VGG-19 Displays Human-like Biases in Statistical Judgment from Visual Graphs

Ruiyi Ding (dingruiyi@outlook.com)
SILC Business School, Shanghai University, Shanghai, China

Yueyuan Zheng (yyzheng@ust.hk)
Division of Social Science, Hong Kong University of Science and Technology, Hong Kong SAR, China

Janet H. Hsiao (jhhsiao@ust.hk)
Division of Social Science and Department of Computer Science & Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, China

Lisheng He (hlisheng@shu.edu.cn)
SILC Business School, Shanghai University, Shanghai, China

Abstract

Convolutional neural networks (CNNs) not only recognize objects with high accuracy, but also acquire from images abstract statistical concepts such as numerosity and correlations. However, it remains unclear whether the CNN architectures implement inductive biases that mimic human biases in statistical judgments. In this paper, we examined whether VGG-19 models, a popular CNN architecture, that are trained to make correlation judgments from scatterplots display human-like biases. In comparisons between model predictions and human data, we found that there was a high correspondence between human biases and machine biases in VGG-19 models. Using explainable AI visualization with saliency maps to unpack the regions on which VGG-19 rely to make correlation judgments, we found that the late layers of the model tend to focus on regions similar to human participants' fixation distributions as captured by eye tracking. We further demonstrate that such models were nearly sufficient to predict human data at an accuracy level rivaling the state-of-the-art model trained on human data in three large-scale correlation discrimination datasets. Our results suggest that VGG-19 models may employ strategies that are similar to those used by human participants for statistical judgments from visual graphs and, therefore, pave the way to address human cognitive biases in visualization-based statistical judgments through the lens of deep neural networks.

Keywords: statistical judgments; cognitive biases; attention; CNN; explainable AI

Introduction

Data visualization is integral to judgment and decision-making in the modern world. It enables decision-makers to rapidly process complex information (Lewandowsky & Spence, 1989; Shao et al., 2022). At times, it also serves as a powerful communication tool, empowering organizations to make informed decisions in dynamic environments through intuitive data representation and analysis (Lurie & Mason, 2007; Vanderplas et al., 2020).

However, our cognitive processing of data visualization is not flawless. Prior work has uncovered various cognitive biases in visualization cognition. For example, consumers may exhibit uneven attention patterns and unsystematic thinking styles when interpreting product rating distributions (Fisher et al., 2018; Lu et al., 2022). Auditor assessments of company performance vary significantly based on how financial indicators are visualized (Beattie & Jones, 2002a, 2002b). Investors' interpretations of market trends can diverge dramatically depending on the visualization tools employed (Spiller et al., 2020). The challenge extends to experts, whereby even data scientists struggle to accurately assess the uncertainty when faced with suboptimal visualizations (Zhang et al., 2023). All in all, many biases observed in non-graphical information processing also manifest in visualization cognition (Dimara, et al., 2018).

Over the past few years, researchers have been making progress in addressing cognitive biases in visualization cognition through innovative applications of artificial intelligence, particularly convolutional neural networks (CNNs). For example, Yang, Ma et al. (2023) trained CNNs on human participants' correlation judgments from scatterplots. Their models not only predict human perceptual patterns but also identify potential areas of bias, paving the way for visualization systems that can actively compensate for known cognitive biases. This work and other similar studies (Jo & Seo, 2019; Ma et al., 2018; Wöhler et al., 2019) have expanded the application of CNNs to analyze various aspects of visualization cognition. Building on these foundations, researchers have developed practical AI-driven solutions that provide real-time feedback during the visualization design process. Bylinskii et al. (2017) showed how these advanced systems can analyze draft visualizations and suggest modifications to minimize potential perceptual biases. This development marks a significant evolution from merely understanding biases to actively preventing them during the design phase.

While these studies demonstrated the effectiveness of artificial intelligence for bias detection, they also revealed crucial gaps. First, while previous studies suggest that neural networks can predict humans' visualization cognition, it remains unclear whether their success in predictions is because of the training data or simply the neural network architectures. Put differently, whether neural networks such as CNNs have human-like cognitive biases in visualization cognition is an unaddressed problem. If the answer to this question is positive, cognitive scientists then can investigate visualization cognition through the lens of of deep neural networks. Second, while related domains have leveraged multimodal data (e.g. a combination of behavioral data and eye-tracking data) for explainable AI (e.g., Qi et al. 2024), the integration of AI and eye tracking data in visualization cognition studies is lacking. The current study tackles the two challenges in one shot. To make the challenges manageable, we focus exclusively on the cognitive biases in correlation

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

judgments from scatterplots, an extremely widely used visualization tool.

