Learning Deep Features for Scene Recognition using Places Database

Bolei Zhou[1], Agata Lapedriza[1,3], Jianxiong Xiao[2], Antonio Torralba[1], and Aude Oliva[1]

[1]Massachusetts Institute of Technology
[2]Princeton University
[3]Universitat Oberta de Catalunya

Abstract

Scene recognition is one of the hallmark tasks of computer vision, allowing definition of a context for object recognition. Whereas the tremendous recent progress in object recognition tasks is due to the availability of large datasets like ImageNet and the rise of Convolutional Neural Networks (CNNs) for learning high-level features, performance at scene recognition has not attained the same level of success. This may be because current deep features trained from ImageNet are not competitive enough for such tasks. Here, we introduce a new scene-centric database called Places with over 7 million labeled pictures of scenes. We propose new methods to compare the density and diversity of image datasets and show that Places is as dense as other scene datasets and has more diversity. Using CNN, we learn deep features for scene recognition tasks, and establish new state-of-the-art results on several scene-centric datasets. A visualization of the CNN layers' responses allows us to show differences in the internal representations of object-centric and scene-centric networks.

1 Introduction

Understanding the world in a single glance is one of the most accomplished feats of the human brain: it takes only a few tens of milliseconds to recognize the category of an object or environment, emphasizing an important role of feedforward processing in visual recognition. One of the mechanisms subtending efficient human visual recognition is our capacity to learn and remember a diverse set of places and exemplars [11]; by sampling the world several times per second, our neural architecture constantly registers new inputs even for a very short time, reaching an exposure to millions of natural images within just a year. How much would an artificial system have to learn before reaching the scene recognition abilities of a human being?

Besides the exposure to a dense and rich variety of natural images, one important property of the primate brain is its hierarchical organization in layers of increasing processing complexity, an architecture that has inspired Convolutional Neural Networks or CNNs [2, 14]. These architectures together with recent large databases (e.g., ImageNet [3]) have obtained astonishing performance on object classification tasks [12, 5, 20]. However, the baseline performance reached by these networks on scene classification tasks is within the range of performance based on hand-designed features and sophisticated classifiers [24, 21, 4]. Here, we show that one of the reasons for this discrepancy is that the higher-level features learned by object-centric versus scene-centric CNNs are different: iconic images of objects do not contain the richness and diversity of visual information that pictures of scenes and environments provide for learning to recognize them.

Here we introduce Places, a scene-centric image dataset 60 times larger than the SUN database [24]. With this database and a standard CNN architecture, we establish new baselines of accuracies on various scene datasets (Scene15 [17, 13], MIT Indoor67 [19], SUN database [24], and SUN Attribute Database [18]), significantly outperforming the results obtained by the deep features from the same network architecture trained with ImageNet[1].

The paper is organized as follows: in Section 2 we introduce the Places database and describe the collection procedure. In Section 3 we compare Places with the other two large image datasets: SUN [24] and ImageNet [3]. We perform experiments on Amazon Mechanical Turk (AMT) to compare these 3 datasets in terms of density and diversity. In Section 4 we show new scene classification performance when training deep features from millions of labeled scene images. Finally, we visualize the units' responses at different layers of the CNNs, demonstrating that an object-centric network (using ImageNet [12]) and a scene-centric network (using Places) learn different features.

2 Places Database

The first benchmark for scene classification was the Scene15 database [13] based on [17]. This dataset contains only 15 scene categories with a few hundred images per class, where current classifiers are saturating this dataset nearing human performance at 95%. The MIT Indoor67 database [19] has 67 categories on indoor places. The SUN database [24] was introduced to provide a wide coverage of scene categories. It is composed of 397 categories containing more than 100 images per category.

Despite those efforts, all these scene-centric datasets are small in comparison with current object datasets such as ImageNet (note that ImageNet also contains scene categories but in a very small proportion as is shown in Fig. 2). Complementary to ImageNet (mostly object-centric), we present here a scene-centric database, that we term the Places database. As now, Places contain more than 7 million images from 476 place categories, making it the largest image database of scenes and places so far and the first scene-centric database competitive enough to train algorithms that require huge amounts of data, such as CNNs.

