A Probabilistic Hard Attention Model For Sequentially Observed Scenes

Samrudhdhi B. Rangrej
samrudhdhi.rangrej@mail.mcgill.ca

James J. Clark
james.j.clark@mcgill.ca

Centre for Intelligent Machines (CIM)
McGill University
Montreal, Quebec, Canada

Abstract

A visual hard attention model actively selects and observes a sequence of subregions in an image to make a prediction. The majority of hard attention models determine the attention-worthy regions by first analyzing a complete image. However, it may be the case that the entire image is not available initially but instead sensed gradually through a series of partial observations. In this paper, we design an efficient hard attention model for classifying such sequentially observed scenes. The presented model never observes an image completely. To select informative regions under partial observability, the model uses Bayesian Optimal Experiment Design. First, it synthesizes the features of the unobserved regions based on the already observed regions. Then, it uses the predicted features to estimate the expected information gain (EIG) attained, should various regions be attended. Finally, the model attends to the actual content on the location where the EIG mentioned above is maximum. The model uses a) a recurrent feature aggregator to maintain a recurrent state, b) a linear classifier to predict the class label, c) a Partial variational autoencoder to predict the features of unobserved regions. We use normalizing flows in Partial VAE to handle multi-modality in the feature-synthesis problem. We train our model using a differentiable objective and test it on five datasets. Our model gains 2-10% higher accuracy than the baseline models when both have seen only a couple of glimpses. code: https://github.com/samrudhdhirangrej/Probabilistic-Hard-Attention.

1 Introduction

A deep feedforward network observes an entire scene to achieve state-of-the-art performance. However, observing an entire scene may not be necessary as many times the task-relevant information lies only in a few parts of the scene. A visual hard attention technique allows a recognition system to attend to only the most informative subregions called *glimpses* [41, 57]. A hard attention model builds a representation of a scene by sequentially acquiring useful glimpses and fusing the collected information. The representation guides recognition and future glimpse acquisition. Hard attention is useful to reduce data acquisition cost [54], to develop interpretable [16], computationally efficient [29] and scalable [45] models.

The majority of hard attention models first analyze a complete image, occasionally at low resolution, to locate the task-relevant subregions [4, 16]. However, in practice, we often do not have access to the entire scene. Instead, we only observe parts of the scene as we attend

to them. We decide the next attention-worthy location at each step in the process based on the partial observations collected so far. Examples include a self-driving car navigating in unknown territory, time-sensitive aerial imagery for a rescue operation, etc. Here, we develop a hard-attention model for such a scenario. While the majority of existing hard attention models pose glimpse acquisition as a reinforcement learning task to optimize a non-differentiable model [41, 57], we use a fully differentiable model and training objective. We train the model using a combination of discriminative and generative objectives. The former is used for class-label prediction, and the latter is used for the content prediction for unobserved regions. We use the predicted content to find an optimal attention-worthy location as described next.

A hard attention model predicts the class label of an image by attending various informative glimpses in an image. We formulate the problem of finding optimal glimpse-locations as Bayesian Optimal Experiment Design (BOED). Starting from a random location, a sequential model uses BOED to determine the next optimal location. To do so, it estimates the expected information gain (EIG) obtained from observing glimpses at yet unobserved regions of the scene and selects a location with maximum EIG. As the computation of EIG requires the content of regions, the model *synthesizes* the unknown content conditioned on the content of the observed regions. For efficiency reasons, the model predicts the content in the feature space instead of the pixel space. Our model consists of three modules, a recurrent feature aggregator, a linear classifier, and a Partial VAE [39]. Partial VAE synthesizes features of various glimpses in the scene based on partial observations. There may exist multiple possibilities for the content of the unobserved regions, given the content of the observed regions. Hence, we use normalizing flows in Partial VAE to capture the multi-modality in the posterior. Our main contributions are as follows.

- We develop a hard attention model to classify images using a series of partial observations. The model estimates EIG of the yet unobserved locations and attends a location with maximum EIG.
- To estimate EIG of unobserved regions, we synthesize the content of these regions. Furthermore, to improve the efficiency of a model, we synthesize the content in the feature space and use normalizing flows to capture the multi-modality in the problem.
- We support our approach with a principled mathematical framework. We implement the model using a compact network and perform experiments on five datasets. Our model achieves 2-10% higher accuracy compared to the baseline methods when both have seen only a couple of glimpses.

