Near-Optimal Glimpse Sequences for Training Hard Attention Neural Networks

William Harvey, Michael Teng, Frank Wood

wsgh@cs.ubc.ca

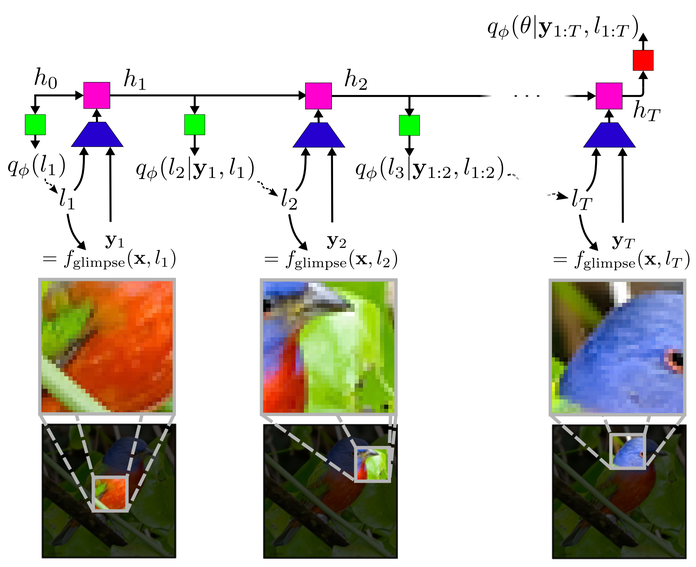
Hard attention

Advantages:

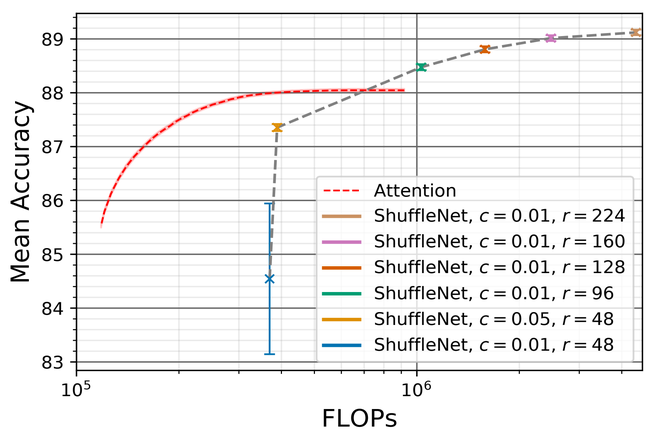
- process only small parts of an image
- can achieve better performance than a CNN when matched for FLOPs

Disadvantages:

- loss is non-differentiable w.r.t. neural network parameters
- training is difficult, especially for long glimpse sequences



Hard attention network architecture along with visualisation of selected glimpses. As in Mnih et al. [1], the hard attention network has an RNN core (pink), glimpse embedder (blue), location network (green) and classifier (red).

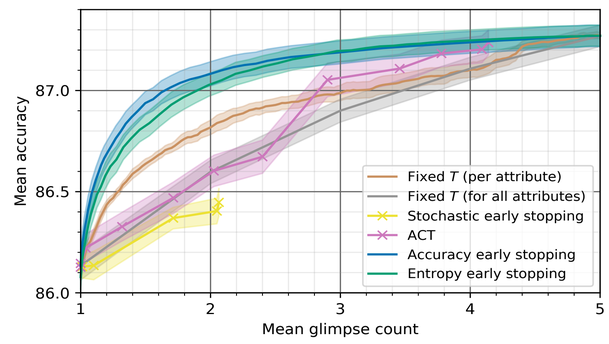


Accuracy for hard attention network compared to best-performing low-power CNNs for image classification on CelebA-HQ dataset. The hard attention network uses adaptive stopping to allow a trade-off between accuracy and computation. It is trained using a variation of our form of partial supervision.

