Momentum Auxiliary Network for Supervised Local Learning

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

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

1 Datasets

We will briefly introduce the four datasets used in our experiments and the data augmentation methods used in each dataset.

CIFAR-10 [6] dataset contains 60,000, 32 × 32 color images, categorized into 10 different classes, often used for image classification tasks in machine learning.The standard data augmentation has been used in the training set, where 4 pixels are padded on each side of samples followed by a 32 × 32 crop and a random horizontal flip.

SVHN [7] dataset consisting of images of house numbers collected from Google Street View. It contains digits 0-9 and is commonly used for digit recognition tasks in machine learning.In the process of augmenting training samples, we add a 2-pixel border around the images and then perform a 32 × 32 crop.

STL-10 [2] is an image dataset with 10 classes, derived from the larger CIFAR-10 dataset. It contains 10,000 labeled images and 100,000 unlabeled images, making it suitable for semi-supervised and self-supervised learning experiments in computer vision.Data augmentation involves applying a random 4 × 4 translation followed by a random horizontal flip.

ImageNet [3] is a large-scale image dataset with 1.28 million labeled images spanning 1,000 different object categories. It is a widely used benchmark in computer vision for task like image classification.For training samples, we use a 224 × 224 random crop with random horizontal flips, while for test samples, we apply a 224 × 224 resize followed by a central crop.

2 More Results

2.1 More Results on Image Classification Benchmarks

When selecting the hyperparameter *momentum* for the Exponential Moving Average (EMA) [5] in the Momentum Auxiliary Network, we conduct multiple

*Equal Contribution.
✉Corresponding Authors: Chengyang Si (chenyang.si.mail@gmail.com) and Dongzhi Guan (guandongzhi@seu.edu.cn).

explorations and comparisons. Ultimately, we determine *momentum* to be 0.995 based on the experimental results. The experimental results of other hyperparameters *momentum* can be seen in Table 1.

If the value of *momentum* is set too high, it will hinder the beneficial information interaction between local blocks. On the other hand, if the value of *momentum* is set too low, the inconsistency between the features and local optimization objectives among local blocks can lead to limited performance improvement.

2.2 More Results on ImageNet

Since the InfoPro method involves a combination of hyperparameters, its source code does not provide the hyperparameter combination for K=8 on the ImageNet dataset. We attempted to use the hyperparameter combination for K=4 to conduct further experiments with K=8, and the results are shown in Table 2.

Table 1: The comparison of experimental results with different hyperparameters *momentum*, where m represents *momentum*. The ***** means addition of our Momentum Auxiliary Network.

Dataset	Method	ResNet-32		ResNet-110	
		K = 8 (Test Error)	K = 16 (Test Error)	K = 32 (Test Error)	K = 55 (Test Error)
CIFAR-10 (E2E(ResNet-32)=6.37, E2E(ResNet-110)=5.42)	PredSim [8]	20.62	22.71	22.08	24.74
	PredSim*(m=0.999)	**14.98**	**15.95**	**17.09**	**18.59**
	PredSim*(m=0.998)	**15.03**	**15.92**	**17.25**	**18.47**
	PredSim*(m=0.990)	**14.25**	**15.60**	**17.11**	**18.01**
	DGL [1]	11.63	14.08	12.51	14.45
	DGL*(m=0.999)	**9.18**	**10.31**	**9.97**	**10.19**
	DGL*(m=0.998)	**9.97**	**9.94**	**9.60**	**10.34**
	DGL*(m=0.990)	**8.96**	**10.15**	**9.94**	**10.03**
	InfoPro [10]	11.51	12.93	12.26	13.22
	InfoPro*(m=0.999)	**9.64**	**9.84**	**9.93**	**10.45**
	InfoPro*(m=0.998)	**9.53**	**9.81**	**9.72**	**9.93**
	InfoPro*(m=0.990)	**9.58**	**9.77**	**9.15**	**9.74**
STL-10 (E2E(ResNet-32)=19.35, E2E(ResNet-110)=19.67)	PredSim [8]	31.97	32.90	32.05	33.27
	PredSim*(m=0.999)	**28.91**	**28.95**	**31.40**	**32.20**
	PredSim*(m=0.998)	**29.76**	**29.35**	**30.92**	**31.71**
	PredSim*(m=0.990)	**29.42**	**30.01**	**30.59**	**31.68**
	DGL [1]	25.05	27.14	25.67	28.16
	DGL*(m=0.999)	**21.43**	**21.47**	**22.95**	**23.93**
	DGL*(m=0.998)	**21.38**	**21.90**	**23.12**	**23.64**
	DGL*(m=0.990)	**21.47**	**21.75**	**22.66**	**22.68**
	InfoPro [10]	27.32	29.28	28.58	29.20
	InfoPro*(m=0.999)	**23.74**	**24.63**	**24.93**	**24.88**
	InfoPro*(m=0.998)	**23.69**	**24.77**	**24.25**	**24.79**
	InfoPro*(m=0.990)	**23.22**	**24.04**	**24.37**	**25.01**
SVHN (E2E(ResNet-32)=2.99, E2E(ResNet-110)=2.92)	PredSim [8]	6.91	8.08	9.12	10.47
	PredSim*(m=0.999)	**5.47**	**6.78**	**7.84**	**8.51**
	PredSim*(m=0.998)	**5.59**	**7.01**	**7.45**	**8.39**
	PredSim*(m=0.990)	**5.69**	**6.97**	**7.19**	**8.43**
	DGL [1]	4.83	5.05	5.12	5.36
	DGL*(m=0.999)	**3.94**	**4.31**	**4.48**	**5.16**
	DGL*(m=0.998)	**4.00**	**4.14**	**4.32**	**4.94**
	DGL*(m=0.990)	**4.03**	**4.09**	**4.42**	**5.01**
	InfoPro [10]	5.61	5.97	5.89	6.11
	InfoPro*(m=0.999)	**4.68**	**5.62**	**5.01**	**4.97**
	InfoPro*(m=0.998)	**4.74**	**5.59**	**4.92**	**5.03**
	InfoPro*(m=0.990)	**4.43**	**5.38**	**4.89**	**4.99**

