UNI FREIBURG

Big Transfer (BiT): General Visual Representation Learning

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver,
Jessica Yung, Sylvain Gelly and Neil Houlsby

What is transfer learning?

(https://www.fox.com/the-simpsons/)

(https://www.fox.com/the-sim

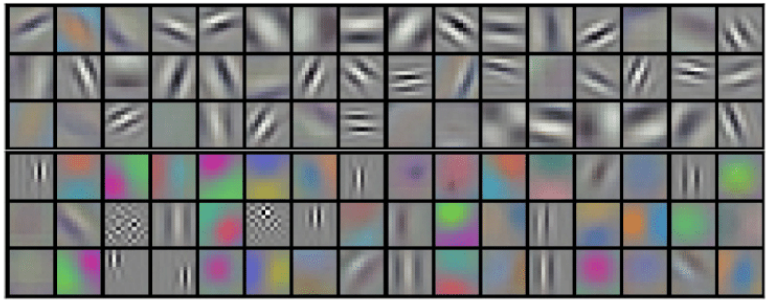
base task

target task

- Mastering the base task, makes learning the target task easier
- Basic representations learned from the base task can be used on the target task

Transfer for representation learning

- Task ~ Dataset
- Most often:
  - Base dataset: large
  - Target dataset: small
- Allows to target small datasets without over-fitting
- But also increases generalization on large datasets
- Transferring even from a distant task is often better than random initialization (Yosinki, et. al., 2014)



(Krizhevsky et. al., 2012)

Terminology warning

base vs. target

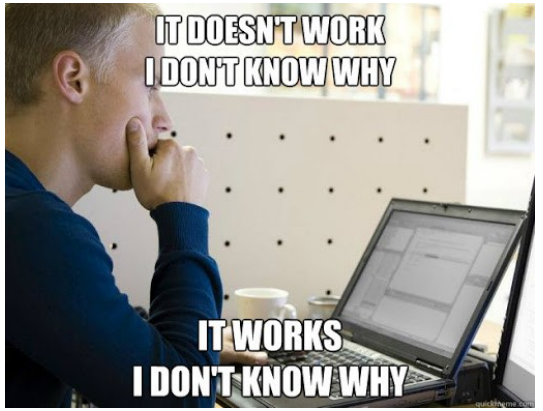
pre-training vs. transfer

upstream vs. downstream

Big Transfer (BiT)

What BiT is NOT about:

- Creating a new component or analysis
- Bringing new theoretical insights
- Getting to bottom of why something works or doesn't



BiT is a recipe for transfer learning where:

Base training

- Very expensive
- But done only once
- Creates a highly adaptable model

Target tuning

- Cheap
- No hyper-parameters to optimize
- Good performance even with a very limited amount of examples

The three BiT ingredients:

For the base training:

1 - Scale:

- Large datasets
- Large networks
- Large computational budgets

2 – Normalization:

- Group normalization
- Weight standardization

For the target tuning:

3 - Pre-optimized hyper-parameters:

- Image scaling
- Number of training steps
- Use of mix-up

Scale

Base datasets:

Model	Base dataset	Size
BiT-S	ILSVRC-2012	1.3 M
BiT-M	ImageNet-21K	14 M
BiT-L	JFT	300 M

Target datasets:

ILSVCR-2012
CIFAR-10
CIFAR-100
Oxford-IIIT Pet
Oxford Flowers-102
VTAB (Visual Task Adaptation Benchmark)

Network architecture:

- ResNet-152x4
- 928M parameters
- Same architecture for all models

`Terminology warning`

Models are not named after their sizes, but after the size of their base datasets.

What if a parameter weighted 1 mg...

1M pars = 1 kg

ResNet-50
(26 kg)

(http://jumbhoanimal.blogspot.com)

Inception-v4
(48 kg)

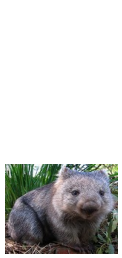
(https://www.thetimes.co.uk)

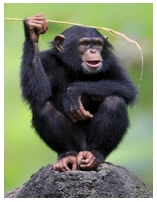
EfficientNet-L2
(480 kg)

(https://oceanwide-expeditions.com)

What if a parameter weighted 1 mg...

