IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 12, 2022, accepted January 27, 2022. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.3149585

An Attention-Based Predictive Agent for Static and Dynamic Environments

MURCHANA BARUAH[1,2], BONNY BANERJEE[1,2], AND ATULYA K. NAGAR[3]
[1]Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, USA
[2]Department of Electrical and Computer Engineering, University of Memphis, Memphis, TN 38152, USA
[3]School of Mathematics, Computer Science and Engineering, Liverpool Hope University, Liverpool L16 9JD, U.K.

Corresponding author: Bonny Banerjee (bonnybanerjee@yahoo.com)

ABSTRACT Real-world applications of intelligent agents demand accuracy and efficiency, and seldom provide reinforcement signals. Currently, most agent models are reinforcement-based and concentrate exclusively on accuracy. We propose a general-purpose agent model consisting of proprioceptive and perceptual pathways. The agent actively samples its environment via a sequence of glimpses. It completes the partial propriocept and percept sequences observed till each sampling instant, and learns where and what to sample by minimizing prediction error, without reinforcement or supervision (class labels). The model is evaluated by exposing it to two kinds of stimuli: images of fully-formed handwritten numerals and alphabets, and videos of gradual formation of numerals. It yields state-of-the-art prediction accuracy upon sampling only 22.6% of the scene on average. The model saccades when exposed to images and tracks when exposed to videos. This is the first known attention-based agent to generate realistic handwriting with state-of-the-art accuracy and efficiency by interacting with and learning end-to-end from static and dynamic environments.

INDEX TERMS Agent, attention, handwriting generation, multimodal, perception, proprioception.

I. INTRODUCTION

Perception and action are inextricably tied together as, in the real world, efficiency is as important as accuracy. Nature has evolved the visual system such that, to minimize resources, it learns to selectively attend to a few locations that provide information for the task at hand. We propose a predictive agent model, which observes its visual environment via a sequence of glimpses. It predicts, learns and acts by minimizing sensory prediction error in a closed loop.

The agent is evaluated on handwriting generation. The model is exposed to images of fully-formed handwritten numerals and alphabets (MNIST, EMNIST datasets) and videos of gradual formation of numerals (SMNIST dataset). This allows evaluation of the agent in static (image) and dynamic (video) environments. In handwriting generation, the agent learns to sequentially sample its visual environment.

A. RELATED WORK

Attention-based models can be hard or soft [1], [2]. Hard-attention models make decisions by processing a part of the data, sampled via a sequence of glimpses. These models are reinforcement-based (e.g., [2], [3]), unsupervised (e.g., [4], [5]) or supervised (e.g., [6]). Soft-attention models process the entire data but weigh the features. Supervised (e.g., [7]) and unsupervised (e.g., [8]) variants of these models have been reported. We propose an unsupervised (no class label) hard-attention model.

A number of models have been proposed for handwriting generation, such as [4], [9]–[16]. Only one of them, DRAW [4], is an unsupervised hard-attention model. In DRAW, attention is explicitly learned. In our model, attention emerges as a consequence of minimizing the prediction error, similar to the model in [16]. However, our prediction error computation function is different from that in [16]. Our function selects the location with maximum information gain at each glimpse. Also, this model is supervised (uses class labels). This model and DRAW have reported results only on images while our model operates on images and videos. Though the role of attention is to foster efficiency, most works on attention-based models, including this model and DRAW, do not report on their efficiency. We evaluate the efficiency of our model with respect to its size (number of trainable parameters) and the number of glimpses,

The associate editor coordinating the review of this manuscript and approving it for publication was Zhouyang Ren.

or equivalently, fraction of scene, required for accurate prediction.

B. NOVELTIES OF OUR AGENT MODEL

(1) It implements the perception-action loop as the optimization of an objective function. Action/attention is modeled as proprioception in a multimodal setting, and is guided by the perceptual prediction error, not by reinforcement. **(2)** The same model can be used for static and dynamic environments. We show applications on image and video. Behaviorally, the agent saccades and tracks when exposed to images and videos respectively. **(3)** This end-to-end model is efficient in terms of size and number of glimpses required for accurate prediction. It learns by sampling locations with maximum information gain at each glimpse. Consequently, it yields state-of-the-art prediction accuracy upon sampling only 22.6% of the scene on average. By the fourth glimpse which corresponds to 11.2% of the scene, the prediction error drops by 60.4%. **(4)** It yields state-of-the-art accuracy in handwriting generation. In particular, it yields 4.9% lower error than the DRAW model on the binarized MNIST benchmark.

II. MODELS AND METHODS

A. PRELIMINARIES

1) AGENT

Anything that perceives from and acts upon its environment using sensors and actuators respectively is called an agent [17].

Perception is the mechanism of interpreting sensory signals from the external environment by an agent [18].

Proprioception is a form of perception in which the agent's environment is its own body. Internal perception of position, movement, and motion of body parts is due to proprioception [18].

2) GENERATIVE MODEL

Given a set of data points x, a generative model p_{model} with parameters θ maximizes the log-likelihood, $\mathcal{L}(x;\theta)$, of the data.

3) EVIDENCE LOWER BOUND (ELBO)

Let the data x be generated by a latent continuous random variable z. Then, computing the log-likelihood requires integrating the marginal likelihood, $\int p_{model}(x,z)dz$, which is intractable [19]. In variational inference, an approximation of the intractable posterior is optimized by defining an evidence lower bound (ELBO) on the log-likelihood, $\mathcal{L}(x;\theta) \leq \log p_{model}(x;\theta)$.