Table 2: Results on the validation set of ImageNet

Network	Method	Top1-Error	Top5-Error
ResNet-101	E2E	22.03	5.93
	InfoPro(K=8) [10]	27.06	9.19
	InfoPro*(K=8)	**23.11(↓3.95)**	**6.69(↓2.50)**
ResNet-152	E2E	21.60	5.92
	InfoPro(K=8) [10]	25.91	8.76
	InfoPro*(K=8)	**22.17(↓3.74)**	**5.99(↓2.77)**
ResNeXt-101, 32 × 8d	E2E	20.64	5.40
	InfoPro(K=8) [10]	25.55	8.44
	InfoPro*(K=8)	**20.93(↓4.62)**	**5.71(↓2.73)**

2.3 More Results on GPU Memory Usage

We further provide the GPU memory usage of CIFAR-10 dataset after integrating our MAN approach in Table 3 for reference.

Table 3: Comparison of GPU memory usage between E2E and other methods with MAN on the CIFAR-10 dataset.

Backbone	Method	GPU Memory(GB)
ResNet-32 (K=16)	E2E	3.37
	PredSim*	**2.33(↓30.9%)**
	DGL*	**2.02(↓40.1%)**
	InfoPro*	**3.08(↓8.7%)**
ResNet-110 (K=55)	E2E	9.26
	PredSim*	**2.47(↓73.3%)**
	DGL*	**2.11(↓77.2%)**
	InfoPro*	**3.22(↓65.2%)**

2.4 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 4. We use ViT-B/16 as the backbone, set the batch size to 1024, and trained for 200 epochs.

Table 4: MAN 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=6)	69.13	3.17(↓ 41.63%)
	DGL+MAN(K=6)	**83.04**	**3.39(↓ 37.56%)**

2.5 Results on Cityscapes

We conduct further experiments on segmentation tasks to verify that our MAN maintains good performance in other computer vision tasks as well.

Table 5: MAN effect on Cityscapes. We train for 40K iterations using DeepLab-V3-R101 as backbone. Crop size is 512×1024, batch size is 8. 'SS' refers to the single-scale inference. 'MS' and 'Flip' are employing the average prediction of multi-scale ([0.5, 1.75]) and left-right flipped inputs during inference.

Method	GPU Memory	mIoU		
		SS	MS	MS+Flip
E2E	38.85GB	79.12%	79.81%	80.02%
InfoPro(K=4)	19.88GB	78.25%	79.14%	79.28%
InfoPro+MAN(K=4)	**20.25GB**	**79.31%**	**80.05%**	**80.40%**

3 Generalization Study

In this section, we study the generalization of our proposed Momentum Auxiliary Network. To this end, we employ the checkpoint trained from the CIFAR-10 [6] training set to perform testing with the STL-10 [2] testing set, which is inspired by [9]. As shown in Table 4, the performance gap between DGL and E2E is substantial. However, with the addition of our proposed MAN, there is a marked performance improvement. This method even surpasses E2E training. These results illustrate that MAN significantly enhances the generalization ability of supervised local learning methods through information exchange between local blocks.

Table 6: Generalization study. Models are trained with 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)
E2E	35.98	36.78
DGL	31.95	33.16
DGL*	**37.55**	**36.92**

References

1. Belilovsky, E., Eickenberg, M., Oyallon, E.: Decoupled greedy learning of cnns. In: International Conference on Machine Learning. pp. 736–745. PMLR (2020)

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. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
6. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
7. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
8. Nøkland, A., Eidnes, L.H.: Training neural networks with local error signals. In: International conference on machine learning. pp. 4839–4850. PMLR (2019)
9. 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)
10. Wang, Y., Ni, Z., Song, S., Yang, L., Huang, G.: Revisiting locally supervised learning: an alternative to end-to-end training. arXiv preprint arXiv:2101.10832 (2021)