周学习总结

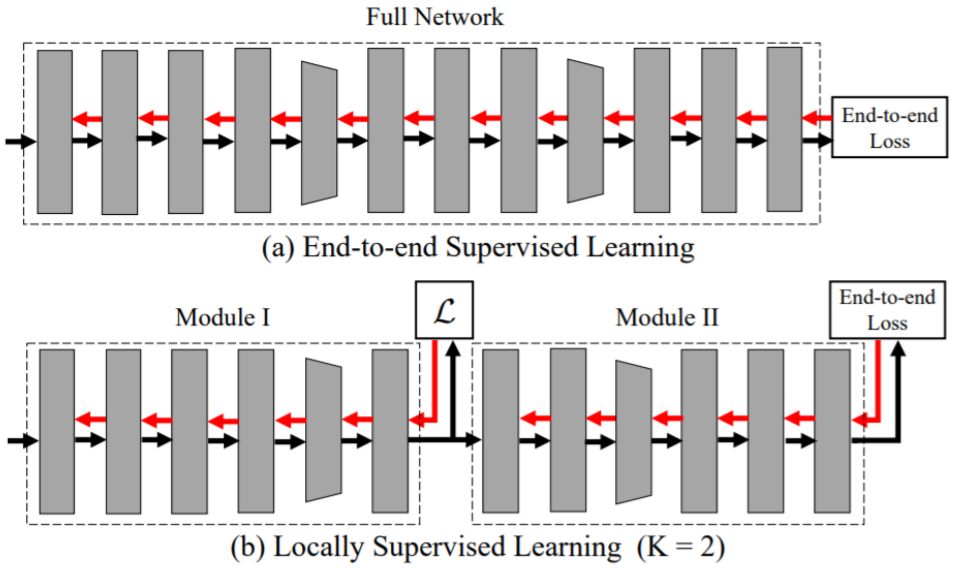
许典

视频学习

- 视频学习到了GAN 和Meta Learning，总体进度(95/119)
- 最近没有合适的设备来做视频后面的实验，暂且搁置

论文

- [Revisiting Locally Supervised Learning: an Alternative to End-to-end Training(ICLR 2021)]
- 局部监督学习：一种端到端训练的替代方法



可以节省显存或者增加并行性

然而这种方法会对模型的效果造成影响

Table 1: Test errors of a ResNet-32 using greedy SL on CIFAR-10. The network is divided into K successive local modules. Each module is trained separately with the softmax cross-entropy loss by appending a global-pool layer followed by a fully-connected layer (see Appendix F for details). "$K=1$" refers to end-to-end (E2E) training.

	$K=1$	$K=2$	$K=4$	$K=8$	$K=16$
Test Error	7.37%	10.30%	16.07%	21.19%	24.59%

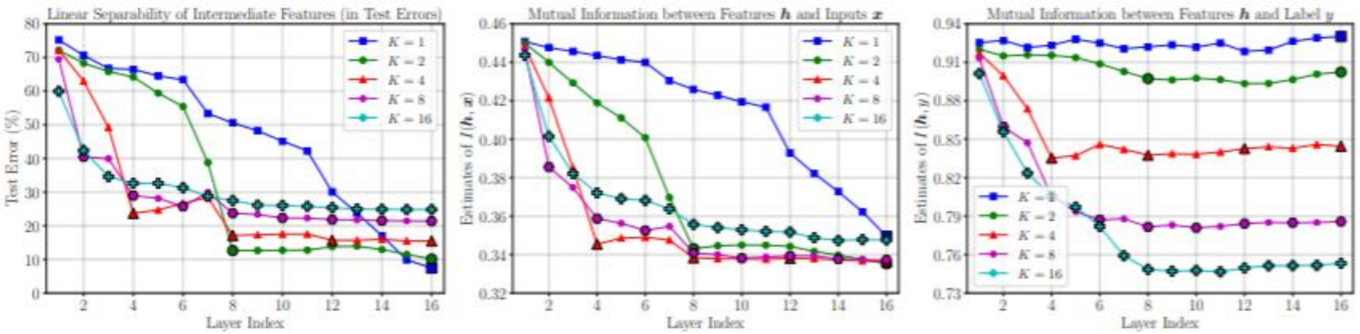


Figure 2: The linear separability (*left*, measured by test errors), mutual information with the input x (*middle*), and mutual information with the label y (*right*) of the intermediate features h from different layers when the greedy supervised learning (greedy SL) algorithm is adopted with K local modules. The ends of local modules are marked using larger markers with black edges. The experiments are conducted on CIFAR-10 with a ResNet-32.

