Retinotopy in CNNs implements Efficient Visual Search

Jean-Nicolas JÉRÉMIE[1] **Emmanuel DAUCÉ**[1,2] **Laurent Udo PERRINET**[1]

[1] *Institut de Neurosciences de la Timone*, CNRS / Aix-Marseille Université, Marseille, France. ,
[2] *École Centrale Marseille*, France.

Retinotopic mapping

The distribution of photoreceptors on the retina defines the organization of the visual field. This is known as the Retinotopic map. In the cortex, visual information is organized around two polar coordinates, azimuth (angle) and eccentricity (distance from the center), with over-representation around the center, the *fovea*, an area of high resolution in comparison to the periphery.

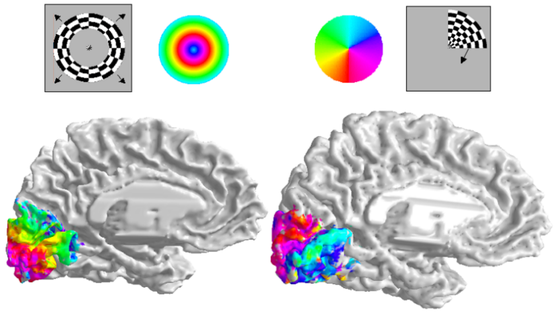

Figure 1: Retinotopy in the human early visual cortex transforms images into a polar reference map with an over-representation of the center (Dougherty et al., 2003).

We transform input images defined in Cartesian coordinates toward Retinotopic coordinates using a log-polar transformation (Araujo and Dias, 1997). The center of fixation is represented as a blue dot, and the blue circle represents the region of maximal magnification under the Retinotopic transformation.

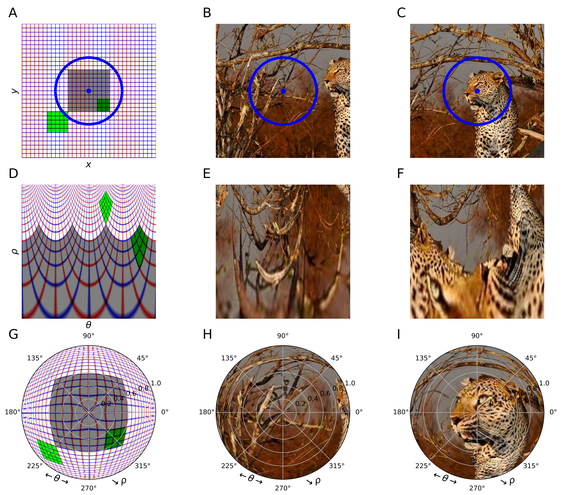

Figure 2: Retinotopy as implemented by a log-polar mapping.

Object Categorization using CNNs

We address the role of the Retinotopic transformation in visual processing, and compare its properties with state-of-the-art (Cartesian) image processing. ResNet 101 (He et al., 2015) Convolutional Neural Networks (CNNs) are trained to identify objects in an image, either with Cartesian or log-polar input, providing around 80% accuracy on ImageNet dataset (Russakovsky et al., 2015) in both cases. Here, we validate the change in response accuracy when rotating the input image. NB: this transformation corresponds to a translation in log-polar space, a transformation for which CNNs are known to be robust.

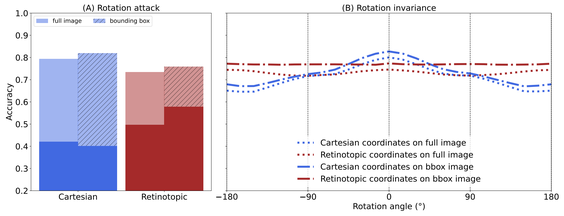

Figure 3: Rotation-based attack and accuracy over the validation set of the ImageNet dataset.

In contrast, the retinotopic transformation is highly sensitive to translations. This explains the need for the fovea to be centered on the object of interest to maximize the classification accuracy. A second generation of CNNs was trained using bounding boxes (containing the object of interest) from the ImageNet dataset. To assess this greater sensibility to translation, a saliency map was calculated using a regular grid ($n \times n$ points of fixation). The likelihood of the image label is then displayed in color code for each position.

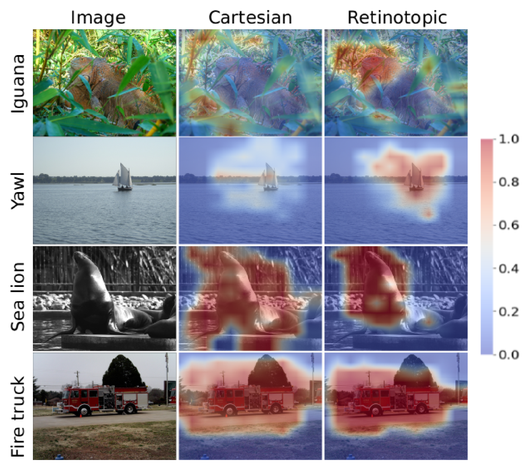

Figure 4: Likelihood as a proxy for saliency map.

Visual Object Localization

By selecting a number of key metrics, we undertake a comparative analysis of the efficiency in the localization of objects for the odels learned in the Retinotopic and Cartesian reference frames, respectively. The results demonstrate that the Retinotopic reference-based network dislays a more pronounced and sustained response to the object of interest in comparison to the state-of-the-art network.