ResNet-50	26M
Inception-v4	48M
EfficientNet-L2	480M
BiT (ResNet-152x4)	928M





BiT-L
(928 kg)

(https://www.wowamazing.com)

The three BiT ingredients:

For the base training:

1 - Scale:

- Large datasets
- Large networks
- Large computational budgets

2 – Normalization:

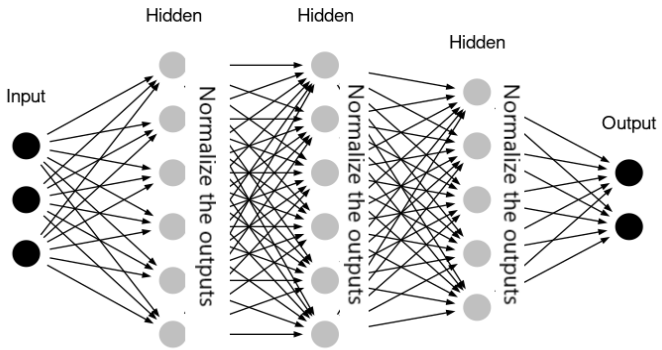
- Group normalization
- Weight standardization

For the target tuning:

3 - Pre-optimized hyper-parameters:

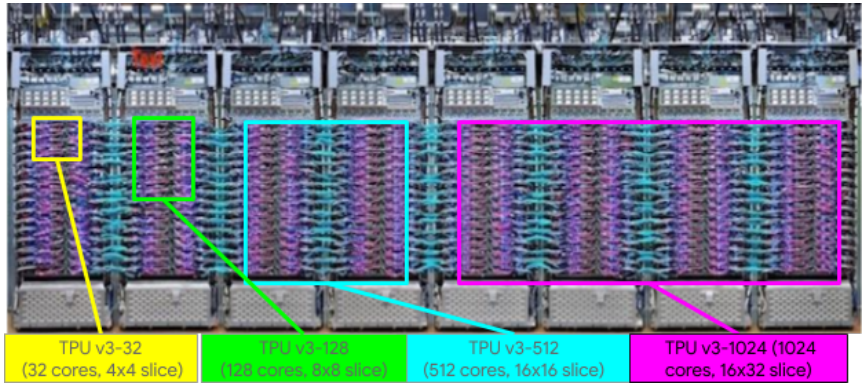
- Image scaling
- Number of training steps
- Use of mix-up

The problem with batch normalization



(deepai.org)

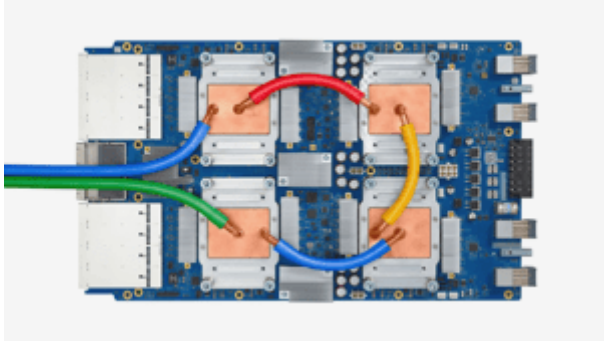
TPU pod **v3-512**:



(cloud.google.com)

BiT batching:

- Batch size: 4096
- Therefore, 8 images per TPU core

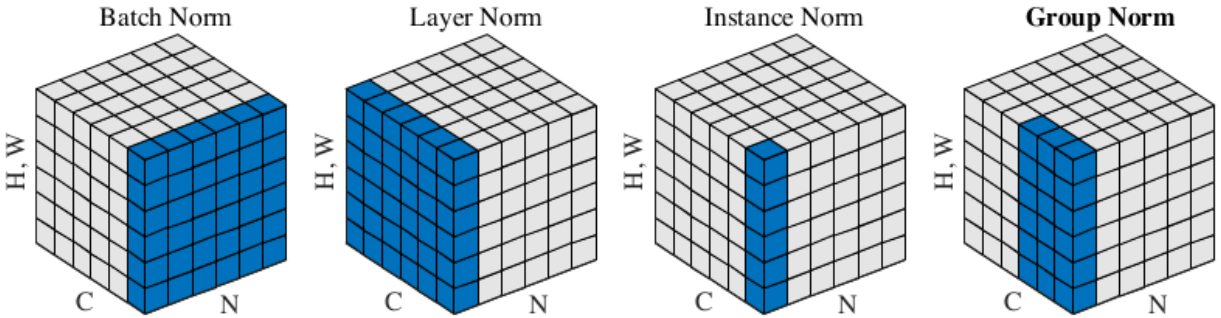


(cloud.google.com)

