence. At present, the deep neural network model compression method have weight pruning, weight quantization, and knowledge distillation and so on, these three methods have their own characteristics, which are independent of each other and can be self-contained, and can be further optimized by effective combination. This paper will construct a deep neural network compression framework based on weight pruning, weight quantization and knowledge distillation. Firstly, the model will be double coarse-grained compression with pruning and network, thereby further accelerating and compressing the model to make the loss of accuracy smaller. The experimental results show that the combination of three algorithms can

Proceedings

References

1. LeCun, Y., Denker, J. S., and Solla, S. A. 1990. Optimal brain damage. NIPS'89 Proceedings of the 2nd International Conference on Neural Information Processing Systems. 598--605.

2. Hassibi, B., and Stork, D. G. 1993. Second order derivatives for network pruning: Optimal brain surgeon. In Advances in neural information processing systems. 164--171.

3. Han, S., Pool, J., Tran, J., and Dally, W. 2015. Learning both weights and connections for efficient neural network. In Advances in neural information processing systems. 1135--1143.

4. Molchanov, P., Tyree, S., Karras, T., Aila, T., and Kautz, J. 2017. Pruning convolutional neural networks for resource efficient inference. In International Conference of Learning Representation. arXiv preprint arXiv:1611.06440

5. He, Y., Liu, P., Wang, Z., and Yang, Y. 2018. Pruning Filter via Geometric Median for Deep Convolutional Neural Networks Acceleration. arXiv preprint arXiv:1811.00250.

6. Singh, P., Verma, V. K., Rai, P., and Namboodiri, V. P. 2018. Leveraging Filter Correlations for Deep Model Compression. arXiv preprint arXiv:1811.10559.

7. Gong, Y., Liu, L., Yang, M., and Bourdev, L. 2014. Compressing deep convolutional networks using vector quantization. arXiv preprint arXiv:1412.6115.

8. Wu, J., Leng, C., Wang, Y., Hu, Q., and Cheng, J. 2016. Quantized convolutional neural networks for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4820--4828.

9. Gupta, S., Agrawal, A., Gopalakrishnan, K., and Narayanan, P. 2015. Deep learning with limited numerical precision. In International Conference on Machine Learning. 1737--1746.

10. Gysel, P., Motamedi, M., and Ghiasi, S. 2016. Ristretto: Hardware-oriented approximation of convolutional neural networks. arXiv preprint arXiv:1605.06402

Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., and Bengio, Y. 2016. Binarized neural
ks: Training deep neural networks with weights and activations constrained to+ 1 or −1. arXiv
it arXiv:1602.02830.

preprint arXiv:1503.02531.

14. Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. 2014. Fitnets: Hints for

Proceedings

15. Yim, J., Joo, D., Bae, J., and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4133--4141.

16. Mishra, A., and Marr, D. 2018. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In International Conference of Learning Representation.

17. Han, S., Mao, H., and Dally, W. J. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149.

18. Oguntola, I., Olubeko, S., and Sweeney, C. 2018. SlimNets: An Exploration of Deep Model Compression and Acceleration. In 2018 IEEE High Performance extreme Computing Conference.1--6.

19. Polino, A., Pascanu, R., and Alistarh, D.2018. Model compression via distillation and quantization. arXiv preprint arXiv:1802.05668

Show Fewer References

Index Terms

Using Distillation to Improve Network Performance after Pruning and Quantization

Computing methodologies

Artificial intelligence

Search methodologies

DL Comment Policy

Comments should be relevant to the contents of this article, (sign in required)

Proceedings

0 Comments

Tweet Share **Sort by Newest**

Nothing in this discussion yet.

View Table Of Contents

Categories

Journals
Magazines
Books
Proceedings
SIGs
Conferences
Collections
People

About

About ACM Digital Library
Subscription Information
Author Guidelines
Using ACM Digital Library
All Holdings within the ACM Digital Library
ACM Computing Classification System

Join

Join ACM
Join SIGs
Subscribe to Publications
Institutions and Libraries

Connect

Contact
Facebook
Twitter
Linkedin

Proceedings