Multi-task self-training for learning general representations

Authors: Golnaz Ghiasi*, Barret Zoph*, Ekin D. Cubuk*, Quoc V. Le, Tsung-Yi Lin
https://arxiv.org/abs/2108.11353

Presented by Eloy Geenjaar and Noah Lewis

Introduction

- Pre-training with a specialized task that is the same as your downstream task works the best

- Gathering a wealth of different types of labels for a dataset is challenging, time-consuming, and therefore expensive
  - We do want and need these types of datasets for the field of multi-task learning

Self-training

- Use a supervised model to generate pseudo labels on unlabeled data

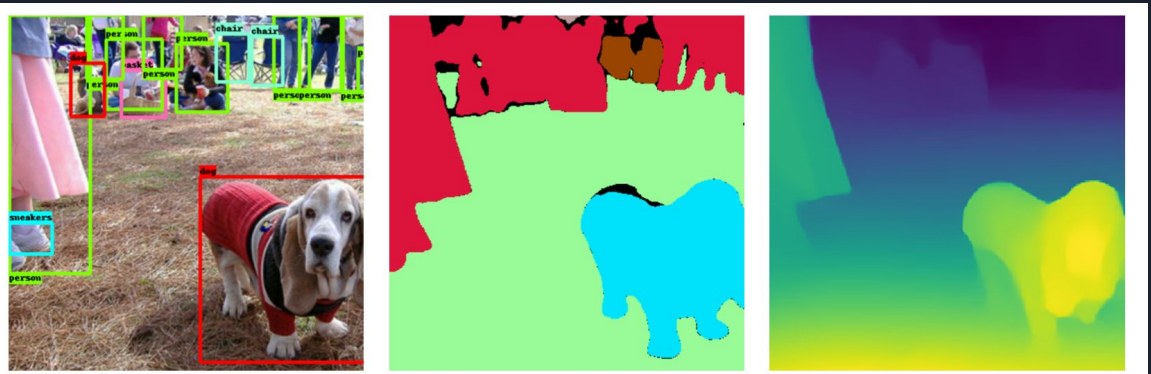

Figure 2. **Examples of pseudo labels on ImageNet. Left:** bounding boxes labeled with an Objects365 teacher model. **Middle:** semantic segmentation labeled with a COCO teacher model. **Right:** depth labeled with a MiDaS teacher model.

Representation learning

- Pre-training on a single task will likely not lead to general representations for other downstream tasks
- Self-supervised learning has become important, but:
  - Self-supervised representations may not be invariant to symmetries in natural images that are important for tasks such as semantic segmentation (e.g. viewpoint invariance).

Idea

- Train multiple teacher models on supervised datasets and use them to create pseudo labels for other datasets

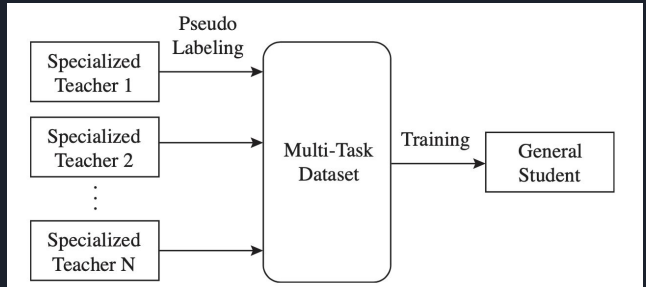

Figure 1. **An overview of Multi-Task Self-Training (MuST).** Specialized Teacher represents a supervised model trained on a single task and dataset (*e.g.*, classification model trained on ImageNet). Specialized Teacher models are trained independently on their own tasks and datasets. They then generate pseudo labels on a shared dataset. Finally, a single General Student model is trained jointly using the pseudo (and supervised) labels on the shared dataset.