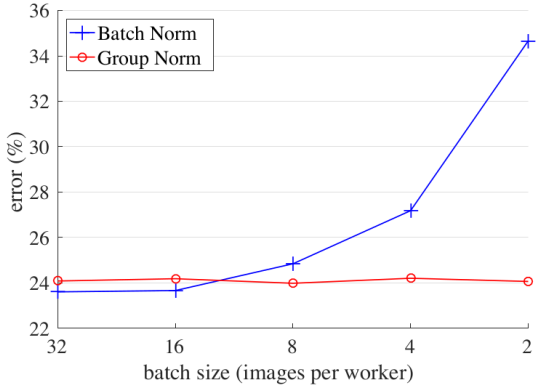
Options:

- Share batch statistics across TPUs → increased latency
- Use (less accurate) local batch statistics

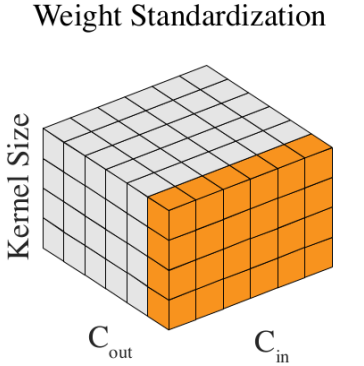
Group normalization & weight standardization



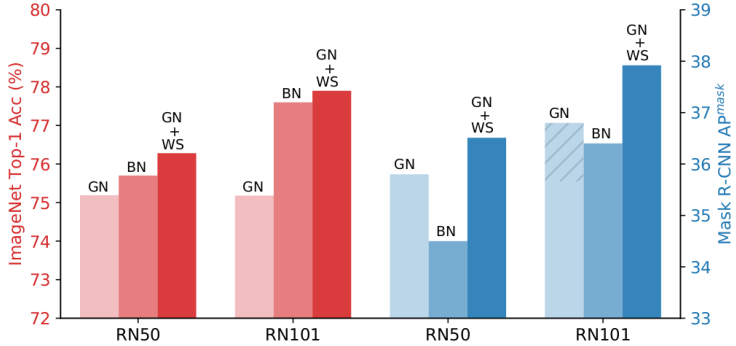
(Wu & He, 2018)



(Wu & He, 2018)



(Qiao et. al., 2019)



(Qiao et. al., 2019)

base training

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

target tuning

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

(Kolesnikov et. al., 2020)

The three BiT ingredients:

For the base training:

1 - Scale:

- Large datasets
- Large networks
- Large computational budgets

2 – Normalization:

- Group normalization
- Weight standardization

For the target tuning:

3 - Pre-optimized hyper-parameters:

- Image scaling
- Number of training steps
- Use of mix-up

Fixed hyper-parameters

SGD

- Momentum: 0.9
- Initial learning rate: 0.03
- LR decay by a factor of 10 at specific epochs
- Batch size: 4096 distributed equally among 512 workers

Weight decay

- wd: 0.0001
- Just during base task training

Architecture

- ResNet152x4
- Replace BN by GN
- Add WS to all convolutional layers

BiT-HyperRule

Image area	Resize to	Random crop
≤ 96 x 96 px	160x160 px	128x128 px
> 96 x 96 px	448x448 px	384x384 px

Number of examples	Training steps	Mix up alpha
≤ 20k	500	0.0
> 20k ≤ 500k	10k	0.1
> 500k	20k	0.1

Mix-up

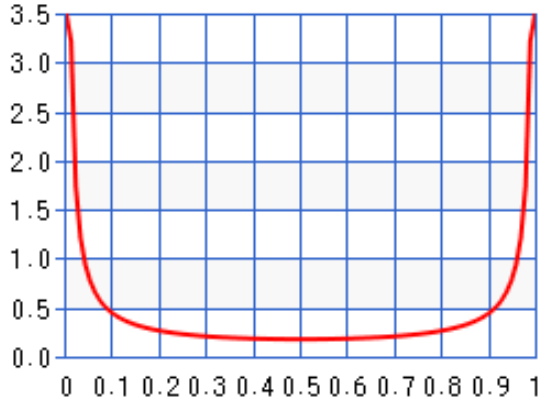
- A data augmentation technique
- From two data points, generates a third one

