Supplement to:
iLab20M: A large-scale controlled object dataset to investigate deep learning

Ali Borji[†] Saeed Izadi[‡] Laurent Itti[§]
[†]Center for Research in Computer Vision, University of Central Florida
[‡]Amirkabir University of Technology, [§] University of Southern California
aborji@crcv.ucf.edu, sizadi@aut.ac.ir, itti@usc.edu

This documents supplements the Borji et al., published in CVPR 2016 [2]. We use t-SNE dimensionality reduction method [3] to visualize the learned representations

Experiment I: category prediction In this experiment, we randomly select 2K samples from 7 categories (boat, bus, f1car, tank, train, ufo, and van) and feed them to a pre-trained CNN model, specifically Alexnet. Having fc7 and pool5 representations of selected samples ready, we use the t-SNE algorithm to reduce their dimensionality to 2D.

In addition, 20K images are randomly selected from all 7 categories and the network is fine-tuned on the provided data for object categorization. The same procedure is carried out on the fine-tuned (FT) network. Fig. 1 depicts the results.

Our results in Fig. 1 show that fc7 representation works remarkably well at recognizing object level categories as they are mutually linearly separable after fine-tuning the network. Furthermore, pool5 representation does not contain discriminative information between object categories compared to fc7. This result is in alignment with Bakry et al., [1]. Fig. 1 also demonstrates the effect of fine-tuning on feature spaces. The distributions of samples for different categories tend to become very compact and concentrated after fine-tuning. Notice that fine-tuning does not add more discriminative power to the pool5 representation.

Experiment II: rotation prediction This experiment makes effort to highlight the power of pool5 layer in representing image variations and discriminating among them. As we discussed in the main paper, our analyses show that pool5 representation gives superior performance for parameter prediction. To confirm this statement, we select 200 samples from the boat category (and instance number 01) while rotation, camera, and lighting parameters are changing. We then label the samples with their rotation values and feed them to the pre-trained Alexnet model. The dimensionality of fc7 and pool5 representations are reduced to 2D using tSNE. The same procedure is carried out using the fine-tuned network to obtain the fc7 and pool5 representations. Results are illustrated in Fig. 2.

It can be seen that fc7 representation is not (fully) capable of discriminating the rotation values, both with and without fine-tuning. The representation by the pool5 layer, in contrast, confirms our findings that pool5 contains information selective to parameters. Samples from 8 different rotation values are perfectly and mutually linearly separable from each other. Fine-tuning tries to improve the discriminability through some sort of transformation.

Experiment III: camera prediction With our success in visualizing the power of pool5 layer in capturing rotation variations, in this experiment we aim to see whether the same judgment is valid for camera prediction. As in the previous experiment, we select 200 samples from the boat category (instance number 01) and label them according to their camera parameter value. 2D feature spaces derived from fc7 and pool5 representations using pre-trained and fine-tuned Alexnet are depicted in Fig. 3.

As before, fc7 representation does not offer useful information regarding separating samples with different camera parameters, both in pre-trained and fine-tuned cases. We observe quite the opposite using the pool5 layer representation. Without fine-tuning the network, we can observe 8 clusters in Fig. 3 (see the up-right panel), each one corresponding to one rotation. For each rotation angle, the representation is surprisingly capable of discriminating different values of camera parameters in five classes (we only use five values for camera parameter here).

Experiment IV: lighting prediction Scrutinizing the behavior of fc7 and pool5 layers should be interesting for lighting prediction as well. Therefore, we follow the previous experiments except that here samples are labeled according to the lighting parameter values. Fig. 4 shows the results for four different cases.

Skipping the poor representation by fc7 layer, pool5 layer again generates reasonable representation which is able to discriminate between different lighting conditions. Eight clusters are observable, each one corresponding to

