Foveation in the Era of Deep Learning

George Killick
2182951k@student.gla.ac.uk
Paul Henderson
paul.henderson@glasgow.ac.uk
Paul Siebert
paul.siebert@glasgow.ac.uk
Gerardo Aragon-Camarasa
gerardo.aragoncamarasa@glasgow.ac.uk

Computer Vision and Autonomous Systems group
School of Computing Science
University of Glasgow
Glasgow, UK

Abstract

In this paper, we tackle the challenge of actively attending to visual scenes using a foveated sensor. We introduce an end-to-end differentiable foveated active vision architecture that leverages a graph convolutional network to process foveated images, and a simple yet effective formulation for foveated image sampling. Our model learns to iteratively attend to regions of the image relevant for classification. We conduct detailed experiments on a variety of image datasets, comparing the performance of our method with previous approaches to foveated vision while measuring how the impact of different choices, such as the degree of foveation, and the number of fixations the network performs, affect object recognition performance. We find that our model outperforms a state-of-the-art CNN and foveated vision architectures of comparable parameters and a given pixel or computation budget. The source code is publicly available at https://github.com/georgeKillick90/FovConvNeXt.

1 Introduction

Many biological vision systems sense the world with a foveated sensor, where the highest resolution processing is limited to only a small central portion of the visual field (the fovea). Computer vision systems have taken inspiration from this aspect of biological vision and incorporated it into visual attention models that learn to sample and process visual scenes actively [2, 25, 31]. The promise of foveated vision is the ability to resolve and process fine details while simultaneously maintaining a wide field of view, which has applications to problems where semantic information can exist over a high-dynamic range of scales. More generally, it is well known that scaling the resolution of inputs to CNNs can reliably improve accuracy in objection recognition problems [33]. Through sparse sampling in the periphery of the field of view, foveated sensors can achieve this with significantly fewer pixels than a uniform sensor, making it an appealing approach to building parsimonious vision systems. This comes with the caveat that a visual attention mechanism must be incorporated into the system, in order to guide high-resolution processing to areas of interest.

In practice, it has proven difficult to reconcile foveated sensors with modern computer vision systems. Due to their space-variant resolution, foveated sensors are not naturally

amenable to standard convolution operators which expect uniformly sampled inputs. Many works propose sensors that can be conveniently mapped back to grids to circumvent this problem [8, 23] but in doing so, inherently change the equivariance properties of the network. There is evidence, such as in the case of the log-polar mapping, that these properties may be sub-optimal for object recognition [23, 26, 35]. Other works propose approximations of foveated sensors by rescaling multiple crops of differing fields of view (FoV) [11, 25]. These methods are again convenient, but raise questions on how features from different crops should interact at intermediate stages of processing. Other challenges arise from training an attention mechanism to guide the sensor. Reinforcement learning has been the method of choice for several works [2, 7, 25], however this is accompanied by additional challenges due to the difficulty of training systems with this method from scratch.

To this end, we propose a foveated convolutional neural network architecture that actively attends to scenes to perform image classification. Our work addresses the aforementioned challenges in the following ways. We introduce a graph convolutional approach for processing foveated images, removing the need for any mapping to a uniform sampling arrangement, in turn affording more flexible control over the equivariance properties of the layer. Additionally, this facilitates the use of a single foveated feature map, allowing foveal and peripheral features to interact at intermediate stages of the network. We adopt a differentiable sampling mechanism based on [15] allowing end-to-end training of the entire system, avoiding reinforcement learning. Additionally, we propose a novel foveated sensor that can be easily configured to control the sampling resolution across the visual field. We verify the efficacy of our model through several experiments. Crucially, we show that our model can be applied effectively to object recognition tasks on challenging real-world image datasets. In summary, we make the following main contributions:

- A novel end-to-end differentiable foveated vision architecture which is able to outperform previous foveated architectures by atleast 1% and up to 3.5% on the Imagenet100 dataset, as well as a state-of-the-art CNN at an equivalent number of input pixels and FLOPs (Section 3.1).
- A novel graph convolution layer designed to process foveated images without requiring them to be mapped to a uniform grid (Section 3.3)
- We show that foveated vision architectures are much better at recognizing objects over a wide range of scales than uniform non-attentive vision architectures when constrained to the same number of input pixels (Section 4).