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j, \quad \text{where } x_i, x_j \text{ are raw input vectors}$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j, \quad \text{where } y_i, y_j \text{ are one-hot label encodings}$$

(Zhang et. al., 2018)

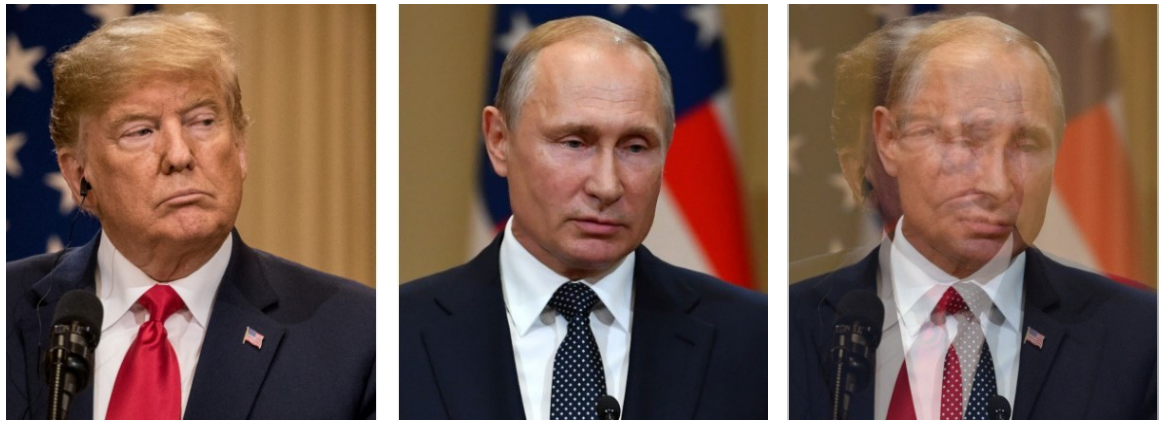
- λ is sampled from a beta distribution

Beta dist. (alpha=0.1) PDF



(keisan.casio.com)

x_i x_j $\tilde{x}$ (λ = 0.4)



(edition.cnn.com)

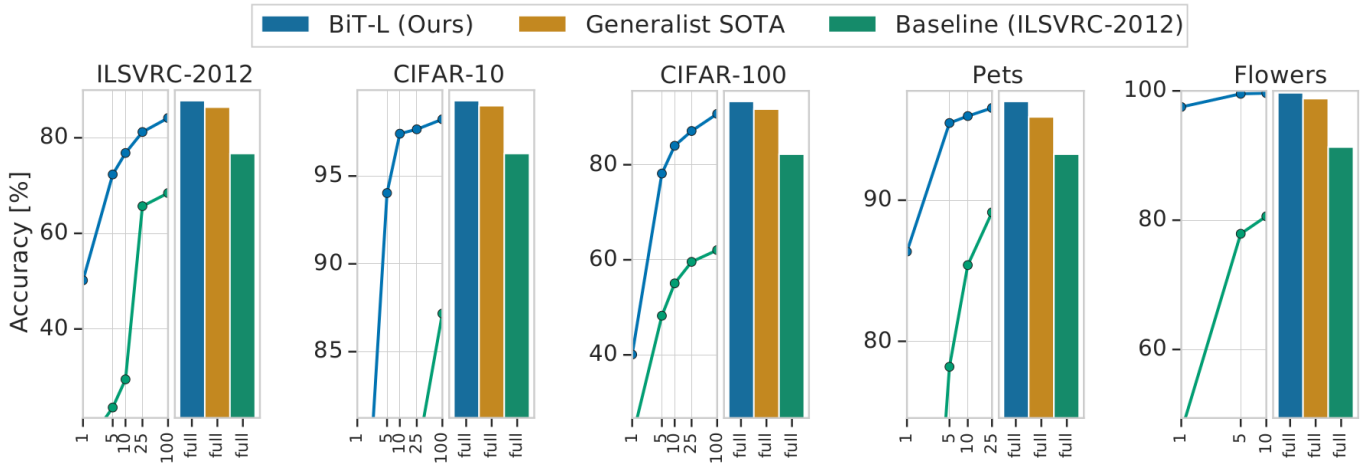
y_i = [1.0, 0.0] y_j = [0.0, 1.0] $\tilde{y}$ = [0.4, 0.6]

