Big Transfer (BiT): General Visual Representation Learning

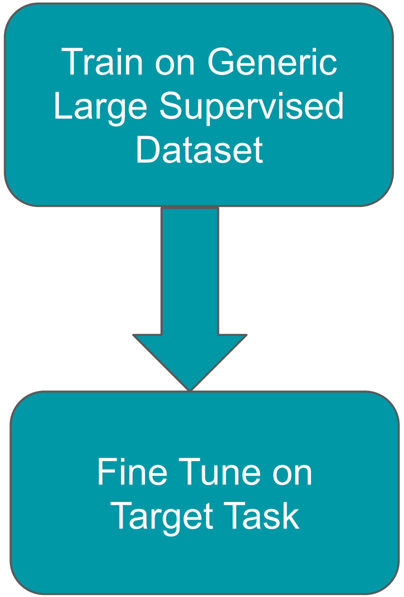
A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, N. Houlsby

Google Research, Brain Team

Outline

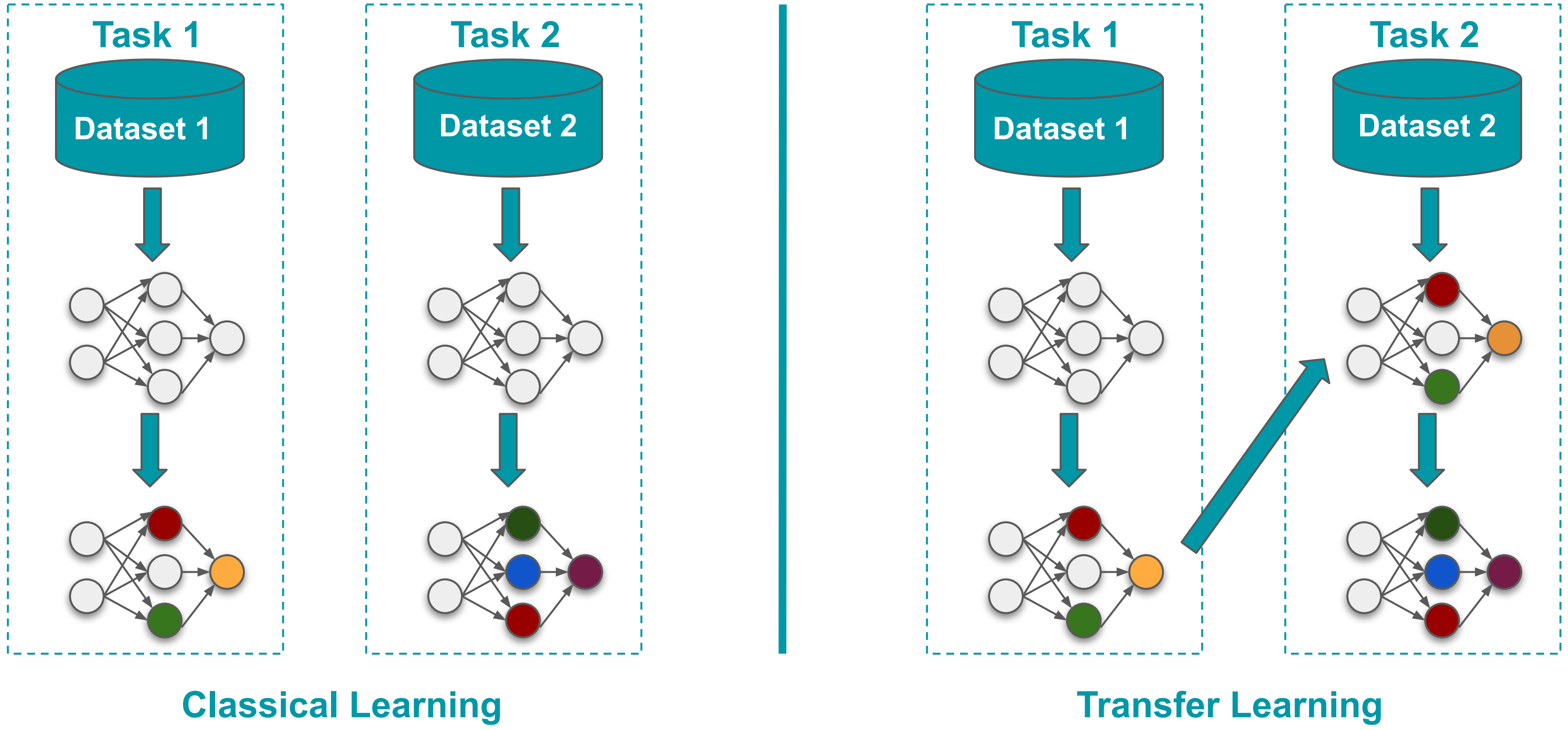
- ❑ Paper Summary
- ❑ Transfer Learning
- ❑ Big Transfer
  - ❑ UpStream Training
  - ❑ Downstream Training
- ❑ Experiments and Results
- ❑ Discussion
- ❑ Quiz questions

