SScene recognition with CNNs: objects, scales and dataset bias

Liangjie Cao

Jun 16, 2018

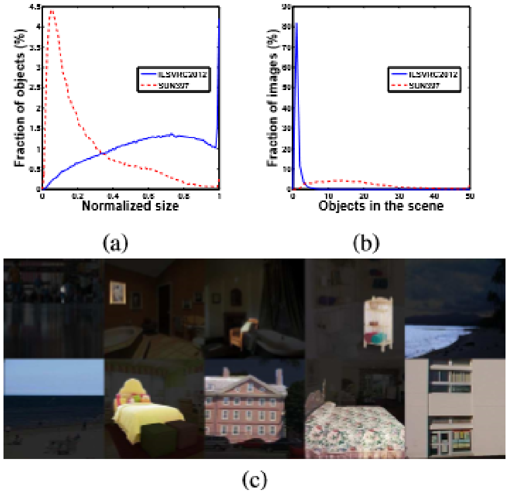
1. Scale sensitivity and object density

They trained a SVM classifier with 50 images per class, and tested on the remaining 50 images. The input feature was the output of the fc7 activation. The results are shown in Fig. 3. The authors use two variants: objects masked and objects with background (see Fig. 2). Regarding objects masked, where the background is removed, they can see that in general the performance is optimal when the object is near full size, above 70-80%. This is actually the most interesting region, with ImageNet-CNN performing slightly better than Places-CNN. This is interesting, since Places-CNN was trained with scenes containing more similar objects to the ones in the test set, while ImageNet-CNN was trained with the less related categories found in ILSVRC2012 (*e.g.* dogs, cats). However, as we saw in Fig. 1a, objects in ILSVRC2012 cover a large portion of the image in contrast to smaller objects in SUN397, suggesting that a more similar scale in the training data may be more important than more similar object categories. As the object becomes smaller, the performance of both models degrades similarly, again showing a limited robustness to scale changes.

Focusing now on the objects with background variant, the performance is worse than when the object is isolated from the background. This behaviour suggests that the background may introduce some noise in the feature and lead to poorer performance. In the range close to full object size, both ImageNet-CNN and Places-CNN have similar performance. However, as the object becomes smaller, and the content is more similar to scenes, Places-CNN has much better performance than ImageNet-CNN, arguably due to the fact it has learn contextual relations between objects and global scene properties. In any case, scales with low accuracy are probably too noisy and not suitable for their purpose.

2. Differences with previous works

The authors' architecture is similar to others proposed in previous multi-scale approaches [2, 8, 1], with the subtle difference of using scale-specific networks in a principled way to alleviate the dataset bias induced by scaling. The main emphasis in these works is on the way multi-scale features are combined, implemented as either VLAD or FV encoding, while leaving the CNN model fixed. While adding a BOW encoding layer can help to alleviate somewhat the dataset bias, the main problem is still the rigid CNN model. In contrast, their method addresses better the dataset bias related with scale and achieves significantly better performance, by simply adapting the CNN model to the target scale, even without relying to sophisticated pooling methods.



(a)

(b)

(c)

Figure 1. **Characteristics of objects in ILSVRC2012 (object data) and SUN397 (scene data): (a) distribution of objects sizes (normalized), (b) number of objects per scene, and (c) examples of objects by increasing normalized size**

They can also regard our approach as a way to combine the training data available in Places and ImageNet. This was explored previously by Zhou *et al.* [9], who trained a HybridCNN using the AlexNet architecture and the combined Places+ImageNet dataset. However, Hybrid-CNN performs just slightly better than Places-CNN on MIT Indoor 67 and worse on SUN397. We believe that the main reason was that this way of combining data from ImageNet and Places ignores the fact that objects found in both datasets in two different scale ranges (as shown in Fig. [1]). In contrast, our architecture combines the knowledge in a scale-adaptive way via either ImageNet-CNN or Places-CNN. Wu *et al.* [5] use Hybrid-CNN on patches at different scales. Again, the main limitation is that the CNN model is fixed, not adapting to the scale-dependent distributions of patches.

3. Experiments on scene recognition

In order to study the behaviour of ImageNet-CNNs and Places-CNNs in object recognition, they need object data extracted from scenes datasets. The authors selected 100 images per category from the 75 most frequent object categories in SUN397, so they can have enough images to train SVM classifiers. They took some precautions to avoid selecting too small objects. In contrast to most object and scene datasets, in this case they have the segmentation of the object within the scene, so they can use it to create variations over the same objects. Then they created four variations (see Fig. 2): original masked, original with background, canonical masked and canonical with background. In particular, to study the response to different scaling, the canonical variant is scaled in the range 10%-100%. Note how scaling the variant with background shifts progressively the content of the crop from object to scene.

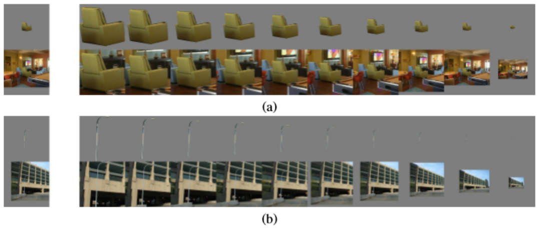


Figure 2. **The two variants used in the object recognition experiments: object masked (top row) and object with background (bottom row) with two examples of (a) armchair and (b) streetlight. Left crops show the object in the original scale in the scene. Right crops show the object scaled progressively from the canonical size (100%) down to 10%. All the images are centered in the object of interest**

4. Discriminability and redundancy

In this section they perform experiments directly over scene data, to evaluate the relation beween scale, training dataset and dataset bias by analyzing the scene recognition performance. Then they combine and evaluate multi-scale architectures.