Current Study Overview

Scatterplots are one of the greatest inventions in the history of data visualization (Friendly & Denis, 2005). They display bivariate relationships in a way that decision makers can easily grasp. Previous work, however, suggests that humans' correlation judgments from scatterplots are not unbiased. Various features in scatterplots, independently of the presented correlation, may modify correlation judgments from scatterplots (Yang et al., 2018; Zhang et al. 2022). In this paper, we tested whether CNNs, without being trained on human data, display human-like biases in correlation judgments from scatterplots.

CNNs are one of the most famous neuro-inspired artificial neural networks (Hubel & Wiesel, 1959; Krizhevsky et al. 2012). A large body of research have shown that CNNs are capable of recognizing and distinguishing highly complex objects with high accuracy (Krizhevsky et al. 2012; Simonyan & Zisserman, 2014). Moreover, they were able to detect statistical concepts such as numerosity without being explicitly trained to do so (Nasr et al., 2019; Testolin et al. 2020). Previous studies have also shown that CNNs were able to predict human correlation judgments from scatterplots (e.g., Yang, Ma et al., 2023) However, they had to be trained on human data before making correlation predictions. Therefore, it is unclear whether their ability to make human-like predictions from data visualizations arises from the inductive biases intrinsic to the architecture of the neural networks.

In this paper we focused on the VGG-19 antialiased model as the signature model. Recent work by Yang, Ma et al. (2023) suggested that, when trained on human data, this model provides the most accurate predictions of human correlation judgments from scatterplots in an out-of-sample manner, among a wide array of deep convolutional neural networks (CNNs). The VGG-19 model, known for its deep convolutional neural network architecture with 19 layers, has been a staple in the field of image recognition since its introduction (Simonyan & Zisserman, 2014). The antialiased model refers to a technique used to reduce the jagged, pixelated edges in images, often applied to neural network inputs to improve the quality of data fed into the model which can enhance the performance of the neural network.

To directly test whether the VGG-19 architecture exhibits human-like inductive biases in statistical judgments from data visualization, we kept the networks blind to human data in the model training phase. Specifically, we trained the neural networks to make correlation judgments from scatterplots. The training labels were the actual Pearson's correlation coefficients (i.e., the ground truth) presented in the graph. To simplify the task, we restricted the presented correlations to the positive domain (i.e. with correlation coefficients between 0 and 1). After training the models on the task, we asked them to make correlation predictions from pre-selected experimental scatterplots, and compared their predictions with human participants' judgments.

We also attempted to interpret the predictive power of the neural networks through explainable AI (XAI). To this end, we generated saliency maps using GradCam (Simonyan et al., 2013), a popular XAI technique, for different layers of the neural networks, and compared them with the attention heatmaps using the eye tracking data of human participants doing the same correlation judgment task. The use of human participants' eye tracking data and the test of alignment between humans' attention patterns and the neural networks' saliency maps presents a more complete test of alignment between human intelligence and machine intelligence.

Finally, we validated the predictive power of such models (without being trained on human data) on three large-scale correlation discrimination data sets, made available by Yang, Ma et al. (2023). This work thus represents an extensive test of the idea of whether VGG-19, one of the most powerful CNN architectures in representing abstract statistical concepts, has human-like inductive biases in visualization cognition.

Methods

Model Training

Models Like most CNNs, VGG-19 was predominantly used for classification tasks. In line with this convention, we first defined a 10-class VGG-19 antialiased model, which has ten discrete outputs activated by a SoftMax function representing a classification of the image content into ten classes. To allow the model to make continuous correlation judgments, a multilayer perceptron with one hidden layer consisting of ten neurons was added to the model, connecting to the existing ten-class output of the VGG-19 antialiased model. The output will then be activated by a sigmoid function. The modification transforms the classification output into a continuous output in [0,1], which can be trained to predict Pearson correlations in the positive domain.

