Big Transfer (BiT): General Visual Representation Learning

Kolesnikov et al.
Google Research, Brain Team

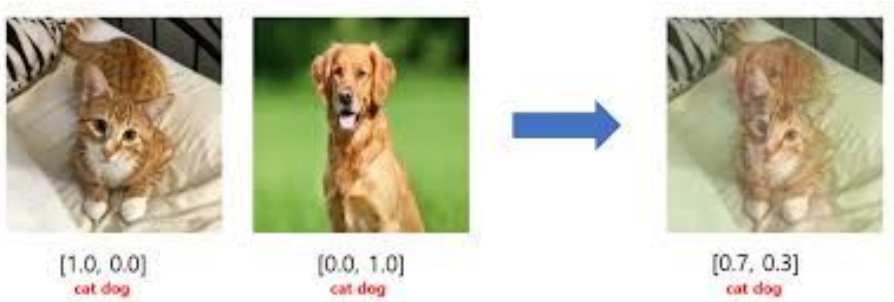
Sungman Cho.

Bit-Hyper Rule

- When training large models with small per-device batches, **BN performs poorly or incurs inter-device synchronization cost.**

- **GN + WS has been shown to improve performance on small-batch training** for ImageNet and COCO.

- **MixUp is not useful for pre-training** BiT. However it is sometimes **useful for transfer.**



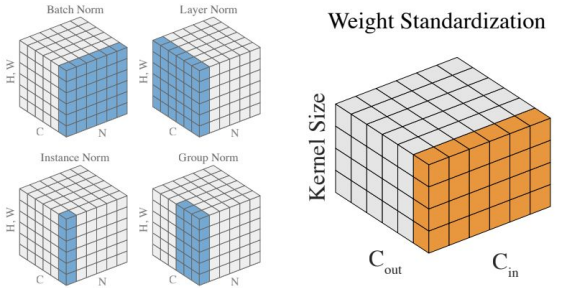
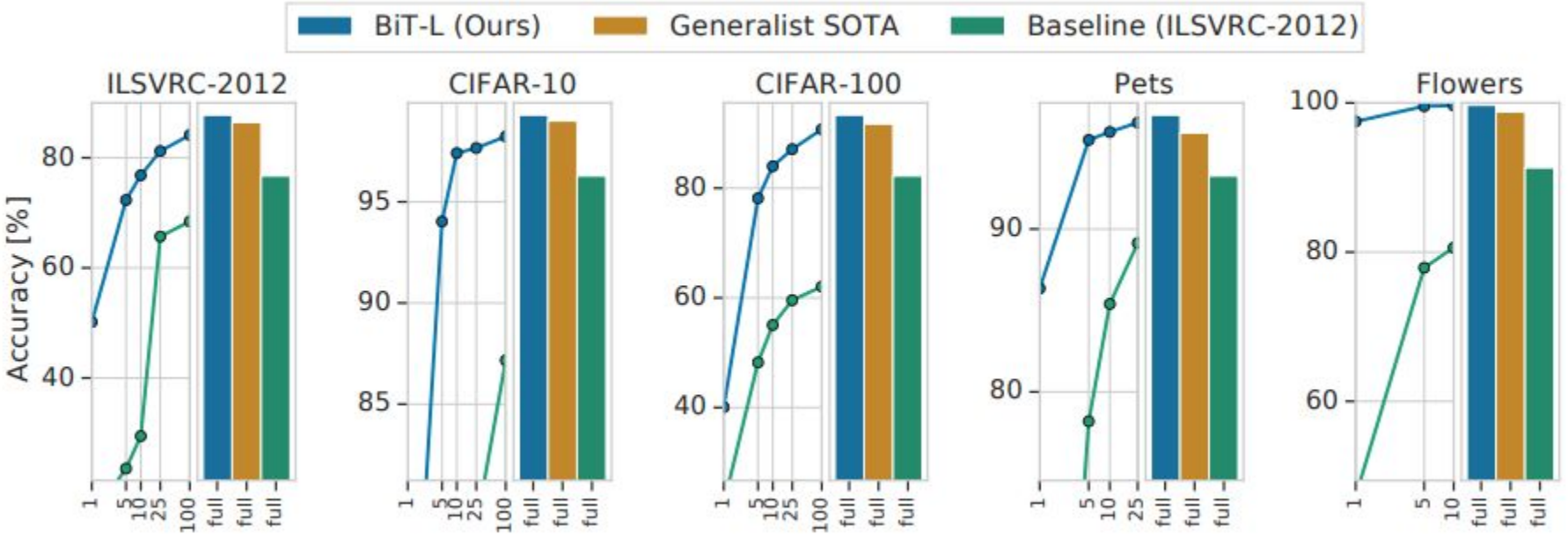


Figure 2. Comparing normalization methods on activations (blue) and Weight Standardization (orange).

