Recurrent Convolutional Neural Network Predicting Early Visual Recognition on Dynamic Handwriting Images

Sungjae Cho (sungjae.cho@umontreal.ca), Eilif B. Muller (eilif.muller@umontreal.ca)
Université de Montréal, CHU Sainte-Justine Research Centre, Mila – Quebec AI Institute, Montréal, Québec, Canada

Abstract

The ability to accumulate evidence and make timely perceptual conclusions in a rapidly changing environment is important in many ecological contexts. Here we propose a recurrent convolutional neural network (RCNN) that can predict human decision times of early multi-class recognition on dynamic handwriting images. We adapted the original RCNN to perform multiple binary decisions and to have two accept and reject thresholds for each class, to model human uncertainty thresholds. With these modifications, our model achieves high classification accuracy while better predicting human decision times than the original RCNN and model lacking recurrence. Moreover, the uncertainty of our model aligns well with human perceptual ambiguities early in the stimulus sequences. Our modeling results thus support the notion that recurrence is an important component in perceptual decision-making models for dynamic visual stimuli.

Keywords: convolutional neural network; recurrent neural network; visual recognition; decision making; dynamical system

Introduction

The ability to accumulate evidence and make timely perceptual conclusions in a rapidly changing environment is important in many ecological contexts. Convolutional neural networks (CNNs) have emerged as standard models of the representational hierarchy (Yamins et al., 2014; Kriegeskorte, 2015). While CNNs process images in a feed-forward manner, visual cortical circuits are highly recurrent, which could play an important role in dynamic object recognition (Kietzmann et al., 2019). Recurrent CNNs (RCNNs) have been proposed that can predict reaction times and better account for dynamics in the visual hierarchy, but still for static images (Spoerer et al., 2019) or a sequence of different static scenes (Sörensen et al., 2023). Here we explore the application of RCNNs for dynamic visual stimuli consisting of naturalistic continuous movements. Specifically, we advance the RCNN to predict human decision times for early multi-class recognition of digits for dynamic handwriting images.

Data

Visual stimuli

We make use of an existing handwriting dataset (Cho & Kim, 2023) to produce visual stimuli. This dataset contains 10 digits from 0 to 9, for which handwriting has been temporally digitized using a stylus and tablet. Samples for each digit are written exclusively using one motor program except 9 includes two variants. There are 1000 samples for each digit, which are divided into training, validation, and test sets with proportions of 80%, 10%, and 10%, respectively. Images are temporally sampled at a rate of 10Hz. The first image is always blank, and no duplicate images appear in an image sequence. The initial stroke of digits starts from the image center to prevent recognition by the starting position. Finally, the maximum duration of the data is 1.1 seconds or 12 time steps. The dataset has two variants: For the Σ-dataset, each sample consists of a sequence of images depicting the tracing of a digit, whereas for the Δ-dataset each image depicts only the newly added pixels between successive frames in the Σ-dataset.

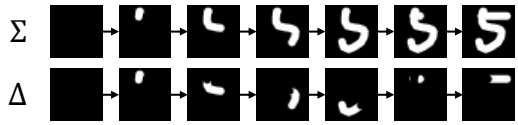

Figure 1: Different sequences of 128-by-128 images of the same handwriting of digit 5 in both the Σ- and Δ-datasets.

Human decision time

Decision time (DT) was collected for samples in the validation and test sets for a human participant. For each sample in the Σ-dataset, the participant viewed an image in the sample for as long as the participant wanted, and could then decide its final digit class or move on to the subsequent image in the sample. The participant was asked to decide the final digit class as early as possible with certainty. **Human DT** is defined as the time index corresponding to the image for which the participant chooses to assign the final digit class. If no decision has been made by the last image, DT becomes the last index of the sample because the final image is not ambiguous.

