Unravelling the relationship between location and categorisation improves convolutional neural networks

Jean-Nicolas Jérémie
Institut de Neurosciences de la Timone
Aix-Marseille Université
CNRS UMR 7289, Marseille, France

Emmanuel Daucé
Institut de Neurosciences de la Timone
Aix-Marseille Université
CNRS UMR 7289 &
École Centrale Méditerranée
Marseille, France

Laurent U Perrinet
Institut de Neurosciences de la Timone
Aix-Marseille Université
CNRS UMR 7289, Marseille, France

Abstract

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. We focus on visual search, which requires the localisation and categorisation of a target object in natural scenes. Due to the inhomogeneity of foveal retinotopy in human visual representations, localisation plays a key role in correctly categorising labels of interest when performing this task. We propose a framework referred to as a 'likelihood map', based on the probability of correctly identifying the target label, which explores prediction by a dedicated network according to the position of the fixation point. Depending on the scenario, it can be guided (or not guided) by the target label in a manner similar to Grad-CAM or DFF. In both scenarios, we demonstrate improved classification performance when the sensor shifts towards the region of interest. Beyond its computational benefits, this framework can be used as an experimental tool to further investigate the neural mechanisms underlying visual processing.

Keywords: Neuromorphic; Convolutional Neuronal Networks; Deep learning; Localisation; Categorisation; ImageNet; Visual search; NeuroAI

Methods

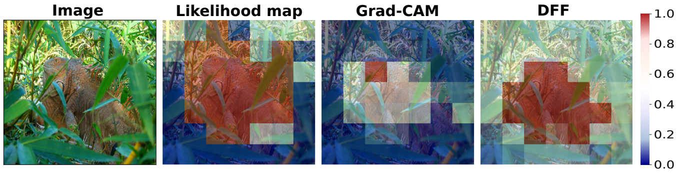

Figure 1: Display of different saliency maps produced by methods using the same CNNs. **From left to right** : Likelihood map, Grad-Cam maps, Deep Feature factorisation (DFF) produced when all methods successfully located the label of interest. Note that for the Likelihood map and Grad-Cam methods, the label is provided to the networks (label displayed : common iguana), whereas only DFF performs a categorisation along the localisation (label predicted : common iguana).

Grad-Cam and Deep Features Factorisation

Class Activation Mapping method (CAM) operates by analyzing the CNN's inner layers in relation to the targeted class, assigning weights to activations in each spatial feature map (Selvaraju et al., 2020). This process generates a heat map, highlighting significant areas of the image based on their contribution to the prediction (see Figure 1) and need to know the label of interest in order to generate the corresponding map, thus performing a localisation task *with cues*. Deep Features factorisation (DFF) is a method capable of localizing similar semantic concepts within an image (Collins et al., 2018). This method is significant because, unlike Grad-CAM, it obtains location information *without* requiring any *cues* of the label of interest, thus enabling simultaneous localisation *and* categorisation (see Figure 1)

From Label map to Likelihood and Information Gain maps

Here 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). The tensor can then be processed to keep the vectors corresponding to the label of interest to produce a $7 \times 7 \times 1$ tensor. For each position u, a hypothesis is formed (inferred) about the visual content of the view. This hypothesis takes the form of a probability distribution $p(k|x_u)$, which assigns a probability to each label $k \in 1,\dots,K$, such that $\sum_k p(k|x_u) = 1$. The logit output of the network f is classically interpreted as a log-probability, so that $f(x_u) \equiv \log p(\cdot|x_u)$. This probability can be interpreted as a confidence score for each label, indicating how certain the network is about its prediction. Two scenarios can arise:

- **If the label is known** (the "visual search" task), the target label k^* is known in advance, hence it is possible to extract, for each position u, the probability $p(k^*|x_u) \in [0,1]$, indicating, for each view, the likelihood that the network identifies the object at that position. The set of views thus produces a *likelihood map* $L^*(u)$, assigning a score to the target label at each spatial position (see Figure 1).
- **If the label is not known** in advance, two approaches can be considered:
  - The most direct extension of the previous method is to select the label with maximum logit score for each position u (see Equation 1).

$$L(u) = \arg\max_k logit(p(k|x_u)) \tag{1}$$

  - In the framework of *active inference*, it has been proposed (see Daucé (2018); Daucé & 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 x_0. This information gain is defined as the reduction in uncertainty regarding the interpretation of the scene, for each possible view. It has been shown that the information gain can be upper-bounded (up to a constant) which is much simpler to compute (see Equation 2).

$$IGUB(u) \equiv \sum_k p(k|x_0,x_u)\log p(k|x_u) \tag{2}$$

This value reflects an *optimistic bias* on the information gain, and is equivalent to combining a consistency term (the IG) and a discovery term (KL divergence of the posterior update), also known as the “Bayesian surprise” (Itti & Baldi, 2009).

