On the Distributional Properties of Adaptive Gradients (Supplementary Material)

Zhang Zhiyi[1] **Liu Ziyin**[2]

[1]School of Statistics, Xi'an University of Finance and Economics
[2]Department of Physics, University of Tokyo

A ADDITIONAL EXPERIMENTS

A.1 DISTRIBUTION ACROSS DIFFERENT RANDOM SEEDS

We also plot compare the distribution over different random initializations in Figure 1. We plot the overlap of three different random seed in dark red, which also agrees well with the prediction. We see that the variance of the distribution across different seeds is relatively small, and all agree well with the theoretical prediction.

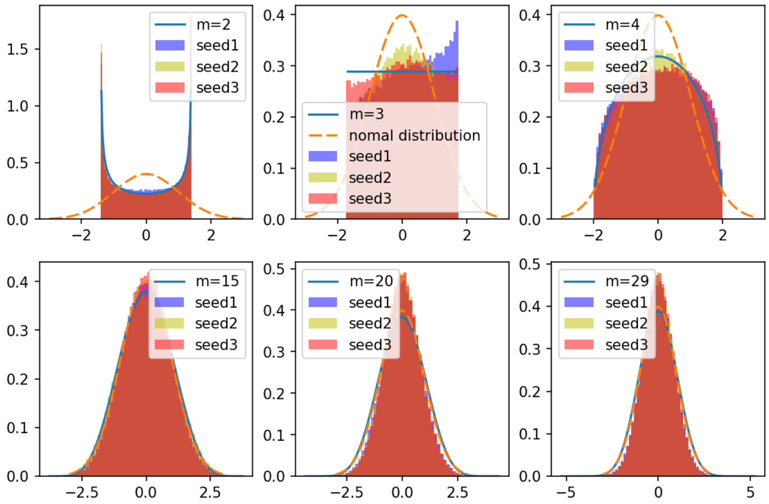

Figure 1: Update distribution for different random seeds.

A.2 OTHER EXPERIMENTS ON CIFAR-10

In this section, we present more experiments to validify our theory. We first plot the distribution of update for a different layer (from what appeared in the main text) RegNetX-200MF trained on CIFAR-10, and excellent agreement between our theory and experiment is observed. We then plot the distribution of updates from random layers (with more than 10^4 parameters) of VGG [Simonyan and Zisserman, 2014], ResNet-18 [He et al., 2016], ShuffleNet-V2 [Zhang et al., 2018], ResNeXt-29 [Xie et al., 2017], MobileNet [Howard et al., 2017], EfficientNet-B0 [Tan and Le, 2019], DenseNet-121 [Huang et al., 2017]. We first plot the variance of θ and $|\theta|$ in Figure 2. While some nets deviate from the theory, others agree quite well. For those that deviate from the prediction, we observe that they all have smaller variances than the predicted value, this agrees with our message that the adaptive gradient methods do not have exploding variance problem at the beginning of training. We then plot the distributions of θ for a single trajectory for each of these models.

