We thank the reviewers for their feedback. We will add the clarification sentences requested. Below are our responses:

R1: no ablation study. We conducted the requested study. Our results show that accuracy is poor if we ablate the attention network or the Saccader cell. We will add the analysis to Fig Supp.3 and state: "The attention network allows the Saccader to better plan for locations to attend to as it has wider receptive field (RF). Note that this wider RF does not affect the interpretability of the model as the classification path RF is still limited to 77x77. Furthermore, removing the Saccader cell (i.e. using the BagNet-77-lowD ordered logits policy) yields poor results compared to the Saccader."

R2: computational cost and/or parameter counts. We will add a table with the parameters count of all models. In summary, our model has 88,554,409 parameters, which is nearly twice the DRAM parameters.

R2: how the final prediction is made. The prediction is made using the averaged logits across the locations.

R2: Does the "location network" ... refer to the attention network, the 1-by-1 conv and the Saccader cell? Yes.

R2 and R4: other datasets. ImageNet is an extremely large and diverse dataset that contains both coarse- and fine-grained class distinctions. To achieve good performance, a method must not only distinguish among superclasses, but also e.g. among the >100 fine-grained classes of dogs. Moreover, most natural image datasets have some class overlap with ImageNet; few datasets are entirely disjoint. Recent work has suggested that ImageNet is a representative benchmark; Kornblith et al. CVPR 2019 (https://arxiv.org/abs/1805.08974) showed that accuracy on ImageNet predicts accuracy on other natural image classification datasets.

R4: This design would fail to apply on ... pedestrain detection… cancer classification. We agree that it would be nice to apply our method to pedestrian detection and cancer classification. However, we want to stress that natural image classification is an important computer vision problem. Recent advances in computer vision started with classification tasks on ImageNet (e.g. Krizhevsky et al., 2012) and then future research extended these methods to other domains. We will add sentences to the results to better motivate the task.

R4: NASNet ... unclear whether this improve comes from the increased model capacity or higher input resolution. We conducted an experiment on ImageNet 224 and using the Saccader-NASNet model. Our results show that the accuracy is better than the Saccader model alone but worse than the Saccader-NASNet model on the high resolution ImageNet 331 (we added this experiment to Figure 6). This finding demonstrates that the accuracy benefits from both the increased capacity as well as the higher input resolution.

R4: pre-training ... This step introduces a strong bias. We agree with the reviewer that pre-training introduces bias. However, we find that this bias is helpful in getting a better final policy. In Figure Supp3, we show that reinforcement learning after pretraining location network enhances accuracy compared to starting learning without this pretraining step. Also, note that the final learned policy is different than the pretraining target policy (i.e., the ordered logits policy). As we show in Figure 3 and 5a, the final learned policy performs much better. Just as SGD biases neural network training toward solutions that generalize well, we find that the pretraining alters the training trajectory in a way that produces a better-performing model.

R4: baseline ... such as Class Activation Map (CAM). In this work, we are concerned with models with hard visual attention. CAM (Zhou et al. 2016) and similar interpretability methods try to provide an explanation of the model decision in a way that relies on a heuristic (e.g., that the spatial localization of features should be preserved in the final feature map) rather than explicitly constraining how the network processes its input. These methods are fragile (see Hooker et al. 2019), and the model's final decision may nonetheless rely on information provided by features with small weights (see Jain and Wallace, 2019). Models with hard attention take a different approach by using a controller that selects parts of the input to be processed by the network, which provides interpretability by design. In our work, the representation network may be regarded as a network to construct CAM, with a guarantee that the receptive field is limited. We will add a citation to Zhou et al. 2016.

R1: no weaknesses ... have been noted. We will add: "Although Saccader outperforms other hard attention models, it still lags behind state-of-the-art feedforward models in terms of accuracy. Future research may extend the Saccader model to achieve even better classification performance while maintaining the interpretability of model decisions."

R1: "what" and "where". We will add: "These are analogous to the ventral ("what") and dorsal ("where") pathways that are involved in object recognition and localization, respectively in human vision (Goodale and Milner 1992)."

Other improvements. We computed errorbars (mean $\pm$ SD) for all plots. In the DRAM, we limited the high resolution to classification and the (high, mid and low) resolutions to initialize the location LSTM state, which encouraged better location exploration. We also extended the DRAM pretraining to two stages on wide and limited receptive fields (120 epochs each), and doubled the LSTM layers size. Despite these changes, the Saccader was still better than the DRAM. We improved the ResNet model accuracy in Fig 4. We corrected the Sobel and Canny baselines plots (accuracy remains poor). Since Canny and Sobel results are similar, we only included the Sobel results to improve the presentation.