How Is Scene Recognition in a Convolutional Network Related to that in the Human Visual System?

Sugandha Sharma and Bryan Tripp(✉)

Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, Canada
{s72sharm,bptripp}@uwaterloo.ca

Abstract. This study is an analysis of scene recognition in a pre-trained convolutional network, to evaluate the information the network uses to distinguish scene categories. We are particularly interested in how the network is related to various areas in the human brain that are involved in different modes of scene recognition. Results of several experiments suggest that the convolutional network relies heavily on objects and fine features, similar to the lateral occipital complex (LOC) in the brain, but less on large-scale scene layout. This suggests that future scene-processing convolutional networks might be made more brain-like by adding parallel components that are more sensitive to arrangement of simple forms.

Keywords: Convolutional neural networks (CNNs) · Scene recognition · Human visual system

1 Introduction

It is remarkable that humans are able to perceive and interpret a complex scene in a fraction of a second, roughly the same time needed to identify a single object. When an image is briefly presented with less than 100 ms of exposure, observers usually perceive global scene information, e.g. whether the image was outdoor or indoor, well above chance. On the other hand, observers perceive details of objects with a couple of 100 ms more exposure time [2]. It has also been found that an exposure of 20–30 ms is enough for categorizing a scene as a natural or urban place [4]. However, it takes twice of that time to determine the basic level category of the scene, e.g. a mountain vs. a beach [3].

Studies in behavioral, computational and cognitive neuroscience suggest two complementary paths of scene perception in humans [7]. First, an object-centered approach, in which components of a scene are segmented and serve as scene descriptors (e.g., this is a street because there are buildings and cars). Second, a space-centered approach, in which spatial layout and global properties of the whole image or place act as the scene descriptors (e.g. this is a street because it is an outdoor, urban environment flanked with tall frontal vertical surfaces with squared patterned textures).

A.E.P. Villa et al. (Eds.): ICANN 2016, Part I, LNCS 9886, pp. 280–287, 2016.
DOI: 10.1007/978-3-319-44778-0_33

Several brain regions responsible for processing different scene properties have been identified, particularly the parahippocampal place area (PPA; a region of the collateral sulcus near the parahippocampal lingual boundary), the retrosplenial complex (RSC; located immediately behind the splenium of the corpus callosum), and the occipital place area (OPA; around the transverse occipital sulcus). PPA and RSC are most studied and respond preferentially to pictures depicting scenes, spaces and landmarks more than to pictures of faces or single movable objects [6].

While both PPA and RSC show selectivity to the spatial layout of the scene in various tasks, the responses of neither of them are modulated by the quantity of objects in the scene i.e., both regions are similarly active when viewing an empty room or a room with clutter [1].

The response of PPA is selective to different views of a panoramic scene, suggesting a view-specific representation in PPA. On the other hand, RSC seems to have a common representation of different views in a panorama, suggesting that RSC may hold a larger representation of the place beyond the current view [9]. However, scene representations in PPA have been found to be tolerant to severe transformations i.e., reflections about the vertical axis [7].

PPA and LOC represent scenes in an overlapping fashion. While PPA confuses scenes with similar spatial boundaries, regardless of the type of content, LOC confuses scenes with the same content, independent of their spatial layout [8]. LOC is not the only brain region involved in object processing, and thus multiple regions may represent different types of content and objects encountered in a scene [7].

Convolutional networks have many structural parallels with the visual cortex. Furthermore, they have recently begun to rival human performance in various vision tasks, including scene recognition as well as object recognition, stereoscopic depth estimation, etc. We would like to understand how similar the decision mechanisms of convolutional networks trained for scene recognition are to the corresponding mechanisms in the human cortex. As a first step, we analyze here the sensitivity of a scene-recognition network to certain input perturbations, to evaluate whether the network is more object-centred or space-centred.

