The Perception-Action Loop in a Predictive Agent

Murchana Baruah (mbaruah@memphis.edu)
Institute for Intelligent Systems, and Department of Electrical and Computer Engineering,
University of Memphis, Memphis, TN 38152, USA

Bonny Banerjee (bbnerjee@memphis.edu)
Institute for Intelligent Systems, and Department of Electrical and Computer Engineering,
University of Memphis, Memphis, TN 38152, USA

Abstract

We propose an agent model consisting of perceptual and proprioceptive pathways. It actively samples a sequence of percepts from its environment using the perception-action loop. The model predicts to complete the partial percept and proprioceptive sequences observed till each sampling instant, and learns where and what to sample from the prediction error, without supervision or reinforcement. The model is exposed to two kinds of stimuli: images of fully-formed handwritten numerals/alphabets, and videos of gradual formation of numerals. For each object class, the model learns a set of salient locations to attend to in images and a policy consisting of a sequence of eye fixations in videos. Behaviorally, the same model gives rise to saccades while observing images and tracking while observing videos. The proposed agent is the first of its kind to interact with and learn end-to-end from static and dynamic environments to generate realistic handwriting with state-of-the-art performance.

Keywords: Agent; Multimodal; Proprioception; Perception; Attention; Saccade; Tracking.

Introduction

An important property of human visual system that fosters efficiency is that one does not tend to process a whole spatiotemporal observation in its entirety at once. Instead humans focus attention selectively, in space and time, on parts of the observation to acquire information when and where it is needed, and combine information from different fixations over time to build up an internal representation of the observation (Rensink, 2000), guiding future eye movements and decision making.

Recently, the problem of handwriting generation has gained interest. Machine learning models for handwriting generation have been reported that incorporate visual attention (e.g., (Gregor et al., 2015)) and ones that do not (e.g., (Gregor et al., 2015; Murray & Salakhutdinov, 2009; Gregor et al., 2013; Salimans et al., 2015; Raiko et al., 2014)), with the former reporting better performance than the latter. In this paper, we propose an agent in the predictive coding framework which observes its visual environment via a sequence of glimpses. The agent is implemented in software; its actions are limited to sampling the visual environment. The predictive coding framework entails that the agent actively makes inferences (predictive and causal), acts and learns by minimizing sensory prediction errors in a perception-action loop (Friston, 2010). Our agent does not require reinforcement or utilities/values of states to learn policies, consistent with predictive coding (Friston et al., 2009).

The novelty of our agent is threefold: (1) the same agent model can interact with static images and dynamic videos; (2) taking into account the past observations and its learned knowledge, the agent completes the perceptual and proprioceptive patterns after each glimpse; and (3) the pattern completion component in our agent is a multimodal generative model where the prediction error in a perceptual modality provides the observation for the proprioceptive modality. To the best of our knowledge, the proposed agent is the first of its kind to interact with and learn end-to-end from static (image) and dynamic (video) environments, with state-of-the-art performance in handwriting generation.

Preliminaries

Definition 1. Agent. An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators (Russell & Norvig, 2002). Such agents, implemented in software, have been reported in our prior work (Banerjee, 2007; Banerjee & Chandrasekaran, 2010a, 2010b; Najnin & Banerjee, 2017; Kapourchali & Banerjee, 2019, 2020; Baruah & Banerjee, 2020) as well as in others'.

Definition 2. Perception. Perception is the mechanism that allows an agent to interpret sensory signals from the external environment (Han et al., 2016).

Definition 3. Proprioception. Proprioception is perception where the environment is the agent's own body. Proprioception allows an agent to internally perceive the location, movement and action of parts of its body (Han et al., 2016).

Definition 4. Generative model. A generative model, p_{model}, maximizes the log-likelihood $\mathcal{L}(x;\theta)$ of the generated data, where θ is a set of parameters and x is a set of data points (Goodfellow, 2016).