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

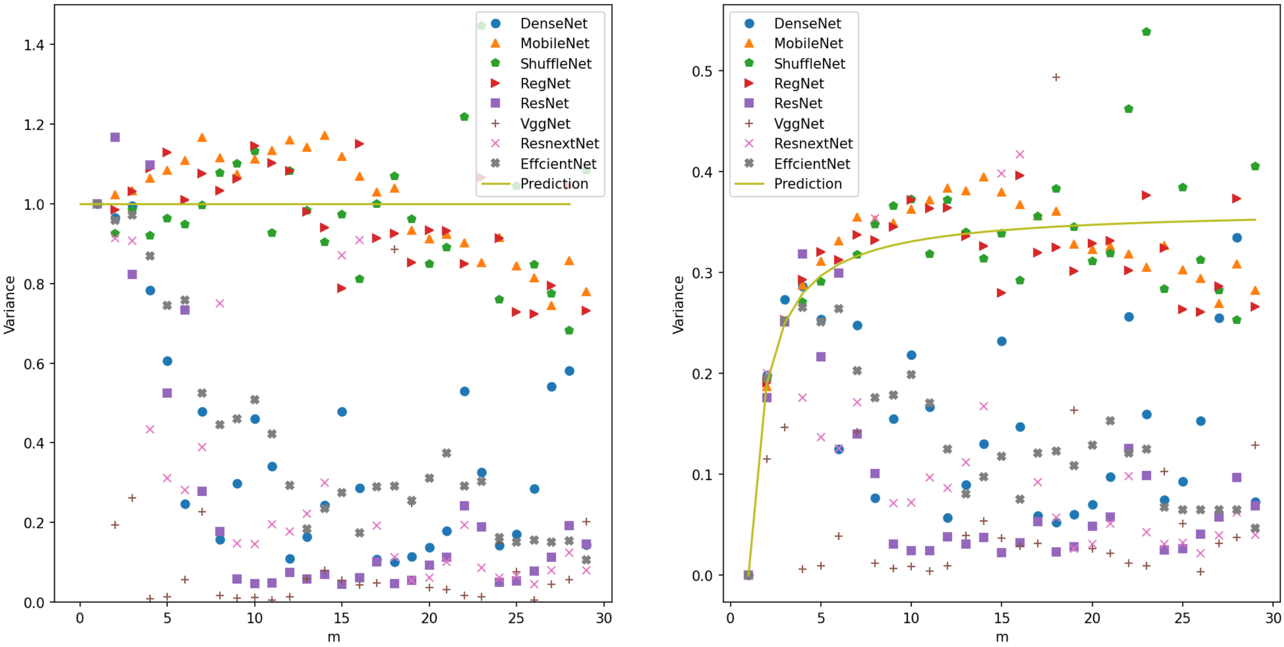

Figure 2: Left: variance of θ. Right: variance of $|\theta|$. We see that, some architectures agree quite well our analysis, while other architectures deviates in a rather consistent way.