2 Related Work

Visual hard attention models employ a dynamic sensor and learn an attention policy to guide the sensor to informative regions of a scene for downstream tasks. Mnih et al.'s seminal work [25] uses reinforcement learning to train a recurrent neural network that guides a foveated sensor (approximated by crops of increasing FoV) to classify MNIST digits in cluttered backgrounds. Ba et al. [2] and Sermanet et al. [31] extend this work to multiple object recognition and fine-grained classification of natural images respectively. Li et al. [20] introduce a dynamic stopping condition to a visual attention model, allowing it to terminate computation early if the model is significantly confident in its prediction, in turn reducing computational costs. This work foregoes a foveated sensor in favour of an adaptive zoomable sampling grid. Elsayed et al. [7] learn an attention policy over the output of a BagNet [4] and

propose a pretraining routine that can alleviate the difficulty of training visual attention models with reinforcement learning. In a similar approach Rangrej *et al.* [27] train a transformer to sequentially select its input patches. Unlike [7], their approach does not require the full image to be processed by the network. They additionally introduce a student-teacher training paradigm, where the teacher model has access to the whole image. These works have been primarily concerned with how to train hard-attention models and their applications to different visual perception tasks but provide relatively few insights into the design of the foveated sensor and feature extractor. We instead, provide greater focus on the underlying feature extraction process and leverage a comparitively simpler end-to-end differentiable approach.

Differentiable image sampling techniques [10, 15] offer a method to train visual attention models without reinforcement learning. Spatial Transformer Networks (STNs) [1, 15] train a neural network to predict affine transformations of the input image to a classifier. Recasens *et al.* [28] and Thavamani *et al.* [34] propose similar extensions to STNs for object recognition and detection respectively. They use adaptive sampling grids that sample an input image at higher resolution based on a saliency map computed by a CNN. Foveated variants of spatial transformers have been proposed. Harris *et al.* [11] use multiple crops, similar to [25], to improve object localization. Polar Transformer Networks [8] use the log-polar mapping to achieve rotation and scale equivariance. Cheung *et al.* [5] train a recurrent neural network to attend to and classify MNIST digits. They introduce a learnable sampling lattice and show the emergence of a foveated structure. Many works avoid explicitly learning an attention mechanism for a particular task. For instance,[17, 19] uses a multi-fov crop to reduce the number of input pixels to a video classification CNN. Ozimek *et al.* [26] use a biologically plausible foveated sensor based on [30] to perform image classification, exploiting camera bias in the dataset to circumvent the need for an attention mechanism. Lukanov *et al.* [23] adopt the foveated sensor proposed in [24] and compute consecutive fixations through argmax over class activation maps. Jonnalagadda *et al.* [16] apply foveated pooling to a small ResNet [12] feature extractor before passing the pooled representations to a vision transformer; subsequent fixations are determined by the maximal activation in the final self-attention layer.

These works encompass a broad range of foveated and adaptive downsampling techniques however relatively few comparisons between different methods has been made. Additionally, many methods report results on simple datasets such as MNIST which may not represent their efficacy on harder natural image datasets. We conduct comparisons across a variety of different approaches on the Imagenet100 dataset. Furthermore, we identify possible aspects of previous approaches to incorporating foveated vision into CNNs, specifically with regard to equivariance properties and the interaction between foveal and peripheral features, that may inhibit their performance. We address this with a novel graph convolution layer for foveated images and show improvements in object recognition performance as a result.

3 Method

3.1 Architecture

We introduce a foveated graph convolutional architecture that learns to attend to salient areas of images to perform object recognition. We consider two variants of this architecture, a sequential model (Figure 1) that can repeatedly attend to the image for many time-steps, and a spatial transformer variant which uses a separate localization network to provide a single location (fixation) for the foveated classifier to attend to.

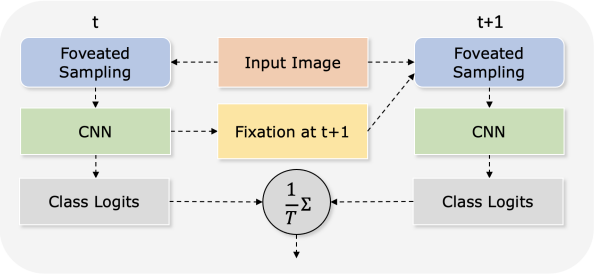

Figure 1: Our pipeline for iterative foveated image classification. The input image is sampled by our foveated sensor and processed by a novel graph convolutional network, which predicts class logits, and optionally a fixation position to which the sensor should attend next

