Supplementary Materials: Multi-Task Self-Training for Learning General Representations

Golnaz Ghiasi*, Barret Zoph*, Ekin D. Cubuk*, Quoc V. Le, Tsung-Yi Lin
Google Research, Brain Team
{golnazg,barretzoph,cubuk,qvl,tsungyi}@google.com

1. Details of Training and Evaluation Datasets

1.1. Training Datasets

In this section, we describe 5 datasets we use to train teacher models.

ImageNet: ImageNet [9] is a classification dataset with 1.2M training images and 1000 unique classes. All of its images are center cropped and have one primary object per image.

Objects365: Objects365 [10] is an object detection dataset that has 365 different classes and 600k training images.

COCO: The COCO dataset [7] contains 118k images that has a variety of different labels (e.g. object detection, instance segmentation, panoptic segmentation). For all experiments we use its pantopic segmentation labels.

MiDaS: The MiDaS depth model [8] that is used for generating our depth pseudo labels is trained on a diverse set of 5 depth datasets. The 5 depth datasets are DIML Indoor [3] (220k images), MegaDepth [5] (130k images), ReDWeb [15] (3600), WSVD [14] (1.5M), and 3D movies (75k). The model is trained to be invariant to the depth range and scale across all datasets, leading to a model that generates robust pseudo labels.

JFT: JFT [12] is a large-scale image multi-label classification dataset with 300M labeled images. This dataset is used to test the scale of MuST and various self-supervised learning algorithms.

1.2. Evaluation Datasets

Next we describe the datasets that all of our representations will be fine-tuned on. We have different datasets with a total of five different tasks. Note the Surface Normal task is never used as a training task to test the task generality of the representations.

CIFAR-100: CIFAR-100 is a classification dataset with 50k images and 100 unique classes.

PASCAL Detection: The Pascal Detection dataset [2] is an object detection dataset with 20 unique classes. We train the models on the `trainval` sets of PASCAL VOC 2007 and PASCAL VOC 2012 which include 16.5k images.

PASCAL Segmentation: The Pascal Segmentation dataset [2] is a semantic segmentation dataset with 20 unique classes. We train the models on the `train` set of the PASCAL VOC 2012 segmentation dataset which has 1.5k images.

NYU Depth V2: The NYU Depth v2 dataset [11] is a depth estimation dataset that contains 47584 train images and 654 validation images.

ADE Segmentation: ADE20k [16] is a segmentation dataset that contains 20k images with 150 object and stuff classes. The dataset contains a wide variety of different indoor and outdoor scenes along with object classes.

DIODE Surface Normal: The DIODE dataset [13] is a depth and surface normal dataset that contains 16884 images. The dataset contains a diverse set of both indoor and outdoor scenes for training and testing. We only make use of the surface normal labels.

2. Implementation Details

2.1. Training Teacher Models

In this section, we introduce the details of training teacher models, which are used to generate pseudo labels in

*Authors contributed equally.

MuST. All the models are trained with a ResNet-152 backbone model.

Objects365 Detection: We use batch size 256 and train for 140 epochs. The image size is 640. We apply scale jittering [0.5, 2.0] (i.e. randomly resample image between 320×320 to 1280×1280 and crop it to 512×512). The learning rate is 0.32 and the weight decay is set as 4e-5. The model is trained from random initialization. The final performance is 26.1 AP.

COCO Segmentation: We use the annotations in COCO panoptic segmentation dataset [4]. We train a semantic segmentation model that only predicts the semantic class for each pixel, instead of also predicting the object instance. We use batch size 256 and train for 384 epochs. The image size is 896. We apply scale jittering [0.5, 2.0]. The learning rate is 0.32 and the weight decay is set as 4e-5. The model is trained from random initialization. The final performance is 53.8 mIoU.

MiDaS Depth: We directly download the pre-trained MiDaS from the github repository and use it as a teacher model to generate pseudo labels.

