Multi-Glimpse Network: A Robust and Efficient Classification Architecture based on Recurrent Downsampled Attention

Sia Huat Tan[1]
csf19@mails.tsinghua.edu.cn
Runpei Dong[2]
runpei.dong@stu.xjtu.edu.cn
Kaisheng Ma[1]
kaisheng@mail.tsinghua.edu.cn

[1] Tsinghua University
[2] Xi'an Jiaotong University

Abstract

Most feedforward convolutional neural networks spend roughly the same efforts for each pixel. Yet human visual recognition is an interaction between eye movements and spatial attention, which we will have several glimpses of an object in different regions. Inspired by this observation, we propose an end-to-end trainable **M**ulti-**G**limpse **N**etwork (**MGNet**) which aims to tackle the challenges of high computation and the lack of robustness based on recurrent downsampled attention mechanism. Specifically, MGNet sequentially selects task-relevant regions of an image to focus on and then adaptively combines all collected information for the final prediction. MGNet expresses higher resistance against adversarial attacks and common corruptions with less computation. Also, MGNet is inherently more interpretable as it explicitly informs us where it focuses during each iteration. Our experiments on ImageNet100 demonstrate the potential of recurrent downsampled attention mechanisms to improve a single feedforward manner. For example, MGNet improves 4.76% accuracy on average in common corruptions with only 36.9% computational cost. Moreover, while the baseline incurs an accuracy drop to 7.6%, MGNet manages to maintain 44.2% accuracy in the same PGD attack strength with ResNet-50 backbone. Our code is available at `https://github.com/siahuat0727/MGNet`.

1 Introduction

Convolutional Neural Networks (CNNs) have achieved promising performance on many visual tasks, such as object detection [12, 49, 50], image segmentation [3, 38] and image captioning [7, 24, 61]. Especially in image classification [27, 31, 66], CNNs can even surpass human performance [17, 19].

However, CNNs are facing various challenges: 1) CNNs are computationally expensive and memory intensive. This increases the difficulty for CNNs to be widely deployed on scenarios like edge-computing; 2) CNNs are vulnerable to adversarial example [13, 44, 57], which is usually an image formed by making a subtle perturbation that leads a trained model to produce an incorrect prediction. This raises major concerns about deploying neural networks in the high-security-demanding systems; 3) CNNs will be confused by many forms of common corruptions [20], such as bad weather, noise, and blur. The lack of robustness is hindering some processes like autonomous vehicle development [32].

Inspiration from the human visual system is a potential hint to solve both the expensive computation and robustness problem. A particularly striking difference between the human visual system and current feedforward convolutional neural networks (FF-Nets) is that the FF-Nets spend enormous and roughly the same amount of computational energy on every single pixel, no matter whether it is essential to the task. Additionally, most FF-Nets process the entire scene just once. The human visual system, by contrast, is not merely feedforward but has various feedback and recurrent connections in the visual cortex [46]. In addition, human beings don't treat an image as a static scene. Instead, cognitive processing is an interaction between attention and eye movements [37]. Specifically, the fovea in the human's eye samples distinct regions of the scene at varying spatial resolutions [59]. The series of fixation on different location and resolution are then collected and integrated to build up an internal representation of the scene [51].

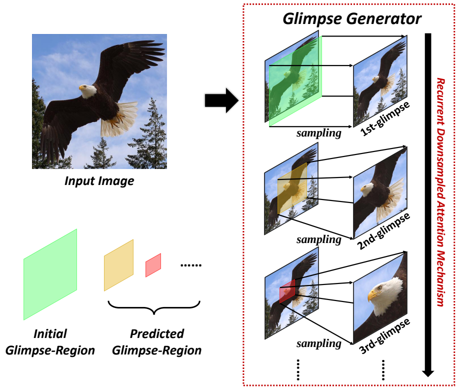

Figure 1: Illustration of the recurrent downsampled attention mechanism. From top to down, the Glimpse Generator sequentially generates glimpses by sampling from the given glimpse-regions in a recurrent manner.

Inspired by the the sequential and variable resolution sampling mechanisms in the human visual system, we propose Recurrent Downsampled Attention (RDA) mechanism and present a novel Multi-Glimpse Network (MGNet) to explore the benefits of deploying RDA in CNNs. Instead of sweeping the entire scene at once, our model sequentially select to focus on some task-relevant regions (illustrated in Figure 1). During each iteration, our model will first apply variable resolution sampling to a various size regions of the original image to produce a much lower dimensionality fixation, which we will refer to as glimpse [33]. Every glimpse will be integrated over time to build up a global internal representation. Since our model only mainly computes on these low dimensionality glimpses, the model can save computational cost. Unlike other model acceleration methods, such as network pruning [15], knowledge distillation [23], quantization [15, 28], and model compacting [53], we break the current paradigm that sweeps the image just once and predicts. By sequentially processing multiple glimpses, we further show that our model is fundamentally more robust against the adversarial attacks and common corruptions.

