Reading List

- Song Han, Huizi Mao, William J. Dally, Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding. ICLR 2016

- Blalock, D., Gonzalez Ortiz, J.J., Frankle, J. and Guttag, J., 2020. What is the state of neural network pruning?. Proceedings of machine learning and systems, 2, pp.129-146.

- Tan, M. and Le, Q., 2019, May. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

- Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J. and Keutzer, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.

- Liu, Z., Sun, M., Zhou, T., Huang, G. and Darrell, T., 2018. Rethinking the value of network pruning. arXiv preprint arXiv:1810.05270.