An Attention-based Predictive Agent for Handwritten Numeral/Alphabet Recognition via Generation

NeurIPS 2023 Gaze Meets ML Workshop, New Orleans, LA. December 16, 2023.

Bonny Banerjee and **Murchana Baruah**

Institute for Intelligent Systems, and Dept. of Electrical & Computer Engineering
University of Memphis, Memphis, TN 38152, USA

bbnerjee@memphis.edu , murchanabaruah@gmail.com

INTRODUCTION

- A number of attention-based models for either classification or generation of handwritten numerals/alphabets have been reported in the literature.
  - However, generation and classification are done jointly in very few end-to-end models.

- We propose a predictive agent model that actively samples its visual environment via a sequence of glimpses.
  - The environment is an image of a handwritten numeral or alphabet.
  - The agent learns to classify handwritten numerals/alphabets from images by generating them.
  - The attention is driven by the agent's sensory prediction (or generation) error.

- This is the first known attention-based agent to interact with and learn end-to-end from images for recognition via generation, with high degree of accuracy and efficiency.

NOVELTY OF THIS WORK

1. The proposed model implements a perception-action loop to optimize an objective function.
   - *The action (attention) is modeled as proprioception in a multimodal setting* and is guided by perceptual prediction error, not by reinforcement.
   - No study has evaluated such a model in comparison to human efficiency.

2. At each sampling instant, the model simultaneously classifies and completes the partial sequence of observations.
   - Pattern completion allows prediction error computation which decides the next sampling location. Thus, attention emerges in our model and does not require learning feature weights.

NOVELTY OF THIS WORK

3. In the model, the pattern completion function maps the partial sequences of perceptual and proprioceptive observations to the class label and completed perceptual pattern.
   - Three variants of this function are proposed. Their accuracies correlate with the number of trainable parameters.

4. The model is more efficient than the human participants in a recently published study (Baruah et al. 2023b).
   - On average, the study participants required 4.2, 4.7 and 4.9 samples to recognize a numeral, uppercase and lowercase alphabet respectively. When exposed to the same stimuli and conditions as the participants, our model requires 2.0, 4.5, 4.2 samples respectively.
   - A highly-cited attention-based reinforcement model (Mnih et al. 2014) falls short of human performance.

M. Baruah, B. Banerjee, A. K. Nagar, and R. Marois. AttentionMNIST: A mouse-click attention tracking dataset for handwritten numeral and alphabet recognition. *Scientific Reports*, 13(1):3305, 2023b.

V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NeurIPS*, pages 2204–2212, 2014. (RAM model)

