,

Are deep CNNs faithful models of the visual system?

An interdisciplinary research project in computational neuroscience spanning machine learning, retinal electrophysiology and human psychophysics. It fits within the SCAI program and its ambition of developing a broad range of interdisciplinary applications of machine learning. Additionally, by reinterpreting the original connection between human and computer vision, this project can open for further development of artificial intelligence.

U. Ferrari (CNRS-CR, Institut de la Vision, HDR) is an expert in computational neuroscience of the retina, with a strongly data-driven approach based on statistical inference.
P. Neri (CNRS-DR, École Normale Supérieure, HDR) is an expert in human psychophysics, with a focus on the visual system.
O. Marre (INSERM-DR, Institut de la Vision, HDR) is an expert in the biology of the visual system, with a focus on electrophysiological recordings from the retina.
The candidate: a motivated student with a strong background in STEM disciplines, outstanding coding skills, and a genuine interest for neuroscience and psychophysics.

Modern deep learning architectures share backbone characteristics with computational models of the brain. This is particularly evident in the case of models of the visual system and convolutional neural networks (CNNs), where the initial inspiration from brain science has led to the incorporation of successful engineering modules that are typically geared towards achieving better accuracy in multiclass classification problems (*1*). More recently, the opposite approach has also been developed in the hope of increasing our understanding of the visual system. In the first instance, a deep CNN with an architecture that mimics the visual system is constructed and trained on visual tasks, to then quantify how accurately (or poorly) its trained states align with recordings from neural structures (*2*) and subsequently score them in terms of their ability to predict neural activity (*3*). Although this approach has shed light on relevant properties of the visual system, several problems persist in filling the gap between how vertebrates and CNNs *see*:

Can we improve how deep CNN's are constructed and trained to better reproduce the behavior of the visual system? Recent results from our groups have highlighted several missing ingredients in previous approaches. First, contemporary CNN's fail to incorporate a deep characterization of the early stages of vision, most notably the retina, the first neuronal tissue responsible for vision. There is compelling evidence (*4*, *5*) that the retina is more sophisticated than a linear spatio-temporal filter, especially in response to complex natural stimuli. State-of-the-art deep models for characterizing the visual stream do not take into account the role of the retina (*6*), and consistently show limited predictions for layers of the primary visual cortex. Furthermore, it has been shown that training a neural network by optimizing only performance on a given task - e.g. object recognition - leads to models that do not reproduce how vertebrates see. We must recognize that the visual process operates at different degrees of *depth* and may express them differently depending on how it is interrogated (*7*). We therefore need to identify meaningful training protocols for artificial networks (*7*–*9*) that bring them closer to the complex behavior exhibited by biological processes

In order to improve over these limitations, this project proposes to

1. exploit data-driven machine learning techniques to introduce an artificial analogue of the retina, **as first layers of a deeper CNN**.
2. constrain outputs to a greater depth of characterization, in the form of **higher-order metrics**, beyond a simple task-performance evaluation.