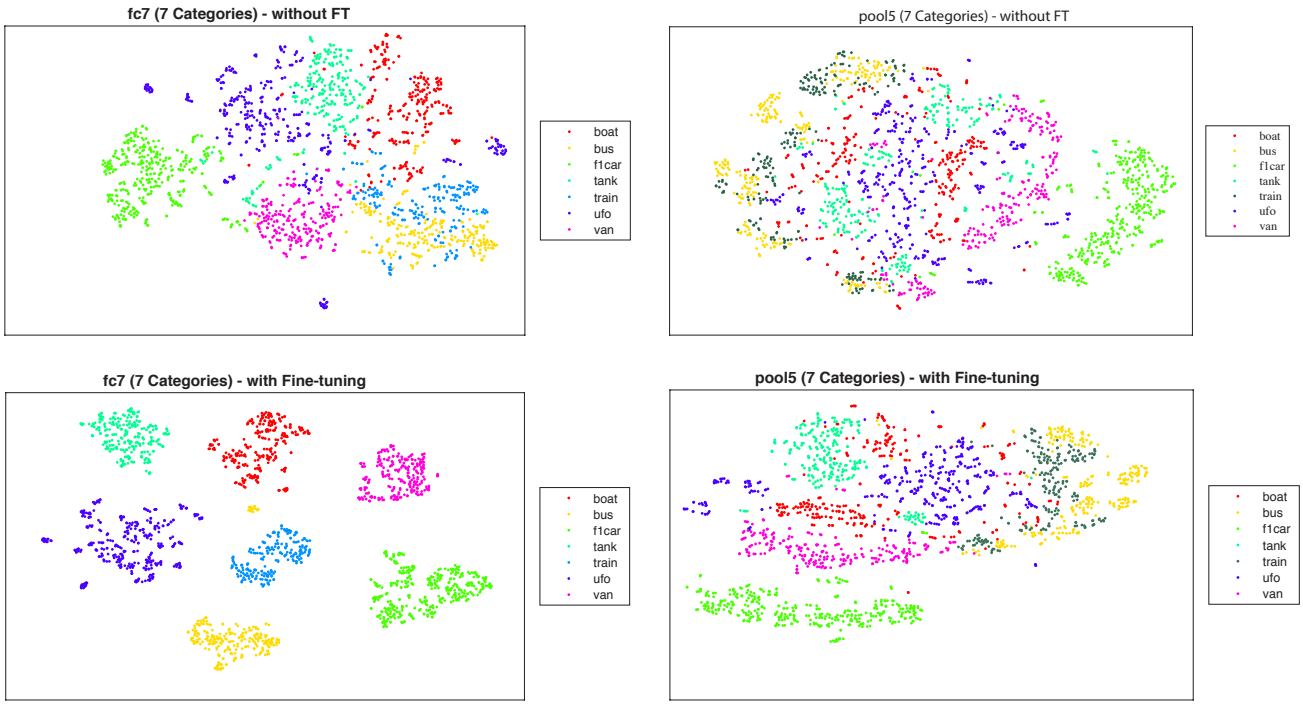

Figure 1. t-SNE representation for category prediction using fc7 and pool5 layers with and without fine-tuning.

