UNIVERSITY OF Nebraska Lincoln

DEPARTMENT OF PSYCHOLOGY

17 February 2021

Editorial Office, *Journal of Vision*

Dear Dr. Wichmann and colleagues,

We would like to start by thanking the editor and reviewers who took the time to carefully review and evaluate our manuscript "Convolutional neural networks can decode eye movement data: A black box approach to predicting task from eye movements" originally submitted to *Journal of Vision* on September 14, 2020. We were encouraged by the enthusiasm for our work that was expressed, and we have now prepared a revised manuscript for your further consideration. Overall, we feel that the changes we have implemented in response to the thoughtful comments and recommendations from the editor and reviewers have improved the quality of the manuscript.

The enclosed document provides a verbatim copy of the initial review comments (written in black text) and our responses recorded in-line (in blue text). We made the effort to respond completely and concisely to each comment, in addition to providing a short description of how the comment or recommendation is addressed in the revised manuscript. In the revision, all changes are shown in red text.

Once again, we are grateful to the editor and reviewers for their time and consideration. We hope that these changes adequately address all issues raised by the reviewers and that you will now find the paper suitable for publication.

Best regards,
Zachary J. Cole and colleagues

238 Burnett Hall / P.O. Box 880308 / Lincoln, NE 68588-0308 / (402) 472-3721 / FAX (402) 472-4637

Reviewer #1

Review on Cole et al. "Convolutional neural networks can decode eye movement data: A black box approach to predicting task from eye movements"

Topic:
The authors trained a CCN classifier for task classification from two eye-movement datasets. The key innovative point is that time-dependent "raw" data of eye positions were used. The authors demonstrate a reliable black box solution to task classification from eye movements.

Comments:
Greene et al. and Borji & Itti investigated the long-standing Yarbus problem before. The current work, the authors claim to have contributed to its solution. The fundamental problem, from my perspective, is related to the definition of the Yarbus challenge. As Greene et al. put it, "Yarbus argued that changing the information that an observer is asked to obtain from an image drastically changes his pattern of eye movements." The "pattern" is clearly meant to be related to series fixations and saccades, not raw data. Therefore, the relation between the current study and the Yarbus problem is rather intransparent and complicated. It is known that fixational eye movements and microsaccades are highly viewer-specific and can be strongly modulated by task. Therefore, it could well be that the current manuscript's results are completely based on aspects of the data that are in a rather indirect connection to the original Yarbus problem.
Two possible solutions seem viable: First, compare results on the raw eye traces with processes data of saccade and fixation sequences. Second, rewrite the manuscript without the attempt to solve the Yarbus problem. The latter solution would make a good contribution. The first is necessary, if the authors that their black box algorithm is classifying "mental state" from eye movement data without discussing the potential bypass from cognition to fixational eye movements to classification, which has little to do with the intended long-standing Yarbus challenge.
EOF

Reviewer #1 has provided a compact statement describing a potential theoretical disconnect between our work and some of the previous literature. Although we do not fully concur with all of the issues raised here on theoretical/semantic grounds, the comments have certainly given us food for thought and an opportunity to clarify and refine some of our discussion.

The fundamental issue raised by Reviewer #1 concerns the definition of the “Yarbus challenge.” The reviewer provides a quote by Greene et al. (2012), in which the reviewer asserts that the word “pattern” is “clearly meant to be related to series of fixations and saccades, not raw data.” For this reason, the reviewer believes that our work is only indirectly related to the “original Yarbus problem.”

Although there is not perfect agreement in the literature regarding how certain terms like the “Yarbus problem” are used, we believe generally our discussion has been consistent with the way the issue is framed in a number of previous studies in this milieu. Still, we have taken this opportunity to reconsider how our undertaking is framed and to attempt to clarify the language used throughout the manuscript. In specific:

(1) Yarbus challenge/problem
The main purpose of our paper was to classify task-at-hand from minimally processed eye movement data, also known as the inverse Yarbus process or inverse Yarbus problem (Haji-Abolhassani & Clark, 2014). The original Yarbus problem (Yarbus, 1967), was to understand how task influences eye

movement data, a definition that follows with the Greene et al. quote presented by the reviewer. This means that the Yarbus problem and the inverse Yarbus problem are two separate (albeit closely related) problems. Of course, Yarbus's work was over fifty years ago and his original investigations were limited by the technology of the time; more recent work has included an increasing number of computationally oriented decoding/classification studies, with that change in focus reflected in the term "inverse Yarbus process" used by Haji-Abolhassani, Clark, and others in the field.

