FOVEATER: FOVEATED TRANSFORMER FOR IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many animals and humans process the visual field with varying spatial resolution (foveated vision) and use peripheral processing to make eye movements and point the fovea to acquire high-resolution information about objects of interest. This architecture results in computationally efficient rapid scene exploration. Recent progress in self-attention-based vision Transformers, an alternative to the traditionally convolution-reliant computer vision systems, allows global interactions between feature locations and increases robustness to adversarial attacks. However, the Transformer models do not explicitly model the foveated properties of the visual system nor the interaction between eye movements and the classification task. We propose Foveated Transformer (FoveaTer) model, which uses pooling regions and eye movements to perform object classification tasks using a Vision Transformer architecture. Using square pooling regions or biologically-inspired radial-polar pooling regions, our proposed model pools the image features from the convolution backbone and uses the pooled features as an input to transformer layers. It decides on subsequent fixation location based on the attention assigned by the Transformer to various locations from past and present fixations. The model uses a confidence threshold to stop scene exploration. It dynamically allocates more fixation/computational resources to more challenging images before making the final image category decision. We construct a Foveated model using our proposed approach and compare it against a Baseline model, which does not contain any pooling. Using five ablation studies, we evaluate the contribution of different components of the Foveated model. We perform a psychophysics scene categorization task and use the experimental data to find a suitable radial-polar pooling region combination. We also show that the Foveated model better explains the human decisions in a scene categorization task than a Baseline model. On the ImageNet dataset, the Foveated model with Dynamic-stop achieves an accuracy of 8% below the Baseline model with a throughput gain of 76%. Using a Foveated model with Dynamic-stop and the Baseline model, the ensemble achieves an accuracy of 0.7% below the Baseline using the same throughput. We demonstrate our model's robustness against PGD adversarial attacks with both types of pooling regions, where we see the Foveated model outperform the Baseline model.

1 INTRODUCTION

Many mammals, including humans, have evolved a locus (the fovea) in the visual sensory array with increased spatial fidelity and use head and eye movements (Land, 2012; Marshall et al., 2014) to orient such locus to regions and objects of interest. The system design allows visual-sensing organisms to accomplish two objectives: fast target detection crucial for survival and savings in computational cost. Computational savings are accomplished by limiting the number of units with high computational costs (i.e., higher spatial resolution processing) to the fovea's small spatial region. Fast target detection is achieved by distributing the remaining computational power across a much larger area in the periphery, with a lower spatial resolution with increasing distance from the fovea. Critical to the design is an efficient algorithm to guide through eye movements the high-resolution fovea to regions of interest using the low-resolution periphery (Hayhoe & Ballard, 2005; Strasburger et al., 2011; Ludwig et al., 2014) and allow optimizing the target detection and scene classification. Various

computational models were proposed to model the search using foveated visual system (Yamamoto et al., 1996; Prince et al., 2005).

Computer vision has evolved from hand-crafted features to data-driven features in modern CNNs. Due to their computational limitations, the objectives of the computer vision systems align well with those of human visual system: to optimize visual detection and recognition with an efficient computational and metabolic footprint. Approaches toward saving computational power can be seen; for example, computer vision systems evolved from using sliding windows to RCNN's (Girshick et al., 2014) use of selective search and Faster-RCNN's (Ren et al., 2015) use of Region Proposal Network (RPN).

A system that mimics human vision by processing the scene with a foveated system and rational eye movements has also been proposed. This approach to exploring the scene can be seen in models like RAM (Mnih et al., 2014) for recognizing handwritten single-digits or detecting objects (Akbas & Eckstein, 2017) where they sequentially process the image and decide what to process next by using the peripheral information. These foveated models approach that of full-resolution models but using a fraction of the computations. Foveated systems have also shown to result in more robustness (Luo et al., 2015; Deza & Konkle, 2020; Kiritani & Ono, 2020; Vuyyuru et al., 2020) against adversarial attacks.

There has been a recent innovation in computer vision using Transformers (Touvron et al., 2020; Dosovitskiy et al., 2020) for object classification tasks that depart from the traditional over-reliance on convolutions. Even after replacing the convolutions with attention modules and multilayer perceptrons, Vision Transformers (Dosovitskiy et al., 2020; Touvron et al., 2020) achieve close to state-of-the-art performance on the ImageNet dataset and provide better robustness against adversarial attacks (Shao et al., 2021).

Due to the flattened architecture of the transformers, it is easier for multi-resolution features to share the same feature channels. Transformers (Vaswani et al., 2017) have the added benefit of self-attention, which facilitates the interaction of various parts of the image irrespective of distance. No papers have evaluated the additional potential gains of incorporating a foveated architecture into Vision Transformers for the task of ImageNet classification.

Here, we evaluate the effect of a foveated architecture and sequential eye movements on a state-of-the-art transformer architecture. We compare the Foveated transformer relative to the Baseline model in terms of classification accuracy and robustness to adversarial attacks. We perform a psychophysics experiment for a scene classification task and evaluate the Foveated model agreement with the human decision against that of the Baseline model. We first perform an object classification task using multiple fixations, moving foveal attention across different parts of the image, and using only a limited portion of the image information at each fixation, thereby reducing the input to the transformer by many folds. The model decides on subsequent fixation location using the self-attention weights accumulated from the previous fixations until the current step. Finally, the model makes the final classification decision.

2 Related work

Transformers have achieved great success in Natural Language Processing since their introduction by Vaswani et al. (2017) for machine translation. Recently, the application of Transformer models in Computer Vision has seen tremendous success. Vision Transformer (ViT) model introduced by Dosovitskiy et al. (2020) achieved remarkable performance on ImageNet (Deng et al., 2009) by using additional data from JFT 300M (Sun et al., 2017) private dataset. Subsequently, the DeiT model (Touvron et al., 2020) introduced knowledge transfer concepts in transformers to leverage the learning from existing models. Using augmentation and knowledge transfer, the DeiT model achieved close to state-of-the-art performance using training data from the ImageNet dataset alone.