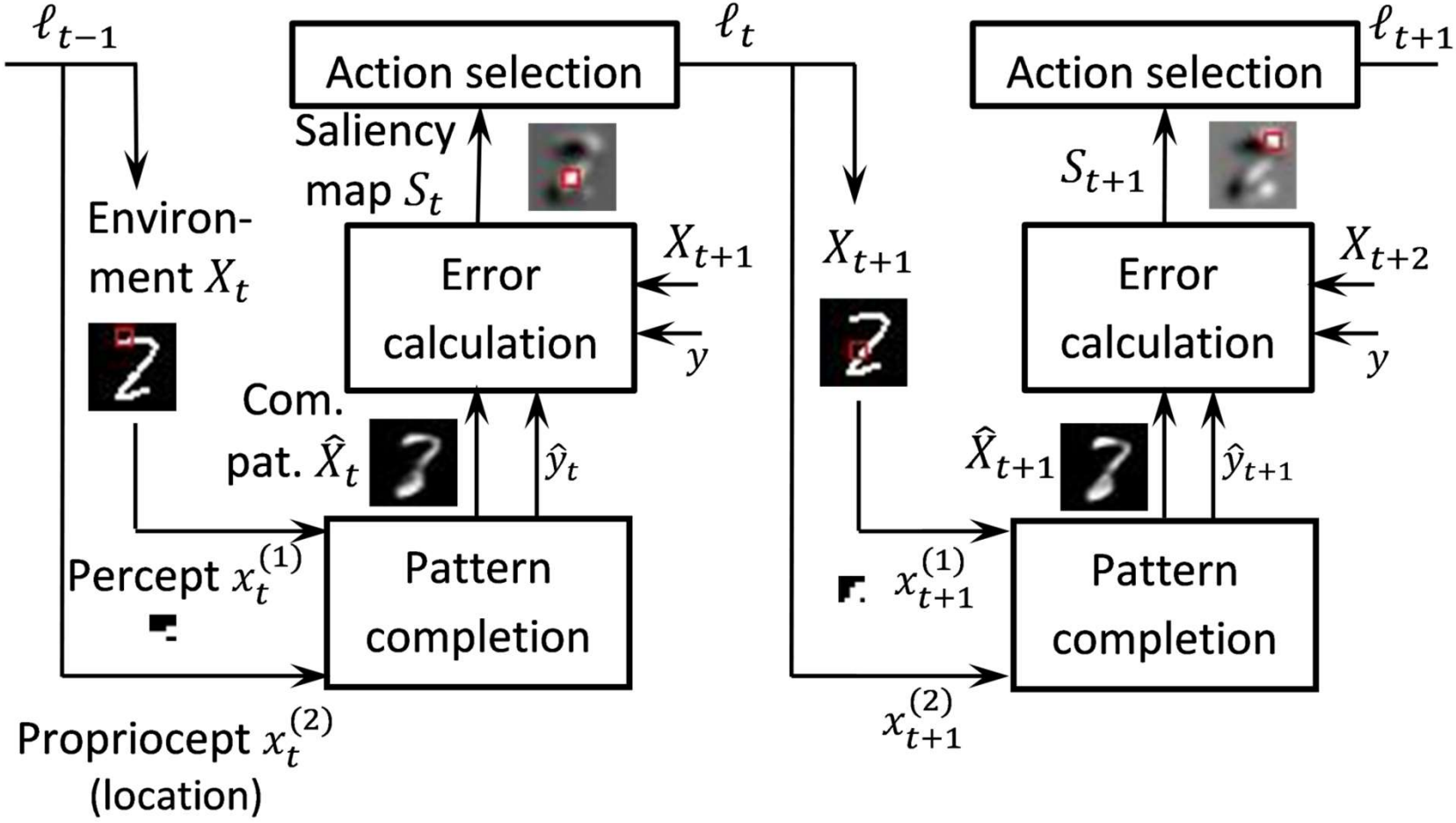
PROBLEM STATEMENT

Let an environment in m modalities be represented by a set of observable variables $\mathbf{X} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(m)}\}$. The variable representing the i-th modality is a sequence: $\mathbf{X}^{(i)} = \langle X_1^{(i)}, X_2^{(i)}, \dots, X_T^{(i)} \rangle$, where T is the sequence length. Let $\mathbf{x}_{\leq t} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}\}$ be a partial observation of $\mathbf{X}$ such that $\mathbf{x}^{(i)} = \langle x_1^{(i)}, x_2^{(i)}, \dots, x_t^{(i)} \rangle$, $1 \leq t \leq T$. Let y represent the class label.

We define *pattern completion and classification* as the problem of accurately generating $\mathbf{X}$ and y from the partial observation $\mathbf{x}_{\leq t}$. Given $\mathbf{x}_{\leq t}$ and a generative model p_θ with parameters θ and latent variables $z_{\leq t}$, the objective for pattern completion and classification at any time t is to maximize the joint log-likelihood of $\mathbf{X}$ and y, i.e.,

$$\underset{\theta}{\operatorname{argmax}} \int \log\big(p_\theta(\mathbf{X}, y | \mathbf{x}_{\leq t}, z_{\leq t}; \theta) p_\theta(z_{\leq t})\big)\, dz$$

