Comparing the Interpretability of the Deep Visual Representations via Network Dissection

Bolei Zhou*, **David Bau***, **Aude Oliva, Antonio Torralba**
CSAIL, MIT
{bzhou, davidbau, oliva, torralba}@csail.mit.edu

Abstract

The success of recent deep convolutional neural networks (CNNs) depends on learning hidden representations that can summarize the important factors of variation behind the data. However, CNNs often criticized as being black boxes that lack interpretability, since they have millions of unexplained model parameters. We propose Network Dissection[1], a general framework to quantify the interpretability of the units inside a deep convolutional neural networks (CNNs). We compare the different vocabularies of interpretable units as concept detectors emerged from the networks trained to solve different supervised learning tasks such as object recognition on ImageNet and scene classification on Places, and self-supervised training tasks. The network dissection is further applied to analyze how the units acting as semantic detectors grow and evolve over the training iterations both in the scenario of the train-from-scratch and in the stage of the fine-tuning between data sources. Our results highlight that interpretability is an important property of deep neural networks that provides new insights into their hierarchical structure.

1 Introduction

Previous efforts to interpret the internals of a convolutional neural network have focused on visualizations, for example, visualizing image patches that maximize individual unit activations Zeiler & Fergus (2014); Zhou et al. (2015); or using optimization to generate patterns and regions salient to a unit Mahendran & Vedaldi (2015); Simonyan et al. (2014); Zeiler & Fergus (2014); Nguyen et al. (2016); or rendering representation space using dimensionality reduction Maaten & Hinton (2008); Jolliffe (2002). Though the visualizations give us the intuition about what image patterns the internal units are trying to detect, the results based on visualization are usually qualitative and unable to be interpreted quantitatively, *i.e.* which human interpretable concept some unit detects and how accurate it is. Therefore it is still an open question on how to quantify the interpretability of the deep visual representations and compare them beyond their classification power.

Recently we propose a framework called *Network Dissection* to quantify the interpretability of any given CNNs Bau et al. (2017). Network dissection quantifies the interpretability of any given network by measuring the degree of alignment between the unit activation and the ground-truth labels in a pre-defined dictionary of concepts. Based on the quantified interpretability, we compare the semantis of units in various networks from supervised training and self-supervised training, and the effect of training iterations and fine-tunining to the internal representations of the networks. Our results highlight that interpretability is an important property of deep neural networks that provides new insights into their deep hierarchical structure.

[1]The complete paper and code are available at `http://netdissect.csail.mit.edu`

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

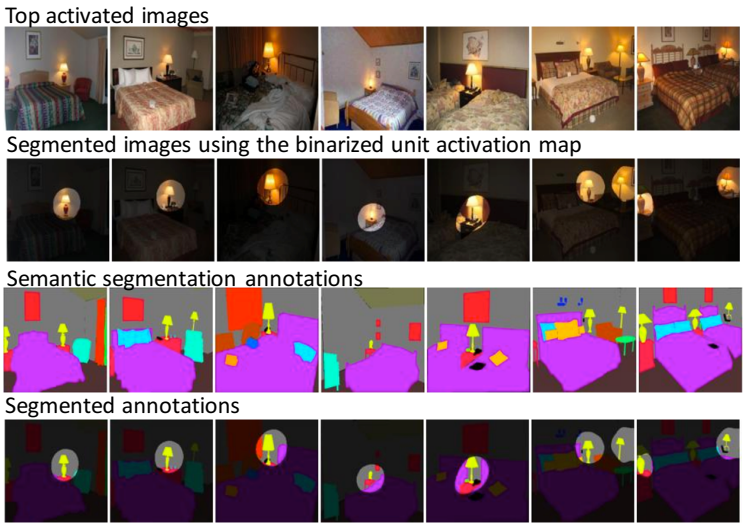

Figure 1: Scoring unit interpretability by evaluating the unit activation for semantic segmentation. Unit activation map is used to segment the top activated images, localizing the favorite image patterns for that unit. The activation map is further used to segment the annotation mask to compute the IoU.

2 Overview of Network Dissection

To measure interpretability, we evaluate the ability of each hidden unit to solve segmentation problems from a dictionary of human-interpretable visual concepts.

