Convolutional neural networks can decode eye movement data: A black box alternative for solving the inverse Yarbus problem

Zachary J. Cole[1], Karl M. Kuntzelman[1], Michael D. Dodd[1], & Matthew M. Johnson[1]

[1] University of Nebraska-Lincoln

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Correspondence concerning this article should be addressed to Zachary J. Cole, 238 Burnett Hall, Lincoln, NE 68588-0308. E-mail: z@neurophysicole.com

Abstract

We learned so deeply we incepted the inferred cogntive processes that underlie the inferred eye movement features.

Keywords: deep learning, eye tracking, convolutional neural network, cognitive state, endogenous attention

Word count: X

Convolutional neural networks can decode eye movement data: A black box alternative for solving the inverse Yarbus problem

Introduction

A goal of many cognitive neuroscientists is to infer mental activity from human eye movements . Foundational work by Yarbus (1967) showed that eye movements patterns appear to differ qualitatively depending on the task at hand. Technological advances and improvements in computing power have allowed researchers to decode and classify the mental state underlying eye movement data, also known as the "inverse Yarbus process" (Haji-Abolhassani & Clark, 2014, p. 127). Current state-of-the-art machine learning and neural network algorithms that are capable of identifying diagnostic patterns in the data appear to trade one inverse problem for another (). As pointed out by Lukander (), "the black box nature of the resulting solution impedes generalizability, and makes applying methods across real life conditions more difficult" (p. 44). To ground these solutions, researchers guide "black box" decoding efforts by providing data and/or models with built-in theoretical assumptions. At this point, there is no clear evidence to support the notion that the commonly applied theoretical constraints actually enhance or clarify "black box" solutions beyond what could be inferred from an unconstrained model .

Consider the case of Greene et al. (), who failed to classify task from commonly used aggregate eye movement features using three separate model architectures. This led Greene et al. to question the robustness of Yarbus' (1967) findings, inspiring a slew of responses that successfully decoded the same dataset using different eye movement features, or different model architectures (e.g.,). The subsequent re-analysis of these data support Yarbus (1967), but lack a unifying theoretical framework that comprehensively explains the failures of Greene et al. and the successes of the subsequent re-analyses.

In general, the selection of tasks, data, and decoding algorithms have been implicated separately in the successful decoding of eye movement data (). Eye movements can only be

differentiated to the extent that the cognitive processes underlying the tasks can be delineated (). Every task is associated with a unique set of cognitive processes (). To distinguish the cognitive processes underlying task invoked eye movements, some studies have chosen to classify tasks that rely on distinct exogenous influences on attention, such as reading text and searching pictures (e.g.,). When the cognitive impetus for the task is the product of exogenous attention, the resulting eye movements thought to reflect complex mental state may actually be confounded by the presence of salient stimulus features (). Additionally, decoding algorithms appear to prioritize these bottom-up patterns in the data over higher-level top-down effects (). For this reason, eye movements associated with exogenously oriented tasks can be reliably decoded (), but do not fit the implied top-down nature of the inverse Yarbus problem.

Despite a lack of objective criteria differentiating the mental activity underlying these endogenously oriented task sets, a comparison of aggregate eye movement features confirmed that eye movements can be differentially influenced by the task at hand (). Decoding of similar eye movement features under similar task conditions has produced classification accuracies typically ranging from chance performance to 59.64% (see Table x). In one case, Coco and Keller () categorized eye movements based on visual or linguistic components of three tasks, resulting in 84% accuracy. A recent follow-up using a different task set categorized four tasks according to two objective spatial and semantic processing dimensions with 51% accuracy (). A closer look at these results showed that the categories within the semantic dimension were consistently mixed up, suggesting that this level distinction may require a more rich dataset, or a more powerful decoding algorithm.

To prepare eye movement data for classification, fixation and saccade statistics are typically aggregated along spatial or temporal dimensions, resulting variables such as fixation density or saccade amplitude (). Implementing these statistical methods is meant to explicitly focus the algorithm on characteristics of the eye movement data that are

representative of theoretically relevant cognitive processes (). In some instances, such as the case of Krol & Krol (), aggregating the data may ground the data structure theoretically while sacrificing resolution. Given the cognitive processes underlying distinct tasks are often overlapping (), decreasing the resolution of the data may actually limit the potential of the algorithm to decode the eye movement data ().