It is somewhat unclear from the reviewer's comments if they are referring to a more originalist interpretation of Yarbus's work or to the more modern usage we intended, but regardless, we have gone through the manuscript again and tried to make sure our usage is clear and consistent wherever possible. Please also see the response to Reviewer #2 for additional/related areas where we clarified and updated some language related to the theoretical framing of our study.

(2) Definition of "pattern" and "raw data"
The reviewer also asserts that the word "pattern" in the quote from Greene et al. was referring specifically to the use of "series fixations and saccades, not raw data." We do not necessarily agree with this interpretation; the original Greene et al. quote does not actually refer to any type of data, only the movement of the eyes. Specifically, Greene et al. say:

> "Yarbus argued that changing the information that an observer is asked to obtain from an image drastically changes his pattern of eye movements. Moreover, the scan paths from this famous figure have been taken as evidence that eye movements can be windows into rather complex cognitive states of mind."

As we can see in fuller context, the phrases "pattern of eye movements" and "this famous figure" are most clearly in reference to a qualitative evaluation of the scan paths in the well-known Yarbus figure named "Unknown Visitor," which is reprinted in Greene et al. (2012) and many other articles. Not coincidentally, these scan paths bear a non-trivial resemblance to the image-based representation of scan paths analyzed in our own paper.

We have re-reviewed the work of Greene et al. as well as the other studies we listed in Table 1 of our manuscript. These studies used a wide variety of features extracted from eye tracking data, including explicit measures of fixations and saccades but also pupil size (MacInnes et al., 2018), eccentricity and screen coverage (Król & Król, 2018), and more. After reviewing the literature, we have not been able to find a clear dividing line in the types of features used. Granted, ours is the only study that used raw timeline data (although, as noted, we also used minimally processed images of scanpaths), but if the only "true" approach to the Yarbus problem is to use exclusively fixation/saccade data, it would appear that several other studies in the literature are equally guilty of using alternative types of features.

After considering these questions thoroughly, we have been forced to conclude that our interpretation of the inverse Yarbus problem simply appears to differ from that of Reviewer #1. We believe that (a) there is an important distinction between the Yarbus problem and the inverse Yarbus problem, and (b) valid solutions are not required to only use eye movement data processed into particular forms. Although there is always room for differing interpretations, we also find that our stances on these positions are well-supported within the relevant literature, and generally consistent with other recent studies in this area of investigation. However, as before, we certainly appreciate the opportunity to reconsider our viewpoints and examine whether we have clearly communicated these ideas, which we hope we have accomplished in the revised manuscript. If there are any particular points that are of significant concern to the reviewer

or editor, we would be happy to continue the dialogue on these matters and address any of those areas of major concern in another revision.

Reviewer #2

In the present manuscript, the authors apply deep learning to predict observer task from eye movement data. They collect multiple datasets where observers viewed images while they had to do one of multiple possible tasks. They test two different modeling approaches. The first one is a DNN operating directly on timelime data of x and y position and pupil size. The second approach first encodes the timeline data into images of gaze traces and then applies a DNN to these images. The authors find that both models can predict observer task with above chance accuracy, however, the image-based model has substantially lower performance than the timeline based model. In ablation analyses they find that x position is most important for task classification, while pupil size is least important.

The effect of tasks on eye movement is very interesting, as can be seen from the long history of research on it. The paper is working on a very relevant question and is written in a very clear and understandable way, which I very much appreciate.

In the following, I am going to list some conceptual and technical issues that I see. Once they are addressed, I think this work can be very relevant for the community.

Conceptual points

1. First, I was a bit puzzled by the use of the terms "task", "cognitive process" and "mental state". In the paper, they seem to be used mainly interchangeably, but I think I would not agree with that. For me, intuitively, the task defines the goal of an observer, the cognitive process is the cognitive part of how the observer tries to reach this goal and the mental state is where in this process the observer is at a given time. Obviously, the task influences the cognitive process and the task will always be part of the mental state, but to me, cognitive processes and mental states carry much more information than the task. To me, all experiments and results in the paper seem to be about observer task, so I think conclusions about cognitive processing and mental states need more discussion.