PROPOSED AGENT MODEL

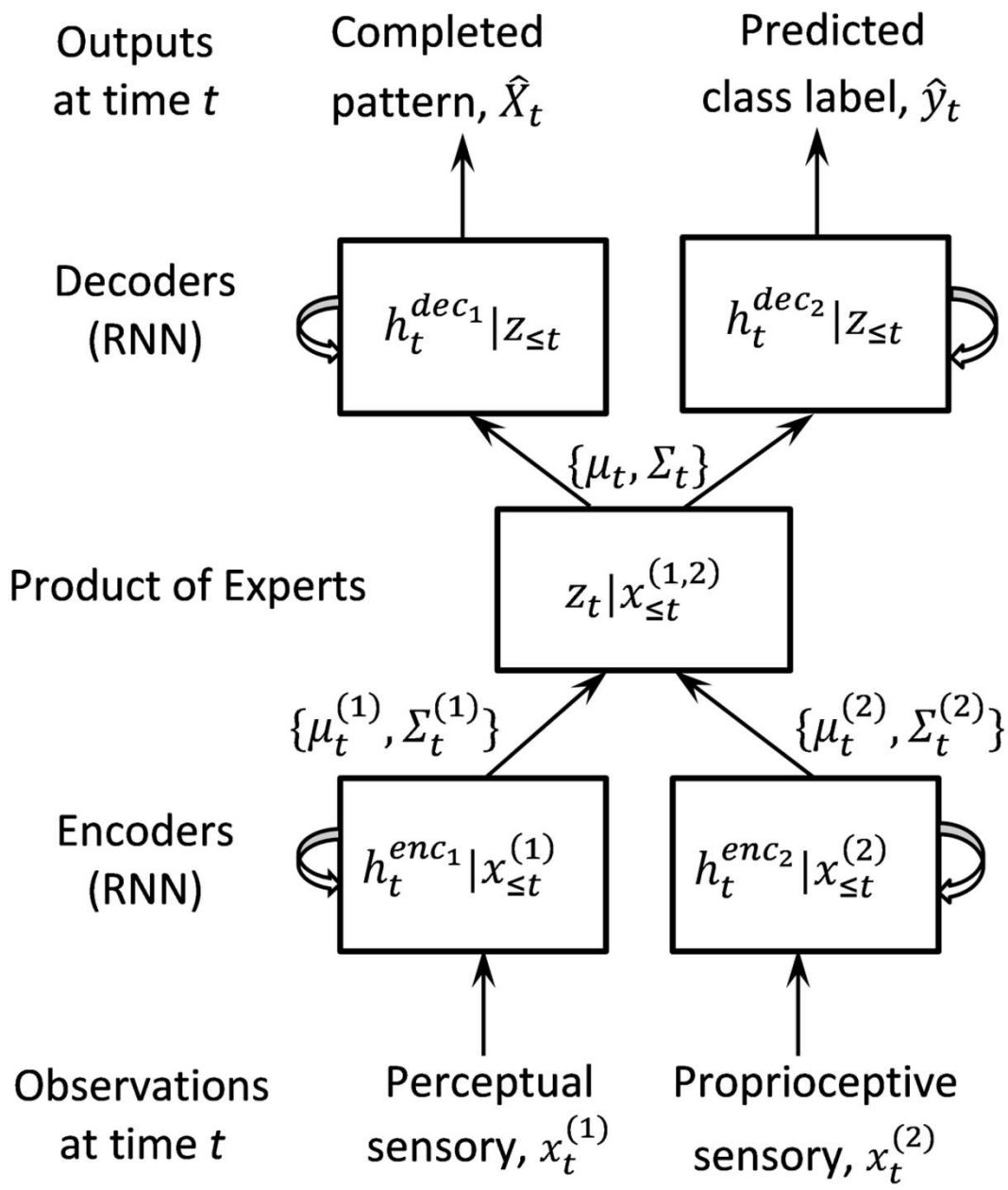


This model is learned end-to-end by maximizing the variational lower bound (ELBO) on the joint log-likelihood of the generated data.

Assumption: X_t and y_t are conditionally independent given the common latent variables and all observations till the current time t.

PATTERN COMPLETION: MODEL M1

The completed pattern and class label are generated from the latent variables.

