Appendix for Local Relation Networks for Image Recognition

Han Hu[1] Zheng Zhang[1] Zhenda Xie[1,2*] Stephen Lin[1]
[1]Microsoft Research Asia [2]Tsinghua University
{hanhu, zhez, v-zhxia, stevelin}@microsoft.com

1. Implementation details

All architectures take a 3×224×224 image as input. The architectures use a skip connection for the shortcut branch of all residual blocks except for across stages where a channel transformation layer followed by batch normalization is used. In *res3*, *res4* and *res5*, downsampling is applied on the 3×3 convolution layer or the local relation layer in the first residual blocks. For fair comparison in ablation experiments, we adapt the bottleneck ratio α to ensure the same FLOPs for different architectures.

In training, the randomly cropped images and employ scale and aspect ratio augmentation. We perform SGD optimization with a mini-batch of 1024 on 16 GPUs for all experiments except for the experiments of adversarial training in which 32 GPUs are used. The initial learning rate is 0.4, with linear warm-up in the first 5 epochs, and decays by 10× at the 30th, 60th and 90th epochs, respectively, following [1]. The total learning period is 110 epochs, with a weight decay of 0.0001 and momentum of 0.9. In inference, we use a single 224×224 center crop from the resized images with a shorter size of 256. Top-1 and top-5 accuracy are reported.

References

[1] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 1

*This work was done when Zhenda Xie was interns at Microsoft Research Asia.