Retinotopic mapping improves the reliability of image classification

Jean-Nicolas Jérémie
jean-nicolas.jeremie@univ-amu.fr
Aix Marseille Univ, CNRS

Emmanuel Daucé
emmanuel.dauce@univ-amu.fr
Aix Marseille Univ, CNRS

Laurent U Perrinet
laurent.perrinet@univ-amu.fr
Aix Marseille Univ, CNRS

Abstract

Many animal species, such as humans, are characterized by a focused vision in which the sensor capturing the light information has a higher resolution around the orientation of gaze. Compared with a regular camera-like model of the eye, this arrangement of sensory inputs is still largely under-exploited in the field of computer vision. We propose to study the advantages of this transformation in the context of image classification. Inspired by this neuroscientific observation, we use a log-polar mapping which can be directly used to transform the input to classical deep learning classification algorithms using Convolutional Neural Networks (CNN). We apply this architecture to the recognition of the presence of an animal in the image and results show an improved accuracy for object recognition with a retinotopic transformation compared to a classical regular grid, but also a more robust object localization with respect to zooms or rotations. Moreover, we find that the retinotopic transformation improves the robustness of the localization of image classification when it is directed towards an isolated object. This opens perspectives for the use of the log-polar mapping in models of visual search, in particular by introducing biologically-inspired saccades in computer vision algorithms to efficiently localize and detect targets.

1. Introduction

A distinctive aspect of the vision of species like primates is that the eye can be directed and that visual information on the retina is concentrated at the center of gaze. It is still an open question as to understand the function of this retinotopic mapping, especially in comparison with other species, like rabbits, which lack this distinctive feature. In that direction, an important task in nature is the detection of objects of interest in a scene, such as an animal. Applied to generic natural scenes, the task is relatively difficult as the animal species is arbitrary and may for instance include birds, insects, or mammals. A further difficulty is due to the large variations in identity, shape, pose, size, and position of the animals that could be present in the scene. Yet, biological visual systems are able to efficiently perform such detection in images which are briefly flashed [13]. This task can be performed very rapidly [4] and robustly to geometrical transforms [9]. With respect to the generality and difficulty of this task, a scientific question is to understand how this is achieved. Here we propose that a retinotopic mapping may be an essential ingredient in that efficiency.

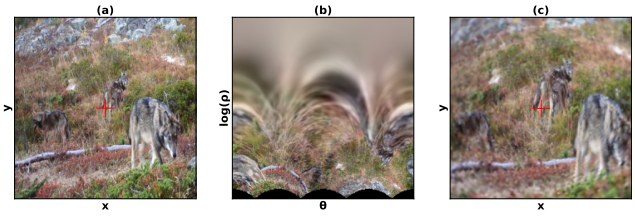


Figure 1. **Log-polar mapping.** *(a)* An example input image with the center of fixation denoted by a red cross. *(b)* Log-polar projection of the coordinates of each pixel of the input image according to: on the x-axis its angle of DEVIATION = θ from the horizontal axis and on the y-axis the logarithm of its eccentricity (or radius) ρ with respect to the fixation point. *(c)* Reconstruction from the log-polar mapping in (b) of the input (a), illustrating that fine details are kept around the point of fixation only.

Serre et al [11] has previously investigated an artificial visual system applied to such a task which compares in efficiency to biological ones. We recently extended such type of architecture on a larger, more generic dataset [3]. In that study, we retrained an artificial neural network to compare its performances with the physiological data on the categorization of an animal in a natural scene. In order to achieve this task, we constructed a dataset built on the IMAGENET database [10]. For this, we defined our dataset's labels based on a large semantic database of English words: WORDNET [2] and used a Transfer Learning method to re-

train the networks [6], a method using an existing network pre-trained on a specific task (here, the VGG16 architecture [12] trained on IMAGENET) and modifies this network by re-training a subset of its weights on a different task. Specifically, we re-trained the CNN, for the categorization task (for instance, "is there an animal in the scene?") as defined by the dataset of supervision pairs.

This method [3] generalized previous results but was applied on regular images constituted by a set of pixels arranged on a regular grid. Here, we will 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. This mapping will be such that each input matrix defined into the two spatial dimensions is transformed into a new matrix with one axis corresponding to the azimuth and the other to the radius with respect to the point of fixation (see Figure 1). As such, this new matrix can be use as the input to a classical CNN. This will allow us to re-train the network using this mapping and quantify the classification abilities of this solution. Moreover, we will expose the contribution of this mapping to the localization of objects by showing how the predicted likelihood of detecting an object may change as a function of the location of gaze.

