Explainable deep learning reveals a behavioral strategy underlying human decisions in a spatial navigation task

Youri Marquise[1], Bilel Abderrahmane Benziane[2], Hachim Bani[1], David Romano-Lambert[1], Youssouf Ismail Cherifi[3] and Denis Sheynikhovich[1]

1- Sorbonne University, INSERM, CNRS, Institut de la Vision, F-75012, Paris, France
2- L@bISEN, Usine du Futur, ISEN Yncrea Ouest and DataLab, Generix Group, France
3- Mines Paris, PSL Research University, Centre for Robotics, CAOR, France

Abstract. This paper uses a set of explainable AI (xAI) methods to study human behavior in a spatial navigation task. First, locomotion and gaze dynamics of human subjects recorded during the task were reproduced in a virtual environment and visual snapshots extracted from this simulation were used as a dataset. Second, the dataset was used to train a deep convolutional network to reproduce human navigation decisions. Third, network strategies used for image classification were analyzed using a combination of three xAI methods. Using this analysis, we discovered a specific oculomotor marker that indicated the behavioral strategy used by human participants in this task. We conclude that xAI is a promising approach to study human behavior in complex real-world tasks.

1 Introduction

Deep Artificial Neural Networks (dANNs) are tools of growing importance in cognitive science research [1]. In current applications, dANNs are often used as neural models of the brain or as data mining applications aiding complex data analyses [2, 3, 4, 5]. This work uses a different approach, whereby we *(i)* train a dANN to reproduce human behavior in a vision-based behavioral task, in which the reason for different behavioral patterns is unclear; *(ii)* we apply a set of explainable AI methods to analyze strategies that the network used to solve the task, allowing us to get insight into subtle behavioral differences underlying human decisions.

To show the potential of this approach we took as an example a recent spatial navigation experiment conducted in our lab [6]. In this experiment, human participants were asked to remember the location of a hidden goal in a large rectangular room with landmarks on the walls (Fig.1 a, b). After several learning trials, the landmark array was rotated (without the participant's knowledge) so as to create a conflict between the landmarks on the walls and geometric room cues. The participants were asked to navigate to the same goal location as before. The results show that about half of the participants followed the rotated landmarks, whereas the other half followed the unchanged geometric cues (Fig.1 c, d). The reason for distinct behavioral patterns in this experiment is unclear, but it is likely that some subtle behavioral differences in visual exploration strategies can provide insights into human behavior. To explore this possibility, we trained AlexNet dANN to classify images extracted from a 3D reconstruction of the human experiment into "landmark" and "geometry" classes. We then applied receptive field analysis [7], layer-wise relevance propagation (LRP) [8] and subsequent heat map analysis with spectral clustering (SpRAy)[9] methods to study our dANN classification strategies.

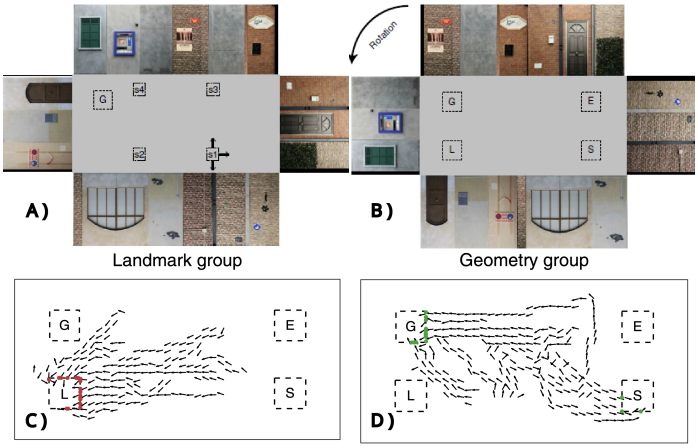


Fig. 1: Behavioral experiment. *a:* Training environment. Disoriented participants placed in a start location (s1 - s4) were asked to locate the unmarked goal G. *b:* Probe trial, after cue rotation. Participants who went to the L corner followed rotated landmarks. Those who went to the G or S corners followed room geometry. *c, d:* Behavior of the participants belonging to the landmark or geometry group.

2 Methods

2.1 Training image set

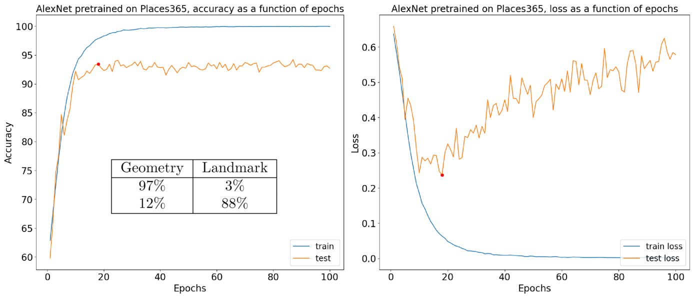
The behavioral experiment was conducted in a well controlled environment equipped with motion capture (MoCap) and eye tracking data [6], allowing us to reproduce locomotion and gaze dynamics of 40 human subjects in a virtual copy of the real environment, created in Unity3D. $N = 65000$ 2D images (visual snapshots) were extracted from Unity (frequency 60 Hz) during the period of visual exploration, between the trial start and the first navigation step taken, for all trials and all subjects. The images were preprocessed by *(i)* resizing the images to 256x256 pixels and *(ii)* normalising the pixels value. Since all people navigated in the same simple environment, many images were extremely similar for both classes. This resulted in many false positives after training. To solve this issue, we removed all nearly identical images from the dataset (using euclidean distance metric). The image set was then randomly separated into training, test and validation sets.