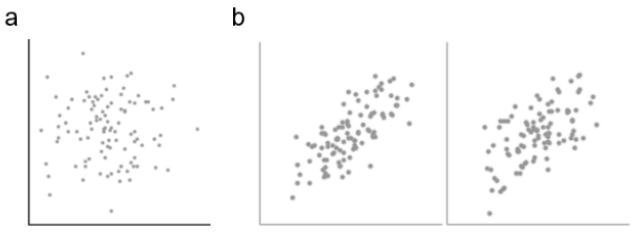

Figure 1: Example trials in the test sets. (a) An example scatterplot for the current model training. (b) An example trial in Yang, Ma et al.'s (2023) YMHTL-100 dataset.

Training Data The training data involved randomly generated scatterplots. Each scatterplot was plotted at the size of 150px×150px. In line with prior work (Yang, Ma et al. 2023), the data points were colored in opaque grey (#999999), against a white background (#FFFFFF). Each scatterplot consisted of 100 data points, with the mean set at 0 and the standard deviation set at 1 for both variables. The Pearson correlations between the two variables in the scatterplots were randomly selected from the uniform distribution between 0 and 0.99, using the

multivariate_random function in python. An example training image can be seen in Figure 1a.

We did a pilot model training exercise on VGG-19 to figure out what training size was required to train the model to make proper correlation predictions. As shown in Figure 2, VGG-19 can achieve desirable predictive accuracy on randomly generated new scatterplots when the training size was relatively small (e.g., 10,000). When the training sample size reached 20,000 or higher, the model's predictive performance appeared to be saturated. To test the robustness of the results, we designed two training sample sizes, 30,000 training samplings (namely 30K) and 40,000 training samplings (namely 40K), and ran each training size with 10 repetitions. All the results reported below are based on the average of the 10 repetitions.

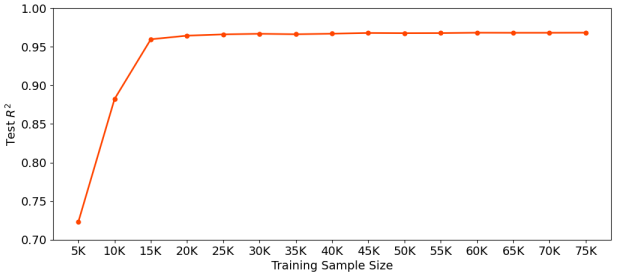

Figure 2: The VGG-19 model's test accuracy of ground-truth correlations (measured with out-of-sample R^2) with different training sizes.

Training Setup The models were trained using function torch.fit(), following Yang, Ma et al. (2023), with the epoch size set at 50. As they were regression models, mean-square error (MSE) was used as the loss function for the fit. We used Cuda Version 12.6, Python Version 3.8.10, Pytorch Version 2.3.1., Pytorch lightning Version 2.3.3, and trained all models on 40GB A100 GPUs.

Testing We used two series of human datasets to assess the emergence of human-like biases in correlation judgments in the VGG-19 antialiased models. The first series of datasets was from He et al. (2025), which involved two laboratory correlation judgment experiments with eye tracking. In each experiment, human participants were asked to make judgments about the correlations displayed in the scatterplots. Scatterplots were displayed one at a time (as in Figure 1a). Participants entered their judgments with a keyboard entry.

Experiment 1 involved 102 participants, each making correlation judgments from 40 randomly generated scatterplots. Experiment 2 involved 139 participants, each making correlation judgments from 48 different scatterplots. The Pearson's correlation coefficients in the scatterplots ranged from 0 to 0.9. In all experiments, human participants observed 1000px × 1000px images. To make the images compatible with the training samples detailed above, we rescaled the images to the size of 150px × 150px for the assessment with the VGG-19 antialiased model. In both experiments, all scatterplots were displayed in the same format as in Figure 1a. Thus, we have a total of 88 scatterplots with both human judgment data and the predictions by the trained VGG-19 models. Because of the similar experimental settings and the focus on group-level patterns, we pooled the two experiments in the analysis.

The availability of eye tracking data in He et al. (2025) made it possible to align human responses and machine responses beyond the behavioral data. Much prior work suggests that human visual attention captured by eye tracking serves as a promising benchmark for explainable artificial intelligence (XAI) (e.g., Mohseni et al., 2021). To illustrate, a convolutional neural network can generate salience maps for any input image, highlighting the regions in the image that are most important for the trained objectives (e.g., telling whether there is a dog in the image). Researchers have compared the sort of saliency maps with the attention heatmaps that they have obtained from human participants using eye tracking (e.g., Qi et al., 2024; Yang, Liu et al., 2023). In a similar vein, we attempted to assess whether the saliency maps generated by the VGG-19 antialiased models were well aligned to the visual attention maps generated by human participants when they were making correlation judgments from the scatterplots.