2.1 Building the Places Database

Since the SUN database [24] has a rich scene taxonomy, the Places database has inherited the same list of scene categories. To generate the query of image URL, 696 common adjectives (messy, spare, sunny, desolate, etc), manually selected from a list of popular adjectives in English, are combined with each scene category name and are sent to three image search engines (Google Images, Bing Images, and Flickr). Adding adjectives to the queries allows us to download a larger number of images than what is available in ImageNet and to increase the diversity of visual appearances. We then remove duplicated URLs and download the raw images with unique URLs. To date, more than 40 million images have been downloaded. Only color images of 200×200 pixels or larger are kept. PCA-based duplicate removal is conducted within each scene category in the Places database and across the same scene category in the SUN database, which ensures that Places and the SUN do not contain the same images, allowing us to combine the two datasets.

The images that survive this initial selection are sent to Amazon Mechanical Turk for two rounds of individual image annotation. For a given category name, its definition as in [24], is shown at the top of a screen, with a question like *is this a living room scene?* A single image at a time is shown centered in a large window, and workers are asked to press a Yes or No key. For the first round of labeling, the default answer is set to No, requiring the worker to actively pick up the positive images. The positive images resulting from the first round annotation are further sent for a second round annotation, in which the default answer is set to Yes (to pick up the remaining negative images). In each HIT(one assignment for each worker), 750 downloaded images are included for annotation, and an additional 30 positive samples and 30 negative samples with ground truth from the SUN database are also randomly injected as control. Valid HITs kept for further analyses require an accuracy of 90% or higher on these control images. After the two rounds of annotation, and as this paper is published, 7,076,580 images from 476 scene categories are included in the Places database. Fig. 1 shows image samples obtained with some of the adjectives used in the queries.

[1] The database and pre-trained networks are available at `http://places.csail.mit.edu`

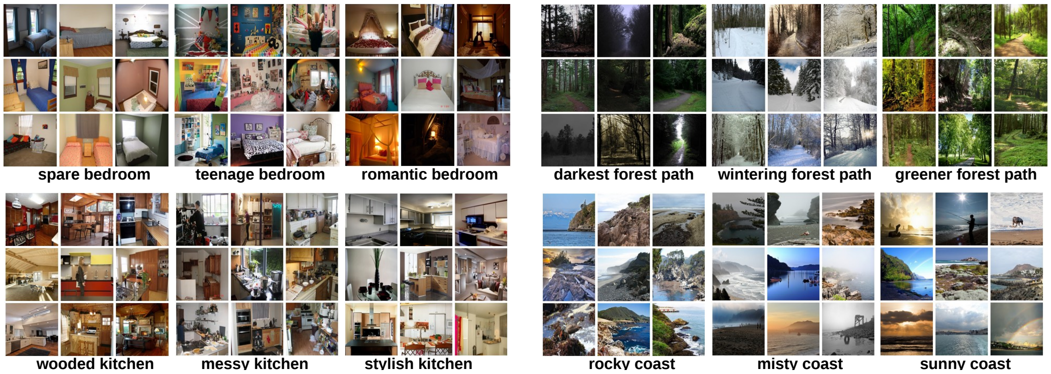

Figure 1: Image samples from the scene categories grouped by their queried adjectives.

