1. Hyperparameters

1.1. CIFAR-100 [4]

Wideresnet-28-8 [7] architecture outputs a feature vector of size 512, which is equally divided between two-level hierarchy and nearly equal as 170, 170 and 172 (coarsest to finest) for three-level of hierarchies to be fed into the hierarchical label classifiers. A batch size of 64 is used.

HIERMATCH (M):

We use random pad and crop, followed by random flip as the data-augmentations techniques. We use the Adam optimizer [3] with a fixed learning rate of 0.002 and β=(0.9, 0.99), with a weight decay of 4e-5. All the models are trained for 500 epochs with one epoch consisting of a total of 1024 iterations. We set unlabeled loss weight λ_u to 150, same as [1].

HIERMATCH (F):

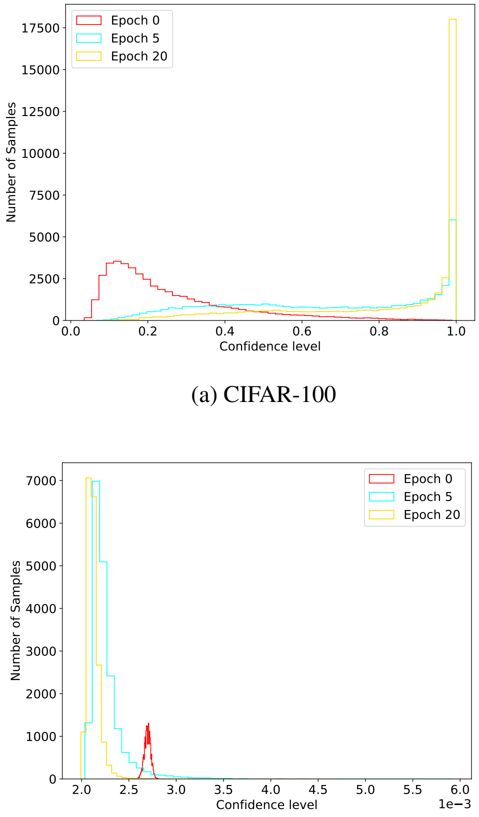
We use random horizontal flip followed by random crop for "weak" augmentation and additionally use RandAugment [2] for "strong" augmentation of the image, following baseline FixMatch [5] across all the experiments. We utilize SGD optimizer with a nesterov momentum of 0.9. Cosine learning rate scheduler with an initial learning rate of 0.03 is used. Unlabeled loss weight λ_u is set to 1 throughout the training. We use unlabeled data ratio of 7 and set confidence threshold as 0.95.

1.2. North-American Birds (NABirds) [6]

HIERMATCH (M):

We augment our data as is standard in the FGVC domain. The images are resized to 256x256 and randomly cropped to 224x224, followed by random horizontal flips and normalization. The backbone outputs a feature vector of size 2048, which is then divided into 682, 682 and 684 features across the 3 hierarchies (coarsest to finest). We use a batch size of 16. For the selection of the best set of hyper-parameters, we perform a sweep over λ_u, backbone learning rate, and classifier learning rate for 25 epochs over 20% of the labeled set. Specifically we try λ_u values $\in$ {25, 75, 100, 150, 200}, backbone learning rates $\in$ {1e-4, 5e-4, 1e-5, 5e-5, 1e-6, 5e-6} and classifier learning rates $\in$ {1e-3, 1e-4, 5e-4, 1e-5, 5e-5, 1e-6}. Our hyperparameter sweep found λ_u of 100, backbone learning rate of 1e-5, and classifier learning rate of 1e-3 as the best set of hyperparameters. We use the Adam optimizer [3] for both backbone and label classifiers. We use a batch size of 16. All the models are trained for 250 epochs.

We set the rest of the MixMatch hyperparameters - temperature sharpening T = 0.5, MixUp beta distribution parameter α = 0.75, and number of augmentations K = 2. These remain the same for all the experiments on both NABirds and CIFAR-100. For all our MixMatch experiments, we linearly ramp up λ_u to its maximum value over the entire training.



(a) CIFAR-100

(b) NABirds

Figure 1: Histogram of confidence values on pseudo-labels of unlabeled data at different epochs on CIFAR-100 (top) and NABirds (bottom) dataset.

HIERMATCH (F):

For labeled data and "weak" augmentation of an image, we use the same augmentations as is used for HIERMATCH (M) for NABirds. For "strong" augmentation, we additionally employ RandAugment similar to CIFAR-100. We use backbone WideResNet-50-2 [7] backbone, pretrained on ImageNet, as our backbone network. We sweep over backbone learning rates $\in$ {1e-4, 5e-4, 5e-5, 1e-5]}, classifier learning rates $\in$ {1e-3, 1e-4, 5e-4, 1e-5, 5e-5, 1e-6}, and threshold values $\in$ {0.5, 0.7, 0.95}. SGD optimizer with nestorov momentum of 0.9 was used.

FixMatch enforces consistency regularisation only when the model is confident. Interestingly, in our experiments of NABirds, we observe that with the best hyperparameters (backbone learning rate of 1e-5 and classifier learning rate of 1e-3), and with a minimum confidence threshold of 0.5, all of the most confident unlabeled samples fall below this threshold. Consequently, the consistency regularization is not imposed at all, implying that the unlabeled loss is zero throughout the training phase while the training accuracy on labeled set overfits the dataset with 100%. The resulting validation accuracy is similar to that of the Fully-supervised setting using limited labeled samples on NABirds. In Figure 1a of CIFAR-100, as the training progresses the confidence values on unlabeled data improve whereas in Figure 1b the confidence values on unlabeled data is too low and therefore, unlabeled data is not used at all while FixMatch training. FixMatch requires both "weak" and "strong" augmentations using RandAugment [2] which requires more careful experimentation to design such augmentations for fine-grained datasets like NABirds. Despite our hyperparameter sweeps, the baseline FixMatch [5] gives poor performance on NABirds [6], so we do not report experimental evaluation of NABirds on FixMatch and HIERMATCH (F).

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[2] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[4] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[5] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020.

[6] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[7] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.