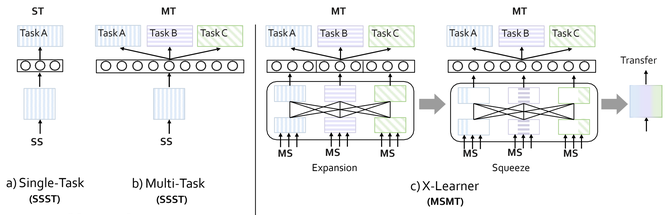
X-Learner: Learning Cross Sources and Tasks for Universal Visual Representation

Yinan He[1*], Gengshi Huang[2*], Siyu Chen[3*], Jianing Teng[4*], Kun Wang[4], Zhenfei Yin[4], Lu Sheng[5], Ziwei Liu[6], Yu Qiao[1] ✉, and Jing Shao[4]

* Indicates equal contribution

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory · SUN YAT-SEN UNIVERSITY · Carnegie Mellon University · ECCV TEL AVIV 2022 · sensetime · BEIHANG UNIVERSITY · NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

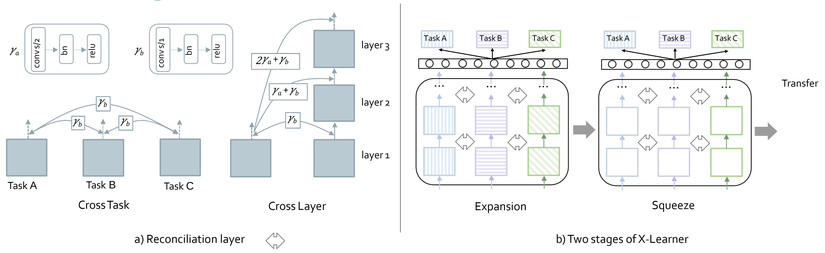
From Single Source to Multi Source



Our contributions:

- strong transfer ability of feature representations;
- several new insights into representation learning and the framework design;
- a new multi-source multi-task learning setting that only requires single-task label per datum;
- a general framework for universal representation learning from supervised multi-source and tasks

Two stage of X-Learner



Expansion Stage

- trains a set of sub-backbones
- combine their representational knowledge via *reconciliation layer*

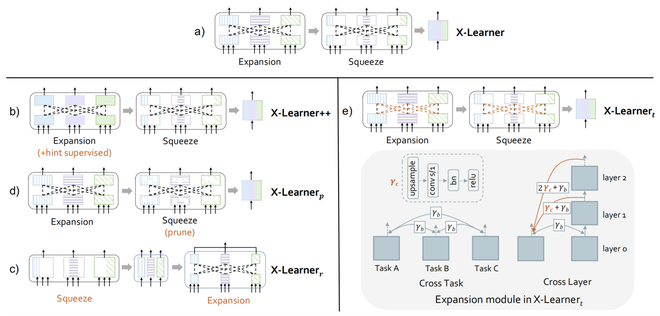
$$F_i^t = \mathcal{E}_i^t + \sum_{\substack{k=1 \\ k\neq t}}^{T} \sum_{j=1}^{i} \gamma_{j\to i}^{k\to t} \left(\mathcal{E}_j^k\right).$$

- Easy to train: use sub-tasks' best hyper-parameters

Squeeze Stage

- highly generalizable for downstream transfer
- Use FitNet for distillation $L_{\text{squeeze}} = \sum_{t=1}^{T} ||F^t - \mathcal{G}^t(\hat{F})||_2^2.$

Variants of X-Learner



(b) supervised by extra hints from single-task single-source pre-trained models.
(c) a Squeeze-Expansion version.
(d) replace the distillation with pruning in the squeeze stage
(e) new reconciliation layer in X-Learner..

Experiments and Observations

O1: Proper multi-task learning promotes collaboration instead of bringing interference

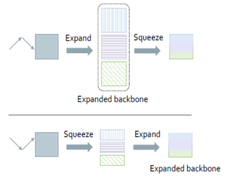
Method	AVG Cls	PASCAL Det	PASCAL Seg
ImageNet [54] Supervised	74.4	81.5	75.7*
SimCLR [10]	74.6	82.9	74.1*
Hard-sharing	73.2	83.7	70.5*
X-Learner	77.1 (+2.7)	84.4 (+2.9)	77.1* (+1.4)
X-Learner++	77.4 (+3.0)	**84.8** (+3.3)	**77.5*** (+1.8)
X-Learner w/ seg	**77.7** (+3.3)	84.3 (+2.8)	**77.6** (+1.9)

O2: Additional sources further improve multitask and multi-source representation learning if task conflicts are well-mitigated.

