Unravelling the relationship between location and categorisation improves convolutional neural networks

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

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

I. Methods

We propose a general framework to locate an object of interest, with or without label, based on CNNs categorisation (here Resnet networks He et al., 2015). We generate a grid of 7×7 fixation points to match the native resolution of Grad-Cam and DFF methods. Thus for a given image, the corresponding batch produces a $7 \times 7 \times 1000$ tensor (the 'label" map), where the last dimension represents the 1000 labels of the ImageNet Challenge Russakovsky et al., 2015.

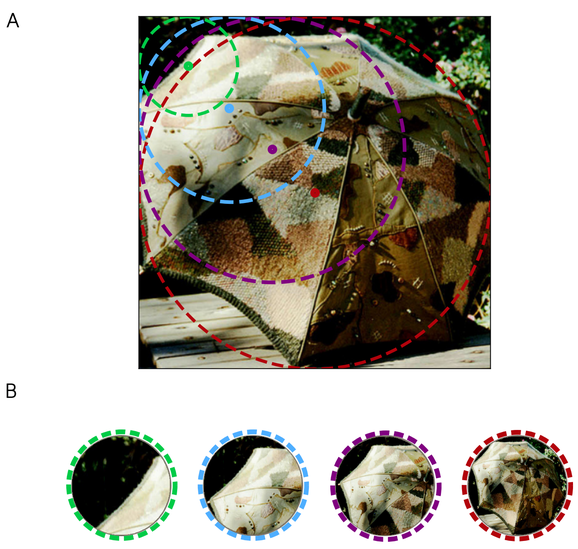

Figure 1: Dissecting an image into different *views*

We choose to implement this transformation for its known properties, indeed it has been shown that this transformation results in an overrepresentation of the central area and a deformation of the visual space, the latter being highly dependent on the viewpoint enhanced the localisation properties of the network while reducing the impact of the background Jérémie, Daucé, and Perrinet, 2024

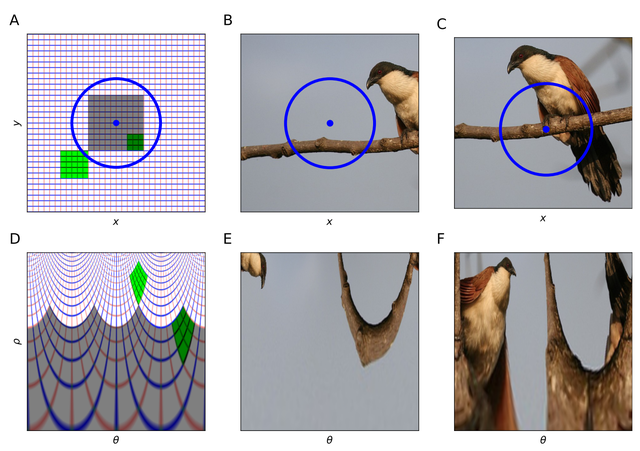

Figure 2: Samples of our Log-polar transform

II. Map samples

Many studies have attempted to enhance the performance of convolutional neural networks (CNNs) by increasing model complexity, adding parameters, or adopting alternative architectures. Our approach differs in that we prioritise ecological plausibility in order to achieve high accuracy with minimal computational cost.

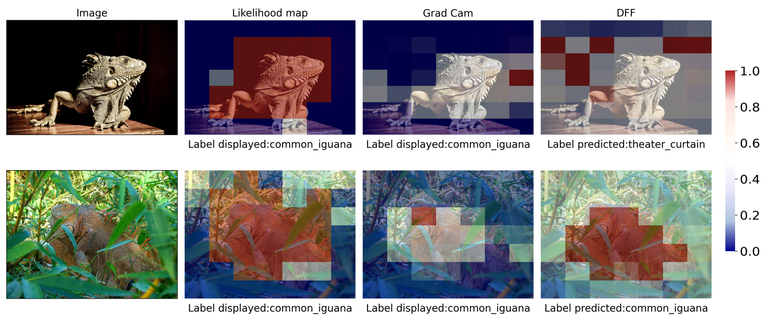

Figure 3: Display of different saliency maps produced by different methods using the same CNNs. **From left to right** : Likelihood map, Grad-Cam maps (Selvaraju et al., 2020), Deep Feature factorisation (DFF; (Collins, Achanta, and Süsstrunk, 2018))

III. Visual search

For each position u, a hypothesis is formed about the visual content of the view, it takes the form of a probability distribution $p(k|\boldsymbol{x}_u)$, which assigns a probability to each label $k \in 1, \ldots, K$, such that $\sum_k p(k|\boldsymbol{x}_u) = 1$. If the label is known : (the "visual search" task), the target label k^* is known in advance, hence it is possible to extract the set of views thus produces a *likelihood map* $L^*(u)$, assigning a score to the target label at each spatial position.

