Examining neural attentional focus in images using model embeddings

Kimberly Nestor
Carnegie Mellon University
kanestor@cmu.edu

Abstract

This research work sought to examine whether participants when presented with stimuli images of various object categories are attending more to specific portions of the image e.g. foreground (main item of the image) or background (view surrounding the image). Neural data (MEG) and stimuli images used to examine this hypothesis were obtained from the THINGS dataset. The original stimuli images were used as input to ResNET50, along with separate inputs to the model of no background (main image item only) and no foreground images (only the scene behind the main item). The second model hidden layer embedding was extracted for the original, no foreground and no background images and was compared to the MEG data using Euclidean distance. There was no distinct difference between the images groups in comparison to the MEG data, indicating participants may be attending to equal portions of the image.

Introduction

Human eye gaze tends to gravitate more towards items of an image that are later written about in greater detail in a free writing task [3]. This indicates participants are able to have greater attention to various portions of an image inherently, and this top-down attention is perpetuated into output produced when asked to recall the image through writing. Though neural network models may not particularly attend to images in the same way as humans, as shown in previous studies [1,3], human attentional data can be used to train the models to view images in a more humanistic manner, using human data.

The initial hypothesis for this work is that the MEG data should be most closely related to the original image model embeddings, with the next closest relation being the no background image embeddings as in theory participants should be focusing more on the main stimulus of the image rather than the background or other distractors. Lastly, the least related to the MEG data should be the no foreground images, as these only contain the background of the image and should be the least attended part of the image by the participants. Though models may not attend to images in the same way as humans though this work hopes to alleviate that issue by separating the parts of the image, that way there is less for the model to attend to in the no background and no foreground images.

Related Work

Humans are inherently predisposed to view more interesting parts of an image, as indicated in prior work that determined humans spend more time viewing the non-blurred portions of an image in comparison to the blurred portion [7]. This prior work is similar to this project as they showed participants an original image as well as subsequent images with non foreground aspects of the image blurred to various degrees. There results determined even when the stimuli images were not blurred participant eye gaze was fixated more on the main stimulus portion of the image compared to the background/ periphery of the image [7]. The premise of this is there is more contextual information in this portion of the image and other non-essential information in the image can be non-attended to.

Another prior work was able to distinguish between the process of attending to and ignoring presented stimuli using neural data. This work determined these processes are controlled in a top-down manner by the parietal and prefrontal cortices and illicit interaction from the sensory cortices in order to facilitate this control [5]. In fact, the frontal and parietal cortices are capable of tagging signals considered to be biases in the goal of visual attention to important stimuli, in order to facilitate the process of attention and ignoring [8].

Methods

Obtaining MEG stimulus images: The THINGS image stimulus dataset contains 1,854 different image categories, each with 14 sample images of those categories, which were all presented to the MRI participants in that portion of the study. The MEG portion of the study selected the first 12 samples of each image category to present to participants [4], and that was done for this research work as well (Fig. 1 left).

Preprocessing stimulus images: No background images were processed (Python, Rembg) to remove the image background and leave a clear white surrounding the main item of the image (Fig. 1 middle). The alpha channels of the no background image were used to create an inverse of this image as well as a mask, which was used to indicate areas in the background that needed to be filled in using inpainting (Python, OpenCV) so there were no blank holes in the background image and the missing portion would look comparable to the rest of the image (Fig. 1 right).

