Retinotopy in CNNs implements Efficient Visual Search

Jean-Nicolas Jérémie (jean-nicolas.jeremie@univ-amu.fr)
Institut de Neurosciences de la Timone (UMR 7289)
CNRS/Aix-Marseille Université, 27, boulevard Jean Moulin
Marseille, 13005 France

Emmanuel Daucé (emmanuel.dauce@univ-amu.fr)
Ecole centrale Marseille, Marseille, France
Institut de Neurosciences de la Timone (UMR 7289)
CNRS/Aix-Marseille Université, 27, boulevard Jean Moulin
Marseille, 13005 France

Laurent U Perrinet (laurent.perrinet@univ-amu.fr)
Institut de Neurosciences de la Timone (UMR 7289)
CNRS/Aix-Marseille Université, 27, boulevard Jean Moulin
Marseille, 13005 France

Abstract

While foveated vision, a trait shared by many animals including humans, is a major contributor to biological visual performance, it has been underutilized in machine learning applications. This study investigates whether retinotopic mapping, a critical component of foveated vision, can enhance image categorization and localization performance when integrated into deep convolutional neural networks (CNN's). Retinotopic mapping was used to transform the inputs of standard off-the-shelf CNN's which were then retrained on the Imagenet task. Surprisingly, the networks with retinotopically-mapped inputs achieved a comparable performance in classification. Furthermore, the networks demonstrated improved classification localization when the foveated center of the transform was moved on the whole image. This replicates a crucial ability of the human visual system that is absent in typical CNN's. These findings suggest that retinotopic mapping may be fundamental to significant preattentive visual processes, in particular the retinotopic version seems to be the best option when applying one of these networks to a visual search task.

Keywords: Foveated vision; Convolutional Neural Networks; Transfer learning; Visual categorization; Neuromorphic transformation

Introduction

The visual system in humans and many mammals is characterized by a substantial resolution disparity between the central area of the visual field (fovea) and the peripheral regions, where the number of photoreceptors decreases exponentially with eccentricity (Polyak, 1941). This particular topographic arrangement, which transform the spatial relationships of visual inputs, is a fundamental component of processing in species such as carnivores or primates (Kaas, 1997), suggesting that they confer evolutionary advantages. Indeed in the visual search task, that consists of finding an object of interest in a visual scene, the fovea is associated with a set of oculomotor behaviors aimed at positioning objects of interest at the center of the retina maximizing access to visual information for those objects. Here, we propose to take advantage of this biologically inspired pre-processing approach, we hypothesize that this biological mapping will improve the performances of CNN's in visual search tasks.

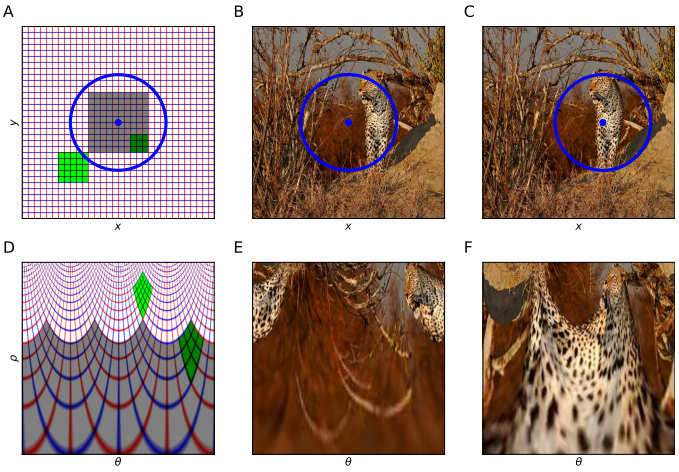

Figure 1: **(A)**, the input image is defined as a regular grid representing the cartesian coordinates (x, y) by vertical (red) and horizontal (blue) lines. As shown in **(D)**, by applying the log-polar transform to this image, the coordinates of each pixel with respect to the fixation point are transformed based on its azimuth angle θ (abscissa) and the logarithm of its eccentricity $\rho = \log(\sqrt{x^2 + y^2})$ (ordinates). When the transformation is applied to a natural image, as shown in **(B)**, there is a noticeable compression of information in the periphery, as shown in **(E)**. As shown in **(F)**, this representation is highly dependent on the fixation point, as indicated by the shift shown in **(C)** when the fixation point is moved to the right and down.

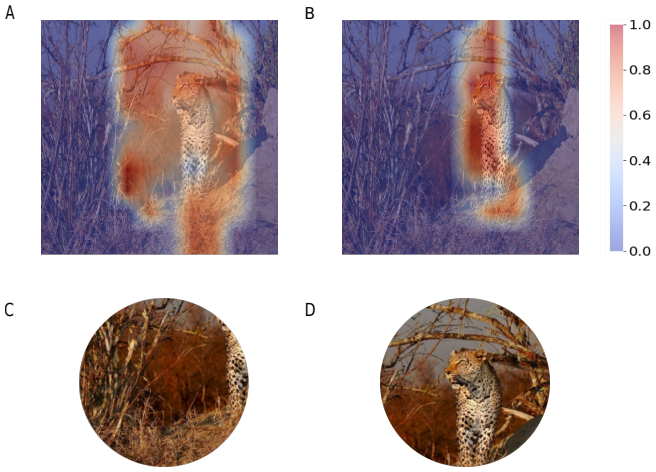

Figure 2: Accuracy maps computed on a sample from the Imagenet data set, using a Resnet101 networks trained and tested either **A** on regular images or **B** on images mapped on the retinotopic space using a log-polar grid. Red predicts the presence of the label : "leopard", blue its absence. Point of view corresponding to the maximum activation of the map in the cartesian **C** or retinotopic **D** referential.