We agree that the terms “task,” “cognitive process,” and “mental state” are all related, but do not carry the same meaning. Potential confusion of these terms is a matter of concern to us as well and an area in which we feel like much of the literature could improve, although as the reviewer notes, we could stand to make these distinctions clearer and more explicit in our own writing as well.

To build off of what the reviewer has said, eye movement features are instantiated by, and can be indicative of, cognitive processes. In the current study, these features are extracted and decoded by the deep learning CNN to determine the task-at-hand. Our own taxonomy is fairly similar to the reviewer’s; we would generally say that cognitive processes are more theoretical constructs that are difficult to isolate, whereas a cognitive task is typically (in an experimental context) a more explicit set of goals and behaviors imposed by the experimenter in an effort to operationalize one or more cognitive processes. A mental state would be another more theoretical term that is a bit more general/generic, which could include a set of ongoing goals and cognitive processes but could also presumably encompass elements like mood or distraction that are not typically explicitly manipulated in tasks of this sort.

We agree with Reviewer #2 that this language can be further clarified in the manuscript. For this reason, we have gone through the manuscript and considered each occasion on which any of these terms (or related terms) are used. Wherever we could find that we were not using the optimal term for our intended meaning, or where we thought the intention might be unclear, we have revised the manuscript to clarify the distinction between the three terms. (See lines 96–102 of the revision for our working definitions of these terms and smaller edits on line 281 and line 412; elsewhere, we felt that the usage matched our working definitions.)

2. The authors do not pass any stimulus information to their models. In the introduction, they argue that "such efforts to not fit the spirit of the inverse Yarbus problem, which is concerned with decoding high-level abstract mental operation that are not dependent on particular stimuli". I'm not sure whether I can agree with this. I would argue that the high-level abstract mental operation operates on the content of the viewed image: to guess how wealthy the people in an image are, I need to figure out where the people are in the image, I will likely inspect their clothing and other attributes and then process this data to come up with a guess of their wealth. Therefore, the eye movements are highly dependent on the stimulus itself. Without access to the stimulus, I don't see how one could expect to decode the mental operation that is going on. Of course there will be differences in the pure eye movements (as can be seen in the above-chance performance of the model), however they might be best explained in combination with the image viewed. For example, if the relevant areas of the image for the task at hand are more scattered over the image, longer saccades might be the result. Another example is given by the authors in the discussion, line 470: pupil size might be mainly affected by stimulus properties. Overall, I'm not sure whether one can expect to decode the "mental operation" or the "cognitive process" purely from the gaze data. However, this might boil down to my first point: Maybe I'm misunderstanding what the authors mean when they talk about mental operations etc.

We agree with the reviewer that there was likely some miscommunication of the message we were trying to get across with regard to the spirit of the inverse Yarbus problem, and with regard to the use of stimulus information in the classification of eye movement data.

In the quote mentioned by the reviewer, we were referring to previous research that has decoded eye movement data collected while participants were carrying out tasks such as reading vs. evaluating an image. In this case, reading requires mostly horizontal (left-to-right) eye movements following lines of text, with a consistent demand to scan left-to-right/top-to-bottom in a manner that would not be efficient for other task types. For this reason, the scan paths of reading are very obviously qualitatively different from memorizing or evaluating an image. These low-level distinctions are so easily differentiated that the more subtle high-level distinctions between these tasks can be ignored. In these cases, the tasks are tied directly to properties of the stimulus rather than properties of the cognitive processes differentiating the tasks-at-hand. This is what we feel doesn't fit the spirit of the inverse Yarbus problem. Obviously, as the reviewer points out, all visual processing relies to some extent on the low-level properties of the image, and where to draw the line is somewhat of a judgment call. In our estimation, though, what separates those more bottom-up-driven studies from studies like ours is that in our study, for example, the same pictures can be used for all of the cognitive tasks; no explicit instruction is given about how the participants should scan the image, and nothing about the conjunction of image and task implies any particular bias in the types of eye movement a participant should make.