Overall, this will provide a tool to qualitatively show how such a retinotopic mapping may be necessary to perform this task. Finally, we will discuss how this work can be beneficial for the conception of new architectures for computer vision biologically inspired by the organization of the visual system.

2. Methods

2.1. 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, it allows us to gain computing time during the training process and to test multiple possible configurations. Here, we use the same protocol exposed in [3] to re-trained the VGGGEN network. The dataset contains a 'train', 'validation' and 'test' folder (2000, 1200 and 1200 images, respectively). Each folder contains a 'target' and a 'distractor' category (50% and 50% of the dataset, respectively). All networks are trained on the 'train' folder and tested during their learning on the corresponding 'validation' folder. Then, we compute the performances using the 'test' folder. Thereafter, we extend the protocol by including a log-polar mapping (see below) to re-train the network (VGGPOLAR). Here we will focus on the F1-score of the networks which is defined as the harmonic mean of the model's precision and recall and thus conveniently combines these measures thus giving a good indication of the ability of the network to detect the presence of a label of interest (here an animal) in an image.

2.2. Retinotopic transformation

We model the retinal transformation by a log-polar projection of the linear space to a space represented on one axis by the azimuth and on the other by the logarithm of the eccentricity. To do this, we start by defining the log polar space. The projection in discrete coordinates can be represented by a disk. Two characteristics are therefore important to take into account: the radius on which we will sample our information and the size of our output matrix. They define a compression index of the information. We can then reshape the pixels composing the image according to these new coordinates. In the case where we apply the log-polar transformation, the radius of the input disk is 240 pixels for an output matrix of 64x64, i.e. a compression ratio of 3.75 between the center and the periphery. This implies a loss of information with the eccentricity with respect to the point of fixation, hence the importance of the choice of the latter (represented by a red cross (see Figure 1)). The Imagenet database appears to have a bias in the positioning of its labels of interest, as these are most often centered in the image. Thus, after exploring training strategies on dataset where we had redefined the center of the image to match the boxes surrounding the objects of interest, or using only these boxes to compose our dataset without observing any real difference between these conditions, we opted for a fixation point at the center of the native image.

3. Results

3.1. Performances on natural scenes

We first tested the network re-traind on our IMAGENET dataset without any transformation. It seems efficient in the categorization on this dataset as it reach about 91% mean accuracy (see Figure 2 *(a)*). Then, we tested the log-polar version of this network, that is re-trained on images in log-polar space. It also reaches accuracies comparable to the network tested on raw images (see Figure 2 *(a)*). Moreover, it seems to exhibit a robustness to rotation and zoom transformation unlike the latter (see Figure 2 *(b)* & *(c)*). These observations are coherent with the characteristics of a CNN like VGG16 as these transformations in the linear space correspond to translation in the log-polar space, a transformation for which this architecture is by definition robust. When we train, then test this network on the reconstruction of the linear image (see Figure 1), we observe a drop in the accuracy of the network, even if we use a version of the dataset IMAGENET with images centered on the label of interest. This drop is probably due to the great variability of the label of interest (here animal) coupled with a low knowledge of the features necessary for this categorization, which makes it difficult to define the point of fixation which

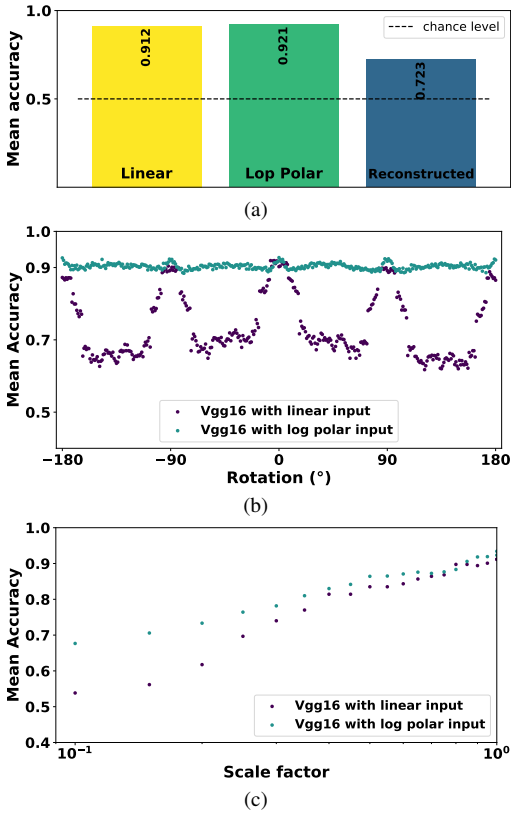
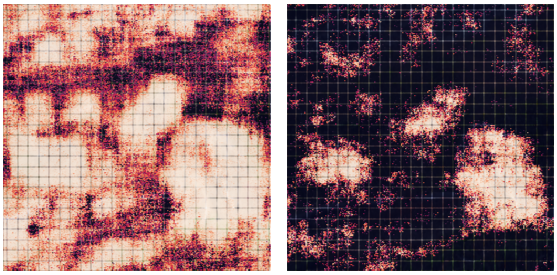