Experiments



Experiments

- Top-1 Acc.

	BiT-L	Generalist SOTA	Specialist SOTA
ILSVRC-2012	**87.54 ± 0.02**	86.4 [57]	88.4 [61]*
CIFAR-10	**99.37 ± 0.06**	99.0 [19]	-
CIFAR-100	**93.51 ± 0.08**	91.7 [55]	-
Pets	**96.62 ± 0.23**	95.9 [19]	97.1 [38]
Flowers	**99.63 ± 0.03**	98.8 [55]	97.7 [38]
VTAB (19 tasks)	**76.29 ± 1.70**	70.5 [58]	-

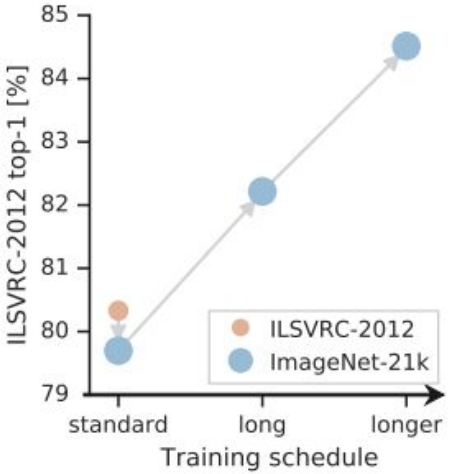
Experiments

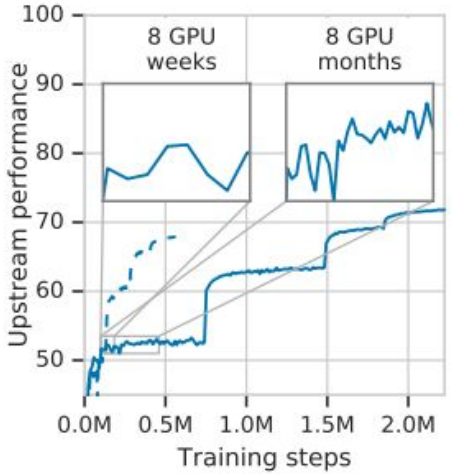
- Pretraining on the ImageNet-21k over the “standard” ILSVRC-2012.
- Models: ResNet152x4.

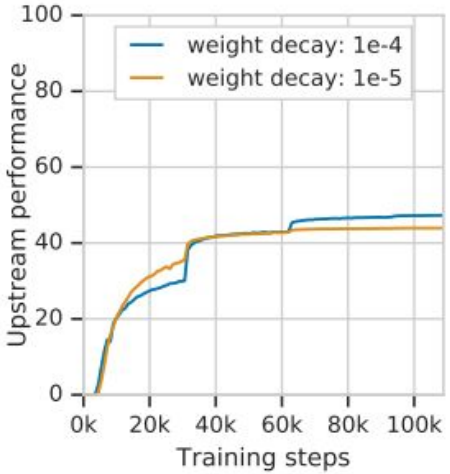
	ILSVRC-2012	CIFAR-10	CIFAR-100	Pets	Flowers	VTAB-1k (19 tasks)
BiT-S (ILSVRC-2012)	81.30	97.51	86.21	93.97	89.89	66.87
BiT-M (ImageNet-21k)	85.39	98.91	92.17	94.46	99.30	70.64
Improvement	+4.09	+1.40	+5.96	+0.49	+9.41	+3.77

Experiments

- Only when we **train longer**, do we see the **benefits of training on the larger datasets.**
- If decays the learning rate too early (dashed), final performance is significantly worse.
- **Higher weight decay** converges more **slow**ly, but **results in a better** final model.