ImageNet Classification: We use batch size 2048 and train for 400 epochs. The image size is 224. The learning rate is 0.8 and weight decay is 4e-5. We apply random augmentation [1] (2L-15M, 2 layers with magnitude 15) and label smoothing (0.1) to regularize the model training. The final performance is 81.6 top-1 accuracy.

2.2. Training Multi-Task Student Models

We use a batch size 256 for training student models in our experiments. The image size is 640. We apply scale jittering [0.5, 2.0] during training. The weight decay is 4e-5 in ImageNet experiments and 3e-6 in JFT experiments. No random augmentation [1] or label smoothing is applied.

2.3. Fine-tuning on Evaluation Datasets

For fine-tuning we initialize the parameters in the ResNet and FPN backbone with a pre-trained model and randomly initialize the rest of the layers. We perform *end-to-end* fine-tuning with an extensive grid search of the combinations of learning rate and training steps to ensure each pre-trained model achieves its best fine-tuning performance. We experiment with different weight decays but do not find it making a big difference and set it to 1e-4. All models are trained with cosine learning rate for simplicity. Below we describe the dataset, evaluation metric, model architecture, and training parameters for each task.

CIFAR-100: We use standard CIFAR-100 train and test sets and report the top-1 accuracy. We resize the image resolution to 256×256. We replace the classification head in the pre-trained model with a randomly initialized linear layer that predicts 101 classes, including background. We use a batch size of 512 and search the combination of training steps from 5000 to 20000 and learning rates from 0.005 to 0.32. We find the best learning rate for SimCLR (0.16) is much higher than the supervised model (0.005). This trend holds for the following tasks.

PASCAL Segmentation: We use PASCAL VOC 2012 train and validation sets and report the mIoU metric. The training images are re-sampled into 512×512 with scale jittering [0.5, 2.0]. We initialize the model from the pre-trained backbone and FPN [6] layers. We remove the pre-trained segmentation head and train from a randomly initialized head. We use a batch size of 64 and search the combination of training steps from 5000 to 20000 and learning rates from 0.005 to 0.32.

PASCAL Detection: We use PASCAL VOC 2007+2012 trainval set and VOC 2007 test set and report the AP_{50} with 11 recall points to compute average precision. The training images are resampled into 896 with scale jittering [0.5, 2.0]. we initialize the model from the pre-trained backbone and FPN [6] layers and randomly initialize the heads. We use a batch size of 32 and search the combination of training steps from 5000 to 20000 and learning rates from 0.005 to 0.32.

NYU Depth: We use NYU depth v2 dataset with 47584 train and 654 validation images. We report the percentage of predicted depth values within 1.25 relative ratio compared to the ground truth. The training images are resampled into 640 with scale jittering [0.5, 2.0]. we initialize the model from the pre-trained backbone and FPN [6] layers and randomly initialize the heads. We use a batch size of 64 and search the combination of training steps from 10000 to 40000 and learning rates from 0.005 to 0.32.

DIODE: We use DIODE outdoor dataset with 16884 train and 446 validation images. We report the percentage of the angle error less than 11.25°. We use the original image resolution 768 for training and evaluation. The training image is applied with scale jittering [0.5, 2.0]. we initialize the model from the pre-trained backbone and FPN [6] layers and randomly initialize the heads. We use a batch size of 32 and search the combination of training steps from 20000 to 80000 and learning rates from 0.01 to 0.16.

3. Visualization of Student Model Predictions

Figure 1 shows more visual examples of the predictions made by a single multi-task student model. The images are sampled from the validation set in ImageNet dataset.

References

[1] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space, 2019. 2

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 1

[3] Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE transactions on Image Processing*, 27(8):4131–4144, 2018. 1

[4] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *CVPR*, June 2019. 2

[5] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 1

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[8] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[10] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 1

[11] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1

[12] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017. 1

[13] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 1

[14] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *2019 International Conference on 3D Vision (3DV)*, pages 348–357. IEEE, 2019. 1

[15] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *CVPR*, 2018. 1

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127(3):302–321, 2019. 1

Figure 1. The visualization of inference on ImageNet dataset made by single multi-task student model.