These architectures share in common four core components, a differentiable foveated sensor, a graph convolutional feature extractor, an attention module and a classifier head. The foveated sensor samples a uniform input image centred on a given fixation coordinate. This foveated image is subsequently processed by a graph convolutional feature extractor to produce a multi-channel feature map. A conventional classifier head consisting of a global average pooling layer and linear layer produces a class prediction. The attention module is similar to that proposed in Polar Transformer Networks [8]. It takes as input a d-dimensional featuremap and applies a 1×1 convolution to produce a single channel saliency map. We apply softmax activation to the saliency map and multiply each element by its corresponding (x,y) coordinate. Finally we sum these values to arrive at a final fixation coordinate. This is equivalent to finding the expected coordinate using the probability of a given location containing salient information.

In the sequential variant, at time step 0, the network receives an initial fixation at the center of the image. A class prediction is computed via the method described above. The feature maps computed by the graph convolutional network are fed into the attention module to compute a fixation for the next time step. This allows the network to leverage rich features already computed by the feature-extractor for classification to inform the next fixation with minimal overhead. The number of fixations the network is allowed to perform is given by the hyperparameter T. After predictions for each time step have been computed, a final prediction is made by averaging predictions over all time steps. In the spatial transformer variant, an initial fixation is computed by applying our attention module to the output of a separate localization network that operates on a uniformly downsampled input image. This fixation is then fed into the foveated classifier and produces a class prediction. Both methods are trained end-to-end via gradient descent through the backpropagation of a standard cross-entropy classification loss.

3.2 Foveated Sensor

We define a foveated arrangement of bilinear sampling kernels over the image plane to perform foveated sampling. We adopt Vogel's model of seed arrangement in a sunflower capitulum [57] to provide an easy way to pack points approximately uniformly in a circle (Fig. 2). Vogel's model is simple and computed at initialisation; it also has the property that the distribution of pixels in a small patch (i.e. the size of a convolutional filter) is approximately isotropic. We adapt this model to a foveated arrangement of points with the highest sampling density in the centre and smoothly decaying resolution as eccentricity increases. We

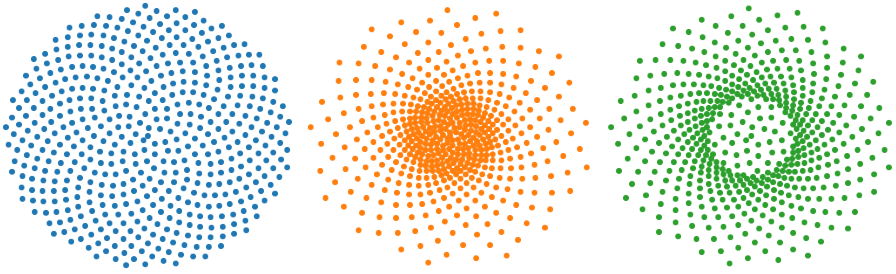

Figure 2: Left to Right: Vogel's model of a sunflower capitulum [37], our foveated adaptation (eq. 1) where fovea sampling density is well parameterised, and our adaptation where the fovea sampling density is poorly parameterised.

achieve this by logarithmically spacing points outside a given radius of the fovea. Formally the position of the i^{th} sampling kernel in polar coordinates (ρ, θ) is given by:

$$\theta_i = 2\pi i\phi, \quad \lambda = r^{\frac{1}{d-N}}, \quad \rho_i = \begin{cases} r\sqrt{\frac{i}{d}}, & \text{if } i < d-1. \\ r\lambda^{i-d}, & \text{otherwise.} \end{cases} \tag{1}$$

where ϕ is the golden ratio, N is the total number of sampling kernels, d is the number of sampling kernels in the fovea, and r is the radius of the fovea. The sampling density of the fovea can be controlled independently from the size of the fovea; however, setting the number of sampling kernels in the fovea too low will result in an excessively sparse fovea. This is undesirable as the sampling resolution will not be able to resolve details in the fovea. (Figure 2). We explore different values for r in Section 4.3.

In order to translate the sensor over the input image, we apply an (x,y) offset to all sampling kernels. Backpropagation with respect to this offset is then possible by accumulating the gradients of each individual sampling kernel output with respect to its position [8, 15]. For cases where sampling kernels extend beyond the image boundary, we use border padding to fill in missing pixels. We found that zero padding causes the network to collapse to predict the same fixation for all images (Supplementary 3).