The second series of human datasets came from Yang, Ma et al. (2023). Their datasets involved three large-scale online behavioral experiments. Here, human participants were asked to make a binary choice between two scatterplots, choosing the one with the highest correlation coefficient (as in Figure 1b). Hence, this test assessed whether the emerging human-like biases trained in one response mode (i.e., separate evaluation) generalize to a different response mode (i.e., joint evaluation in a binary choice). In every choice, each scatterplot was at the size of 150 pixels × 150 pixels, the same as the image size of training data in the current study. The data points were colored with opaque grey (#999999) against a white background.

In each of Yang, Ma et al.'s (2023) experiments, 210 human participants were recruited from Prolific Academic, each making 96 binary choices. Therefore, each experiment involved 20160 binary choice data. The key differences between experiments were the distributional properties of the scatterplots. In Experiment 1, denoted by YMHTL-100, each scatterplot had 100 data points randomly generated using bivariate Gaussian distributions with a given correlation coefficient. In Experiment 2, denoted by YMHTL-200, each scatterplot had 200 data points randomly generated using bivariate Gaussian distributions with a given correlation coefficient. In Experiment 3, denoted by YMHTL-95+5, each scatterplot had 100 data points. The data points were initially randomly generated in the same way as in Experiment 1, followed by five data points being replaced by outliers that resided approximately 3.5 standard deviations away from the mean in one of the two dimensions.

Results

He et al.'s (2025) Judgment Datasets

Behavioral Data We ran a sanity check to see whether the models had been trained properly to predict the Pearson correlation of scatterplots. To this end, we used the trained VGG-19 models to predict the correlations presented in the 88 stimuli across the two experiments in He et al. (2025). We found that the model predicted the correlations with high accuracy, comparable to human participants' accuracy

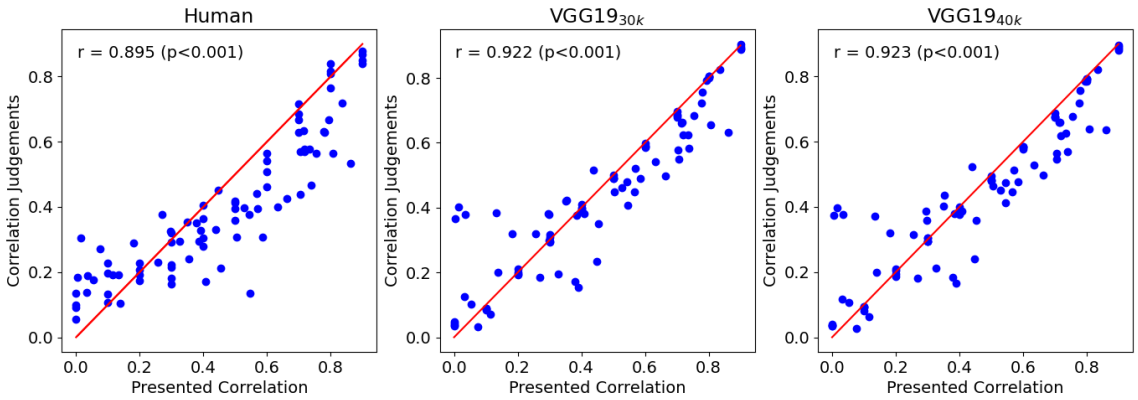

Figure 3: Correlation judgments from scatterplots by human participants and the trained VGG-19 models. The red line is the identity line (y=x). Each point represents a trial in the experiments.

level (see Figure 3). Therefore, the models were properly trained to make correlation judgments from scatterplots.

Like human participants, the trained VGG-19 models showed strong underestimation biases with presented correlations at intermediate-to-high levels (i.e. presented $R \geq 0.3$) (humans: $t_{60} = -9.40$, $p < .001$; VGG-19$_{30K}$: $t_{60} = -5.27$, $p < .001$; VGG-19$_{40K}$: $t_{60} = -5.92$, $p < .001$). In the meantime, human participants and the VGG-19 models displayed significant overestimation biases when the presented correlations were relatively low (i.e. presented $R < 0.3$) (humans: $t_{26} = 4.51$, $p < .001$; VGG-19$_{30K}$: $t_{26} = 2.85$, $p = .009$; VGG-19$_{40K}$: $t_{26} = 2.74$, $p = .011$).