Definition 5. Evidence lower bound (ELBO). If z is a latent continuous random variable generating the data x, computing log-likelihood requires computing the integral of the marginal likelihood, $\int p_{model}(x,z)dz$, which is intractable (Kingma & Welling, 2013). Variational inference involves optimization of an approximation of the intractable posterior by defining an evidence lower bound (ELBO) on the log-likelihood,

$$\mathcal{L}(x;\theta) \leq \log p_{model}(x;\theta) \tag{1}$$

Definition 6. Variational autoencoder (VAE). VAE is a deep generative model that assumes the data consists of independent and identically distributed samples, and the prior, $p_\theta(z)$, is an isotropic Gaussian. VAE maximizes the ELBO given by (Kingma & Welling, 2013):

$$\mathcal{L}(x;\theta) \leq \mathbb{E}_{q_\phi z|x}[\log p_\theta(x|z)] - \mathrm{KL}[q_\phi(z|x), p_\theta(z)] \quad (2)$$

where $q_\phi(z|x)$ is a recognition model, $p_\theta(x|z)$ is a generative model, $\mathbb{E}$ denotes expectation, and KL denotes Kullback-Leibler divergence.

Definition 7. Saliency. Saliency is a property of each location in a predictive agent's environment. The attention mechanism is a function of the agent's prediction error (Spratling, 2012; Banerjee & Dutta, 2014; Najnin & Banerjee, 2017; Kapourchali & Banerjee, 2019, 2020). Other definitions of saliency (e.g., (Dutta & Banerjee, 2015; Dutta, Banerjee, & Reddy, 2016)) are not relevant to this paper.

Problem Statement

Let $\mathbf{X} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(n)}\}$ be a set of observable variables representing an environment in n modalities. The variable representing the i-th modality is a sequence: $\mathbf{X}^{(i)} = \langle X_1^{(i)}, X_2^{(i)}, \ldots, X_T^{(i)} \rangle$, where T is the sequence length. Let $\mathbf{x}_{\leq t} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$ be a partial observation of $\mathbf{X}$ such that $\mathbf{x}^{(i)} = \langle x_1^{(i)}, \ldots, x_t^{(i)} \rangle$, $1 \leq t \leq T$. We define *pattern completion* as the problem of generating $\mathbf{X}$ as accurately as possible from its partial observation $\mathbf{x}_{\leq t}$.

Given $\mathbf{x}_{\leq t}$ and a generative model p_θ with parameters θ and latent variables $z_{\leq t}$ (see Def. 4, 5), the generative process of $\mathbf{X}$ is given as:

$$p_\theta(\mathbf{X}|\mathbf{x}_{\leq t}) = \int p_\theta(\mathbf{X}|\mathbf{x}_{\leq t}, z_{\leq t}; \theta) p_\theta(z_{\leq t}) dz \quad (3)$$

At any time t, the objective for pattern completion is to maximize the log-likelihood of $\mathbf{X}$, i.e. $\arg\max_\theta \int log(p_\theta(\mathbf{X}|\mathbf{x}_{\leq t}, z_{\leq t}; \theta) p_\theta(z_{\leq t})) dz$.

Agent Architecture

The proposed predictive agent architecture comprises of five components: environment, observation, pattern completion, action selection, and learning. See block diagram in Fig. 1(a).

1. **Environment.** The environment is the source of sensory data. Two kinds of environment are considered: static (images) and dynamic (videos).

2. **Observation.** The agent interacts with the environment via a sequence of glimpses. The observations, sampled from the sensory data, are in two modalities: perception and proprioception.

   Perceptual sensory observation. Perceptual sensory reports the visual observation at a location generated by the proprioception, as in (Friston et al., 2012). Mathematically, $\mathbf{x}^{(1)} = \langle x_1^{(1)}, \ldots, x_T^{(1)} \rangle$, where $x_t^{(1)} \in \{0,1\}^{n \times n}$ is a patch. If an image is of size $M \times M$ pixels, $n \leq M$.

   Proprioceptive sensory observation. Proprioceptive sensory reports the activations of oculomotor muscles due to fixation. In this paper, it is represented by the 2D coordinates of the fixation location in the environment. Mathematically, $\mathbf{x}^{(2)} = \langle x_1^{(2)}, \ldots, x_T^{(2)} \rangle$, $x_t^{(2)} \in [0,1]^2$.

3. **Pattern completion.** The pattern is completed for both perceptual and proprioceptive modalities from all their past observations.