Alternative Foveated Sensors. We consider two alternative methods for producing foveated images in our experiments. Firstly, a sensor comprised of multiple crops of increasing field of view, sampled at the resolution of the smallest crop [2, 11, 25]. For models processing this type of image, we process each crop independently and average class predictions over crops at the end. We experimented with stacking each crop in the channel dimension before being processed by the CNN but found the former to work better. We also test the log-polar transform as a foveation method [8, 36]. In this case, convolutions use circular padding in the angular dimension of the image to address the discontinuity introduced in the log-polar transform [8, 18].

3.3 Graph Convolutions

The arrangement of foveated sampling kernels does not follow a regular grid and, as such, it poses a challenge for downstream processing since typical 2D convolution operators cannot be applied here. Unlike standard images, local patches of pixels on a foveated image may not have a canonical ordering or spatial arrangement of points. This necessitates a convolution

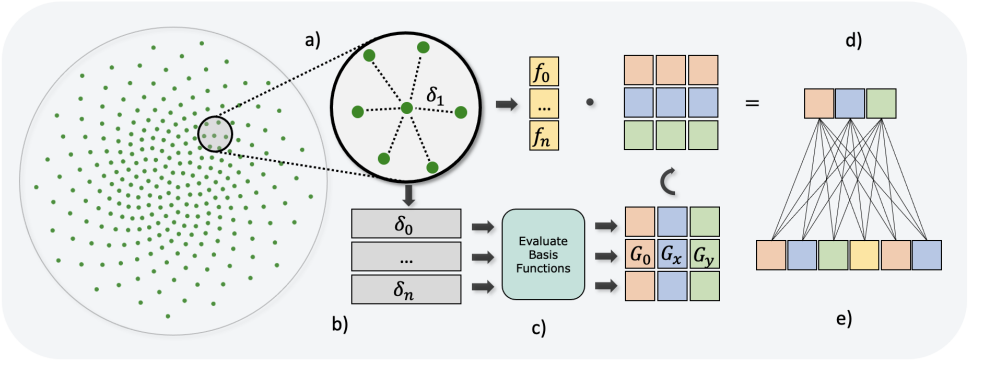

Figure 3: Our novel graph CNN for processing foveated images. Pixels are sampled non-uniformly (a) and for each, its nearest neighbours are found and their relative displacements δ_i are calculated (b). Gaussian Derivative basis functions are evaluated using the δ_i offsets to obtain basis filter weights for the current patch (c). A weighted sum between basis filter weights and features is computed to obtain a vector of responses to the basis functions (d). A learned linear combination of basis function responses computes final feature vectors for a given patch (e).

operator that can produce filter responses that are largerly invariant to changes in these properties. We overcome this by leveraging a graph convolution operator (Fig. 3) that is invariant to the permutation of pixels in the receptive field, and computes filter weights based on their relative position to the receptive field's center.

Graph Construction. Our convolution layer takes as input a set of feature vectors, each associated with a vertex $u \in U$ of a 2D planar graph and a position (x,y) that defines the feature's spatial position within the feature map. Similarly, we define a set of vertices V pertaining to output features with an (x,y) coordinate and a feature vector $\hat{f}$ that we aim to compute. For each output vertex $v_i \in V$, we define its receptive field R_i as its K nearest spatial neighbours in U and connect them to form a bipartite graph $G = (U,V,E)$. We label each edge in the graph with the spatial offset between the two vertices it connects. For example, the edge e_{ij} connecting an input vertex $u_i \in U$ to output vertex $v_j \in V$ is labelled as $\delta_{ij} = (x_i - x_j, y_i - y_j)$. We additionally normalize the δ offsets by the mean δ offset for a given patch. This scales filters as sampling density becomes more sparse in the periphery. Scaling filters exponentially with eccentricity produces feature maps that are equivariant to global scale transformations of the input. This is seen in applying convolutions to log-polar images [8]. However, our method differs in that our convolution maintains a consistent filter orientation with respect to the cartesian axes.