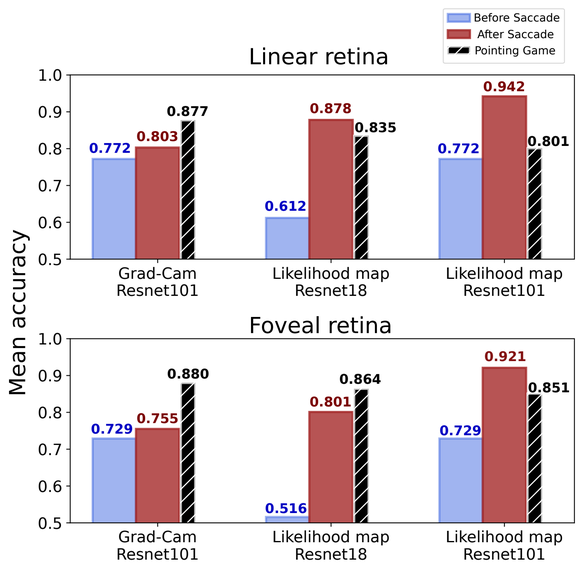

Figure 4: Label-driven visual search. Average accuracy when processing all images from the ImageNet *validation* data set.

If the label is not known :The most direct extension of the previous method is to select the label with maximum logit score for each position u :

$$L(u) = \arg\max_k logit(p(k|\boldsymbol{x}_u))$$

In the framework of *active inference*, it has been proposed (see Daucé (2018) and Daucé and Perrinet (2020)) to consider the *information gain* as a measure of the relevance of a view, relative to the initial glimpse of the scene, denoted $\boldsymbol{x}_0$. It has been shown that the information gain can be upper-bounded (up to a constant) which is much simpler to compute :

$$IGUB(u) \equiv \sum_k p(k|\boldsymbol{x}_0, \boldsymbol{x}_u) \log p(k|\boldsymbol{x}_u)$$

This value reflects an *optimistic bias* on the information gain.

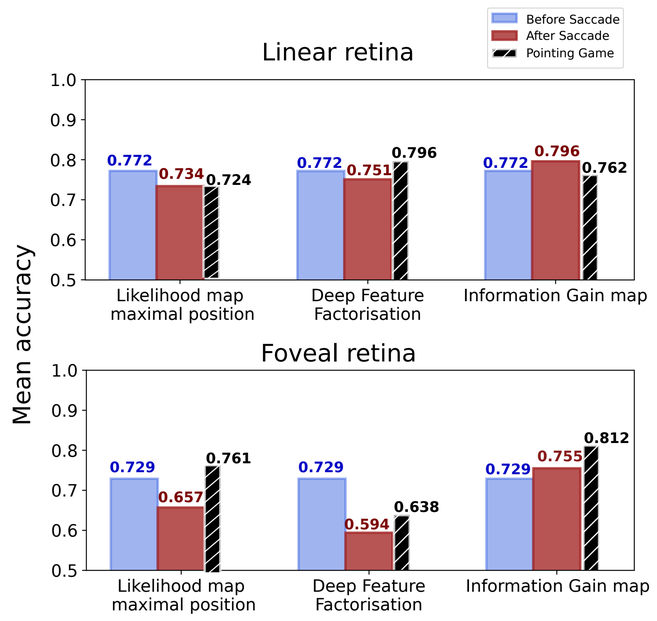

Figure 5: Uncued visual search. Average accuracy when processing all images from the ImageNet *validation* data set (same data set used in Figure 4).

IV. Active vision

The aim of this study was to explore the relationship between localisation and categorisation, with the ultimate goal of identifying the optimal viewpoint at which a given network's categorisation accuracy is maximised.