Sequential processing provides three main advantages in computer vision. Larochelle & Hinton (2010) proposed a model based on the Boltzmann machine that uses foveal glimpses and can make eye movements. First, it can limit the amount of information processed at a given instant to be constant, i.e., the ability to keep computations constant irrespective of the input image size. Second, sequential models can help model human eye movement strategies and help transfer that information to build better computer vision systems. RAM (Mnih et al., 2014) introduced a sequential model

capable of making a sequence of movements across the image to integrate information before classification. In addition, the hard-attention mechanism, implemented using reinforcement learning, was used to predict the sequence of fixation locations. Ba et al. (2015) extended these ideas to recognize multiple objects in the images on a dataset constructed using MNIST. Third, sequential processing requires fewer parameters than a model using full-resolution image input. Other models (Xu et al., 2015) have proposed image captioning models based on both hard-attention and soft-attention. Additionally, the spatial bias introduced into CNNs due to padding (Alsallakh et al., 2021) can be overcome using sequential models (Tsotsos, 2011). On the flip side, sequential models might suffer longer processing times due to sequential processing and slow convergence times for reasons similar to RNNs (Pascanu et al., 2013).

Computational models of categorization and eye movements have been proposed for rapid categorization in terms of low-level properties such as spatial envelopes (Oliva & Torralba, 2001) and texture summary statistics (Rosenholtz et al., 2012). Saliency-based models (Koch & Ullman, 1987; Itti et al., 1998; Itti & Koch, 2000) traditionally tried to model eye movements by identifying bottom-up properties in the image that will capture attention. Torralba et al. (2006) showed how saliency could be combined with contextual information to guide eye movements. Low-resolution periphery and high-resolution central fields are integrated with saliency to predict human-like eye movements (Wloka et al., 2018). Data-driven scan path prediction models (Kümmerer et al., 2022) train on image content and human fixations to predict the fixations under a free viewing but do not consider decision accuracy in specific tasks after multiple fixations. Goal-directed attention control (Zelinsky et al., 2021) showed the dependency of search patterns on target features and scene context. Akbas & Eckstein (2017) implemented a biologically-inspired foveated architecture (Freeman & Simoncelli, 2011) with a deformable parts model to build a foveated object detector on PASCAL dataset (Everingham et al., 2014), whose accuracy was close to a full-resolution model but using a fraction of the computations. Spatial transformer networks (Jaderberg et al., 2015), an older technique different from the proposed Vision Transfomers, were used on CIFAR-10 dataset (Krizhevsky, 2009), with foveation to improve object localization using foveated convolutions (Harris et al., 2019) and achieve better eccentricity performance (Dabane et al., 2021) on MNIST dataset (LeCun et al., 1998).

FoveaTer combines biologically-inspired foveated architecture with a Vision Transformer Network. Unlike the previous architectures (Akbas & Eckstein, 2017; Mnih & Gregor, 2014), we do not scale the image and thereby retain the parallelism with biological mechanisms. We apply our model to real-world images from the ImageNet dataset for image classification. In contrast, the previous works were mainly limited to datasets with small image sizes or a smaller number of output classes. They did not extend to large-scale real-world databases like ImageNet, which has 1000 class labels. We also evaluate the functional roles of various components through ablation studies, including the memory of foveal and peripheral information from previous fixations, inhibition of return, and eye movement guidance algorithms.

A novel aspect of the proposed work is that the model also learns that all images are not equally difficult to classify, adapting the exploration of eye movements to different images and thus varying computational resources used to classify different images successfully. The model implements this idea using a confidence threshold to restrict the scene exploration to the necessary fixations to classify the image.

Also novel is an evaluation of the adversarial robustness of our model to understand the contributions of the foveated architecture and that of sequential fixations towards defense against adversarial attacks. We use the projected gradient descent method (Kurakin et al., 2017; Madry et al., 2018), which iteratively computes the adversarial image. The architectural changes may not be trivially transferable to a new architecture. End-to-End training and hyper-parameter settings might be needed to adapt to the architectural differences.

3 MODEL

The model consists of three components, as shown in Figure 1 - convolution backbone, foveation module, and transformer layers. Interactions between different feature locations are limited to local regions in the convolution backbone. The Foveation module performs non-uniform pooling on the input features, reducing feature dimensionality. The Foveation module can contain two types of

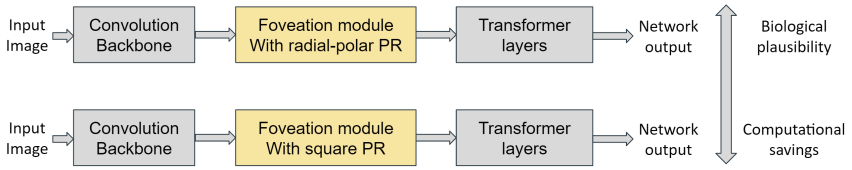
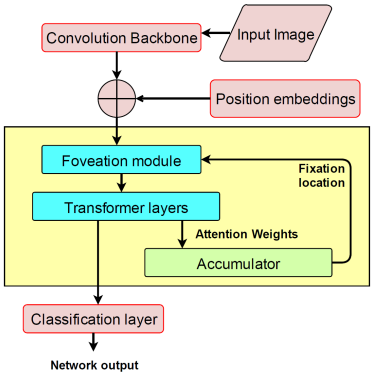
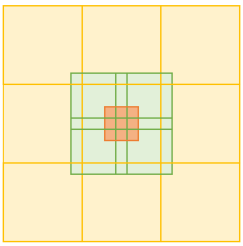


Figure 1: PR refers to the pooling regions. Foveation with radial-polar pooling regions is more biologically plausible than the square pooling regions but computationally slower and vice-versa.



(a) **FoveaTer architecture:** The foveation module performs fixation-dependent pooling. *Accumulator* uses the attention weights from the last transformer layer of past and present fixations to predict the next fixation location. Model blocks within the yellow region are executed for each fixation.