2.1 Broden: Broadly and Densely Labeled Dataset

As a dictionary of visual concepts, we construct the Broadly and Densely Labeled Dataset (**Broden**), which unifies several densely labeled image data sets: ADE Zhou et al. (2017), OpenSurfaces Bell et al. (2014), Pascal-Context Mottaghi et al. (2014), Pascal-Part Chen et al. (2014), and the Describable Textures Dataset Cimpoi et al. (2014), containing a broad range of labeled classes of objects, scenes, object parts, textures, and materials, with most examples labeled at the pixel level.

2.2 Scoring Unit Interpretability

Let c denote any concept within the Broden dataset and let k denote any convolutional unit in a CNN. Network dissection defines the quality of the interpretation c for unit k by quantifying the ability of k to solve the segmentation problem given by c using this IoU score:

$$IoU_{k,c} = \frac{\sum |M_k(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum |M_k(\mathbf{x}) \cup L_c(\mathbf{x})|}, \tag{1}$$

In the above, $\mathbf{x}$ represents an image in the Broden dataset, $L_c(\mathbf{x})$ is the set of pixels labeled with concept c, and $M_k(\mathbf{x})$ is binary mask selecting those pixels that lie within areas of highest activation of unit k. M_k is computed by (bilinearly) upsampling the activation of k on input $\mathbf{x}$, and applying a threshold T_k that selects a fixed quantile (0.5%) of the pixels over the entire dataset. Because the data set contains some categories of labels (such as textures) which are not present on some subsets of inputs, the sums are computed only on the subset of images that have at least one labeled concept of the same category as c. Figure 1 gives one example of computing the IoU over the top activated images with semantic segmentation annotations.

The value of $IoU_{k,c}$ is the accuracy of unit k in detecting concept c. In our analysis, we consider a unit k as a detector for concept c if $IoU_{k,c} > 0.04$, and when a unit detects more than one concept, we choose the top scoring label. To quantify the interpretability of a layer, we count the distinct concepts detected, i.e., the number of *unique detectors*.

3 Experiments

3.1 The emergent concept detectors across different networks

Network dissection is applied to the last convolutional layer of different networks (the details of each network are available at the project page). Figure 2 shows the histogram of units identified as concept detectors in each network. Each concept class might have several units as its detectors. For example, for the networks trained on ImageNet, the most frequent detectrors are dog detectors. For the networks trained on Places, the most frequent detector in AlexNet is water detector, while the most frequent detector in ResNet is airplane detector. We can see that the emergent detectors vary across training supervisions and network architectures. Figure 3 shows some exemplar detectors from different networks grouped by some object classes. We can see that deeper networks such as DenseNet and ResNet are able to capture more compact shapes of the object.

3.2 The emergence of concepts over training iterations

Figure 4 plots the interpretability of snapshots of the baseline model (AlexNet trained on Places205) at different training iterations along with the accuracy on the validation set. We can see that object detectors and part detectors start emerging at about 10,000 iterations (each iteration processes a batch of 256 images). Meanwhile, we do not find the evidence of transitions across different concept levels during training. For example, units in `conv5` does not turn into texture or material detectors before converging into object or part detectors. Besides, we see there is strong correlation between the validation accuracy and the emergence of high-level object detectors thus the interpretability might help debug the network during the training.

In Figure 5, we keep track of four units over different training iterations. We observe that the units start converging to the semantic concept at early stage. For example, in the second row the unit starts detecting mountain from iteration 5000. Meanwhile, some units have interesting transition over concepts, for example the unit in the first row detects road first before it detects car.

3.3 The evolution of units in transfer learning

Fine-tuning the pre-trained network such as ImageNet-CNN to another target dataset is a commonly used technique in transfer learning. It makes the training converge faster, while it leads to better accuracy in the case that there is not enough training data at the target dataset. Here we observe that the interpretation of the internal units evolves over different stages of training in the transfer learning.

Given a well trained Places-AlexNet and ImageNet-AlexNet respectively, we fine-tune the Places-AlexNet on ImageNet and fine-tune the ImageNet-AlexNet on Places respectively. The interpretability results of the snapshots of the networks over the fine-tuning iterations are plotted in Figure 6. We can see that the training indeed converges faster compared to the network trained from scratch on Places in Figure 4. The semantics of units also change over fine-tuning. For example, the number of unique object detectors first drop then keep increasing for the network trained on ImageNet being fine-tuned to Places365, while it is slowly dropping for the network trained on Places being fine-tuned to ImageNet.