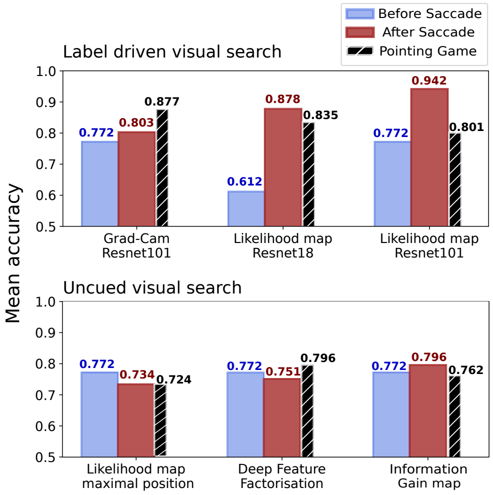

Figure 2: **Visual search** Average accuracy when processing all images from the ImageNet *validation* data set (only images with bounding box ground truth, approximately 98% of the $50,000$ images). **(Blue)**: Average accuracy for all images (centered viewpoint). **(Red)**: Average accuracy at the most salient viewpoint for the corresponding saliency map. **(Black)**: Average Pointing Game scores (a measure of correct localisation when the most salient point falls within the bounding boxes provided with the ImageNet data set) for the corresponding method. Top row represent the label-driven visual search methods while the bottom row represent the uncued visual search methods

Results

Label-driven visual search

First, we study the contribution of localisation when the label of interest is known (*with cue*). Contrary to expectations, when relying on the Grad-Cam method to determine the position of the next saccade accuracy increases only slightly (from 77.2% to 80.3%), despite very accurate localisation (with 87.75% of the most salient points falling within the bounding boxes). However, when selecting the most salient point from the likelihood map, the network's accuracy increases substantially (from 77.2% to 94.2%), even though the localisation appears less accurate compared to Grad-Cam. While this measure comes with a caveat, since our methods already involve the networks “exploring” all possibilities to make a decision, the result remains significant as it demonstrates that there exists an *optimal viewpoint* within the scene that can elevate the networks' accuracy to a competitive level (see Figure 2 top). We also evaluated a shallower network: ResNet-18, although it started with weaker categorisation performance (i.e., 61.3% accuracy), when the highest saliency point was used, the accuracy rose above expectations (i.e., 87.8% accuracy, see Figure 2 top).

Uncued visual search

Here we investigate whether saliency maps for categorisation can be obtained by a method that does not rely on the label of interest, *without cue*. Surprisingly, when using the DFF method, post-saccadic accuracy decreases compared to pre-saccadic accuracy (i.e., from 77.2% to 75.1%). When using a likelihood map method, if we simply use the most salient point among all fixation points and all labels, categorisation performance is also degraded (i.e., from 77.2% to 73.4%). However, if we rely on the IGUB to select the best position, post-saccadic accuracy increases significantly (from 77.2% to 79.6%). Although the Information Gain maps show better accuracy than the DFF, they remains still far from the optimal accuracy (i.e., 94.2%). In contrast, the pointing game score (i.e. localisation) remains lower in the IGUB than in the DFF, illustrating again a subtle trade-off between both objectives (see Figure 2 bottom).

Discussion

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 (i.e. 94.2% for the ResNet-101 network). This is even more notable given that shallower networks tend to be more competitive at this position (i.e. 87.8% for the ResNet-18 network). Various CAM-based methods can efficiently identify a label of interest. However, many of these locations do not achieve the maximum classification rate. The challenge lies in determining the optimal gaze location when the label of interest is unknown, with the aim of achieving the maximum classification rate. Once again, pure localisation methods, such as the DFF, struggle to maintain decent categorisation accuracy, although they do provide correct localisation scores. Here, we present an initial adaptation of the Information Gain framework for visual search on natural images. These methods are crucial for enhancing the classification rate and are the only ones that can improve accuracy by 2–3% in this study when the sensor is positioned correctly and there is no prior knowledge. The phenomenon of misalignment between classification and localisation observed in this study may be due to various factors, including the perception that an object of interest in a visual scene is in a 'winner takes all' competition with other objects present (Lee et al., 1999). This effect is even more pronounced in natural scenes, where the background can resemble the object of interest. Classical and contemporary findings — from goal-directed attention (Yarbus, 1961) to target-absent search behaviour (Yang et al., 2022) — highlight the importance of understanding how humans explore visual scenes. Further exploration of this 'optimal' fixation point could deepen our understanding of the clustering effect and clarify the features on which the CNN architecture relies for categorisation.

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”. 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. For the purpose of open access, authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Collins, E., Achanta, R., & Süsstrunk, S. (2018). *Deep feature factorization for concept discovery* (No. arXiv:1806.10206). arXiv. Retrieved 2025-02-18, from http://arxiv.org/abs/1806.10206 doi: 10.48550/arXiv.1806.10206

Daucé, E. (2018). Active fovea-based vision through computationally-effective model-based prediction. *Frontiers in neurorobotics*, *12*, 76. doi: 10.3389/fnbot.2018.00076

Daucé, E., & Perrinet, L. U. (2020). Visual search as active inference. In *Iwai 2020.* Retrieved from https://laurentperrinet.github.io/publication/dauce-20-iwai doi: 10.1007/978-3-030-64919-7_17

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

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision research*, *49*(10), 1295–1306.

Lee, D., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. , *2*(4), 375–381. Retrieved 2025-02-20, from https://www.nature.com/articles/nn0499_375 doi: 10.1038/7286

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. doi: 10.1007/s11263-015-0816-y

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. , *128*(2), 336–359. Retrieved from http://arxiv.org/abs/1610.02391 doi: 10.1007/s11263-019-01228-7

Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., & Samaras, D. (2022). *Target-absent human attention* (No. arXiv:2207.01166). arXiv. Retrieved 2025-04-08, from http://arxiv.org/abs/2207.01166 doi: 10.48550/arXiv.2207.01166

Yarbus, A. (1961). Eye movements during the examination of complicated objects. , *6(2)*, 52–56.