Response to reviewers

We are grateful to the reviewers for their encouraging remarks and useful comments which were essential in improving our manuscript. In addition to the revised version of the manuscript, we detail here our response to the different points being made while showing some screenshots of the revised manuscript when useful. A "revised tracked changes" is also available as a supplementary file to see for all changes being made since the original revision.

As a reference, the original revision is available at https://www.biorxiv.org/content/10.1101/725879v2.full.pdf

Reviewer #1:

This paper proposes a foveated visual search model. The model implements the What vs. Where separation in a focal accuracy seeking policy (i.e. accuracy driven action selection). The model is experimentally evaluated on a search task for handwritten digits on cluttered backgrounds. Model performance is evaluated and analyzed depending on SNR, eccentricity of the target and the number of saccades performed. Performance as a function hyper parameters is also analysed. The authors made their source code for re-producing the results publicly available. Overall, I think this would be a good contribution to the JOV.

I recommend the acceptance of the paper after the following minor issues are addressed:

- The discussion on the time efficiency of the proposed model over the exhaustive scan (i.e. classical computer vision) approach can be made more convincing. First of all, I think it deserves to be discussed in the main text instead of a footnote (as in page 10). Here is my thinking: suppose that the cost of foveal processing is C. Then, the cost of the exhaustive scan model would be n times C, where n is the number of all pixels (or rather all the locations where fovea will be evaluated). On the other hand, the cost of the proposed model is f times C + f times P, where f is the number of fixations and P is the cost of the log-polar processing model (the Where pathway). The relation between P and n is obviously P = k logn (where k is some constant). Assuming that f is typically much smaller than n, the proposed model seems to be more time efficient than exhaustive search.

We thank the reviewer to propose this very useful addition to the discussion and allowing us to improve the manuscript on this point. O(n) is a lower bound for the processing of n pixels (read at least every pixel once). A full convolutional scan of an image with a window of size C and a stride of 1 is effectively O(nC) but this is rather like an upper bound. In practice, the processing is expected to stand between O(n) and O(nC). With a tight optimization, modern image processing methods are really close to O(n). In our case, a log-polar encoded image contains O(log n) coefficients. One dual processing step is O(C) + O(log n). If k is a bounded number of steps, and C considered as a constant, the total processing cost is O(k C + k log n)) = O(log n) (constant values do not change the order). We have added a specific section "3.2 Analysis" in page 10 of the revised text:

3.2 Analysis

This full covering of the 128×128 image range is done at a much lesser cost than what would be done by a systematic image scan, as in classic computer vision. Taking n the number of pixels in the original image (in our case $n = 128 \times 128 = 16384$), our log-polar encoding provides $O(\log n)$ log-polar visual features by construction. The total visual data processed is the addition of the C pixels processed at the fovea and the $O(\log n)$ log-polar visual features processed at the periphery. The total processing cost is thus $O(C + \log n)$. Taking C as a constant, the total processing cost can be said $O(\log n)$ (for constant processing times do not change the order). In the case of multiple saccades (see next section), the total cost is $O(k \times (C + \log n))$ with k the number of saccades. If the number of saccades k is bounded by a constant K, this allows to estimate the processing cost as $O(K \times (C + \log n))$ in the worst case, that also resumes to $O(\log n)$. This is to be contrasted, for instance, with the linear cost obtained with a full convolutional scan with a window of size C and a stride of 1, that is precisely $O(C \times n)$. Various optimizations can of course be considered, of which the well-known max-pooling principle used in deep learning, but anyway image processing (without compression loss) is generally considered as linear in the size of the visual data

processed [SKE06]

Our sub-linear processing time thus justifies a strategy that may have been chosen in a variety of natural vision systems. The compromise between the urgency to detect and the need to be accurate may justify the different balances which may exist in different species. In particular, this may justify the differences observed between preys (with a less sparse cone density at the periphery) and predators (with a tendency toward denser foveal regions).

Thanks to this point, we removed some redundancies in our manuscript and have made some effort to shorten, and simplify the prose at several places. The most important changes are listed in this response and all changes are highlighted in the tracked changes PDF.

Minor and more specific comments

- What does "Full-scale" mean in L502? Do you mean high-and-uniform resolution? Please clarify.

Indeed, this formulation was not clear. We have redefined it more clearly (page 15) as the images which are typically used in computer vision:

Full-scale images. We call "full-scale images", input images which correspond to a discretized, rectangular sampling (pixels) of the visual field, such that which are usually used in computer vision. These input images are set to a size of 128 × 128 pixels in which we embed the target. Each target location is drawn at random in this large image. To enforce isotropic generation (at any direction from the fixation point), a centered circular mask covering the image (of radius 64 pixels) is defined/ Also, the target's location is such that the embedded sample fits entirely into that circular mask.