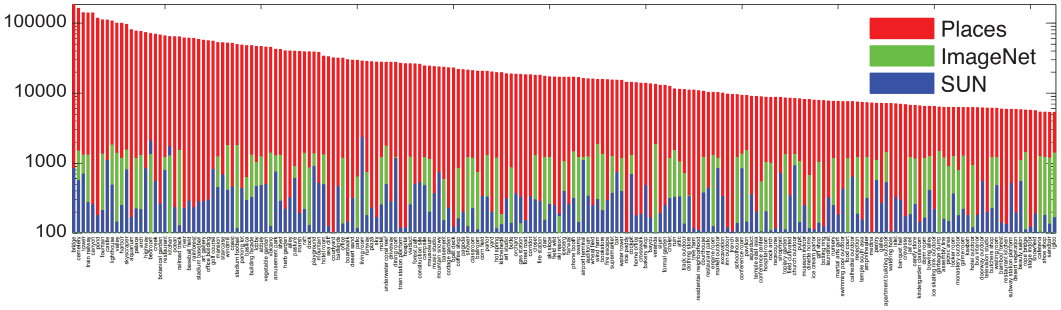

Figure 2: Comparison of the number of images per scene category in three databases.

We made 2 subsets of Places that will be used across the paper as benchmarks. The first one is Places 205, with the 205 categories with more than 5000 images. Fig. 2 compares the number of images in Places 205 with ImageNet and SUN. Note that ImageNet only has 128 of the 205 categories, while SUN contains all of them (we will call this set SUN 205, and it has, at least, 50 images per category). The second subset of Places used in this paper is Places 88. It contains the 88 common categories with ImageNet such that there are at least 1000 images in ImageNet. We call the corresponding subsets SUN 88 and ImageNet 88.

3 Comparing Scene-centric Databases

Despite the importance of benchmarks and training datasets in computer vision, comparing datasets is still an open problem. Even datasets covering the same visual classes have notable differences providing different generalization performance when used to train a classifier [23]. Beyond the number of images and categories, there are aspects that are important but difficult to quantify, like the variability in camera poses, in decoration styles or in the objects that appear in the scene.

Although the quality of a database will be task dependent, it is reasonable to assume that a good database should be dense (with a high degree of data concentration), and diverse (it should include a high variability of appearances and viewpoints). Both quantities, density and diversity, are hard to estimate in image sets, as they assume some notion of similarity between images which, in general, is not well defined. Two images of scenes can be considered similar if they contain similar objects, and the objects are in similar spatial configurations and pose, and have similar decoration styles. However, this notion is loose and subjective so it is hard to answer the question *are these two images similar?* For this reason, we define relative measures for comparing datasets in terms of density and diversity that only require ranking similarities. In this section we will compare the densities and diversities of SUN, ImageNet and Places using these relative measures.

3.1 Relative Density and Diversity

Density is a measure of data concentration. We assume that, in an image set, high density is equivalent to the fact that images have, in general, similar neighbors. Given two databases A and B, relative density aims to measure which one of the two sets has the most similar nearest neighbors. Let a_1 be a random image from set A and b_1 from set B and let us take their respective nearest neighbors in each set, a_2 from A and b_2 from B. If A is denser than B, then it would be more likely that a_1 and a_2 are closer to each other than b_1 and b_2. From this idea we define the relative density as $\text{Den}_B(A) = p\left(d(a_1, a_2) < d(b_1, b_2)\right)$, where $d(a_1, a_2)$ is a distance measure between two images (small distance implies high similarity). With this definition of relative density we have that A is denser than B if, and only if, $\text{Den}_B(A) > \text{Den}_A(B)$. This definition can be extended to an arbitrary number of datasets, $A_1, ..., A_N$:

$$\text{Den}_{A_2,...,A_N}(A_1) = p(d(a_{11}, a_{12}) < \min_{i=2:N} d(a_{i1}, a_{i2})) \tag{1}$$

where $a_{i1} \in A_i$ are randomly selected and $a_{i2} \in A_i$ are near neighbors of their respective a_{i1}.

The quality of a dataset can not be measured just by its density. Imagine, for instance, a dataset composed of 100,000 images all taken within the same bedroom. This dataset would have a very high density but a very low diversity as all the images would look very similar. An ideal dataset, expected to generalize well, should have high *diversity* as well.