Results

Results: Classic benchmarks

	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]	-

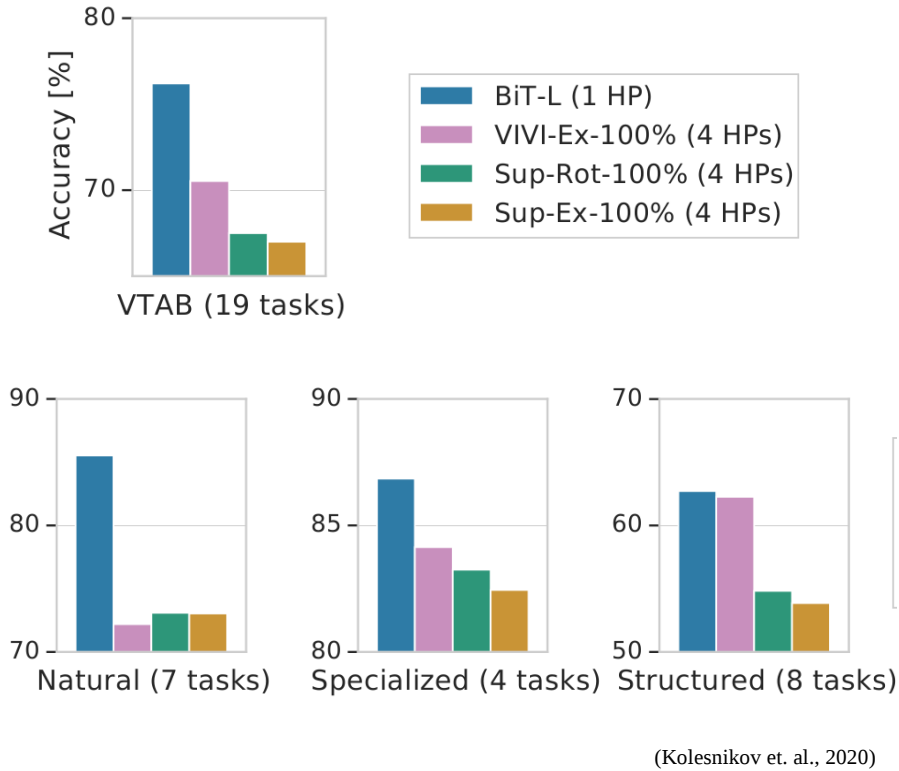
(Kolesnikov et. al., 2020)



(Kolesnikov et. al., 2020)

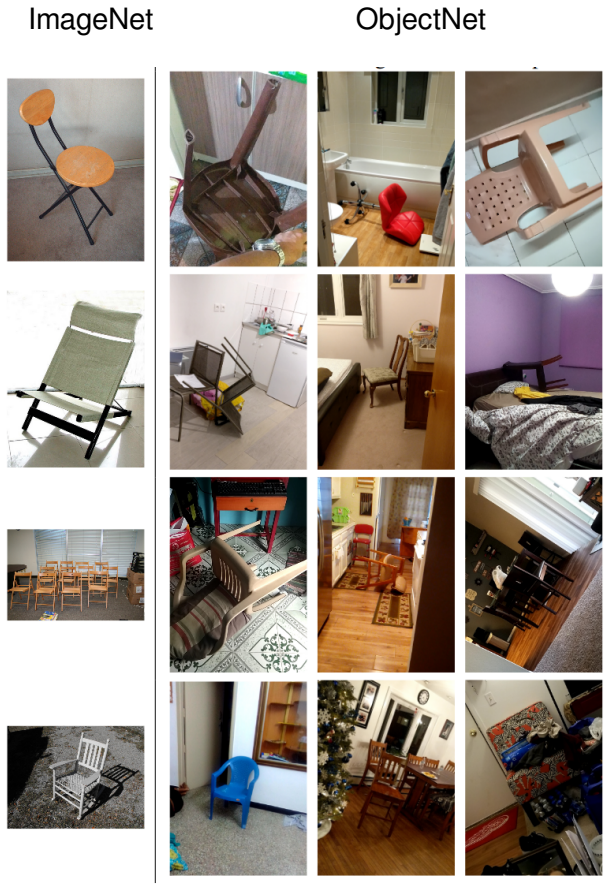
Results: VTAB

Natural	Specialized	Structured
Caltech101	Camelyon	Clevr-Count
CIFAR-100	EuroSAT	Clevr-Dist
DTD	Resisc45	DMLab
Flowers102	Retinopathy	dSpr-Loc
Pets		Dspr-Ori
Sun397		KITTI-Dist
SVHN		sNORB-Azim
		SNORB-Elev