Edge Conditioned Filter Weights. In order to allow our model to learn filters that are as expressive as ordinary convolutions, our convolution layer produces filter weights as a function of edge labels (δ offsets). This allows the relative spatial positions of features to be considered, permitting the learning of filters such as oriented edge detectors. Graph convolution operators of this form, often termed edge-conditioned graph convolutions [32], use a parameterised differentiable function approximator to map edge labels to filter weights [3, 13, 39]. We instead decompose the filter weights into a parameterised linear combination of Gaussian derivative basis functions (Figure 1 in supplementary material), which has been shown to work well for learning convolution filters [14, 21] uniform images. This has two

major benefits for our approach. Firstly, we can explicitly control the maximum frequency of our filters by changing the σ of the Gaussian derivatives or truncating the higher-order derivatives. Secondly, the basis is known a priori and fixed, meaning we can precompute all basis filters at network initialisation. The convolution operation for a single-channel input and single-channel output is defined in the supplementary material (Section 2).

4 Experiments

We conduct our main experiments on ImageNet-100, a 100-class subset of the ImageNet-1K dataset [6], to test the efficacy of our method against alternative and similar methods in the literature (Section 4.1). We conduct additional experiments on an augmented MNIST dataset, to evaluate our model's ability to classify very large and very small objects under a highly constrained number of pixels (Section 4.2). Finally, we perform ablation studies (Section 4.3) on our model using the Imagewoof dataset [9], a 10-class fine-grained classification subset of ImageNet-1K comprising different breeds of dogs.

4.1 Effectiveness of foveation on image classification tasks

We perform experiments on the ImageNet-100 dataset, comprised of 130,000 images across 100 categories. We constrain the number of input pixels to the feature extractors to be 112^2 for all models. The reduced sampling resolution requires the models to adapt their sampling grid to salient regions of the input image that would ordinarily be lost through uniform downsampling. Additionally, the feature extractors must be able to extract rich features from the images in order to perform the classification task effectively.

Baselines. For all baseline methods, we use the atto variant of ConvNeXt as our feature extraction backbone [22, 38]. Models that receive an input of fewer than 224^2 pixels have their final downsampling layer in the convolution stages removed to maintain approximately 7^2 pixels in the final feature maps. We consider three non-attentive baselines, the default ConvNeXt at 224×224 input resolution, a downsampled input variant at 112×112 input resolution, and a foveated variant using our sensor and graph convolution layers. We consider 6 variants of Spatial Transformer Networks (STN). For the localisation networks, we use a truncated ConvNeXt atto, which has all layers after the penultimate convolution stage removed. All localization networks operate on a downsampled 112×112 input image. For the full affine spatial transformer, we append a 2 hidden layer MLP, with 128 neurons in each layer, batch normalization and ReLU activation, that regresses the affine transformation matrix. For the foveated variants, we consider the Polar Transformer Network [8] (PTN), a multi-fov crop STN [11], an STN with the FCG sensor [24] and our foveated sensor with our graph convolution operator. In these variants, we append our attention module (Section 3.1) to compute fixations. Finally, we compare against the Learning to Zoom model. This method remains largely unchanged from its original proposal except from updating the localization network and feature extractor from ResNet [12] to ConvNeXt [38].

Results and discussion.

We report top-1 accuracy on the Imagenet-100 test set in Table 1, along with the number of parameters and GFLOPs. Even without attention, we find that our foveated graph ConvNeXt outperforms a uniform ConvNeXt by 2.5%. We posit that this performance improvement is possible due to the tendency of relevant objects to be centred in the frame. Notably, despite not having attention, this model also outperforms all alternative foveated spatial transformer methods by atleast 0.7% and up to 1.8% in the case of the log-polar sensor. When incorporating attention into our method, further increase over other foveated

Method	Operator	Sensor	# Input Pixels	# Fixations	Params (M)	GFLOPs	Accuracy (%)
ConvNeXt	Conv	Uniform	50176	-	3.7	0.55	78.4
ConvNeXt	Conv	Uniform	12544	-	3.7	0.20	70.0
Ours (non-attentive)	Graph Conv	Our Sensor	12544	-	3.7	0.20	72.5
STN	Conv	Uniform	12544	1	4.8	0.32	72.7
PTN	Conv	Log-Polar	12800	1	4.8	0.33	70.7
FCG-STN	Conv	FCG	12544	1	4.8	0.33	71.0
Fov STN	Conv	Multi-FoV Crops	12800	1	4.8	0.33	71.8
Fov STN (ours)	Graph Conv	Our Sensor	12544	1	4.8	0.32	74.2
Learning to Zoom	Conv	Deformable Grid	12544	1	4.8	0.32	75.8
Sequential	Conv	FCG	12800	2	3.7	0.41	70.2
Sequential	Conv	Log-Polar	12800	2	3.7	0.41	70.4
Sequential	Conv	Multi-FoV Crops	12800	2	3.7	0.41	72.8
Sequential (Ours)	Graph Conv	Our Sensor	12544	2	3.7	0.41	73.8
Sequential (Ours)	Graph Conv	Our Sensor	12544	3	3.7	0.61	76.5