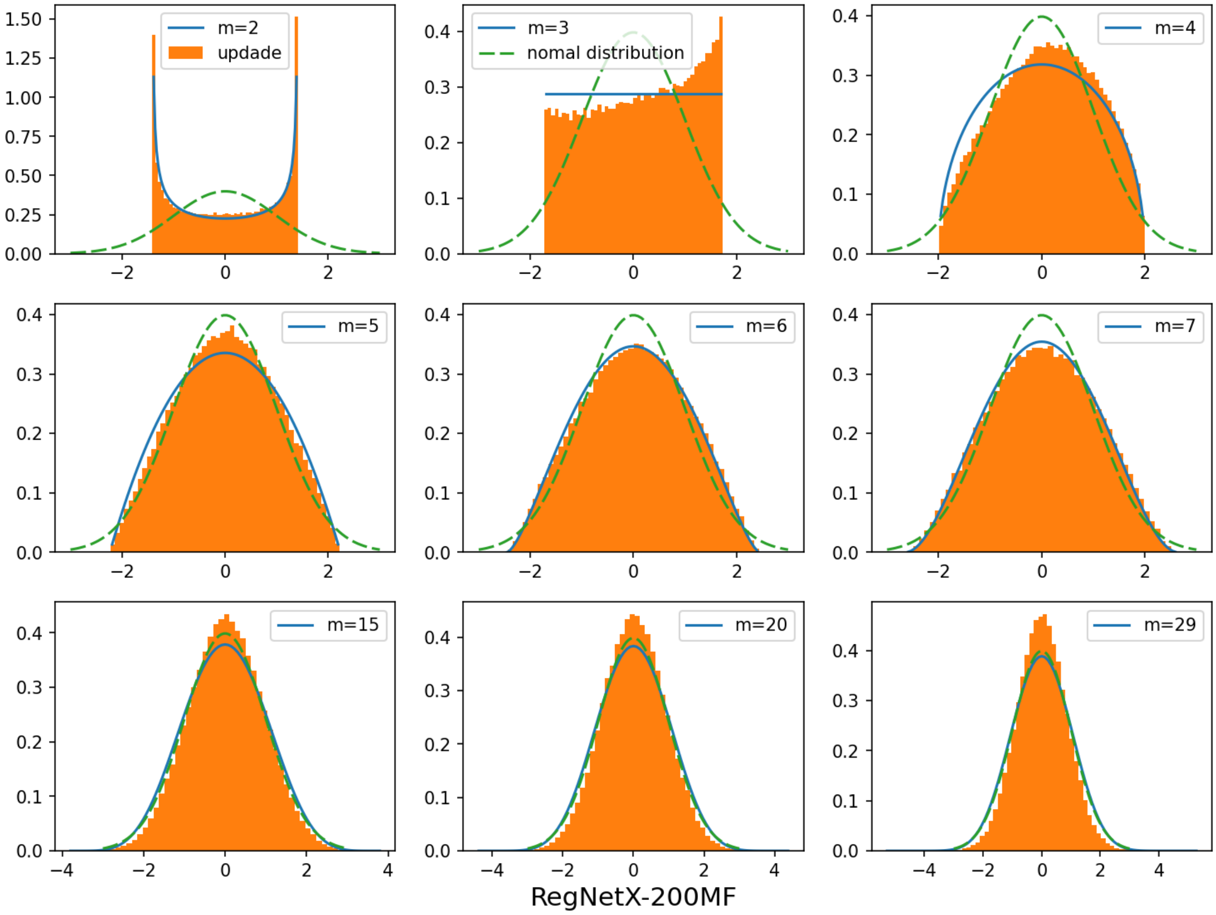

Figure 3: Distribution of the update distribution of another non-hand-picked layer of a RegNetX-200MF trained on CIFAR-10.

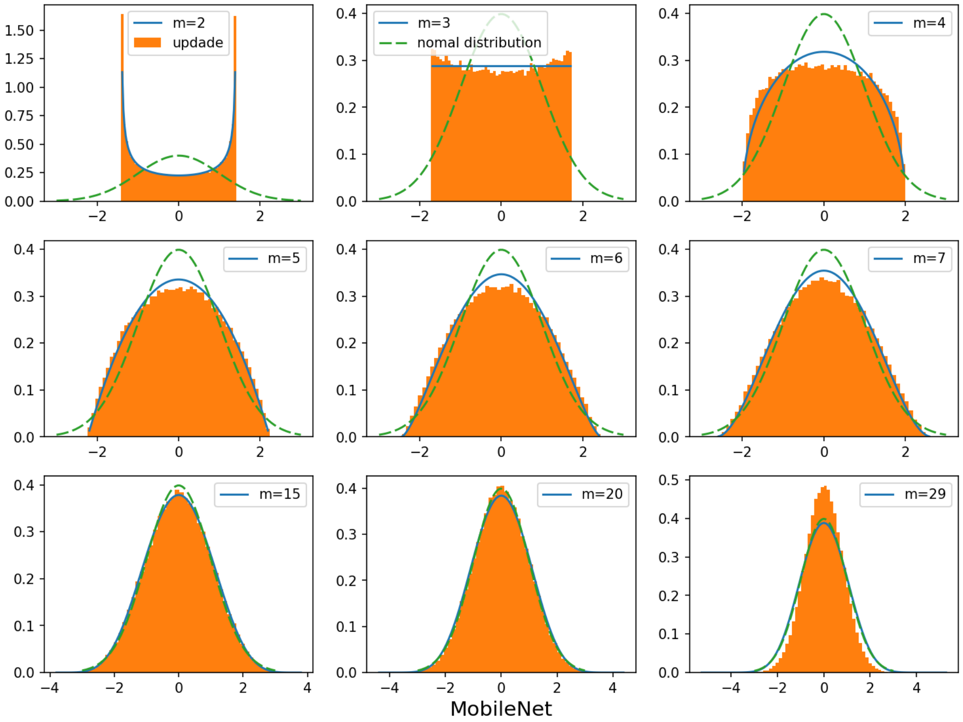

Figure 4: Distribution of the update distribution of another non-hand-picked layer of a MobileNet trained on CIFAR-10.