2 Methods

We used the Places CNN [10], a network that has been previously trained for scene recognition on the Places205 dataset. The network has the same structure as [5]. It receives an image of a scene as input (e.g. the bedroom image in Fig. 1A). It has 205 outputs, corresponding to different scene categories. It is trained to output a high value for the category to which a given input image belongs (e.g. *bedroom*) and low values for other categories (e.g. *assembly line*, etc.).

2.1 Occlusion

We systematically occluded parts of the image in order to gauge how important different parts of the scene were for the network's prediction. To find which parts of an image were most important for the network, we slid a square occlusion window over an image. We set pixel values within the square to zero, passed the occluded image through the network, and recorded the output of the softmax output unit that corresponded to the correct category. In order to study the effect of objects of various sizes in the image, this procedure was repeated with squares of 9, 23, 39, 51, 87 and 113 pixels on a side. The full images had a fixed resolution of 227×227.

2.2 Blurring

We randomly selected 50 images from different categories in the Places205 test set and blurred them with a Gaussian filter of standard deviation varying between 0 and 13, in steps of 0.5. Thus there were 26 filtered images for each image in the original set of 50 images, leading to a total of 1300 images which were fed to the network. The output probabilities for correct predictions were normalized by dividing them by their maximum values across blur levels (typically the maximum occurred with zero blur). This was done to map the predictions for all the images to the same scale.

2.3 Spatial Boundaries

As discussed in Sect. 1, in the human visual system, PPA confuses scenes with similar spatial boundaries, regardless of the type of content, whereas the LOC makes the opposite errors, i.e. confusing scenes with the same content, independent of their spatial layout [7].

We conducted an experiment to explore whether the network resembles either PPA, LOC or both of them in terms of the kind of mistakes it makes. First, two categories having similar spatial boundary were selected, 'forest path' and 'corridor'. Ten images of each of these categories were selected, and the average predicted probability (average of probability that it's a forest/corridor over 10 images) for both categories was recorded. Then two categories having similar content were selected, 'classroom' and 'conference room', and the average predicted probability for both categories was also recorded. All the images for this experiment were taken from a Google images search, i.e. not from the Places205 dataset.

2.4 Panoramic Scenes

As discussed in Sect. 1, PPA is selective for different views of a panoramic scene, while the response of RSC has a common representation of different views in a panorama [9]. Motivated by this, we conducted an experiment to see whether the response of the network was selective for different views in a panoramic

scene. We collected 100 images from 12 different scene categories (images were obtained from a Google Images search), and split them up into left and right segments. These segments were then passed through the network and the correlation between the unit activations for the left and right segments were averaged over each layer and plotted as a function of layer number.

3 Results

3.1 Occlusion

Figure 1 shows heatmaps of occlusion effects over an image of a bedroom, for six different occlusion-window sizes. It is clear from the heatmaps that the bed is the most important object in the scene on which the model prediction is based. Moreover, occluding small parts of the bed has little impact on the model prediction, but occluding large areas has a large impact. This result was consistent with other experiments (not shown) in which we occluded various parts of the scene with unrelated pictures.

3.2 Blurring

Figure 2A visually shows the amount of blurring caused by the range of standard deviations used, on one of the sample images. Figure 2B shows the effect of

