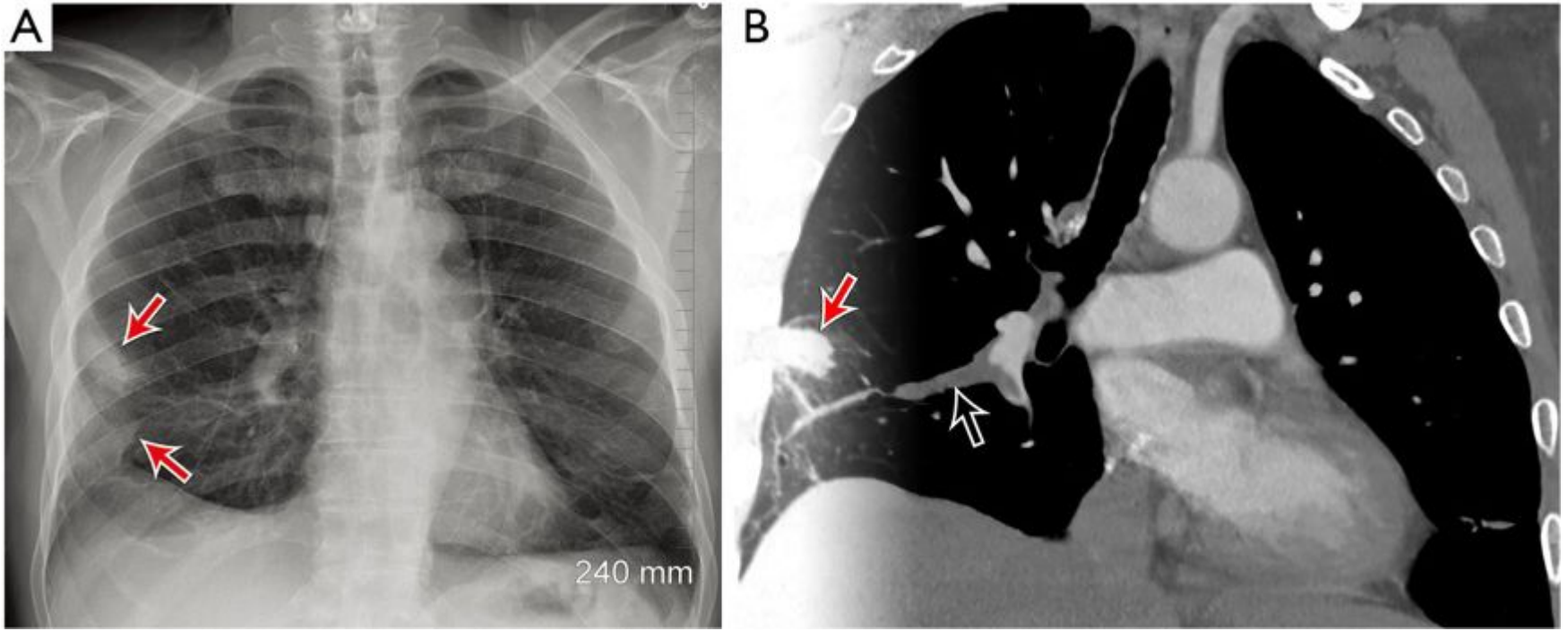
Big Transfer (BiT): General Visual Representation Learning

Anouar Meziou

Example:



Problems faced:

- Current computer vision training generally involves a pre-trained model due to lack of labeled data for computer vision tasks.
- This has been a common problem for computer vision scientists to collect and train models with a large set of generic data.

Solution:

- A common approach to mitigate the lack of labeled data for computer vision tasks is to use models that have been pre-trained on generic data (e.g., ImageNet).
- The idea is that visual features learned on the generic data can be re-used for the task of interest.
- They revisit the paradigm of pre-training on large supervised datasets and fine-tuning the model on a target task. They scale up pre-training, and propose a simple recipe that we call Big Transfer (BiT).

Three tested BiT models:

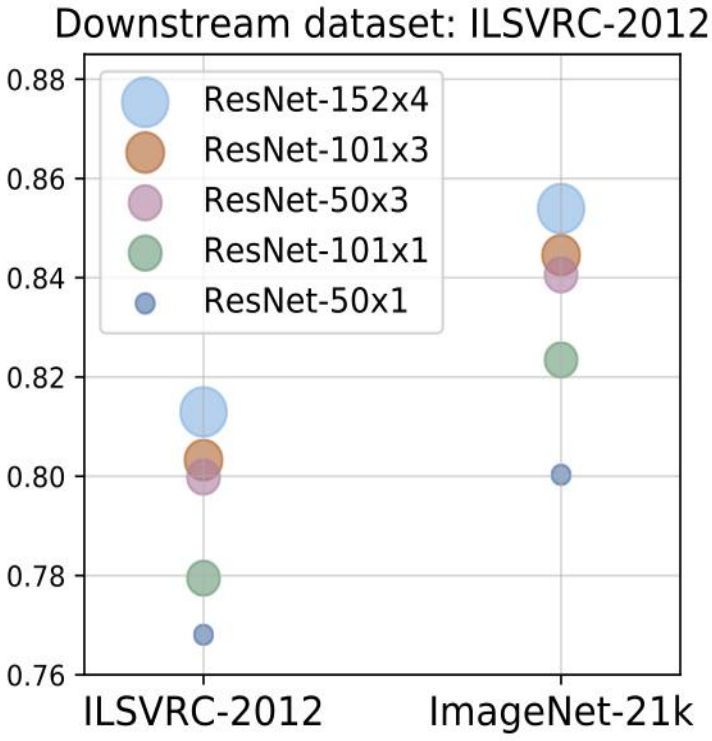
The first component is scale. It is well-known in deep learning that larger networks perform better on their respective tasks. Further, it is recognized that larger datasets require larger architectures to realize benefits, and vice versa.

- BiT-L: The largest, BiT-L is trained on the JFT-300M dataset which contains 300 M noisily labelled images with 1.26 labels per image on average.
- BiT-M: For practitioners, we will release the performant BiTM model trained on ImageNet-21k which contains 14.2 M images and 21K classes.
- BiT-S: It is trained on the ILSVRC-2012 variant of ImageNet, which contains 1.28 million images and 1000 classes. Each image has a single label.

From ResNet to BiT:

All of our BiT models use a vanilla ResNet-v2 architecture , except that we replace all Batch Normalization layers with Group Normalization and use Weight Standardization in all convolutional layers.

We train ResNet-152 architectures in all datasets, with every hidden layer widened by a factor of four (ResNet152x4).



Accuracy Comparison:

	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]	-

- Generalist SOTA: Is the training on imagenet for imagenet classification for example. At this moment you do not know, what task you are pre-training for (i.e. cifar classification / mnist classification / cat and dog detection). Then you use pretrained weights to fine-tune for the concrete task.
- Specialist SOTA: Is a more complex process. Generally this means that you know what task are you pre-training for, when you pretrain (i.e. during imagenent pretraining you already know, that you need this pretrain for cifar classification).

Thank you