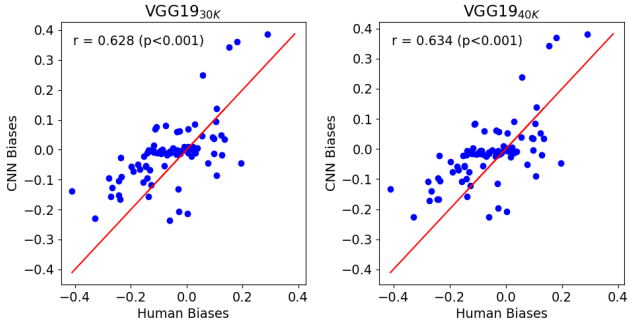

Figure 4: Human biases and CNN biases on the trial level. The red line is the identity line indicating y=x. Each point represents a trial in the experiments.

The alignment between human correlation judgments and the neural networks' correlation judgments at the behavioral level can be seen more directly in Figure 4, which shows that the human judgment biases (i.e., the difference between the judged correlations and the ground truth) over trials can be accurately predicted by the biases generated by the trained VGG-19 models (VGG-19$_{30K}$: $r = .628$, $p < .001$; VGG-19$_{40K}$: $r = .634$, $p < .001$), although these models were blind of any human data. Again, the high correlations in Figure 4 are driven by the underestimation biases in the intermediate-to-high levels and the overestimation bias in the low levels of presented correlations by both humans and CNNs.

Eye Tracking Data The above analyses using behavioral data suggest that the VGG-19 antialiased models were able to mimic human participants' behavioral patterns in visualization cognition, even though the models were not trained by any human data. Here we present a stronger test of the alignment between humans and neural networks in visualization-based statistical judgments. Specifically, we used the saliency map technique to identify the regions of the scatterplots that the neural networks focused on when making their predictions, and compared them with the regions that human participants focused on when making correlation judgments from scatterplots.

We did this for each of the 16 convolutional layers of VGG-19 and the results were summarized in Figure 5 (left panels). The saliency maps in the early and middle layers (until the first convolutional layer of Block 5) display low-to-intermediate levels of similarity with the attention heatmaps generated by human participants. However, we see a big leap in cosine similarity in the late layers (e.g., the final two convolutional layers), suggesting that the late layers approach to focusing on the regions that human participants attended to in the same task. Statistical analysis reveals that the cosine similarities for the last two convolutional layers were significantly higher than the baseline level (see Figure 5 caption for details).

To unpack what happened in the early and middle layers, we generated the reversed saliency maps in each layer by subtracting each pixel's saliency value from the maximum saliency value in that layer. We then calculated the cosine similarity with human participants' attention heatmaps in the same way as for the original saliency maps. Figure 5 (right panels) shows that the reversed saliency maps of middle layers (e.g., Block4_Conv1) have relatively high similarity with attention heatmaps. This can be seen in Figure 6. While the middle layers (e.g., Block4_Conv1) focused predominantly on the blank regions, the late layers (e.g., Block5_Conv3) switched attention to the point cloud regions in a way akin to humans' attention distributions. Collectively, these results suggest that, although the neural networks had not been trained on any human data, the regions that their late layers relied on to make correlation predictions largely overlapped with what human participants attended to when making correlation judgments from scatterplots.

Yang, Ma, et al.'s (2023) Choice Datasets

The alignment between humans' and neural networks' correlation judgments was further confirmed by three

large-scale correlation discrimination datasets made available by Yang, Ma et al.'s (2023). In Yang, Ma et al. (2023), human participants were presented with pairs of scatterplots and were asked to choose in each pair the scatterplot that presented the highest correlation coefficient (see Figure 1b). Our trained VGG-19 models can be conveniently adapted to such a task. We simply ask the models to make correlation predictions for each of the two scatterplots respectively, $Corr_{left}$ and $Corr_{right}$. The models' predicted choice for a pair would be:

$$Choice = \begin{cases} Right, & \text{if } Corr_{left} < Corr_{right} \\ Left, & \text{if } Corr_{left} > Corr_{right}. \end{cases}$$

The predictive accuracy of the trained VGG-19 models was obtained by comparing model predictions with human participants' choices. In the very rare cases in which $Corr_{left} = Corr_{right}$, the model's accuracy in predicting that pair was set at 0.5.