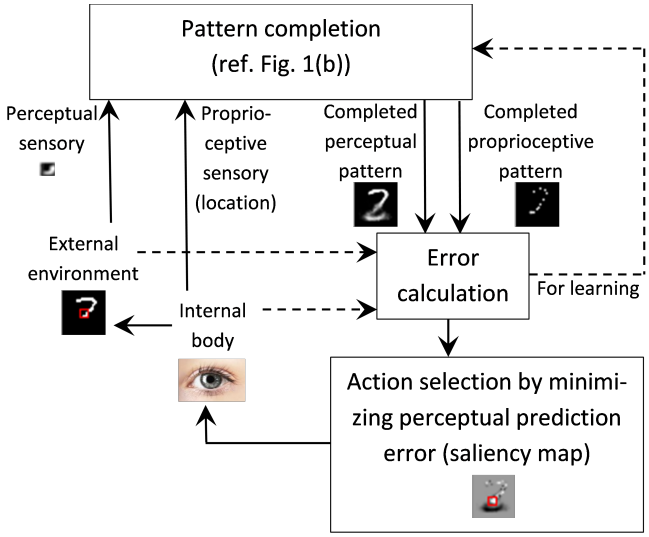
   Perceptual pattern completion. The completed perceptual pattern, $\mathbf{X}^{(1)}$, at any time t is the fully generated handwritten numeral or alphabet expected to be observed from $t = 1$ through $t = T$. We represent the perceptual modality as: $\mathbf{X}^{(1)} = \langle X_1^{(1)}, \ldots, X_T^{(1)} \rangle$, where $X_t^{(1)} \in \{0,1\}^{M \times M}$.

   Proprioceptive pattern completion. The completed proprioceptive pattern, $\mathbf{X}^{(2)}$, at any time t is the expected sequence of actions for sampling the observations from $t = 1$ through $t = T$. We represent the proprioceptive modality as: $\mathbf{X}^{(2)} = \langle X_1^{(2)}, \ldots, \mathbf{X}_T^{(2)} \rangle$, where $X_t^{(2)} \in [0,1]^{2 \times T}$.

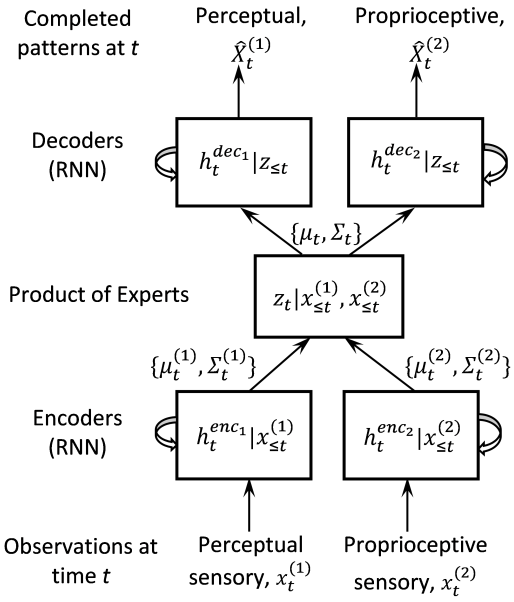
   A multimodal variational recurrent neural network (RNN) (Fig. 1(b)) is used for completing the pattern for the two modalities. Recognition and generation are the two processes involved in a variational RNN (Chung et al., 2015).

   Recognition (Encoder). The recognition model, $q_\phi(z_t|\mathbf{x}_{\leq t})$, is a probabilistic encoder (Kingma & Welling, 2013). Given the observations $\mathbf{x}_{\leq t}$, it produces a Gaussian distribution over the possible values of the code z_t from which the observations $\mathbf{x}_{\leq t}$ could have been generated. The recognition model consists of two RNNs, each with one layer of long short-term memory (LSTM) units. Each RNN generates the parameters for the approximate posterior distribution for each modality. The parameters from each modality are combined using product of experts (PoE), as in (Wu & Goodman, 2018), to generate the joint distribution parameters (see Fig. 1(b)) for the approximate posterior $q_\phi(z_t|x_{\leq t}^{(1)}, x_{\leq t}^{(2)})$. The prior is sampled from a standard normal distribution $p_\theta(z_t) \sim \mathcal{N}(0,1)$, as in (Gregor et al., 2015).