2   Related Works

Hard Attention. A hard attention model prioritizes task-relevant regions to extract meaningful features from an input. Early attempts to model attention employed image saliency as a priority map. High priority regions were selected using methods such as winner-take-all [26, 27, 33], searching by throwing out all features but the one with minimal activity [1], and dynamic routing of information [43]. Few used graphical models to model visual attention. Rimey and Brown [49] used augmented hidden Markov models to model attention policy. Larochelle and Hinton [35] used a Restricted Boltzmann Machine (RBM) with third-order connections between attention location, glimpse, and the representation of a scene. Motivated by this, Zheng et al. [53] proposed an autoregressive model to compute exact gradients, unlike in an RBM. Tang et al. [53] used an RBM as a generative model and searched for in-

formative locations using the Hamiltonian Monte Carlo algorithm. Many used reinforcement learning to train attention models. Paletta et al. [44] used Q-learning with the reward that measures the objectness of the attended region. Denil et al. [13] estimated rewards using particle filters and employed a policy based on the Gaussian Process and the upper confidence bound. Butko and Movellan [7] modeled attention as a partially observable Markov decision process and used a policy gradient algorithm for learning. Later, Butko and Movellan [8] extended this approach to multiple objects.

Recently, the machine learning community uses the REINFORCE policy gradient algorithm to train non-differentiable hard attention models [4, 16, 41, 45, 57]. Other recent works use EM-style learning procedure [47], wake-sleep algorithm [5], a voting based region selection [2], and spatial transformer [17, 20, 28]. Among the recent methods, [4, 5, 47] look at the low-resolution gist of an input at the beginning, and [16, 17, 20, 57] consume the whole image to predict the locations to attend. In contrast, our model does not look at the entire image at low resolution or otherwise. Moreover, our model is fully differentiable.

Image Completion. Image completion methods aim at synthesizing unobserved or missing image pixels conditioned on the observed pixels [23, 46, 55, 56, 58, 59]. Image completion is an ill-posed problem with multiple possible solutions for the missing image regions. While early methods use deterministic models, [9, 50, 51] used stochastic models to predict multiple instances of a complete image. Many used probabilistic models to ensure that the completions are generated according to their actual probability [18, 19, 39, 52, 62]. We use Partial VAE [39] – a probabilistic model – to predict the content of the complete image given only a few glimpses. However, unlike other approaches that infer image content in the pixel space, we predict content in the feature space.

Patch Selection. Many computer vision methods select attention-worthy regions from an image, e.g., region proposal network [48], multi-instance learning [24], top-K patch selection [3, 11], attention sampling [29]. These approaches observe an entire image and find all attention-worthy patches simultaneously. Our model observes an image only partially and sequentially. Furthermore, it selects a single optimal patch to attend at a given time.

3 Model

In this paper, we consider a recurrent attention model that sequentially captures glimpses from an image x and predicts a label y. The model runs for time $t = 0$ to $T - 1$. It uses a recurrent net to maintain a hidden state h_{t-1} that summarizes glimpses observed until time $t-1$. At time t, it predicts coordinates l_t based on the hidden state h_{t-1} and captures a square glimpse g_t centered at l_t in an image x, i.e. $g_t = g(x, l_t)$. It uses g_t and l_t to update the hidden state to h_t and predicts the label y based on the updated state h_t.

3.1 Building Blocks

As shown in Figure 1, the proposed model comprises the following three building blocks. A recurrent feature aggregator (F and R) maintains a hidden state h_t. A classifier (C) predicts the class probabilities $p(y|h_t)$. A normalizing flow-based variational autoencoder (S and D) synthesizes a feature map of a complete image given the hidden state h_t. Specifically, a flow-based encoder S predicts the posterior of a latent variable z from h_t, and a decoder D uses z to synthesize a feature map of a complete image. A feature map of a complete image can also be seen as a map containing features of all glimpses. The BOED, as discussed in section 3.2, uses the synthesized feature map to find an optimal location to attend at the next time step.
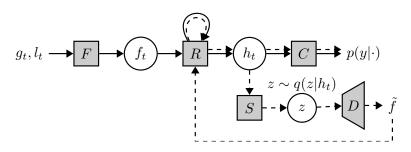
Figure 1: A recurrent attention model sequentially observes glimpses from an image and predicts a class label. At time t, the model actively observes a glimpse g_t and its coordinates l_t. Given g_t and l_t, the feed-forward module F extracts features f_t, and the recurrent module R updates a hidden state to h_t. Using an updated hidden state h_t, the linear classifier C predicts the class distribution $p(y|h_t)$. At time $t+1$, the model assesses various candidate locations l before attending an optimal one. It predicts $p(y|g,l,h_t)$ ahead of time and selects the candidate l that maximizes $D_{KL}[p(y|g,l,h_t)||p(y|h_t)]$. The model synthesizes the features of g using a Partial VAE to approximate $p(y|g,l,h_t)$ without attending to the glimpse g. The normalizing flow-based encoder S predicts the approximate posterior $q(z|h_t)$. The decoder D uses a sample $z \sim q(z|h_t)$ to synthesize a feature map $\tilde{f}$ containing features of all glimpses. The model uses $\tilde{f}(l)$ as features of a glimpse at location l and evaluates $p(y|g,l,h_t) \approx p(y|\tilde{f}(l),h_t)$. Dashed arrows show a path to compute the lookahead class distribution $p(y|\tilde{f}(l),h_t)$.