Variational autoencoder (VAE) is a multilayered generative model. It assumes an isotropic Gaussian prior, $p_\theta(z)$, and i.i.d. data samples. VAE maximizes the following ELBO [19]:

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\mathrm{KL}}[q_\phi(z|x), p_\theta(z)] \tag{1}$$

where $p_\theta(x|z)$ and $q_\phi(z|x)$ are generative and recognition models respectively, $\mathbb{E}$ denotes expectation, and D_{KL} denotes Kullback-Leibler divergence. The first and second terms capture accuracy and complexity respectively. The negative of this ELBO is also known as *variational free energy*, minimization of which has been hypothesized as a general principle guiding brain function [20].

Saliency lies in the eyes of an agent. Saliency of a location in an environment is a function of its neighborhood and an agent's internal model (see [21], [22]).

B. PROBLEM STATEMENT

Let an environment in m modalities be represented by a set of observable variables $X = \{X^{(1)}, X^{(2)}, \ldots, X^{(m)}\}$. The variable representing the i-th modality is a sequence: $X^{(i)} = \langle X_1^{(i)}, X_2^{(i)}, \ldots, X_T^{(i)} \rangle$, where T is the sequence length. Let $x_{\leq t} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ be a partial observation of X such that $x^{(i)} = \langle x_1^{(i)}, \ldots, x_t^{(i)} \rangle$, $1 \leq t \leq T$. As in [23], we define *pattern completion* as the problem of accurately generating X from its partial observation $x_{\leq t}$. Given $x_{\leq t}$ and a generative model p_θ with parameters θ and latent variables $z_{\leq t}$, the generative process of X is given as $p_\theta(X|x_{\leq t}) = \int p_\theta(X|x_{\leq t}, z_{\leq t}; \theta) p_\theta(z_{\leq t}) dz$. The objective for pattern completion at any time t is to maximize the log-likelihood of X, i.e. $\arg\max_\theta \int log(p_\theta(X|x_{\leq t}, z_{\leq t}; \theta) p_\theta(z_{\leq t}))dz$.

C. AGENT ARCHITECTURE

As shown in the block diagram in Fig. 1, environment, observation, pattern completion, action selection, and learning are the five components of the proposed agent architecture.

1) ENVIRONMENT

Two kinds of environment, or source of sensory data, are considered: static (images) and dynamic (videos).

2) OBSERVATION

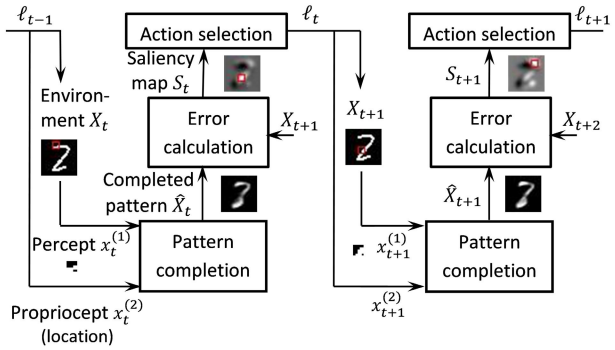
Our agent sequentially samples its environment in two modalities: visual perception and visual proprioception. The 2D coordinates of the fixation location in the environment constitutes the proprioceptive observation while the visual stimuli at that location constitutes the corresponding perceptual observation, as in [24]. See Table 1 for variable dimensions.

TABLE 1. Variable dimensions as used in this paper. Here $(.)^{(1)}$, $(.)^{(2)}$ refer to visual perception and visual proprioception respectively; T is maximum number of glimpses, t is glimpse index or time, $n \times n$ is patch size, $M \times M$ is image size.

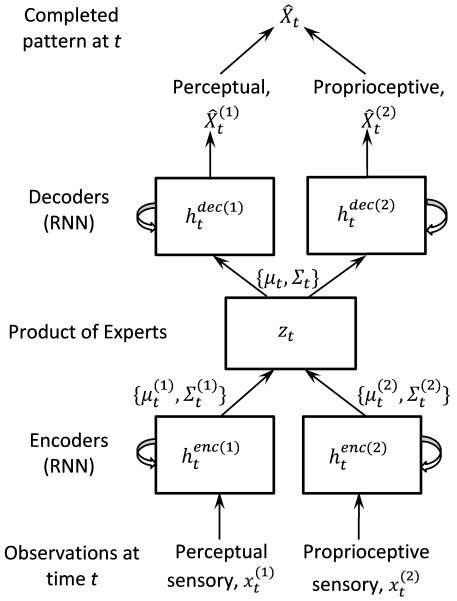
$x_t^{(1)}$	$x_t^{(2)}$	$X_t^{(1)}$	$X_t^{(2)}$	$S_t^{(1)}$
$\{0,1\}^{n \times n}$	$\mathbb{R}^2$	$\{0,1\}^{M \times M}$	$\mathbb{R}^{2 \times T}$	$\mathbb{R}^{M \times M}$

3) PATTERN COMPLETION

Patterns in the two modalities are completed using a multimodal variational recurrent neural network (VRNN). Recognition and generation are the two processes involved in the operation of a VRNN [25].



(a) Predictive agent architecture. The computation within the pattern completion block is shown below.



(b) Pattern completion model.

FIGURE 1. Different components of the proposed agent.

Recognition (Encoder). The recognition model, $q_\phi(z_t|x_{\leq t})$, is a probabilistic encoder [19]. It produces a Gaussian distribution over the possible values of the code z_t from which the given observations $x_{\leq t}$ could have been generated. Two RNNs, each with one layer of long short-term memory (LSTM) units, constitute the recognition model. Each RNN generates the parameters for the approximate posterior distribution for each modality. The parameters for all modalities are combined using product of experts (PoE) [26] to generate the joint distribution parameters for the approximate posterior $q_\phi(z_t|x_{\leq t})$. The prior can be sampled from a standard normal distribution $p_\theta(z_t) \sim \mathcal{N}(0, 1)$ as in [4]. The function of the encoder is shown in Lines 1–5 in Algorithm 2, where RNN_ϕ^{enc} represents the function of a LSTM unit, φ^{enc} is a function that returns the mean and the logarithm of the standard deviation as a linear function of the hidden state, as in [25].