Paper Summary



- Scale Up Pre-Training
  - Train ResNet152x4 on JFT 300M dataset.
  - Shows how to train models at such scale.
- Fine Tune this model to different tasks (20)
  - Cheap fine-tuning
  - Only few hyper-params need to be tuned.
- Fine Tuned models perform very well.

Transfer Learning

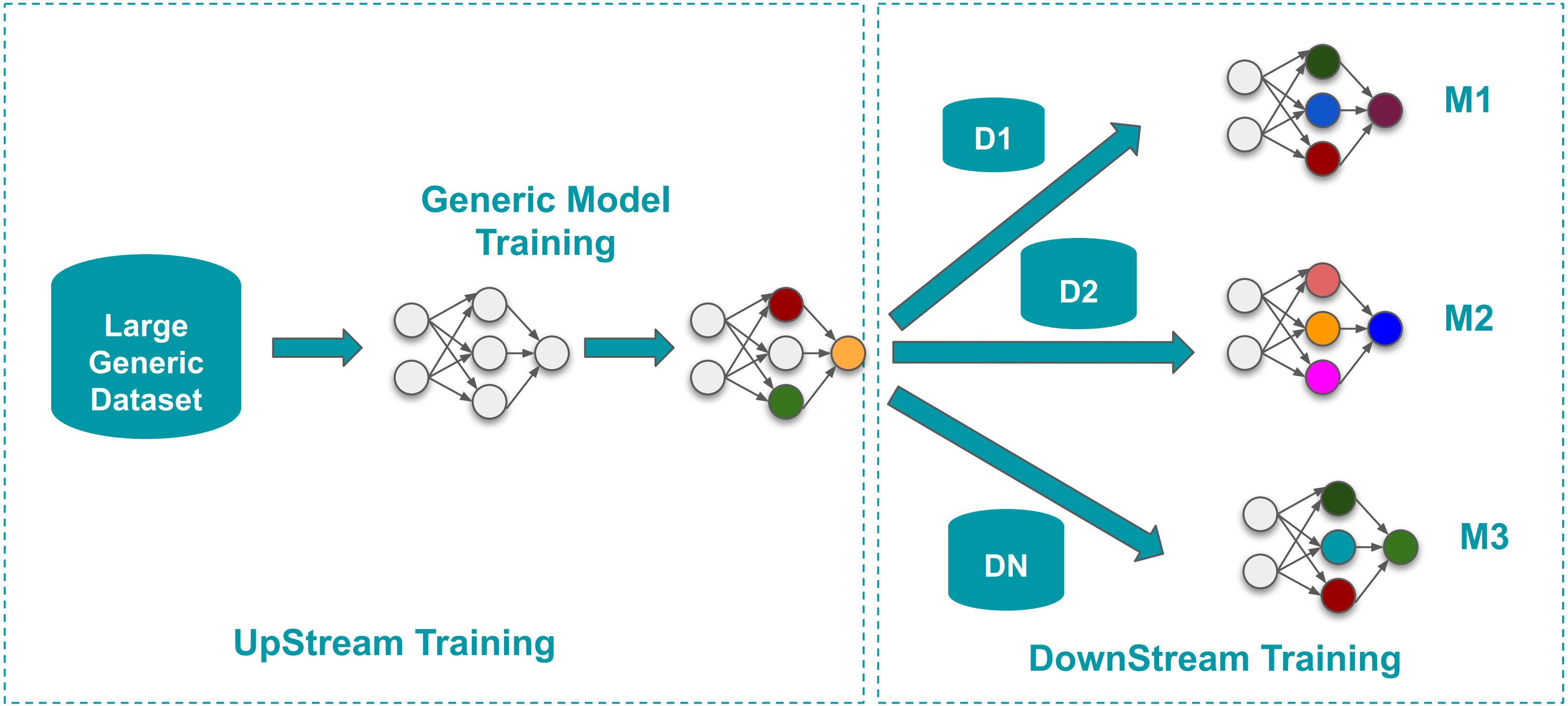


Why Transfer Learning?

- Scarcity of Labelled Data
- Training Models for every task is expensive and time consuming
- There is redundant work in training

- Train Just one model.
- Fine tuning it to other tasks take less data and less compute.
- Promotes Reuse.