They evaluate the proposed architectures with three widely used scene benchmarks. 15 scenes [3] is a small yet popular dataset with 15 natural and indoor categories. Models are trained with 100 images per category. MIT Indoor 67 [4] contains 67 categories of indoor images, with 80 images per category available for training. Indoor scenes tend to be rich in objects, which in general makes the task more challenging, but also more amenable to architectures using ImageNet-CNNs on patches. SUN397 [6, 7] is a larger scene benchmark (at least considered as such before Places) containing 397 categories, including indoor, man-made and natural categories. This dataset is very challenging, not only because of the large number of categories, but also because the more limited amount of training data (50 images per category) and a much larger variability in objects and layout properties. It is widely accepted as the reference benchmark for scene recognition. We consider seven scales in our experiments, obtained by scaling images between 227x227 and 1827x1827 pixels.

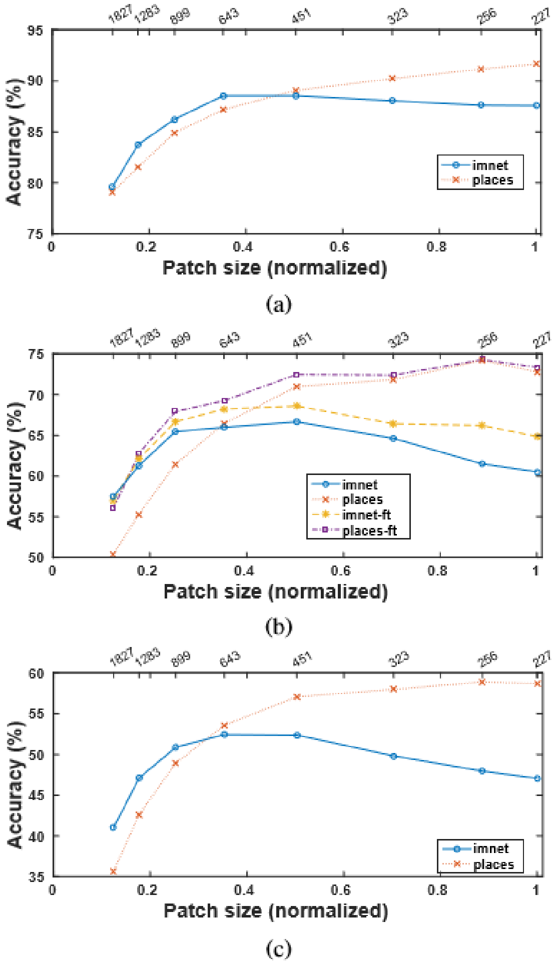


Figure 3. **Scene recognition accuracy for different scales: (a) 15 scenes, (b) MIT Indoor 67, and (c) SUN397**

5. Single scale

Average accuracy is a reasonable metric to evaluate a deep representation in the context of a classification task. For the different scales, they extracted fc7 activations locally in pacthes as features, and then trained SVMs. In addition to the seven scales evaluated, they included 256x256 pixels as a baseline, since off-the-shelf ImageNet-CNN and PlacesCNN are trained on this scale. The results for the three datasets are shown in Fig. 3, with similar patterns. PlacesCNN achieves the best performance when is applied globally at scene level (227x227 or 256x256), while rapidly degrades for more local scales. ImageNet-CNN exhibits a very different behaviour, with a more modest performance at global scales, and achieving optimal performance on patches at intermediate scales, and outperforming PlacesCNN at most local scales. These curves somewhat represent the operational curve of CNNs and the scale. In particular, the performance of ImageNet-CNN can be increases notably just by using an appropriate scale.

6. Accumlation of SCNN

Specifically, suppose we have a three-dimensional tensor K, in which $K_{i,j,k}$, i, j, k are recorded as the weights between the elements of channel i in the last slice and the elements of channel j in the current slice, and the offset between these two elements is k column. Similarly, $X_{i,j,k}$ is recorded as elements of tensor x, where i, j, k respectively refer to channels, rows, and columns. The forward calculation of SCNN is:

If j=1:

$$X^{'}_{i,j,k} = X_{i,j,k} \tag{1}$$

If j=2,3,...,H

$$X^{'}_{i,j,k} = X_{i,j,k} + f(\Sigma_m \Sigma_n X_i, j-1,^{'} k+n-1 \times K_{m,i,n}) \tag{2}$$

References

[1] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos. Scene classification with semantic fisher vectors. In *CVPR*, 2015.

[2] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, 2014.

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

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

[5] R. Wu, B. Wang, W. Wang, and Y. Yu. Harvesting discriminative meta objects with deep cnn features for scene classification. In *ICCV*, 2015.

[6] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. Sun database: largescale scene recognition from abbey to zoo. In *CVPR*, 2016.

[7] Y. Xiao, J. Wu, and J. Yuan. mCENTRIST: a multichannel feature generation mechanism for scene categorization. *IEEE TIP*, 2014.

[8] D. Yoo, S. Park, J. Lee, and I. S. Kweon. Multi-scale pyramid pooling for deep convolutional representation. In *CVPR*, 2015.

[9] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.