Scaling Supervised Local Learning with Augmented Auxiliary Networks

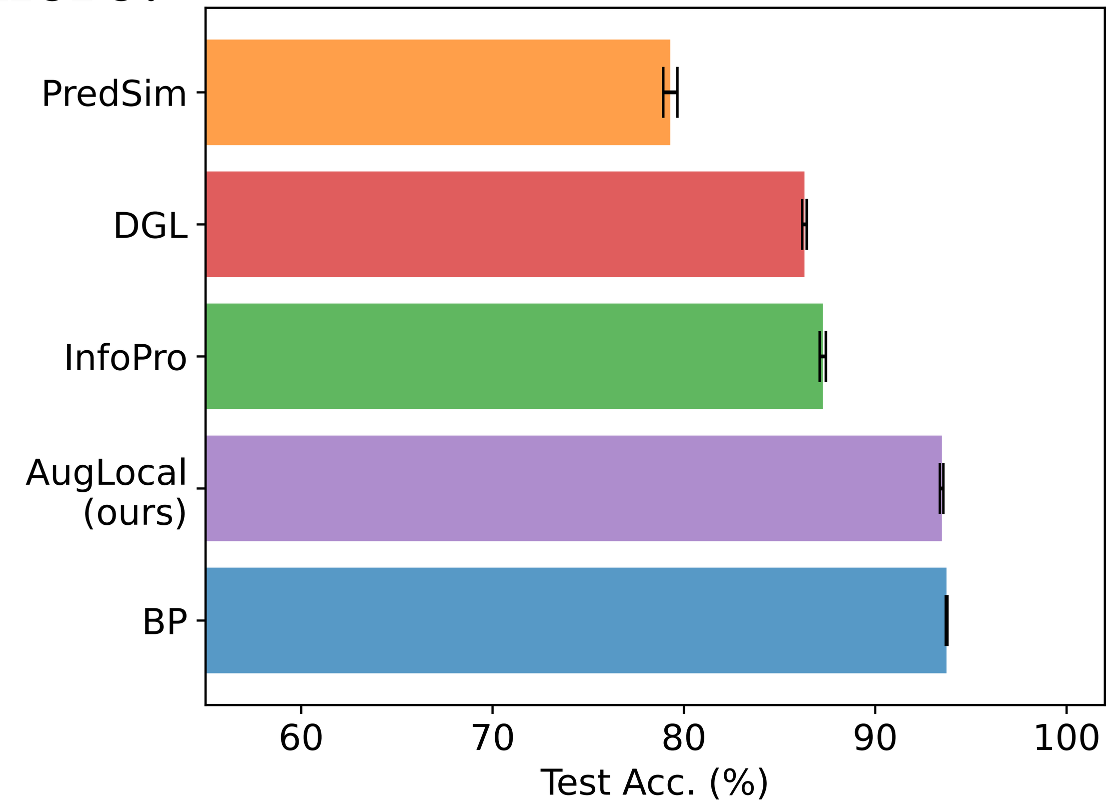
Chenxiang Ma[1], Jibin Wu[1]✉, Chenyang Si[2], Kay Chen Tan[1]

[1]The Hong Kong Polytechnic University, Hong Kong SAR, China

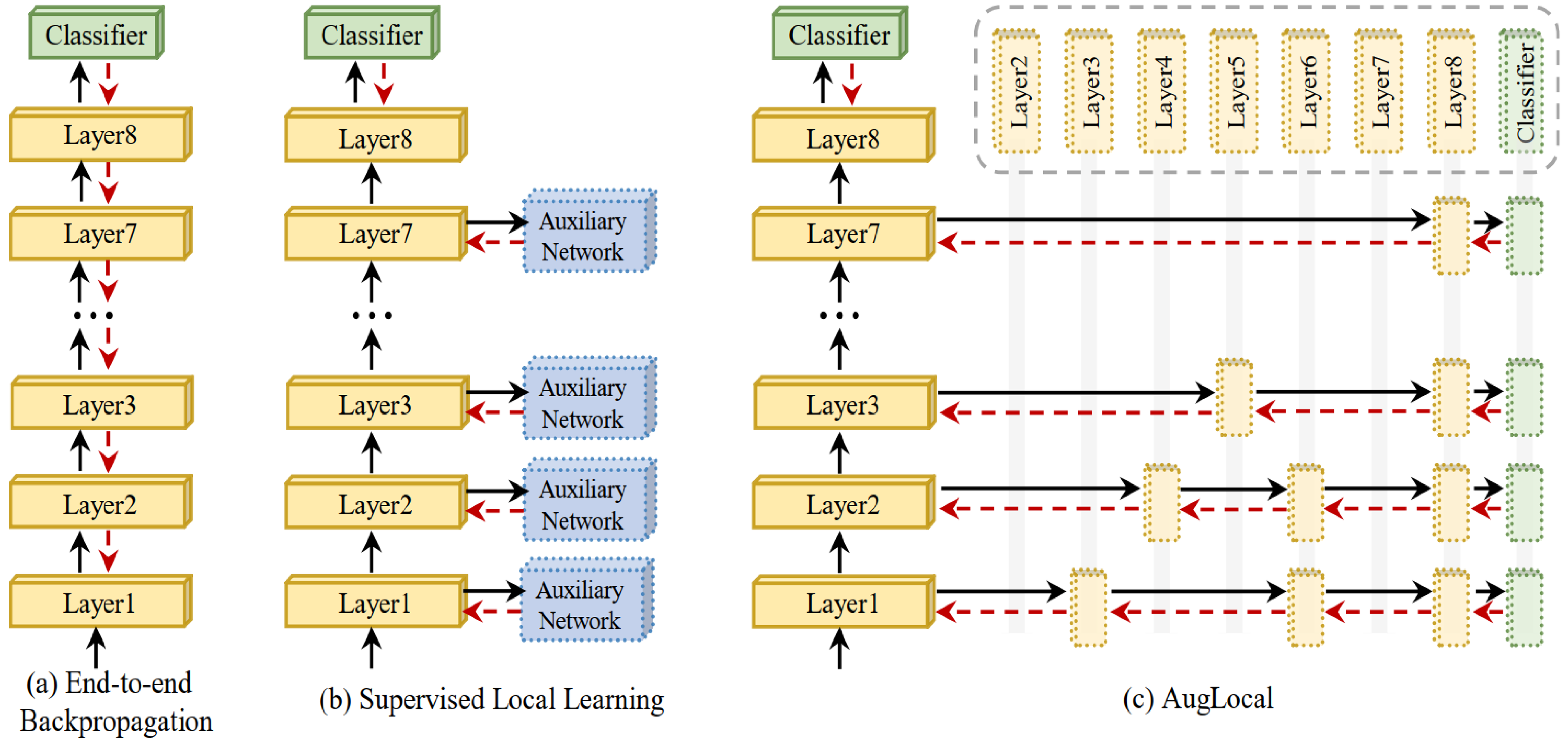
[2]Nanyang Technological University, Singapore