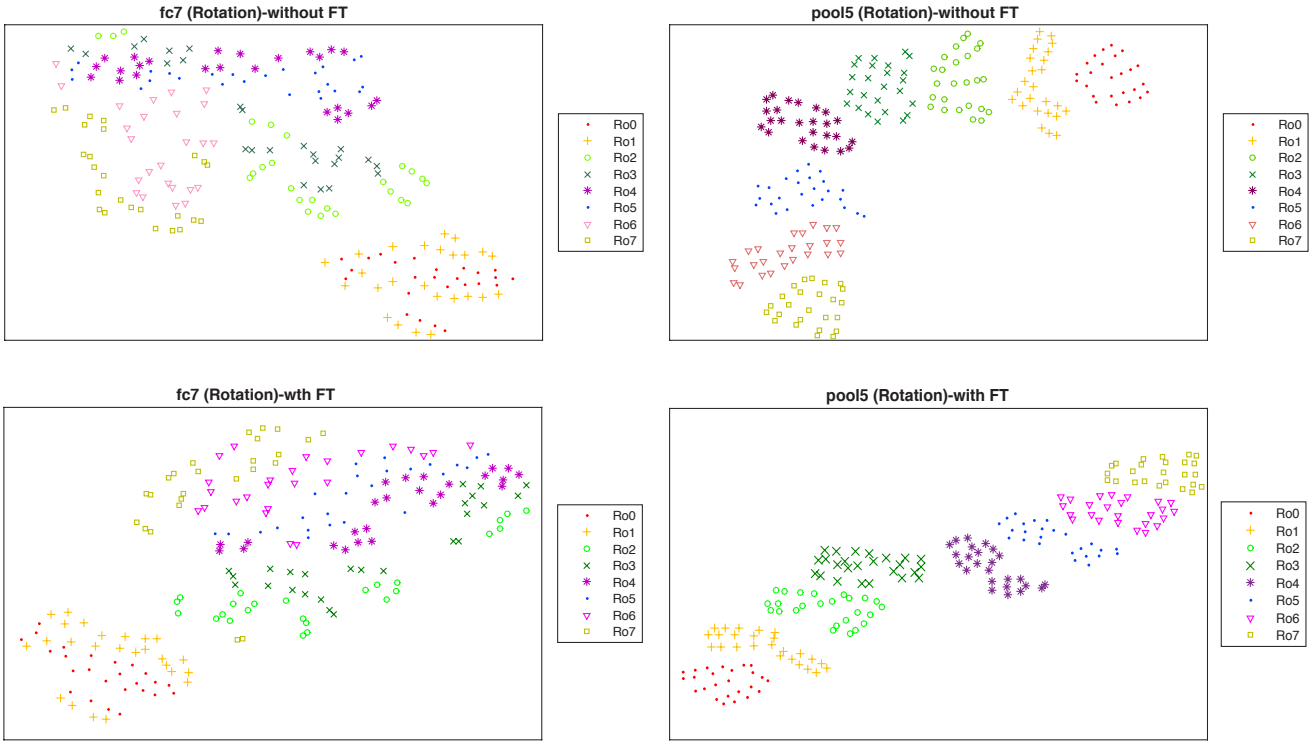

Figure 2. t-SNE representation for lighting prediction using fc7 and pool5 layers with and without fine-tuning.

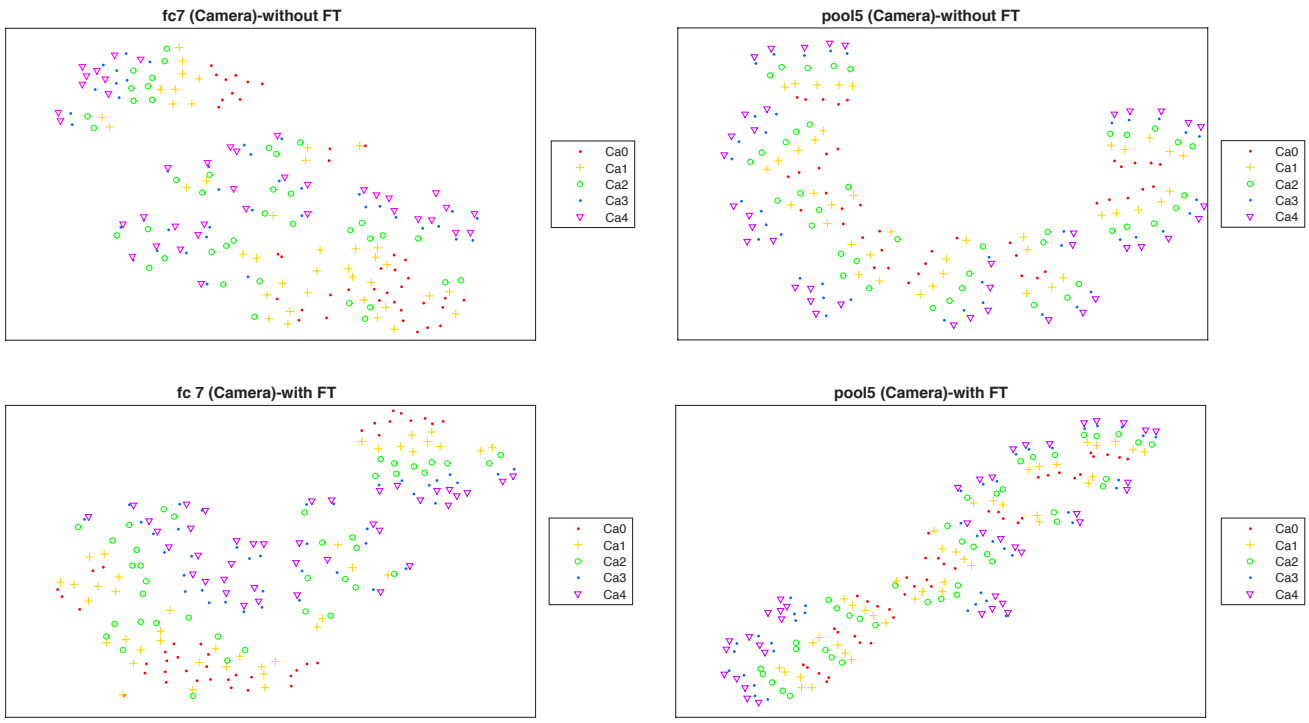

Figure 3. t-SNE representation for camera prediction using fc7 and pool5 layers with and without fine-tuning.

one rotation angle. In each cluster, samples with different lighting parameters are discriminant which again supports our previous statement regarding the capability of the pool5 layer in parameter prediction.

