Hyperparameters

Type	Hyperparameters	ImageNet	CIFAR10	CIFAR100
Training	Epochs	15		
	Batch size	256		
	AdamW ε	1e-8		
	AdamW β	(0.9, 0.999)		
	Weight decay	0.01		
	Base learning rate	1e-3		
	Layer Decay	0.81		
	Learning rate schedule	Cosine		
	Warmup steps	10% of total steps		
	Hardware	1 × RTX 4090 24GB		
Mixup [1]/**Cutmix** [2]	Mixup prob.	0.8		
	Cutmix prob.	1.0		
	Prob.	0.9		
	Switch prob.	0.5		
	Label smooting	0.1		
RandAugment [3]	Magintude	9		
	Magnitude std.	0.5		
	Magnitude inc.	1		
	# ops	2		
RandomErase [4]	Prob.	0.25		
	Mode	pixel		
	# erase	1		

Table 1: Hyperparameters used for the ImageNet-1K [5], CIFAR10 [6], and CIFAR100 [6] finetuning of the distilled Data2Vec image model. We refer to the respective papers for details on the augmentation techniques [1]–[4].

Bibliography

[1] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[2] S. Yun, D. Han, S. Chun, S. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2019, pp. 6022–6031. doi: 10.1109/ICCV.2019.00612.

[3] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 3008–3017. doi: 10.1109/CVPRW50498.2020.00359.

[4] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random Erasing Data Augmentation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 13001–13008, Apr. 2020, doi: 10.1609/aaai.v34i07.7000.

[5] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[6] A. Krizhevsky, “Learning multiple layers of features from tiny images,” 2009.