Paul Aggarwal

Big Transfer

QUESTION 5:

For object detection, they use the COCO-2017 dataset and train a top-performing object detector, RetinaNet (simple one-stage object detector), using pre-trained BiT models (BiT-S, BiT-M, BiT-S) as backbones (share feature extraction layer). Due to memory constraints, they use the ResNet-101x3 architecture for all their BiT models. They fine-tune the detection models on the COCO-2017 train split and do not use BiT-HyperRule but stick to the standard RetinaNet training protocol. It demonstrates that BiT models outperform standard ImageNet pretrained models. They can see clear benefits of pre-training on large data beyond ILSVRC-2012: pretraining on ImageNet-21k results in a 1.5 point improvement in Average Precision (AP), while pretraining on JFT-300M further improves performance by 0.6 points. (Refer to 201912.11370.pdf)

RetinaNet, named for its dense sampling of object locations in an input image. Its design features an efficient in-network feature pyramid and use of anchor boxes. RetinaNet is a single, unified network composed of a backbone network (in this case BiT models: BiT-S, BiT-M, BiT-S) and two task-specific subnetworks. The backbone is responsible for computing a convolutional feature map over an entire input image and is an off-the-self convolutional network. The first subnet performs convolutional object classification on the backbone's output; the second subnet performs convolutional bounding box regression. While there are many possible choices for the details of these components, most design parameters are not particularly sensitive to exact values as shown in the experiments. RetinaNet can be trained with stochastic gradient descent (SGD). (Refer to 1708.02002.pdf)

Upstream Pre-Training they use BiT models as backbone: All of their BiT models use a vanilla ResNet-v2 architecture, except that they replace all Batch Normalization layers with Group Normalization and use Weight Standardization in all convolutional layers. 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. They study the effectiveness of scale (during pre-training) in the context of transfer learning, including transfer to tasks with very few datapoints. They investigate the interplay between computational budget (training time), architecture size, and dataset size. For this, they train three BiT models on three large datasets: ILSVRC-2012 which contains 1.3M images (BiT-S), ImageNet-21k which contains 14M images (BiT-M), and JFT which contains 300M images (BiT-L). The second component is Group Normalization (GN)and Weight Standardization (WS). Batch Normalization (BN)is used in most state-of-the-art vision models to stabilize training. However, they find that BN is detrimental to Big Transfer for two reasons. First, when training large models with small per-device batches, BN performs poorly or incurs inter-device synchronization cost. Second, due to the requirement to update running statistics, BN is detrimental for transfer. GN, when combined with WS, has been shown to improve performance on small-batch training for ImageNet and COCO. Here, they show that the combination of GN and WS is useful for training with large batch sizes and has a significant impact on transfer learning.

Transfer to Downstream Fine-Tuning Tasks they use RetinaNet protocol: Train all of their models for 30 epochs using a batch size of 256 with stochastic gradient descent, 0.08 initial learning rate, 0.9 momentum and 10−4 weight decay. They decrease the initial learning rate by a factor of 10 at epochs

number 16 and 22. They did try training for longer (60 epochs) and did not observe performance improvements. The input image resolution is 1024 × 1024. During training they use a data augmentation scheme as in [34, refer to 1405.0312.pdf]: random horizontal image flips and scale jittering. They set the classification loss parameters α to 0.25 and γ to 2.0, see [33, refer to 1708.02002.pdf] for the explanation of these parameters.

RetinaNet (one-stage detector) was designed to be more efficient than two-stage detectors. One-stage detectors that are applied over a regular, dense sampling of possible object locations have the potential to be faster and simpler, but previously have trailed the accuracy of two-stage detectors. It was discovered that the extreme foreground-background class imbalance encountered during training of dense detectors is the central cause. To address this class imbalance by reshaping the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples, Focal Loss method was used which focuses training on a sparse set of hard examples and prevents the vast number of easy negatives from overwhelming the detector during training. RetinaNet was used to evaluate the Focal Loss method. Results show that when trained with the focal loss, RetinaNet is able to match the speed of previous one-stage detectors while surpassing the accuracy of all existing state-of-the-art two-stage detectors. (Refer to 1708.02002.pdf)

Applying the BiT models as backbone to RetinaNet showed clear benefits of one-stage detectors and the benefits BiT models can be to object detectors by using standard training protocols.