The current study aims to maximize the potential of the data by submitting unprocessed x-coordinate, y-coordinate, and pupil size data to a convolutional neural network (CNN) model. CNNs have a natural propensity to develop low level feature detectors similar to primary visual cortex (). For this reason, CNNs are commonly implemented for image classification (). To test the possibility that the image data are more suited to the CNN classifier, the data will be decoded in raw timeline and image formats. To our knowledge, no study has attempted to address the inverse Yarbus problem using: (1) non-aggregate data, (2) image data format, or (3) a CNN architecture. Given that CNN classification performance is robust to multidimensional, non-structured data (), we expect the non-theoretically constrained CNN architecture to decode both data types at levels consistent with the current state-of-the-art. Furthermore, we expect that despite the claims that "black box" approaches to the inverse Yarbus problem are unreliable (), our initial findings will replicate when tested on an entirely separate dataset.

Methods

Participants

Two separate datasets were used to develop and test the deep CNN architecture. The two datasets were collected from two separate experiments referred to as exploratory and confirmatory. The participants for both datasets consisted of college students (Exploratory $N = 124$; Confirmatory $N = 77$) from a large Midwestern university who participated in exchange for class credit. Participants who took part in the exploratory experiment did not participate in the confirmatory experiment.

Materials and Procedures

Each participant viewed scene images (see Figure x) while carrying out a search, memorization, or rating task. For the search task, participants were instructed to find a "Z" or "N" embedded in the image. If the letter was found, the participants were instructed to press a button which terminated the trial. For the memorization task, participants were instructed to memorize the image for a test that would take place when the task was completed. Memory was tested by asking participants to select which of two images they had seen during the task. For the rating task, participants were asked to think about how they would rate the image on a scale from 1 (very unpleasant) to 7 (very pleasant). The participants were prompted for their rating immediately after viewing the image. The same materials were used in both experiments with a minor variation in the procedures. In the confirmatory experiment, participants were directed as to where search targets might appear in the image (e.g., on flat surfaces). No such instructions were provided in the exploratory experiment. In both experiments, trials were presented in one mixed block, and three separate task blocks. For the mixed block, the trial types were randomly intermixed within the block. For the three separate task blocks, each block consisted entirely of one of the three tasks (search, memorize, rate).

Apparatus

Eye movements were recorded using an SR Research EyeLink II eye tracker with a sampling rate of 1000Hz. On some of the search trials, a probe was presented on the screen at six seconds. To equate the data from all three conditions, only the first six seconds of each trial was analyzed. Trials that were missing data were excluded before analysis. For both datasets, the trials were pooled across participants. After removing bad trials, the exploratory dataset consisted of 12,177 trials, and the confirmatory dataset consisted of 9,301 trials.

Datasets

Data were extracted from the two experimental datasets in timeline and plot image formats, then classified. The raw x-coordinate, y-coordinate, and pupil size data collected at every sampling time point in the trial were used as inputs to the deep learning classifier.

For the plot image datasets, the timeline data for both experiments were converted into scatterplot diagrams. The x and y coordinates and pupil size were used to plot each sample collected by the eye tracker on a scatterplot diagram (e.g., see Figure X). The coordinates were used to plot the location of the dot, pupil size was used to determine the relative size of the dot, and shading of the dot was used to indicate the time-course of the eye movements throughout the trial. The background of the plot images and first data point was white. The final data point was black. Each subsequent data point in between was incrementally darker until final data point was reached. To ensure every data point was fully represented within the scatterplot image, the following equation was used to adjust the size of the data points: dot size = (pupil size)/10 + 1. The plots were sized to match the dimensions of the data collection monitor (1024 x 768 pixels) then shrunk to (240 x 180 pixels) in an effort to limit file size.

