Supplementary Materials for
Positive-Congruent Training: Towards Regression-Free Model Updates

Sijie Yan[*] Yuanjun Xiong Kaustav Kundu Shuo Yang Siqi Deng
Meng Wang Wei Xia Stefano Soatto

AWS/Amazon AI

yysijie@gmail.com, {yuanjx, kaustavk, shuoy, siqdeng, mengw, wxia, soattos}@amazon.com

Approach	Error Rate (%)		NFR	Rel. NFR	#params
	ϕ^{old}	ϕ^{new}	(%)	(%)	
No Treatment	35.34	23.81	3.56	23.14	25M
Naive	35.69	23.30	3.12	20.82	25M
FD-KD	35.69	28.56	2.34	12.74	25M
FD-LM	35.69	27.90	2.44	13.58	25M
Ensemble	31.94	21.98	2.71	18.13	100M

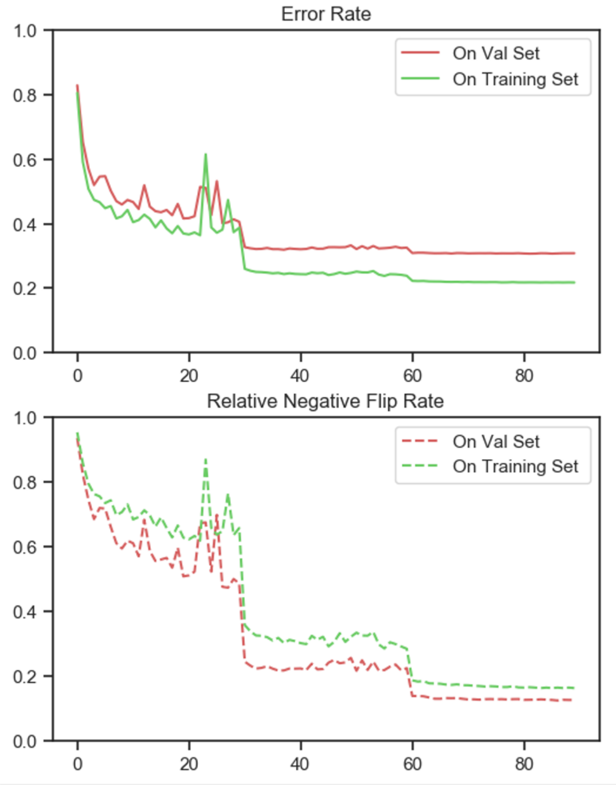
Table 1: Change in architecture + increase in #samples

1. Two Changes in Model Updates

Model updates can have two or more changes together. Here we present preliminary results on model updates with both (a) change in model architecture, and (b) increase in training samples. We train a Resnet-18 [1] model on 50% of the classes of the ILSVRC12 training dataset [2] for the old model. For the new model, we train a Resnet-50 [1] on the entire training set. The results are shown in Table 1. Though the accuracy of the models improve by ∼12%, 3.56% of the samples undergo negative flips. The different approaches of PC training behave consistently with the pattern we had observed for the different settings in the main paper.

2. The Special Case of Fine-Tuning

In model updates only involving data changes but not any architecture change, there is a special case that we can build the new model by finetuning the old model on the new training data. We analyze this special case and compare our PC training approaches in it. We use the same setting as in Sec. 5.3 of the main paper. The results of both scenarios of data changes are summarized in Table 2. The odd rows denote the normal cases where train the new model from scratch. The even rows (in gray) denotes the special cases of initializing the new model with the weights of the old model.


Error Rate

Relative Negative Flip Rate

Figure 1: Relative NFRs and error rates of the new model on both the training and the validataion sets. The new model is trained with focal distillation. We plot the error rates and NFRs after every training epoch to visualize the evolution of the values during training.

We observe even though the new models started with the weights of the old models, they still suffers from negative flips. Applying our PC training approaches help reduce NFR in these special cases similar to that for the normal new models.

[*]Currently at The Chinese University of Hong Kong. Work conducted while at AWS.

Approach	**Fine-tuning**	**Increase in #samples**				**Increase in classes**				**#params**
		Error Rate (%)		NFR	Rel. NFR	Error Rate (%)		NFR	Rel. NFR	
		ϕ^{old}	ϕ^{new}	(%)	(%)	ϕ^{old}	ϕ^{new}	(%)	(%)	
No Treatment	✗	28.58	23.65	3.98	19.47	19.30	23.65	8.05	41.90	25M
	✓	28.58	24.92	3.28	18.42	19.30	25.31	7.96	38.97	25M
Naive	✗	28.45	24.46	3.27	18.68	19.28	23.74	7.70	40.20	25M
	✓	28.45	24.50	2.78	15.85	19.28	23.69	7.14	37.33	25M
FD-KD	✗	28.45	25.20	2.89	15.84	19.28	24.14	7.07	36.29	25M
	✓	28.45	25.27	2.42	13.38	19.28	24.36	6.94	35.29	25M
FD-LM	✗	28.45	24.77	2.85	16.09	19.28	25.11	7.37	36.35	25M
	✓	28.45	24.91	2.39	13.41	19.28	25.05	7.17	35.46	25M
Ensemble	✗	26.39	22.09	2.75	16.88	17.53	21.98	6.72	37.06	100M
	✓	26.39	22.81	2.18	12.98	17.53	23.20	6.70	35.02	100M

Table 2: The special case of finetuning the old model to obtain the new model when there is only data change. This setting does not apply to model updates where the model architecture is changed.

3. Training Set Negative Flips

By definition, we assess the sample-wise regression problem on the testing samples. It is worth to note that there are also negative flips in the training samples. In Fig. 1 we present the evolution of NFRs and error rates during training of the new model on both the training and the validation sets of ILSVRC12 [2]. The new model is PC trained with focal distillation. Interestingly, the NFR on the training is even higher than that on the validation set.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 1, 2