To distinguish the synthesized feature map from an actual one, let us call the former $\tilde{f}$ and the latter f. Henceforth, we crown any quantity derived from the synthesized feature map with a (˜). Next, we provide details about the three building blocks of the model, followed by a discussion of the BOED in the context of hard attention.

Recurrent Feature Aggregator. Given a glimpse g_t and its location l_t, a feed-forward module extracts features $f_t = F(g_t, l_t)$, and a recurrent network updates a hidden state to $h_t = R(h_{t-1}, f_t)$. We define $F(g,l) = \mathcal{F}_g(g) + \mathcal{F}_l(l)$ and $R(h,f) = LeakyReLU(\mathcal{F}_h(h) + \mathcal{F}_f(f))$. $\mathcal{F}_g$ is a small CNN with receptive-field equal to size of g. $\mathcal{F}_l, \mathcal{F}_h, \mathcal{F}_f$ are shallow networks with one linear layer.

Classifier. At each time step t, a linear classifier predicts the distribution $p(y|h_t) = C(h_t)$ from a hidden state h_t. As the goal of the model is to predict a label y for an image x, we learn a distribution $p(y|h_t)$ by minimizing $D_{KL}[p(y|x)||p(y|h_t)]$. Optimization of this KL divergence is equivalent to the minimization of the following cross-entropy loss.

$$\mathcal{L}_{CE}(t) = -p(y|x)\log(p(y|h_t)) \tag{1}$$

Partial Variational Autoencoder. We adapt a variational autoencoder (VAE) to synthesize the feature map of a complete image from the hidden state h_t. A VAE learns a joint distribution between the feature map f and the latent variable z given h_t, $p(f,z|h_t) = p(f|z)p(z|h_t)$. An encoder approximates the posterior $q(z|f,h_t)$, and a decoder infers the likelihood $p(f|z)$. The optimization of VAE requires calculation of $D_{KL}[q(z|f,h_t)||p(z|h_t)]$ [30]. As the hard attention model does not observe the complete image, it cannot estimate $q(z|f,h_t)$. Hence, we cannot incorporate the standard VAE directly into a hard attention framework and use the

following approach.

At the time t, let's separate an image x into two parts, o_t — the set of regions observed up to t, and u_t — the set of regions as yet unobserved. Ma et al. [39] observed that in a VAE, o_t and u_t are conditionally independent given z, i.e. $p(x|z) = p(u_t|z)p(o_t|z)$. They synthesize u_t independently from the sample $z \sim q(z|o_t)$, while learning the posterior $q(z|o_t)$ by optimizing ELBO on $\log(p(o_t))$. They refer to the resultant VAE as a Partial VAE.

The BOED, as discussed in section 3.2, requires features of o_t and u_t. Ma et al. [39] consider o_t and u_t in pixel space. Without the loss of generality, we consider o_t and u_t in the feature space. Synthesizing o_t and u_t in the feature space serves two purposes. First, the model does not have to extract features of o_t and u_t for the BOED as they are readily available. Second, Partial VAE does not have to produce unnecessary details that may later be thrown away by the feature extractor, such as the exact color of a pixel. Recall that the features $f_{1:t}$ and the hidden state h_t calculated by our attention model correspond to the glimpses observed up to t, which is equivalent to o_t, the set of observed regions. Hence, we write $q(z|o_t)$ as $q(z|h_t)$ and $p(o_t|z)$ as $p(f_{1:t}|z)$ in the ELBO of Partial VAE.

$$\mathcal{L}_{\text{PVAE}}(t) = -ELBO = -\left\{ \mathbb{E}_{q(z|o_t)} \log(p(o_t|z)) - D_{\text{KL}}[q(z|o_t)||p(z)] \right\}$$
$$= -\left\{ \mathbb{E}_{q(z|h_t)} \log(p(f_{1:t}|z)) - D_{\text{KL}}[q(z|h_t)||p(z)] \right\} \quad (2)$$

In equation 2, the prior $p(z)$ is a Gaussian distribution with zero mean and unit variance. To obtain expressive posterior $q(z|h_t)$, we use normalizing flows in Partial VAE [52]. Specifically, we use auto-regressive Neural Spline Flows (NSF) [15]. Between the two flow layers, we flip the input [14] and normalize it using ActNorm [31]. Refer to supp. material for brief overview of NSF. In Figure 1, the flow-based encoder S infers the posterior $q(z|h_t) = S(h_t)$.