Teacher models

1. Classification
2. Semantic segmentation
3. Object box detection
4. Depth estimation

Types of tasks

1. Classification
2. Object detection: object box detection
3. Pixel-wise prediction: semantic segmentation, depth estimation, and normal prediction

Architecture

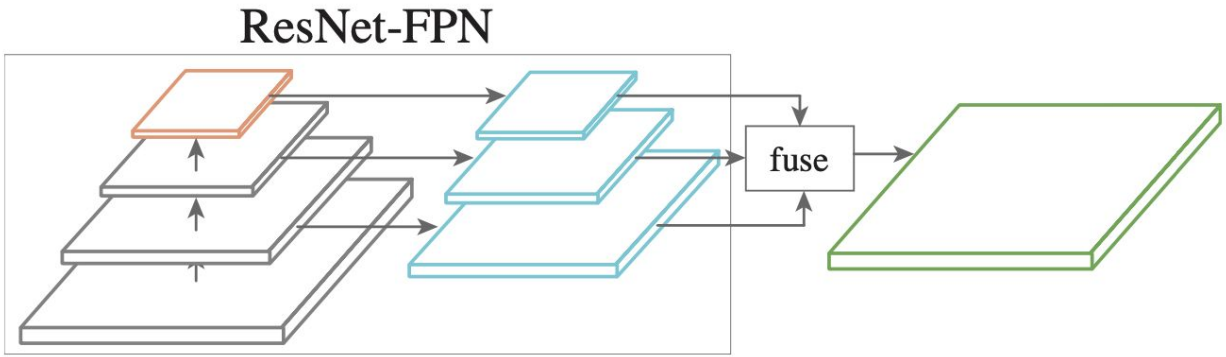

Figure 3. **The ResNet-FPN backbone architecture for multi-task learning.** **Orange:** the top-level features for classification. **Cyan:** multi-scale features for box detection and instance segmentation. **Green:** the high resolution features for pixel-wise tasks (*e.g.*, segmentation, depth, and surface normal estimation.)

Loss function

$$L = \sum_i w_i L_i$$

For ImageNet

$$w_i = \frac{b_s lr_{it}}{b_{it} lr_s}$$

Difference with other models

- Supervised and pseudo-labels are treated equally
- Uniformly sample from each dataset (because all data is (psuedo-)labeled for each task)

Datasets/fine-tuning

- End-to-end finetuning
- CIFAR-100: classification
- Pascal detection: object box detection
- Pascal: semantic segmentation
- NYU: depth prediction

Datasets/fine-tuning

Training Datasets			Evaluation Datasets		
Name	Task	Num Images	Name	Task	Num Images
ImageNet [47]	Classification	1.2M	CIFAR-100 [31]	Classification	50k
Objects365 [49]	Detection	600k	Pascal [13]	Detection	16.5k
COCO [37]	Segmentation	118k	Pascal [13]	Segmentation	1.5k
MiDaS [44]	Depth	1.9M	NYU V2 [50]	Depth	47k
JFT [51]	Classification	300M	ADE [66]	Segmentation	20k
			DIODE [54]	Surface Normal	17k

Table 1. **Datasets using for MuST and for downstream fine-tuning evaluation.**

Results: pre-training

Settings		Transfer Learning Performance					
Method	Epochs	CIFAR-100 Classification	Pascal Detection	Pascal Segm.	NYU Depth	ADE Segm.	DIODE Normal
Self-supervised (SimCLR [5])	800	**87.1**	83.3	72.2	83.7	41.0	**52.8**
ImageNet Supervised	90	85.4	79.3	70.6	81.0	39.8	48.9
+ Multi-task Pseudo Labels	90	86.3 (+0.9)	**85.1** (+5.8)	**80.6** (+10.0)	**87.8** (+6.8)	**43.5** (+3.7)	**52.7** (+3.8)

Table 2. **Multi-Task Self-Training (MuST) outperforms supervised and self-supervised representations on ImageNet**. We compare MuST to state-of-the-art self-supervised and supervised learning using the same pre-training dataset (ImageNet). MuST learns more general features and achieves the best performance on $4/6$ downstream fine-tuning tasks. The performance differences show the impact of different training objectives.

