

**A Foveated Vision-Transformer Model for Scene Classification**

Aditya Jonnalagadda¹, Miguel P. Eckstein¹²

¹Dept. of Electrical & Computer Engineering, UC Santa Barbara, CA, USA, ²Dept. of Psychological & Brain Sciences, UC Santa Barbara, CA, USA

**INTRODUCTION**

- Humans can rapidly categorize scenes (R. Vanluitten & S. Thorpe, 2001), even using peripheral vision (Larson & Loschky, 2009).
- Various computational models have been proposed for rapid scene categorization in terms of low-level properties such as spatial envelopes (Oliva & Torralba, 2001) and texture summary statistics (ITM, Rosenholtz et al., 2012). Yet, these models do not explicitly model the foveal properties of the visual system or the interaction between eye movements and the scene categorization task.
- We propose a model with a foveated visual system and eye movements that can predict the dependence of human categorization performance across fixations.
- Full-resolution models do not model the foveated properties of the visual system.
- Thirty categories from the places365 dataset (Zhou et al., 2018) are used to create the dataset for the scene classification experiment.

**HUMAN EXPERIMENT**

- A gaze-contingent display was used to randomly interrupt the display after 1, 2, 3, or 4 fixations.
- The task is to classify each image into one of the 30 categories.
- Each of the twenty-two participants classified 240 images.
- Sixty images are present for each condition.
- Image subtended 22.7 x 22.7 degrees visual angle.
- Initial fixation was at the bottom-center or top-center of the images.

**MODEL**

- The model (Jonnalagadda et al., 2021) combines square pooling regions with the computer vision-Transformer architecture (Dosovitskiy et al., 2020, Touvron et al., 2020) and makes multiple fixations to maximize classification using the technique of self-attention (Parikh et al., 2016, Bhadiraju et al., 2015).
- For computational economy, the Foveation module uses squared pooling regions to convert the full-resolution features into pooled features.
- Accumulator uses the self-attention weights of the last layer to decide the next fixation location.
- The highlighted block is executed at each time step.

**RESULTS**

- There is no significant improvement in performance after the 2nd fixation (Δ correct categorization=0.015; p=0.4729), unlike performance for object search (Koehler and Eckstein, 2017).
- The model correctly predicts modest classification improvements for free-viewing fixations (Δ=0.011).

**RESULTS (cont.)**

- We focus on a single eye movement condition (highest accuracy increase across fixations)
- Three algorithms for the model: Random fixation, Guided fixation using self-attention weights, Fixation guided by human data
- Accuracy of the model guided by human eye movements or self-attention model is higher than the random fixation.
- The model with self-attention is in greater agreement (class labels) with human observers than the random fixation model.
- The model’s agreement with human decisions is further improved by using the fixations made by humans (rather than the self-attention fixations).

**CONCLUSIONS**

- Our findings suggest that human categorization of scenes within a single fixation can be explained by the spatially global distribution of the visual information in the scene and their availability even through the bottlenecks of the visual periphery.
- Guided fixations by the self-attention model resulted in higher accuracy and decision agreement than random fixations but lower than a model that utilized measured human fixations.
- For very brief presentations and a single fixation, a foveated model better-predicted the observer accuracies across scene classes than a full-resolution scene classification model.
- The newly proposed hybrid approach using biologically based modeling and Transformers can flexibly be applied to various naturalistic tasks and stimuli.