Big Transfer (BiT)



BiT Components (Ingredients)

UpStream Components

- ❑ Large Scale Dataset and Model
- ❑ Group Normalization
- ❑ Weight Standardization

DownStream Components

- ❑ Task Specific Dataset
- ❑ Fine-Tuning Protocol
- ❑ Bit-HyperRule

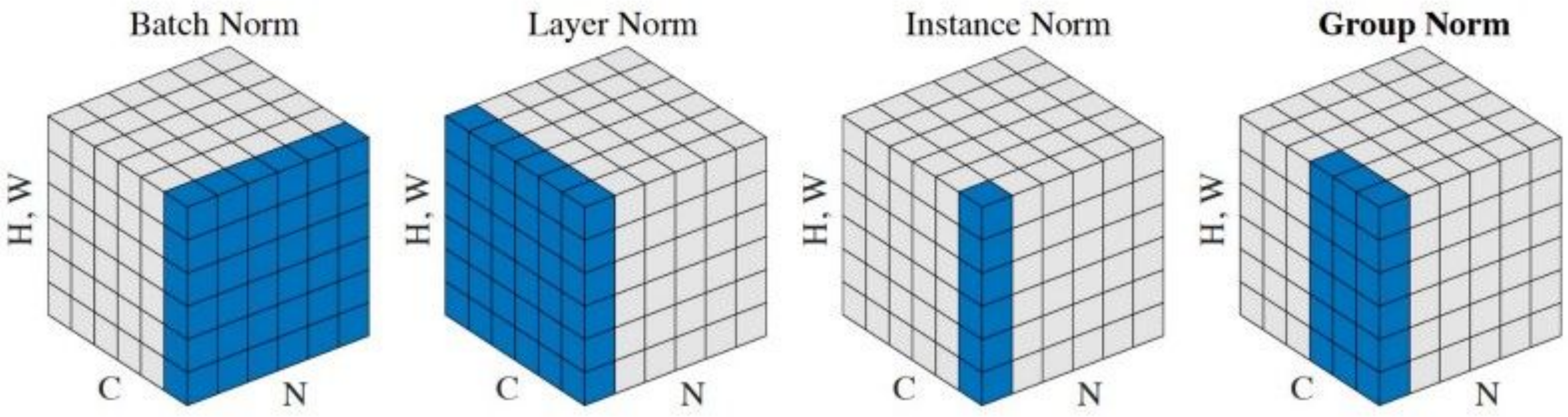
Upstream Training

Data for Upstream Training

Model	Data Set	Remarks
BiT-S	ILSVRC-2012 variant of ImageNet	1.28M images, 1000 classes, 1 label/image
BiT-M	ImageNet-21k	14.2M images, 21k classes
BiT-L	JFT-300M	300M images, 1.26 labels/image, 18291 classes, 20% noisy labels due to automatic annotations