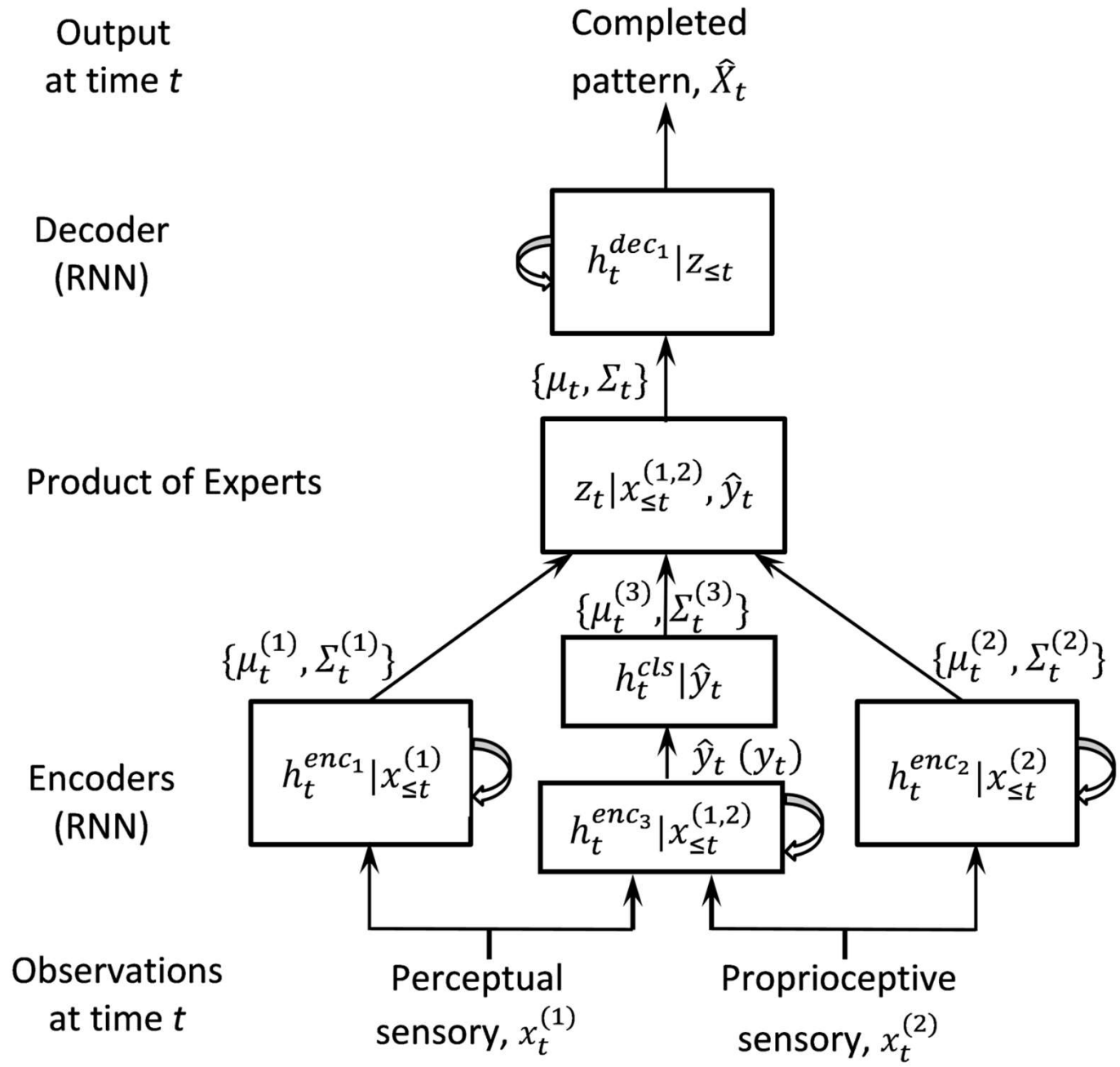


Slide prepared by Dr. Bonny Banerjee.

PATTERN COMPLETION: MODEL M2

The class label is inferred from the