Background & Motivation

- Backpropagation (BP)
  - Biological implausibility
  - Update locking
  - Huge memory requirement
- Local learning:
  - Independent training of each layer
  - Addressing of all the above issues of BP
  - **Lower accuracy compared to BP**
  - **Short-sightedness problem:** Hidden layers learn representations for their local targets rather than benefiting subsequent layers
- Comparison of local learning rules and BP on CIFAR-10. ResNet32, with 16 local layers, is used here.



Method



- To strengthen the synergy between local layers and their subsequent layers, we propose an augmented local learning rule, namely AugLocal, which builds each local layer's auxiliary network using a uniformly sampled small subset of its subsequent layers.
- To reduce the additional computational cost of auxiliary networks, we propose a pyramidal structure that linearly decreases the depth of auxiliary networks as the local layer approaches the output.

Results

- Results on image classification datasets

Network	Method	CIFAR-10	SVHN	STL-10
ResNet-32 ($L=16$)	BP	93.73±0.04	97.01±0.03	80.80±0.17
	PredSim (Nøkland & Eidnes, 2019)	79.29±0.37	92.02±0.35	70.67±0.39
	InfoPro (Wang et al., 2021)	87.26±0.16	93.30±0.73	70.85±0.14
	DGL (Belilovsky et al., 2020)	86.30±0.12	95.14±0.09	73.13±1.08
	AugLocal ($d=2$)	91.12±0.24	95.86±0.04	78.58±0.66
	AugLocal ($d=3$)	92.26±0.20	96.43±0.02	79.79±0.27
	AugLocal ($d=4$)	93.08±0.10	96.79±0.08	80.65±0.16
	AugLocal ($d=5$)	93.38±0.11	**96.87±0.03**	80.73±0.18
	AugLocal ($d=6$)	**93.47±0.09**	96.85±0.08	**80.95±0.55**
ResNet-110 ($L=55$)	BP	94.61±0.18	97.10±0.05	80.41±0.74
	PredSim (Nøkland & Eidnes, 2019)	74.95±0.36	89.90±0.76	68.91±0.48
	InfoPro (Wang et al., 2021)	86.95±0.46	92.26±0.59	70.61±0.50
	DGL (Belilovsky et al., 2020)	85.69±0.32	95.12±0.06	72.27±0.51
	AugLocal ($d=2$)	90.98±0.05	95.92±0.10	78.29±0.37
	AugLocal ($d=3$)	92.62±0.22	96.45±0.08	79.30±0.26
	AugLocal ($d=4$)	93.22±0.17	96.74±0.11	**80.77±0.33**
	AugLocal ($d=5$)	93.75±0.20	96.85±0.05	80.20±0.56
	AugLocal ($d=6$)	**93.96±0.15**	**96.96±0.01**	80.40±0.17

- Results on ImageNet

Network	Method	Top-1 Acc.	Top-5 Acc.
VGG13 ($L=10$)	BP	71.59	90.37
	DGL	67.32	87.81
	AugLocal	**70.92**	**90.13**
ResNet-34 ($L=17$)	BP	74.28	91.76
	AugLocal	**73.95**	**91.70**
ResNet-101 ($L=34$)	BP	77.34	93.71
	AugLocal	**76.70**	**93.29**

- GPU Memory Efficiency

Dataset	Network	Method	GPU Memory (GB)
CIFAR-10	ResNet-32 ($L=16$)	BP	3.15
		AugLocal	1.67 (↓ **47.0%**)
	ResNet-110 ($L=55$)	BP	9.27
		GC	3.03 (↓ 67.3%)
		AugLocal	1.72 (↓ **81.5%**)
ImageNet	ResNet-34 ($L=17$)	BP	42.95
		AugLocal	29.04 (↓ **32.4%**)
	ResNet-101 ($L=34$)	BP	157.12
		AugLocal	97.65 (↓ **37.9%**)

- Representation Similarity Analysis