Reconstruction-as-feedback serves as an effective attention mechanism for object recognition and grouping

Seoyoung Ahn[1], Hossein Adeli[1], Gregory J. Zelinsky[1,2]
[1]Stony Brook University, Department of Psychology, Stony Brook, NY, [2]Stony Brook University, Department of Computer Science, Stony Brook, NY

Humans are able to partially reconstruct visual information, as evidenced by our ability to imagine and dream, yet it is debated whether a reconstruction process is functionally used for online visual perception. Here we built an iterative encoder-decoder system which generates hypotheses about an object's shape and appearance — made explicit as an object reconstruction. The model uses these object reconstructions as a top-down attentional bias for efficiently routing relevant spatial and feature information of the object.

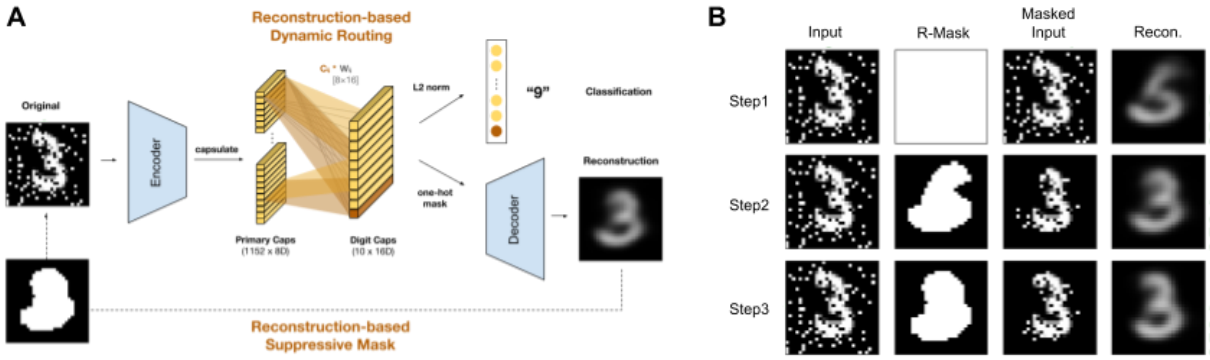

Figure1. (A) Model Architecture. Our model iteratively recognizes an image using both local and long-range reconstruction-generated attentional feedback. (B) Model input and output for every iteration step. The object reconstruction from the model's most likely object hypothesis changes from 5 to 3 through 3 steps (Groundtruth class is 3).

Figure 1A shows our model architecture. Our reconstruction-based attention operates on two levels. First, the model has a long-range projection that inhibits irrelevant spatial regions based on the mask generated from the most likely object reconstruction. Second, the model dynamically changes its feature routing weights through local recurrence, where part-whole connection is modulated based on the reconstruction error for each hypothesized object (represented as a slot). This formulation loosely implements biased-competition theory, where the reconstruction error biases a competition between object slots for the visual parts. We tested this model using the challenging out-of-distribution digit recognition task, MNIST-C, where 15 different types of corruption (e.g., noise, blur, occlusion, affine transformation etc) are applied to handwritten digit images. Our model outperformed other models that are especially designed to deal with generalization, e.g., capsule network, adversarially trained models. Ablation studies also confirmed that using reconstruction-based attention significantly increases the model's robustness (>12%) compared to when it is not used. Finally, we compared our model behavior with actual human recognition responses on the same testing dataset. Although the results are mixed across different types of corruption, our model generally showed more human-like behavior (e.g., response time, the types of error made) compared to a simple CNN. This was especially true when complex grouping operations are required (e.g., when the digit is occluded with randomly generated splotches or superimposed with a zigzag pattern).