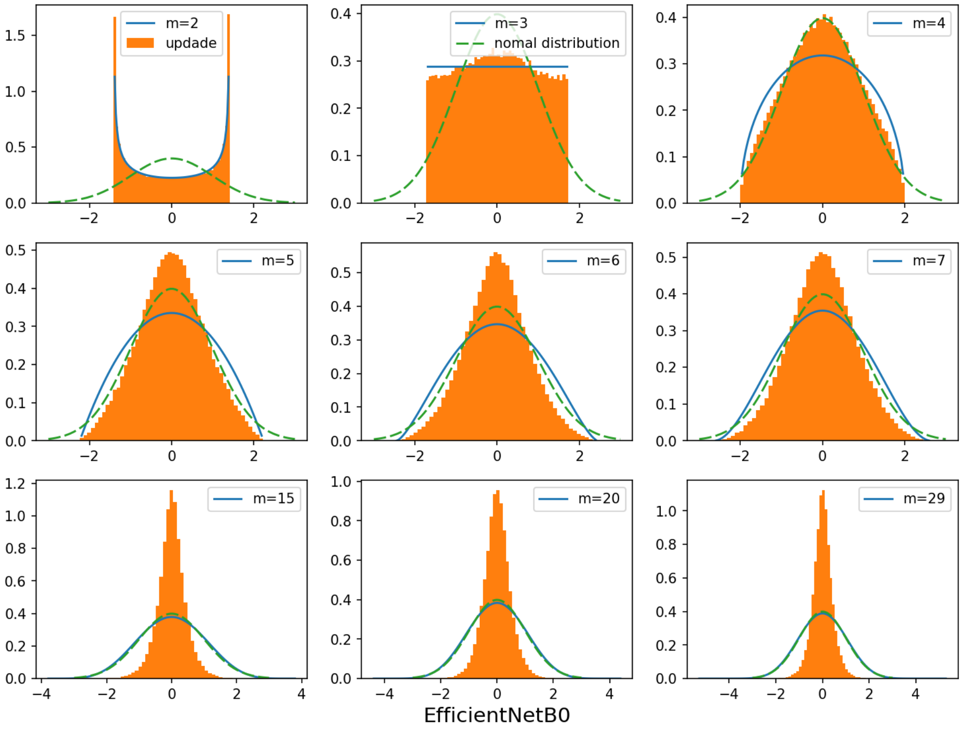

Figure 5: Distribution of the update distribution of another non-hand-picked layer of a EfficientNet trained on CIFAR-10.