Experiment V: instance prediction In the last experiment, we aim to inspect the capacity of fc7 and pool5 layers of CNNs for instance prediction. We randomly choose 2K samples from the boat category. The samples are passed through the network up to pool5 and fc7 layers. The obtained representations are visualized after dimensionality reduction using the tSNE. The same procedure is repeated with the fine-tuned network. Fig. 5 show the results.

The fc7 representation, without fine-tuning, is remarkably capable to separate samples from different instances. Fine-tuning the network dramatically boosts this discrimination power by making clusters more compact. A representation is invariant to varying parameters if it ignores variations and treats samples with different parameters equally, i.e., it makes the representations of similar samples as close as possible in the feature space. This is exactly what we see in the in the representation space provided by fc7.

Despite the reasonable parameter separability, the pool5 layer does not force different instances to be clustered. This is the place where difference between pool5 and fc7 layers can be seen in practice. This result indicates that the fc7 layer seeks to produce invariant representations (by collapsing manifolds), while the pool5 layer tries to preserve manifolds as much as possible.

References

[1] A. Bakry, M. Elhoseiny, T. El-Gaaly, and A. Elgammal. Digging deep into the layers of cnns: In search of how cnns achieve view invariance. *arXiv:1508.01983*, 2015. 1

[2] A. Borji, S. Izadi, and L. Itti. What can we learn about cnns from a large scale controlled object dataset? *arXiv preprint arXiv:1512.01320*, 2015. 1