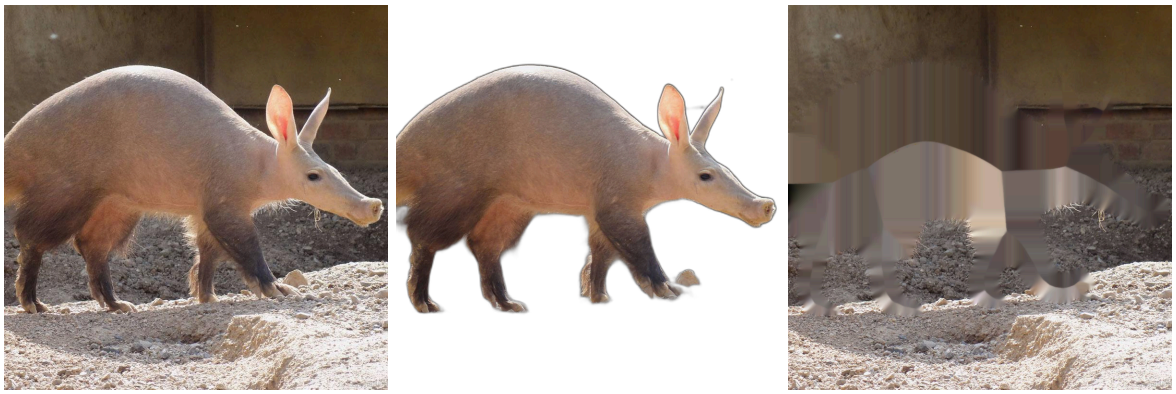

Fig. 1 showing original THINGS stimulus image (left), no background (middle) and no foreground (right) images after image processing.

Obtaining model hidden layer embeddings: The three image samples for each image category (original, no background, no foreground) were then input separately into ResNet50 (Pytorch) [2] and the second hidden layer was obtained for each image group). The second hidden layer was selected as the first would be most closely representative of the original image, while the second layer would indicate more semantic meaning of the image. The second hidden layer with more semantic meaning should have more relation to image processing in the brain. The ResNet50 model was selected for this work because it is trained on the COCO dataset [6], which is a dataset similar to the THINGS dataset, of various everyday items categories.

MEG data processing and comparison: Lanczos resampling (Python, OpenCV) was used to match the shape of the MEG data to that of the model hidden layer embeddings. The Euclidean distance (L2 norm) was then computed between the MEG data and the model hidden embeddings for original, no background and no foreground images.

Results

The initial hypothesis indicated least attended part of the image i.e. the background and in this work the no foreground or background only images should be the least like the original MEG data. While the original images and the foreground only images should be more closely related to the MEG data. The results of this work however contradict the original hypothesis as the Euclidean distances obtained (Fig. 2) from examining the pairings of model embeddings to the MEG data are all about the same. This indicates there is no difference between participant attention to the foreground versus the background of the image.

Though the results are largely the same, there is very minute difference between relation to the MEG data, as expected the MEG data compared to the original image model embeddings are the most closely related, with the no background image embeddings having slightly less relation in comparison. An interesting observation is the MEG data compared to the no foreground image embeddings is slightly closer in relation to the original MEG data than the MEG data compared to the no background images. This is an interesting observation of the results, but as previously stated the distance measurements have minute differences and are not enough to make inferences in regard to the data and participant attentiveness.

