Rethinking ImageNet Pre-training

Kaiming He, Ross Girshick, Piotr Dollar

Motivation and objectives

Universal assumption

When little data available, pre-trained ImageNet models *better* than from scratch

What this paper shows

On-par results when trained from scratch on MS-COCO dataset
(for object detection, instance segmentation and keypoint detection)

Contributions

1. ImageNet pre-training *speeds up* convergence

2. ImageNet pre-training does *not* automatically give better regularization

3. ImageNet pre-training does *not* show benefit when the task is different (e.g. going from *classification* to *localization*)

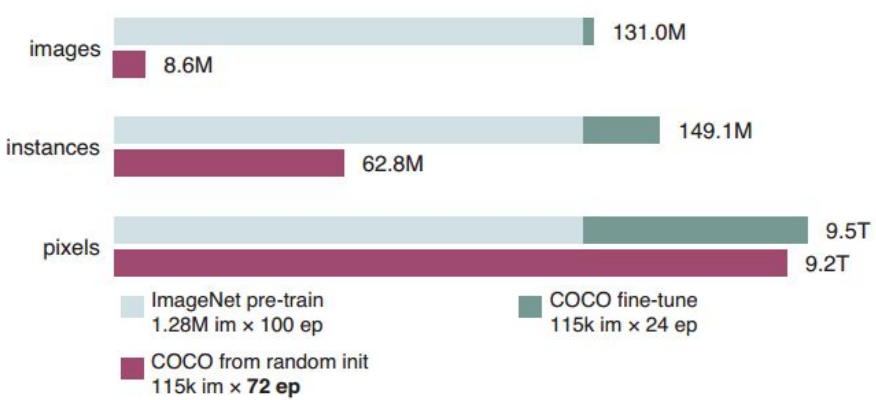
Methodology -- List of dark magic hacks

1. Replace Batch Normalization because of small batch size (n=2 per GPU) by
   - Group normalization (GN)
   - or Synchronized Batch Normalization (SyncBN)

2. Longer training for fair comparison
3. Learning rate schedule + linear warm-up of the learning rate
4. He initialization

5. Based on Detectron and Mask R-CNN

Number of epochs

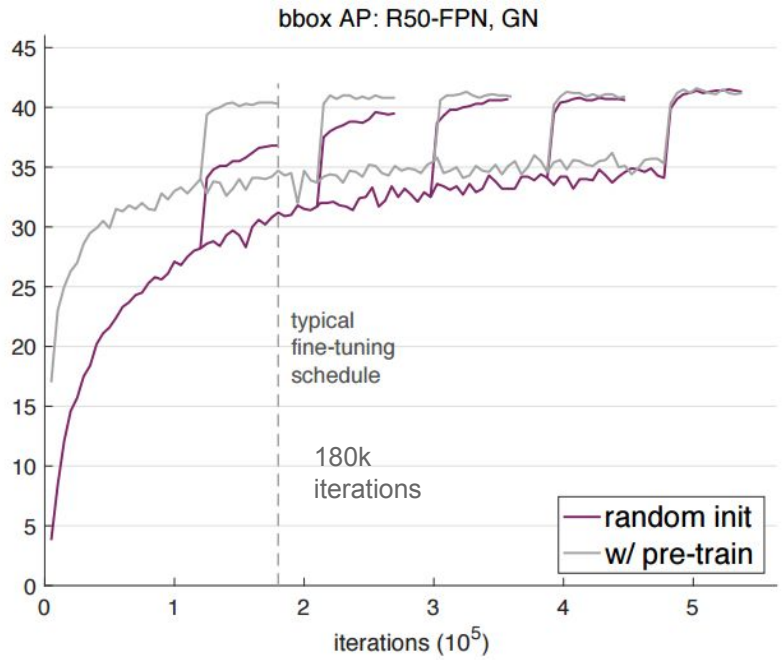


Blue: ImageNet pre-training

Green: COCO fine-tuning

Purple: COCO scratch

Pre-trained vs. from scratch -- Object detection



One schedule corresponds to 90k iterations

Learning rate is reduced by 10x in the last 60k and 20k

(similar results on instance segmentation and keypoint detection)

Pre-trained converges faster

Both are on par

Pre-trained vs. from scratch -- Ablation

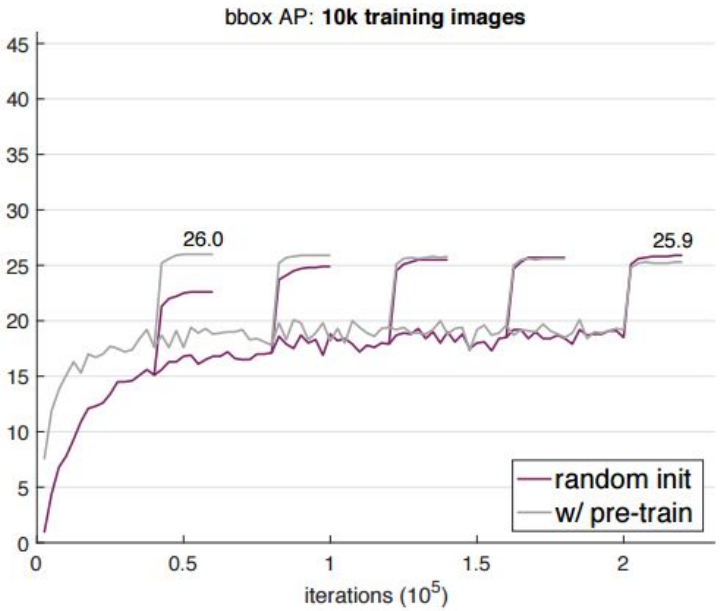
1. **GN and SyncBN.** Both show training from scratch is possible

2. **Detection metrics.** Consistent results with AP50 or AP75

3. **Data augmentation.** Consistent results when applying horizontal flipping, cascade R-CNN, test-time augmentation.

4. **Other NN models.** Consistent results with ResNeXt and VGG16 (no normalization!)

ImageNet is not necessarily a regularizer



Low data regime means overfitting

Train with *1/3* or *1/10* of MS-COCO

(Need some hyper-parameter changes, Grid search done on pre-trained then applied on from scratch for fair comparisons)

Training from scratch is no worse than pre-training in low-regime

Implications

Is ImageNet pre-training necessary? No, if enough target data and computation.

Is ImageNet helpful? Yes, allows to iterate quickly thanks to faster convergence.

Trained from scratch should be a mandatory baseline in self-supervised papers.