The models' predictive accuracy is summarized in Figure 7. Here, we compared our trained VGG-19 models (VGG-19_{30K} and VGG-19_{40K}) with two benchmarks from Yang, Ma et al. (2023). One was the baseline model that makes perfect correlation judgments (i.e., always choosing the scatterplot with the highest correlation coefficient). Comparisons between VGG-19_{30K} and the baseline model on predicting human choices suggest that the VGG-19_{30K} constantly outperformed the baseline model that makes perfect correlation judgments across the three datasets (YMHTL-100: $t = 9.89$, $p < .001$; YMHTL-200: $t = 5.78$, $p < .001$; YMHTL-95+5: $t = 3.84$, $p = .003$). Likewise, comparisons between VGG-19_{40K} and the baseline model suggests that the VGG-19_{40K} also always outperformed the baseline model (YMHTL-100: $t = 15.03$, $p < .001$; YMHTL-200: $t = 8.85$, $p < .001$; YMHTL-95+5: $t = 3.29$, $p = .009$). That means the models made biased correlation judgments in a way similar to what human participants did, even though they had not seen any human data. It is noteworthy that the models had not even seen scatterplots like those in YMHTL-200 and YMHTL-95+5, let alone human data on those scatterplots. Thus, this makes a case for a strong generalization test.

The other benchmark model was the VGG-19 antialiased architecture trained on human choice data by Yang, Ma et al. (2023). In the array of over 30 neural network architectures tested, this model emerged as the most successful model in predicting human participants' correlation discrimination tasks after being trained on human data. Therefore, this model serves as the state-of-the-art in predicting the choice data in Yang, Ma et al. (2023). For convenience, we call this model VGG-19_{Human} hereafter (i.e., the yellow bars in Figure 7).

Figure 5: Cosine similarities between the saliency maps (or their reverse) at different convolutional layers and the human participants' attention heatmaps. Error bars represent standard errors. The red horizontal lines represent the baseline cosine similarity (i.e., mean cosine similarities between human participants' attention heatmaps and randomly generated heatmaps). On each layer, we ran a paired t-test between the observed cosine similarities and the baseline cosine similarities (* < .05; ** < .01; *** < .001).

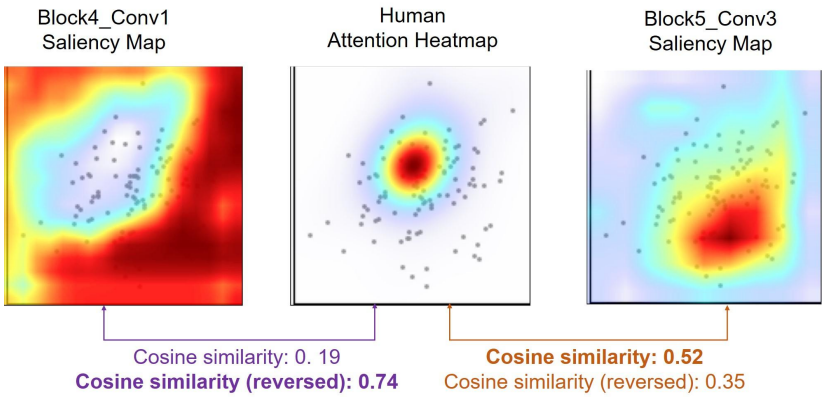

Figure 6: Human participants' attention heatmap and VGG-19$_{40K}$ saliency maps in an example trial in He et al. (2025) datasets.

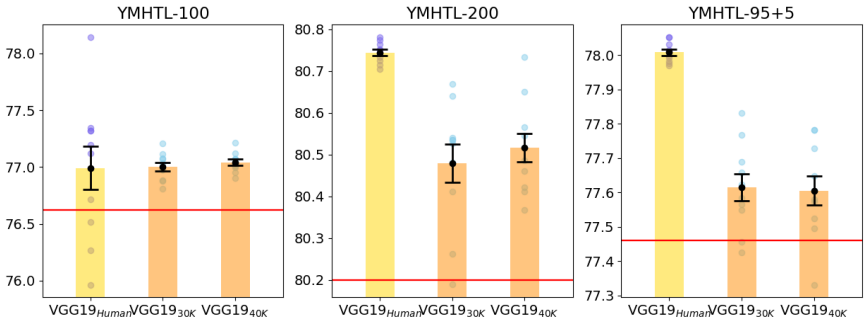

Figure 7: Models' predictive accuracy (%) of the three correlation discrimination datasets from Yang, Ma et al. (2023). The horizontal red lines represent the accuracy of the baseline model that always selected the scatterplot with the highest correlation. VGG19$_{Human}$ represents Yang, Ma, et al.'s (2023) model trained on human data. Error bars denote standard errors.