In a Partial VAE, $p(f|z) = p(f_{1:t}|z)p(f_{1:t}^c|z)$; where $f_{1:t}^c$ are the features of the glimpses other than the ones observed up to t and $f = f_{1:t} \cup f_{1:t}^c$. We implement a decoder D that synthesizes a feature map containing features of all glimpses in an image given the sample $z \sim q(z|h_t)$, i.e. $\tilde{f} = D(z)$. Let m_t be a binary mask with value **1** for the glimpses observed by the model up to t and **0** otherwise; hence, $f_{1:t} = m_t \odot f$, where $\odot$ is an element-wise multiplication. We assume a Gaussian likelihood and evaluate the log-likelihood in equation 2 using the mask m_t as follows.

$$-\log(p(f_{1:t}|z)) \propto \frac{1}{2} \sum \frac{|m_t \odot \tilde{f} - m_t \odot f|^2}{\sigma^2} + \log(\sigma^2) \quad (3)$$

Where σ is a model parameter. The BOED uses $\tilde{f}$ to find an optimal location to attend.

3.2 Bayesian Optimal Experiment Design (BOED)

The BOED evaluates the optimality of a set of experiments by measuring information gain in the interest parameter due to the experimental outcome [10]. In the context of hard attention, an experiment is to attend a location l and observe a corresponding glimpse $g = g(x, l)$. An experiment of attending a location l is optimal if it gains maximum information about the class label y. We can evaluate optimality of attending a specific location by measuring several metrics such as feature variance [22], uncertainty in the prediction [40], expected Shannon information [58]. For a sequential model, information gain $D_{\text{KL}}[p(y|g, l, h_{t-1})||p(y|h_{t-1})]$ is an ideal metric. It measures the change in the entropy of the class distribution from one time step to the next due to observation of a glimpse g at location l [6, 39].

The model has to find an optimal location to attend at time t before observing the corresponding glimpse. Hence, we consider an expected information gain (EIG) over the generating distribution of g. An EIG is also a measure of Bayesian surprise [25, 50].

$$EIG(l) = \mathbb{E}_{p(g|l,h_{t-1})} D_{KL}\big[p(y|g,l,h_{t-1})||p(y|h_{t-1})\big] \tag{4}$$

$$= \mathbb{E}_{p(f(l)|h_{t-1})} D_{KL}\big[p(y|f(l),h_{t-1})||p(y|h_{t-1})\big] \tag{5}$$

Where $f(l)$ are features of a glimpse located at l, i.e. $f(l) = F(g,l)$. Inspired by [21], we define $p(f(l)|h_{t-1})$ as follows.

$$p(f(l)|h_{t-1}) = \mathbb{E}_{q(z|h_{t-1})} p(f(l)|z) = \mathbb{E}_{q(z|h_{t-1})} \delta(D(z)(l)) = \mathbb{E}_{q(z|h_{t-1})} \delta(\tilde{f}(l)) \tag{6}$$

Here, $\delta(\cdot)$ is a delta distribution. As discussed in the section 3.1, the flow-based encoder S predicts the posterior $q(z|h_{t-1})$ and the decoder D predicts the feature map containing features of all glimpses in an image, $\tilde{f} = D(z)$. Combining equation 5 and equation 6 yields,

$$EIG(l) = \mathbb{E}_{q(z|h_{t-1})} D_{KL}\big[p(y|\tilde{f}(l),h_{t-1})||p(y|h_{t-1})\big] \tag{7}$$

To find an optimal location to attend at time t, the model compares various candidates for l_t. It predicts $EIG(l)$ for each candidate l and selects an optimal candidate as l_t, i.e. $l_t = \arg\max_l EIG(l)$. When the model is considering a candidate l, it uses $\tilde{f}(l)$ to calculate $\tilde{h}_t = R(h_{t-1}, \tilde{f}(l))$ and $p(y|\tilde{h}_t) = C(\tilde{h}_t)$. It uses the distribution $p(y|\tilde{h}_t) = p(y|\tilde{f}(l), h_{t-1})$ to calculate EIG in equation 7. We refer to $p(y|\tilde{h}_t)$ as the *lookahead* class distribution computed by anticipating the content at the location l ahead of time. In Figure 1, the dashed arrows show a lookahead step. Furthermore, to compute EIG for all locations simultaneously, we implement all modules of our model with convolution layers. The model computes EIG for all locations as a single activation map in a single forward pass. An optimal location is equal to the coordinates of the pixel with maximum value in the EIG map.