Normalization

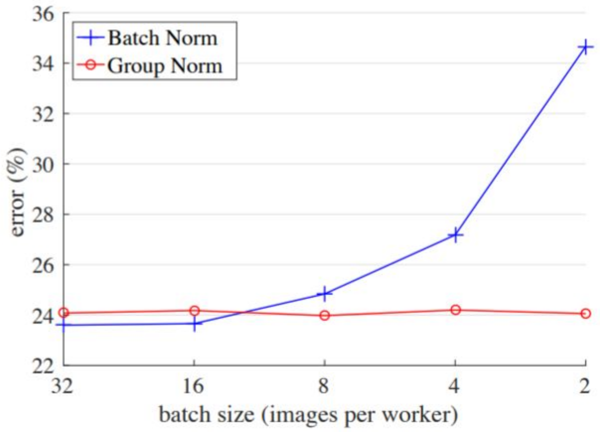
- Normalize activations along subset of (N,C,H,W) dimensions.
- Faster and stable training of NNs
- Makes Loss function smooth and hence optimization is easier.



https://arxiv.org/pdf/1803.08494.pdf

Group Normalization

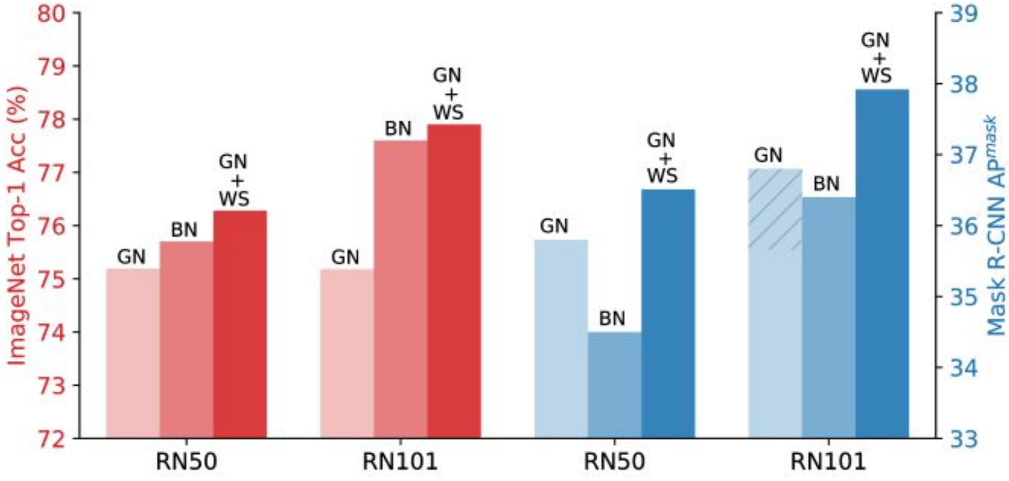
- Normalize over groups of channels. Not all channels are equally important.
- Layer Normalization and Instance Normalization are special cases of GN.
- More effective then BN when batch size is very small. But BN is better with bigger batch sizes



https://arxiv.org/pdf/1803.08494.pdf

Weight Standardization

- ❑ Normalizes weights instead of activations.
- ❑ Helps in smoothing the loss landscape.
- ❑ Works well in conjunction with GN in low batch size regime.



https://arxiv.org/pdf/1903.10520.pdf

Summary of Upstream Training

Model

- ResNet 152 x4
- Each hidden layer widened by x4
- 928 Million params
- Same model for all datasets

Data Parallel Training

- Global BS = 4096
- Train on TPUv3-512
- 8 img/chip
- Use GN + WS

Optimization

- SGD with Momentum (0.9), weight Decay(1e-4)
- LR=0.03 and reduce by factor of 10 after 10, 23,30, 37 epochs. (BiT-L)
- Train for 40 epochs
- Linear LR warmup for first 5K opt. Steps

DownStream Training

DownStream Components

Goal : Cheap fine-tuning

BiT-HyperRule

- Most Hyper-Params need not be changed.
- Depending on dataset size and image resolution set the following,
  - Training Schedule Length
  - Image Resolution
  - MixUp Regularization
- Small (~ 20K), Medium (~500K), Large(> 500K)

Data Processing

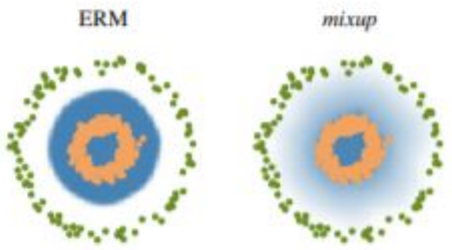
- Random Crops and Horizontal Flips (all tasks)
- Smaller than 96x96 => 160x160 => random crop 128x128
- Larger, => 448x448 => random crop 384x384

