Author Response for "Stochastic Normalization"

We thank all reviewers for insightful and professional comments. In general, reviewers find the paper well written, the topic important, and the method novel. Major concerns are addressed here, which will be incorporated to the revision.

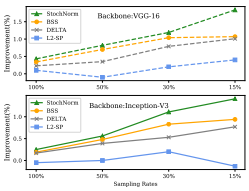
Reviewer #1

Question 1: Inconsistent experimental setup. Some results were omitted due to space limit rather than cherry-picking. We agree with the reviewer that a consistent presentation of the results is important, and further present these detailed results in Table 1 and Fig. (a) below. In Fig. (a), the improvements over vanilla fine-tuning are averaged across three datasets to keep the plots simple. These consistent results confirm that StochNorm works well for a variety of backbones and datasets with different sampling rates. We will further provide the training/test splits of all datasets.

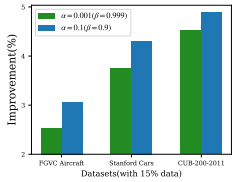
Table 1: Accuracy (%) of different methods on Chest X-ray, CUB-200-2011, and FGVC Aircraft datasets (backbone: ResNet-50).

Method	Chest X-ray		
	5%	10%	15%
vanilla	70.37	75.85	76.64
L^2-SP	70.02	72.73	75.71
DELTA	70.99	74.35	75.97
BSS	69.86	73.27	76.10
StochNorm	72.50	76.48	77.01

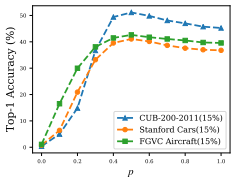
Method	CUB-200-2011				FGVC Aircraft			
	15%	30%	50%	100%	15%	30%	50%	100%
L^2-SP	45.08	57.78	69.47	78.44	39.27	57.12	67.46	80.98
L^2-SP+**StochNorm**	49.92	60.48	70.47	79.24	42.57	60.16	69.16	81.12
DELTA	46.83	60.37	71.38	78.63	42.16	58.60	68.51	80.44
DELTA+**StochNorm**	49.27	62.86	72.78	79.72	44.10	60.13	70.12	81.03
BSS	47.74	62.03	72.56	78.85	40.41	59.23	69.19	81.48
BSS+**StochNorm**	50.67	64.10	73.01	79.91	43.89	60.25	69.41	81.50



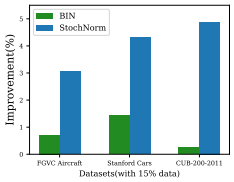
(a) Backbone



(b) Value of α



(c) Value of p



(d) Compare with BIN

Reviewer #2

Question 1: The moving statistics hyperparameter α (PyTorch's α = 1 − Reviewer's β). α is a hyperparameter of BatchNorm rather than of StochNorm. In **PyTorch**, $\alpha = 0.1$ by default. We are curious to the reviewer's question, and compare $\alpha = 0.1$ with $\alpha = 0.001$ in Fig. (b) above, showing the improvements over vanilla BN. Empirically we find that smaller α (larger β) leads to slower convergence and slightly worse results. A more comprehensive study on α will be included. We will clarify this detail as well as tone down the arguments related to Dropout.

Question 2: Table 4 with other datasets. Please see Table 1 above. StochNorm yields consistently better accuracies.

Reviewer #3

Question 1: How to avoid collapse when normalized by moving statistics. This is an **insight** found this paper: in fine-tuning, collapse will occur only if **all** features are normalized by moving statistics. In StochNorm, some randomly selected channels are normalized by moving statistics while others by mini-batch statistics. This successfully stabilizes fine-tuning without relying on Batch Renormalization.

Question 2: Ablation study of selection probability p. As stated in the paper (Line 223), $p = 0.5$ works well for most experiments. Fig. (c) above presents an ablation study of p. Here $p = 1$ is BatchNorm, while $p = 0$ is to normalize **all** features with moving statistics (leads to collapse as pointed out by the reviewer). The best p sits around 0.5.

Question 3: Over-claim the orthogonality. Orthogonality means that StochNorm is architecture-based while L2-SP, DELTA and BSS are regularizer-based. In many cases they are complementary and an integration leads to better results. As such improvements are general but not absolute, we will tone down this claim.

Question 4: Possible theoretical analysis. Theoretical analyses for normalization are generally not easy. The papers of BatchNorm, GroupNorm, InstanceNorm neither provide theoretical analyses. Thanks for pointing it as future work.

Reviewer #4

Question 1: Comparison with other normalization approaches. Fig. 3(b) in the paper compares StochNorm with BatchNorm (BN). There are no publicly available models pre-trained on ImageNet with instance normalization (IN) and weight normalization (WN). Only pre-trained models with BatchNorm are available. Batch-Instance Normalization (BIN) can be compared because it is based on BatchNorm. We present the absolute improvements over BN in Fig. (d) above, which shows that BIN is slightly better than BN but substantially inferior to StochNorm.

Question 2: Comparison with data augmentation. Data augmentations for all experiments (both StochNorm and other compared methods) are the same, as stated in supplementary material (Line 5). They are used by default in the computer vision community, and out of the scope of this paper, so we do not conduct ablations for data augmentations.