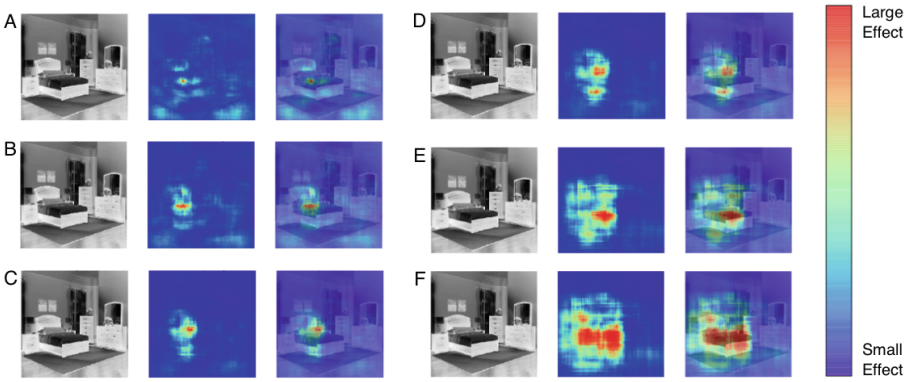

Fig. 1. Heatmaps of the effect of occlusion on the bedroom scene. The right image in each panel shows a heatmap superimposed on a black and white negative of the image. The image and the heatmap are also shown individually for clarity (left and centre; the left image is the same in each case). The red areas in the heatmap show the areas in the scene which are important for classification (the plotted values are the probabilities output by the *bedroom* node, with occlusion centred at the corresponding pixels; red is the lowest probability, or highest "effect" of occlusion). **A**–**F**: heat maps obtained by using occlusion windows of 9, 23, 39, 51, 87 and 113 pixels, respectively. (Color figure online)

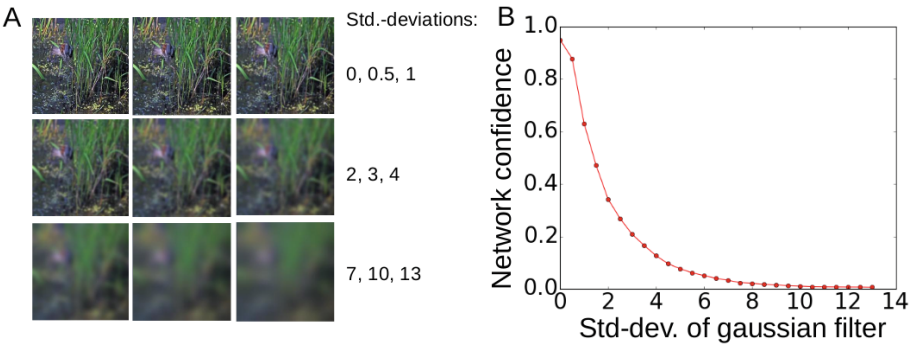

Fig. 2. Results of the blurring experiment. **A**: Effect of blurring on a sample image (shown for visual comparison). **B**: Effect of blurring on the confidence level of the network. The vertical axis shows the confidence level of the network normalized to lie within [0,1], averaged over 50 example images. The horizontal axis shows the standard deviation of the Gaussian filter used to blur the image (in pixels). (Color figure online)

blurring on the confidence level of the network (averaged over 50 randomly selected images). The confidence level of the model falls quickly with an increase in the standard deviation of the Gaussian filter. This shows that the model is not able to make predictions based on only the global features of the scenes, if it can't extract the local scene properties. This implies that the predictions of the network are based on local scene properties.

Table 1. Results from the spatial boundaries experiment. Integers indicate category indices (e.g. "Forest" is category 78).

	Forest [78] (Opponent: corridor)	Corridor [54] (Opponent: forest)	Classroom [44] (Opponent: conference room)	Conference room [51] (Opponent: classroom)
Categories predicted	[78, 78, 78, 78, 78, 78, 79, 78, 78]	[54, 54, 54, 54, 54, 54, 54, 54, 54, 54]	[51, 44, 44, 44, 44, 44, 51, 44, 44, 51]	[51, 51, 51, 51, 51, 51, 51, 51, 51, 51]
Avg. probability	0.570 (grayscale: 0.562)	0.892 (grayscale: 0.929)	0.612 (grayscale: 0.606)	0.733 (grayscale: 0.581)
Avg. probability (opponent)	2.935e-05 (greyscale: 1.327e-04)	8.343e-06 (greyscale: 4.869e-06)	0.159 (greyscale: 0.115)	0.018 (greyscale: 0.025)
Top 5 probability	[(0.570, 'forest_path 78'), (0.288, 'forest_road 79'), (0.043, 'rainforest 149'), (0.0362, 'bamboo_forest 16'), (0.0129, 'tree_farm 186')]	[(0.892, 'corridor 54'), (0.033, 'locker_room 144'), (0.025, 'lobby 113'), (0.015, 'hospital 94'), '(0.007, 'jail_cell 105')]	[(0.6122, classroom 44'), (0.159, 'conference_room 51'), (0.0586, 'conference_center 50'), (0.0448, 'cafeteria 37'), (0.0375, 'auditorium 12')]	[(0.733, 'conference_room 51'), (0.064, 'Conference_center 50'), (0.029, 'banquet_hall 17'), (0.025, 'dinette/home 70'), (0.021, 'office 129')]