We do agree with the reviewer that stimulus information could improve the decodability of the processed images. This is actually mentioned in our Discussion section as a potential future direction to take with this line of research. In reality, as noted above, visual processing of images is achieved via an interaction of eye movements and low-level stimulus information, and we suspect that the most accurate classification possible

would involve giving the classifier the same kind of input we give our visual systems by moving our eyes across an image. However, for whatever reason, most of the studies we review and much of the Yarbus-inspired literature of the past fifty years have focused largely on properties of the eye movements themselves, such as the fixations and saccades mentioned by Reviewer #1. The properties of the visual stimuli may implicitly be reflected in those eye movement properties – for example, if the task is to judge the age of a person, saccades may be shorter on average than if the task were to judge broader properties of the room they are standing in – but most of the "inverse Yarbus problem" literature has allowed that reflection to remain implicit by focusing their efforts on recorded eye movements, rather than including explicit stimulus information.

To more clearly communicate our intention to distinguish between studies where the pattern of eye movements is primarily driven directly by the structure of the stimuli, versus studies where the stimuli and task-at-hand are independent (with the subject deciding on their own how to scan the stimulus based on the task-at-hand), we have revised the passage in question (lines 116–120). As always, if there is any continued disagreement on the matter, we are happy to continue the conversation further and consider additional revisions.

3. A major result of the paper is the comparison of the timeline model and the image model, where the authors find that the timeline model works much better for task classification. I'm not sure I fully understand what the authors hope to show with this. In theory, both models have access to exactly the same data (except for a few occlusions that should be inferable), so with sufficient computational capacity and training data, both models should reach exactly the same performance. Therefore, the differences in performance are based in different inductive biases of the different architectures and in training and overfitting problems. If I'm not mistaken, the image model has more than 100 times as many parameters as the timeline model and therefore can easily suffer more from overfitting. In addition, it is trained from scratch. I could imagine that adding a ImageNet pretrained backbone such as ResNet or DenseNet can improve performance (I'm aware that the input images are quite far from natural images, but even on out-of-domain data, deep models often still encode surprisingly useful features). However, besides all of those points, I can imagine that it is very hard to bring the image based model to the same performance as the timeline model. After all, the features that so far have been most successful in predicting task (see Table 1 of the manuscript) are much easier to compute from timeline data than they are to compute from the image data. Of course, if stimulus information should be included in the models, then the image-based model makes a lot of sense. Overall, I think the paper would profit if the authors state more clearly what they hope to learn from comparing these two models.

Deep learning CNNs are known for their proficiency in decoding image data. We were originally planning to process the data into images in an attempt to take advantage of the CNN's ability to decode data in this form. Some other studies (e.g., Bashivan et al., 2016) have found success in transforming timeline data (in that case, EEG data) into an image-based format before classification. In addition, earlier studies by Yarbus (1967) and others presented eye movement traces in a similar image representation as the original justification for an association between eye movements and cognitive states, so there is significant historical precedent for thinking of that format as a meaningful representation of eye movement patterns. We also classified the data in the timeline format because it is the native and more raw form of eye movement data, and to provide a relative baseline for the image classification. Of course, as it turned out, the timeline data ended up outperforming the image-based representation.

Although it is true that "In theory, both models have access to exactly the same data," we and others have previously found that transforming data into another representation (even if it is formally equivalent, or at least approximately equivalent) can help or hinder the performance of certain classifiers. For example, if

relevant features are encoded in the frequency domain more than the time domain, it can be useful to pre-process data with a Fourier transform before classification, even if it is theoretically possible for a neural network to learn that operation on its own given sufficient time and training data. However, it is not always obvious what format best represents the relevant features of the data without putting it to the test empirically. Hence another motivation for explicitly comparing the two representational formats.

In terms of parameters, the differences were not as large as they may have seemed; the timeline data classifier had 16,946 trainable parameters and the image data classifier had 18,525 trainable parameters (only a 9% difference, which is close to negligible, given the possible ranges of neural network sizes).

In the original submission, we mentioned some of our rationale for investigating the image data format, but did not explicitly state the reason for the comparison of these model types. Clearly, the points mentioned by the reviewer have helped us realize that it would be useful to include more of this information and rationale in the paper. Thus, we have added more explicit information on complexity (parameter counts) of the models in the Figure 3 caption, and we have revised lines 179–190 to include more of our justification for considering both of these formats.

