Near-Optimal Glimpse Sequences for Training Hard Attention Neural Networks

William Harvey[1], **Michael Teng**[2], **Frank Wood**[1]
[1]University of British Columbia, [2]University of Oxford

1 Introduction

Hard visual attention mechanisms control the allocation of limited sensor resources to observe the environment. Their most prominent example is the human visual system where photoreceptors cover a 210 degree field of view [14] and are most dense in the centre of the retina [2]. Eye movements, or saccades, are necessary to apply these sensors to the most salient parts of a scene. Inspired by this, artificial hard attention mechanisms can solve certain tasks using orders of magnitude less computation and sensor bandwidth than the alternatives [8, 11]. However, their training is a reinforcement learning problem and has proven to be difficult for many real-world applications.

We introduce methodology from Bayesian optimal experimental design [3] for generating information-theoretically near-optimal sequences of glimpse locations for hard attention neural networks. We use these sequences to partially supervise the training of such neural networks, and show improvements in training speed and considerable differences in the learned policies. The faster and lower-variance training we enable makes our approach particularly applicable to architecture search [5] for hard attention networks. We use this to find a hard attention architecture which achieves higher accuracy than low-power convolutional neural network (CNN) architectures which process the full-image.

2 Hard attention

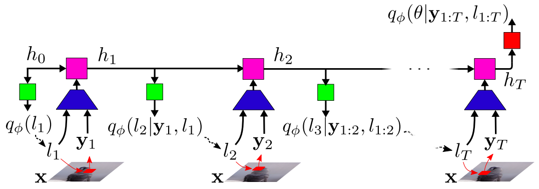

Figure 1: The hard attention network we consider. At each $t = 1, \ldots, T$, the location network (green) outputs a distribution from which l_t is sampled. A glimpse of the image at l_t is fed into the glimpse embedder (blue), along with l_t. The resulting embedding is input into the RNN (pink), updating the hidden state to h_t. At $t = T$, the classifier outputs a classification distribution.

Throughout this paper, we consider image classification: given an image, $\mathbf{x}^{(\mathbf{i})}$, we infer its label, $\theta^{(i)}$. For brevity, we shall from now on refer to the image as $\mathbf{x}$ and label as θ, with the index implicit. We consider the model of recurrent hard attention shown in Figure 1, which corresponds closely to that of Mnih et al. [10]. At each non-final time t, the neural network outputs a distribution over possible image *locations* to attend to. We denote the location sampled from this l_t. A contiguous square of pixels, which we call a *glimpse* and denote $\mathbf{y}_t = f_{\text{glimpse}}(\mathbf{x}, l_t)$, is extracted from the image at this location and fed into the neural network. At some constant final time T, the network outputs a classification distribution, $q_\phi(\theta|\mathbf{y}_{1:T}, l_{1:T})$. A loss calculated with this output is used to optimise all network parameters, ϕ. However, it is typically impossible to differentiate the glimpse, $\mathbf{y}_t$, with respect to l_t, and so the attention mechanism cannot be trained by standard backpropagation. REINFORCE [10, 15] is commonly used instead, with a sparse reward given by task success (e.g. a 0-1 classification loss). The resulting high variance gradient estimates have made it difficult to scale hard attention to complex datasets. Some previous attempts have attempted to avoid the problem of learning long glimpse sequences by computing glimpse locations with a (possibly downsampled) version of the full image [4, 8, 13]; using large glimpses which can

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

Algorithm 1 Experimental design pipeline

```
1:  procedure BOED($\mathbf{y}_{1:t-1}, l_{1:t-1}$)
2:      for $n \leftarrow 1, \dots, N$ do
3:          $\mathbf{x}^{(n)} \sim r_{\text{img}}(\mathbf{x}|\mathbf{y}_{1:t-1}, l_{1:t-1})$
4:      for $l_t \in L$ do
5:          for $n \leftarrow 1, \dots, N$ do
6:              $\mathbf{y}_t^{(n)} \leftarrow f_{\text{glimpse}}(\mathbf{x}^{(n)}, l_t)$
7:              $\text{PE}_{l_t}^{(n)} \leftarrow \mathcal{H}\left[g(\theta|\mathbf{y}_{1:t}^{(n)}, l_{1:t})\right]$
8:          $\text{EPE}_{l_t} \leftarrow \frac{1}{N}\sum_{n=1}^{N} \text{PE}_{l_t}^{(n)}$
9:      $l_t^* \leftarrow \text{argmin}_{l_t} \text{EPE}_{l_t}$
10:     return $l_t^*$
```