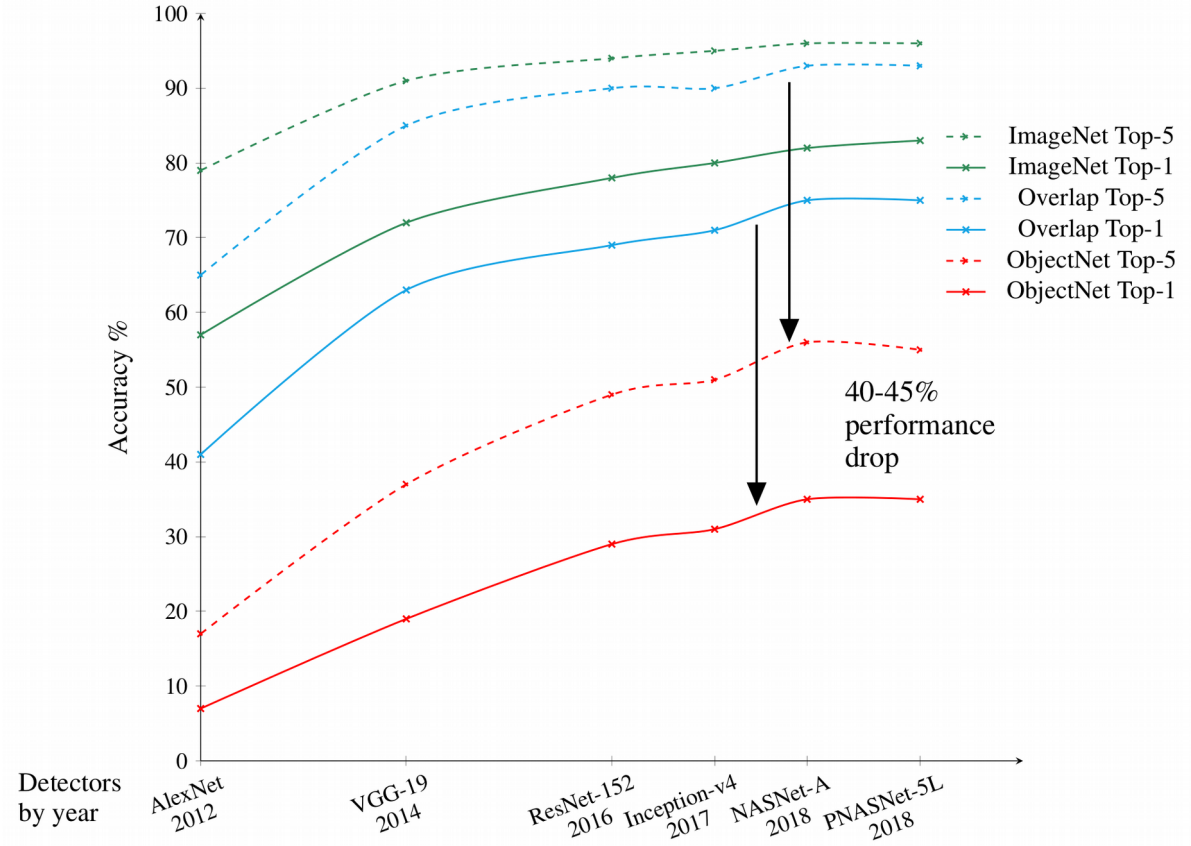


(Kolesnikov et. al., 2020)