Our main contributions can be summarized as follows:

- We propose Multi-Glimpse Network, which is end-to-end trainable in one-stage while not requiring any supervised spatial guidance or hand-crafted pre-training method.
- With the same amount of computational cost, we demonstrate that MGNet outperforms FF-Nets with various backbones. Additionally, as the network is shared over

iterations, it can decide to early-exit on-the-fly without adding any overhead.

- We show that MGNet is intrinsically more robust against adversarial attacks and common corruptions. For example, accuracy is improved by 4.76% in common corruptions with 36.9% computational requirement in average.

2 Related Work

Robustness. Szegedy *et al.* [57] first show that a carefully perturbed image can fool a trained model entirely in high confidence. Goodfellow *et al.* [13] propose FGSM to generate adversarial examples. Madry *et al.* [41] study the adversarial robustness of neural networks and propose a robust minimax optimization called PGD adversarial training. The research direction in studying adversarial attack and defense method is in the progress [43, 47, 65]. Besides, Hendrycks and Dietterich [20] consider common real-world corruptions and propose a benchmark to measure general robustness. Recently, various data augmentation techniques [11, 21, 22] are introduced to improve the general robustness.
Computational Efficiency. Many research work have been proposed to reduce the computational cost of deep neural networks. As there are considerable redundant parameters in neural networks, some focus on pruning the non-essential connections to reduce computational cost [15, 55, 58]. Another approach is quantization, which focuses on compressing the bit-width of weights for floating-point operations and memory usage reduction [4, 48]. Hinton *et al.* [23] propose knowledge distillation where the student learns to mimic the teacher's prediction results. This technique has been widely used to transfer the knowledge from larger models into compact models [39, 52]. Recent works further reduce computation by designing efficient network architectures [25, 26, 40].
Recurrent Attention Model. Recurrent attention mechanism has been explored in many fields, such as reinforcement learning [16, 45], machine translation [1, 2, 10], image classification [30, 42, 68] and generative models [8, 36, 67]. In the vision task, Mnih *et al.* [42] first propose a recurrent visual attention model to control the amount of computation on the augmented MNIST dataset [35]. While the model is not differentiable, it is trained using reinforcement learning. Gregor *et al.* [14] propose differentiable attention mechanisms to generate images sequentially. Jaderberg *et al.* [29] show that meaningful object parts can be discovered automatically with only image labels. Fu *et al.* [9] propose a recurrent attention model to learn region-based feature representation at multiple scales in fine-grained image classification. Zoran *et al.* [68] show that an adversarially trained sequential attention network is significantly more robust than a feedforward model.

Since each recurrent attention-related work has a different focus, most of them are designed experimentally using multiple model capacity or computational cost or both. In this work, with the proposed RDA and MGNet, we aim to answer a question: *given the same model capacity and computational cost, is it beneficial to introduce recurrent mechanism in CNNs?* Our experiments further show that MGNet is intrinsically more robust against adversarial attacks, and a low-dimensionality glimpse is crucial to improve general robustness.

3 Approach

In this section, we present an overview of our proposed MGNet, as illustrated in Figure 2. Instead of blindly carrying out a large amount of computation for every single pixel of an

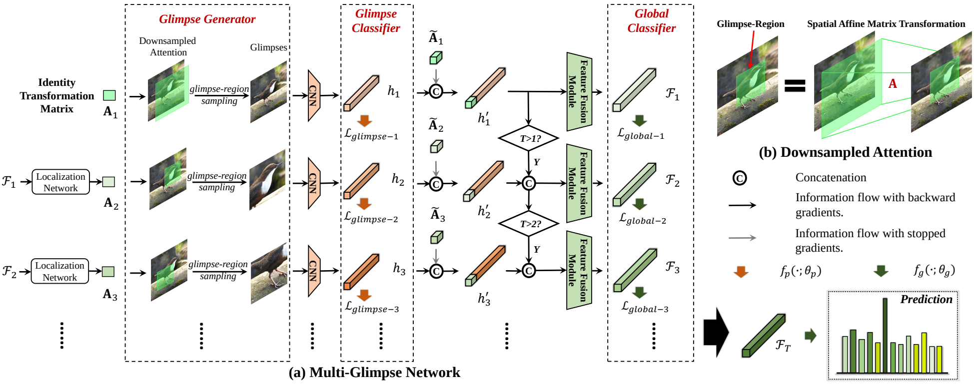

Figure 2: Details of our method: (a) The framework of MGNet. Glimpses are generated by sequentially sampling the image from the glimpse-region. The Glimpse Classifier guides to make every glimpse count and will be dropped after training. Multi glimpse features are integrated by Feature Fusion Module into a global feature. The global feature will be decoded to predict the label and the next glimpse-region. Note that we share all the parameters during the iterations. (b) Illustration of the downsampled attention. *Best viewed in color.*

image, our model will sequentially generate T glimpses to be processed and fuse all the glimpses for the final prediction.