Table 1: Top-1 Accuracy on the Imagenet100 test set. We split the table into three sections. Top: non-attentive models. Middle: Spatial Transformer like models. Bottom: Sequential Models.

architectures is observed. Our sequential and spatial transformer variants perform atleast 1.0% and 2.4% better than their next best foveated counterparts. In both cases, the next best method is the Multi-FoV crop method which shares similarities with our method in that filters are only translated and scaled over the visual field, however, unlike our method, foveal and peripheral features are maintained in separate featuremaps. We find that performance of our model trails slightly behind the learning to zoom method and 'full' resolution ConvNeXt. By allowing the sequential model to perform more fixations we can achieve higher accuracy than Learning-to-Zoom (76.5% vs 75.8%) albeit at an increased number of FLOPs. Similarly, given the performance improvement of foveated sampling over uniform downsampling for non-attentive models, we speculate that a similar increase in accuracy could be obtained for a foveated sensor with 224^2 pixels.

4.2 Performance with Diverse Object Scales

A key benefit of foveated sampling is the ability to resolve fine details yet also reason about a wide field of view, with fewer pixels than uniform sampling. We analyse this claim by performing experiments using two augmented variants of the MNIST dataset, where we closely control the scale distribution of digits and avoid any bias regarding the positions of objects. This property is not common in image classification datasets (e.g. Imagenet [6]), which typically exhibit camera bias – a tendency for objects of interest to be centred in the frame and relatively large in proportion to the full image. Our 2 MNIST variants, namely S-MNIST and ST-MNIST (scaled and scaled-translated, respectively), consist of randomly scaled MNIST digits between 1 and 8 times their original size. In S-MNIST, we place the scaled digits at the centre of a 224×224 pixel canvas, whereas in ST-MNIST, we also randomly translate the digit. Crucially, while the images for this task are 224×224 in size, we enforce a constraint that all networks aggressively downsample the input to approximately 28^2 pixels to significantly reduce the computational overhead of the system. In such a scenario, small-scaled digits will have a large portion of salient information about their class destroyed, and attentive methods must learn to adapt their sensors to resolve and classify small digits. Full details of network architectures and implementation details are given in the supplementary material.

Results and discussion. Our results show that adaptive sampling methods can accurately solve the classification task and significantly improve over the uniform downsampling approach (Table 2). In particular, our active approach achieves an 11% performance improve-

Method	Sensor	Separate Localisation Network	Accuracy (%)	
			S-MNIST	ST-MNIST
CNN	Uniform	✗	86.6	86.9
STN	Uniform	✓	93.7	90.1
PTN	Log-Polar	✓	98.4	97.5
Fov-STN	Multi-FoV Crops	✓	98.2	96.9
Fov-STN (Ours)	Our Sensor	✓	98.4	97.6
Sequential (Ours)	Our Sensor	✗	98.4	97.9
Graph-CNN (Ours	Our Sensor	✗	98.3	79.2

Table 2: Accuracy on our MNIST variants with high-scale variation. We show that adaptive and foveated sampling methods significantly improve the models ability to recognize objects over a wide range of scales compared to static uniform methods (CNN)

ment over the uniform downsampling baseline on ST-MNIST and does so with only 66 additional parameters. We show that the adaptive methods can nearly match their ST-MNIST performance showing they accurately learn to locate and attend to small objects in the scene. For additional context, we evaluate our approach while only fixating on the image centre. This model achieves only 79.2% accuracy, showing that attention is indeed needed to solve this task. The full affine spatial transformer performs worse than other adaptive methods by $\sim 7\%$ on ST-MNIST and $\sim 5\%$ on S-MNIST. We observed that training would frequently collapse for spatial transformers, we include results for runs that did not collapse. Notably, the full affine spatial transformer is the only method that adapts its sampling grid by regressing transformation parameters through an MLP. This suggests that a principled computation of transformation parameters performed in the other models is significantly easier to optimise. All the adaptive methods are approximately on par with each other.