Parcellations. To systematically assess the predictive value of the data provided by each of the three eye movement dimensions (x-coordinates, y-coordinates, pupil size), plots were made with each of the dimensions removed. Plots were also made only using x-coordinate data, y-coordinate data, and pupil size data (see Figure X). For each of these separate sets of plots, a raw timeline dataset of the corresponding image dimensions (i.e., x-coordinate only, y-coordinate only, pupil size only) was also developed.

Classification

Deep CNNs model architectures were implemented to classify the trials into categories of search, memorize, or rate. Each model split the data into 70% training, 15% validation, and 15% testing. Each network was run through 10 iterations of the data. The same

decoding models were run on the raw timeline data, and the image data.

The exploratory models consisted of convolutional layers, fully connected layers. To maximize the accuracy of the exploratory model, model parameters were adjusted and tested using the exploratory dataset. Adjustments consisted of changing the kernel size, stride rate, and the number of filters. In total, 16 models were tested (see Table x). The same models were used to decode the raw timeline data and the images. The most accurate model is shown in Figure x. This model consisted of two convolutional layers, and one fully connected layer (see Figure x). This final model was validated on the confirmatory dataset.

Analysis

Results for the CNN architecture that resulted in the highest accuracy on the exploratory dataset are reported below. For every dataset tested, a one-sample t-test was used to compare the CNN accuracies against chance (33%). The Shapiro-Wilks test of normality was conducted to test the normality for each dataset. When normality was assumed, the mean accuracy for that dataset was compared against chance using Student's t-test. When normality could not be assumed, the median accuracy for that dataset was compared against chance using Wilcoxon's Signed Rank test.

To determine the relative value of the three components of the eye movement data, the parcellated datasets were compared within the timeline and plot image data types. If classification accuracies were lower when the data was parcellated, the component that was removed was assumed to have some diagnostic contribution that the model was using to inform classification decisions. To determine the relative value of the contribution from each component, the accuracies from each parcellation with one dimension of the data removed were compared to the accuracies for the non-parcellated dataset using a one-way between-subjects Analysis of Variance (ANOVA). To further evaluate the decodability of each component independently, the accuracies from each parcellation containing only one

dimension of the eye movement data were compared within a separate one-way between-subject ANOVA. All post-hoc comparisons were corrected using Tukey's *HSD*.

Results

Timeline Data Classification

Exploratory. Classification accuracy for the timeline data were well above chance (M = .526, SD = .018; $t(9)$ = 34.565, $p < .001$). Accuracy for classifications of the systematically parsed data were all better than chance (see Table x).

There was a difference in classification accuracies for the non-parcellated dataset and those that had the pupil size, x-coordinate, and y-coordinate data systematically removed ($F(3, 36)$ = 47.471, $p < .001$, *eta^{2}* = 0.798). Post-hoc comparisons showed that when compared to the full dataset, removing the pupil size ($t(18)$ = -1.635, p = .372) and y-coordinate ($t(18)$ = 2.645, p = .056) did not affect classification accuracy, suggesting that these data were not informing classification judgments made by the CNN. Classification for the dataset with the x-coordinates removed was lower than classification for the full dataset ($t(18)$ = 9.420, $p < .001$), showing that these data were relatively important criterion in classification decisions.

There was also a difference in the classification accuracies for the parcellated datasets containing only the x-coordinate, y-coordinate, and pupil size data ($F(2, 27)$ = 75.145, $p <$ .001, *eta^{2}* = 0.848). Post-hoc comparisons show that classification accuracies for the the pupil size dataset were lower than the x-coordinate ($t(18)$ = -12.213, $p < .001$) and y-coordinate ($t(18)$ = -7.026, $p < .001$) datasets, and accuracies for classification of the x-coordinate dataset were higher than for the y-coordinate dataset ($t(18)$ = 5.187, $p < .001$). These findings suggest that pupil size is the least decodable criterion informing classification decisions, while the x-coordinate data was the most decodable.

Confirmatory. Classification accuracy for the confirmatory timeline dataset was well above chance ($M = .537$, $SD = 0.036$, $t(9) = 17.849$, $p < .001$). Accuracy for classifications of the systematically parsed data were also all better than chance (see Table x). Overall, there were some discrepancies in the pattern of results describing the relative contribution of the x- and y-coordinate data to the model, but the general trend showing that pupil size was the eye tracking data component least informative to the model remained stable in both datasets (see Table x).