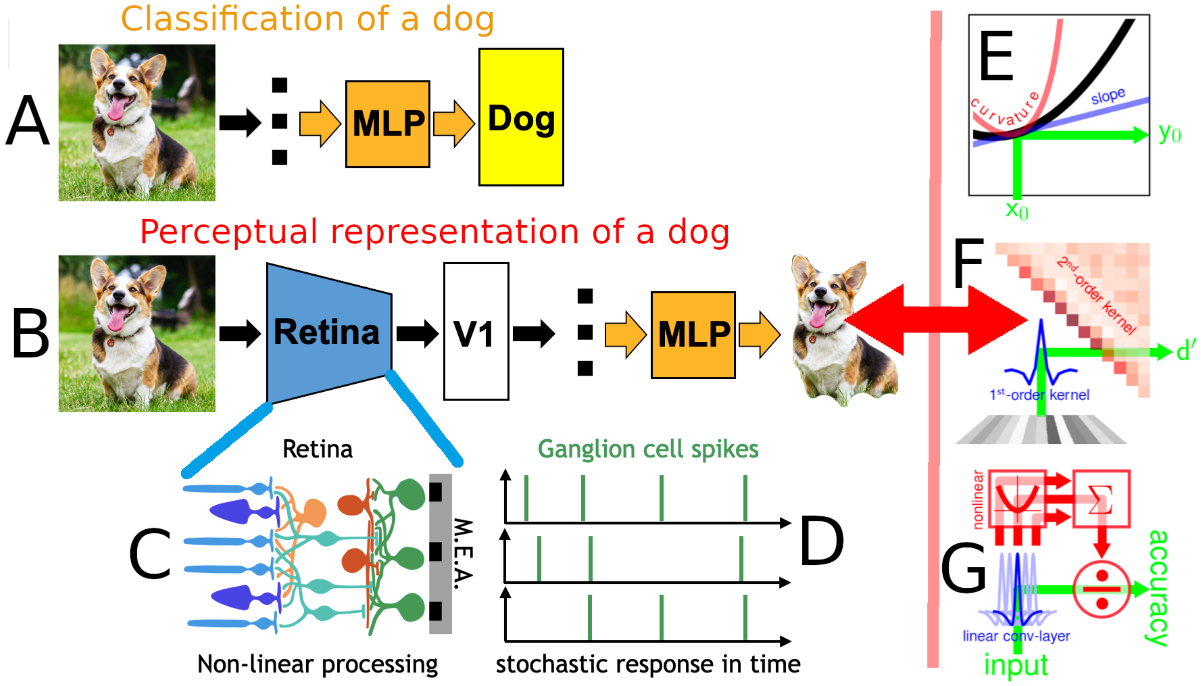

Figure 1: Conventional CNN's may successfully classify an object (**A**) but do not reproduce the full repertoire of human behavior. The behaviour of a human observer or a neural network can be characterized to different degrees of depth (intuitively similar to the Taylor expansion in **E**), from low-order (sensitivity/accuracy e.g. d', green in **F**) to higher-order (perceptual kernels of first-/second-order, blue/red). It is possible to establish a connection between first-order kernels (blue in **F**) and the linear portion of small network models such as convolutional layers (blue in **G**). The addition of biologically inspired modules such as the retina in the form of a bottleneck block of layers within deep CNNs (**B**) may re-establish the connection with complex behavior (red double arrow between **B** and **F**). The retinal network (**C**) shares canonical computations with models of human perceptual processing (**G**): light signals are absorbed by photoreceptors (dark/light blue neurons) and then processed throughout the retinal multi-layer architecture. Each neuronal layer consists of many different cell-types, for a total of at least 50 functionally-different convolutional kernels. The spikes of ganglion cells (green in **D**) represent the retinal output signal, and can be recorded with multi-electrode arrays. Because visual information must first go through the retina, the manner in which it is formatted by the retina must reverberate across the entire visual system all the way to behavior and associated metrics (**F**).

Research Plan

A primary focus of this proposal is to develop a satisfactory model of the retina, which is a physiologically dense network of neurons with horizontal and feedforward connections (Fig. **1C**): an efficient biological implementation of a recurrent deep CNN (*10*) that has benefited from millions of years of natural selection. In order to quantify and understand the role of horizontal/recurrent connections, we will pursue a data-driven approach by constraining machine learning models with data collected by O. Marre's team (*11*): these recordings characterize the response of retinal output neurons (Fig. **1D**) to a wide range of visual stimuli. Similar approaches on shallow CNN's have been tested by our group in (*5*, *11*, *12*). In this project we aim to extend this line of research to deeper and more realistic architectures.

For the purpose of testing whether our models can serve as adequate approximations for the full behavioral repertoire exhibited by biological processes, we will consider more articulate descriptors of the output following the approach developed by P. Neri's team(*7*, *8*). We can outline three different levels of characterization (see Fig. **1E-G**) for the output of a neural network: a *zeroth-order* level based on scalar metrics such as accuracy (green in Fig. **1E-G**); a *first-order* level associated with the concept of linear filter (blue in Fig. **1E-G**), as in the overall feature selectivity of a convolutional layer - already characterized in (*5*); a *second-order* level meant to capture outputs that are not adequately represented by lower order descriptors (red in Fig. **1E-G**). An intuitive way of understanding these three levels is to draw a parallel with Taylor expansion (Fig. **1E**).

This project is *feasible* because all necessary datasets are already available and ready to analyze. From the retinal perspective, the long-standing collaboration between U. Ferrari and O. Marre has made it possible to gather massive datasets and to develop preliminary models of the retina as CNN. From the visual system perspective and its impact on behavior, the large human psycophysical dataset collected by P. Neri's group will enable us to immediately test novel metrics and guide/improve the match between animal and CNN vision.

Perspectives

Bridging the gap between artificial models and biological evidence in vision can lead to several advantages in both domains: on the one hand, more explainable and robust - e.g. to adversarial attacks - deep learning models are inevitably essential for successful deployment of real-world AI applications; on the other hand, because small circuit models are insufficient to explain the structure of existing neural recordings, resorting to deep data-driven models could clarify the nature of low-level transformations in biological sensory processing. In order to achieve these goals, we need to operate a fundamental paradigm shift towards using higher-order descriptors of outputs and learning how to use brains to regularize machines. This is necessary to instantiate a productive feedback loop between the two research domains, and will be instrumental in establishing a general framework that may extend to computations in other sensory areas of the brain.

References

1. O. Russakovsky *et al.*, *International journal of computer vision* **115**, 211–252 (2015).
2. D. L. Yamins, J. J. DiCarlo, *Nature neuroscience* **19**, 356–365 (2016).
3. M. Schrimpf *et al.*, *Neuron* **108**, 413–423 (2020).
4. T. Gollisch, M. Meister, *Neuron* **65**, 150–164 (2010).
5. M. A. Goldin *et al.*, *bioRxiv* (2021).
6. J. Dapello *et al.*, *Advances in Neural Information Processing Systems* **33**, 13073 (2020).
7. P. Neri, *Neural Networks* (under review).
8. P. Neri, *Chaos* **20**, 045118 (Dec. 2010).
9. P. Neri, *Journal of Vision* **4**, 2–2 (2004).
10. H. Tanaka *et al.*, *Advances in neural information processing systems* **32** (2019).
11. S. Deny *et al.*, *Nature communications* **8**, 1–17 (2017).
12. G. Mahuas, G. Isacchini, O. Marre, U. Ferrari, T. Mora, *Advances in neural information processing systems* **33**, 5070–5080 (2020).