Figure 7 shows the evolution of the six units in the network fine-tuned from ImageNet to Places365 and reversely. The top associated interpretation for each unit keep evolving during the fine-tuning process. For example, in the network fine-tuned from ImageNet to Places365, the first unit which detects the white dog, evolves to detect the waterfall; the third unit which detects the green concept, evolves to detect the baseball field. On the other hand, in the network fine-tuned from Places365 to ImageNet, units detecting different concepts converge to detect dog-relevant concepts such as ear and dog head. Interestingly though those units evolve to detect different concepts, many of them still remain to have similarity in colors or textures.

4 Conclusion

Based on the network dissection, we compare the interpretability of deep visual representations for a range of networks trained from different supervisions and training conditions. We show that the

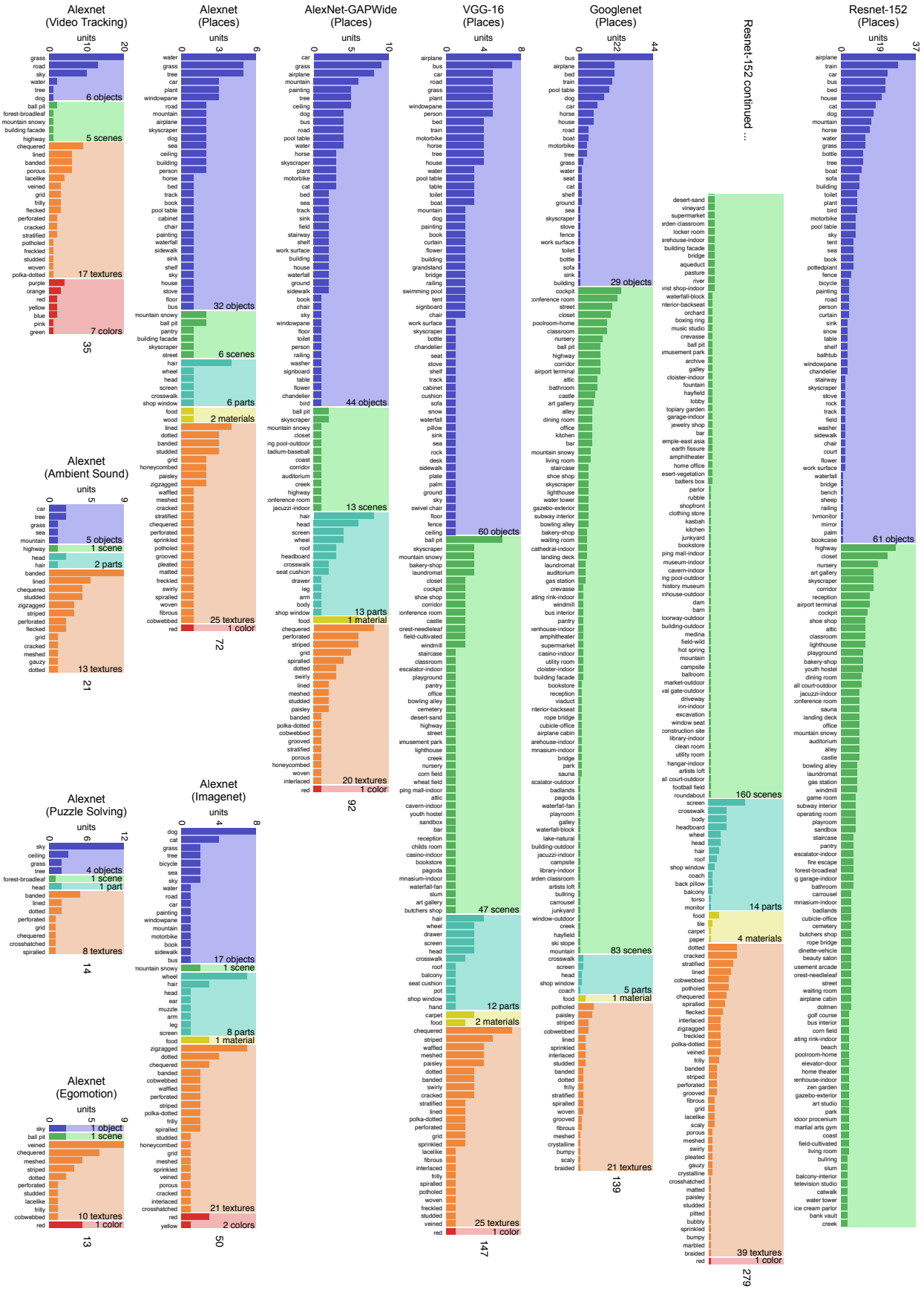

Figure 2: Comparison of unique detectors of all types on a variety of network architectures and supervision tasks. More results are available at the project page.

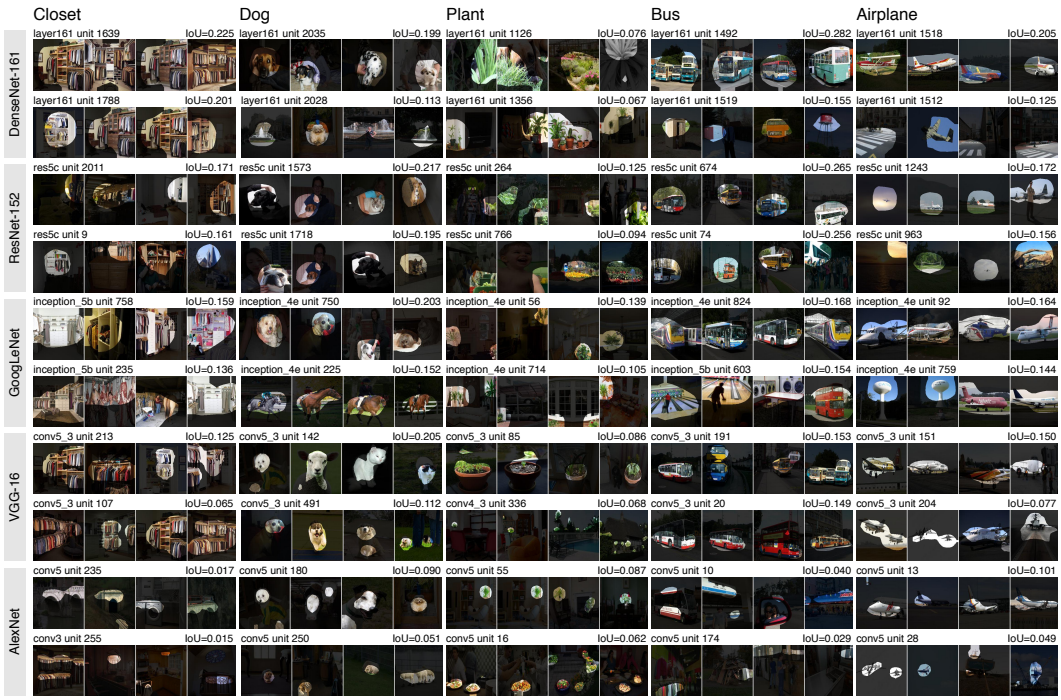

Figure 3: Comparison of several visual concept detectors identified by network dissection in DenseNet, ResNet, GoogLeNet, VGG, and AlexNet. Each network is trained on Places365. The two highest-IoU matches among convolutional units of each network is shown. The four maximally activated Broden images segmented by unit activation map are shown as the visualization of each unit.