(b) **Square pooling:** Input feature map is pooled to generate the pooled feature vectors. The Fovea is shown in red containing 9 feature vectors. The first level of pooling regions is of size 5×5 with stride 4 (green). The second pooling region level is size 7×7 with stride 7 (orange).

Figure 2: Network architecture

pooling regions, square pooling regions which provide computationally fast processing, or biologically plausible radial-polar pooling regions, (Freeman & Simoncelli, 2011). Under this non-uniform average-pooling model, locations closer to the fixation location use smaller neighborhoods for pooling than locations far from the fixation location. The last component consisting of the transformer layers contributes in three ways - 1. They allow global interactions, which allows the possibility of using context-based decision-making. 2. They eliminate the need to design convolution layers on top of non-uniform sampled features from the Foveation module. 3. Self-attention weights of the transformer layers can be helpful in fixation guidance.

For the square pooling regions, the input image is first passed through the convolution backbone resulting in a feature vector of size $[384, 14, 14]$. After adding the sinusoidal position embedding and performing fixation-dependent average-pooling using the Foveation module, the feature size reduces to $[384, 22]$. Pooled features of size $[384, 22]$ are passed through the transformer layers, followed by the classification layer resulting in a logits vector. We use the self-attention weights from the last transformer layer to predict the subsequent fixation location. We make five fixations on each image during the model training. This choice keeps the computational cost relatively the same as the Baseline model. Model architecture is shown in Figure 2a.

The convolution backbone consists of six convolution layers and is structured similarly to the initial layers of the ResNet-18 model. Square pooling regions can exploit the fast average-pooling library functions, whereas the pooling in the radial-polar pooling regions needs custom implementation. Four architectural changes make it possible for the FoveaTer model to perform serial processing, achieve throughput improvements and retain information across fixations. Firstly, the Foveation module is a plug-in module that can be preceded or succeeded by the transformer layers. However, additional changes would be required for the convolution layers to follow the foveation module. Secondly, the periphery (i.e., the pooling regions other than Fovea) pools the feature vectors and, as a result, reduces the number of features processed by the subsequent layers. Thirdly, the attention-based fixation guidance mechanism (FGM) helps predict the subsequent fixation location using the attention values of current and past fixations. Lastly, the features from the past fixation's foveal locations are retained and processed along with the foveal and peripheral features of the current fixation. Thus, allowing the model to access memory.

Retention of foveal features: The number of feature vectors processed by the Foveation module varies across fixations due to the retention of foveal features from past fixations. For each fixation, if the number of peripheral and foveal features are A and B, the number of features processed by the foveation module at Nth fixation equals A+NB.

Initial fixation for the Foveated model: The input feature map to the Foveation module has a spatial size of 14×14 for the condition of square pooling regions. All locations except the last and first row/column are potential fixation locations, resulting in 144 locations. We select a random location as the initial fixation point during training, and the model guides subsequent fixations.

Loss function: We use Cross-entropy for computing the classification loss. Loss from all fixations is incorporated to get the mini-batch loss, $loss = \sum_{i=1}^{N} L_{CE}(O_i, y)$ Where i corresponds to the fixation index, $N = 5$ for the Foveated model & $N = 1$ for the Baseline model, i.e., single-pass, y corresponds to the target label, O_i corresponds to the network output for fixation i, and L_{CE} correspond to cross-entropy loss.

3.1 FOVEATION MODULE

The mean feature vector corresponding to each pooling region is computed using $P = (1/M) \sum_{j=0}^{M-1} E_j$, Where E_j is a feature vector belonging to that pooling region, and M is the number of feature vectors in that pooling region.

We use square pooling regions for computational speed-up. Each image in a mini-batch has a corresponding fixation location. The fixation location represents the center of the visual field, allowing us to align the input image/feature map with the visual field. After aligning the input feature map with the visual field, features falling within a pooling region are average-pooled, and the resultant pooled vector represents that pooling region. We use pooling regions with receptive field sizes 1×1, 5×5, 7×7 blocks on a feature map of size 14×14 blocks, as shown in Figure 2b. Each block corresponds to a $[16, 16]$ pixel region in the input image of width and height 224. Central 3×3 red block represents the high-resolution Fovea, where there is no average pooling. The next ring of pooling regions, where the pooling region is green, has a receptive field of 5×5 which translates to an average pooling of 25 feature vectors to generate the representative feature vector for that pooling region. Similarly, the rings of orange-colored pooling centers have receptive fields of 7×7.

3.2 ACCUMULATOR

Accumulator uses the self-attention weights from the last transformer layer for fixation guidance. Using the attention weights of the current and past fixations along with inhibition of return, the Accumulator (see below) predicts the subsequent fixation location. A confidence map (CM_N) is constructed based on the fixation point location by putting these weights back on a 14×14 map at the corresponding pooling region's location, where 14×14 corresponds to the size of the input feature map. Inhibition of return (IoR) (Dukewich & Klein, 2015) refers to a tendency in human observers not to attend to previously attended or fixated regions. Old accumulated attention map (AM_{N-1}) is weighted by 0.5 and added to the current confidence map to create the new accumulated attention map: $AM_N = 0.5 * AM_{N-1} + CM_N$. The inhibition of return (IOR_N) map is initialized with zeros and is the same size as the feature map. Locations corresponding to the current fovea location are changed to 16. After subtracting the IOR map from the accumulated attention map, max location of the resultant map is used as the next fixation location $Fix_{N+1} = \arg\max(AM_N - IOR_N)$.

3.3 DYNAMIC-STOP OF FIXATION EXPLORATION:

Due to various factors such as occlusion, camera angle, and brightness, the difficulty of making a classification decision varies across object classes and images. To achieve higher computational efficiency in our Foveated model during inference, we stop exploring the images with fixations when the predicted class with the highest probability reaches a pre-defined threshold corresponding to that class. We compute the threshold from the training dataset's set of all the correct prediction probabilities. The model stops if the top prediction is above the 50th percentile of probabilities for that class and the second-best prediction is below the 5th percentile for that respective class.