   Generation (Decoder). The generative model, $p_\theta(\mathbf{X}_t|\mathbf{x}_{\leq t}, z_{\leq t})$, 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, $X_t^{(1)}|x_{\leq t}^{(1)}, z_{\leq t}$ is sampled from a Bernoulli distribution (as the perceptual observation is binary) with means generated by the first RNN, and $X_t^{(2)}|x_{\leq t}^{(2)}, z_{\leq t}$ is sampled from a Gaussian distribution (as the proprioceptive observation is real) with means and variances as output of the second RNN (see Fig. 1(b)).



(a) Predictive agent architecture.



(b) Pattern completion model.

Figure 1: Different components of the proposed agent. In (a), the dashed arrows indicate direction of signal flow for learning while firm arrows indicate direction of signal flow for inference.

For time-varying data, such as videos, $X_1^{(1)}, X_2^{(1)}$ represent two adjacent frames; for static data, such as images, they represent the same image.

4. **Action selection.** In the proposed agent model, action selection is to decide which location in the environment to sample from. If the environment is an image of size $M \times M$ pixels, there are M^2 possible locations in that image.

   An action at time t is generated as a function of the saliency map. We denote the saliency map at time t as $S_t \in \mathbb{R}^{M \times M}$ and the value of the saliency map at location ℓ as $S_{t,\ell}$. The saliency map is a function of the prediction error computed as $S_t = \mathcal{N}(.,\sigma) * [X_t^{(1)} - \hat{X}_t^{(1)}]^+$, where $\hat{X}_t^{(1)}$ is the generated data, $\mathcal{N}(.,\sigma)$ is a Gaussian filter with standard deviation σ, $*$ is the convolution operator, $[.]^+ \equiv max(0,.)$. In our experiments, $\sigma = 2$. The salient location ℓ^* is the location with the highest value in the saliency map; $\ell^* = \underset{\ell \in \{1,2,\ldots,M^2\}}{\text{argmax}} \; S_{t,\ell}$.

   The salient location ℓ at any time t is the proprioceptive observation $x_{t+1}^{(2)}$ for time $t+1$. Therefore, the salient locations at $t = 0, 1, 2, \ldots, T-1$ constitutes the proprioceptive pattern $\mathbf{X}^{(2)}$. Hence, prediction error (saliency) guides the sampling of the observations 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 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 (first term in Eq. 4 for $i = 2$). 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 most informative location 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 (see Def. 5) for the multimodal variational RNN. This objective function, obtained by modifying the objective for multimodal VAE (Eq. 2 in (Wu & Goodman, 2018)) with variational RNN (Eq. 1 in (Chung et al., 2015)), is as follows:

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

   where the first term for $i = 1, 2$ is the expected negative prediction error for the two modalities. The KL-divergence is a regularizer to prevent overfitting during training.

   The negative of the ELBO is also referred to as negative log-likelihood (NLL). In this paper, we refer to the negative

of the first term in Eq. 4 for $i = 1$ and 2 as perceptual NLL and proprioceptive NLL respectively.

Experimental Evaluations

Datasets. The proposed model is evaluated on three datasets:
(1) MNIST (LeCun et al., 1998): It consists of 60,000 training and 10,000 test images (28×28 pixels) of handwritten numerals $\{0, 1, \ldots 9\}$.
(2) EMNIST (Cohen et al., 2017): It consists of 124,800 training and 20,800 test images (28×28 pixels) of handwritten English alphabets in uppercase and lowercase, forming a balanced set.
(3) MNIST stroke sequence dataset (SMNIST) (de Jong, 2016): It was created for sequence learning from the original MNIST dataset. It consists of the MNIST images and a sequence of locations of how the numeral might be formed for each image. We select an equal number of equidistant locations and create a video for each image such that it shows the gradual formation of the numeral. Each video frame is 28×28 pixels.

Experimental setup. The number of hidden units for recognition and generative models are 256 and 512 respectively for each modality. The number of latent variables is 20, minibatch size is 100, and maximum number of glimpses $T = 12$. The parameters $\beta, \lambda_1, \lambda_2$ are all fixed at 1. The model is trained end-to-end using backpropagation and Adam optimization (Kingma & Ba, 2014) with a learning rate of 0.001.