Figure 2. **Comparisons for animal categorization with different mappings to the input.** *(a)* Mean accuracy of the CNN VGG16 re-trained with linear images (yellow), with a log-polar transformation to the input (green), or with a reconstructed images (blue) as shown in Figure 1. *(b)* Mean accuracy over the test dataset of the re-trained networks for different rotations of the input image. The rotation is applied around the fixation point with an angle ranging from -180° to +180° (step=1°) with raw images (purple) or with a log-polar mapping to the input (green), showing the invariance to rotation characteristic to log-polar mappings. *(c)* Mean accuracy over a centered test dataset of the re-trained networks for different scaling of the input image. The zoom is applied around the fixation point with a scale factor ranging from 0.1 to 1 (step=0.1) with raw images (purple), with a log-polar transformation to the input (green).

could maximize the contribution of this transformation.

3.2. Implementing the saliency map protocols

A saliency map corresponds, in our case, to a map indicating the positions of the images for which the prediction of the categorization of the presence of an animal by our network is above $p = 0.5$. To obtain these maps, we performed a subsampling of a 240×240 window on a 2400×2400 resolution image with a step size of 9, for a total of 58081 subsamples. Each of these samples is then sent to the input of the network, providing a matrix of 241×241 predictions for each of these positions. We then superposed this saliency maps on the raw image dataset with or without



(a) (b)

Figure 3. **Retinotopic saliency map compared with linear saliency map** (a) Heat map computed with the CNN VGG16 re-trained with linear images on raw inputs. (b)Heat map computed with the CNN VGG16 re-trained with linear images on reconstructed inputs.

applying a retinotopic transformation (see Figure 1) during the process. Foremost, in both case it allows us to extract the regions of interest including the key features necessary for the categorization of an animal by our network, with a finer contour around the area of interest for the heat map generated with a reconstructed input and this even if the network used is trained to recognize an animal in a linear space (see Figure 3). Thus this example is qualitative, but we observed similar results on several images for several species of animals in different contexts (not shown here).

4. Discussion

In summary, we have shown that a retinotopic mapping can be applied to a classical CNN and that when we re-trained networks using transfer learning, it achieved human-level accuracies to on an ecological task. Furthermore, the robustness of the categorization is comparable to those found in psycho-physical data as the categorization of the dedicated networks are robust to transformation like rotation, reflection or grayscale filtering [9].

A surprising fact is the conservation of categorization despite training with a log-polair transformation resulting in the degradation of textures outside the area of interest. These results are promising because in addition to being consistent with physiological data, they allow us to pursue a research direction where we could implement training of a retinotopic map with information compression in the periphery. The use of saliency maps allowed us to highlight the fact that the categorization of this kind of network can be modulated by the application of a retinotopic log-polar transformation, which seem to allow a more accurate localization of the object of interest, here an animal. Furthermore, it gives us an insight on what features our networks actually rely on, in order to categorize the presence of an animal. Of course, a study on a larger dataset is necessary to validate these results.