Generation (Decoder). The generative model, $p_\theta(X_t|x_{<t}, z_{\leq t})$, generates the data from the latent variables, z_t, at each time step. The generative model has two RNNs with one layer of hidden LSTM units. Each RNN generates the parameters of the distribution of the sensory data for a modality. The sensory data is sampled from this distribution, which can be multivariate Gaussian or Bernoulli. In our model, $X_t^{(1)}$ is sampled from a multivariate Bernoulli distribution (as the perceptual observation is binary) with means generated by the perceptual decoder RNN, and $X^{(2)}$ is sampled from a multivariate Gaussian distribution (as the proprioceptive observation is real) with means and variances as output of the proprioceptive decoder RNN (see Fig. 1b). The pattern, $p_\theta(X|x_{<t}, z_{\leq t})$, is completed at every time step. In order to generate the perceptual data at any time step, the output from the perceptual RNN at the previous time step is added to the current perceptual RNN output before applying the sigmoid function, as in [4] (ref. Line 8 of Algorithm 2). The decoder equations are shown in Lines 7–11 of Algorithm 2, where RNN_θ^{dec} and φ^{dec} are same as RNN_ϕ^{enc} and φ^{enc}.

4) ACTION SELECTION

In our model, action selection is to decide the location in the environment to sample from. At any time t, a saliency map S_t is computed which assigns a salience score $S_t^{(\ell)}$ to each location ℓ:

$$S_t^{(\ell)} = D_{KL}(p(X_{t+1,\ell}^{(1)})||p_\theta(X_{t+1,\ell}^{(1)}|z_{\leq t}, x_{\leq t})) \tag{2}$$

where $p(X_{t+1,\ell}^{(1)})$ is the true data distribution at location ℓ and is sampled from a Bernoulli distribution. KL divergence, also known as *relative entropy*, is a measure of information gain achieved by using the true distribution, $p(X_{t+1,\ell}^{(1)})$, instead of the predicted distribution, $p_\theta(X_{t+1,\ell}^{(1)}|z_{\leq t}, x_{\leq t})$. Thus, the saliency map is a function of the prediction error. The most salient location is computed from this saliency map, which constitutes the sampling location.

The saliency map is smoothed using a Gaussian kernel $\mathcal{N}(., \sigma)$. The sampling location is chosen as:

$$\ell_t = \underset{\ell_t \in \{1,2,\ldots,M^2\}}{\text{argmax}} \ \text{conv}(\mathcal{N}(., \sigma), S_t) \tag{3}$$

where $\sigma = 2$. Each sample is a $n \times n$ patch centered at ℓ_t.

The salient location ℓ_t at any time t is the proprioceptive observation $x_{t+1}^{(2)}$ for time $t+1$. Therefore, the salient locations at $t = 1, 2, \ldots, T$ constitutes the proprioceptive pattern $X^{(2)}$. Hence, prediction error (saliency) guides the sampling of the scenes in our model. Unlike typical multimodal models, the two modalities in our model interact at the observation level as the perceptual prediction error provides the observation for the visual proprioceptive modality. The agent learns a policy to generate the proprioceptive pattern or the sequence of expected salient locations by minimizing the proprioceptive prediction error. This error, at any time, is a function of the difference between predicted fixation location from the learned policy and the most salient location in the scene. The most salient location is the location that yields the maximum

Algorithm 1 Learning the Proposed Network

1: Initialize parameters of the generative model θ, recognition model ϕ, sequence length T.
2: Initialize optimizer parameters: $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\eta = 0.001$, $\epsilon = 10^{-10}$.
3: Initialize $x_1^{(1)} \leftarrow F(X_1^{(1)}, \ell_0)$, $x_1^{(2)} \leftarrow g_3(\ell_0)$, where ℓ_0 is the initial sampling location (ref. Experimental setup in Section III), g_3 is an identity function (ref. Action selection in Section II-C), and the function F extracts a sample $x^{(1)}$ (e.g., 5×5 patch) from the environment $X^{(1)}$ (e.g., 28×28 image) at location ℓ (e.g., center of the image).
4: **while** true **do**
5: **for** $\tau \leftarrow 1$ *to* T **do**
6: $\hat{X}_\tau^{(1:2)} \leftarrow PatternCompletion(x_{1:\tau}^{(1:2)})$
Saliency Computation
7: $S_\tau \leftarrow g_1(X_{\tau+1}^{(1)}, \hat{X}_\tau^{(1)})$ [ref. Eq. 2]
8: $\ell_\tau \leftarrow g_2(S_\tau)$ [ref. Eq. 3]
9: $x_{\tau+1}^{(2)} \leftarrow g_3(\ell_\tau)$
10: $x_{\tau+1}^{(1)} \leftarrow F(X_{\tau+1}^{(1)}, \ell_\tau)$
Learning
11: Update $\{\theta, \phi\}$ by maximizing Eq. 4.
12: **end for**
13: **end while**