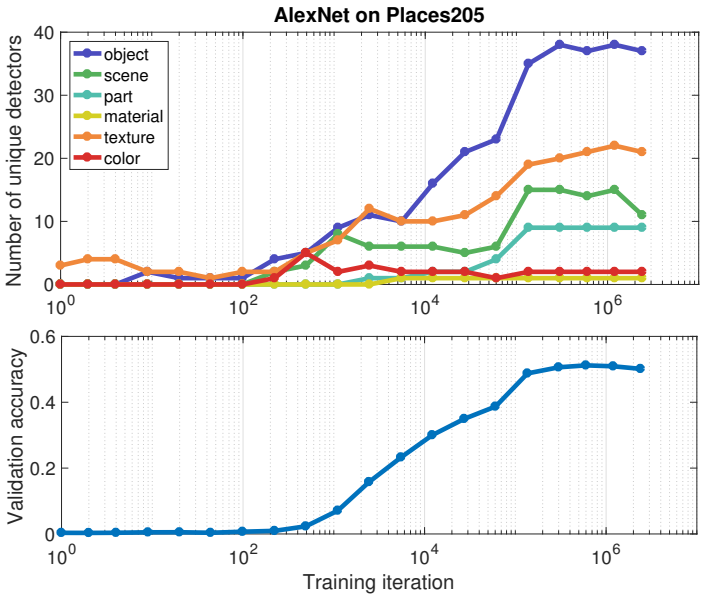

Figure 4: The interpretability of the units at conv5 layer of the baseline model over 300,000 training iterations. The validation accuracy is plotted below.

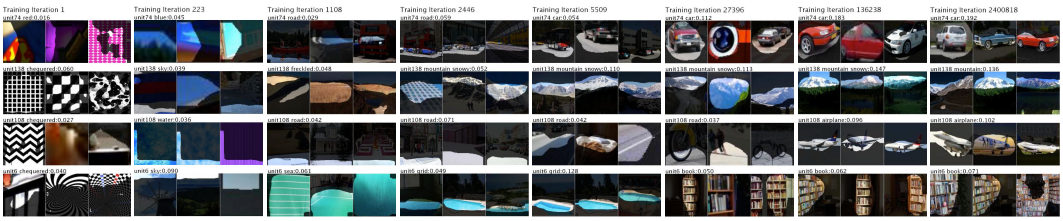

Figure 5: The interpretation of four units evolves at different training iterations.

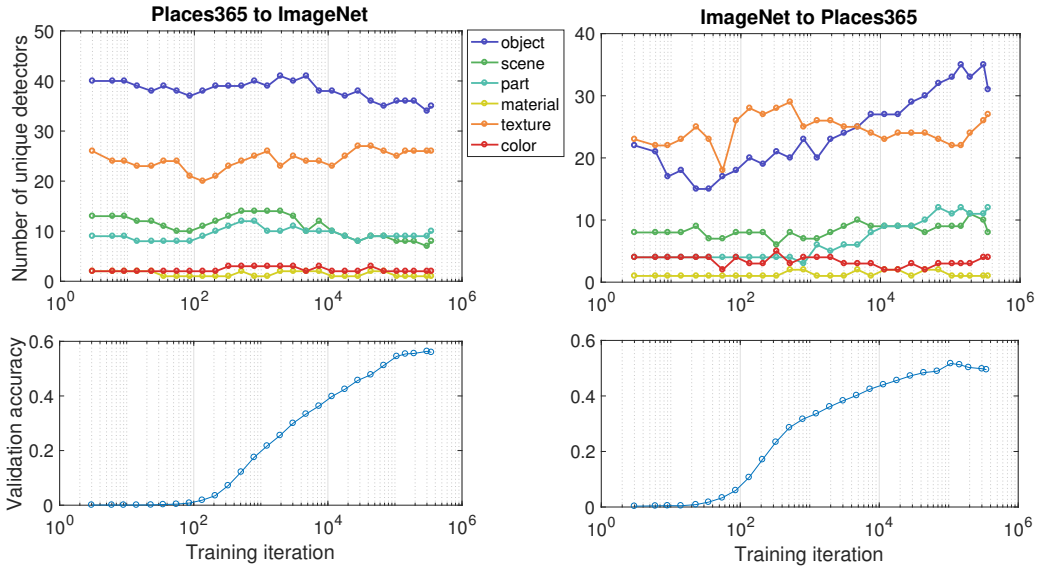

Figure 6: The network interpretability under the fine-tuning between Places and ImageNet. The validation accuracy is plotted below. The network architecture is the same as AlexNet.

interpretability based on unit-concept alignment is an important property of deep neural networks that could be used to compare networks beyond their classification accuracy.

References

Bau, David, Zhou, Bolei, Khosla, Aditya, Oliva, Aude, and Torralba, Antonio. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

Bell, Sean, Bala, Kavita, and Snavely, Noah. Intrinsic images in the wild. *ACM Trans. on Graphics (SIGGRAPH)*, 2014.