partial observation.

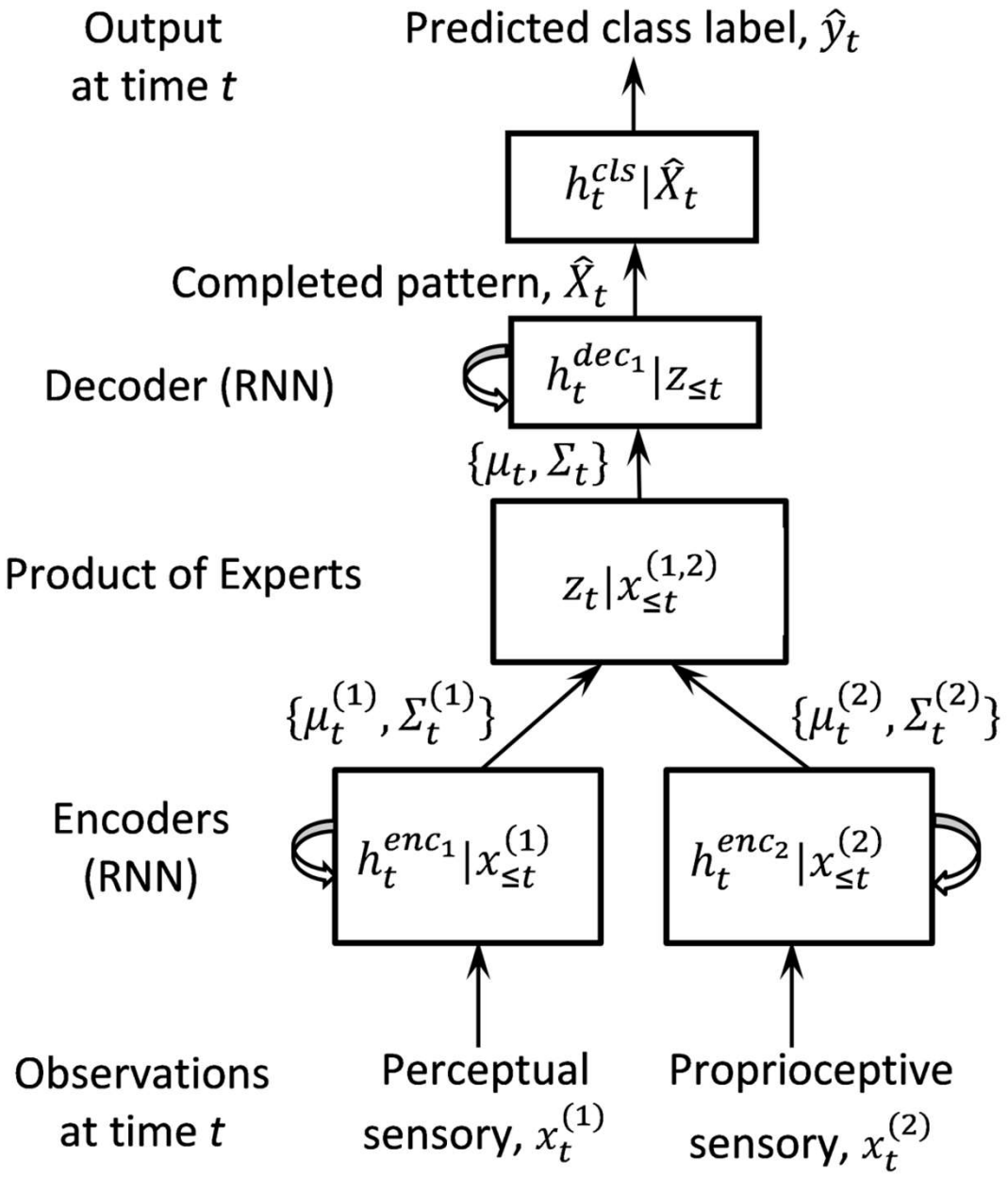
The latent variables are inferred from the class label and partial observation.