Contributions

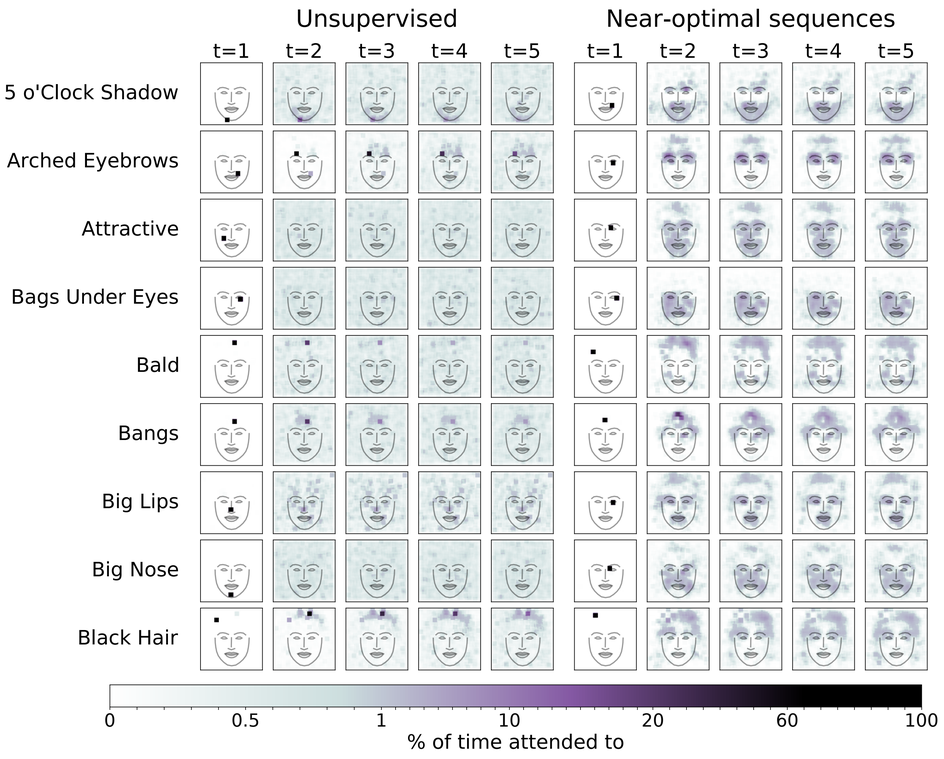
- Bayesian experimental design pipeline to generate sequences of near-optimal glimpse locations for attention neural networks
- Supervision with such sequences improves hard attention training
- Introduced a new form of adaptive stopping for hard attention

Adaptive stopping

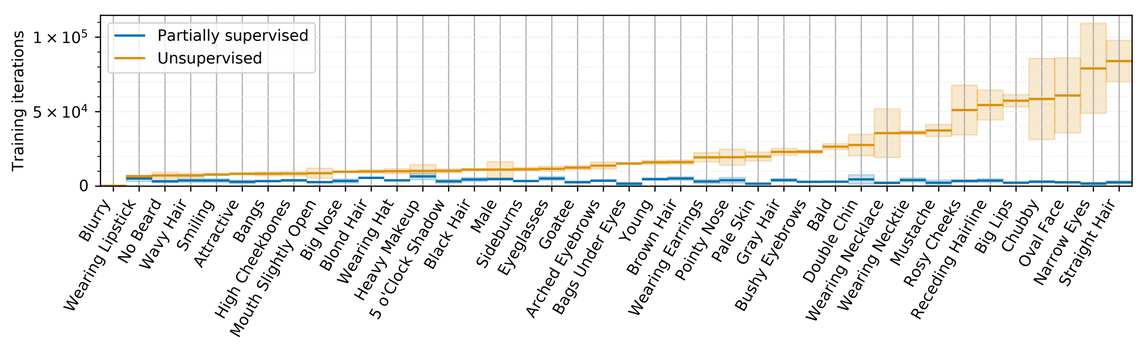


The hard attention network is trained to predict the expected information gain before each glimpse, and stops if this is below some threshold. This threshold is varied in the plot to trade off error and computational cost, achieving higher accuracy for a given budget than all baselines, including Adaptive Computation Time [2].