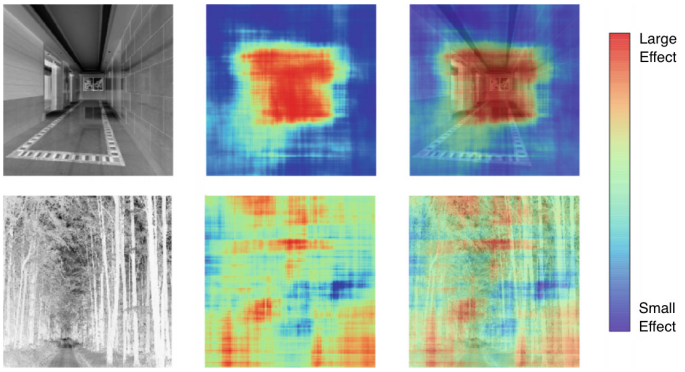

Fig. 3. Visualization of forest (bottom) and corridor (top) categories using a heatmap. The two rows show the negative of the scene on the left, heatmap in the middle and the heatmap superimposed on the scene on the right. The red areas are the most important for scene prediction. The important areas include distinguishing features of objects (e.g. tree trunks). (Color figure online)

3.3 Spatial Boundaries

We examined the extent to which the network confused scene categories with similar boundaries (specifically, forest paths and corridors) and categories with similar contents (classrooms and conference rooms).

The results are shown in Table 1. The network classified 3 of the 10 classrooms as conference rooms. However, it did not confuse forest paths and corridors. The average probability of the opponents is low for both forest and corridor, but higher for classroom and conference room. This suggests that the model confuses scenes with similar content but not the scenes with similar spatial boundaries. This was confirmed by looking at the top-5 predictions of the model. For example, for the 'forest' category, all top-5 predictions contained trees, but spatial boundaries varied (e.g. forest path vs. tree farm). Figures 3 and 4 also show the heatmaps for the four categories chosen in this experiment. The heatmaps suggest that the network is using objects to make its predictions. For example, in the classroom tables and chairs are important.

To test the extent to which colour differences accounted for the lack of confusion between forest paths and corridors, we repeated the tests with greyscale images. The results were similar to those with colour images (Table 1).

3.4 Panoramic Scenes

Figure 5 shows the correlations between the unit activations of the left and right segments of the panoramic scenes averaged over the units in each layer, over 100 different images. As expected, the average correlation is low for the input layers and increases for higher level layers.

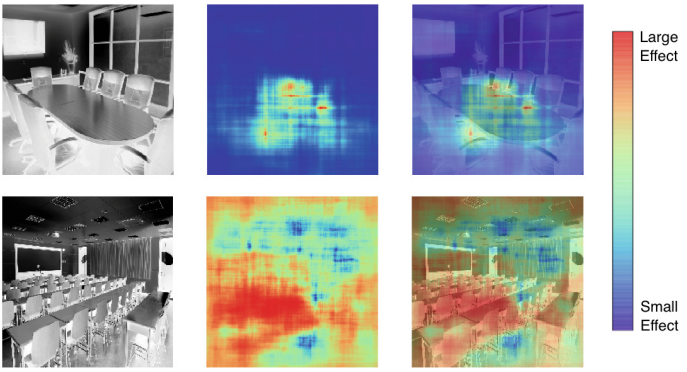

Fig. 4. Visualization of classroom (bottom) and conference room (top) categories using a heatmap. The two images show the negative of the scene on the left, heatmap in the middle and the heatmap superimposed on the scene on the right. The red areas are the most important for scene prediction. (Color figure online)