Optimization

- SGD with Momentum (0.9), weight Decay(1e-4)
- LR=**0.003** and reduce by factor of 10 in later epochs
- Epochs:
  - Small: 500
  - Medium: 10K
  - Large: 20K

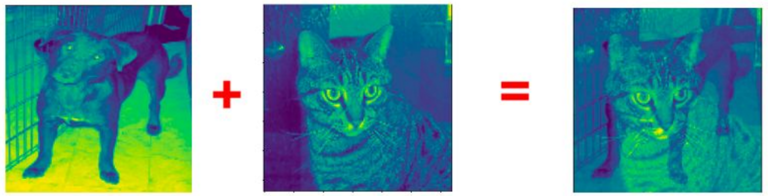
MixUp Regularization

Introduce new samples which are convex combination of existing samples.



2

1



- Improves Generalization
- Reduces memorization of corrupt labels.
- Increases Robustness to adversarial examples.
- Used mixup with alpha=0.1 for large and medium tasks.

1. https://towardsdatascience.com/2-reasons-to-use-mixup-when-training-yor-deep-learning-models-58728f15c559
2. https://arxiv.org/pdf/1710.09412.pdf

Experiments

Downstream Tasks

Benchmarks

- ILSVRC-2012
- CIFAR 10/100
- Oxford-IIIT Pet
- Oxford Flowers-102

Datasets differ in

- Total number of images
- Input resolution
- Nature of categories
  - ImageNet and CIFAR (general)
  - Pets and Flowers (fine-grained)

Results reporting

- BiT fine-tuned on official training split
- Report results on official test split if available else use validation split

Further assessment

- VTAB benchmark
- To assess generality of representations learned by BiT
- 19 tasks, 1000 training samples each
- Three groups of tasks - natural, special, structured

Hyperparameter Details

Upstream Pre-Training	Downstream Fine-Tuning
● ResNetv2 architecture, each hidden layer widened by factor of 4 (ResNet152x4) ● BN layers replaced by GN, WS in all conv layers ● SGD with momentum(0.9) ● Initial LR - 0.03 - decayed in all 3 models by factor of 10 in later epochs ● Batch size - 4096, Linear learning rate warmup for 5000 steps, weight decay of 0.0001	● BiT - HyperRule ● Resolution - < 96x96 160x160 - then random 128x128 crop, Larger images resize to 448x448 then 384x384 crop ● Schedule - - Small - <20k ex, tune 500 steps, - Medium - <500k ex, tune 10k steps - Large - tune for 20k steps ● MixUp - α = 0.1, for medium and large tasks

Results

Top-1 accuracy for BiT-L

	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]	-

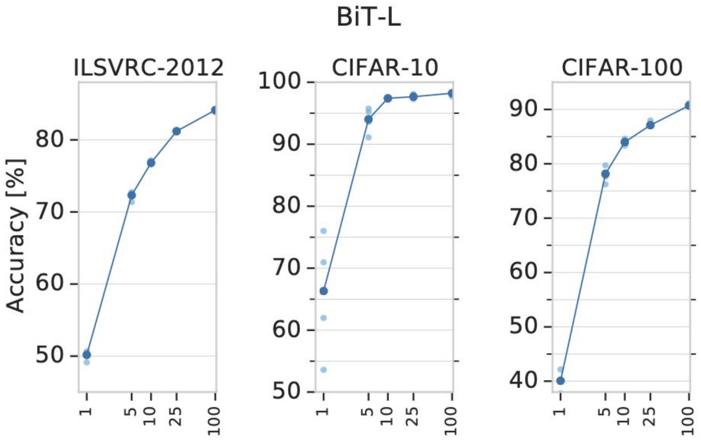
The entries show median ± standard deviation across 3 fine-tuning runs.