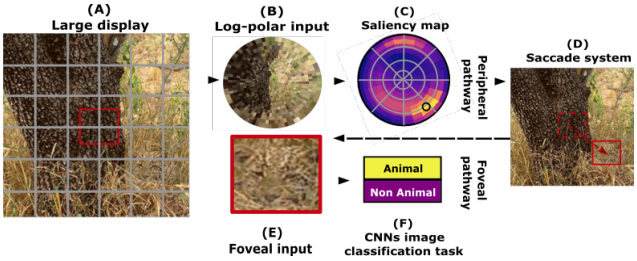


Figure 4. Model build over the anatomical visual processing pathways observed in mammals, namely the "What" and the "Where" pathways : Peripheral pathway (top row) is applied to a large display from a natural scene *(A)*: It is first transformed into a retinotopic log-polar input *(B)* and we then learn to return a "saliency map" *(C)*. The latter infers, for different positions in the target, the predicted accuracy value that can be reached by the foveal pathway, mimicking the "Where" pathway used for global localization. The position with the best accuracy will feed a saccade system *(D)*, adjusting the fixation point at the input of the foveal pathway (bottom row). It takes a subsample *(E)* of the large display *(A)*, over which a categorization is done *(F)*, mimicking the "What" pathway.

5. Perspectives

One of the main goals of this study was to provide a comparison on an ecological and well studied task used in primate visual neuroscience. Although this study focuses on the analysis of categorization, it is a necessary step for a well-known task in the field of vision: visual search. This task consists of the simultaneous localization and detection of a visual target of interest. Applied to the case of natural scenes, searching for example for an animal (either a prey, a predator or a partner) constitutes a challenging problem due to large variability over numerous visual dimensions. Previous models managed to solve the visual search task by dividing the image into sub-areas. This is at the cost, however, of computer-intensive parallel processing on relatively low-resolution image samples [5] [8]. Taking inspiration from natural vision systems [7], we develop here a model that builds over the anatomical visual processing pathways observed in mammals, namely the "What" and the "Where" pathways [1]. It operates in two steps, one by selecting a region of interest, before knowing their actual visual content, through an ultra-fast/low resolution analysis of the full visual field, and the second providing a detailed categorization over the detailed "foveal" selected region attained with a saccade (see Figure 4). Modeling this dual-pathways architecture allows offering an efficient model of visual search as active vision. In particular, it allows filling the gap with the shortcomings of CNNs with respect to physiological performances. In the future, we expect to apply this model to better understand visual pathologies in which there would exist a deficiency of one of the two pathways [14] while contributing to the field of computer vision.

Acknowledgments

Authors received funding from ANR project N° ANR-20-CE23-0021 ("AgileNeuroBot"). Authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version.

References

[1] E. Daucé & L. Perrinet. Visual search as active inference. In Tim Verbelen, Pablo Lanillos, Christopher L. Buckley, & Cedric De Boom, editors, *Active Inference*, volume 1326, pages 165–178. Springer International Publishing, Cham, 2020. 4

[2] C. Fellbaum. *WordNet: An Electronic Lexical Database*. Language, Speech, & Communication. A Bradford Book, 1998. 1

[3] J-N. Jérémie & L. Perrinet. Ultra-fast image categorization in vivo & in silico. (arXiv:2205.03635). 1, 2

[4] H. Kirchner & S. Thorpe. Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, 46(11):1762–1776, may 2006. 1

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, Cheng-Y. Fu, & Alex&er C. Berg. SSD: Single Shot MultiBox Detector. *arXiv:1512.02325 [cs]*, 9905:21–37, 2016. 4

[6] A. Mari, T. R. Bromley, J. Izaac, M. Schuld, & N. Killoran. Transfer learning in hybrid classical-quantum neural networks. *Quantum*, 4:340, Oct. 2020. 2

[7] M. Mishkin & L. Ungerleider. Object vision & spatial vision: Two cortical pathways. *Trends in Neurosciences*, 6:414–7, 1983. 4

[8] S. Ren, K. He, R. Girshick, & J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv:1506.01497 [cs]*, Jan. 2016. 4

[9] G. A. Rousselet, M. J.-M. Macé, & M. Fabre-Thorpe. Is it an animal? Is it a human face? Fast processing in upright & inverted natural scenes. *Journal of Vision*, 3(6):440–455, 2003. 1, 3

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, & Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

[11] T. Serre, A. Oliva, & T. Poggio. A feedforward architecture accounts for rapid categorization. *PNAS*, 104(15):6424–6429, 2007. 1

[12] K. Simonyan & A. Zisserman. Very Deep Convolutional Networks for Large-Scale. *arXiv:1409.1556 [cs]*, Apr. 2015. 2

[13] S. Thorpe, D. Fize, & C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996. 1

[14] E. Wiecek, L. Pasquale, J. Fiser, S. Dakin, & P. Bex. Effects of Peripheral Visual Field Loss on Eye Movements During Visual Search. *Frontiers in Psychology*, 3, 2012. 4