There are several measures of diversity, most of them frequently used in biology to characterize the richness of an ecosystem (see [9] for a review). In this section, we will use a measure inspired by Simpson index of diversity [22]. Simpson index measures the probability that two random individuals from an ecosystem belong to the same species. It is a measure of how well distributed are the individuals across different species in an ecosystem and it is related to the entropy of the distribution. Extending this measure for evaluating the diversity of images within a category is non-trivial if there are no annotations of sub-categories. For this reason, we propose to measure relative diversity of image datasets A and B based on this idea: if set A is more diverse than set B, then two random images from set B are more likely to be visually similar than two random samples from A. Then, the diversity of A with respect to B can be defined as $\text{Div}_B(A) = 1 - p(d(a_1, a_2) < d(b_1, b_2))$, where $a_1, a_2 \in A$ and $b_1, b_2 \in B$ are randomly selected. With this definition of relative diversity we have that A is more diverse than B if, and only if, $\text{Div}_B(A) > \text{Div}_A(B)$. For an arbitrary number of datasets, $A_1, ..., A_N$:

$$\text{Div}_{A_2,...,A_N}(A_1) = 1 - p(d(a_{11}, a_{12}) < \min_{i=2:N} d(a_{i1}, a_{i2})) \tag{2}$$

where $a_{i1}, a_{i2} \in A_i$ are randomly selected.

3.2 Experimental Results

We measured the relative densities and diversities between SUN, ImageNet and Places using AMT. Both measures used the same experimental interface: workers were presented with different pairs of images and they had to select the pair that contained the most similar images. We observed that different annotators are consistent in deciding whether a pair of images is more similar than another pair of images.

In these experiments, the only difference when estimating density and diversity is how the pairs are generated. For the diversity experiment, the pairs are randomly sampled from each database. Each trial is composed of 4 pairs from each database, giving a total of 12 pairs to chose from. We used 4 pairs per database to increase the chances of finding a similar pair and avoiding users having to skip trials. AMT workers had to select the most similar pair on each trial. We ran 40 trials per category and two observers per trial, for the 88 categories in common between ImageNet, SUN and Places databases. Fig. 3a shows some examples of pairs from one of the density experiments.The pair selected by AMT workers as being more similar is highlighted.

For the density experiments, we selected pairs that were more likely to be visually similar. This would require first finding the true nearest neighbor of each image, which would be experimentally costly. Instead we used visual similarity as measured by using the Euclidean distance between the Gist descriptor [17] of two images. Each pair of images was composed from one randomly selected image and its 5-th nearest neighbor using Gist (we ignored the first 4 neighbors to avoid

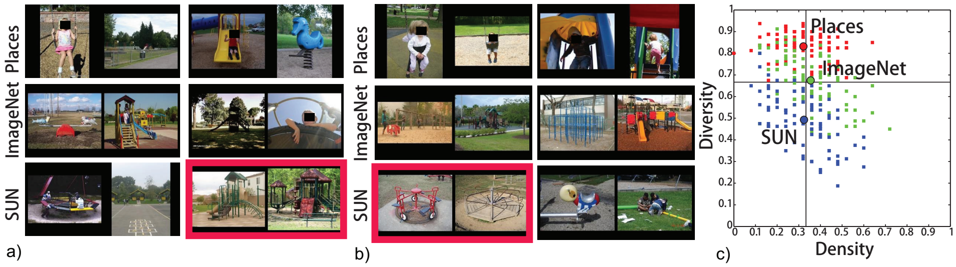

Figure 3: a) Examples of pairs for the diversity experiment. b) Examples of pairs for the density experiment. c) Scatter plot of relative diversity vs. relative density per each category and dataset.