Algorithm 2 $PatternCompletion(x_{1:\tau}^{(1:2)})$

1: **Recognition Model**
2: **for** $i \leftarrow 1$ *to* 2 **do**
3: $h_\tau^{enc(i)} \leftarrow RNN_\phi^{enc}(x_{1:\tau}^{(i)}, h_{\tau-1}^{enc(i)})$
4: $[\mu_\tau^{(i)}; \sigma_\tau^{(i)}] \leftarrow \varphi^{enc}(h_\tau^{enc(i)})$
5: **end for**
Product of Experts
6: $z_\tau \sim \mathcal{N}(\mu_\tau, \Sigma_\tau)$, where
$$\Sigma_\tau \leftarrow \Big(\sum_{i=1}^{2} \Sigma_\tau^{(i)-2}\Big)^{-1}, \ \mu_\tau \leftarrow \Big(\sum_{i=1}^{2} \mu_\tau^{(i)} \Sigma_\tau^{(i)-2}\Big)\Sigma_\tau$$
Generative Model
For perceptual modality:
7: $h_\tau^{dec(1)} \leftarrow RNN_\theta^{dec}(z_\tau, h_{\tau-1}^{dec(1)})$
8: $\hat{X}_\tau^{(1)} \leftarrow f_\sigma(h_\tau^{dec(1)}, \hat{X}_{\tau-1}^{(1)})$
For proprioceptive modality:
9: $h_\tau^{dec(2)} \leftarrow RNN_\theta^{dec}(z_\tau, h_{\tau-1}^{dec(2)})$
10: $[\mu_{x^{(2)},\tau}^{(2)}; \sigma_{x^{(2)},\tau}^{(2)}] \leftarrow \varphi^{dec}(h_\tau^{dec(2)})$
11: $\hat{X}_\tau^{(2)} \leftarrow \mu_{x^{(2)},\tau}^{(2)}$

information gain in the environment. These are the locations where the agent's prediction error is the highest given all the past observations. The agent attends to these locations to update its internal model.

5) LEARNING

The recognition and generative model parameters are jointly learned by maximizing the ELBO for the multimodal variational RNN. This objective, obtained by modifying the objective for multimodal VAE [26] with variational RNN [25], is to maximize

$$\mathbb{E}_{q_\phi(z_{\le T}|x_{\le T})}\Big[\sum_{t=1}^{T}\sum_{i=1}^{2}\lambda_i \log p_\theta(X_t^{(i)}|z_{\le t}, x_{<t})\Big] - \sum_{t=1}^{T}\beta D_{KL}\big(q_\phi(z_t|x_{\le t}), p_\theta(z_t)\big) \quad (4)$$

where λ_1, λ_2, β are the weights balancing the terms. See Appendix for derivation.

We assume a one-to-one mapping between the agent's body and its environment, i.e. between the oculomotor muscles to the locations in the image/video frame. This assumption allows us to map from the perceptual space ℓ to the proprioceptive space $x^{(2)}$ using a simple function g_3 (ref. Line 9 in Algorithm 1).

III. EXPERIMENTAL RESULTS

Our model is implemented using TensorFlow 1.3 in Python 3.5.4. All experiments are carried out in HPC using PowerEdge R740 GPU nodes equipped with Tesla V100 PCIE 16GB.

A. DATASETS

Three datasets are used to evaluate our model:

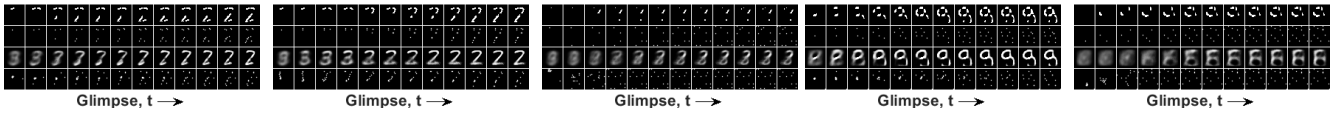
(1) MNIST [27] is a dataset of handwritten numerals $\{0, 1, \ldots 9\}$, consisting of 60,000 training and 10,000 test images (28×28 pixels).

(2) EMNIST [28] is a balanced dataset of handwritten English alphabets in uppercase and lowercase, consisting of 124,800 training and 20,800 test images (28×28 pixels).

(3) MNIST stroke sequence dataset (SMNIST) [29] was designed to learn sequences from MNIST images. It consists of a sequence of locations forming each MNIST numeral. We create a video for each image by selecting an equal number of more or less equidistant locations. Each frame is 28×28 pixels. Such videos show the gradual formation of numerals.

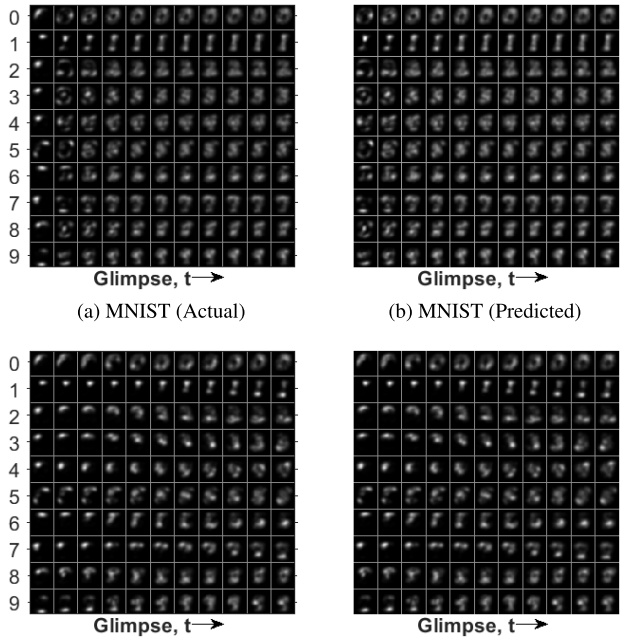
B. EXPERIMENTAL SETUP

For each modality, the generative and recognition models consist of 512 and 64 hidden units respectively. The latent variable dimension is 10. These parameters are estimated experimentally, as shown in Fig. 5. Maximum number of glimpses $T = 12$, and minibatch size is 100. The parameters β, λ_1, λ_2 are fixed to 1. The model is learned end-to-end using backpropagation and Adam optimization [30] with a learning rate of 0.001. These hyperparameters are estimated via cross-validation using 10,000 images or videos from the training set. The first observation is sampled from the starting pixel of the numeral in a SMNIST video as obtained from [29], and the center pixel of an image in MNIST and EMNIST. Fixing the first observation (or origin) on an object