Slide prepared by Dr. Bonny Banerjee.

PATTERN COMPLETION: MODEL M3

The class label is inferred from the completed pattern which is generated from the latent variables.

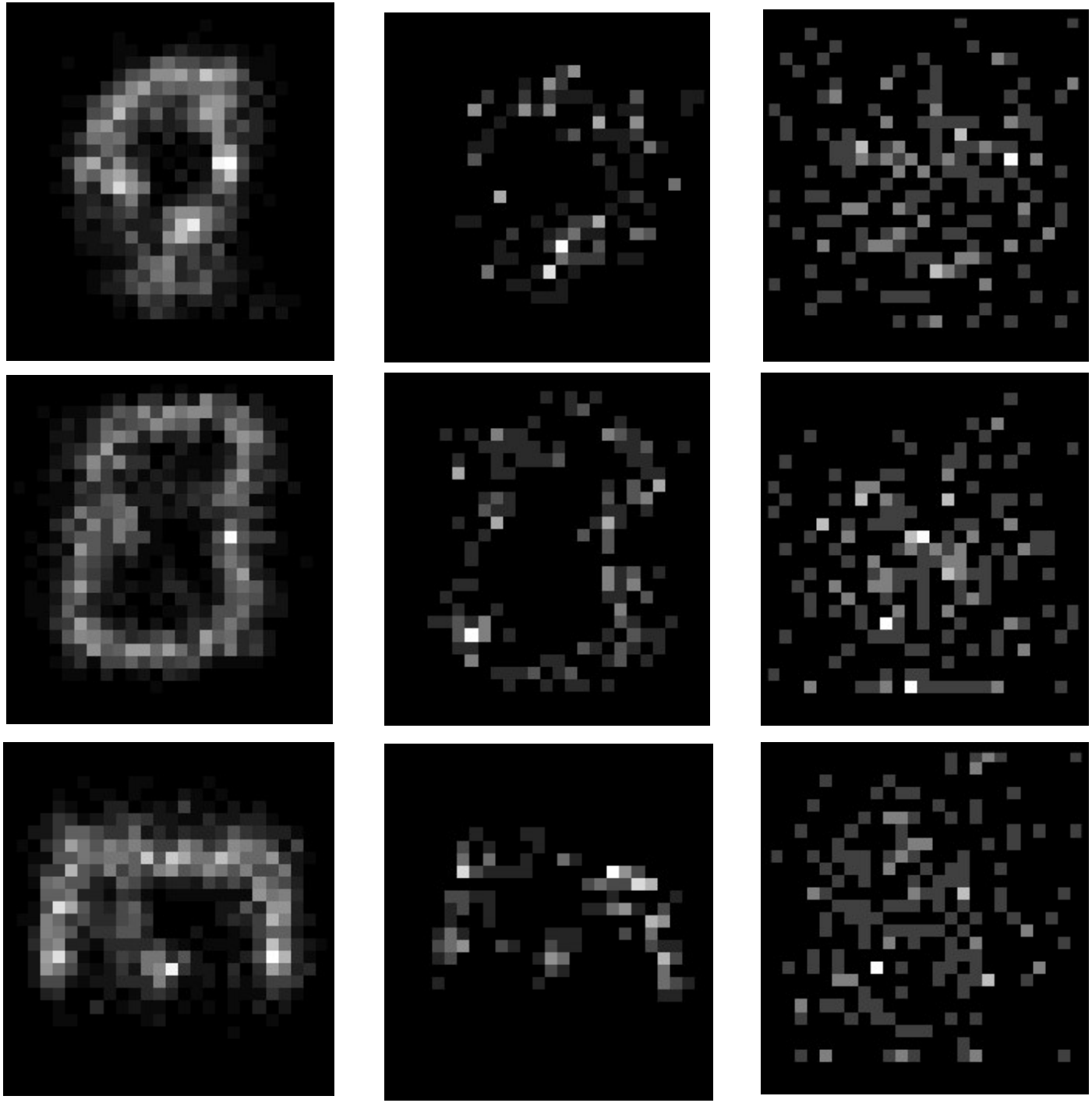


EXPERIMENTAL SETUP

- Datasets
  - **MNIST** (LeCun et al. 1998): Images from 10 numerals, 60000 training examples, 10000 test examples.
  - **EMNIST** (Cohen et al. 2017): Images from 26 alphabets (uppercase and lowercase), 124800 training examples, 20800 test examples
  - **AttentionMNIST** (Baruah et al. 2023b): Sequence of time-stamped samples from MNIST and EMNIST datasets are collected from participants using MTurk. Each sample consists of: (1) the location in the image selected by the participant, (2) the class(es) selected by the participant, and (3) the time taken by the participant to register the current sample. This data is recorded from 15 distinct stimuli from each class for MNIST, EMNIST uppercase, and EMNIST lowercase letters. The dataset is collected from 382 distinct participants. It consists of 1736 samples from MNIST, 4431 samples from EMNIST uppercase, and 4315 samples from EMNIST lowercase, and 169.1 responses per class on average..

- Hyperparameters are estimated via cross-validation using 10,000 images from the training set.

Figure 3: Comparison of the distribution of the sequence of fixations over a class for different cases; classes ‘9’, ‘B’, ‘m’ are shown in rows 1 to 3 respectively. The fixations are scattered in case of RAM, our model shows similar pattern with the participants data.



Participants | Our model M1 | RAM

Slide prepared by Dr. Bonny Banerjee.

EXPERIMENTAL RESULTS

Table 2: Evaluation of fixation maps from RAM and our model (Model 1) for the stimuli presented in the MTurk experiments, averaged over all classes and samplings. Standard deviations are included in parenthesis.

Metric	MNIST		EMNIST uppercase		EMNIST lowercase	
	Our model (M1)	RAM	Our model (M1)	RAM	Our model (M1)	RAM
KL	22.44(7.50)	22.50(7.48)	22.90(7.55)	22.96(7.24)	22.30(7.37)	22.23(7.16)
CC	0.02(0.01)	0.01(0.00)	0.02(0.01)	0.01(0.00)	0.02(0.01)	0.01(0.00)
SIM	0.18(0.11)	0.17(0.09)	0.16(0.10)	0.16(0.07)	0.18(0.10)	0.18(0.09)

Metrics (Bylinskii et al. 2018)**:**

- KL divergence (**KL**) between two image distributions (fixation maps). Lower KL indicates higher similarity.
- Pearson correlation coefficient (**CC**) evaluates the linear relationship between two fixation maps. Higher CC indicates higher similarity.
- Similarity (**SIM**) is another measure of similarity between two fixation maps. Higher SIM indicates higher similarity.

Conclusion: Between our model (M1) and RAM, the fixation maps generated by the former are more similar to those generated by the participants in (Baruah et al. 2023b).

EXPERIMENTAL RESULTS

Table 3: Classification accuracy and NLL on the test set reported after the final glimpse.

Dataset	Variants of the proposed model	Accuracy (%)	NLL ($\leq$)
MNIST	M1	96.3	76.5
	M2	92.3	107.0
	M3 (pretrained)	94.6	76.1
	M4 (not end-to-end)	82.9	76.1
EMNIST	M1	90.2	125.8
	M2	80.4	82.6
	M3 (pretrained)	88.5	78.9
	M4 (not end-to-end)	75.4	78.9

In model M4, the generative model is trained as in M3, and then an RNN with LSTM units is used to classify the data from the latent variables. M3 utilizes a CNN-based classifier.

Conclusion: Model M1 yields the highest classification accuracy followed by M3. However, M3 yields the best generation accuracy, and so does M4.