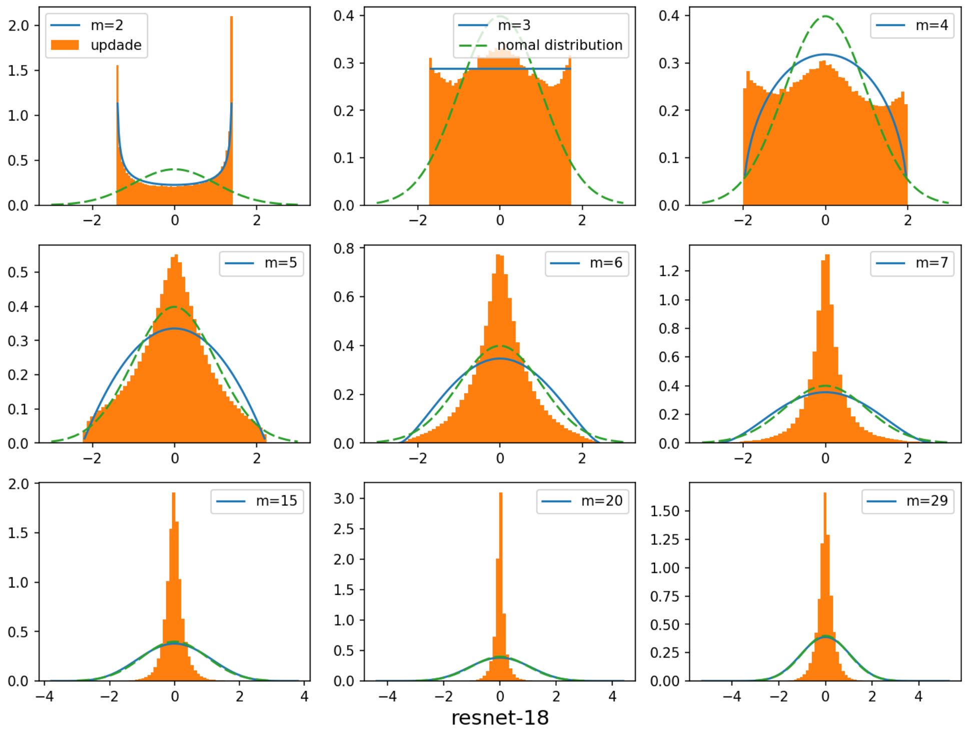

Figure 6: Distribution of the update distribution of another non-hand-picked layer of a ResNet-18 trained on CIFAR-10.

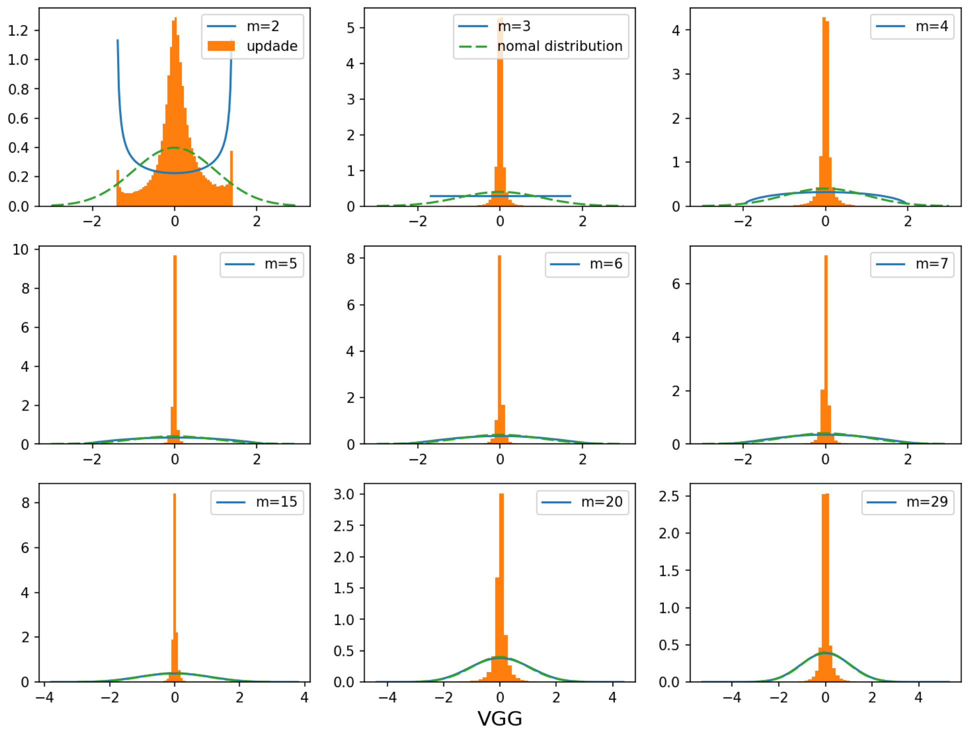

Figure 7: Distribution of the update distribution of another non-hand-picked layer of a VGG trained on CIFAR-10.