Table 1: **Ablation Studies:** Four network components are considered, and the percentage accuracy drop after five fixations with respect to the Benchmark model is reported in the last row. Checkmark (✓) indicates that the model includes the component, while the dashed-line (—) indicates that the component has been removed.

Network component	Benchmark	Study 1	Study 2	Study 3	Study 4
Foveation	✓	✓	✓	✓	✓
Peripheral features	✓	—	✓	✓	✓
Foveal features	✓	✓	—	✓	✓
Retention of foveal features	✓	✓	—	—	✓
Inhibition of return	✓	✓	✓	✓	—
Accuracy@1	76.29	62.85	72.60	75.29	75.23
Percentage drop		17.6	4.8	1.3	1.4

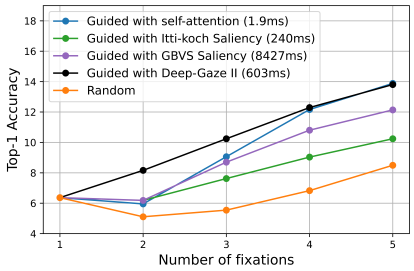


Figure 3: **Study 5:** Self-attention guidance outperforms the random fixations by 63%. Initial fixation at the top-left corner. Time taken for computing five fixations is shown in brackets.

4 Ablation Studies

We study the contribution of different network components to model performance in five ablation studies. The model was trained on ImageNet for 300 epochs and fine-tuned for 30 epochs for each ablation study. The first two studies asses the importance of peripheral and foveal features. Studies three and four assess the importance of memory provided by past foveal features and IoR, respectively. Lastly, we look at the contributions of the fixation guidance mechanism. Results are shown in Table 1 and Figure 3.

Study 1: Contribution of peripheral features: Peripheral features are essential because they contribute to image classification and help decide the subsequent fixation location. There is a sharp 17.6% drop in the network performance by removing the peripheral features.

Study 2: Contribution of foveal features: Foveal features provide high-resolution information. By removing access to foveal features of current and past fixations, the model loses access to all full-resolution information. There is a 4.8% drop in the network performance by removing the foveal features.

Study 3: Retention of the foveal features: We incorporate memory by retaining the past foveal features and processing them along with the foveal and peripheral features of the current fixation. Even without this network component, the network has some memory as the model makes fixations to more informative locations using guided fixations. In this experiment, we remove the usage of foveal features from past fixations, and as a result, the model performance drops by 1.3%.

Study 4: Contribution of Inhibition of Return: By limiting the model's ability to revisit the fixation locations of the past, we force the model to explore rather than get stuck at one location. We only see a slight drop in performance of 1.4% without the IoR, signifying that the model can operate well without IoR. The results suggest that the model can learn not to revisit locations without explicitly implementing IoR.

Study 5: Effectiveness of Fixation guidance mechanism: Objects in the ImageNet dataset often occupy a large part of the image. As a result, image classification might be possible by fixating anywhere on a large percentage of the image. The importance of guided fixation is best illustrated when a few image regions are informative. To identify that subset of images, we separate the testing images into two groups, one with moderate difficulty and the other with too few or too many informative locations. To identify these two groups of images, we run our model under a one-fixation condition at each possible fixation location and calculate the percentage of locations (PoL) with the correct classification in that image. We use this as the metric for image difficulty, i.e., higher PoL signifies less difficulty and vice versa. As there are 144 locations, PoL ranges from 0 to 144. We label all the images where the PoL is more than one-eighth the maximum value, i.e., greater than 18, as too easy. Similarly, images with a PoL of zero are labeled as too difficult. After removing the images labeled as too easy or too difficult, approximately 8% images fall in the middle, i.e., moderately difficult category. Figure 3 shows the comparison of random and guided fixations on this subset of images, and guided fixations have approximately 63% improvement over random fixations. We also compare the fixation guidance using the Itti-Koch, Graph Based Visual Saliency (Harel et al., 2006) and the DeepGaze-II model, where they take the image without foveation as input. Fixations guided by self-attention outperform the fixations guided by the Itti-Koch model and are as effective as those

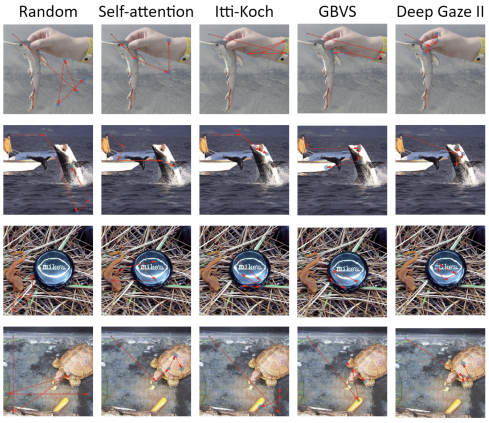


Figure 4: Fixation guidance by different models

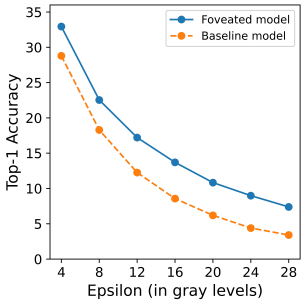


Figure 5: PGD attack: Strength of the attack is represented in terms of equivalent gray levels.

guided by the Deep-Gaze II for later fixations. Lower performance than Deep-Gaze II in the first fixations is not surprising since Deep-Gaze II is predicting the most likely regions to be fixated by humans (not the order).

Comparing the time taken for fixation prediction, our fixation guidance is the fastest as we leverage the model's internal attention weights rather than running a separate model. Time taken for computing five fixations - Guided by self-attention (1.9ms) $<$ Itti-Koch (240ms) $<$ Deep-Gaze (603ms) $<$ GBVS (8427ms). Sample image fixations are shown in Figure 4.

5 Accuracy and Robustness on ImageNet

In the following sub-sections, we compare the performance, computational complexity, and adversarial robustness of the Foveated model against the Baseline. The Foveated model is trained for five fixations, although it can work with any desired number of fixations at test time. Baseline and Foveated models have the same 24M parameters.