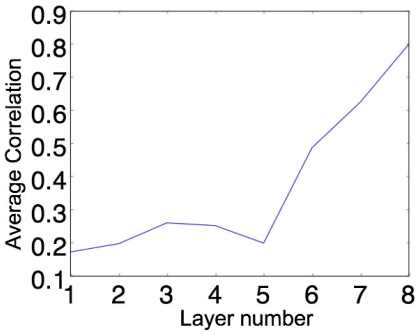

Fig. 5. Average correlation for left and right segments plotted as a function of layer number. Correlations were calculated between the activations of each unit in response to left and right parts of the panoramic images. The correlations of all units within a layer were then averaged to compute the average correlation for each layer. The later layers respond similarly to different views of each scene, similar to RSC.

4 Discussion

Our experiments suggest that the network is more object-centered (reliant on objects or local scene properties for its predictions) than space-centered (reliant on global scene properties). Its performance is impaired by occlusion of specific objects. It is sensitive to small amounts of blur (whereas humans can categorize scenes using very low spatial frequencies). This suggests that it is not able to make accurate predictions based only on the global scene properties, if it can't extract the local scene properties. Additionally, the network confuses scenes with similar content (objects, e.g. chairs etc.), but it does not confuse scenes with similar spatial boundaries but different textures. This further emphasizes the

importance of objects in a scene for accurate predictions, and suggests the relative insignificance of spatial layout for distinguishing different scenes. It would be worthwhile in future work to more specifically compare the effects of the same image manipulations on human and network performance.

It may be possible to make convolutional networks for scene recognition more robust, or at least more similar to the human visual system, by adding parallel components that are specifically trained to encourage space-centered representations. One possible approach would be to train such a parallel network on blurred images. The parallel networks might then complement each other in a way that is similar to the multiple scene processing regions in the human brain.

Acknowledgments. Supported by CFI & OIT infrastructure funds, the Canada Research Chairs program, NSERC Discovery grants 261453 and 296878, ONR grant N000141310419, AFOSR grant FA8655-13-1-3084 and OGS.

References

1. Epstein, R., Kanwisher, N.: A cortical representation of the local visual environment. Nature **392**(6676), 598–601 (1998)
2. Fei-Fei, L., Iyer, A., Koch, C., Perona, P.: What do we perceive in a glance of a real-world scene? J. Vis. **7**(1), 10–10 (2007)
3. Greene, M.R., Oliva, A.: The briefest of glances: the time course of natural scene understanding. Psychol. Sci. **20**(4), 464–472 (2009)
4. Greene, M.R., Oliva, A.: Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cogn. Psychol. **58**(2), 137–176 (2009)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25**, 1097–1105 (2012)
6. MacEvoy, S.P., Epstein, R.A.: Constructing scenes from objects in human occipitotemporal cortex. Nature Neurosci. **14**(10), 1323–1329 (2011)
7. Oliva, A.: Scene perception. In: Werner, J.S., Chalupa, L.M. (eds.) The New Visual Neurosciences, pp. 725–732. MIT Press, Cambridge (2014)
8. Park, S., Brady, T.F., Greene, M.R., Oliva, A.: Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. J. Neurosci. **31**(4), 1333–1340 (2011)
9. Park, S., Chun, M.M.: Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception. Neuroimage **47**(4), 1747–1756 (2009)
10. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. Adv. Neural Inf. Process. Syst. **27**, 487–495 (2014)