- In Fig 4, the fonts and graphics are blurry. Perhaps, it is the result of using a raster format instead of vector graphics. Same with Figs 5 and 6.

Done, thanks. This was indeed caused by the low resolution of the images. We have uploaded high-definition and vectorized versions of the figures.

- Please consider providing a label for the y-axis in Fig 4.

Done, thanks.

- In the text, figures are cited as "Figure", "figure" or "fig". Please be consistent and follow the journal's style.

Done, thanks. In addition, we have improved the overall readability and had the manuscript intensively proof-red. Please see the tracked changes' PDF that highlights all the changes we have done on the manuscript.

Reviewer #2:

This paper presents an interesting study at the intersection of neural networks and human vision, specifically involving visual search.

This work is hard to categorize as it sits somewhere between the two aforementioned fields, making direct comparisons to either studies in visual search involving CNNs, or studies involving human participants a challenge. To address this, a suitable paradigm is introduced which involves presentation of the classic MNIST digits at different degrees of contrast over noise. This allows assessment of the degree to which a model that foveates regions of the image (modeled by a log-polar transform) can effectively localize and identify targets of interest.

For me the key finding of this work is that a sub-linear optimized spatial search is useful and effective in localizing and identifying targets of the type chosen.

We thank reviewer #2 for his encouraging and valuable comments. In our revision, we have tried to put forward this strong point and render them more visible to the readers. This point was also the object of an observation of Reviewer 1.

I do have some suggestions and questions relating to the manuscript as follows:

i. The overall presentation could be tightened up a little in terms of grammar and sentence structure

In this revision, we have improved the overall readability and had the manuscript intensively proof-red. Please see the tracked changes' PDF that highlights all the changes we have done on the manuscript.

ii. Figure 4 references orange bars, but they appear to be brown to me.

We refer this as transparent light orange now and reworded the caption to make this point clearer.

iii. The notion of optimal strategies for exploration could be expanded upon. E.g. The discussion of Najemnik and Geisler's work is a good fit, one could also include Bayesian Surprise (The IK reference is inappropriate - this was Baldi and Itti), or other information seeking strategies - the AIM or SUN models)

We are grateful to the reviewer to have raised our attention to that elements in the literature of visual search. We have now more precisely described the notion of optimal strategies and included these further references.

iv. Central to the model is the decision between foveation and identification. The paper states: "If the predicted accuracy in the output of the "Where" network is higher than that predicted in the "What" network, the position of maximal activity in the "Where" pathway serves to generate a saccade which shifts the center of gaze. Else, we interrupt the visual search and classify the foveal image using the "What" pathway such as to give the answer (ANS)." It is not clear that these quantities are on the same scale or comparable. More detail on this particular mechanism would be welcomed.

Indeed, this is a crucial point. When we compare the output accuracies of the “What” and the “Where” models, we get scalars which have the same unit and scale as they represent an accuracy. But both models output predict these values thanks to the supervised learning scheme. The “What” pathway uses the CrossEntropy Loss as it is trained on the classification of the digits from the MNIST dataset. The output of the “What” pathway can then be interpreted as a probability, that gives the chance of correct classification for each possible choice. Similarly, the “Where” pathway is trained on an accuracy map that predicts the chance of correct classification for each counterfactual saccade. This was precised in the caption to Figure 2.

v. I wasn't able to discern whether the network is trained piecewise (e.g. with the saccade decision part done manually in code), or whether the entire network is trained with the BCE and argmax end-to-end. If slightly more detail could be provided on the nature of the training procedure and how one constructs the model in such a way that it performs both selection and identification while training, this would also be welcomed.

You are right, the learning is sequential, first the “What”, then the “Where”. We have detailed this in the text and in particular page 8:

> In practice, the “What” and ‘Where” networks are both implemented in `pytorch` [PGM+19], and trained with gradient descent over multiple layers. Each network is trained and tested separately. Because the training of the “Where” pathway depends on the accuracy given by the “What” pathway (and not the reverse), we trained the latter first, though a joint learning also yielded similar results. Finally, these are evaluated in a coupled, dynamic vision setup.

Moreover, we have tried to improve this description by doing a number of changes to the text. Please see the tracked changes’ PDF that highlights all the changes we have done on the manuscript.