Accuracy improvement with ImageNet-21k

	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

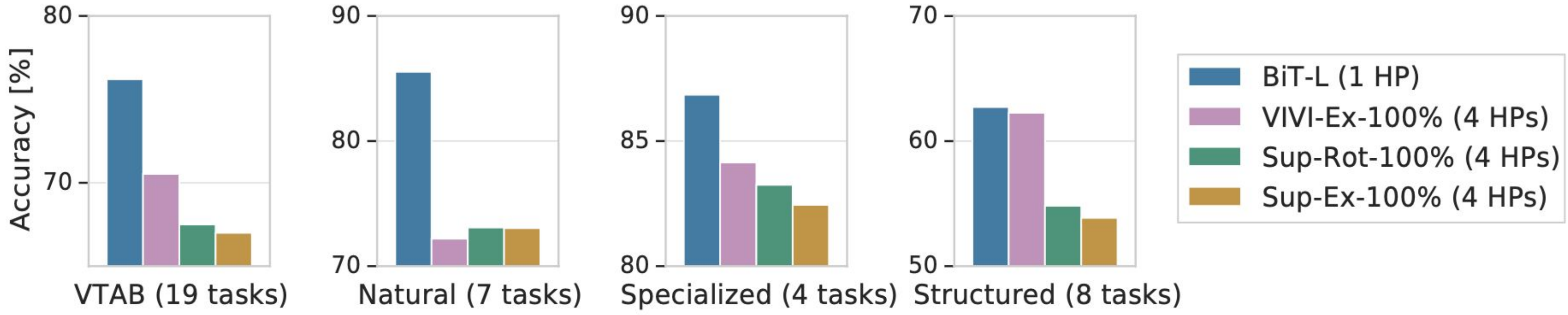
Top-1 accuracy is reported above. Both models are ResNet152x4

Few-Shot Learning



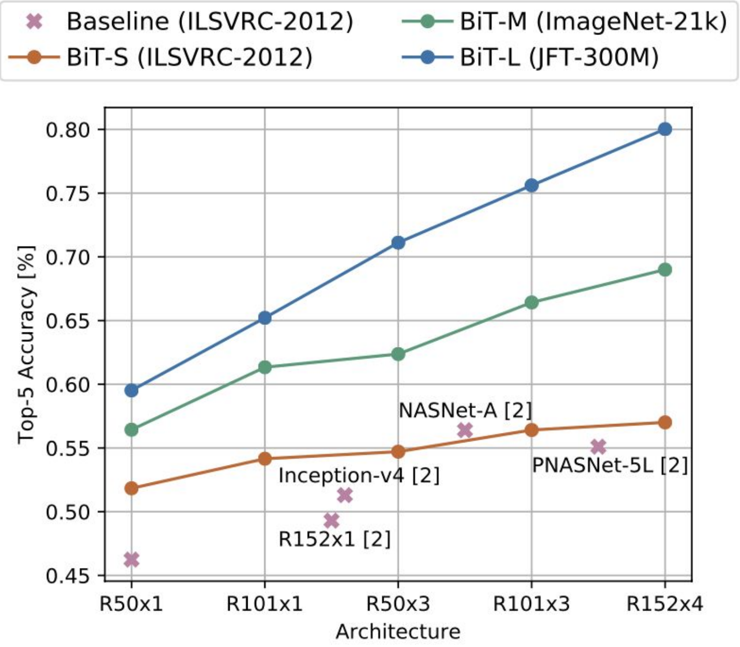
ILSVRC-2012 - Top-1 accuracy of 72% with 5 samples/class, 84.1% with 100 samples/class
CIFAR-100 - Top-1 accuracy of 82.6% with just 10 samples per class.