To test generalizability of the model to other eye tracking data, classification accuracies for the non-parcellated exploratory and confirmatory timeline datasets were compared. The Shapiro-Wilk test for normality indicated that the exploratory ($W = 0.937$, $p = .524$) and confirmatory ($W = 0.884$, $p = .145$) were normally distributed, but Levene's test indicated that the variances were not equal, $F(1, 18) = 8.783$, $p = .008$. Welch's unequal variances t-test did not show a difference the two datasets, $t(13.045) = -0.907$, $p = .381$, Cohen's $d = -0.406$. These findings suggest that the confirmatory dataset classifications were less precise , but the deep learning model decoded the exploratory and confirmatory timeline datasets equally well.

Plot Image Classification

Exploratory. Classification accuracy for the plot image data type was better than chance ($M = .436$, $SD = .020$, $p < .001$), but was less accurate than the classifications for the exploratory timeline data ($t(18) = -10.813$, $p < .001$). Accuracy for the classifications for all parcellations of the plot image data except the Pupil Size Only dataset were better than chance (see Table x). The parsed plot image dataset classification accuracies were not compared to the parsed timeline dataset classification accuracies.

There was a difference in classification accuracies for the non-parcellated dataset and those that had the pupil size, x-coordinate, and y-coordinate data systematically removed

($F(4, 45) = 7.093$, $p < .001$, *eta^{2}* = .387). Post-hoc comparisons showed that when compared to the full dataset there was no effect of removing the pupil size ($t(18) = -0.474$, $p = .989$) or x-coordinate ($t(18) = -1.792$, $p = .391$) data, but classification accuracy was worse when the y-coordinate data were removed ($t(18) = 2.939$, $p = .039$).

There was also a difference in the classification accuracies for the parcellated datasets containing only the x-coordinate, y-coordinate, and pupil size data ($F(2, 17.993) = 228.137$, $p < .001$, *eta^{2}* = .899). Because Levene's test revealed unequal variances between the groups ($F(2, 27) = 3.815$, $p = .035$), the Welch correction was used to interpret the findings of this omnibus ANOVA. Post-hoc comparisons showed that there was no difference in the classification accuracies for the x-coordinate and y-coordinate datasets ($t(18) = 0.423$, $p = .906$), but classification for the pupil size dataset were lower than the x-coordinate ($t(18) = -13.569$, $p < .001$) and y-coordinate ($t(18) = -13.235$, $p < .001$) datasets.

Confirmatory. Classification accuracy for the confirmatory image dataset remained well above chance ($M = .449$, $SD = 0.012$, $t(9) = 31.061$, $p < .001$), but was less accurate than the classifications for the confirmatory timeline data ($t(18) =$, $p < .001$). Accuracy for classifications of the systematically parsed data were also all better than chance (see Table x). As with the timeline data, there were discrepancies in the pattern of results describing the relative contribution of the x- and y-coordinate data to the model, but the general trend showing that pupil size was the eye tracking data component least informative to the model remained stable in both datasets (see Table x).

To test the generalizability of the model, the classification accuracies for the non-parcellated exploratory and confirmatory plot image datasets were compared. The independent samples *t*-test showed that the deep learning model did equally well at classifying the exploratory and confirmatory plot image datasets, $t(18) = -1.777$, $p = .092$, Cohen's $d = -0.795$.

Discussion

The results supported our hypothesis that the CNN would decode the timeline data with state-of-the-art accuracy. The image data were decoded above chance, but not to the standard set by the state-of-the-art, as hypothesized. The pattern of findings between the exploratory and confirmatory datasets also supported our hypothesis that the results will be reliable and replicated between datasets. To probe the resolution of the data, we compared accuracies between the timeline and image datasets and compared relative value of the raw components of the eye movement data (x-coordinates, y-coordinates, pupil size). The timeline dataset was more accurately decoded than the image dataset. When comparing the parcellated datasets, pupil size was the least informative component of the eye movement data. The implications of these findings are discussed further below.