Expriments	Methods	Pre-train ImageNet	iNat2021	Places	*Cars*	*Dogs*	COCO	Objects365	*FACE*	Transfer AVG Cls	PASCAL Det
Base	Hard-sharing	75.0	75.3	53.0	–	–	35.5	17.4	–	73.2	83.7
	X-Learner	77.3	79.7	54.4	–	–	39.9	22.2	–	77.1	84.4
+ Cls Sources	Hard-sharing	73.7	73.6	52.3	98.5	85.3	35.4	17.6	–	77.5	83.1
	X-Learner	77.3	77.9	54.4	98.4	86.9	40.5	22.6	–	80.6	84.3
+ Cls & Det Sources	Hard-sharing	73.6	73.6	52.0	98.4	85.4	34.9	16.5	31.5	77.1	83.2
	X-Learner	76.9	78.6	54.6	98.6	85.9	40.1	22.1	33.6	80.5	84.3

O3: Expansion-Squeeze is better than Squeeze-Expansion.

Experiment	Sub-Backbone	Expansion	Squeeze	Pre-Distillation	Parameters
Hard-sharing	ResNet-50	×	×	×	→
X-Learner	ResNet-50	√	D	×	↘↘
X-Learner$_r$	HalfResNet-50	√	D	×	↘↘
X-Learner$_t$	ResNet-50	√	D	×	↘↘
X-Learner$_p$	ResNet-50	√	P	×	↗↘
X-Learner++	ResNet-50	√	D	√	↗↘
X-Learner w/o Rec.	ResNet-50	×	D	×	↘

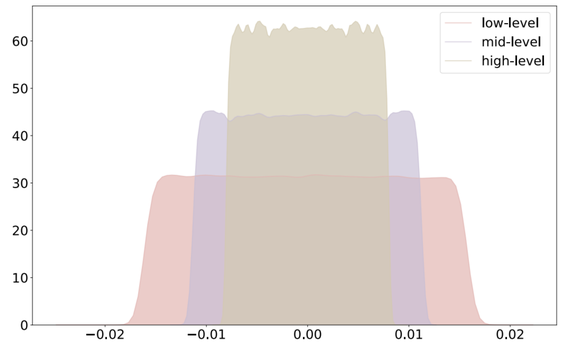


Observation 4: Reconciliation layers should receive information from lower levels.

Observation 5: Pruning may replace distillation in Squeeze Stage.

	AVG Cls	PASCAL Det
X-Learner w/o Rec	74.8	83.9
X-Learner	**77.1**	**84.4**

Method	Pre-train ImageNet	iNat2021	Places	COCO	Objects365	Transfer AVG Cls	PASCAL Det
Hard-sharing	75.0	75.3	53.0	35.5	17.4	73.2	83.7
X-Learner	77.3	79.7	54.4	39.9	22.2	77.1	84.4
X-Learner$_r$	73.9	76.6	52.5	41.1	21.7	73.9	84.1
X-Learner$_t$	76.3	79.9	53.3	42.5	22.0	74.5	83.5
X-Learner$_p$	76.1	78.6	53.5	42.4	23.4	77.2	83.1
X-Learner++	77.2	80.4	54.6	40.1	22.4	**77.4**	**84.8**



Observation 6: X-Learner has high data-efficiency.

Method	Backbone	Pre-training Settings	CIFAR-100 [33]	PASCAL Det [15]	PASCAL Seg [15]	NYU-Depth V2 [58]
MuST [19]	ResNet-152	ImageNet + DET. + SEG. + DEP.	86.3	85.1	80.6	87.8
MuST [19]	ResNet-152	JFT300M + DET. + SEG. + DEP.	88.3	87.9	**82.9**	89.5
X-Learner++	ResNet-50	ImageNet + DET.	87.0 (+0.7)	87.3 (+2.2)	78.8* (-1.8)	89.0 (+1.2)
X-Learner$_{R152}$	ResNet-152	ImageNet + OBJ365 + COCO	**88.7** (+2.4)	**88.5** (+3.4)	**81.4** (+0.8)	**91.2***(+3.4)
X-Learner$_{R152}$	ResNet-152	ImageNet + DET. + SEG.	**89.7** (+3.4)	**88.6** (+3.5)	**82.6** (+2.0)	**91.3***(+3.5)

Method	Backbone	Amount of Data	AVG Cls	CIFAR-100	PASCAL Det	PASCAL Seg	NYU-Depth V2
JFT-supervised	R152	300M	/	88.6	84.9	79.7	79.7
MuST	R152	302M	/	88.3	87.9	**82.9**	89.5
X-Learner	R152	**1.9M**	/	**88.7**	**88.5**	81.4	**91.2**
MoCo	R50	1B	/	/	82.2	73.6	/
BYOL	R50	300M	72.7	/	/	75.8	84.4
DnC [1]	R50	300M	76.3	/	/	76.9	86.1
X-Learner	R50	**13M**	**77.4**	**87.0**	**87.3**	**78.8**	**89.0**