Supplementary of
Towards Inadequately Pre-trained Models in Transfer Learning

Andong Deng[1,†], **Xingjian Li**[2,†], **Di Hu**[3,*], **Tianyang Wang**[4], **Haoyi Xiong**[2] **Cheng-Zhong Xu**[5],
[1]University of Central Florida [2]Baidu Research [3]Renmin University of China
[4]University of Alabama at Birmingham [5]University of Macau

1. Datasets

CIFAR10 [6] and **CIFAR100** [6] are two fundamental datasets in computer vision community. Both of them contain 50,000 training samples and 10,000 test samples, and all the samples are evenly distributed in each category. CIFAR10 consists of 10 common classes of objects. CIFAR100 includes 10 superclasses and each superclasses is made up of 10 fine-grained categories, and the size of each sample is 32×32. **Food-101** [1] is a challenging food classification dataset, which consists of 101 categories. There are 250 clean test images for each class and 750 training images containing some noisy labels. **FGVC Aircraft** [7] is a fine-grained dataset for aircraft classification. It contains 10,000 images of 100 categories of aircraft, and the training set is 2/3 of the whole dataset. **Stanford Cars** [5] contains 196 classes of fine-grained cars, and there are 8,144 and 8,041 samples in the training set and test set, respectively. **CUB-200-2011** [11] is a fine-grained bird classification dataset containing 200 species. There are 11,788 training samples and 5,894 test samples. Annotation of the bounding box, rough segmentation, and attributes are provided. **Oxford 102 Flowers** [8] contains 200 common species of flowers in United Kingdom. Each of the categories has 40 up to 258 images. There are 2,040 training samples as well as 6,149 test samples. **MIT Indoor 67** [9] contains 67 indoor scene categories with in total of 15,620 images, and 80% images are used for training. **MURA** [10] is a dataset of musculoskeletal radiographs, containing 40,561 X-ray images from 14,863 patient studies. The goal is to distinguish normal musculoskeletal examples from abnormal ones. We follow the common setting to perform binary classification on each image.

2. Experimental setting of ResNet50

Pre-training We borrow the official PyTorch implementation for ImageNet training using ResNet50. The total number of training epochs is set to 90. Stochastic gradient descent with a momentum of 0.9 is used to update the model parameters. The initial learning rate is 0.1 and is multiplied by 0.1 every 30 epochs. The weight decay is 1e-4. The pre-training performance is shown in Table 1.

Table 1. Pre-training performance of ResNet50 on ImageNet.

Epoch	20	30	40	50	60	70	80	90
Acc%	52.38	55.13	70.60	70.44	70.78	75.63	75.76	76.06

Transfer learning In transfer learning, we use different training configurations to adapt to different datasets. For CIFAR10 and CIFAR100, in both FE and FT, the total training epoch is set as 150. The initial learning rate is 0.1 and is decayed by 10 times every 50 epochs. The optimizer is Adam[4]. For the rest natural datasets, we run 6,000 iterations and 9,000 iterations for FE and FT, respectively; the learning rate is set to 0.1 for FT and 0.01 for FE.

3. Experimental setting of T2T-ViT_t-14

Pre-training We perform pre-training following the official codes of T2T-ViT [12]. Specifically, we train T2T-ViT_t-14 on ImageNet for 300 epochs. The final model achieves a Top-1 accuracy of 81.55% on the ImageNet validation set. We choose checkpoints on epoch [20,40,60,80,100,120,150,200,250,300] for transfer learning experiments.

Transfer learning For transfer learning, we perform the same data processing pipeline as used in ResNet50. For sufficient adaptation, the initial learning rate is set to 0.05 and decayed by a cosine annealing strategy, as suggested by the T2T-ViT paper.

4. Additional results

Additional FE evaluation results on DTD [2] and Caltech256 [3] are shown in table 2, which are also consistent with our claims. To further consolidate our conclusion, we

also add the FE results of Swin-T in Table 3. We train Swin-T for 300 epochs following the default settings in the official repository. The results of Swin-T clearly validate our conclusion that the best feature extractors are those inadequately pre-trained models.

Table 2. FE evaluation of DTD and Caltech256.

Pre-training Epoch	40	50	60	70	80	90
DTD (%)	69.36	69.15	69.36	**70.43**	68.88	69.84
Caltech256 (%)	79.58	80.06	79.58	80.92	**81.29**	80.89

Table 3. Pre-training and FE performance on Swin-T.

Epoch	240	250	260	270	280	290	300
pre-train %	79.89	80.20	80.51	80.88	81.00	81.10	**81.21**
CIFAR100 %	74.95	**75.72**	75.30	75.02	75.43	75.33	75.58
DTD %	**68.99**	68.46	68.51	68.88	68.83	68.88	68.46
Flower102 %	**86.39**	86.24	85.95	85.93	85.75	86.37	85.61
Aircraft %	44.47	45.47	43.97	**45.92**	44.69	45.53	43.94
Indoor %	79.81	79.28	80.31	**81.76**	80.19	81.26	80.04

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 1

[3] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[5] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 1

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[7] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[8] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 1

[9] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009. 1

[10] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. *Hand*, 1(602):2–215. 1

[11] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1

[12] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021. 1