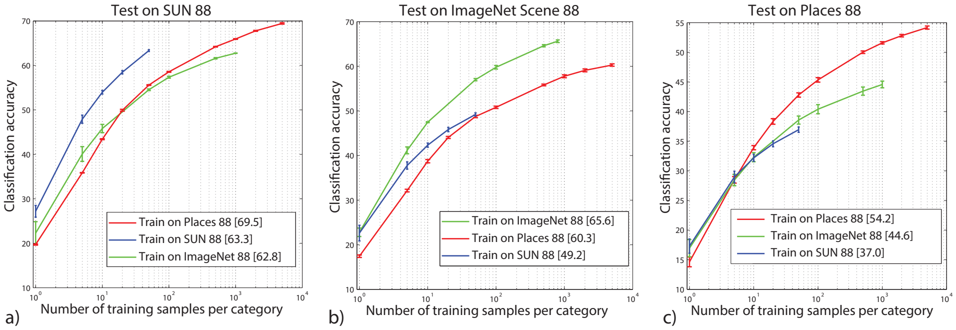

Figure 4: Cross dataset generalization of training on the 88 common scenes between Places, SUN and ImageNet then testing on the 88 common scenes from: a) SUN, b) ImageNet and c) Places database.

near duplicates, which would give a wrong sense of high density). In this case we also show 12 pairs of images at each trial, but run 25 trials per category instead of 40 to avoid duplicate queries. Fig. 3b shows some examples of pairs per one of the density experiments and also the selected pair is highlighted. Notice that in the density experiment (where we computed neighbors) the pairs look, in general, more similar than in the diversity experiment.

Fig. 3c shows a scatter plot of relative diversity vs. relative density for all the 88 categories and the three databases. The point of crossing between the two black lines indicates the point where all the results should fall if all the datasets were identical in terms of diversity and density. The figure also shows the average of the density and diversity over all categories for each dataset.

In terms of density, the three datasets are, on average, very similar. However, there is a larger variation in terms of diversity, showing Places to be the most diverse of the three datasets. The average relative diversity on each dataset is 0.83 for Places, 0.67 for ImageNet and 0.50 for SUN. In the experiment, users selected pairs from the SUN database to be the closest to each other 50% of the time, while the pairs from the Places database were judged to be the most similar only on 17% of the trials. The categories with the largest variation in diversity across the three datasets are *playground*, *veranda* and *waiting room*.

3.3 Cross Dataset Generalization

As discussed in [23], training and testing across different datasets generally results in a drop of performance due to the dataset bias problem. In this case, the bias between datasets is due, among other factors, to the differences in the density and diversity between the three datasets. Fig. 4 shows the classification results obtained from the training and testing on different permutations of the 3 datasets. For these results we use the features extracted from a pre-trained ImageNet-CNN and a linear SVM. In all three cases training and testing on the same dataset provides the best performance for a fixed number of training examples. As the Places database is very large, it achieves the best performance on two of the test sets when all the training data is used. In the next section we will show that a CNN network trained using the Places database achieves a significant improvement over scene-centered benchmarks in comparison with a network trained using ImageNet.

Table 1: Classification accuracy on the test set of Places 205 and the test set of SUN 205.

	Places 205	SUN 205
Places-CNN	**50.0%**	**66.2%**
ImageNet CNN feature+SVM	40.8%	49.6%

4 Training Neural Network for Scene Recognition and Deep Features

Deep convolutional neural networks have obtained impressive classification performance on the ImageNet benchmark [12]. For the training of Places-CNN, we randomly select 2,448,873 images from 205 categories of Places (referred to as Places 205) as the train set, with minimum 5,000 and maximum 15,000 images per category. The validation set contains 100 images per category and the test set contains 200 images per category (a total of 41,000 images). Places-CNN is trained using the Caffe package on a GPU NVIDIA Tesla K40. It took about 6 days to finish 300,000 iterations of training. The network architecture of Places-CNN is the same as the one used in the Caffe reference network [10]. The Caffe reference network, which is trained on 1.2 million images of ImageNet (ILSVRC 2012), has approximately the same architecture as the network proposed by [12]. We call the Caffe reference network as ImageNet-CNN in the following comparison experiments.