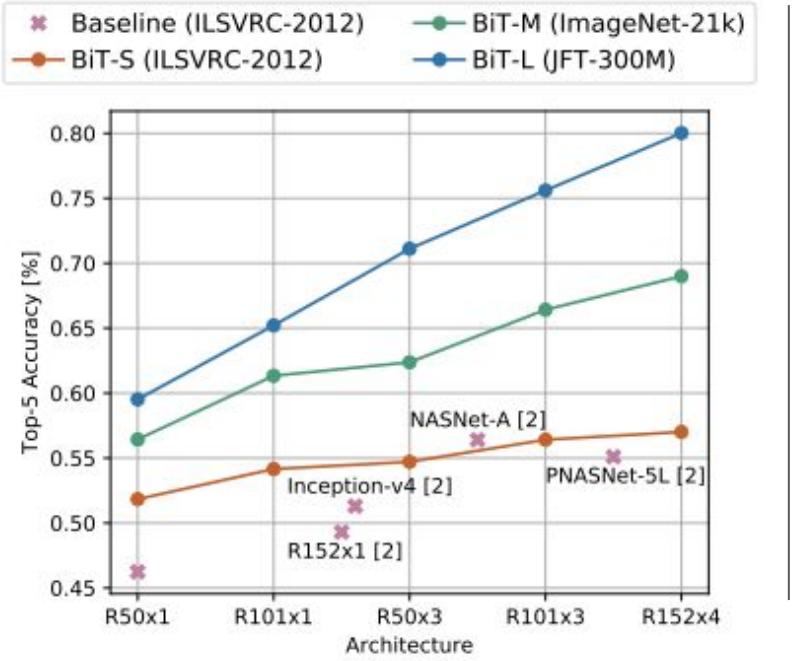






Experiments

- Recognition

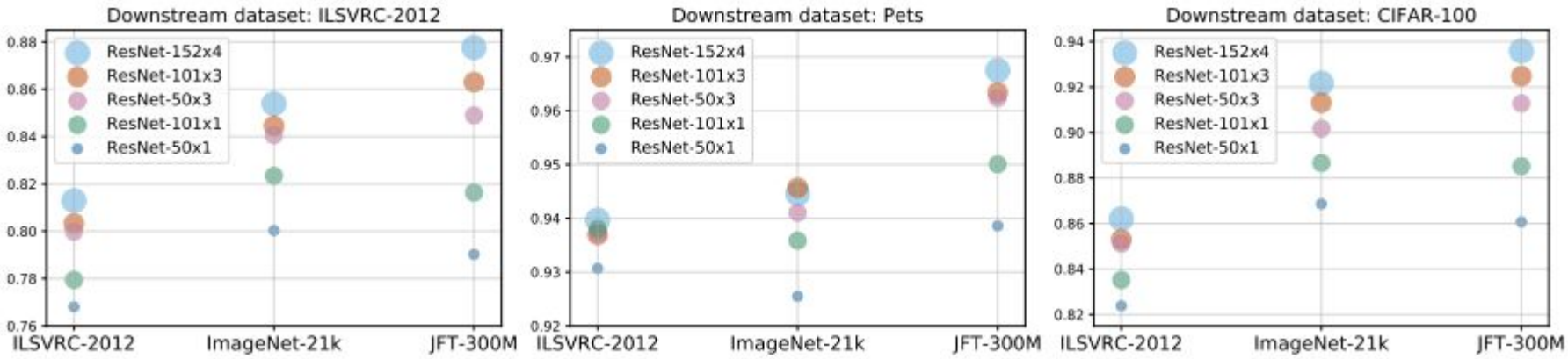


- Detection

Model	Upstream data	AP
RetinaNet [33]	ILSVRC-2012	40.8
RetinaNet (BiT-S)	ILSVRC-2012	41.7
RetinaNet (BiT-M)	ImageNet-21k	43.2
RetinaNet (BiT-L)	JFT-300M	**43.8**