Model

Our model is based on the RCNN (Spoerer et al., 2019), for which we propose several modifications. Specifically, we reduced the number of filters at each RCNN layer by a factor of 8. We added one additional 256-unit dense layer before the readout layer for accumulating features, and the final sigmoid layer performs 10 binary decisions upon the accumulation. As such, the accumulation layer collects evidence (Gold & Shadlen, 2007), rather than the 10 decisions of the softmax inputs as in the original RCNN. Batch normalization was not included, as we found it resulted in instability. The Δ-dataset is provided as input to the model, as it has no mechanism to attend to new signals as humans can. The model is trained to minimize the binary cross-entropy (BCE) $L_b = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{10}[p_i(t)\ln\hat{p}_i(t) + (1-p_i(t))\ln(1-\hat{p}_i(t))]$, given time steps $T = 11$ except the first step, one-hot targets $(p_1(t),...,p_{10}(t))$ and predictions $(\hat{p}_1(t),...,\hat{p}_{10}(t))$. The loss

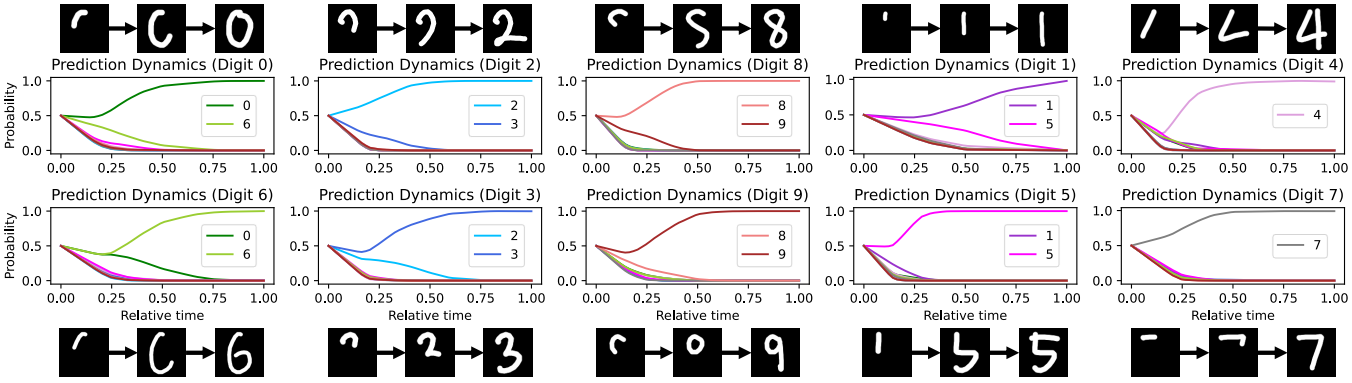

Figure 2: Our model's prediction dynamics for each digit for the test Δ-dataset. The model decreases the probabilities for less probable digits and maintains higher probabilities for more probable ones. Probabilities for digits with similar early shapes — (0,6), (2,3), (8,9), and (1,5) — remain higher. Predictions for the digit 7 diverge from the others due to its unique early shape.

Table 1: Comparison of the test performances of our model (Model 3) to those of other models under different settings.

Model	1	2	**3 (ours)**	4
Output	softmax	sigmoid	sigmoid	sigmoid
Loss	CCE	BCE	BCE	BCE
#Thresholds	1	1	20	20
Recurence	O	O	O	X
	Performance (mean ± STD; 30 seeds/model)			
DT MAE	.636 ±.058	.578 ±.045	**.493** ±.032	.668 ±.029
Accuracy	.950 ±.042	.957 ±.049	**.994** ±.003	.989 ±.006

of the original RCNN with the softmax output is the categorical cross-entropy (CCE): $L_c = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{10}[p_i(t)\ln\hat{p}_i(t)]$. For the first blank image at time $t=0$, the loss $L_{KL} = \frac{1}{2}\sum_{i=1}^{10}[\ln[\frac{1}{2}/\hat{p}_i(t)]+\ln[\frac{1}{2}/(1-\hat{p}_i(t))]]$ is added to either L_b or L_c to set the equal initial probabilities. The same optimization procedure as the original RCNN was applied with a batch size of 8 for 30 epochs. Model parameters are selected at the end of the epoch where the mean absolute error (MAE) between human and model DTs is the minimum for the validation set.