We use the Patchconvnet (Touvron et al., 2021) architecture. We initialize the convolution backbone with the weights from ResNet-18 (He et al., 2016) model, and the transformer layers are initialized with the weights of the DeiT-small (Touvron et al., 2020) model and trained for 300 epochs with an initial learning rate of $5e-4$ and a minimum learning rate of $1e-5$. We use AdamW (Kingma & Ba, 2014; Loshchilov & Hutter, 2019) optimizer with a decay of $1e-8$ and a cosine learning rate schedule. We use ImageNet (Deng et al., 2009) dataset for the results shown in the following sub-sections. We use RTX A6000 GPUs for training and testing purposes. We report the number of inferences completed by the GPU during a one-second time interval to compare the computational complexity of different models during inference time.

5.1 Top-1 Accuracy:

For the Dynamic-stop, we first compute the throughputs of the Foveated model for each of the one to five fixation conditions, followed by the number of images belonging to each of those five fixation conditions. The throughput of the Dynamic-stop model is computed as the weighted Harmonic mean of the throughputs of individual fixation models. Ensemble refers to a model composed of both the Foveated and Baseline models. When the Dynamic-stop is applied, and the model cannot make a decision even after the maximum number of fixations, the Ensemble model transfers the responsibility of making a decision to the Baseline model.

We present the results on the ImageNet dataset in Table 2. The Deit-Small model has a throughput of 1699 and Top-1 accuracy of 79.83. The Baseline, which has the same architecture as the Foveated model except for the foveation module, has a throughput of 1229 and an accuracy of 81.90. Since the first level of the pooling region is of size 5×5, we construct a pooled version of the baseline model using 5×5 average-pooling. We compare this with the Foveated model with two fixations, with approximately the same throughput. The Foveated model with two fixations outperforms the uniformly pooled Baseline model, as shown in row 6. Dynamic-stop and Ensemble performances are shown in the last two rows. The performance of the ensemble model reaches close to the Baseline model in terms of throughput and accuracy.

Table 2: Throughput and Accuracy on ImageNet: We compare our models against the baseline model using Top-1 accuracy and Image throughput. (*Uniform pool* - uniform 5×5 pooling, *CF* - initial fixation at image center, *Rand* - random initial fixation)

Model	Pooling type	Fixations	Type	Throughput	Acc@1
DeiT-Small				1699	79.83
Baseline				1229	81.90
			Uniform pool	2506	70.90
Foveated	Square	Rand-1		3820	69.80
		CF-1		3820	72.80
		CF-2		2307	74.70
		CF-3		1506	75.40
		CF-5		923	76.30
		CF-3	Dynamic Stop	2169	75.30
		CF-3	Ensemble	1236	81.30

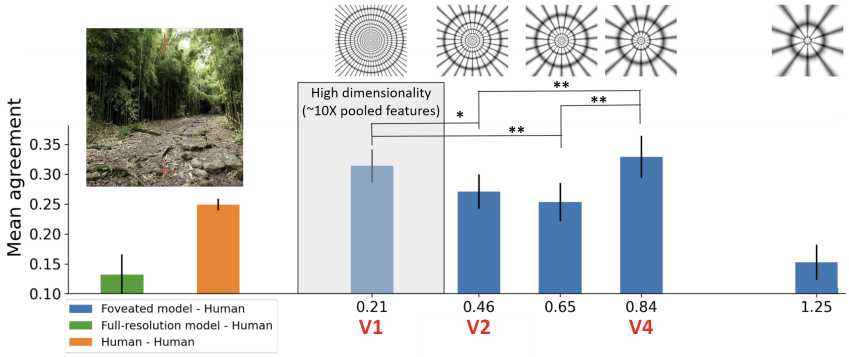


Figure 6: Mean agreement values of the Baseline and the Foveated models with human decisions (correct/incorrect). Error bars refer to the standard error across 22 participants. Paired t-test p values indicate statistical significant agreement differences across scales, $^{**}p < 0.01$, $^{*}p < 0.05$.

5.2 Robustness against adversarial attacks

We consider the Projected Gradient Descent (PGD) attack to compare the robustness of Foveated and Baseline models. PGD uses ten iterations with a step-size of $\epsilon/5$ and l-infinity norm. We use Cleverhans library (Papernot et al., 2018) for implementing the adversarial attacks. Figure 5 shows the model accuracy after attacking the input image with the adversarial attack. Epsilon (ϵ) represents the strength of the attack. Foveated model displays strong defense as compared to the Baseline model. Foveated model consistently outperforms the Baseline model. A comparison with existing models, showing the robustness of the foveated systems against adversarial attacks, is demonstrated in Appendix A.3.

6 Biologically plausible FoveaTer

Radial-Polar pooling makes the model more biologically plausible. Through psychophysics experiments of image discrimination and modeling, Freeman & Simoncelli (2011) showed that different layers of the visual cortex correspond to different *scales* where the *scale* parameter determines how many radial and polar pooling regions are present in that configuration. We use this model to predict human decisions in a scene classification psychophysics task while maintaining fixation and calibrate the scaling parameters of the pooling regions of FoveaTer. Figure 6 shows examples of various configurations.

6.1 Calibration of radial-polar pooling regions

We used thirty scene categories from the places365 dataset (Zhou et al., 2018) to create the experiment dataset. The task was to classify each image into one of the 30 categories. Sixty images were presented, with each image subtending 22.7 x 22.7 degrees visual angle, and observers fixated at the bottom-center or top-center within the images (2.2 degrees from the top or bottom edges of the image, Figure 6). Real-time infra-red video eye tracking allowed for interruption of the displayed image when observers made an eye movement.

Table 3: Throughput and Accuracy on ImageNet using radial-polar pooling regions with *Scale* 0.84: All foveated models made three fixations. (*CF* - initial fixation at image center)

Model	Type	Throughput	Acc@1
Baseline		1229	81.90
Foveated (square)	Dynamic Stop	1506	75.40
	Ensemble	2169	75.30
		1236	81.30
Foveated (Radial-Polar)	Dynamic Stop	117	76.69
	Ensemble	198	76.65
		186	81.52

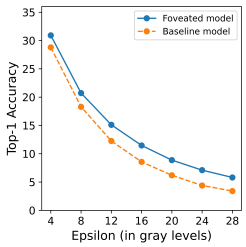


Figure 7: PGD attack on Foveated model with radial-polar pooling regions.