Results on VTAB



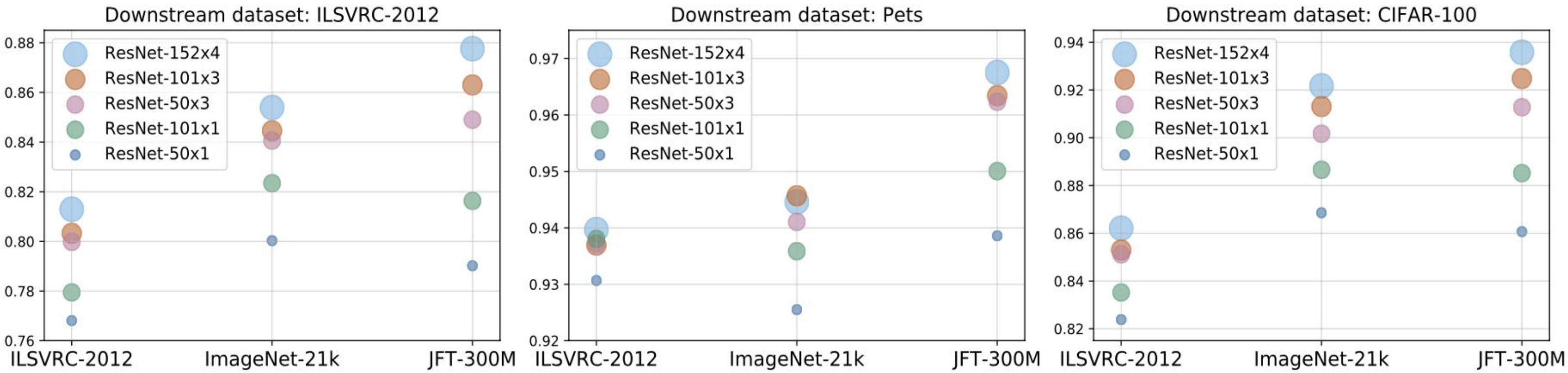
VTAB (19 tasks) with 1000 examples/task, and the current SOTA.

ObjectNet & Object Detection

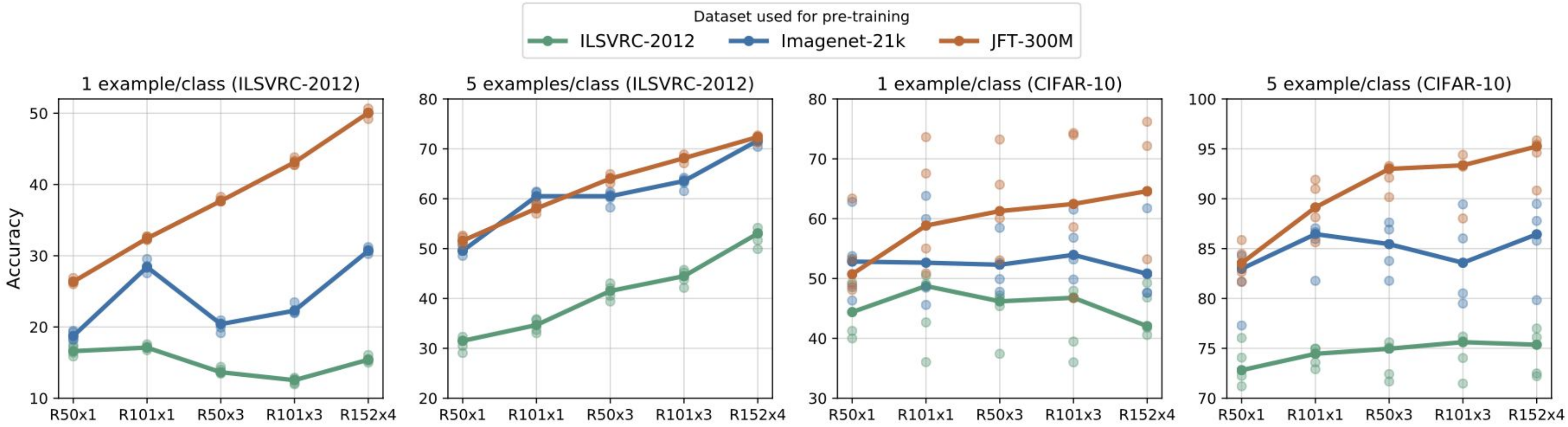


Model	Upstream data	AP
RetinaNet [33]	ILSVRC-2012	40.8
RetinaNet (BiT-S)	ILSVRC-2012	41.7
RetinaNet (BiT-M)	ImageNet-21k	43.2
RetinaNet (BiT-L)	JFT-300M	**43.8**