Other points

1. In the introduction, lines 150-163 the authors argue that due to different datasets used, it is hard to compare the different approaches used for classifying task in the past. I fully agree, however, I think this would be even more reason to check at least one or two of the better performing methods from Table 1 on the collected dataset. The blackbox DNN should outperform them since it has access to the full raw data (see also in the discussion, lines 421-423).

We agree with the reviewer's suggestion that it would be helpful to clarify the efficacy of our approach relative to others. At this point, it would be hard to say whether access to the full raw data would actually lead to better performance of our approach compared to another. This is because the other approaches use data that is processed into a different structure, and it could be that data processed in the format of the other approaches is more indicative of the task-at-hand and more differentiable between tasks.

To address this issue, we obtained the code used by Coco and Keller (2014) to process and classify their eye movement data. Using our Exploratory dataset, we processed the data into seven of the features used by Coco and Keller (i.e., initiation time, number of fixations, mean entropy, mean saccade amplitude, mean fixation duration, percent area fixated, and mean fixation saliency). These features were classified using all three model approaches used by Coco and Keller (SVM, LASSO, MM). As suggested by the reviewer, our approach outperformed Coco and Keller's, which was essentially at chance in our dataset. The drastic drop in performance of the Coco and Keller approach between their data and ours is likely due to the explicit efforts made by Coco and Keller to differentiate the tasks in a way that would not be dependent on the analysis approach, which they demonstrated by using Greene et al.'s approach to successfully classify their data (Greene et al.'s approach failed on their own data). Our dataset contains tasks that are capable of being distinguished (as we, and others using the same tasks, have demonstrated), but these tasks are not as easily distinguishable as the tasks used by Coco and Keller. For this reason, to successfully classify our dataset may require a more complex analysis approach.

We would have liked to run more comparisons using the approaches listed in Table 1, and made a good-faith attempt to do so, but were unable to run the others either due to unavailability of public, up-to-date code or

data, or incompatibilities between our dataset and others' approaches (or between their datasets and our approach). This helped us realize the importance of making our own dataset and code public so that future studies might be able to perform more explicit comparisons between approaches, using our data and/or code as a baseline. Thus, in our revised discussion we have included a mention of our attempt to replicate Coco and Keller's approach as well as a public link to our dataset and code, hosted on the Open Science Framework (see lines 451–459 in the revision).

2. In the ablation studies, the authors use datasets where one or more components have been removed (X, Y, P). From the manuscript, I'm not sure whether the models where retrained on the new data, or whether the already trained models where evaluated on the reduced data. For assessing the relevance of the different components, I would argue that the models should be retrained. Zeroing out some components introduces a substantial domain shift and model performance might just drop because of this.

We agree that zeroing out components without retraining the models would have been a problem; in the original manuscript, the models were indeed retrained. However, we thank the reviewer for pointing out that this was not clear in the paper; we have revised lines 306–309 to make that explicit.

3. On a related note, if I see it correctly, for the generalization test from exploratory to confirmatory data, the models are completely retrained. Here, I think I would have gone for not retraining the models since the distribution shift is much smaller and it is actually interesting to see how well the confirmatory data can be predicted from the explanatory data (as opposed to other confirmatory data). But I don't request that the authors do that, I just wanted to point out that here, both approaches make sense. They are just answering different questions.

We can appreciate the reviewer's comment on the decision to retrain the confirmatory model using the new dataset. We chose to retrain the confirmatory model because we felt like this was the best way to validate the efficacy of our overall approach, not just the model we used. While the model is certainly an important aspect of the study, we feel that the unique method of decoding the scan paths as images, and using the minimally processed datasets is the most interesting contribution that our manuscript brings to the field. However, we agree that testing the generalizability of a specific trained model between datasets/experiments is an interesting separate question of its own, and one we may pursue in future studies.

4. Training/test split (line 270): If I understand this correctly, for each iteration, a new random training/test split was sampled. This seems a bit unusual to me, in deep learning usually either a fixed train/test split is used multiple times to assess the variance of the different random initializations of the model, or, sometimes, a full k-fold crossvalidation is used to make best use of all available data (but then error bars might be less relevant). Now the error bars will be partially due to the different initializations and partially due to the fact that different datasets were evaluated, that might be slighly different in their difficulty.