	Cartesian	Retinotopic
Activation ratio	+ 23%	**+ 41%**

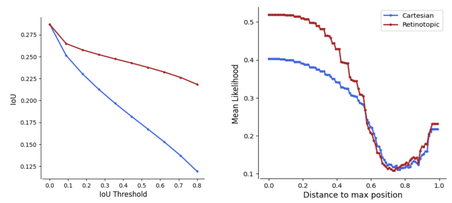

Figure 5: Quantifying the localization performance of models.

In comparison to the accuracy map generated with images in Cartesian space (see Figure 6-A & B), the accuracy maps in Retinotopic space provide a more focused localization of the object of interest. While the position of the leopard is clearly discernible in both maps, the Retinotopic version exhibits less noise than the Cartesian version. This is illustrated Figure 6-C & 6-D, which depicts the maximum activation corresponding to the leopard in the Retinotopic version. To quantify this effect, we employ the Pointing Game metric.

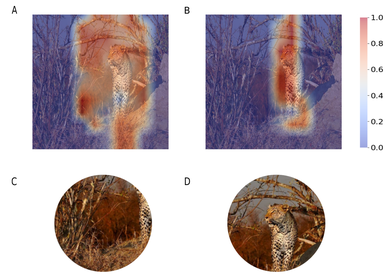

Figure 6: Sample saccade to maximum likelihood position.

	Cartesian	Retinotopic
Pointing Game	41%	**58%**
Before saccade	70%	**71%**
Saccade with no prior	64%	**71%**
Saccade with a prior	94%	**95%**

The mean network accuracy is evaluated as a function of the fixation point, with the fixation point (saccade) selected for categorization. This is achieved either by focusing on the most salient item associated with the target label (priors) or by focusing on the most salient item that is not associated with the target label (no priors). The results demonstrate that networks utilizing the Retinotopic reference frame appear to optimize prediction following the movement of the fixation point, as evidenced by the mean accuracy being superior in this test.

Toward a new neuromorphic model

Taking inspiration from natural vision systems Mishkin and Ungerleider, 1983, we will develop a model that builds over the anatomical visual processing pathways observed in mammals, namely the "What" and the "Where" pathways Daucé and L. Perrinet, 2020. It operates in two steps, one by providing a detailed categorization over the detailed "foveal" selected region attained as described in this work ("What"), and the second by selecting a region of interest ("Where"), before knowing its actual visual content, through an ultrafast/low resolution analysis of the full visual field which allow the model to afford a saccade (Yarbus, 1961).

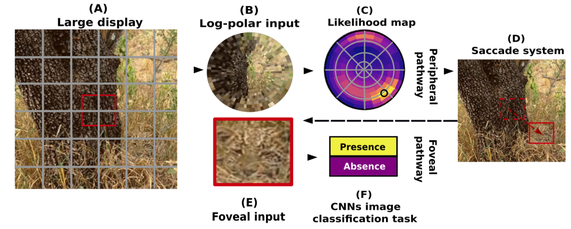

Figure 7: Modelling the dual visual processing pathways observed in primates, namely the "What" and the "Where" pathways by allowing the model to saccade.

References

Araujo, H. and J.M. Dias (1997). "An introduction to the log-polar mapping". In: *Proceedings II Workshop on Cybernetic Vision* 1. 00000, pp. 139–144. DOI: 10.1109/CYBVIS.1996.629454.

Daucé, Emmanuel and Laurent Perrinet (2020). "Visual Search as Active Inference". en. In: *Active Inference*. Ed. by Tim Verbelen et al. Communications in Computer and Information Science. 00001. Cham: Springer International Publishing, pp. 165–178. DOI: 10.1007/978-3-030-64919-7_17.

Dougherty, Robert F. et al. (Oct. 24, 2003). "Visual field representations and locations of visual areas V1/2/3 in human visual cortex". In: *Journal of Vision* 3.10, p. 1. DOI: 10.1167/3.10.1.

He, Kaiming et al. (Dec. 2015). "Deep Residual Learning for Image Recognition". In: 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.

Jérémie, Jean-Nicolas, Emmanuel Daucé, and Laurent U. Perrinet (Feb. 23, 2024). *Retinotopic Mapping Enhances the Robustness of Convolutional Neural Networks*. DOI: 10.48550/arXiv.2402.15480. arXiv: 2402.15480 [cs, q-bio]. preprint.

Mishkin, Mortimer and Leslie Ungerleider (1983). "Object vision and spatial vision: Two cortical pathways". In: *Trends in Neurosciences* 6. 03342 tex.date-modified: 2021-06-17 16:11:48 +0200, pp. 414–7. DOI: 10.1016/0166-2236(83)90190-x.

Russakovsky, Olga et al. (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115, pp. 211–252.

Yarbus, A (1961). "Eye Movements during the Examination of Complicated Objects". In: *Biofizika* 6(2), pp. 52–56.

Poster B159 presented at Cognitive Computational Neuroscience - August 6–9, 2024 - Boston, Massachusetts. Authors received funding from the ANR project N° ANR-20-CE23-0021 ("AgileNeuroBot"). Get more material at https://laurentperrinet.github.io/publication/jeremie-24-ccn/.