(a) MNIST (b) SMNIST (c) MNIST-random (d) EMNIST (e) EMNIST-random

FIGURE 2. Pattern completion for (a) a random example ('2') from MNIST test set, (b) same example from SMNIST, (d) a random example ('g') from EMNIST; (c) and (e) correspond to examples (a) and (d) respectively when the observations are sampled randomly and not from the saliency map. Here patch size is 5 × 5. Each column in subfigures a–e corresponds to time or glimpse number. Rows 1, 2 show the perceptual and proprioceptive observation till the current glimpse in 28 × 28 space. Rows 3, 4 show the perceptual and proprioceptive pattern completion after each glimpse.



(a) MNIST (Actual) (b) MNIST (Predicted)

(c) SMNIST (Actual) (d) SMNIST (Predicted)

FIGURE 3. Distribution of salient locations for MNIST and SMNIST datasets, averaged over all examples of a class in the test set. Actual salient locations are obtained from the saliency map and predicted salient locations are the predictions for visual proprioception. Each row and column correspond to a class and a glimpse number respectively.

(egocentric reference) allows learning a position-invariant representation of the object.

The quality of generated images is measured using negative log-likelihood (NLL), as in [4]. Efficiency of the model is evaluated with respect to number of trainable parameters and number of glimpses required for accurate prediction.

C. EVALUATION FOR ACCURACY

At the initial time steps, the completed patterns are of poor quality (ref. Figs. 2a, b, d) as the agent samples from the latent distribution of multiple classes. Within a few glimpses, the predictions improve significantly. The examples in Figs. 2c, e show that when the agent samples the input space randomly, it may sample uninformative locations and will require more observations to determine the true class and generate the data accurately.

For static environment (image), the actual sequence of salient locations, $x^{(2)}_{1:T}$, and that predicted by our model, $\hat{X}^{(2)}$,

TABLE 2. Increase in number of trainable parameters with patch size in our model. Baseline patch size is 5 × 5 pixels.

Patch size	9 × 9	13 × 13	17 × 17	21 × 21
# additional parameters	57344	147453	270336	425984

TABLE 3. Comparison of generation accuracy at the final time step (T) between variants of the proposed model. Perceptual (Perc.) and proprioceptive (Prop.) modalities are shown separately for each dataset. Best results are highlighted.

Dataset	Variants of proposed model	Perc. NLL	Prop. NLL
MNIST	w/ prop.	**1125.3**	-761.6
	w/o prop.	1794.2	
EMNIST	w/ prop.	**1331.1**	-694.5
	w/o prop.	2332.0	
SMNIST	w/ prop.	**1439.25**	-745.11
	w/o prop.	2028.2	

are randomly distributed over the shape of an object (numeral or alphabet). Hence, with increase in number of glimpses, the distribution of salient locations for an object resembles its shape (ref. Figs. 3a, b). For dynamic environment (video), the sequence of salient locations, both actual and predicted, follow the motion. For example, in Figs. 3c, d, the salient location for '0' starts from top-left and ends at top after traversing in clockwise and anticlockwise directions as both formations of '0' are present in SMNIST. Thus, an interesting behavior emerges in our agent—it saccades while observing images and tracks the formation of objects while observing videos.

For both static and dynamic cases, the actual and predicted proprioceptive pattern distributions for each object class, obtained by averaging the actual and predicted salient locations from the test set, are quite similar. Thus in both cases, the distribution of salient locations is learned by the agent from its own behavior.

Our model's prediction accuracy, reported at the final time step as in [4], for MNIST is higher than the existing state-of-the-art (ref. Table 4). In this comparison, we have considered all recent attentional and non-attentional models that have reported prediction accuracy on the binarized version of MNIST dataset [31] in terms of NLL. Under similar conditions, the NLL for EMNIST is 76.6. NLL on the

TABLE 4. Prediction error (negative log-likelihood or NLL) comparison on binarized MNIST dataset [31]. Baseline refers to the case where the entire image is sampled by our model at any glimpse, i.e. it observes 100% of the ground truth.

Attention models		Non-attention models									
Ours	DRAW [4]	Baseline (Ours w/o attn.)	DARN 1hl [9]	Pixel CNN [10]	Info-VAE [11]	Row LSTM [10]	Diagonal BiLSTM [10]	BIVA [12]	NVAE [15]	Pixel-VAE++ [14]	MAE [13]
$\leq$ **76.99**	$\leq$80.97	$\leq$79.5	$\approx$84.13	81.3	$\leq$80.76	80.54	79.2	$\leq$78.59	$\leq$78.01	$\leq$78	$\leq$77.98

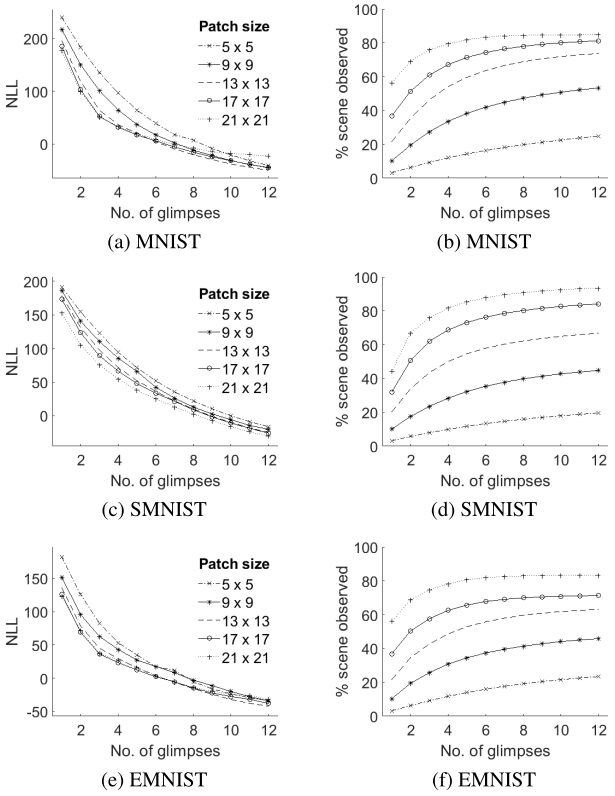


FIGURE 4. Comparison of prediction error (a, c, e) and efficiency (b, d, f) for different patch sizes.