4.1 Visualization of the Deep Features

Through the visualization of the responses of the units for various levels of network layers, we can have a better understanding of the differences between the ImageNet-CNN and Places-CNN given that they share the same architecture. Fig.5 visualizes the learned representation of the units at the Conv 1, Pool 2, Pool 5, and FC 7 layers of the two networks. Whereas Conv 1 units can be directly visualized (they capture the oriented edges and opponent colors from both networks), we use the mean image method to visualize the units of the higher layers: we first combine the test set of ImageNet LSVRC2012 (100,000 images) and SUN397 (108,754 images) as the input for both networks; then we sort all these images based on the activation response of each unit at each layer; finally we average the top 100 images with the largest responses for each unit as a kind of receptive field (RF) visualization of each unit. To compare the units from the two networks, Fig. 5 displays mean images sorted by their first principal component. Despite the simplicity of the method, the units in both networks exhibit many differences starting from Pool 2. From Pool 2 to Pool 5 and FC 7, gradually the units in ImageNet-CNN have RFs that look like object-blobs, while units in Places-CNN have more RFs that look like landscapes with more spatial structures. These learned unit structures are closely relevant to the differences of the training data. In future work, it will be fascinating to relate the similarity and differences of the RF at different layers of the object-centric network and scene-centric network with the known object-centered and scene-centered neural cortical pathways identified in the human brain (for a review, [16]). In the next section we will show that these two networks (only differing in the training sets) yield very different performances on a variety of recognition benchmarks.

4.2 Results on Places 205 and SUN 205

After the Places-CNN is trained, we use the final layer output (Soft-max) of the network to classify images in the test set of Places 205 and SUN 205. The classification result is listed in Table 1. As a baseline comparison, we show the results of a linear SVM trained on ImageNet-CNN features of 5000 images per category in Places 205 and 50 images per category in SUN 205 respectively. Places-CNN performs much better. We further compute the performance of the Places-CNN in the terms of the top-5 error rate (one test sample is counted as misclassified if the ground-truth label is not among the top 5 predicted labels of the model). The top-5 error rate for the test set of the Places 205 is **18.9**%, while the top-5 error rate for the test set of SUN 205 is **8.1%**.

4.3 Generic Deep Features for Visual Recognition

We use the responses from the trained CNN as generic features for visual recognition tasks. Responses from the higher-level layers of CNN have proven to be effective generic features with state-of-the-art performance on various image datasets [5, 20]. Thus we evaluate performance of the