Although several aggregate eye movement features have been tested as task predictors (), to our knowledge no other study has assessed the predictive value of the data format (viz., data in the format of an image). Our results suggest that although CNNs are robust image classifiers, eye movement data is decoded in the standard timeline format more effectively than in image format. This may be a consequence of the inherent resolution of these data formats. Over the span of the trial (six seconds), the eye movements occasionally overlapped. When there was an overlap in the image data format the more recent data points overwrote the older data points. This resulted in some data loss that did not occur when the data was represented in the standard timeline format. Despite the loss of overwritten data, the image format was still classified with better than chance accuracy. To further examine the viability of classifying task from eye movement image datasets, future research might consider decoding 3-dimensional data formats, or more complex color combinations capable of representing overlapping data points.

Datasets with lower classification accuracies confused the memorization condition with the search and rate conditions. This suggests that the eye movements associated with the

memorization task are likely indicative of underlying cognitive processes that are shared by the search and rate tasks. Previous research (i.e., Krol & Krol) has attributed the inability to differentiate one condition from the others to a lack of clarity in the data. This attribution is supported in the data by evidence that the parcellated datasets, with fewer defined variables, classified the memorization task less definitively than the other tasks. In cases when the parcellations were decoded as well as the main dataset, the memorize condition was classified as accurately as the other conditions.

In determining the relative contributions of the the eye movement features used in this study (x-coordinates, y-coordinates, pupil size), pupil size data was consistently the least valuable. When pupil size was removed from the exploratory and confirmatory timeline and image datasets classification accuracy remained stable (c.f., non-parcellated dataset). Furthermore, when pupil size data was classified on its own, classification was the lowest of all of the parcellations, and, in one instance, was no better than chance ().

The findings from the current study support the notion that black box CNNs are a viable solution to the inverse Yarbus problem. The inconsistent pattern of results for x- and y-coordinate parcellation comparisons are a confirmation of Lukander's () assertion that implementing black box solutions to the inverse Yarbus problem can lead to unreliable results. On the other hand, the consistent pattern of results for the non-parcellated timeline and image datasets suggest that black box solutions can be successfully replicated when the resolution of both datasets are equivalent. In reality, the purpose of decoding mental state from eye movement data is often to advance technology to improve educational outcomes , strengthen the independence of physically and mentally handicapped individuals , or improve human human-computer interfaces . To this end, the use of consistently effective and efficient black box solutions can be justified.

Moving forward, variations improvements to the image data may have the potential to advance the current state-of-the-art. If the goal is application, including stimulus feature

information may benefit classification. According to Bulling et al. (), incorporating stimulus feature information into the dataset may provide diagnostic information in addition to spatial location. According to Borji & Itti (), accounting for salient features in the the stimulus might overcome the classifier, leaving little room for the additional consideration of endogenous effects. In this case we know there is the potential for improved classification of the image dataset because the same algorithm consistently classified the timeline data more accurately. Thus, overlying scanpath plot images onto the trial stimulus could potentially improve the performance of the model used in this study.

Conclusion

This study was the first to provide a solution to the inverse Yarbus problem using (1) non-aggregated eye tracking data (x-coordinates, y-coordinates, pupil size), (2) timeline and image data formats (see Figure X), and (3) a "black box" CNN architecture. The CNN was able to decode the image and timeline data better than chance, although only the timeline datasets were rich enough to be decoded with state-of-the-art accuracy. Datasets that were not as rich were not able to differentiate the cognitive processes underlying the memorization task from those underlying the search and rate tasks. Decoding parcellations of the data revealed that pupil size was the least informative component of the eye movement data.

Given the questionable reliability and generalizability surrounding the "black box" nature of CNN classification, the models were first tested on an exploratory dataset, then confirmed using a second unrelated dataset. The findings appear stable and generalizable. Although the timeline data outperformed the image data format, future studies that incorporate stimulus features have the potential to provide a solution to the inverse Yarbus problem that surpasses the current state-of-the-art.