2.2 Neural network training and analysis

The convolutional dANN AlexNet was first pre-trained on the Places365 image set [10] to classify scene images into 400+ scene categories. AlexNet was chosen for its simplicity and due to the fact that most of the xAI methods used for the analysis were developed and tested using this network architecture. Moreover, there was no reason to look for more complicated architecture, as the network already displayed good performance (Fig.2). Next, we set the weights of the last convolutional layer to 0 and retrained the network to classify images in our dataset into "geometry" and "landmark" classes, as is often done in transfer learning problems [11].

To analyze the network strategies, we first used the method by Zhou et al. [7], to obtain "receptive fields" of the neurons in the last convolutional layer. Briefly, top-10 images preferred by each neuron in this layer were first extracted. These images were then locally modified in order to see which image pixels induced the largest change in the neuron's activity.

Second, the LRP method [8] was used to generate, for each image in the dataset, a heat map of pixel importance for its classification into one or the other class. In contrast to neurons receptive fields, which evaluate the importance of different image pixels for activating *a particular neuron*, heat maps evaluate the image pixels that are important for *the whole network* for correct classification.



Geometry	Landmark
97%	3%
12%	88%

Fig. 2: Accuracy (left) and loss function (right) of the network during learning with the training data in blue and testing data in orange. *Table:* Confusion Matrix of the selected epoch. *Red dot:* selected epoch.

Third, we used the SpRAy method [9] to understand which strategies the network used for classification. More specifically, a distance matrix was created from the ensemble of heatmaps. Next, spectral clustering was used on the distance matrix to separate heat maps into clusters. Each cluster corresponds to a particular image classification strategy learned by the network.

3 Results

Perhaps surprisingly, the dANN was able to classify images with a very good accuracy of 92.6% (Fig. 2). The biological interpretation of this result is that

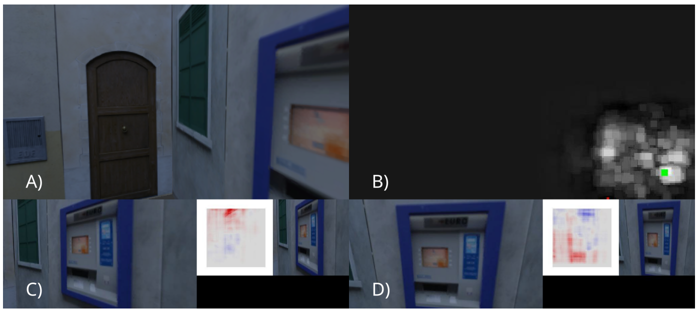


Fig. 3: Receptive fields and LRP heatmaps. *a:* An image from the training dataset. *b:* Receptive field of the 252th neuron. *c-d:* Examples of LRP heatmaps from the geometry (c) and landmark (d) classes. In each plot: left - the original image; right - resized image; middle- the LRP heatmap with red pixels important for the classification into the geometry class and blue pixels important for the classification into the landmark class.

only one image from a person visually exploring the environment (even before a person makes a first goal-directed movement) is sufficient for the network to tell whether a person will follow landmark cues or room geometry when looking for a hidden goal. The rest of this section describes the approach we took in order to understand image-processing strategies adopted by the network, hoping that it will provide insights into human behavior.

We obtained the receptive fields of 256 neurons in the last convolutional layer (see Fig. 3a,b for an example). This information was used to separate images into clusters according to different strategies as follows: To obtain strategies used by the network, we used LRP to generate heat maps of pixel importance for all images in the dataset (see Fig. 3c,d for examples). These heatmaps were then separated into different image clusters, corresponding to different network strategies, using spectral clustering [9]. This method has determined 555 different strategies for the geometry class and 291 strategies for the landmark class. Most image clusters contained different variations of the same environmental feature, such as ATM. The strategies used by the network corresponded to similar environmental features in both the landmark and geometry classes. For example, the ATM was the main environmental feature present in one cluster of the landmark class and one cluster on the geometry class (Fig. 3c,d). How does the network decide to put these very similar images into two different classes? We noticed that representative images of many clusters in this category had only one main difference between the two classes: the main feature was viewed from an oblique angle in the geometry class and straight ahead in the landmark class. We tested this hypothesis by computing the angle between the gaze and the wall along the Y axis in the virtual environment, on all images in the identified clusters (Fig. 4a). The distribution of gaze vector angles in Fig. 4a shows that the participants in the geometry group indeed looked at image features with more oblique angles. One reason for this behavior could be that looking along the wall, rather than straight to it, could provide more information about the wall length – a geometric cue.