The initial observation is always sampled from the origin of the image for each dataset. The origin for each image in the MNIST and SMNIST datasets is the starting pixel of the numeral, which is obtained from (de Jong, 2016) and the center pixel of the image for EMNIST. Fixing the origin as the starting pixel of the numeral allows to learn a position-invariant representation of the numerals. In our experiments, $n = 5$.

Generative models reported in the literature are evaluated using NLL ($-\log p$) involving the perceptual modality only. In order to compare the performance of the proposed agent with reported models, we define two variants of our model, Prop1 and Prop2, that generate only the perceptual modality. The proposed model is abbreviated as Prop.

Prop: Observation and generation consist of both perceptual and proprioceptive modalities, as in the objective in Eq. 4.

Prop1: Observation and generation consist of perceptual modality only; $i = 1$ in the objective in Eq. 4.

Prop2: Observation consists of perceptual and proprioceptive modalities. Generation consists of perceptual modality only; $i = 1$ in the objective in Eq. 4.

Evaluation results. Figs. 2a–e show that the agent can complete the visual observation very close to the true pattern within a few glimpses. In the initial few time steps, the completed perceptual patterns (third row of Figs. 2a–e) are blurred images as the agent samples from the latent distribution of multiple classes. In the completed proprioceptive patterns (bottom row of Figs. 2a–e), during the initial steps, the locations are concentrated at a small region. This difference between the perceptual and proprioceptive modalities is due to the difference in their dimensions and assumptions in their distributions.

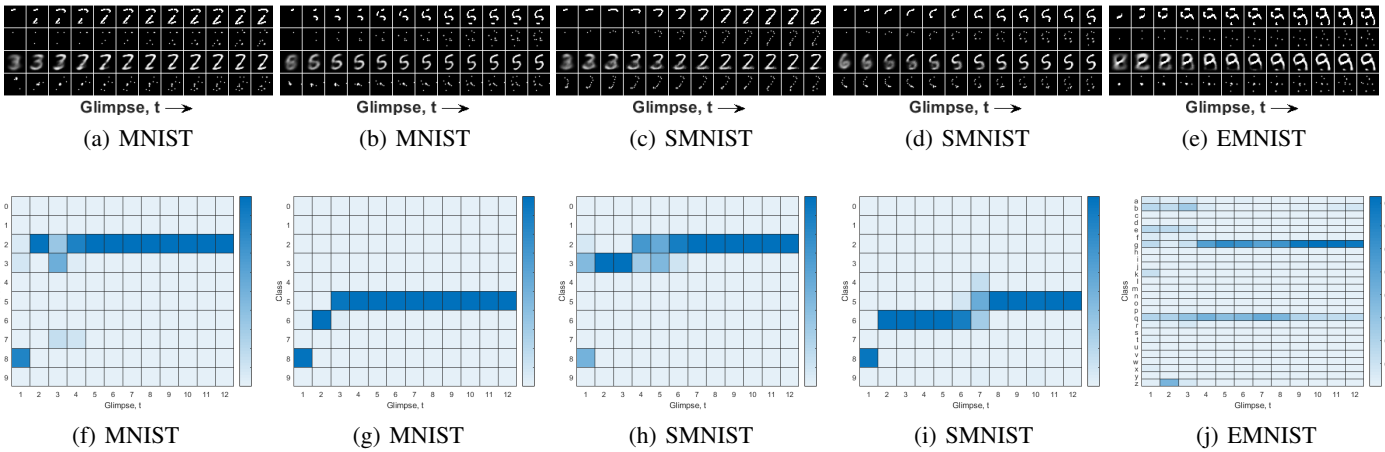
The agent can infer an object class within a few glimpses, as can be seen from the class distribution of the completed patterns at each time step (Figs. 2f–j). It, however, requires more observations to refine the style within a class. From the example in Fig. 2, inferring a class takes less time in images (Fig. 2b) than in videos (Fig. 2d). The agent figures out that it is a '5' from the first three observations from the image in Fig. 2b, but takes eight observations to do the same from the video in Fig. 2d. This is because, in our model, the entire image is given which is being observed via a sequence of glimpses, but the entire video is not given. Hence, the salient locations can occur anywhere in an image which allows these locations to be more discriminative towards object classes. In contrast, the salient locations in a video follow the trajectories of the formation of a numeral; thus our model has to wait for the discriminative locations to present themselves before a class can be inferred.