Experiments

- Scaling Models and Datasets



Experiments

- Top-1 accuracy of ResNet-50 on ILSVRC-2012 with batch-size of 4096

Scratch

	Plain Conv	Weight Std.
Batch Norm.	75.6	75.8
Group Norm.	70.2	**76.0**

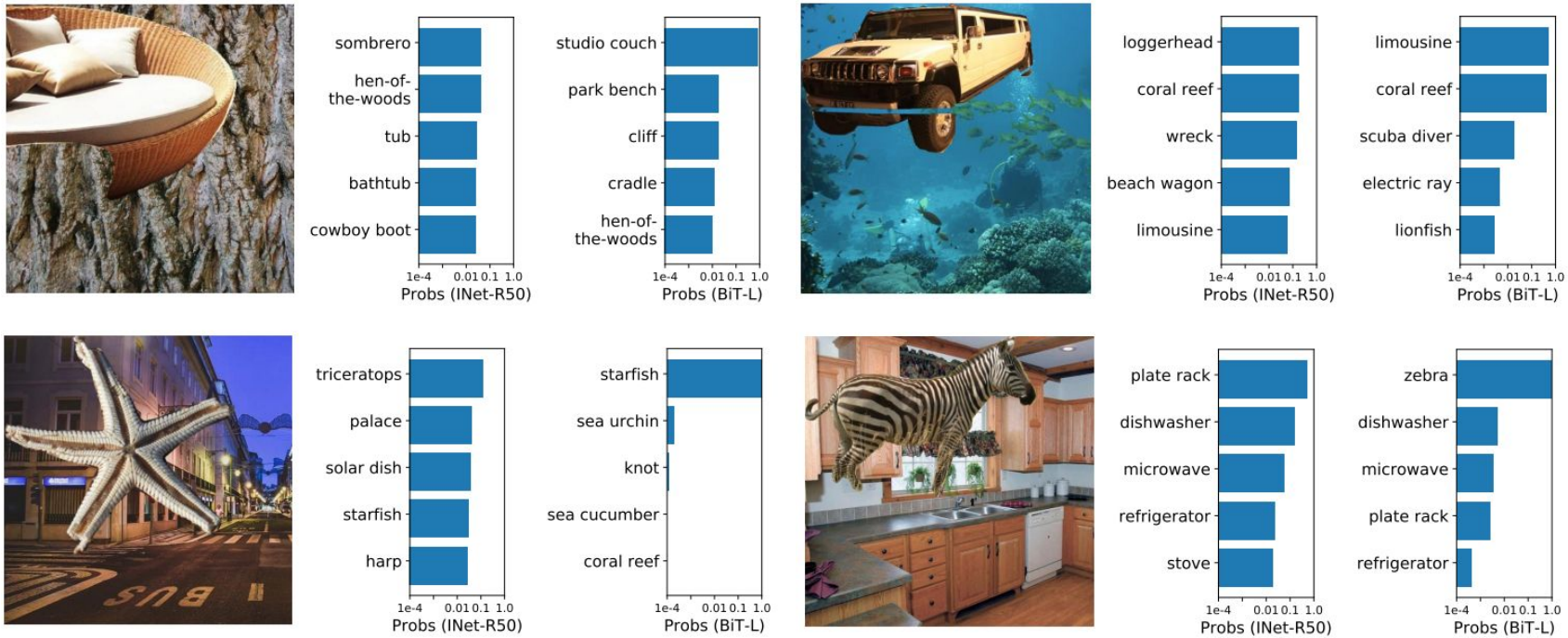
Fine-tuned to the 19 VTAB-1k

	Plain Conv	Weight Std.
Batch Norm.	67.72	66.78
Group Norm.	68.77	**70.39**

BN performs worse when the number of images on each accelerator is too low

Experiments

- Top 5 prediction produced by an ILSVRC-2012 model.



Experiments

- Performance (when dup.)

	From JFT			From ImageNet21k			From ILSVRC-2012		
	Full	Dedup	Dups	Full	Dedup	Dups	Full	Dedup	Dups
ILSVRC-2012	87.8	87.9	6470	84.5	85.3	3834	80.3	81.3	879
CIFAR-10	99.4	99.3	435	98.5	98.4	687	97.2	97.2	82
CIFAR-100	93.6	93.4	491	91.2	90.7	890	85.3	85.2	136
Pets	96.8	96.4	600	94.6	94.5	80	93.7	93.6	58
Flowers	99.7	99.7	412	99.5	99.5	335	91.0	91.0	0

Experiments

- Detected duplicates between the ILSVRC-2012 train/test.