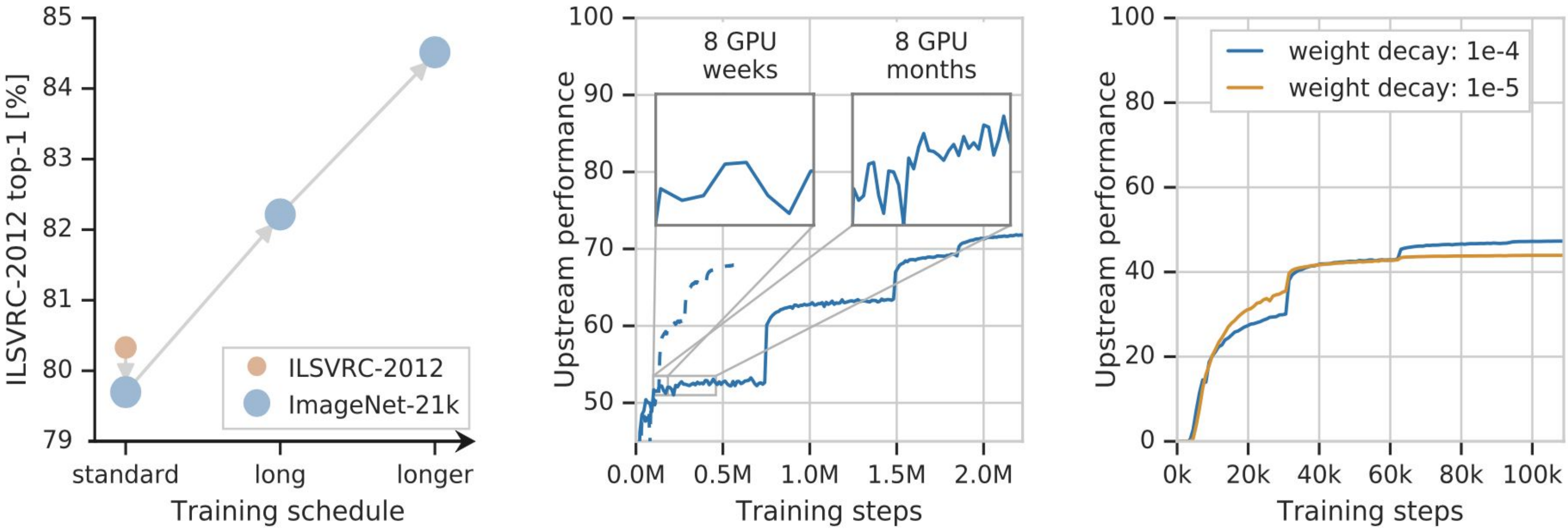
Scaling Models and Datasets



Scaling Models and Datasets



Optimization for large datasets



Large Batches, Group Normalization, Weight Standardization

Table 4: Top-1 accuracy of ResNet-50 trained from scratch on ILSVRC-2012 with a batch-size of 4096.

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

Table 5: Transfer performance of the corresponding models from Table 4 fine-tuned to the 19 VTAB-1k tasks.

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

Criticism and Future work

- Upstream Training is expensive, requires lot of resources (GPU etc.)
- These models may be poisonous or may contain backdoors ?
-

Thank you!

Discussion

Quiz

Question 1

1. The authors find Batch Normalization to be detrimental for Big Transfer. Which other techniques are suggested instead for upstream pre-training?
    a. Group Normalization
    b. Weight Standardization
    c. Dropout
    d. MixUp regularization

Answers :

a. Group Normalization

b. Weight Standardization

Question 2

2. Which of the following statements are true?
   a. BiT uses extra unlabelled in-domain data.
   b. Lower weight decay results in a highly performant final model.
   c. BiT has 928 million parameters.
   d. Decaying learning rate too early leads to sub-optimal model.

Answers :

c. BiT has 928 million parameters.

d. Decaying learning rate too early leads to sub-optimal model.

Question 3

3.

Statement I : The authors perform random horizontal flipping or cropping of training images during fine tuning, irrespective of the type of downstream task.

Statement II : For fine-tuning BiT-L needs more samples per class.

a. Statement I is false, Statement II is true
b. Statement I is true, Statement II is false
c. Both Statement I and II are true
d. Both Statement I and II are false

Answers :

a. Both Statement I and II are false