Chen, Xianjie, Mottaghi, Roozbeh, Liu, Xiaobai, Fidler, Sanja, Urtasun, Raquel, and Yuille, Alan. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proc. CVPR*, 2014.

Cimpoi, Mircea, Maji, Subhransu, Kokkinos, Iasonas, Mohamed, Sammy, and Vedaldi, Andrea. Describing textures in the wild. In *Proc. CVPR*, 2014.

Jolliffe, Ian. *Principal component analysis*. Wiley Online Library, 2002.

Maaten, Laurens van der and Hinton, Geoffrey. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Mahendran, Aravindh and Vedaldi, Andrea. Understanding deep image representations by inverting them. *Proc. CVPR*, 2015.

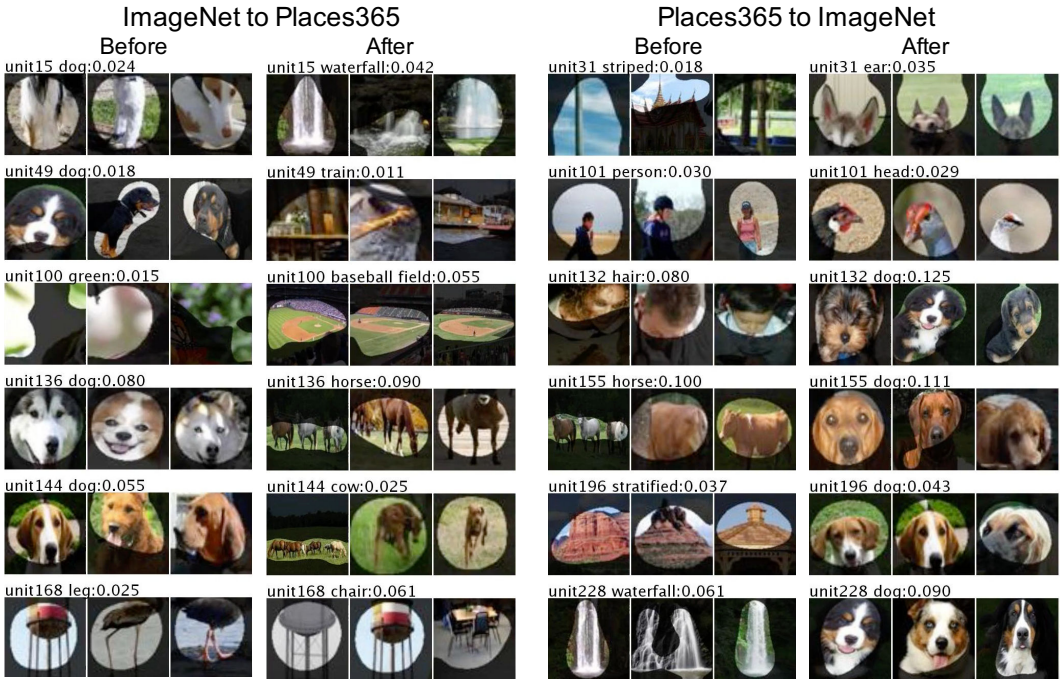

Figure 7: The interpretation of six units before and after fine-tuning between ImageNet and Places.

Mottaghi, Roozbeh, Chen, Xianjie, Liu, Xiaobai, Cho, Nam-Gyu, Lee, Seong-Whan, Fidler, Sanja, Urtasun, Raquel, and Yuille, Alan. The role of context for object detection and semantic segmentation in the wild. In *Proc. CVPR*, 2014.

Nguyen, Anh, Dosovitskiy, Alexey, Yosinski, Jason, Brox, Thomas, and Clune, Jeff. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *In Advances in Neural Information Processing Systems*, 2016.

Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. *International Conference on Learning Representations Workshop*, 2014.

Zeiler, Matthew D and Fergus, Rob. Visualizing and understanding convolutional networks. *Proc. ECCV*, 2014.

Zhou, Bolei, Khosla, Aditya, Lapedriza, Agata, Oliva, Aude, and Torralba, Antonio. Object detectors emerge in deep scene cnns. *International Conference on Learning Representations*, 2015.

Zhou, Bolei, Zhao, Hang, Puig, Xavier, Fidler, Sanja, Barriuso, Adela, and Torralba, Antonio. Scene parsing through ade20k dataset. *Proc. CVPR*, 2017.