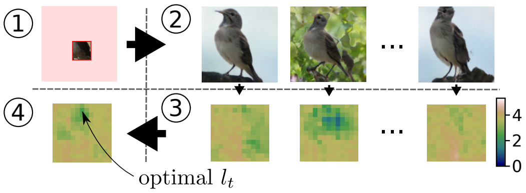

Figure 2: Experimental design for $t = 2$, corresponding to Alg. 1. Panel 1 shows the previous glimpse, $\mathbf{y}_1$, at l_1. Panel 2 shows image completions, sampled on line 3 of Alg. 1. Line 7 estimates the posterior entropy for each l_2 and sampled image (panel 3). Line 8 averages these to give the expected posterior for each location (panel 4). The minimum is returned.

solve the task in a single time-step (e.g. several 96 × 96 pixel patches at different resolutions [13]); or taking many glimpses simultaneously [8]. While these techniques have achieved impressive results, they make impossible some of the gains in computational efficiency promised by hard attention, indicating a problematic trade-off between efficiency at inference time and the practicality of training.

3 Bayesian optimal experimental design

We consider the allocation of attention as designing an experiment to infer some parameter, θ. To clarify the connection, we denote the design of the experiment l and the observed outcome $\mathbf{y} \sim p(\mathbf{y}|l, \theta)$. In Bayesian experimental design [3, 6], an optimal design is considered to be one which, on expectation, leads to the least possible uncertainty about θ. This is quantified by the expected Shannon entropy in the posterior over θ, $\text{EPE}(l) = \mathbb{E}_{p(\mathbf{y}|l)}[\mathcal{H}[p(\theta|\mathbf{y}, l)]]$. We consider sequential experimental design, where a series of experiments are performed and the design of each can depend on all previous experiments. We therefore select l_t at each step t to minimise the expected posterior entropy, given the designs and outcomes of previous experiments, $l_{1:t-1}$ and $\mathbf{y}_{1:t-1}$. Both the posterior over θ and the distribution over $\mathbf{y}_t$ are conditioned on these, giving

$$\text{EPE}_{\mathbf{y}_{1:t-1}, l_{1:t-1}}(l_t) = \mathbb{E}_{p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, l_{1:t})}\left[\mathcal{H}\left[p(\theta|\mathbf{y}_{1:t}, l_{1:t})\right]\right]. \tag{1}$$

In the computer vision setting which we consider, we do not typically have an explicit joint distribution over images and labels (i.e. $p(\theta, \mathbf{x})$). We instead use the dataset to create approximations to the distributions required to evaluate the expected posterior entropy (Equation 1):

- $g(\theta|\mathbf{y}_{1:t}, l_{1:t}) \approx p(\theta|\mathbf{y}_{1:t}, l_{1:t})$. This is parameterised by a convolutional neural network (CNN) which receives the embedding of $\mathbf{y}_{1:t}$ and $l_{1:t}$ shown in Figure 4 and outputs a categorical distribution over θ. The CNN is trained to minimise an expectation of $D_{\text{KL}}[p(\theta|\mathbf{y}_{1:t}, l_{1:t})||g(\theta|\mathbf{y}_{1:t}, l_{1:t})]$ over t, $\mathbf{y}_{1:t}$ and $l_{1:t}$, which are distributed according to some uniform prior distribution.
- $r_{\text{glimpse}}(\mathbf{y}_t|\mathbf{y}_{1:t-1}, l_{1:t}) \approx p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, l_{1:t})$. This is sampled from by sampling a full image, $\mathbf{x}$, from $r_{\text{img}}(\mathbf{x}|\mathbf{y}_{1:t-1}, l_{1:t-1}) \approx p(\mathbf{x}|\mathbf{y}_{1:t-1}, l_{1:t-1})$ and extracting $\mathbf{y}_t$ from $\mathbf{x}$ at location l_t, effectively marginalising out all pixels not in l_t. Sampling from $r_{\text{img}}(\mathbf{x}|\mathbf{y}_{1:t-1}, l_{1:t-1})$ is essentially

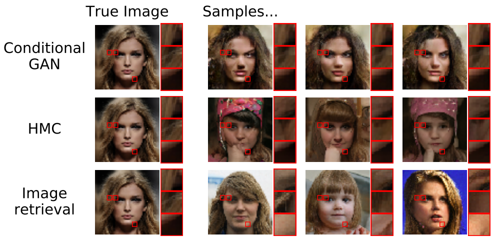

Figure 3: Image samples from various techniques, conditioned on the glimpses in the left column.