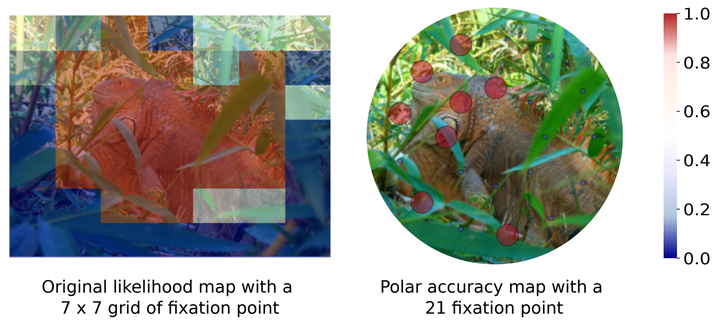

Figure 6: Sample of the new polar binary fixation point data

The discovery of the *optimal point of fixation* provided a crucial mechanism for generating training targets. Each supervision pair included a polar retinotopic input and a matrix encoding the ventral CNN's categorisation accuracy at various fixation points. This process allowed us to establish an active vision framework, which is instrumental in investigating the efficiency of saccades.

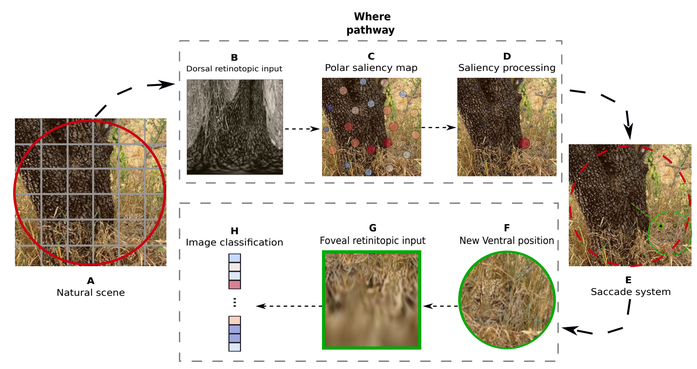

Figure 7: Pipeline of the proposed dual-pathway visual search model : *A*: Original full-resolution input image, with a red circle denoting the region of the polar transformation. *B*: Grayscale dorsal input in a retinotopic (polar) representation. *C*: Likelihood map generated by the dorsal pathway, indicating salience at each polar fixation coordinate. *D*: Fixation point selection based on either maximum likelihood or refined strategies such as the information gain protocol. *E*: Implementation of a simulated saccade mechanism to reposition the ventral input. *F*: Newly selected ventral input window. *G*: Transformation of this window into the foveal retinotopic referential. *H*: Final output of the ventral network, displaying the predicted category at the chosen fixation point.

V.

- Collins, Edo, Radhakrishna Achanta, and Sabine Süsstrunk (2018). *Deep Feature Factorization For Concept Discovery*. DOI: 10.48550/arXiv.1806.10206. arXiv: 1806.10206[cs].
- Daucé, Emmanuel (2018). "Active fovea-based vision through computationally-effective model-based prediction". In: *Frontiers in neurorobotics* 12, p. 76. DOI: 10.3389/fnbot.2018.00076.
- Daucé, Emmanuel and Laurent U Perrinet (Dec. 17, 2020). "Visual search as active inference". In: *IWAI 2020*. DOI: 10.1007/978-3-030-64919-7_17.
- He, Kaiming et al. (2015). "Deep Residual Learning for Image Recognition". In: *arXiv:1512.03385 [cs.CV]*. DOI: 10.1109/CVPR.2016.90. arXiv: 1512.03385 [cs.CV].
- Jérémie, Jean-Nicolas, Emmanuel Daucé, and Laurent U Perrinet (2024). *Foveated Retinotopy Improves Classification and Localization in CNNs*. DOI: 10.48550/arXiv.2402.15480. arXiv: 2402.15480[cs].
- Russakovsky, Olga et al. (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115, pp. 211–252. DOI: 10.1007/s11263-015-0816-y.
- Selvaraju, Ramprasaath R. et al. (Feb. 2020). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *International Journal of Computer Vision* 128.2, pp. 336–359. DOI: 10.1007/s11263-019-01228-7. arXiv: 1610.02391[cs].

Poster presented at CCN 2025 in Amsterdam, The Netherlands (Paper #A145). 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". This work was granted access to the HPC resources of Aix-Marseille Université financed by the project Equip@Meso (ANR-10-EQPX-29-01) of the program "Investissements d'Avenir" supervised by the Agence Nationale de la Recherche.