In YMHTL-100, which had the scatterplots most similar to the scatterplots used for our VGG-19 model training, we found that VGG-19$_{30K}$ and VGG-19$_{40K}$ (VGG-19$_{30K}$: t = -0.26, p = 0.80; VGG-19$_{30K}$: t = -0.06, p = 0.95), achieved an accuracy level that was comparable to the state-of-the-art, VGG-19$_{Human}$, even though the former were simply trained to make correlation judgments without introducing any human-like biases in the training process. In the remaining two datasets, YMHTL-200 and YMHTL-95+5, these two models lagged behind the state-of-the-art (ps < .001). That was likely because the scatterplots in the test sets were new to the two models. They still outperformed the baseline model, suggesting that they display human-like biases in the two datasets.

Discussion

Data visualization has been indispensable in many walks of life. While humans can grasp statistical information quickly from visualizations, those judgments can be systematically biased. Here we ask: Do artificial intelligence architectures such as convolutional neural networks (CNNs) have human-like inductive biases? Using correlation judgments from scatterplots as an example, we show that CNNs such as VGG-19, without being trained on human data, are able to generate human-like biases on correlation judgments from scatterplots. The behavioral findings were consolidated by the similarity between XAI saliency maps of CNNs and human participants' eye fixation heatmaps. Moreover, the generalizability of the finding was established using three additional large-scale correlation discrimination datasets. The alignment between neural networks and humans suggests that behavioral biases in human participants may arise in a way similar to that in neural networks, opening up the possibility of studying human statistical judgments from visualizations through the lens of neural networks.

That said, gaps between human responses and CNN predictions remain. For example, VGG-19 models did not perfectly predict human biases (see Figure 4). Saliency maps display slightly different attention distributions from human visual heatmaps (see Figure 6). Figure 7 also shows limitations in generalization abilities, where human data were still needed to enhance the models' ability to predict human decisions, especially when the stimuli were relatively new to the model. Future studies should investigate the gaps between VGG-19 predictions and human behaviors.

Finally, although we only tested the idea with VGG-19, we believe that human-like biases may emerge in other neural network architectures. We look forward to future work that extends the current work by testing additional neural networks (e.g, ViT, Dosovitskiy et al., 2020), on different visualizations (e.g., pie charts and histograms) and other judgment and decision tasks (Ciccione & Dehaene, 2021; Ciccione et al., 2023).

Acknowledgments

The authors acknowledge financial support from the National Natural Science Foundation of China (No. 72101156) and appreciate the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System (No. 19DZ2252600) for providing the computing resources and technical support.

References

Beattie, V., & Jones, M. J. (2002a). The impact of graph slope on rate of change judgments in corporate reports. *Abacus, 38*(2), 177-199.

Beattie, V., & Jones, M. J. (2002b). Measurement distortion of graphs in corporate reports: an experimental study. *Accounting, Auditing & Accountability Journal, 15*(4), 546-564.

Bobko, P., & Karren, R. (1979). The perception of Pearson product moment correlations from bivariate scatterplots. *Personnel Psychology, 32*(2), 313-325.

Bylinskii, Z., Kim, N. W., O'Donovan, P., Alsheikh, S., Madan, S., Pfister, H., Durand, F., Russell, B., & Hertzmann, A. (2017). Learning visual importance for graphic designs and data visualizations. In *Proceedings of the 30th Annual ACM symposium on user interface software and technology*, 57-69

Ciccione, L., & Dehaene, S. (2021). Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots. *Cognitive Psychology, 128*, 101406.

Ciccione, L., Dehaene, G., & Dehaene, S. (2023). Outlier detection and rejection in scatterplots: Do outliers influence intuitive statistical judgments? *Journal of Experimental Psychology: Human Perception and Performance, 49*(1), 129-144.

Cleveland, W. S., Diaconis, P., & McGill, R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science*, *216*(4550), 1138-1141.