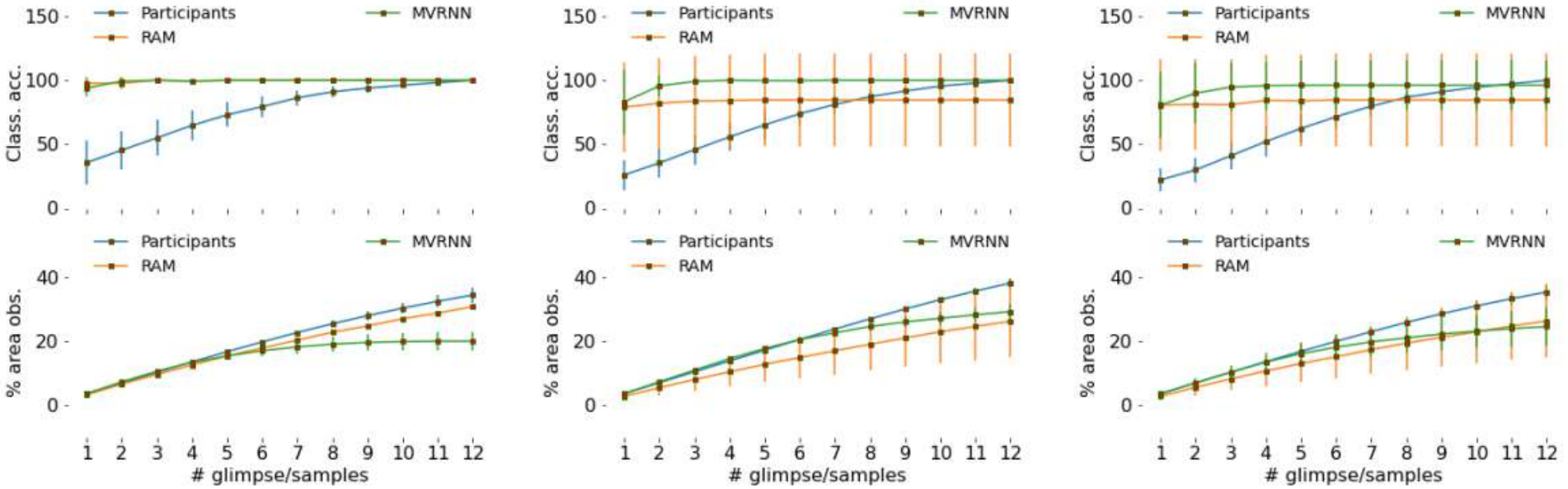
EXPERIMENTAL RESULTS

Table 4: Classification accuracy and NLL on the stimuli presented to the participants in (Baruah et al., 2023b), reported after the final glimpse.

Dataset	Variants of the proposed model	Accuracy (%)	NLL ($\leq$)
MNIST	M1	100	71.3
	M2	96	102.5
	M3 (pretrained)	98.7	71.8
	M4 (not end-to-end)	20.7	71.8
EMNIST upp.	M1	98.7	129.7
	M2	90.2	91.7
	M3 (pretrained)	98.7	83.9
	M4 (not end-to-end)	76.9	83.9
EMNIST low.	M1	95.6	111.0
	M2	85.4	66.8
	M3 (pretrained)	96.9	62.3
	M4 (not end-to-end)	74.9	62.3

Conclusion: Model M1 yields the highest classification accuracy followed by M3. However, M3 yields the best generation accuracy, and so does M4.