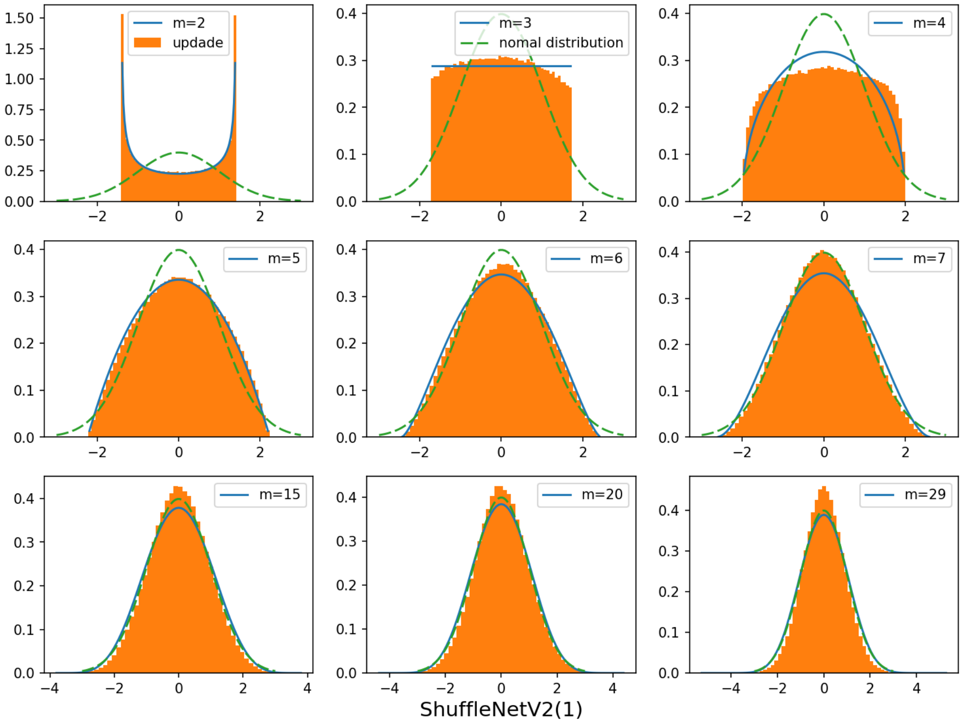

Figure 8: Distribution of the update distribution of another non-hand-picked layer of a ShuffleNet trained on CIFAR-10.

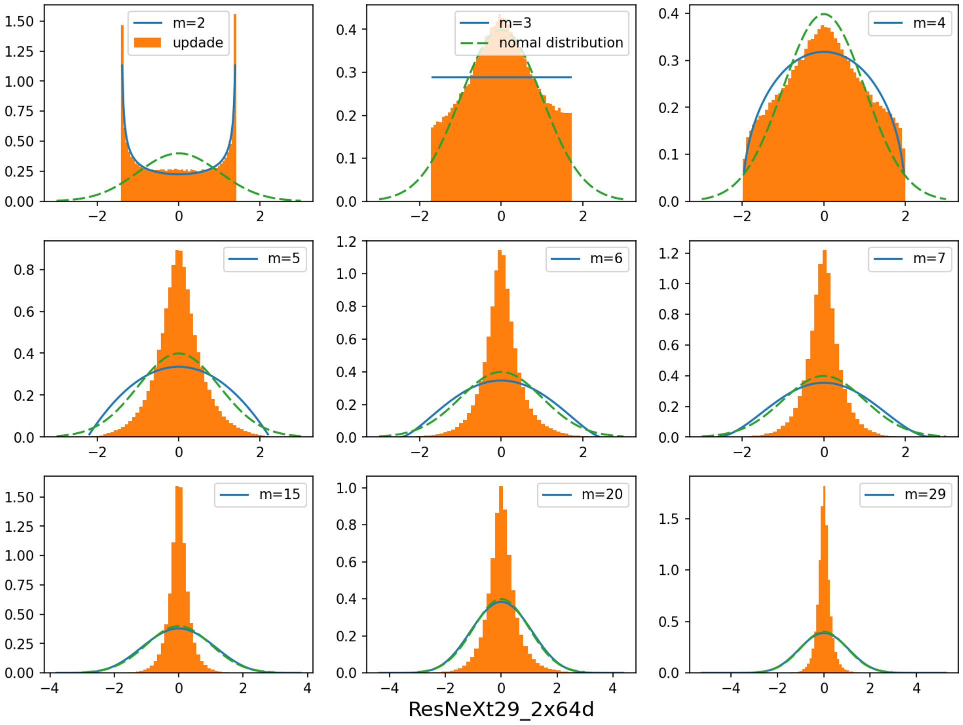

Figure 9: Distribution of the update distribution of another non-hand-picked layer of a ResNeXt29 trained on CIFAR-10.

References

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.

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.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.