We tweaked the last convolutional layer so that the convolution backbone of the model outputs a $56 \times 56 \times 384$ feature map instead of a $14 \times 14 \times 384$ feature map, thus allowing us to apply the pooling regions on a higher resolution feature map. We train multiple models with different *scale* values for spatial pooling. For each *scale*, Foveated model is trained for 60 epochs after initializing the weights with the square-pooled Foveated model trained on ImageNet.

Error consistency metric (Geirhos et al., 2020) produces the normalized decision agreement between two observers, where the normalization is a function of the accuracy of both observers. We computed the mean agreement between the human decisions and the Foveated model for a set of *scales* as shown in Figure 6. We also computed the mean agreement between human decisions and the Baseline model (independent of *scale*). For *scales* corresponding to V2 (*scale*-0.46) and V4 (*scale*-0.84) layers of the visual cortex, we observe a significant difference between the mean agreement of humans with Foveated and Baseline models. Although the accuracy of the Baseline model (0.93) is higher than the Foveated model (0.86), human decisions with a mean accuracy of 0.83 are in better agreement with the Foveated model. The fixation at the top-center or the bottom-center limited the image information accessible to the human observers, which the Full-resolution fails to model. Our findings suggest that human categorization of scenes within a single fixation can be better predicted with FoveaTer with pooling regions that scale according to properties of the visual cortex (V1 and V4).

6.2 Accuracy and robustness on ImageNet

We evaluated FoveaTer's accuracy and robustness using pooling parameters (scaling 0.84, V4) that predicted human scene classification decisions and were computationally efficient (relative to V1). Results are shown in Table 3. The throughput of the Foveated model with radial-polar pooling regions is very low due to the lack of library functions implementing the radial-polar pooling. As the specialized hardware performing neuro-foveal pooling becomes available in the future, throughput gaps will disappear, and the Foveated model will become competitive with the Baseline models. Adversarial robustness of Foveated model against PGD attack with radial-polar pooling regions is illustrated in Figure 7. As with the square pooling regions, Foveated model with radial-polar pooling regions is also more adversarial robust than the Baseline model.

7 Conclusion

We provided a comprehensive framework for using foveal processing and fixation exploration on a Vision Transformer architecture for image classification. The proposed architecture introduces a way to limit computations required to process an image by flexibly adjusting the required number of fixations, providing robustness to adversarial attacks, and giving us a model that can allocate computational resources based on the difficulty of an image. Our ablation studies highlight the importance of peripheral processed features, how the self-attention guiding eye movements learn to inhibit revisits and results in accuracy similar to a model guided by predictions of human fixation (DeepGaze). We also implemented a more biologically plausible implementation with radial polar pooling and showed that pooling parameters corresponding to visual cortical areas V1 and V4 could explain human scene categorization decisions better than the Baseline non-foveated model. In conclusion, we leveraged the most recent Vision Transformer architecture and combined it with ideas from foveated vision to come up with a model which has multiple knobs in terms of the number of fixations to be executed and limits on the computations performed so that the end-user will have the flexibility to fine-tune depending on their needs.

REFERENCES

Emre Akbas and Miguel P. Eckstein. Object detection through search with a foveated visual system. *PLOS Computational Biology*, 13(10):1–28, 10 2017. doi: 10.1371/journal.pcbi.1005743. URL https://doi.org/10.1371/journal.pcbi.1005743.

Bilal Alsallakh, Narine Kokhlikyan, Vivek Miglani, Jun Yuan, and Orion Reblitz-Richardson. Mind the pad – {cnn}s can develop blind spots. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=m1CD7tPubNy.

Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.7755.

Ghassan Dabane, Laurent Perrinet, and Emmanuel Daucé. What You See Is What You Transform: Foveated Spatial Transformers as a bio-inspired attention mechanism. 9 2021. doi: 10.36227/techrxiv.16550391.v1. URL https://www.techrxiv.org/articles/preprint/What_You_See_Is_What_You_Transform_Foveated_Spatial_Transformers_as_a_bio-inspired_attention_mechanism/16550391.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Arturo Deza and Talia Konkle. Emergent properties of foveated perceptual systems. *ArXiv*, abs/2006.07991, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kristie R. Dukewich and Raymond M. Klein. Inhibition of return: A phenomenon in search of a definition and a theoretical framework. *Attention, Perception, & Psychophysics*, 77(5):1647–1658, Jul 2015. ISSN 1943-393X. doi: 10.3758/s13414-015-0835-3. URL https://doi.org/10.3758/s13414-015-0835-3.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98–136, 2014.

Jeremy Freeman and Eero P. Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14: 1195 – 1201, 2011.

Robert Geirhos, Kristof Meding, and Felix Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *ArXiv*, abs/2006.16736, 2020.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014. doi: 10.1109/CVPR.2014.81.

Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NIPS*, 2006.

Ethan William Albert Harris, Mahesan Niranjan, and Jonathon Hare. Foveated convolutions: improving spatial transformer networks by modelling the retina. In *Shared Visual Representations in Human and Machine Intelligence: 2019 NeurIPS Workshop*, December 2019. URL https://eprints.soton.ac.uk/441204/.

Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188—194, April 2005. ISSN 1364-6613. doi: 10.1016/j.tics.2005.02.009. URL https://doi.org/10.1016/j.tics.2005.02.009.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. doi: 10.1109/34.730558.

Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10):1489–1506, 2000. ISSN 0042-6989. doi: https://doi.org/10.1016/S0042-6989(99)00163-7. URL https://www.sciencedirect.com/science/article/pii/S0042698999001637.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

Taro Kiritani and Koji Ono. Recurrent attention model with log-polar mapping is robust against adversarial attacks. *ArXiv*, abs/2002.05388, 2020.