Table 1: Negative log-likelihood (NLL) at the final time step, perceptual (Percep.) NLL, proprioceptive (Proprio.) NLL, and average (Avg.) NLL comparison of the proposed model (Prop) and its variants (Prop1, Prop2) for all datasets. Avg. NLL is mean over all glimpses. Best results are highlighted.

Dataset	Model	NLL (T) ($\leq$)	Percep. NLL	Proprio. NLL	Avg. NLL $\leq$
MNIST	Prop		**1107.6**	-631.8	44.01
	Prop1	156.5	1988.7		167.5
	Prop2	**79.20**	1407.2		120
EMNIST	Prop		**1119.5**	-628.7	46.2
	Prop1	184.8	2318		196.7
	Prop2	**65.7**	1468.8		125.9
SMNIST	Prop		**1183.4**	-711.8	43.7
	Prop1	153.6	1999.7		168.4
	Prop2	**63.0**	1242.3		106.4

The salient locations occur somewhat randomly for the static case (image) but follows a sequence for the dynamic case (video). Consequently, the agent saccades while observing images and tracks while observing videos. For each object class, the actual and predicted proprioceptive pattern distributions, obtained by averaging the actual and predicted salient locations from the test set, are very similar (Fig. 3a–d) for both static and dynamic cases. Thus, our agent can learn the distribution of salient locations from its own behavior in both cases.

For all datasets, the perceptual NLL is lowest for our model, followed by its variants Prop1 and Prop2 (ref. Table 1). This is because the pattern completion model learns a richer representation of the environment by maximizing perceptual and proprioceptive log-likelihood (proposed model)



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

(f) MNIST (g) MNIST (h) SMNIST (i) SMNIST (j) EMNIST

Figure 2: Pattern completion for two randomly chosen images from MNIST test set (a, b), the same examples from SMNIST (c, d), and a random example from EMNIST (e). Rows 1, 2 show the perceptual and proprioceptive observation till the current glimpse in $M \times M$ space. Rows 3, 4 show the perceptual and proprioceptive pattern completion after each glimpse. Each column in subfigures a–j corresponds to time or glimpse number, increasing from left to right. For each case a–e, the probability distribution over the class of the generated image is shown below in subfigures f–j. The vertical axis denotes the class labels in f–j. The probability distribution is generated by training a multilayer perceptron classifier with a softmax output layer.

Table 2: Performance comparison of the proposed model for MNIST. Best results from our model are highlighted. The other results are from [1] (Salakhutdinov & Hinton, 2009), [2] (Murray & Salakhutdinov, 2009), [3] (Uria et al., 2014), [4] (Raiko et al., 2014), [5] (Rezende et al., 2014), [6] (Salimans et al., 2015), [7] (Gregor et al., 2013), [8] (Gregor et al., 2013), [9] (Gregor et al., 2015), [10] (Oord et al., 2016).

Model	NLL
DBM 2hl [1]	≈ 84.62
DBN 2hl [2]	≈ 84.55
NADE [3]	88.33
EoNADE 2hl (128 orderings) [3]	85.10
EoNADE-5 2hl (128 orderings) [4]	84.68
DLGM [5]	≈ 86.60
DLGM 8 leapfrog steps [6]	≈ 85.51
DARN 1hl [7]	≈ 84.13
MADE 2hl (32 masks) [8]	86.64
DRAW [9]	≤ 80.97
DRAW without attention [9]	≤ 87.4
PixelCNN [10]	81.30
Row LSTM [10]	80.54
Diagonal BiLSTM (1 layer, h=32) [10]	80.75
Diagonal BiLSTM (7 layers, h=16) [10]	79.20
Prop1	≤ 156.5
Prop2	$\leq$ **79.20**