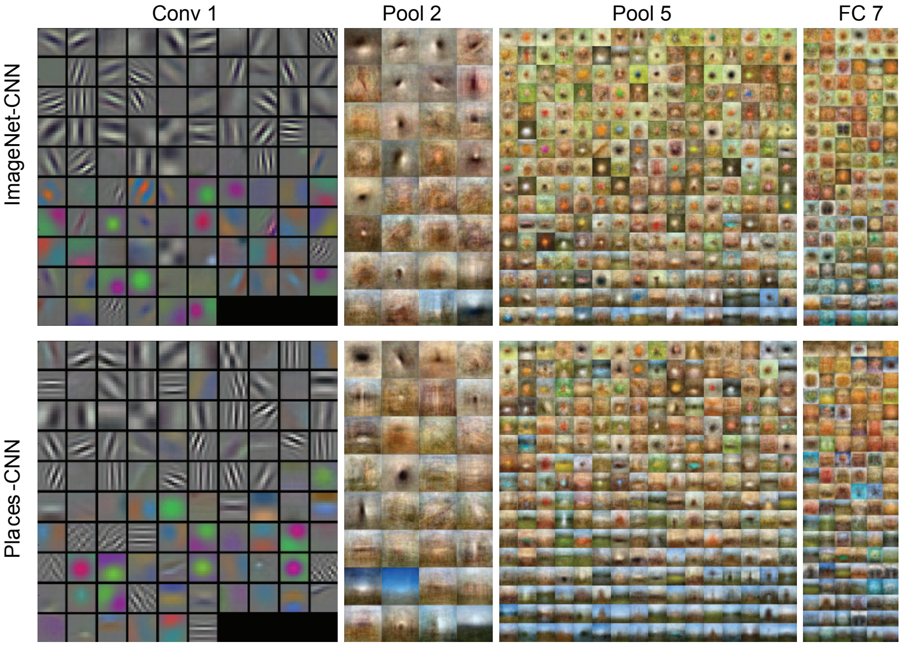

Figure 5: Visualization of the units' receptive fields at different layers for the ImageNet-CNN and Places-CNN. Conv 1 units contains 96 filters. The Pool 2 feature map is 13×13×256; The Pool 5 feature map is 6×6×256; The FC 7 feature map is 4096×1. Subset of units at each layer are shown.

Table 2: Classification accuracy/precision on scene-centric databases and object-centric databases for the Places-CNN feature and ImageNet-CNN feature. The classifier in all the experiments is a linear SVM with the same parameters for the two features.

	SUN397	MIT Indoor67	Scene15	SUN Attribute
Places-CNN feature	**54.32±0.14**	**68.24**	**90.19±0.34**	**91.29**
ImageNet-CNN feature	42.61±0.16	56.79	84.23±0.37	89.85

	Caltech101	Caltech256	Action40	Event8
Places-CNN feature	65.18±0.88	45.59±0.31	42.86±0.25	94.12±0.99
ImageNet-CNN feature	**87.22±0.92**	**67.23±0.27**	**54.92±0.33**	**94.42±0.76**

deep features from the Places-CNN on the following scene and object benchmarks: SUN397 [24], MIT Indoor67 [19], Scene15 [13], SUN Attribute [18], Caltech101 [7], Caltech256 [8], Stanford Action40 [25], and UIUC Event8 [15]. All the experiments follow the standards in those papers [2].

As a comparison, we evaluate the deep feature's performance from the ImageNet-CNN on those same benchmarks. Places-CNN and ImageNet-CNN have exactly the same network architecture, but they are trained on scene-centric data and object-centric data respectively. We use the deep features from the response of the Fully Connected Layer (FC) 7 of the CNNs, which is the final fully connected layer before producing the class predictions. There is only a minor difference between the feature of FC 7 and the feature of FC 6 layer [5]. The deep feature for each image is a 4096-dimensional vector.

Table 2 summarizes the classification accuracy on various datasets for the ImageNet-CNN feature and the Places-CNN feature. Fig.6 plots the classification accuracy for different visual features on SUN397 database and SUN Attribute dataset. The classifier is a linear SVM with the same default parameters for the two deep features (C=1) [6]. The Places-CNN feature shows impressive performance on scene classification benchmarks, outperforming the current state-of-the-art methods for SUN397 (47.20% [21]) and for MIT Indoor67 (66.87% [4]). On the other hand, the ImageNet-CNN feature shows better performance on object-related databases. Importantly, our comparison

[2] Detailed experimental setups are included in the supplementary materials.

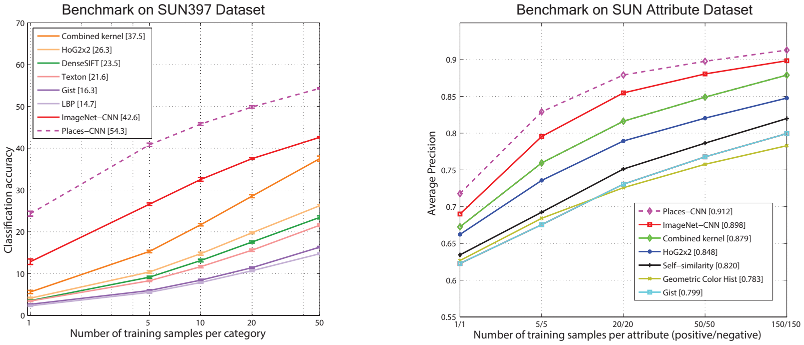

Figure 6: Classification accuracy on the SUN397 Dataset and average precision on the SUN Attribute Dataset with increasing size of training samples for the ImageNet-CNN feature and the Places-CNN feature. Results of other hand-designed features/kernels are fetched from [24] and [18] respectively.