The original RCNN makes a decision if $-\sum_{i=1}^{10}\hat{p}_i\ln\hat{p}_i \le \theta_s$. Our model makes a decision if one of the 10 units satisfies $\ln\hat{p}_i \ge u_i$, and all the other 9 units satisfy $\ln(1-\hat{p}_{j\neq i}) \ge l_{j\neq i}$, reflecting that it should be certain that the digit is one unique class and not the other 9. There exist 20 thresholds of 10 upper and lower thresholds, u_i and l_i, respectively. If both models have not decided, the prediction at the terminal time step is read. **Model DT** is defined as the time step when the decision is made. The thresholds are computed using the following procedure. At the human DT of all validation samples, read out the prediction of the model, then find correct predicted samples X. Then, compute $\theta_s := \mathbb{E}_X[-\sum_{i=1}^{10}\hat{p}_i\ln\hat{p}_i]$. For all correct predicted samples X_i of each digit i, compute $u_i := \mathbb{E}_{X_i}[\ln\hat{p}_i]$ and $l_{j\neq i} := \mathbb{E}_{X_i}[\ln(1-\hat{p}_j)]$. These multiple thresholds are based on our hypothesis that a decision is made according to different levels of uncertainty with respect to each class and its rejection and acceptance.

Results and Conclusions

We find our model performs better at predicting DTs and classes than previous designs as shown in Tab. 1. The sigmoid outputs with the BCE loss (Model 2) predicted DT better than the softmax outputs with the CCE loss (Model 1), which the original RCNN uses. The softmax for Model 1 often resulted in drastic changes in predictions since exploding activity occurs to win against competing classes, and the CCE does not directly train output probabilities to decrease, whereas the BCE does. Decision-making with the 20 thresholds (Model 3) contributed to better prediction on DT than that with the single threshold (Model 2). Removing recurrence deteriorated DT prediction (Model 4). This supports the notion that recurrence is an important component to model the dynamics of biological vision (Kietzmann et al., 2019). Our model revealed that model DT had a linear relationship with human DT (Fig. 3a). Moreover, 95% of model DTs were within one time step of human DTs. (Fig. 3b). The uncertainty of our model in response to ambiguous early stimuli aligned well with human perceptual ambiguities early in the stimulus sequences (Fig. 2).

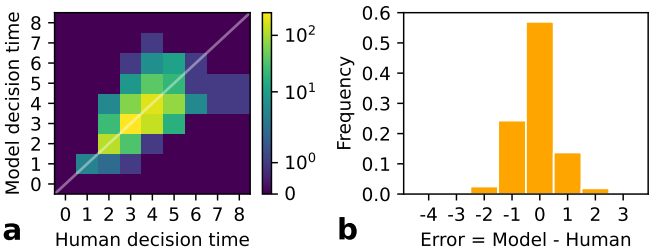

Figure 3: Human and model DTs for all correct test samples. (a) The distribution of human and model DTs on an exponential scale. (b) The distribution of errors. In 57% of model DTs, there is no error, and 95% have absolute errors less than 2.

Therefore, our model, which involves the sigmoid outputs, BCE loss, and two upper and lower thresholds for each class, predicts human DT better than the models with the previous strategies: the CCE loss and single thresholding. We also found that including recurrence in the model allowed a significant improvement in predicting human DT.

For future work, our model should be assessed for data that include more participants, who have different levels of certainty in decision-making. It could also be applied to more naturalistic tasks, such as activity recognition from videos.

Acknowledgement

This research is supported in part by the FRQNT Strategic Clusters Program (Centre UNIQUE - Centre de rercherche Neuro-IA du Québec) to S.C., an NSERC Discovery Grant to E.B.M., and the CHU Sainte-Justine Research Centre (CHUSJRC). Compute infrastructure was supported through a grant of computing time to E.B.M. from the Digital Research Alliance of Canada. E.B.M. was further supported the Fonds de Recherche du Québec–Santé (FRQS), the Canada CIFAR AI Chairs Program, the Quebec Artificial Intelligence Institute (Mila), and Google.

References

Cho, S., & Kim, T. (2023). Developing a neural network model generating handwriting motor sequences within realistic spatiotemporal scales and its biological implications. *2023 Conference on Cognitive Computational Neuroscience*.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual review of neuroscience*, *30*, 535-74.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 21854 - 21863.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modelling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Sörensen, L. K. A., Bohté, S. M., de Jong, D., Slagter, H. A., & Scholte, H. S. (2023). Mechanisms of human dynamic object recognition revealed by sequential deep neural networks. *PLOS Computational Biology*, *19*.

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2019). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, *16*.

Yamins, D., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*, 8619 - 8624.