Results: more tasks -> general features?

Settings	Transfer Learning Performance					
Method	CIFAR-100 Classification	Pascal Detection	Pascal Segm.	NYU Depth	ADE Segm.	DIODE Normal
ImageNet Supervised	85.4	79.3	70.6	81.0	39.8	48.9
+ Depth Pseudo Labels	84.4(-1.0)	79.3(+0.0)	71.0(+0.4)	86.0(+5.0)	39.5(-0.3)	51.3(+2.4)
+ Depth / Segm. Pseudo Labels	85.3(-0.1)	81.6(+2.3)	78.6(+8.0)	87.2(+6.2)	41.5(+1.7)	52.4(+3.5)
+ Depth / Segm. / Detection Pseudo Labels	**86.3**(+0.9)	**85.1**(+5.8)	**80.6**(+10.0)	**87.8**(+6.8)	**43.5**(+3.7)	**52.7**(+3.8)

Table 3. **Multi-Task Self-Training (MuST) benefits from increasing the number of different pseudo label tasks.** We add depth, segmentation, and detection pseudo labels in addition to supervised ImageNet classification labels and test the representational quality. The results reveal that adding pseudo labels from more tasks leads to more general pre-trained models. All models are trained for 90 epochs on ImageNet.

Results: transfer learning of teacher model

Settings		Performance		Transfer Learning Performance					
Task	Train Dataset	Obj365 Detection	COCO Segm.	CIFAR-100 Classification	Pascal Detection	Pascal Segm.	NYU Depth	ADE Segm.	DIODE Normal
		Teacher Model							
Detection	Objects365	26.1	—	84.0	87.6	78.8	90.1	46.0	55.6
Segmentation	COCO	—	53.8	80.8	82.2	80.2	86.6	42.8	51.0
		Student Model							
Detection	ImageNet	20.6	—	83.2	86.0	78.5	88.5	44.7	55.2
Detection	JFT	20.7	—	85.2	87.7	79.5	89.6	45.4	55.0
Segmentation	ImageNet	—	55.5	82.3	80.5	79.2	86.3	41.8	51.2
Segmentation	JFT	—	49.0	83.1	82.8	78.2	86.6	41.9	51.6

Table 4. **Models trained on supervised data or pseudo labeled data have similar transfer learning performance.** Results comparing how representations transfer if they are trained on supervised data or on pseudo labels that are generated by the supervised model. Pseudo labels effectively compress the knowledge in a supervised dataset. The performance of student models increases with the size of the unlabeled dataset. As the unlabeled dataset size increased, the performance of student model increases. This reveals the scalability of MuST. All student models are trained for the same training iterations (90 ImageNet epochs and 0.36 JFT epochs).

Results: across datasets

Settings	Transfer Learning Performance					
Method	CIFAR-100 Classification	Pascal Detection	Pascal Segm.	NYU Depth	ADE Segm.	DIODE Normal
Supervised Multi-Task	85.3	85.1	82.1	87.6	43.9	53.4
Supervised Multi-Task + Pseudo Labels	**86.3** (+1.1)	**86.2** (+1.1)	**82.3** (+0.2)	**88.2** (+0.6)	**45.4** (+1.5)	**54.7** (+1.3)

Table 5. **Comparing Multi-Task Training versus Multi-Task Self-Training.** We compare MuST against a baseline of doing supervised multi-task training on the union of all teacher datasets. We use three datasets: ImageNet, COCO and Objects365. Supervised model is jointly trained on the supervised labels of these three datasets. MuST trains jointly on all three supervised and pseudo labels generated by the teacher models. The transfer learning performance gets strong improvements by incorporating pseudo labels into every image.