Table 3: Classification accuracy/precision on various databases for Hybrid-CNN feature. The numbers in bold indicate the results outperform the ImageNet-CNN feature or Places-CNN feature.

SUN397	MIT Indoor67	Scene15	SUN Attribute	Caltech101	Caltech256	Action40	Event8
53.86±0.21	**70.80**	**91.59±0.48**	**91.56**	84.79±0.66	65.06±0.25	**55.28±0.64**	94.22±0.78

shows that Places-CNN and ImageNet-CNN have complementary strengths on scene-centric tasks and object-centric tasks, as expected from the benchmark datasets used to train these networks.

Furthermore, we follow the same experimental setting of train and test split in [1] to fine tune Places-CNN on SUN397: the fine-tuned Places-CNN achieves the accuracy of 56.2%, compared to the accuracy of 52.2% achieved by the fine-tuned ImageNet-CNN in [1]. Note that the final output of the fine-tuned CNN is directly used to predict scene category.

Additionally, we train a Hybrid-CNN, by combining the training set of Places-CNN and training set of ImageNet-CNN. We remove the overlapping scene categories from the training set of ImageNet, and then the training set of Hybrid-CNN has 3.5 million images from 1183 categories. Hybrid-CNN is trained over 700,000 iterations, under the same network architecture of Places-CNN and ImageNet-CNN. The accuracy on the validation set is 52.3%. We evaluate the deep feature (FC 7) from Hybrid-CNN on benchmarks shown in Table 3. Combining the two datasets yields an additional increase in performance for a few benchmarks.

5 Conclusion

Deep convolutional neural networks are designed to benefit and learn from massive amounts of data. We introduce a new benchmark with millions of labeled images, the Places database, designed to represent places and scenes found in the real world. We introduce a novel measure of density and diversity, and show the usefulness of these quantitative measures for estimating dataset biases and comparing different datasets. We demonstrate that object-centric and scene-centric neural networks differ in their internal representations, by introducing a simple visualization of the receptive fields of CNN units. Finally, we provide the state-of-the-art performance using our deep features on all the current scene benchmarks.

Acknowledgement. Thanks to Aditya Khosla for valuable discussions. This work is supported by the National Science Foundation under Grant No. 1016862 to A.O, ONR MURI N000141010933 to A.T, as well as MIT Big Data Initiative at CSAIL, Google and Xerox Awards, a hardware donation from NVIDIA Corporation, to A.O and A.T., Intel and Google awards to J.X, and grant TIN2012-38187-C03-02 to A.L. This work is also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7211 to A.T. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

[1] P. Agrawal, R. Girshick, and J. Malik. Analyzing the performance of multilayer neural networks for object recognition. In *Proc. ECCV*. 2014.

[2] Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2009.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

[4] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *In Advances in Neural Information Processing Systems*, 2013.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. 2014.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. 2008.

[7] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 2007.

[8] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[9] C. Heip, P. Herman, and K. Soetaert. Indices of diversity and evenness. *Oceanis*, 1998.

[10] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. `http://caffe.berkeleyvision.org/`, 2013.

[11] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva. Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psych Science*, 2010.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *In Advances in Neural Information Processing Systems*, 2012.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2006.

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1989.

[15] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Proc. ICCV*, 2007.

[16] A. Oliva. Scene perception (chapter 51). *The New Visual Neurosciences*, 2013.

[17] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int'l Journal of Computer Vision*, 2001.

[18] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proc. CVPR*, 2012.

[19] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Proc. CVPR*, 2009.

[20] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014.

[21] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *Int'l Journal of Computer Vision*, 2013.

[22] E. H. Simpson. Measurement of diversity. *Nature*, 1949.

[23] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proc. CVPR*, 2011.

[24] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR*, 2010.

[25] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Proc. ICCV*, 2011.