

Large-scale distributed training of deep neural networks suffers from the generalization gap caused by the increase in the effective mini-batch size. Previous ap