We appreciate that this is a slightly less common cross-validation approach than others the reviewer may have seen, although it is not without precedent and conceptually is not that different from a k-fold cross-validation. K-fold cross-validation can be particularly useful in cases where the classification requires only a training and test set (e.g., SVMs). However, here we also needed a validation set for determining when to stop training the deep learning model, which complicates the combinatorial nature of the k-fold approach. (How best to approach cross-validation in deep learning for psychology and cognitive neuroscience data is an interesting side question of its own; we actually have an ongoing project exploring this issue more explicitly, although it is still too early in the analysis/writeup process to cite in this paper.) The approach we used essentially achieves the same effect as k-fold cross-validation insofar as it allows all data to be used for

both training and test (and for validation), but it does so with randomly resampled folds rather than with a one-time fold division that is then iterated through. The advantage is that the training/validation/test sets can all be different sizes and the process is overall simpler to implement than trying to integrate a validation set into the traditional k-fold cross-validation paradigm. In case other readers have similar questions, we have added a little more detail on this procedure to our methods in lines 291–295.

Minor points

* please elaborate a bit on the relative computational capacities of the two models, i.e. the parameter counts. Was overfitting a problem for the image data model?

As noted above, details regarding the parameter counts have have been added to the model description in the manuscript (Figure 3 caption). Overfitting is somewhat difficult to demonstrate conclusively, but during our model exploration process on the Exploratory dataset, we attempted to find architectures that performed well on the held-out test data subset and did not appear to be excessively overfitting during training. We have made this point more explicit in the revision in lines 301–302.

* I'm missing some information on the used learning rates. Was there no learning rate decay at all? I would expect the models to profit from using some sort of learning rate schedule (even when using ADAM).

Actually, there was a decay in learning rate; we used the default values for the Adam optimizer and originally did not include that level of detail in the paper, but this comment has convinced us that it would be useful to some readers to include all of the optimizer (hyper)parameters. This information has been added to the model description in the revised manuscript (Figure 3 caption).

* "To determine the relative value of the contribution from each component", [ANOVA was performed] (lines 296): This is really only a subtle point, but I would argue that ANOVA only determines whether there is any effect at all but not the relative value of each component. The relative value could be measured, e.g., by the differences in performance. Obviously, a statistical test is required to assess that these differences are meaningful, but ultimatively, in my opinion, it is more interesting to see how large the effect is.

We see that our initial phrasing was slightly inaccurate, as pointed out; the ANOVA indeed demonstrates the existence of an effect, but not the size of the effect as the phrase “relative value” seems to imply. We have revised this phrasing in the line mentioned and in a few other places where similar phrasing occurred. We agree that the actual sizes of the effects could be interesting as well, but if we understand the question correctly, it seems like estimating those sizes could turn into a relatively complex investigation of its own. For example, suppose that pupil size is partially redundant with X and partially redundant with Y; capturing this relationship quantitatively might involve some fairly sophisticated comparisons between not only the XYP and XYØ data subsets, but also between XØØ and XØP, YØØ and YØP, and so on. One might get at this question by examining the covariance structure of the data components, but that raw covariance structure may or may not relate directly to the ultimate redundancies in information content that are revealed after classification. For this reason, our judgment for now has been that a detailed investigation might lie outside the scope of the current paper and could be better suited to a separate, standalone quantitative paper on that topic; however, if the reviewer disagrees, we are happy to consider the issue further.

* Figure 4: the dataset without pupil data seems to result in slightly larger performance than the full dataset. This is within the margin of error, but I wonder whether it indicates any overfitting problems (although it

should not reduce the effective capacity of the model substantially). It might be interesting to discuss this result at least in a few words.

We agree that this is a potentially interesting phenomenon. We are unsure of what might be driving it; overfitting is indeed one possibility. However, the effect was not significant or particularly near statistical significance, as reported in Table 2 ($p=.372$ for Exploratory dataset and $p=.288$ for Confirmatory). As such, it is also quite possible that this could simply be due to sampling fluctuations, and thus we did not feel comfortable speculating about it in the original submission. For the same reason, we have not yet added any discussion of the matter in the revision, although we have made note of this internally to keep an eye on in future studies; and if the reviewer strongly feels that we should still discuss the effect, we are happy to revisit the topic.