4.3 Ablation Experiments

We conduct additional experiments on Imagewoof, a 10-class subset of Imagenet, to probe various aspects of our model. This dataset contains 10 different species of dogs with approximately 900 images per class in training and 400 per class for testing. We take 10% of the training set to create a validation set which we used to tune hyperparameters and select the best checkpoint. The fine-grained nature of this dataset means it remains challenging, while its small size allows us to rapidly perform experiments in a variety of settings. We examine three different aspects of our model – how the radius of the fovea, the number of fixations and different policies for attention affect classification accuracy. We maintain the same experimental set-up and training routine as the Imagenet-100 experiments, except we change the learning rate from 0.004 to 0.0025.

Effect of Fovea Radius. In order to ascertain whether performance increase over a uniform baseline is due to foveation, we sweep over a range of fovea radii. All models are trained with three fixations. We consider 3 variants of our model for this experiment, a 112^2 pixel input variant, a 56^2 pixel variant, and a narrow 56^2 pixel variant with half the number of filters in each layer. We find optimal fovea radius values of 60%, except in the case of the narrow model which is 40%. The benefit of foveation is more pronounced for the scaled down models. Sandler *et al*. [29] observe that, due to the common architectural design choices of CNNs, resolution and width scaling are approximately the same. This can explain why foveation becomes a more important factor when scaling the model down with respect to these dimensions and suggests foveation can provide an increasingly beneficial role in more lightweight CNN architectures.

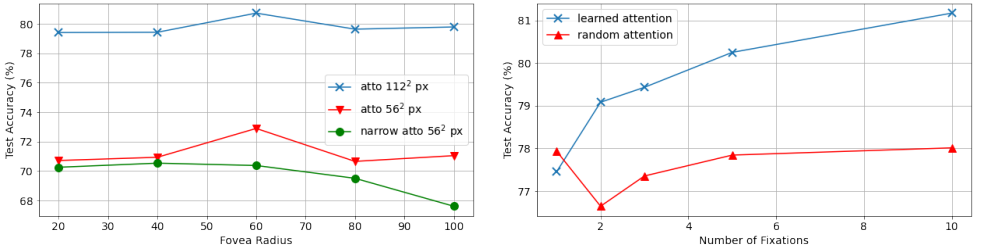

Figure 4: **Left:** Performance of our model with varying radius of fovea (100 corresponds to negligible foveation). **Right:** Performance with varying numbers of fixations; we see that performance increases with more fixations when the attention policy is learned

Effect of Number of Fixations. We investigate how our model scales as the number of fixations increases (Figure 4, right). We consider two different attention policies, our learned attention, and random attention. We show that under a learned attention policy our network continues to result in improved classification accuracy as the number of fixations increases. Noticeably there is an approximately 4% accuracy boost from 1 to 10 fixations. By contrast, the network does not derive significant benefits from extra fixations under a random policy, and even degrades when the number of fixations is low. Given the intuition that object recognition improves when the foveated sensor is centered on the object of interest, we should expect it to decrease when it is not. This can explain the drop in performance with a random policy as our model simply averages predictions across timesteps.

5 Conclusion

In this work, we have presented an end-to-end differentiable active vision model. Our model actively attends to visual scenes and aggregates information over multiple timesteps through a space-variant foveated sensor. We show that our model can outperform a non-attentive CNN at classifying natural images by $\sim 2\%$ with 1 fixation using a 112^2 pixel input and performs better than alternative foveated CNNs. We further show that foveated architectures have particular application to recognizing objects over a high dynamic range of scales. Ablation experiments show that foveation becomes increasingly beneficial for smaller architectures suggesting they can fulfil a useful role in lightweight computer vision systems. There are many avenues for future work with regard to this architecture. More powerful methods for integrating information from multiple fixations could prove beneficial (e.g. self-attention). Additionally, allowing the network to remember previously attended locations and prevent returning to them may help in maximizing the information the system gains when fixating over many time steps.

References

[1] Baptiste Angles, Yuhe Jin, Simon Kornblith, Andrea Tagliasacchi, and Kwang Moo Yi. Mist: Multiple instance spatial transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2412–2422, 2021.

[2] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. In Yoshua Bengio and Yann LeCun, editors, *3rd International*

Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL http://arxiv.org/abs/1412.7755.

[3] Alexandre Boulch. Convpoint: Continuous convolutions for point cloud processing. *Computers & Graphics*, 88:24–34, 2020.

[4] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SkfMWhAqYQ.