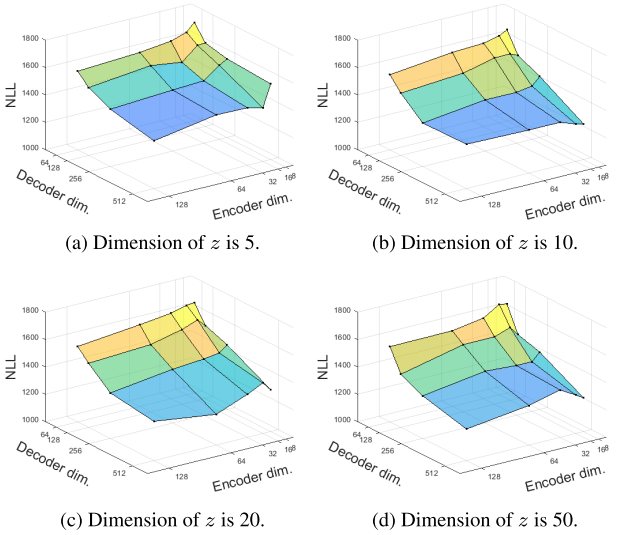


FIGURE 5. Prediction error (NLL) decreases up to a certain extent with increase in model size (i.e. encoder, decoder and latent variable (z) dimensions) for the MNIST dataset. In most cases, prediction does not improve beyond encoder dimension 64 and z dimension 10.

EMNIST dataset is being reported for the first time in this work.

The results in Table 4 are obtained with our model's encoder and latent variable dimensions as 128 and 20 respectively. It yields $(1 - 76.99/80.97) = 4.9\%$ lower error than the DRAW model. Our model (with attention) observes at most 23.4% of the ground truth (ref. Fig. 4b). When encoder and latent variable dimensions are 64 and 10 respectively, our model's NLL is $\leq$ 79.25. Comparison to [16] was not possible since they did not report prediction error or accuracy. Our NLL was lower when saliency was computed using KL divergence (ref. Eq. 2) as compared to rectification or Euclidean norm, as used in [16].

There are two key differences between the DRAW [4] and our model:

(1) At any instant, prediction error of our model drives its attention (sampling location). In DRAW, attention weights are learned explicitly which are not driven by the model's prediction error alone.

(2) Our model considers the patch and its location as separate modalities, perception and proprioception respectively, resulting in two input modalities which are combined. DRAW considers the patch and its location together as input in one modality.

Our improvement in generation accuracy (NLL) suggests that these differences are playing an important role.

D. ABLATION STUDY

Here we evaluate the contribution of proprioceptive modality in our model. We define a variant of our model by eliminating the proprioceptive modality at input (observation) and output (generation), keeping rest of the model unchanged. That is, $x_{<t} = \{x^{(1)}_{<t}\}$ and $i = 1$ in Eq. 4. For all datasets, the NLL is lower when the proprioceptive modality is used (ref. Table 3). Thus, the proprioceptive modality facilitates more accurate pattern completion.

Intuitively, our agent senses its body (via proprioception) in addition to sensing its environment (via perception). This allows it to learn the relations between its perceptual and proprioceptive signals, which is the key to its accuracy. In recent years, artificial intelligence and related areas have been flooded with attention-based models for numerous applications. Our work is unique as it models action/attention

as proprioception, similar to perception, and validates its role in attaining state-of-the-art accuracy.

E. EVALUATION FOR EFFICIENCY

The number of trainable parameters increases exponentially with patch size (ref. Table 2). The results in Table 4 are obtained with 5×5 patch size which allows our model to yield high accuracy while being efficient in model size.

Our experiments show that prediction accuracy (NLL) improves exponentially with increase in number of glimpses and our model starts yielding high accuracy within a few glimpses (ref. Figs. 4a, 4c, 4e). On average, by the fourth glimpse which corresponds to 11.2% (std. 1.1) of the scene, the prediction error drops by 60.4% (std. 10.1).

Figs. 4b, 4d, 4f show that for 21×21 patch size, more than 80% of the scene is viewed within the sixth glimpse for all three datasets (MNIST, EMNIST, SMNIST). In contrast, for 5×5 patch size, less than 25% of the scene is viewed till the last (12^{th}) glimpse for all the datasets. However, the prediction accuracy for 5×5 is only slightly lower than that for 21×21 (ref. Figs. 4a, 4c, 4e). This is because our model learns by sampling locations with maximum information gain at each glimpse. It yields state-of-the-art accuracy upon viewing only 22.6% of the scene on average over all three datasets (std. 2.7).

The proposed model yields state-of-the-art accuracy while being size and sample efficient. However, there is still room for improving its accuracy and efficiency. Our future work will include applying this model to other kinds of data, and learning this model using class labels in addition to perceptual and proprioceptive inputs.

