HPFF: Hierarchical Locally Supervised Learning with Patch Feature Fusion

Junhao Su[1*], Chenghao He[2*], Feiyu Zhu[3*], Xiaojie Xu[4*]
Dongzhi Guan[1], and Chenyang Si[5✉]

[1] Southeast University
[2] East China University of Science and Technology
[3] University of Shanghai for Science and Technology
[4] The Hong Kong University of Science and Technology
[5] Nanyang Technological University

1 Datasets

In this section, we provide a brief introduction to the four image classification datasets that we utilized, as well as the data augmentation methods employed.

SVHN [6] is a real-world image dataset used for developing machine learning and object recognition algorithms, especially for recognizing digits in visual objects. It is derived from Google Street View data and contains over 600,000 images of digits, covering 10 classes (0-9). Each image is a 32×32 pixel color image.

CIFAR-10 [5] dataset contains 60,000, 32×32 color images, divided into 10 classes, with 6,000 images per class. These classes include airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. The dataset is divided into 50,000 training images and 10,000 test images.

STL-10 [2] dataset is designed to evaluate unsupervised feature learning and self-learning algorithms. It is inspired by the CIFAR-10 dataset but has some changes. It includes 10 classes, each with 500 training images and 800 test images. All images are 96×96 color images. In addition, an unlabeled dataset is provided, containing 100,000 additional images.

ImageNet [3] is a large-scale visual database composed of over 10 million high-resolution images with detailed labels. These images cover more than 20,000 categories, with the number of images per category ranging from a few hundred to tens of thousands. The goal of ImageNet is to provide researchers with an easily accessible, large-scale image database to assist them in their research in computer vision and other fields.

*Equal Contribution.
✉Corresponding Author: Chengyang Si (chenyang.si.mail@gmail.com).

2 More Results

2.1 More Results on ImageNet

Due to InfoPro involving combinations of hyperparameters, and the source code only provides combinations for K=2 and 4, we use the hyperparameter combination for K=4 to conduct experiments with K=8, and results are presented in Table 1.

Table 1: More results on the ImageNet. Results are obtained from a training of 90 epochs.

Network	Method	Top1-Error	Top5-Error
ResNet-101	E2E	22.03	5.93
	InfoPro(K=8)	27.06	9.19
	InfoPro+HPFF(K=8)	**22.87**	**6.54**

2.2 Results on ViT

We attempt to apply supervised local learning to the computation of Vision Transformer [4] and have displayed the detailed results in Table 2. We use ViT-B/16 as the backbone, set the batch size to 1024, and trained for 200 epochs.

Table 2: HPFF effect on ViT. We conduct training for 200 epochs with a batch size of 1024.

Dataset	Method	ViT-B/16	
		ACC(%)	GPU Memory(GB)
CIFAR-10	E2E	88.99	5.43
	DGL(K=12)	61.48	2.69(↓ 50.48%)
	DGL+HPFF(K=12)	**74.65**	**2.64(↓ 51.38%)**

2.3 Abalation study on PFF

We conduct a further ablation study on the PFF module to investigate its impact on performance. As seen in Table 3, setting n to 2 achieves the optimal balance between GPU memory usage, performance, and training time.

Table 3: A detailed ablation study of the PFF. We use the ResNet-32 (K=16) as backbone on the CIFAR-10 dataset, and use a batch size of 1024, with the total training time amounting to 400 epochs.

Method	GPU Memory	Training Time	Test Error
InfoPro	2.67GB	143min	12.93
InfoPro+PFF(n=2)	**1.62GB**	**164min**	**11.17**
InfoPro+PFF(n=4)	**1.41GB**	**413min**	**11.83**
InfoPro+HPFF(n=2)	**2.31GB**	**177min**	**8.99**

2.4 Results on Cityscapes

We further validate the performance of our HPFF on segmentation tasks. Results can be seen in Table. 4. As can be seen, our HPFF significantly enhances the performance of supervised local learning methods on segmentation tasks, even surpassing that of BP.

Table 4: HPFF effect on Cityscapes. We train for 40K iterations using DeepLab-V3-R101 as the backbone. The batch size is 8. 'SS' refers to the single-scale inference. 'MS' and 'Flip' denote employing the average prediction of multi-scale ([0.5, 1.75]) and left-right flipped inputs during inference.

Crop Size	Method	mIoU		
		SS	MS	MS+Flip
512×1024	E2E	79.12%	79.81%	80.02%
	InfoPro(K=4)	78.25%	79.14%	79.28%
	InfoPro+HPFF(K=4)	**80.04%**	**80.62%**	**81.13%**

3 Generalization Study

In this section, we study the generalization of our proposed HPFF. We directly use the checkpoints trained on the CIFAR-10 [5] dataset for testing on the STL-10 [2] dataset, which is inspired by [7].

From Table 5, we can observe a significant difference in the generalization abilities between DGL and BP. However, after adding our HPFF method, the test accuracy improved significantly, even surpassing BP. Based on these results, we can infer that HPFF, by facilitating information interaction between local modules, enhances the generalization ablity of supervised local learning method.

Table 5: Generalization study. Checkpoints are trained on the CIFAR-10 dataset and tested on the STL-10 dataset. The data in the table represents the test accuracy.

Method	ResNet-32 (K=16)	ResNet-110 (K=55)
BP	35.98	36.78
DGL [1]	31.95	33.16
DGL*	**39.06**	**40.62**

References

1. Belilovsky, E., Eickenberg, M., Oyallon, E.: Greedy layerwise learning can scale to imagenet. In: International conference on machine learning. pp. 583–593. PMLR (2019)

2. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215–223. JMLR Workshop and Conference Proceedings (2011)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
6. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
7. Qu, Z., Jin, H., Zhou, Y., Yang, Z., Zhang, W.: Focus on local: Detecting lane marker from bottom up via key point. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14122–14130 (2021)