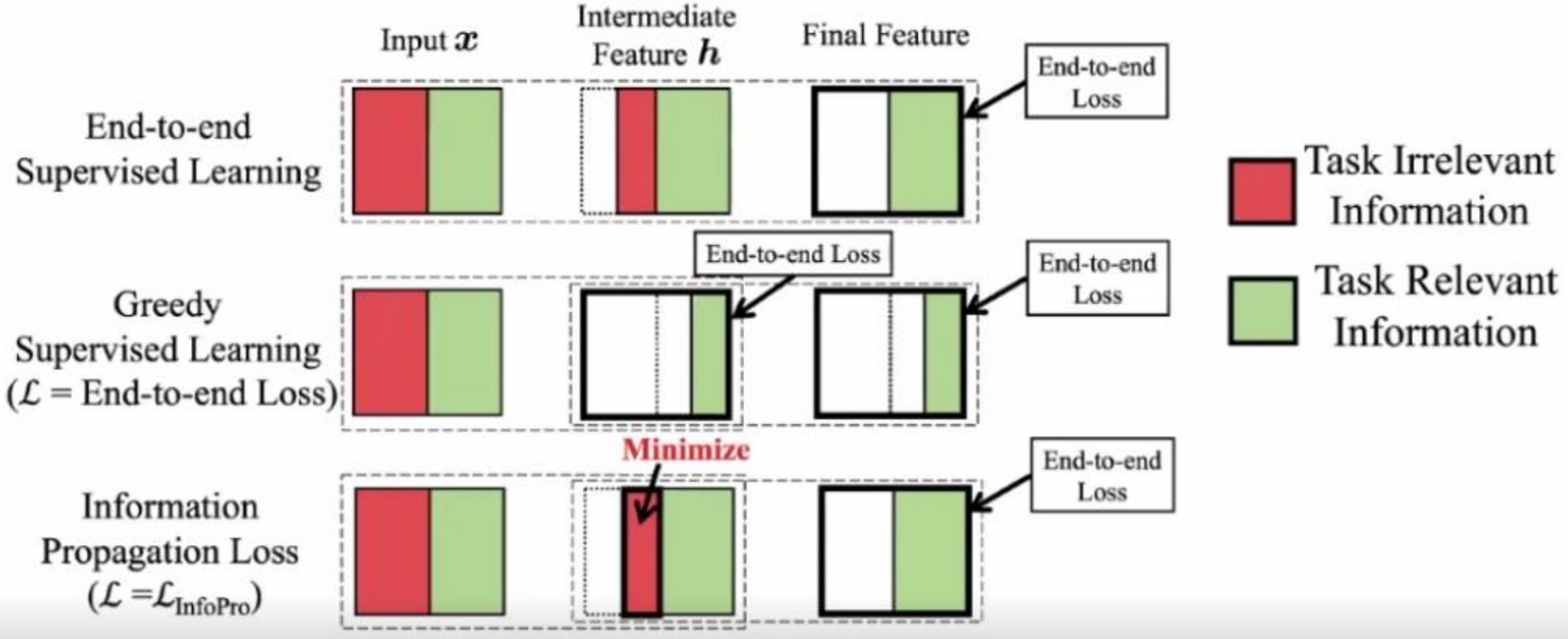
Method – *Information Propagation* (InfoPro) *Loss*

- Definition of Information Propagation (InfoPro) Loss

$$\mathcal{L}_{\text{InfoPro}}(\boldsymbol{h}) = \alpha[-\boxed{I(\boldsymbol{h},\boldsymbol{x})} + \beta\boxed{I(r^*,\boldsymbol{h})}],\ \alpha,\beta \geq 0, \quad s.t.\ r^* = \operatorname*{argmax}_{r,\ I(r,\boldsymbol{x})>0,\ I(r,y)=0} I(r,\boldsymbol{h}),$$

Propagate all information

Discard task-irrelevant information



(c) Comparison of Three Training Approaches

实验效果

Table 3: Trade-off between GPU memory footprint during training and test errors. Results of training ResNet-110 on a single Nvidia Titan Xp GPU are reported. 'GC' refers to gradient checkpointing (Chen et al., 2016).

Methods	CIFAR-10 (batch size = 1024)			STL-10 (batch size = 128)		
	Test Error	Memory Cost	Computational Overhead (Theoretical / Wall Time)	Test Error	Memory Cost	Computational Overhead (Theoretical / Wall Time)
E2E Training	$6.50 \pm 0.34\%$	9.40 GB	–	$22.27 \pm 1.61\%$	10.77 GB	–
GC (Chen et al., 2016)	$6.50 \pm 0.34\%$	3.91 GB ($\downarrow$58.4%)	32.8% / 27.5%	$22.27 \pm 1.61\%$	4.50 GB ($\downarrow$58.2%)	32.8% / 27.0%
InfoPro*, $K = 2$	$\mathbf{6.41 \pm 0.13\%}$	5.38 GB ($\downarrow$42.8%)	**1.3% / 1.1%**	$\mathbf{20.95 \pm 0.57\%}$	6.15 GB ($\downarrow$42.9%)	**1.3% / 1.7%**
InfoPro*, $K = 3$	$6.74 \pm 0.12\%$	4.22 GB ($\downarrow$55.1%)	3.3% / 7.5%	$21.00 \pm 0.52\%$	4.96 GB ($\downarrow$53.9%)	3.3% / 7.0%
InfoPro*, $K = 4$	$6.93 \pm 0.20\%$	**3.52 GB** ($\downarrow$62.6%)	5.9% / 13.4%	$21.22 \pm 0.72\%$	**4.08 GB** ($\downarrow$62.1%)	5.9% / 11.4%

Table 4: Single crop error rates (%) on the validation set of ImageNet. We use 8 Tesla V100 GPUs for training.

Models	Methods	Batch Size	Top-1 Error	Top-5 error	Memory Cost (per GPU)	Computational Overhead (Theoretical / Wall Time)
ResNet-101	E2E Training	1024	22.03%	5.93%	19.71 GB	–
	InfoPro*, $K = 2$	1024	**21.85%**	**5.89%**	**12.06 GB** ($\downarrow$38.8%)	5.7% / 11.7%
ResNet-152	E2E Training	1024	21.60%	5.92%	26.29 GB	–
	InfoPro*, $K = 2$	1024	**21.45%**	**5.84%**	**15.53 GB** ($\downarrow$40.9%)	3.9% / 8.7%
ResNeXt-101, 32×8d	E2E Training	512	20.64%	5.40%	19.22 GB	–
	InfoPro*, $K = 2$	512	**20.35%**	**5.28%**	**11.55 GB** ($\downarrow$39.9%)	2.7% / 5.6%

FPGA

- 简单了解了OpenCL
- 在虚拟机中用qemu虚拟机测试XRT和OpenCL
- 解决了实验时读取外部存储的数据文件的问题，而又不用自己写Linux驱动