IV. CONCLUSION

A predictive agent model is proposed that sequentially samples and interacts with its environment. At each instant, it samples the location with maximum information gain to minimize its sensory prediction error in a greedy manner. The agent operates as a closed-loop system involving perceptual ('what') and proprioceptive ('where') pathways which are learned end-to-end, without supervision (class labels) or reinforcement. The same model can be used for static and dynamic environments. Experiments on handwriting generation reveal that the model is sample and size efficient, and yields state-of-the-art accuracy. Conceptually, this work is unique due to its modeling action/attention as proprioception, using it with perception in a multimodal setting, and experimentally validating its role in yielding state-of-the-art accuracy in an end-to-end model.

APPENDIX LOSS FUNCTION DERIVATION

Here we derive the objective function in Eq. 4. The generative and recognition models are factorized as:

$$p_\theta(X_{\leq T}, z_{\leq T}|x_{\leq T}) = \prod_{t=1}^{T} p_\theta(X_t|z_{\leq t}, x_{<t}) p_\theta(z_t)$$

$$q_\phi(z_{\leq T}|x_{\leq T}) = \prod_{t=1}^{T} q_\phi(z_t|x_{\leq t})$$

The variational lower bound (ELBO) on the log-likelihood of the generated data, $\log p_\theta(X_{\leq T}|x_{\leq T})$, is derived as:

$$\begin{aligned}
&\mathbb{E}_{q_\phi(z_{\leq T}|x_{\leq T})}\Big[\log p_\theta(X_{\leq T}|x_{\leq T}) \frac{q_\phi(z_{\leq T}|x_{\leq T})}{q_\phi(z_{\leq T}|x_{\leq T})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|x_{\leq T})}\Big[\log \frac{p_\theta(X_{\leq T}, z_{\leq T}|x_{\leq T})}{p_\theta(z_{\leq T}|x_{\leq T})} \frac{q_\phi(z_{\leq T}|x_{\leq T})}{q_\phi(z_{\leq T}|x_{\leq T})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|x_{\leq T})}\Big[\sum_{t=1}^{T} \log \frac{p_\theta(X_t|z_{\leq t}, x_{<t}) p_\theta(z_t)}{p_\theta(z_t|x_{\leq t})} \frac{q_\phi(z_t|x_{\leq t})}{q_\phi(z_t|x_{\leq t})}\Big] \\
&= \mathbb{E}_{q_\phi(z_{\leq T}|x_{\leq T})}\Big[\sum_{t=1}^{T} \Big[\log p_\theta(X_t|z_{\leq t}, x_{<t}) \\
&\quad - \log \frac{q_\phi(z_t|x_{\leq t})}{p_\theta(z_t)} + \log \frac{q_\phi(z_t|x_{\leq t})}{p_\theta(z_t|x_{\leq t})}\Big]\Big] \\
&\geq \mathbb{E}_{q_\phi(z_{\leq T}|x_{\leq T})}\Big[\sum_{t=1}^{T} \log p_\theta(X_t|z_{\leq t}, x_{<t})\Big] \\
&\quad - \sum_{t=1}^{T} D_{KL}\big(q_\phi(z_t|x_{\leq t}), p_\theta(z_t)\big)
\end{aligned}$$

We assume, the modalities $X_t^{(1)}$ and $X_t^{(2)}$ are conditionally independent given the common latent variables [26] and all observations till the current time. Therefore,

$$\log p_\theta(X_t|z_{\leq t}, x_{\leq t}) = \sum_{i=1}^{2} \log p_\theta(X_t^{(i)}|z_{\leq t}, x_{<t})$$

Thus,

$$\begin{aligned}
&\log p_\theta(X_{\leq T}|x_{\leq T}) \\
&\quad \geq \mathbb{E}_{q_\phi(z_{\leq T}|x_{\leq T})}\Big[\sum_{t=1}^{T}\sum_{i=1}^{2} \log p_\theta(X_t^{(i)}|z_{\leq t}, x_{<t})\Big] \\
&\qquad - \sum_{t=1}^{T} D_{KL}\big(q_\phi(z_t|x_{\leq t}), p_\theta(z_t)\big) \\
&\quad \geq \mathbb{E}_{q_\phi(z_{\leq T}|x_{\leq T})}\Big[\sum_{t=1}^{T}\sum_{i=1}^{2} \lambda_i \log p_\theta(X_t^{(i)}|z_{\leq t}, x_{<t})\Big] \\
&\qquad - \sum_{t=1}^{T} \beta D_{KL}\big(q_\phi(z_t|x_{\leq t}), p_\theta(z_t)\big)
\end{aligned}$$

where λ_1, λ_2, β are the weights balancing the terms.

REFERENCES

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.

[2] G. Elsayed, S. Kornblith, and Q. V. Le, "Saccader: Improving accuracy of hard attention models for vision," in *Proc. NIPS*, 2019, pp. 702–714.

[3] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2204–2212.

[4] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," 2015, *arXiv:1502.04623*.

[5] S. M. A. Eslami, N. Heess, T. Weber, Y. Tassa, D. Szepesvari, K. Kavukcuoglu, and G. E. Hinton, "Attend, infer, repeat: Fast scene understanding with generative models," 2016, *arXiv:1603.08575*.