[3] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008. 1

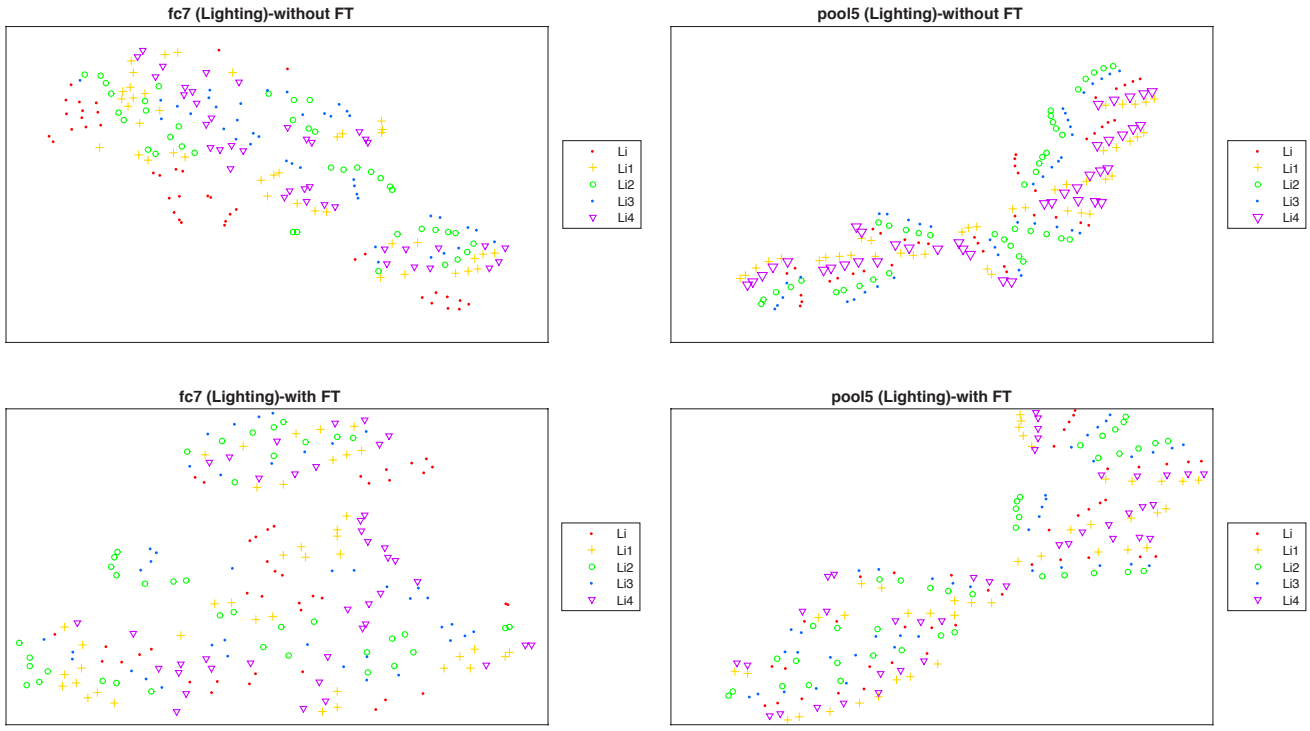

Figure 4. t-SNE representation for lighting prediction using fc7 and pool5 layers with and without fine-tuning.

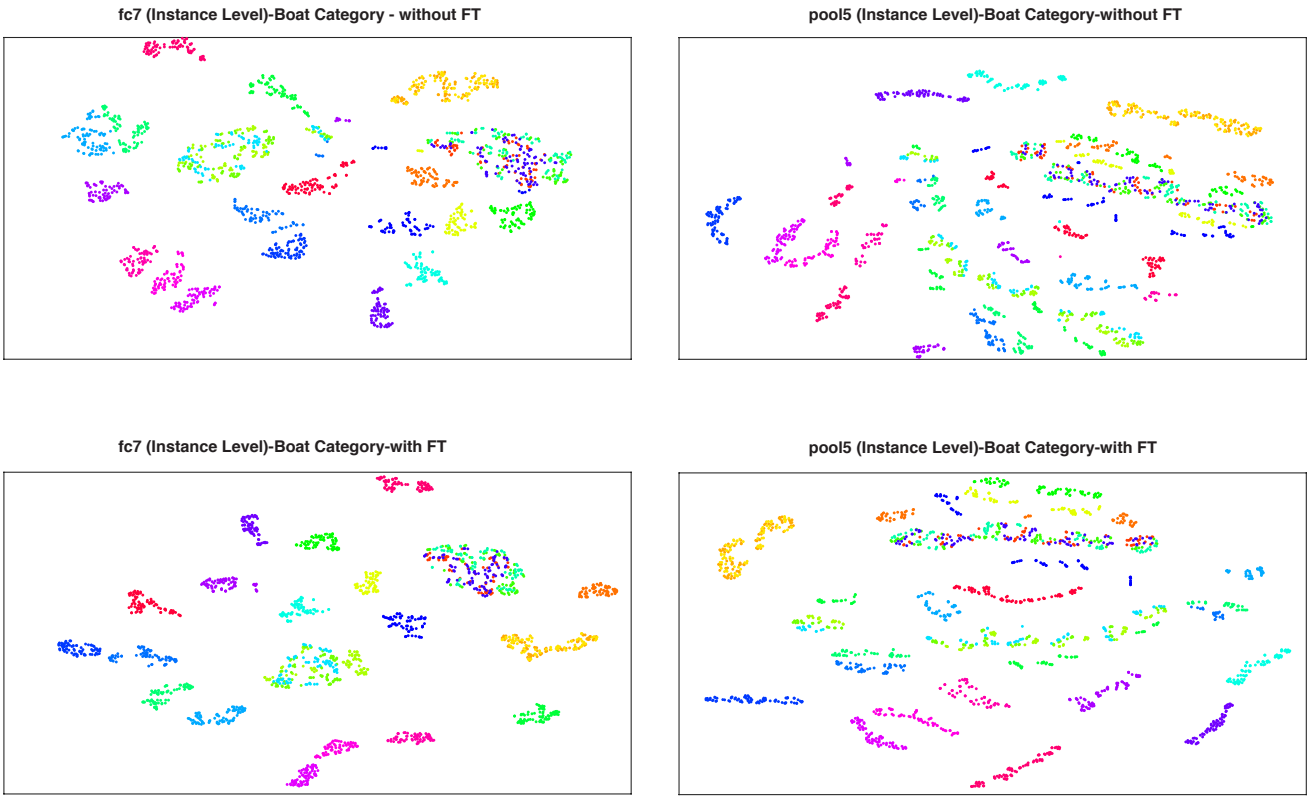

Figure 5. t-SNE representation for instance prediction using fc7 and pool5 layers with and without fine-tuning.