Results: ObjectNET



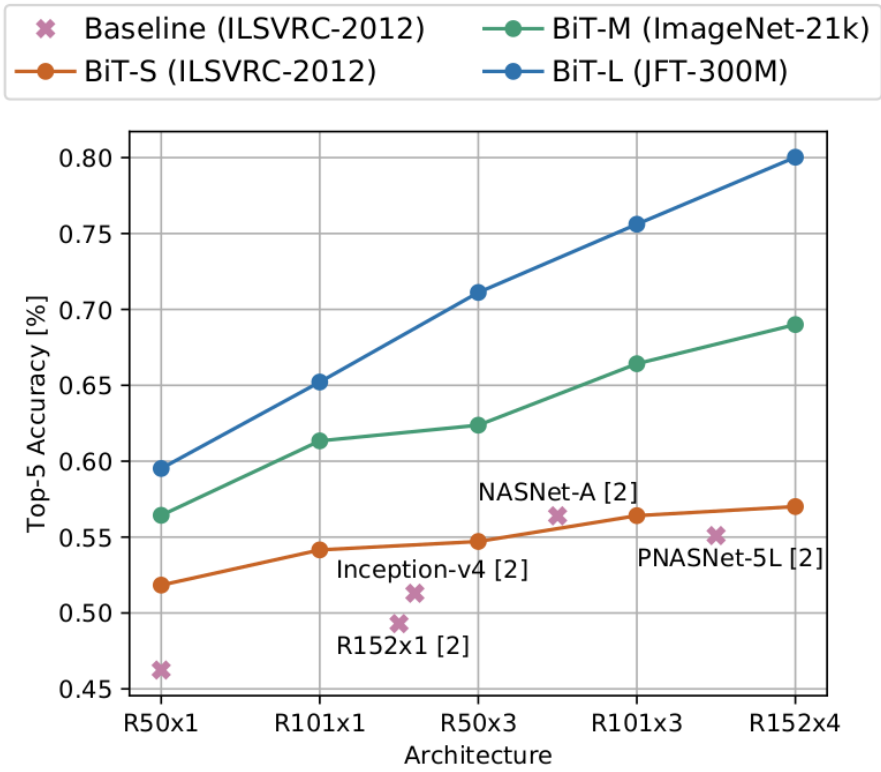
(Barbu & Mayo et. al, 2019)



(Barbu & Mayo et. al, 2019)

Results: ObjectNET

(Barbu & Mayo et. al., 2019)

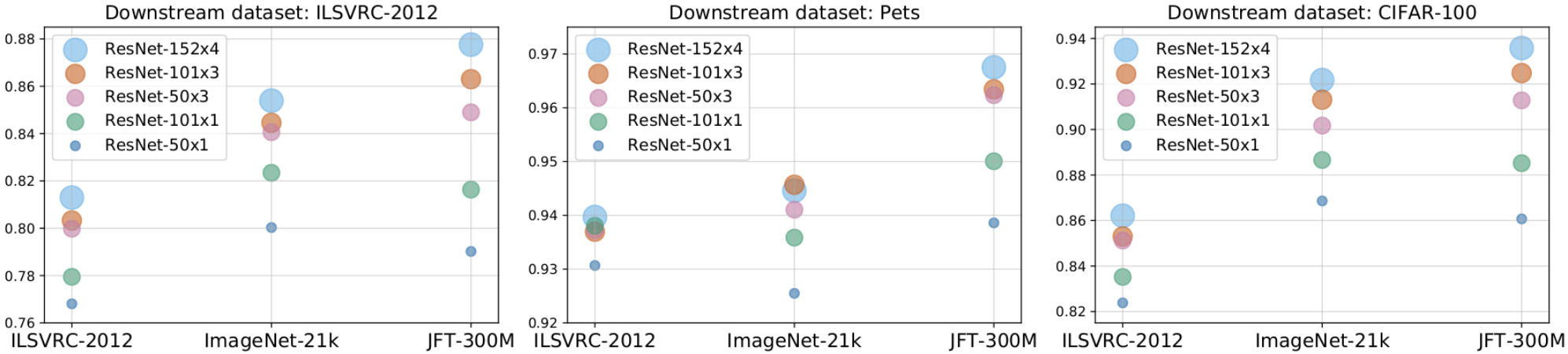


(Kolesnikov et. al., 2020)

Network and dataset sizes

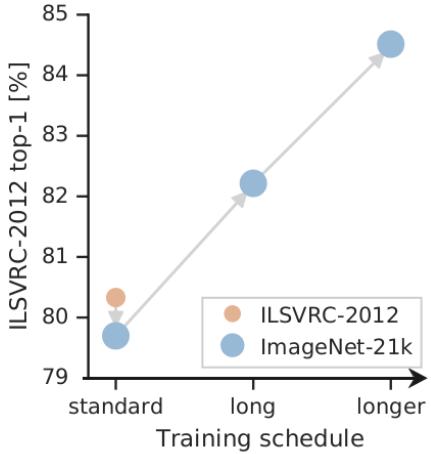
	ILSVRC-2012	CIFAR-10	CIFAR-100	Pets	Flowers	VTAB-1k (19 tasks)
BiT-S (ILSVRC-2012)	81.30	97.51	86.21	93.97	89.89	66.87
BiT-M (ImageNet-21k)	85.39	98.91	92.17	94.46	99.30	70.64
Improvement	+4.09	+1.40	+5.96	+0.49	+9.41	+3.77

