Appendix for Enhancing Representation Learning with Deep Classifiers in Presence of Shortcut

Amirhossein Ahmadian and Fredrik Lindsten

A Additional implementation and hyperparameter details

What follows is additional details about the *proposed method*, for both CIFAR-10 and CelebA experiments. The learning rate used for the classifier network and lens network are 0.0002 and 0.002 respectively, and batch size is 128 (in the CIFAR-10 case, the size is 128×4 in effect due to the rotations). The higher learning rate for the lens is because we observed that in practice the lens cannot keep up with the classifier during the adversarial training (game) if it has the same learning rate and number of iterations as the classifier (possibly due to the longer gradient backpropagation path compared to the classifier). The first 2 iterations (on mini-batches) of the training are considered as warm-up iterations. During these iterations, the lens is ignored, and the classifier is trained only on the real data. It should be emphasized that the ResNet model we train on the upstream data (for our method as well as the other compared ones) is not pre-trained on any dataset, and is initialized randomly.

In the vanilla baseline (ERM) method, the learning rate is equal to the value used for the classifier in our method, to eliminate any effect of learning rate. In general, when choosing the hyperparameters for the other compared methods, we try to use the values suggested in their papers or corresponding code when a similar dataset (CIFAR-10 or CelebA) has been used in those papers.

For the *Automatic Shortcut Removal* method, the learning rate is 0.0001 for both of the networks and the datasets, and the batch size is 128. In the arrow shortcut experiment, a suggested λ was already available since this problem is experimented in their paper as well. For the light gradient and CelebA experiments, we found the best λ by probing the downstream test accuracy with $\lambda \in \{1, 10, 160, 320, 640, 1280\}$.

For the *Spectral Decoupling* method, we followed the λ and learning rate values that is used in their paper/code. This means the learning rate 0.001 (with batch size 50) for CIFAR-10 and 0.0001 (with batch size 128) for CelebA experiments.

For the *Diverse Ensemble* method, the ResNet network is used as the shared feature extractor, where its pre-logits output is fed to each head. The heads are feedforward networks with 2 layers of 512 hidden units, ReLU activation functions, and softmax output, which is the same network used in their paper. The best value of λ was determined by probing the downstream test accuracy with $\log_{10} \lambda \in \{-1, 0, 1, 2, 3, 3.6\}$. The learning rate and batch size for this method are 0.0002 and 128.

The downstream classifier is always trained with ADAM optimizer, learning rate 0.001, and batch size 64.