Improved training with partial supervision



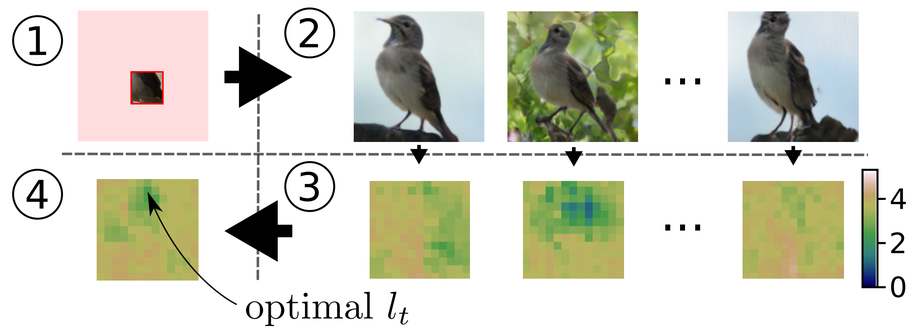
Locations attended to on the test set for attention networks trained with and without supervision. The colour of each pixel corresponds to the proportion of time it is attended to. The network trained without supervision typically learns a good first glimpse location, but does not learn where to attend at later time steps. Training with supervision yields a reasonable attention policy at every time step.



Number of training iterations until near the best validation accuracy for networks trained both with and without supervision. Supervision gives 0.4% higher mean test accuracy whilst making training almost 7× faster.

Bayesian optimal experimental design

Select the $\overbrace{\text{parameters}}^{\text{location}}$ of an $\overbrace{\text{experiment}}^{\text{glimpse}}$ to maximise the information about θ gained from the $\underbrace{\text{outcome.}}_{\substack{\text{observed} \\ \text{image patch}}}$

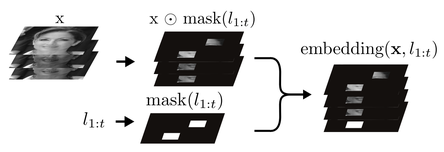


Overview of experimental design for next glimpse location.

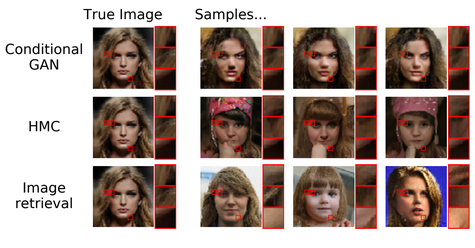
Select l_t to minimise the *expected posterior entropy* given previous glimpses:

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

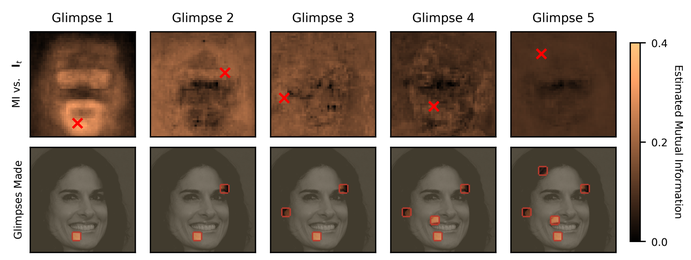
We use learned approximations of both $p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, l_{1:t})$ and $p(\theta|\mathbf{y}_{1:t}, l_{1:t})$.



Our novel embedding of multiple glimpses for the attentional variational posterior CNN. This allows us to approximate $p(\theta|\mathbf{y}_{1:t}, l_{1:t})$.



Various image completion techniques considered for approximating $p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, l_{1:t})$.



Top row: estimate of mutual information of each possible attention location with the image label. The maximum is selected as the next glimpse location. **Bottom row:** Glimpses taken.

References

[1] Mnih V, Heess N, Graves A. Recurrent models of visual attention. In Advances in neural information processing systems 2014 (pp. 2204-2212).

[2] Graves A. Adaptive computation time for recurrent neural networks. arXiv preprint arXiv:1603.08983. 2016 Mar 29.

PλAI

THE UNIVERSITY OF BRITISH COLUMBIA

UNIVERSITY OF OXFORD