Christof Koch and Shimon Ullman. *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry*, pp. 115–141. Springer Netherlands, Dordrecht, 1987. ISBN 978-94-009-3833-5. doi: 10.1007/978-94-009-3833-5_5. URL https://doi.org/10.1007/978-94-009-3833-5_5.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Matthias Kümmerer, Matthias Bethge, and Thomas S. A. Wallis. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22, 2022.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ArXiv*, abs/1607.02533, 2017.

M. Land. Oculomotor behaviour in vertebrates and invertebrates. *The Oxford Handbook of Eye Movements*, 01 2012. doi: 10.1093/oxfordhb/9780199539789.013.0001.

Hugo Larochelle and Geoffrey E Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/677e09724f0e2df9b6c000b75b5da10d-Paper.pdf.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278–2324, 1998.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Casimir J. H. Ludwig, J. Rhys Davies, and Miguel P. Eckstein. Foveal analysis and peripheral selection during active visual sampling. *Proceedings of the National Academy of Sciences*, 111 (2):E291–E299, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1313553111. URL https://www.pnas.org/content/111/2/E291.

Yan Luo, Xavier Boix, Gemma Roig, Tomaso A. Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *ArXiv*, abs/1511.06292, 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018.

N Marshall, M. Land, and T Cronin. Shrimps that pay attention: Saccadic eye movements in stomatopod crustaceans. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369:20130042, 02 2014. doi: 10.1098/rstb.2013.0042.

Andriy Mnih and Karol Gregor. Neural Variational Inference and Learning in Belief Networks. In *Proceedings of ICML*, 2014.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 2204–2212, Cambridge, MA, USA, 2014. MIT Press.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001. ISSN 1573-1405. doi: 10.1023/A:1011139631724. URL https://doi.org/10.1023/A:1011139631724.

Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *ICML*, 2013.

Simon Prince, James H. Elder, Yuqian Hou, Mikhail Sizintsev, and Yevgen Olevskiy. Statistical cue integration for foveated wide-field surveillance. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2:603–610 vol. 2, 2005.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

Ruth Rosenholtz, Jie Huang, and Krista Ehinger. Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology*, 3:13, 2012. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00013. URL https://www.frontiersin.org/article/10.3389/fpsyg.2012.00013.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *ArXiv*, abs/2103.15670, 2021.

Hans Strasburger, Ingo Rentschler, and Martin Jüttner. Peripheral vision and pattern recognition: A review. *Journal of vision*, 11:13, 05 2011. doi: 10.1167/11.5.13.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 843–852, 2017. doi: 10.1109/ICCV.2017.97.

Antonio Torralba, Aude Oliva, Monica Castelhano, and John Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological review*, 113:766–86, 11 2006. doi: 10.1037/0033-295X.113.4.766.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Piotr Bojanowski, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Herv'e J'egou. Augmenting convolutional networks with attention-based aggregation. *arXiv preprint arXiv:2112.13692*, 2021.

John K. Tsotsos. A computational perspective on visual attention. 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Manish Reddy Vuyyuru, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2135–2146. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/17256f049f1e3fede17c7a313f7657f4-Paper.pdf.

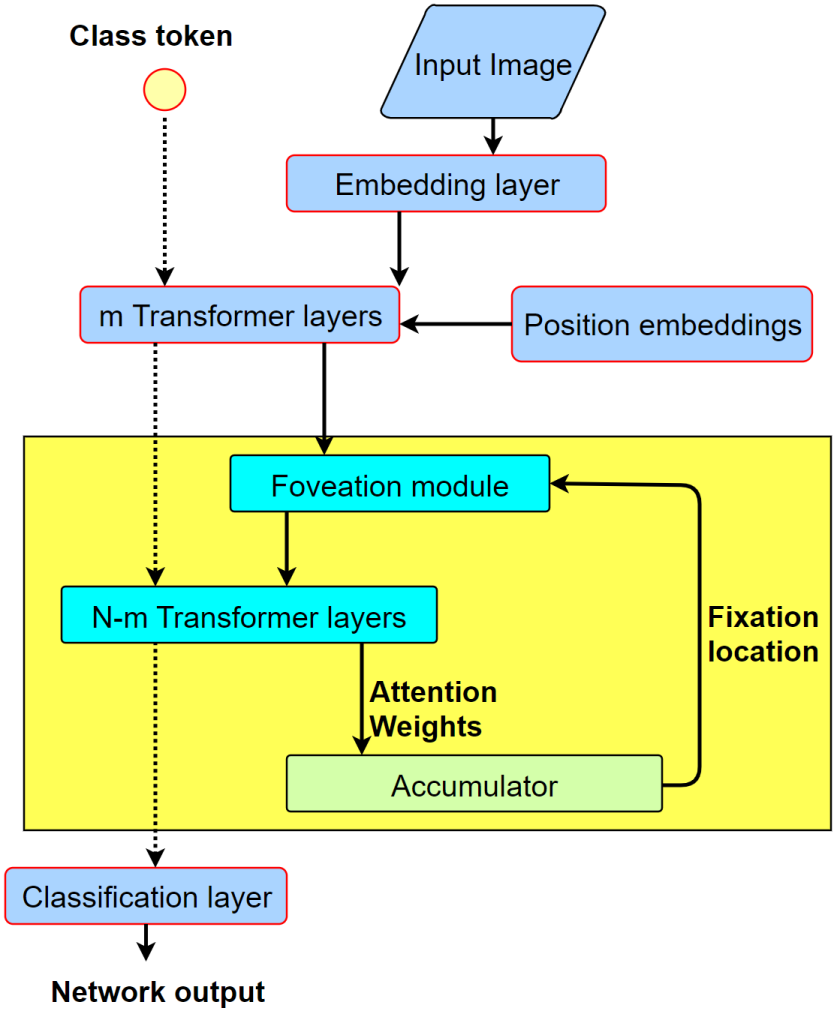
Calden Wloka, Iuliia Kotseruba, and John K. Tsotsos. Active fixation control to predict saccade sequences. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3184–3193, 2018. doi: 10.1109/CVPR.2018.00336.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/xuc15.html.