[5] Brian Cheung, Eric Weiss, and Bruno A. Olshausen. Emergence of foveal image sampling from learning to attend in visual scenes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJJKxrsgl.

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

[7] Gamaleldin Elsayed, Simon Kornblith, and Quoc V Le. Saccader: Improving accuracy of hard attention models for vision. *Advances in Neural Information Processing Systems*, 32, 2019.

[8] Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HktRlUlAZ.

[9] fastai. Imagenette and imagewoof. URL https://github.com/fastai/imagenette.

[10] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International conference on machine learning*, pages 1462–1471. PMLR, 2015.

[11] 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*, 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Pedro Hermosilla, Tobias Ritschel, Pere-Pau Vázquez, Àlvar Vinacua, and Timo Ropinski. Monte carlo convolution for learning on non-uniformly sampled point clouds. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018.

[14] Jorn-Henrik Jacobsen, Jan Van Gemert, Zhongyu Lou, and Arnold WM Smeulders. Structured receptive fields in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2610–2619, 2016.

[15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[16] Aditya Jonnalagadda, William Yang Wang, B.S. Manjunath, and Miguel Eckstein. Foveater: Foveated transformer for image classification, 2023. URL https://openreview.net/forum?id=Gy8vD-zGQqH.

[17] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[18] Jinpyo Kim, Wooekun Jung, Hyungmo Kim, and Jaejin Lee. Cycnn: A rotation invariant cnn using polar mapping and cylindrical convolution layers. *arXiv preprint arXiv:2007.10588*, 2020.

[19] Xin Li, Zequn Jie, Wei Wang, Changsong Liu, Jimei Yang, Xiaohui Shen, Zhe Lin, Qiang Chen, Shuicheng Yan, and Jiashi Feng. Foveanet: Perspective-aware urban scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 784–792, 2017.

[20] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu. Dynamic computational time for visual attention. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1199–1209, 2017.

[21] Tony Lindeberg. Scale-covariant and scale-invariant gaussian derivative networks. *Journal of Mathematical Imaging and Vision*, 64(3):223–242, 2022.

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11966–11976. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01167. URL https://doi.org/10.1109/CVPR52688.2022.01167.

[23] Hristofor Lukanov, Peter König, and Gordon Pipa. Biologically inspired deep learning model for efficient foveal-peripheral vision. *Frontiers in Computational Neuroscience*, 15:746204, 2021.

[24] José Martínez and Leopoldo Altamirano Robles. A new foveal cartesian geometry approach used for object tracking. *SPPRA*, 6:133–139, 2006.

[25] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2204–2212, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/09c6c3783b4a70054da74f2538ed47c6-Abstract.html.

[26] Piotr Ozimek, Nina Hristozova, Lorinc Balog, and Jan Paul Siebert. A space-variant visual pathway model for data efficient deep learning. *Frontiers in cellular neuroscience*, 13:36, 2019.

[27] Samrudhdhi B Rangrej, Chetan L Srinidhi, and James J Clark. Consistency driven sequential transformers attention model for partially observable scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2518–2527, 2022.

[28] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–66, 2018.

[29] Mark Sandler, Jonathan Baccash, Andrey Zhmoginov, and Andrew Howard. Non-discriminative data or weak model? on the relative importance of data and model resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[30] Eric L Schwartz. Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision research*, 20(8):645–669, 1980.

[31] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.7054.

[32] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3693–3702, 2017.

[33] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[34] Chittesh Thavamani, Mengtian Li, Nicolas Cebron, and Deva Ramanan. Fovea: Foveated image magnification for autonomous navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15539–15548, 2021.

[35] Parsa Torabian, Ronak Pradeep, Jeff Orchard, and Bryan Tripp. Comparison of foveated downsampling techniques in image recognition. *Journal of Computational Vision and Imaging Systems*, 6(1):1–3, 2020.

[36] V Javier Traver and Alexandre Bernardino. A review of log-polar imaging for visual perception in robotics. *Robotics and Autonomous Systems*, 58(4):378–398, 2010.

[37] Helmut Vogel. A better way to construct the sunflower head. *Bellman Prize in Mathematical Biosciences*, 44:179–189, 1979.

[38] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext V2: co-designing and scaling convnets with masked autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pages 16133–16142. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01548. URL https://doi.org/10.1109/CVPR52729.2023.01548.

[39] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 9621–9630, 2019.