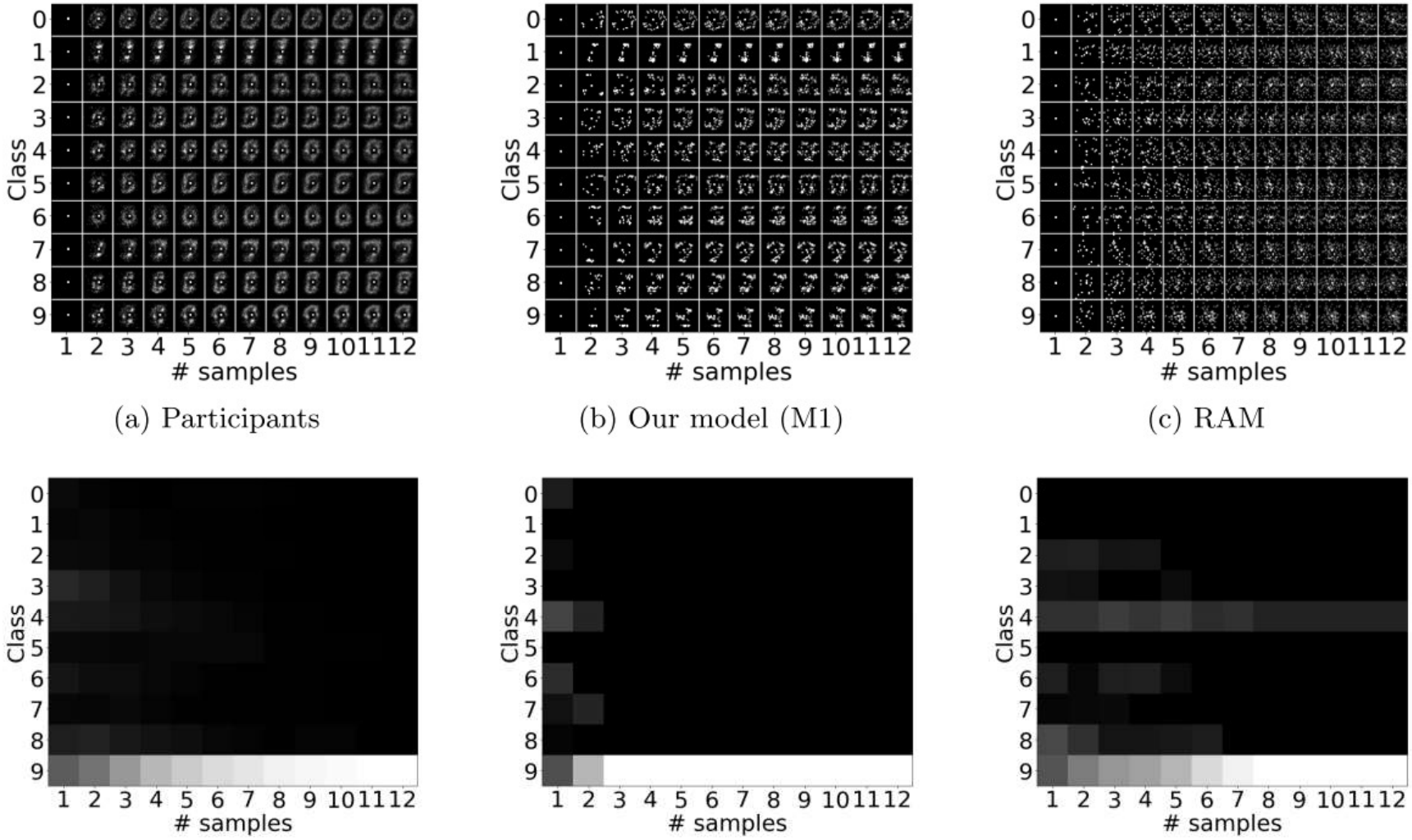
EXPERIMENTAL RESULTS



(a) MNIST (b) EMNIST uppercase (c) EMNIST lowercase

Figure 4: Errorbar plot showing the change in classification accuracy and percentage of image area observed by the participants in (Baruah et al., 2023b), RAM (Mnih et al., 2014) and our model (M1, MVRNN) with number of glimpses or samples.

Conclusion: In order to yield the same accuracy, our model (M1) requires fewer glimpses than RAM and the participants. Hence, our model is more efficient.



(a) Participants (b) Our model (M1) (c) RAM

(d) Participants (e) Our model (M1) (f) RAM

Figure 5: (a)–(c) Distribution of sampling locations (or fixation maps) for each numeral and each sampling instant. (d)–(f) Class distribution for class '9'. Qualitatively, the participants' fixation maps are more similar to our model's than RAM's. The distributions are obtained by averaging the responses over all stimuli presented from each class. Each row corresponds to a class, and each column corresponds to a sampling instant which increases from left to right. Also see Figs. A1 and A2 in Appendix B, which show similar results for uppercase and lowercase alphabets respectively.

CONCLUSIONS

- We proposed an attention-based agent model for handwritten numeral/alphabet recognition via a sequence of glimpses.
  - Three variants of this model are evaluated on benchmark datasets. Their accuracies are comparable and correlate with the model size.
  - Our experiments reveal that the proposed model is more data-efficient in handwritten numeral/alphabet recognition than human participants as well as a highly-cited attention-based reinforcement model, under the same conditions and stimuli.
  - Qualitatively, the participants' fixation maps are more similar to our model's fixation maps than the reinforcement model's.

- To the best of our knowledge, this is the first attention-based end-to-end agent of its kind for recognition via generation, with high degree of accuracy and efficiency.

Thank You!

For more information, please feel free to contact us:

Bonny Banerjee, bbnerjee@memphis.edu

Murchana Baruah, murchanabaruah@gmail.com