Results: scaling

Settings		Transfer Learning Performance					
Method	Epochs	CIFAR-100 Classification	Pascal Detection	Pascal Segm.	NYU Depth	ADE Segm.	DIODE Normal
Self-Supervised with JFT images (SimCLR [5])	1	85.6	82.4	71.0	83.7	41.4	54.4
Self-Supervised with JFT images (SimCLR [5])	2	85.8	83.7	73.3	84.3	42.2	55.3
Self-Supervised with JFT images (SimCLR [5])	5	86.1	84.1	74.9	84.8	43.0	56.0
JFT supervised	3	87.7	84.6	78.2	86.0	43.4	50.7
JFT supervised	5	88.6	84.9	79.7	86.1	44.3	51.1
JFT supervised	10	**89.6**	85.2	80.4	86.5	45.7	53.1
Multi-Task Pseudo Labels	2.5	87.6	87.8	82.2	**89.8**	47.0	56.2
JFT supervised + Multi-Task Pseudo Labels	2	88.3(+0.5)	**87.9**(+0.1)	**82.9**(+0.7)	89.5(-0.3)	**47.2**(+0.2)	**56.4**(+0.2)

Table 6. **Scaling Multi-Task Self-Training to 300M images.** We repeat the experiments in Table 2 on the JFT dataset (300M images with classification labels). The supervised learning benefits more from the additional images and annotations compared to the self-supervised SimCLR algorithm.

Results: bootstrap pre-trained models

Settings	Transfer Learning Performance			
Method	Pascal Detection	Pascal Segm.	NYU Depth	ADE Segm.
Previous SoTA	**89.3** [14]	**90.0** [67]	90.4 [43]	**54.1** [8]
ALIGN [24]	86.2	86.6	91.1	54.0
MuST w/ [24]	88.2	**89.8**	**91.9**	**54.3**

Table 7. **MuST checkpoints are versatile and achieve competitive performance compared to state-of-the-art models.** MuST improves the transfer learning performance of the ALIGN EfficientNet-L2 checkpoint on these four downstream tasks.

Visualizations

Figure 4. **Visualization of the predictions generated by a multi-task student model.** The MuST student model not only learns general feature representations, but also makes high quality visual predictions with a single model.

Transfer learning improvement

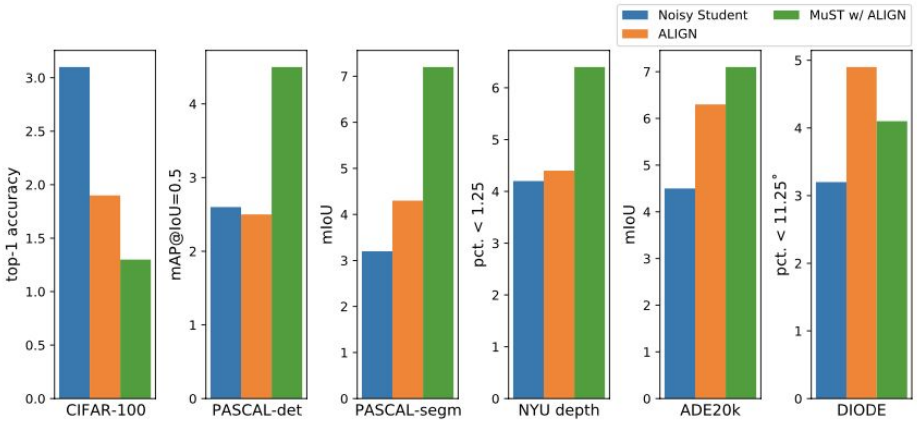

Figure 5. **Relative transfer learning performance gains over the ImageNet pre-trained model [52].** Checkpoints trained with more data or labels typically provide gains on transfer learning to downstream tasks. Fine-tuning the EfficientNet-B7 ALIGN checkpoint with MuST can further improve transfer learning performance for 4/6 downstream tasks.

Discussion

- Self-supervised learning can now outperform supervised pre-training
- Self-training is complementary and currently performs better
    - Combining the two is an important future field

Conclusion

- It works