Ultra-rapid visual search in natural images using active deep learning

Jean-Nicolas JEREMIE Emmanuel DAUCÉ Laurent PERRINET

Institut de Neurosciences de la Timone

I. Ultra-fast vision

Visual search, that is, the simultaneous localization and detection of a visual target of interest, is a vital task. Biological visual systems are able to perform such detection efficiently [1]. A distinctive aspect of the vision of species like primates is that it is foveal. To solve the visual task, it seems to rely on two parallel streams known to provide information about "What" they are looking at or "Where" to look. We infer that this may be one essential ingredient in that efficiency.

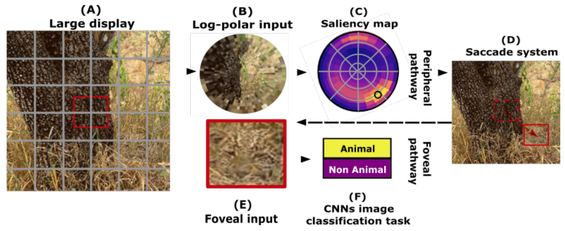

Figure 1: A dual-pathway model implementing saccades in natural images.

Taking inspiration from natural vision systems, we develop here a model 1 that builds over the anatomical visual processing pathways observed in mammals, namely the "What" and the "Where" pathways [2] to solve a ecological visual task.

II. Transfer Learning

Transfer Learning is a method that takes advantage of the knowledge accumulated on a problem to *transfer* it to a different but related problem.

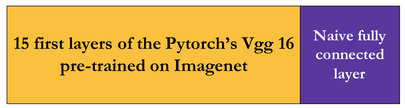

Figure 2: Network's architecture with VGG16 [3] backbone

We have shown that when we could re-train VGG networks using transfer learning, so that it can be applied to an ecological task [4]. The network achieves accuracies similar to those found in psycho-physics and we found the categorization of the networks to be robust to transformation like rotation, reflection or grayscale filtering [5].

III. Retinotopic mapping

Here, we define a retinotopic log-polar mapping transforming the regular pixel grid into a grid resembling that found in some animal species and such that visual information is concentrated in the center of gaze.

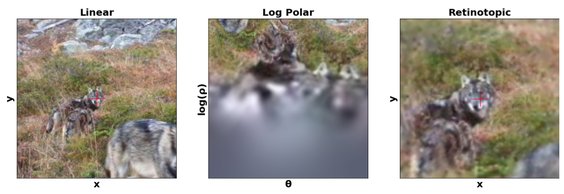

Figure 3: (Linear) An example input image with the center of fixation denoted by a red cross. (Log-polar) projection of the coordinates of each pixel of the input image according to its angle of azimuth θ from the horizontal axis on the x-axis and the logarithm of its eccentricity (or radius) ρ with respect to the fixation point on the y-axis. (Retinotopic) reconstruction from the log-polar mapping.

IV. Saliency maps

We define a saliency map as the positions of different fixation point in the image for which the prediction of the categorization of the presence of an animal is above $p = 0.5$.

Figure 4: Saliency maps with or without applying the retinotopic mapping transformation (see Figure 3) to inputs.

In both case, these maps allow us to extract the regions of interest including the key features necessary for the categorization of an animal by our network. Note the finer contour around the area of interest for the heat map generated with a reconstructed input and a network trained to recognize an animal in a linear space (see Figure 3).

V. Collicular saliency maps

To define the collicular saliency maps, we followed the same protocols as in Figure 4 with a retinotopic-like spacing of the fixation points. For each image, we can compute a vector (16×16) of the predictions of our model 5

Figure 5: Examples of 16×16 collicular saliency maps obtained with the label "animal".

Each image is then associated to a 16×16 matrix where each element represents the prediction of an animal by the network to compose a dataset dedicated to the training of the "Where" network.

VI. Training the "Where" network

Based on the maps produced by our categorization network, we train a 4-layer convolutional network producing a 16×16 output from a retinotopic input with a central fixation point.

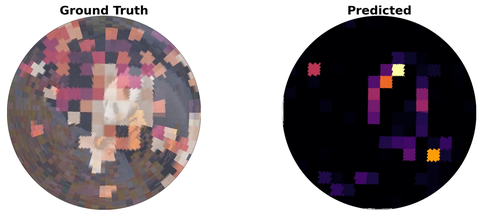

Figure 6: Example of data pair used to supervised the training of the "Where" network on a simplified task. (Ground truth) Collicular saliency map of the "What" network exposing its predictions for different fixation points in the image, projected into retinotopic space. (Predicted) Output from a fully connected "Where" network after 10 training epochs.

Preliminary results on the task with natural images seem promising even if some adjustments of the network architecture and the protocol used for its training seem necessary.

VII. Task dependence

For each image, the saliency maps depend on the categorization for which the "What" network is trained.

Figure 7: The saliency maps depend on the network used. (A) networks train to recognize the label "animal" . (B) networks train to recognize the label "dog". (C) networks train to recognize the label "cat" in an image.

This allows us to model visual search paths based on the symbol of interest. The human gaze maybe influenced by attention mechanisms [6]. These networks could be a efficient tools in the characterization of the different strategies used by humans to efficiently solve this task.

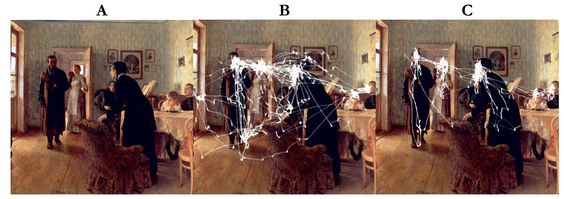

Figure 8: (A) The painting "An Unexpected Visitor". Scan paths in different conditions: (B) while the subjects freely explore the scene, (C) while asking the observer to asses ages of characters (from [6]).

VIII. References

1. Thorpe, S. & Fabre-Thorpe, M. Neuroscience. Seeking Categories in the Brain. *Science (New York, N.Y.)* **291,** 260–263. doi:10.1126/science.1058249 (Jan. 2001).
2. Daucé, E., Albigès, P. & Perrinet, L. U. A dual foveal-peripheral visual processing model implements efficient saccade selection. *Journal of Vision* **20,** 22–22. doi:10.1167/jov.20.8.22 (June 5, 2020).
3. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale. *arXiv:1409.1556 [cs]*. doi:https://doi.org/10.48550/arXiv.1409.1556. arXiv: 1409.1556 [cs] (Apr. 2015).
4. Jérémie, J.-N. & Perrinet, L. U. Ultra-Fast Image Categorization in Vivo and in Silico. doi:10.48550/arXiv.2205.03635. arXiv: 2205.03635 [cs, q-bio] (May 12, 2022).
5. Rousselet, G. A., Macé, M. J.-M. & Fabre-Thorpe, M. Is It an Animal? Is It a Human Face? Fast Processing in Upright and Inverted Natural Scenes. *Journal of Vision* **3,** 440–455. doi:10.1167/3.6.5 (2003).
6. Yarbus, A. Eye Movements during the Examination of Complicated Objects. *Biofizika* **6(2),** 52–56 (1961).

Poster presented at AREADNE 2002 in Santorini, Greece. Authors received funding from the ANR project N° ANR-20-CE23-0021 ("AgileNeuroBot"). Get more material at https://laurentperrinet.github.io/publication/jeremie-22-areadne/.