[6] Y. Zheng, R. S. Zemel, Y.-J. Zhang, and H. Larochelle, "A neural autoregressive approach to attention-based recognition," *Int. J. Comput. Vis.*, vol. 113, no. 1, pp. 67–79, May 2015.
[7] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10705–10714.
[8] H.-F. Sang, Z.-Z. Chen, and D.-K. He, "Human motion prediction based on attention mechanism," *Multimedia Tools Appl.*, vol. 79, nos. 9–10, pp. 5529–5544, Mar. 2020.
[9] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, and D. Wierstra, "Deep autoregressive networks," in *Proc. ICML*, 2014, pp. 1242–1250.
[10] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," 2016, *arXiv:1601.06759*.
[11] S. Zhao, J. Song, and S. Ermon, "InfoVAE: Information maximizing variational autoencoders," 2017, *arXiv:1706.02262*.
[12] L. Maaloe, M. Fraccaro, V. Liévin, and O. Winther, "Biva: A very deep hierarchy of latent variables for generative modeling," in *Proc. NIPS*, 2019, pp. 6551–6562.
[13] X. Ma, C. Zhou, and E. Hovy, "MAE: Mutual posterior-divergence regularization for variational AutoEncoders," 2019, *arXiv:1901.01498*.
[14] H. Sadeghi, E. Andriyash, W. Vinci, L. Buffoni, and M. H. Amin, "PixelVAE++: Improved PixelVAE with discrete prior," 2019, *arXiv:1908.09948*.
[15] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," 2020, *arXiv:2007.03898*.
[16] K. Standvoss, S. C. Quax, and M. A. Van Gerven, "Visual attention through uncertainty minimization in recurrent generative models," *BioRxiv*, Feb. 2020. [Online]. Available: https://www.biorxiv.org/content/biorxiv/early/2020/02/14/2020.02.14.948992.full.pdf
[17] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2020.
[18] J. Han, G. Waddington, R. Adams, J. Anson, and Y. Liu, "Assessing proprioception: A critical review of methods," *J. Sport Health Sci.*, vol. 5, no. 1, pp. 80–90, Mar. 2016.
[19] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[20] K. Friston, "The free-energy principle: A unified brain theory?" *Nature Rev. Neurosci.*, vol. 11, no. 2, pp. 127–138, 2010.
[21] M. W. Spratling, "Predictive coding as a model of the V1 saliency map hypothesis," *Neural Netw.*, vol. 26, pp. 7–28, Feb. 2012.
[22] K. J. Friston, J. Daunizeau, and S. J. Kiebel, "Reinforcement learning or active inference?" *PLoS ONE*, vol. 4, no. 7, Jul. 2009, Art. no. e6421.
[23] M. Baruah and B. Banerjee, "The perception-action loop in a predictive agent," in *Proc. CogSci*, 2020, pp. 1171–1177.
[24] K. Friston, R. A. Adams, L. Perrinet, and M. Breakspear, "Perceptions as hypotheses: Saccades as experiments," *Frontiers Psychol.*, vol. 3, p. 151, May 2012.
[25] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Proc. NIPS*, 2015, pp. 2980–2988.
[26] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *Proc. NIPS*, 2018, pp. 5575–5585.
[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[28] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: Extending MNIST to handwritten letters," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2921–2926.
[29] E. D. de Jong, "Incremental sequence learning," 2016, *arXiv:1611.03068*.
[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[31] R. Salakhutdinov and I. Murray, "On the quantitative analysis of deep belief networks," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 872–879.

MURCHANA BARUAH received the B.E. degree in electronics and telecommunication engineering from Assam Engineering College, India, in 2013, and the M.Tech. degree in signal processing and communication from the Gauhati University Institute of Science and Technology, India, in 2015. She is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Memphis, USA. From 2019 to 2020, she was an Intern at ZF Friedrichshafen AG, working on data analysis and machine learning to improve the performance of autonomous driving. Her research interests include artificial intelligence, machine learning, and cognitive science, focusing on predictive modeling and attention models with applications to multimodal data.

BONNY BANERJEE received the M.S. degree in electrical engineering and the Ph.D. degree in computer and information science from The Ohio State University, Columbus, USA. In 2019, he was a Visiting Researcher at Samsung Research Institute, working on social robots. Currently, he is a tenured Associate Professor of electrical and computer engineering at the University of Memphis, USA, with the joint appointment in the highly interdisciplinary research-intensive Institute for Intelligent Systems. Just after graduating with Ph.D., he spent three and half years leading the research at a startup, which resulted in seven patents, substantial investor funding, and launch of a commercial product for the end-user. The product was covered widely by major news and television channels. The intellectual property was acquired by the leading company in the field. He has published over 60 peer-reviewed articles in top journals and conference proceedings in the areas of artificial intelligence, machine learning, data mining, and cognitive science. His research has been funded by the U.S. National Science Foundation and the City of Memphis.

ATULYA K. NAGAR received the B.Sc. (Hons.), M.Sc., and M.Phil. (Hons.) degrees in mathematical physics from the MDS University of Ajmer, India, and the D.Phill. degree in applied nonlinear mathematics from the University of York, U.K., in 1996. He currently holds the Foundation Chair as a Professor of Mathematical Sciences and the Pro Vice-Chancellor (Research) at Liverpool Hope University, U.K. He is responsible for developing sciences and engineering and has been the Head of the School of Mathematics, Computer Science and Engineering, which he established at the university. Prior to joining Liverpool Hope University, he was with Brunel University London, London. He is an internationally respected scholar working at the cutting edge of nonlinear mathematics, theoretical computer science, and systems engineering. He has edited volumes on intelligent systems and applied mathematics. He is well published with over 450 publications in prestigious publishing outlets. He has an extensive background and experience of working in universities at U.K., and India. He has been an expert reviewer for the Biotechnology and Biological Sciences Research Council (BBSRC) grants peer-review committees for Bioinformatics Panel and the Engineering and Physical Sciences Research Council (EPSRC) for High Performance Computing Panel, and served on the Peer-Review College of the Arts and Humanities Research Council (AHRC) as a Scientific Expert Member. He sits on the JISC Research Strategy Group and he is a fellow of the Institute of Mathematics and Its Applications (FIMA) and the Higher Education Academy (FHEA). He received a prestigious Commonwealth Fellowship for pursuing his doctorate (D.Phil.) degree.