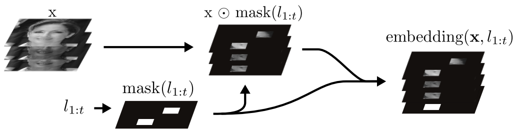

Figure 4: Embedding of $\mathbf{y}_{1:t}$ and $l_{1:t}$. A mask is created which is zero at unobserved pixels. It is applied to the image and then concatenated as an additional channel.

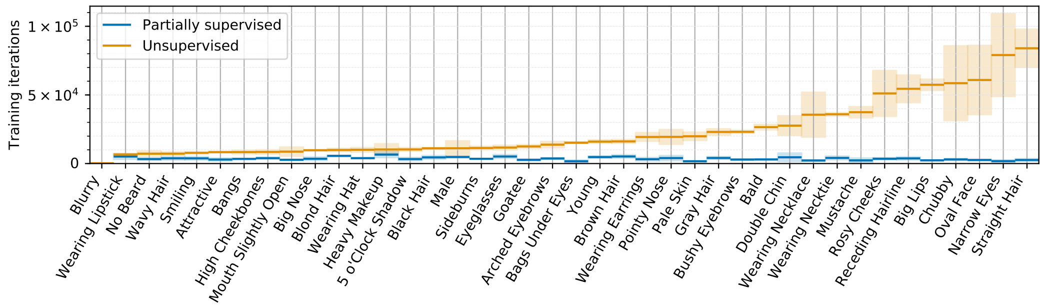

Figure 5: Training iterations before reaching within 0.01 of the best achieved validation accuracy, with and without supervision sequences.

image completion. Figure 3 shows several techniques we considered for this: a GAN mapping from $\mathbf{y}_{1:t}$ and $l_{1:t}$ to a full image; HMC in the latent space of a pretrained GAN, conditioned on the observations under a Gaussian likelihood; and probabilistic image retrieval from a synthetic database of 1.5×10^6 images. We used probabilistic image retrieval as only it gave sufficient sample diversity.

Using these approximations and an N-sample Monte Carlo estimate of the expectation gives

$$\text{EPE}_{\mathbf{y}_{1:t-1},l_{1:t-1}}(l_t) \approx \frac{1}{N}\sum_{n=1}^{N} \mathcal{H}\big[g(\theta|\{\mathbf{y}_{1:t-1}, \mathbf{y}_t^{(n)}\}, l_{1:t})\big] \tag{2}$$

where $\mathbf{y}_t^{(n)} \sim r_{\text{glimpse}}(\mathbf{y}_t|\mathbf{y}_{1:t-1}, l_{1:t})$. The optimal next attention location is found with a grid search over l_t. This process is summarised in Algorithm 1 and Figure 2. It is done for each $t = 1, \dots, T$ to generate a sequence of locations.

4 Training with supervision

We use the generated glimpse sequences to partially supervise the training of a hard attention network. Taken together, the training procedures for supervised and unsupervised examples can be interpreted as maximising the joint log-likelihood, $\mathbf{L}$, of the class labels and glimpse sequences,

$$\mathbf{L} = \sum_{i \in \text{sup.}} \overbrace{\log q_\phi(\theta^i, l^i_{1:T}|\mathbf{x}^i)}^{\text{supervised objective}} + \sum_{i \in \text{unsup.}} \overbrace{\log q_\phi(\theta^i|\mathbf{x}^i)}^{\text{unsupervised objective}} . \tag{3}$$

where ‘sup’ is the subset of training indices with supervision sequences, and ‘unsup’ is the subset without. On unsupervised examples, a lower bound on the unsupervised objective is maximised by minimising a standard hard attention loss [1, 10] with REINFORCE gradient estimates. For examples with supervision, the network's glimpse locations are fixed to those in the supervision sequence and it is trained to maximise the log-likelihood of both the class label and the supervision locations.

5 Experiments

Improved training We create 600 near-optimal sequences for each of the 40 classification tasks on the CelebA-HQ dataset [7], and compare training with and without partial supervision. The hard attention networks use five glimpses, each of 16×16 pixels. Figure 5 shows the effect on the number of iterations taken before convergence. We find that supervision with near-optimal sequences reduces the number of iterations by a factor of 6.8 on average compared to training without supervision. It also gives a more than $5\times$ average reduction in variance, and a mean increase in accuracy of 0.4%.

Qualitative comparison of attention policies Figure 6 contrasts the attention policies of the networks trained with and without supervision. While both forms appear to learn reasonable first glimpse locations, supervision is required to learn to make use of the later time steps.