4 Experiments

Datasets. We evaluate our model on SVHN [42], CINIC-10 [12], CIFAR-10 [34], CIFAR-100 [34], and TinyImageNet [36] datasets. These datasets consist of real-world images categorized into 10, 10, 10, 100 and 200 classes respectively. Images in TinyImageNet are of size 64×64 and images in the remaining dataset are of size 32×32.

Training and Testing. We train our model in three phases. In the first phase, we pre-train modules F, R, and C with a random sequence of T glimpses using $\sum_{t=0}^{T-1} \mathcal{L}_{CE}(t)$ as a training objective. In the second phase, we introduce S and D. We pre-train S and D while keeping F, R, C frozen. Again, we use a random sequence of T glimpses and train S and D using $\sum_{t=0}^{T-1} \mathcal{L}_{PVAE}(t)$ training criterion. To produce the target f used in equation 3, we feed a complete image and a grid of all locations to the pretrained F, which computes features of all glimpses as a single feature map. Pre-training (F,R,C) and (S,D) separately ensures that the Partial VAE receives a stable target feature map f in equation 3. Finally, in the third phase, we fine-tune all modules end-to-end using the training objective $\mathcal{L} = \sum_{t=0}^{T-1} \alpha \mathcal{L}_{PVAE}(t) + \beta \mathcal{L}_{CE}(t)$, where α and β are hyperparameters. In the finetuning stage, we sample an optimal sequence of glimpses using the BOED framework. We use only one sample $z \sim q(z|h_t)$ to estimate the EIG map during the training, which leads to exploration. In the test phase, we achieve exploitation by using P samples of z to estimate the EIG accurately. The test procedure is shown in Algorithm 1. Refer to supp. material (SM) for more details.

Algorithm 1 Test procedure for the model shown in Figure 1

1: Randomly sample l_0; Capture g_0 at l_0, compute f_0, h_0 and $p(y|h_0)$
2: **for** $t \in \{1, \ldots, T-1\}$ **do**   ▷ T is the time budget
3:     Sample $z_i \sim q(z|h_{t-1})$ and predict $\tilde{f}^i$; $i \in \{0, \ldots, P-1\}$ ▷ P is the sample budget
4:     Compute $\tilde{h}_t^i$, $p(y|\tilde{h}_t^i)$ and $EIG = \frac{1}{P}\sum_i D_{\mathrm{KL}}[p(y|\tilde{h}_t^i)||p(y|h_{t-1})]$ ▷ equation 7
5:     $l_t = \arg\max\{EIG\}$
6:     Capture g_t at l_t; Compute f_t, h_t and $p(y|h_t)$
7: **end for**

Hyperparameters. We implement our model with a small number of parameters, and it runs for $T = 7$ time steps. It senses glimpses of size 16×16 overlapping with stride $n = 8$ for TinyImageNet and senses glimpses of size 8×8 overlapping with stride $n = 4$ for the remaining datasets. The Partial VAE predicts $\tilde{f}$ and EIG for a set of glimpses separated with stride equal to n. We do not allow our model to revisit glimpses attended in the past. We choose the hyperparameters α and β such that the two loss-terms contribute equally. The sample budget P is 20 for all experiments. Refer to SM for more details.

4.1 Baseline Comparison

We compare our model with four baselines in Figure 2(a-e). RAM is a state-of-the-art hard attention model that observes images partially and sequentially to predict the class labels [41]. We implement RAM using the same structure as our model. Instead of the Partial VAE, RAM has a controller that learns a Gaussian attention policy. Mnih et al. [41] minimize $\mathcal{L}_{CE}$ at the end of T steps. Following [37], we improve RAM by minimizing $\mathcal{L}_{CE}$ at all T steps. We refer to this baseline as RAM+. We also consider a baseline model that attends glimpses on random locations. The Random baseline does not have a controller or a Partial VAE. Our model and the three baselines described so far observe the image only partially through a series of glimpses. Additionally, we train a feed-forward CNN that observes the entire image to predict the class label.

For the SVHN dataset, the Random baseline outperforms RAM for initial time steps. However, with time, RAM outperforms the Random baseline by attending more useful glimpses. RAM+ outperforms RAM and the Random baselines at all time steps. We observe a different trend for non-digit datasets. RAM+ consistently outperforms RAM on non-digit datasets; however, RAM+ falls behind the Random baseline. In RAM and RAM+, the classifier shares latent space h_t with the controller, while the Random baseline dedicates an entire latent space to the classifier. We speculate that the dedicated latent space in the Random baseline is one of the reasons for its superior performance on complex datasets.