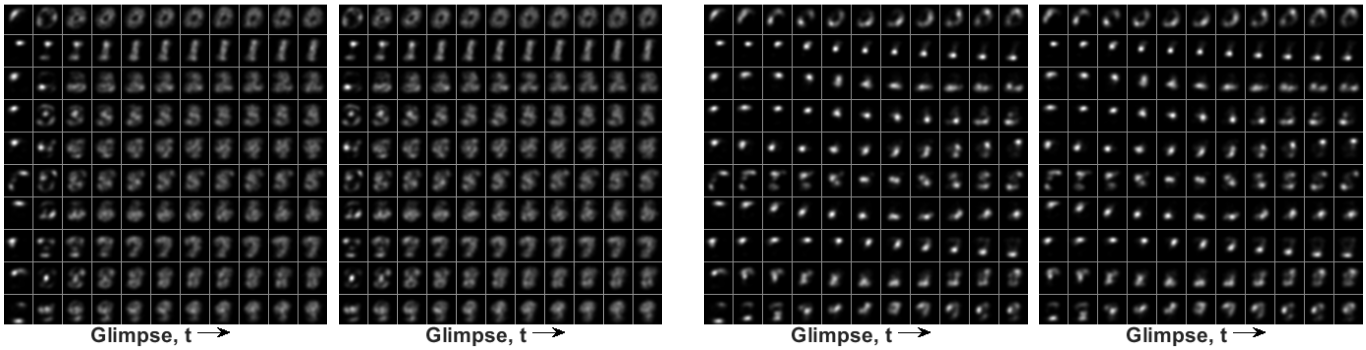
than by maximizing perceptual log-likelihood alone (Prop1, Prop2). In the proposed model and Prop2, the posterior is approximated from perceptual and proprioceptive observations. In the proposed model, the proprioceptive NLL (ref. Table 1) is lower for SMNIST than for MNIST or EMNIST dataset because salient locations occur in a sequence in SMNIST but somewhat randomly in the others, thereby making it easier to learn from SMINST.

The NLL (reported at the final time step) for MNIST from Prop2 is better than all models and comparable to the state-of-the-art (ref. Table 2). As the nature of the EMNIST data is similar to MNIST, the NLL from EMNIST is comparable to that from MNIST using our model. We are the first to report NLL on the EMNIST dataset.

As expected, the NLL drops significantly with increase in glimpses (ref. Fig. 4). The NLL at the final time step is much less than the average NLL for our model and its variants (ref. Table 1). Thus, by sampling the observations greedily based on saliency, the proposed agent improves its generations as it sees more, resulting in very realistic generations after the final glimpse.

Conclusions

A predictive agent with perceptual and proprioceptive pathways is proposed. It completes the observed pattern for perceptual and proprioceptive modalities after each glimpse. The perceptual prediction error provides the observation for the proprioceptive modality. Experimental results using our agent for handwriting generation are comparable to the state-of-the-art. Behaviorally, the agent saccades while observing images and tracks while observing videos. This is the



(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. The actual distribution is obtained from the salient locations in the saliency map while the predicted distribution is obtained from the salient locations predicted by the model. Actual locations are from glimpse 1–11 and predicted locations from glimpse 2–12. Each row represents a class. Each column corresponds to time or glimpse number, increasing from left to right.

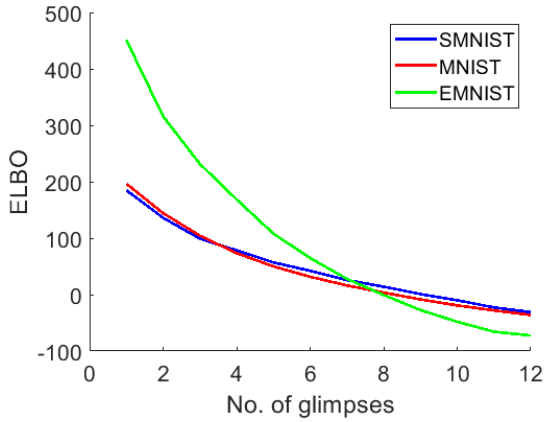


Figure 4: In our model, ELBO decreases steadily with glimpses for all datasets. Best viewed in color.

first work on an attention-based agent for handwriting generation that actively samples its environment, static or dynamic, guided by prediction error.