Dimara, E., Franconeri, S., Plaisant, C., Bezerianos, A., & Dragicevic, P. (2018). A task-based taxonomy of cognitive biases for information visualization. *IEEE transactions on visualization and computer graphics*, *26*(2), 1413-1432.

Doherty, M. E., Anderson, R. B., Angott, A. M., & Klopfer, D. S. (2007). The perception of scatterplots. *Perception & Psychophysics*, *69*, 1261-1272.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., & Uszkoreit, J. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fisher, G. (2021). Intertemporal choices are causally influenced by fluctuations in visual attention. *Management Science, 67*(8), 4961-4981.

Friendly, M., & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences, 41*(2), 103-130.

He, L., Wang, H., Bian, Y., Zhang, X., & Bhatia, S. (2025) Information sampling and Bayesian belief formation in statistical judgments. *Under Review*.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology, 148*(3), 574-591.

Jo, J., & Seo, J. (2019). Disentangled representation of data distributions in scatterplots. 2019 IEEE Visualization Conference (VIS), Vancouver, Canada.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1-9.

Lewandowsky, S., & Spence, I. (1989). Discriminating strata in scatterplots. *Journal of the American Statistical Association, 84*(407), 682-688.

Lu, T., Yuan, M., Wang, C., & Zhang, X. (2022). Histogram distortion bias in consumer choices. *Management Science, 68*(12), 8963-8978.

Lurie, N. H., & Mason, C. H. (2007). Visual representation: Implications for decision making. *Journal of Marketing, 71*(1), 160-177.

Ma, Y., Tung, A. K., Wang, W., Gao, X., Pan, Z., & Chen, W. (2018). Scatternet: A deep subjective similarity model for visual analysis of scatterplots. *IEEE Transactions on Visualization and Computer Graphics, 26*(3), 1562-1576.

Mohseni, S., Block, J. E., & Ragan, E. (2021). Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In *26th international conference on intelligent user interfaces*, 22-31.

Nasr, K., Viswanathan, P., & Nieder, A. (2019). Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances, 5*(5), eaav7903.

Qi, R., Zheng, Y., Yang, Y., Cao, C. C., & Hsiao, J. H. (2024). Explanation strategies in humans versus current explainable artificial intelligence: Insights from image classification. *British Journal of Psychology*, 1-24.

Rensink, R. A., & Baldridge, G. (2010). The perception of correlation in scatterplots. *Computer Graphics Forum, 29*(3), 1203-1210.

Shao, C., Yang, Y., Juneja, S., & GSeetharam, T. (2022). IoT data visualization for business intelligence in corporate finance. *Information Processing & Management*, *59*(1), 102736.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Spiller, S. A., Reinholtz, N., & Maglio, S. J. (2020). Judgments based on stocks and flows: Different presentations of the same data can lead to opposing inferences. *Management Science*, *66*(5), 2213-2231.

Testolin, A., Zou, W. Y., & McClelland, J. L. (2020). Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Developmental Science, 23*(5), e12940.

Vanderplas, S., Cook, D., & Hofmann, H. (2020). Testing statistical charts: What makes a good graph?. *Annual Review of Statistics and Its Application, 7*(1), 61-88.

Wöhler, L., Zou, Y., Mühlhausen, M., Albuquerque, G., & Magnor, M. (2019). Learning a Perceptual Quality Metric for Correlation in Scatterplots. *VMV 2019-Vision, Modeling and Visualization*, 55-62.

Yang, F., Harrison, L. T., Rensink, R. A., Franconeri, S. L., & Chang, R. (2018). Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics, 25*(3), 1474-1488.

Yang, A., Liu, G., Chen, Y., Qi, R., Zhang, J., & Hsiao, J. (2023). Humans vs. AI in detecting vehicles and humans in driving scenarios. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45, No. 45), 1832-1839.

Yang, F., Ma, Y., Harrison, L., Tompkin, J., & Laidlaw, D. H. (2023, April). How Can Deep Neural Networks Aid Visualization Perception Research? Three Studies on Correlation Judgments in Scatterplots. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-17.

Zhang, S., Heck, P. R., Meyer, M. N., Chabris, C. F., Goldstein, D. G., & Hofman, J. M. (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences*, *120*(33), e2302491120.

Zhang, X., He, L., & Bhatia, S. (2022). Information sampling explains Bayesian learners' biases in correlation judgment. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44), 431-438.