Our model consistently outperforms all attention baselines on all datasets. The performance gap between the highest performing baseline and our model reduces with many glimpses, as one can expect. Predicting an optimal glimpse-location is difficult for early time-steps as the models have access to minimal information about the scene so far. Compared to the highest performing baseline at $t = 1$, our model achieves around 10% higher accuracy on SVHN, around 5-6% higher accuracy on CIFAR-10 and CINIC-10, and around 2-3% higher accuracy on CIFAR-100 and TinyImageNet. Note that CIFAR-100 and TinyImageNet are more challenging datasets compared to SVHN, CINIC-10, and CIFAR-10. Similar to RAM and RAM+, the classifier and the Partial VAE share a common latent space in our model. Hence, our model achieves a lower gain over the Random baseline for complex datasets. We attribute the small but definite gain in the accuracy of our model to a better
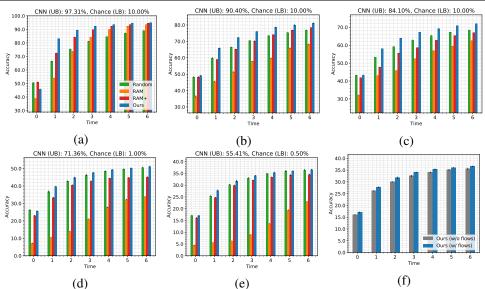
Figure 2: *Baseline Comparison on (a) SVHN (b) CIFAR-10 (c) CINIC-10 (d) CIFAR-100 (e) TinyImageNet. Ablation study on normalizing flows on (f) TinyImageNet.* (a-e) We compare various methods for $t = 0$ to 6. All results are averaged over ten different runs. Accuracy of a CNN and chance accuracy, displayed on top of each plot, serve as upper and lower bounds for the accuracy of glimpse-based methods. A CNN observes an entire image, whereas glimpse-based methods observe $< 43.75\%$ area of the image by $t = 6$. When compared with baseline methods, our method achieves the highest accuracy. (f) The use of normalizing flows improves accuracy.
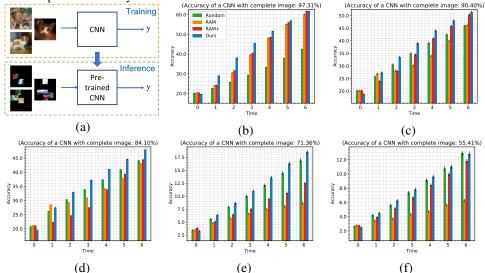

Figure 3: *Accuracy of a CNN when only the attended glimpses are observed.* (a) Experiment setup: we train a CNN using complete images and test it on masked images with only the glimpses attended by various methods made visible. Results on (b) SVHN (c) CIFAR-10 (d) CINIC-10 (e) CIFAR-100 (f) TinyImageNet. All results are averaged over three runs.
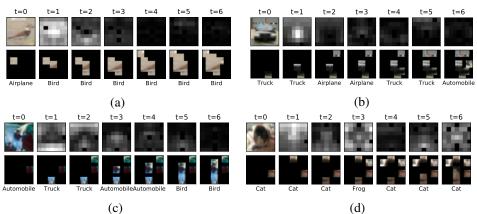
Figure 4: *Visualization of the EIG maps and the glimpses observed by our model on CIFAR-10 images.* The top rows in each plot show the entire image and the *EIG* maps for $t = 1$ to 6. The bottom rows in each plot show glimpses attended by our model. The model observes the first glimpse at a random location. Our model observes a glimpse of size 8×8. The glimpses overlap with the stride of 4, resulting in a 7×7 grid of glimpses. The EIG maps are of size 7×7 and are upsampled for the display. We display the entire image for reference; our model never observes the whole image. (a-c) success cases (d) failure case.

selection of glimpses.

The CNN has the highest accuracy as it observes a complete image. The accuracy of the CNN serves as the upper bound for the hard attention methods. Unlike CNN, the hard attention methods observe less than half of the image through small glimpses, each uncovering only 6.25% area of an image. On the CIFAR-10 dataset, the CNN predicts correct class labels for approximately 9 out of 10 images after observing each image completely. Remarkably, our model predicts correct class labels for 8 out of 10 images after observing less than half of the total area in each image.

4.1.1 Comparison of Attention Policies using a common CNN

Above, we compared the attention policies of various methods using their respective classifiers. However, each model attains different discriminative power due to different training objectives. While RAM, RAM+, and our model are trained jointly for two different tasks, i.e., classification and glimpse-location prediction or feature-synthesis, the Random baseline is trained for only one task, i.e., classification. Consequently, the Random baseline attains higher discriminative power than others and achieves high accuracy despite using a sub-optimal attention policy. To make a fair comparison of the attention policies irrespective of the discriminative power of the models, we perform the following experiment.