References

Banerjee, B. (2007). *Spatial problem solving for diagrammatic reasoning*. Unpublished doctoral dissertation, Dept. of Computer Science & Engineering, The Ohio State University, Columbus.

Banerjee, B., & Chandrasekaran, B. (2010a). A constraint satisfaction framework for executing perceptions and actions in diagrammatic reasoning. *J. Artif. Intell. Res.*, 373–427.

Banerjee, B., & Chandrasekaran, B. (2010b). A spatial search framework for executing perceptions and actions in diagrammatic reasoning. In *Diagrammatic Representation and Inference, LNAI* (Vol. 6170, pp. 144–159). Springer, Heidelberg.

Banerjee, B., & Dutta, J. K. (2014). SELP: A general-purpose framework for learning the norms from saliencies in spatiotemporal data. *Neurocomputing*, *138*, 41–60.

Baruah, M., & Banerjee, B. (2020). A multimodal predictive agent model for human interaction generation. In *CVPR Workshop*.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., & Bengio, Y. (2015). A recurrent latent variable model for sequential data. In *NIPS* (pp. 2980–2988).

Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017). EMNIST: An extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373*.

de Jong, E. D. (2016). Incremental sequence learning. *arXiv preprint arXiv:1611.03068*.

Dutta, J. K., & Banerjee, B. (2015). Online detection of abnormal events using incremental coding length. In *AAAI* (pp. 3755–3761).

Dutta, J. K., Banerjee, B., & Reddy, C. K. (2016). RODS: Rarity based outlier detection in a sparse coding framework. *IEEE Trans. Knowl. Data Eng.*, *28*(2), 483–495.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.*, *11*(2), 127.

Friston, K., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Front. Psychol.*, *3*, 151.

Friston, K., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PloS One*, *4*(7), e6421.

Goodfellow, I. (2016). NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015). DRAW: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*.

Gregor, K., Danihelka, I., Mnih, A., Blundell, C., & Wierstra, D. (2013). Deep autoregressive networks. *arXiv preprint arXiv:1310.8499*.

Han, J., Waddington, G., Adams, R., Anson, J., & Liu, Y.

(2016). Assessing proprioception: A critical review of methods. *J. Sport Health Sci.*, *5*(1), 80–90.

Kapourchali, M. H., & Banerjee, B. (2019). State estimation via communication for monitoring. *IEEE Trans. Emerg. Topics Comput. Intell.*.

Kapourchali, M. H., & Banerjee, B. (2020). EPOC: Efficient perception via optimal communication. In *AAAI*.

Kingma, D. P., & Ba, J. (2014). ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*, *86*(11), 2278–2324.

Murray, I., & Salakhutdinov, R. R. (2009). Evaluating probabilities under high-dimensional latent variable models. In *NIPS* (pp. 1137–1144).

Najnin, S., & Banerjee, B. (2017). A predictive coding framework for a developmental agent: Speech motor skill acquisition and speech production. *Speech Commun.*, *92*, 24–41.

Oord, A. v. d., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*.

Raiko, T., Li, Y., Cho, K., & Bengio, Y. (2014). Iterative neural autoregressive distribution estimator nade-k. In *NIPS* (pp. 325–333).

Rensink, R. A. (2000). The dynamic representation of scenes. *Vis. Cogn.*, *7*(1-3), 17–42.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

Russell, S., & Norvig, P. (2002). *Artificial Intelligence: A Modern Approach* (2nd ed.). Prentice Hall.

Salakhutdinov, R., & Hinton, G. (2009). Deep Boltzmann machines. In *AISTATS* (pp. 448–455).

Salimans, T., Kingma, D., & Welling, M. (2015). Markov chain monte carlo and variational inference: Bridging the gap. In *ICML* (pp. 1218–1226).

Spratling, M. W. (2012). Predictive coding as a model of the V1 saliency map hypothesis. *Neural Netw.*, *26*, 7–28.

Uria, B., Murray, I., & Larochelle, H. (2014). A deep and tractable density estimator. In *ICML* (pp. 467–475).

Wu, M., & Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. In *NIPS* (pp. 5575–5585).