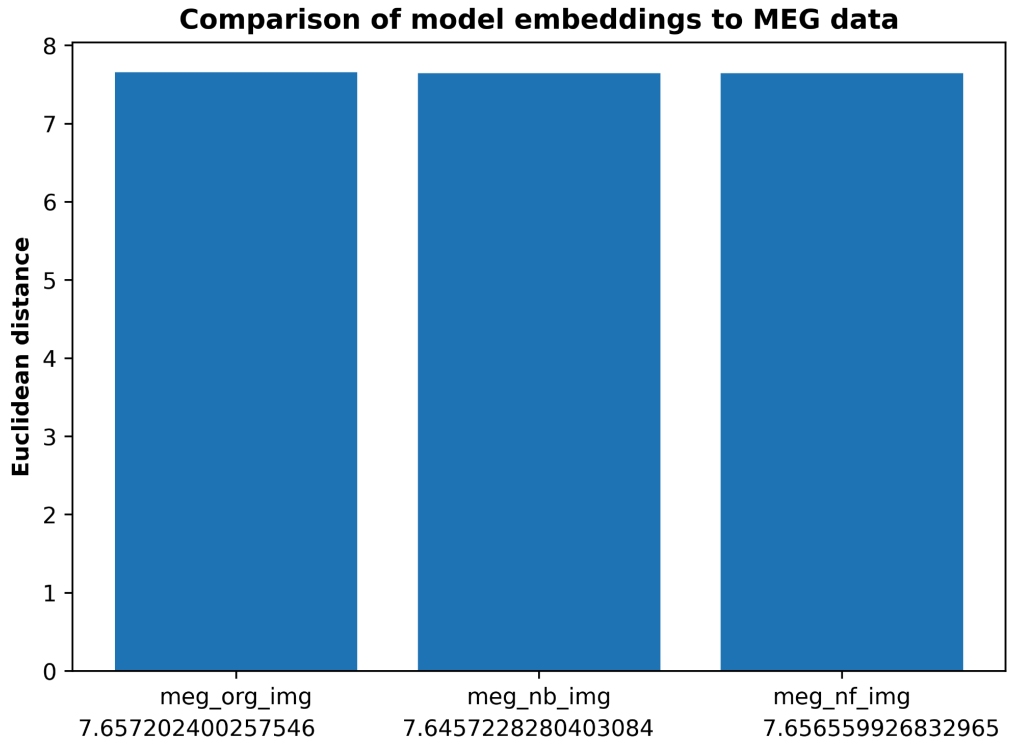

Fig. 2 showing results obtained from computing the Euclidean distance between the neural MEG data and the ResNet50 second layer model embeddings with the original images (left), the no background images (middle) and no foreground images (right).

Discussion

As stated prior the project aim was to determine whether participants attend more to specific parts of an image i.e. foreground versus background, and to examine this by comparing model hidden layer embeddings to MEG data where participants viewed images from various categories. The original hypothesis was the original image model embeddings would be most closely related to the MEG data, with the next close relation being the foreground only model embeddings and lastly the no foreground i.e. background only model embeddings. This was hypothesized because the participant would attend to the main item in the image more than the periphery which is deemed non essential to item categorization and semantic understanding. The results contradict the hypothesis however and the distance scores comparing the three image groups model embeddings to the MEG data are essentially the same.

The results obtained in this work could be because the participants are not attending to specific portions of the image, though this also contradicts prior work indicating participants can attend to

only critical image portions due to top-down attentive processing and blurring of non-essential stimuli. There is the additional confound that the THINGS image stimuli images presented to participants were cropped versions of the images to highlight only the stimulus item in the image. These cropped versions were used in the image processing of this work and as such many of the images do not have a background to speak of, or an interesting background to attend to. This lack of interesting background could be a potential limitation of this work.

References

[1] Das, A., Agrawal, H., Zitnick, L., Parikh, D., & Batra, D. (2017). Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? Computer Vision and Image Understanding, 163, 90–100. https://doi.org/10.1016/j.cviu.2017.10.001

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition (arXiv:1512.03385). arXiv. http://arxiv.org/abs/1512.03385

[3] He, S., Tavakoli, H. R., Borji, A., & Pugeault, N. (2019). Human Attention in Image Captioning: Dataset and Analysis. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 8528–8537. https://doi.org/10.1109/ICCV.2019.00862

[4] Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., & Baker, C. I. (2023). THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. eLife, 12, e82580. https://doi.org/10.7554/eLife.82580

[5] Lenartowicz, A., Simpson, G. V., Haber, C. M., & Cohen, M. S. (2014). Neurophysiological Signals of Ignoring and Attending Are Separable and Related to Performance during Sustained Intersensory Attention. Journal of Cognitive Neuroscience, 26(9), 2055–2069. https://doi.org/10.1162/jocn_a_00613

[6] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). Microsoft COCO: Common Objects in Context (arXiv:1405.0312). arXiv. http://arxiv.org/abs/1405.0312

[7] Smith, W. S., & Tadmor, Y. (2013). Nonblurred regions show priority for gaze direction over spatial blur. Quarterly Journal of Experimental Psychology, 66(5), 927–945. https://doi.org/10.1080/17470218.2012.722659

[8] Ungerleider, S. K. A. L. G. (2000). Mechanisms of Visual Attention in the Human Cortex. Annual Review of Neuroscience, 23(1), 315–341. https://doi.org/10.1146/annurev.neuro.23.1.315