We then checked to what extent this behavior is true in the whole set of recorded images, not just in a particular cluster. While this did not appear to be the case in the full set of images (Fig. 4d), it was in the images from the first training trial (Fig. 4b for vertical direction and Fig. 4c for the horizontal direction). The first trial is different from all other trials, because the subject sees the environment for the first time. This result indicated that human participants from the geometry group express a particular type of behavioral strategy, absent from the participants in the landmark group.

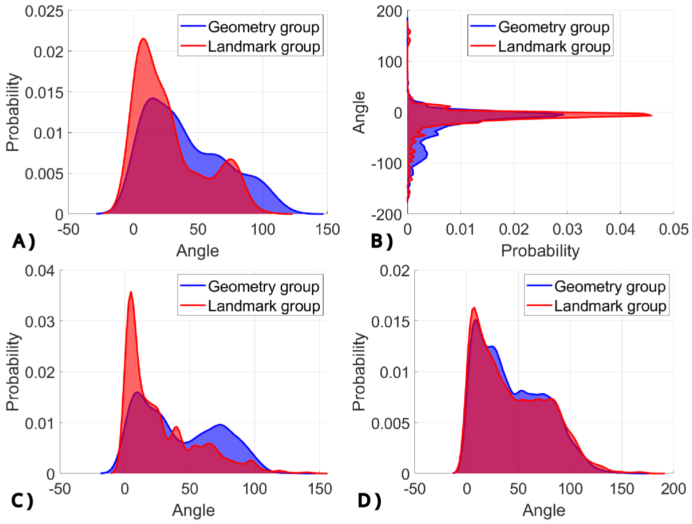


Fig. 4: *a:* Distribution of angle between gaze vector and wall normal around Z-axis for chosen clusters. *b:* Distribution of angle between gaze vector and wall normal around x-axis for first trial of all participants. *c:* Distribution of angle between gaze vector and wall normal around Z-axis for first trial of all participants. *d:* Distribution of angle between gaze vector and wall normal around Z-axis for all data.

4 Conclusion

In this article, we demonstrated the use of a new AI-based approach to study human behavior. Instead of using a network directly as a model of human brain or as an analysis tool for data mining, we used it to discover behavioral patterns in a complex behavioral task. By asking a neural network to solve a problem similar to what human participants were asked to do in the experiment, and then using xAI methods for analysis, we obtained new insights on human behavior that participants *were not able to verbalize.* This method allowed us to shed a new light on experimental data and characterize subtle behavioral patterns specific to a particular subject group. In particular, our results extend the previous findings by Becu et al. 2020 by suggesting that geometry-related preference in navigating humans is expressed not only by looking at the floor (proposed as a general strategy in their work), but also rely on looking at a wider range of angles oblique to the walls. It helps to resolve a puzzling observation that very often geometry-preferring subjects looked at landmarks, an unnecessary behavior within a suggested floor-based strategy. A subtle difference in the looking angle, discovered by our method, proposed a novel explanation for this observation and suggested a simple oculomotor marker of geometry-based strategies which are principally used by aged navigators (Becu et al 2020, 2022).

We believe that this method is not limited only to behavioral data, as the same approach we used to analyze oculomotor and behavioral data can in principle be used to study behavioral correlates of neural activities, recorded by fMRI or EEG.

References

[1] R. M. Cichy and D. Kaiser. Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, 23(4):305–317, April 2019. Publisher: Elsevier.

[2] D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci*, 19(3):356–365, March 2016. Publisher: Nature Publishing Group.

[3] N. Kriegeskorte and P. K Douglas. Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, 55:167–179, April 2019.

[4] U. Guclu and M. A. J. van Gerven. Modeling the Dynamics of Human Brain Activity with Recurrent Neural Networks. *Front. Comput. Neurosci.*, 11, February 2017. Publisher: Frontiers.

[5] A. Abrol, Z. Fu, M. Salman, R. Silva, Y. Du, S. Plis, and V. Calhoun. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat Commun*, 12(1):353, January 2021.

[6] M. Becu, D. Sheynikhovich, G. Tatur, C. Persephone A., L. L. Bologna, J.-A. Sahel, and A. Arleo. Age-related preference for geometric spatial cues during real-world navigation. *Nat Hum Behav*, 4(1):88–99, January 2020. Publisher: Nature Publishing Group.

[7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object Detectors Emerge in Deep Scene CNNs, April 2015. arXiv:1412.6856 [cs].

[8] G. Montavon, A. Binder, A. Lapuschkin, W. Samek, and K. Muller. Layer-Wise Relevance Propagation: An Overview. In W. Samek, G. Montavon, Andrea Vedaldi, Lars Kai Hansen, and K. Muller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 193–209. Springer International Publishing, Cham, 2019.

[9] A. Lapuschkin, S. Waldchen, A. Binder, G. Montavon, W. Samek, and K. Muller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun*, 10(1):1096, March 2019. Publisher: Nature Publishing Group.

[10] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, June 2018. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[11] S. Jialin Pan. Transfer Learning. In *Data Classification*. Chapman and Hall/CRC, 2014. Num Pages: 34.