We mask all image regions except for the ones observed by the attention model so far and let the baseline CNN predict a class label from this masked image (see Figure 3). RAM+ consistently outperforms RAM, suggesting that the former has learned a better attention policy than the latter. As RAM+ is trained using $\mathcal{L}_{CE}$ at all time-steps, it achieves higher accuracy, and ultimately, higher reward during training with REINFORCE [41]. RAM and RAM+ outperform the Random baseline for the SVHN dataset. However, they fall short on natural image datasets. In contrast, our method outperforms all baselines with a significant margin on all datasets, suggesting that the glimpses selected by our model are more informative about the image class than the ones chosen by the baselines.

RAM and RAM+ struggle on images with many objects and repeated structure [51], as often the case with natural image datasets. For example, TinyImageNet includes many images with multiple objects (e.g. beer bottle), repeated patterns (e.g. spider web), and dispersed items (e.g. altar). Note that a random policy can perform competitively in such scenarios, especially when the location of various objects in an image is unknown due to partial observability. Yet, our method can learn policies that are superior to the Random baseline. Refer to SM for additional analyses.

4.2　Ablation study on Normalizing Flows

We inspect the necessity of a flexible distribution for the posterior $q(z|h_t)$ and, therefore, the necessity of normalizing flows in the encoder S. To this end, we model the posterior with a unimodal Gaussian distribution and let S output mean and diagonal covariance of a Gaussian. We do not use flow layers in this case. Figure 2(f) shows the result for TinyImageNet dataset. We observe that modeling a complex posterior using normalizing flows improves accuracy. Ideally, the Partial VAE should predict all possibilities of $\tilde{f}$ consistent with the observed region o_t. When the model observes a small region, a complex posterior helps determine multiple plausible feature maps. A unimodal posterior fails to cover all possibilities. Hence, the EIG estimated with the former is more accurate, leading to higher performance. Refer to SM for visualization of $q(z|h_t)$ estimated with and without normalizing flows.

4.3　Visualization

We visualize a few interesting examples of sequential recognition from the CIFAR-10 dataset in Figure 4. Refer to SM for additional examples. In Figure 4(a), activity in the EIG map reduces as the model settles on a class 'Bird'. In Figure 4(b), the model takes a long time to decide the true class of the image. Though incorrect, it predicts classes that are types of vehicles. After discovering parts like headlight and rear-view mirror, it predicts the true class 'Automobile' at $t = 6$. Figure 4(c) shows a difficult example. The model decides the true class 'Bird' after covering parts of the bird at $t = 5$. Notice a high amount of activity in the EIG maps up to $t = 5$ and reduced activity at $t = 6$. Figure 4(d) shows a failure case with enduring activity in EIG maps. Finally, observe that the EIG maps are often multimodal.

5　Conclusions

We presented a hard attention model that uses BOED to find the optimal locations to attend when the image is observed only partially. To find an optimal location without observing the corresponding glimpse, the model uses Partial VAE to synthesize the content of the glimpse in the feature space. Synthesizing features of unobserved regions is an ill-posed problem with multiple solutions. We use normalizing flows in Partial VAE to capture a complex distribution of unobserved glimpse features, which leads to improved performance. The synthesized features enable the model to evaluate and compare the expected information gain (EIG) of various candidate locations, from which the model selects a candidate with optimal EIG. The predicted EIG maps are often multimodal. Consequentially, the attention policy used by our model is multimodal. Our model achieves superior performance compared to the baseline methods that use unimodal attention policy, proving the effectiveness of multimodal policies in hard attention. When all models have seen only a couple of glimpses, our model achieves 2-10% higher accuracy than the baselines.

References

[1] Subutai Ahmad. Visit: a neural model of covert visual attention. In *Advances in neural information processing systems*, pages 420–427, 1992.

[2] Bogdan Alexe, Nicolas Heess, Yee W Teh, and Vittorio Ferrari. Searching for objects driven by context. In *Advances in Neural Information Processing Systems*, pages 881–889, 2012.

[3] Baptiste Angles, Simon Kornblith, Shahram Izadi, Andrea Tagliasacchi, and Kwang Moo Yi. Mist: Multiple instance spatial transformer network. *arXiv preprint arXiv:1811.10725*, 2018.

[4] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.

[5] Jimmy Ba, Russ R Salakhutdinov, Roger B Grosse, and Brendan J Frey. Learning wake-sleep recurrent attention models. In *Advances in Neural Information Processing Systems*, pages 2593–2601, 2015.

[6] José M Bernardo. Expected information as expected utility. *the Annals of Statistics*, pages 686–690, 1979.

[7] Nicholas J Butko and Javier R Movellan. I-pomdp: An infomax model of eye movement. In *2008 7th IEEE International Conference on Development and Learning*, pages 139–144. IEEE, 2008.