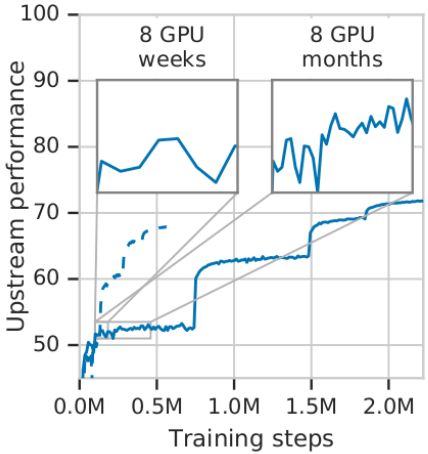
(Kolesnikov et. al., 2020)

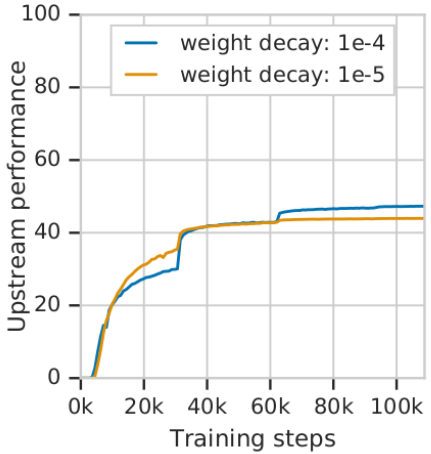


(Kolesnikov et. al., 2020)

Training budget





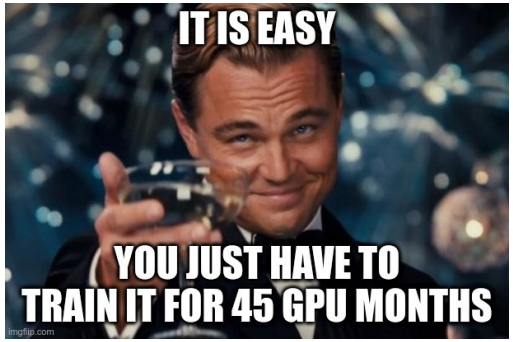


(Kolesnikov et. al., 2020)

Cloud TPU v3 Pod	Evaluation Price / hr	1-yr Commitment Price (37% discount)	3-yr Commitment Price (55% discount)
32-core Pod slice	$32 USD	$176,601 USD	$378,432 USD

(cloud.google.com)

÷ 12 x 16 = 235.468 USD
(monthly cost of the v3-512)



Conclusion

1. Bigger was better
   - Big models, datasets and computers
   - But they all have to be scaled up simultaneously

2. Normalization was essential
   - But the technique has to be appropriate to the hardware

3. It was possible to get solid target performance, with little HPO
   - A pre-tuned hyper-parameter lookup table worked fine

References

- Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., ... & Katz, B. (2019). **Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models**. In Advances in Neural Information Processing Systems (pp. 9453-9463).

- Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2019). **Big transfer (BiT): General visual representation learning**. arXiv preprint arXiv:1912.11370.

- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). **Imagenet classification with deep convolutional neural networks**. In Advances in neural information processing systems (pp. 1097-1105).

- Qiao, S., Wang, H., Liu, C., Shen, W., & Yuille, A. (2019). **Weight standardization**. arXiv preprint arXiv:1903.10520.

- Wu, Y., & He, K. (2018). **Group normalization**. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). **How transferable are features in deep neural networks?.** In Advances in neural information processing systems (pp. 3320-3328).

- Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., ... & Beyer, L. (2019). **A large-scale study of representation learning with the visual task adaptation benchmark**. arXiv preprint arXiv:1910.04867.

- Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). **mixup: Beyond empirical risk minimization**. arXiv preprint arXiv:1710.09412.

Acknowledgments

Sudhanshu Mittal

Alexander Kolesnikov

Lucas Beyer

Thank you