Methods

Retinotopy and Transfer Learning

We implement retinotopic mapping with the well known log-polar referential (Araujo & Dias, 1997). By applying a transformation from the regular cartesian pixel grid to a log-polar grid (see Figure 1). We choose transfer learning (Mari, 2020) to retrain two version of state-of-the-art CNN's Resnet50 and Resnet101 (He, Zhang, Ren, & Sun, 2015). In order to initiate those networks for the visual search task, we generated a data set of fixation point defined as the center of the bounding box of the label of interest from the Imagenet (Russakovsky et al., 2015) data set. This new dataset was then used to train both retinotopic and cartesian networks. Networks then are tested for localization against the original Imagenet dataset.

Likelihood map protocol

We sub sampled each image with a grid of equidistant viewpoints at a resolution of 11×11. At each viewpoint, the largest possible sample is cropped. Thus a minimum size of 224×224 at the border and the whole image at the center. From the cartesian or retinotopic reference frame, this sample is then resized to a 224×224 resolution to match the optimal size for the CNN before processing. We selected some key metrics to compare the retinotopic or cartesian referential. An indicator of the correct map activation position is the Energy-Based Pointing Game accuracy (Selvaraju et al., 2019) where the localization is a success when the peak activation of the heat map of a given label is located inside the ground true box. In addition, we chose to track the ratio of the activation inside to outside the box, the higher, the greater is the contrast of the heat map.

Table 1: Accuracy and mean activation ratio over the Imagenet validation data set. A saccade correspond to the selection of another fixation point in the grid. Before the saccade the networks process the whole image.

	ResNet50		ResNet101	
	Cartesian	Retinotopic	Cartesian	Retinotopic
Ratio Activation	1.33	**1.45**	1.23	**1.41**
Pointing Game	0.47	**0.60**	0.41	**0.58**
Before saccade	0.67	**0.68**	0.70	**0.71**
Saccade no prior	0.65	**0.69**	0.64	**0.71**
Saccade prior	0.93	**0.94**	0.94	**0.95**

Results

Accuracy maps as a proxy for saliency

Compared to the accuracy map generated with images in cartesian space (see Figure 2-A & B), the accuracy maps in retinotopic space provide a more focused localization of the object of interest. Although the leopard's position is clearly visible on both maps, the retinotopic version is less noisy than the cartesian version. This is highlighted Figure 2-C & 2-D, where we can see that the maximum activation corresponds to the leopard in the retinotopic version. However, when comparing metrics, whether in terms of positioning with the pointing game, or response contrast with the ratio of activation, networks exploiting the properties of the retinotopic space outperform those in cartesian space. Surprisingly Resnet50 tend to perform better then Resnet101 on these indicators.

Visual search task

Here we compared the average network accuracy as a function of the fixation point. By selecting the fixation point (saccade), either the most salient to the target label ('priors') or the most salient independent to the target label ('no priors'). The results indicate that networks using the retinotopic reference frame appear to maximize prediction after moving the fixation point, as the mean accuracy is better in this test. Note that since the networks are trained on the boxes, their pre-saccadic acuity is a test in itself, as they are presented with the full image. Although cartesian networks perform better in validation during training, they lose their advantage on the full image.

Conclusion

In this study, we have shown promising computational results for the localization of visual objects, In particular, it shows that we can use a network retrained with a retinotopic map with high information compression in the periphery to perform categorization and localization tasks. Finally, the implementation of this refined localization of a label of interest could allow us to extend this study to a more complex task (i.e.,visual search), the accuracy maps could provide the underlying pre-attentive mechanisms on which its effectiveness seems to depend and that can be compared with physiological data (Crouzet, 2011).

Acknowledgments

Authors received funding from the ANR project number ANR-20-CE23-0021 (”AgileNeuroBot”) and from the french government under the France 2030 investment plan, as part of the Initiative d’Excellence d’Aix-Marseille Université – A*MIDEX grant number AMX-21-RID-025 “Polychronies”. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Araujo, H., & Dias, J. (1997). An introduction to the log-polar mapping. *Proceedings II Workshop on Cybernetic Vision*(1), 139–144. Retrieved from http://ieeexplore.ieee.org/document/629454/ (00000) doi: 10.1109/CYBVIS.1996.629454

Crouzet, S. M. (2011). What are the visual features underlying rapid object recognition? *Frontiers in Psychology*, *2*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December). Deep Residual Learning for Image Recognition. Retrieved 2023-07-20, from http://arxiv.org/abs/1512.03385 (336 citations (INSPIRE 2023/7/20) 336 citations w/o self (INSPIRE 2023/7/20) arXiv:1512.03385 [cs.CV]) doi: 10.1109/CVPR.2016.90

Kaas, J. H. (1997, January). Topographic Maps are Fundamental to Sensory Processing. *Brain Research Bulletin*, *44*(2), 107–112. Retrieved 2023-09-27, from https://www.sciencedirect.com/science/article/pii/S0361923097000944 doi: 10.1016/S0361-9230(97)00094-4

Mari, B. T. R. I. J. S. M. K. N., Andrea. (2020, October). Transfer learning in hybrid classical-quantum neural networks. *Quantum*, *4*, 340. Retrieved 2021-05-18, from http://arxiv.org/abs/1912.08278 doi: 10.22331/q-2020-10-09-340

Polyak, S. L. (1941). The retina.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, *115*, 211–252.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. , *128*(2), 336–359. Retrieved 2024-03-14, from http://arxiv.org/abs/1610.02391 doi: 10.1007/s11263-019-01228-7