[8] Nicholas J Butko and Javier R Movellan. Optimal scanning for faster object detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2009.

[9] Weiwei Cai and Zhanguo Wei. Piigan: Generative adversarial networks for pluralistic image inpainting. *IEEE Access*, 8:48451–48463, 2020.

[10] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[11] Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. Differentiable patch selection for image recognition. *arXiv preprint arXiv:2104.03059*, 2021.

[12] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

[13] Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. Learning where to attend with deep architectures for image tracking. *Neural computation*, 24(8):2151–2184, 2012.

[14] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[15] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, pages 7511–7522, 2019.

[16] Gamaleldin Elsayed, Simon Kornblith, and Quoc V Le. Saccader: improving accuracy of hard attention models for vision. In *Advances in Neural Information Processing Systems*, pages 702–714, 2019.

[17] SM Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems*, pages 3225–3233, 2016.

[18] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713, 2018.

[19] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

[20] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, 2015.

[21] William Harvey, Michael Teng, and Frank Wood. Near-optimal glimpse sequences for improved hard attention neural network training. *arXiv preprint arXiv:1906.05462*, 2019.

[22] Sheng-Jun Huang, Miao Xu, Ming-Kun Xie, Masashi Sugiyama, Gang Niu, and Songcan Chen. Active feature acquisition with supervised matrix completion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1571–1579, 2018.

[23] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[24] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[25] Laurent Itti and Pierre F Baldi. Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, pages 547–554, 2006.

[26] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.

[27] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[28] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[29] Angelos Katharopoulos and Francois Fleuret. Processing megapixel images with deep attention-sampling models. In *International Conference on Machine Learning*, pages 3282–3291, 2019.

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[31] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.

[32] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.

[33] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[34] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[35] Hugo Larochelle and Geoffrey E Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *Advances in neural information processing systems*, pages 1243–1251, 2010.

[36] Fei-Fei Li, Andrej Karpathy, and Justin Johnson. Tiny imagenet challenge. http://cs231n.stanford.edu/2016/project.html, 2016.

[37] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu. Dynamic computational time for visual attention. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1199–1209, 2017.

[38] Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

[39] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*, 2018.

[40] Prem Melville, Maytal Saar-Tsechansky, Foster Provost, and Raymond Mooney. Active feature-value acquisition for classifier induction. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 483–486. IEEE, 2004.

[41] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[42] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *Advances in neural information processing systems*, 2011.

[43] Bruno A Olshausen, Charles H Anderson, and David C Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700–4719, 1993.

[44] Lucas Paletta, Gerald Fritz, and Christin Seifert. Q-learning of sequential attention for visual object recognition from informative local descriptors. In *Proceedings of the 22nd international conference on Machine learning*, pages 649–656, 2005.

[45] Athanasios Papadopoulos, Paweł Korus, and Nasir Memon. Hard-attention for scalable image classification. *arXiv preprint arXiv:2102.10212*, 2021.

[46] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[47] Marc'Aurelio Ranzato. On learning where to look. *arXiv preprint arXiv:1405.5488*, 2014.

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 91–99, 2015.

[49] Raymond D Rimey and Christopher M Brown. Controlling eye movements with hidden markov models. *International Journal of Computer Vision*, 7(1):47–65, 1991.

[50] Philipp Schwartenbeck, Thomas FitzGerald, Ray Dolan, and Karl Friston. Exploration, novelty, surprise, and free energy minimization. *Frontiers in psychology*, 4:710, 2013.

[51] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. *arXiv preprint arXiv:1412.7054*, 2014.

[52] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015.

[53] Charlie Tang, Nitish Srivastava, and Russ R Salakhutdinov. Learning generative models with visual attention. In *Advances in Neural Information Processing Systems*, pages 1808–1816, 2014.

[54] Burak Uzkent and Stefano Ermon. Learning when and where to zoom with deep reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12345–12354, 2020.

[55] Miao Wang, Yu-Kun Lai, Yuan Liang, Ralph R Martin, and Shi-Min Hu. Biggerpicture: data-driven image extrapolation using graph matching. *ACM Transactions on Graphics*, 33(6), 2014.

[56] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2019.

[57] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[58] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10561–10570, 2019.

[59] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[60] Lingzhi Zhang, Jiancong Wang, and Jianbo Shi. Multimodal image outpainting with regularized normalized diversification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3433–3442, 2020.

[61] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5741–5750, 2020.

[62] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

[63] Yin Zheng, Richard S Zemel, Yu-Jin Zhang, and Hugo Larochelle. A neural autoregressive approach to attention-based recognition. *International Journal of Computer Vision*, 113(1):67–79, 2015.