H. Yamamoto, Y. Yeshurun, and M. Levine. An active foveated vision system: Attentional mechanisms and scan path covergence measures. *Comput. Vis. Image Underst.*, 63:50–65, 1996.

Gregory J. Zelinsky, Seoyoung Ahn, Yupei Chen, Zhibo Yang, Hossein Adeli, Lihan Huang, Dimitrios Samaras, and Minh Hoai. Predicting goal-directed attention control using inverse-reinforcement learning. *Neurons, behavior, data analysis and theory*, 2021, 2021.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1452–1464, 2018.



(a) **FoveaTer architecture:** Solid black arrows denote the flow of image-related features. N is the total number of transformer layers. The foveation module performs fixation-dependent pooling. *Accumulator* uses the attention weights from the last transformer layer of past and present fixations to predict the next fixation location. Model blocks within the yellow region are executed for each fixation.

Figure 8

A Appendix

A.1 Alternate model

We present a less biologically plausible Foveated model in this section. With this architecture, the ensemble model can outperform the Baseline model.

DeiT-Small Touvron et al. (2020): The DeiT-Small architecture begins with a convolution embedding layer that transforms the $[3, 224, 224]$ input image into a $[384, 14, 14]$ representation whose spacial size is 14×14, followed by a series of twelve transformer blocks, each sized for a 384-dimensional embedding.

Foveated model: Model architecture is shown in Figure 8a. The Foveation module can be plugged-in at any stage of the transformer architecture. The first *m* transformer layers process full-resolution features, and the last *(N-m)* transformer layers process the pooled features from the foveation module. The input image is first passed through the embedding layer resulting in a feature vector of size

Table 4: Throughput and Accuracy on ImageNet: We compare our models against the baseline model using Top-1 accuracy and Image throughput. (*DS* - Dynamic stop, *Ens* - Ensemble, *Pool* - uniform 5×5 pooling, *CF* - central fixation)

Condition	Model	Fixations	Type	Throughput	Acc@1
Baseline	DeiT-Small		Baseline	323	79.83
Pooled	DeiT-Small		Pool	592	73.64
	Model	CF-1		564	75.2
Upper bound	Oracle	CF-3	DS	507	80.36
	Oracle	CF-3	Ens	388	84.27
Optimal		CF-3	DS	**489**	78.31
		CF-3	Ens	**348**	**79.99**

$[384, 14, 14]$. After adding the position embedding and flattening the spatial size of the embedding layer output, the resultant full-resolution feature vector of size $[384, 196]$ is passed through the *m* transformer blocks along with a learnable vector of size 384 values, called a class token. As the same size is maintained at the input and output of the transformer layer, a feature vector of size $[384, 196]$ is obtained at the input of the Foveation module. Then, we perform fixation-dependent average-pooling using the Foveation module, resulting in features of size $[384, 22]$. Under this non-uniform average-pooling model, locations closer to the fixation location use smaller neighborhoods for pooling than locations far from the fixation location. Pooled features of size $[384, 22]$ along with the class token are passed through the remaining *(N-m)* transformer layers. We use the self-attention weights corresponding to the class token from the last transformer layer to predict the next fixation location. Finally, the classification layer transforms the class token into a logits vector. During training, the total number of fixations is limited to five fixations.

We use a 6-6 configuration, i.e., six transformer layers before the Foveation module and six transformer layers after it. We present the results on the ImageNet dataset in Table 4. The original full-resolution model is referred to as 'Baseline', which has a throughput of 323 and Top-1 accuracy of 79.83. Since the first level of the pooling region is of size 5×5, we construct a pooled version of the baseline model using 5×5 average-pooling. We compare this with the foveated model with one fixation at the image center, with approximately the same throughput. The foveated model with single fixation outperforms the pooled baseline model, as shown in row 3. 'Oracle' refers to the model with perfect Dynamic-stop, i.e., it knows the ground truth and stops the model when the prediction matches the ground truth. Since 'Oracle' has the perfect stopping rule, it provides the upper bound on the performance of the Dynamic-stop model. Dynamic-stop and Ensemble performance is computed. Finally, the Foveated model's ensemble model outperforms the Baseline model in terms of throughput and accuracy.

A.2 SCENE CATEGORIES USED FOR PSYCHOPHYSICS EXPERIMENT

Classes present in the scene classification task,

1. airport terminal
2. amphitheater
3. assembly line
4. bamboo forest
5. banquet hall
6. basement
7. beach
8. boxing ring
9. bus interior
10. canal natural
11. canyon
12. classroom
13. cliff
14. corn field
15. department store
16. desert sand
17. dining room
18. forest path
19. glacier
20. greenhouse indoor
21. gymnasium indoor
22. jail cell
23. museum indoor
24. phone booth
25. railroad track
26. sauna
27. subway station platform
28. water park
29. wind farm
30. zen garden

A.3 COMPARISON OF FOVEATER WITH EXISTING MODELS

	Luo (2016)	**Reddy (2020)**	**Ours**
Dataset	ImageNet	CIFAR10, ImageNet	ImageNet, Places365 subset
Baseline Architecture	CNN (AlexNet, VGG, GNT)	CNN (ResNet)	Vision Transformer (deit)
Image scaling	Yes	No	No
Adversarial attacks	BFGS, sign method	FGSM, PGD	PGD
Resource usage (N fix)	1x	Retinal - Nx, Cortical - 1x	0.8x for 3 fix
Foveation Location	Input image	Input image	can plug-in anywhere

Table 5: Comparison with existing models

Comparison with existing models, which show the robustness of the foveated systems against adversarial attacks, is demonstrated in Table 5. Our model is based on Vision transformer architecture compared to the other models on CNN architectures. Our model can also be extended to have a convolution backbone, as shown in the supplementary material. We do not perform any image scaling. Our resource usage is $0.8\times$ that of the full resolution model. We allow the possibility of applying foveation to an intermediate feature map rather than restricting it to be applied only to the input image.