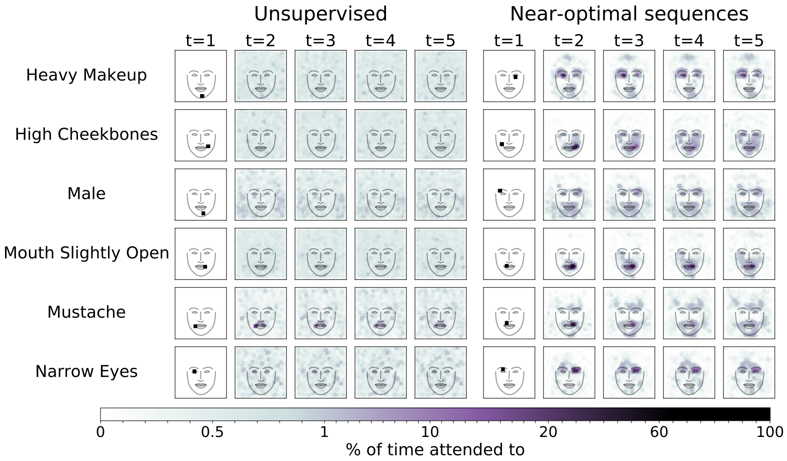

Figure 6: Glimpse locations on the test set for several classification tasks. The colour of each pixel at each time step corresponds to the proportion of glimpses that attended to a location covering the pixel, averaged over the test set. The face outlines are given by averaging coordinates from a face detector. The networks trained without supervision typically attend to a single location at the first time step, but very broad distributions at later time steps. The network trained with supervision also attends to a single initial location but learns the salient regions at later time steps as well.

Comparison with CNN baselines Figure 7 compares the accuracy achieved by a hard attention network against various CNNs. We considered CNNs with the ShuffleNet [9] and MobileNet [12] architectures, which are designed to use a minimal number of floating point operations (FLOPs). We varied the number of FLOPs in each by searching over scalar multipliers, c, for the number of channels in each layer and different resolutions, r, to downsample the input to. Only the best performing architectures are shown. The hard attention network uses adaptive stopping to trade off computation and accuracy. This is implemented by an additional network module which maps from the hidden state to an estimate of the expected increase in accuracy for a prediction made after the next glimpse, versus a prediction made now. The network stops if this is below some threshold, which is varied along the line. Additionally, we found that training with either no supervision or with supervision by near-optimal sequences led to overfitting on the small datasets we use. We therefore used a heuristic to create supervision sequences involving sampling a location independently at each time step from a categorical distribution with log probability proportional to each location's negative expected posterior entropy from the first glimpse (i.e. $\text{EPE}_{\emptyset,\emptyset}(l_t)$). The hard attention architecture can be seen to improve on the accuracy achievable by CNNs with less than 0.7 MFLOP, but future work is required to investigate if hard attention is beneficial for higher computational budgets.

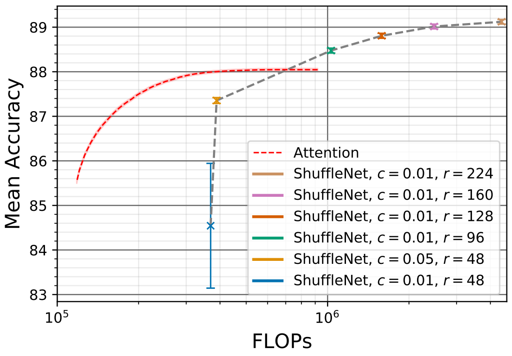

Figure 7: Accuracy versus computation for various low-power CNNs and a hard attention architecture with a variable stopping time.

6 Discussion and conclusions

We have demonstrated a novel Bayesian experimental design pipeline for generating near-optimal glimpse sequences. This pipeline can significantly improve hard attention training, which is particularly beneficial for applications such as neural architecture search. Our framework could also be extended to tasks such as question answering where the latent variable of interest is more richly structured.

References

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

[2] Mark F Bear, Barry W Connors, and Michael A Paradiso. *Neuroscience*, volume 2. Lippincott Williams & Wilkins, 2007.

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

[4] Gamaleldin F Elsayed, Simon Kornblith, and Quoc V Le. Saccader: Improving accuracy of hard attention models for vision. *arXiv preprint arXiv:1908.07644*, 2019.

[5] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018.

[6] Adam Foster, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Tom Rainforth, and Noah Goodman. Variational estimators for bayesian optimal experimental design. *arXiv preprint arXiv:1903.05480*, 2019.

[7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[8] Angelos Katharopoulos and Francois Fleuret. Processing megapixel images with deep attention-sampling models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3282–3291, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/katharopoulos19a.html`.

[9] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018.

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

[11] Ronald A Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000.

[12] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

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

[14] Harry Moss Traquair. *An introduction to clinical perimetry*. Mosby, 1949.

[15] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.