Given an image $x \in \mathbb{R}^{H \times W}$, H and W respectively denote the height and width of the image. For the t-th iteration, the Glimpse Generator g will apply affine transformation to the input image and perform sampling to produce a glimpse $x_t^g = g(x, \mathbf{A}_t; M)$, where $x_t^g \in \mathbb{R}^{\frac{H}{M} \times \frac{W}{M}}$, M is a downsampling factor and $\mathbf{A}_t$ is the t-th affine transformation matrix. The downsampling factor M is fixed and greater than 1 to reduce the amount of computation. $\mathbf{A}_t$ is generated by the Localization Network, except for the initial matrix $\mathbf{A}_1$ which we set as an identity transformation matrix. Therefore, the first glimpse will be a low-resolution version of the original image. We will introduce the Glimpse Generator in Section 3.1.

The glimpse x_t^g is first encoded by a CNN backbone (including global average pooling) to produce a glimpse feature h_t. Each glimpse feature will be decoded by a glimpse classifier $f_p(\cdot;\theta_p)$ into class logits to make every glimpse count. The affine transformation matrix $\mathbf{A}_t$ will be flattened as $\tilde{\mathbf{A}}_t$ and appended to the glimpse feature h_t. We stop the gradient on $\tilde{\mathbf{A}}_t$ as it is a positional encoding that can help the model understands where the feature comes from. Then all glimpse features will be integrated by a Feature Fusion Module to produce global internal representation $\mathcal{F}_t$ of the image during the t-th iteration. This module will be introduced in Section 3.2.

With a fully-connected layer $f_g(\cdot;\theta_g)$ as the global classifier, we decode $\mathcal{F}_t$ into class logits iteratively to produce T classification results. Note that the decoded result of $\mathcal{F}_T$ will be the final prediction. $\mathcal{F}_t$ will also be fed into the Localization Network to generate the next glimpse-region (if needed), and more details can be found in Section 3.3.

3.1 Glimpse Generator

This module aims to generate low-dimensionality glimpses. The non-differentiability of cropping and resizing makes it difficult to learn where to look, which can be addressed with reinforcement methods such as policy gradient [42]. We will briefly introduce a differentiable affine transformation operation proposed by Jaderberg *et al.* [29], making it possible to be trained end-to-end with SGD.

We first generate a 2D flow field (we call it glimpse-region) by applying a parameterized sampling grid with an affine transformation matrix $\mathbf{A}$. Since we only consider cropping, translation, and isotropic scaling transformations, $\mathbf{A}$ is more constrained and requires only 3 parameters,

$$\mathbf{A} = \begin{bmatrix} a^s & 0 & a^x \\ 0 & a^s & a^y \end{bmatrix}, \tag{1}$$

where a^s, a^x, and a^y are the output of the Localization Network (details in Section 3.3).

To generate a glimpse x^g, we first perform a pointwise transformation

$$\begin{pmatrix} x_{t_x} \\ x_{t_y} \end{pmatrix} = \mathbf{A} \begin{pmatrix} x^g_{t_x} \\ x^g_{t_y} \\ 1 \end{pmatrix}, \tag{2}$$

where (x_{t_x}, x_{t_y}) are the coordinates of the regular grid in the input image x, and $(x^g_{t_x}, x^g_{t_y})$ are the coordinates that define the sample points. Then we apply a bilinear sampling to generate a glimpse $x^g \in \mathbb{R}^{\frac{H}{M} \times \frac{W}{M}}$. Especially, for the first glimpse, we let a^s equal to 1 and a^x, a^y equal to 0, which denote an identity transformation. Since the downsampling factor M is greater than 1, the first glimpse represents a low-resolution version of the input image. The differentiability of this affine transformation allows our model to learn the task-relevant regions with backpropagation.

3.2 Feature Fusion Module

It is crucial to integrate the information of every glimpse to make the final prediction. In this section, we introduce our Feature Fusion Module, using attention mechanism [60] with a single attention head, to integrate all the glimpse features $h_1, h_2, \cdots, h_t$ into a global internal feature $\mathcal{F}_t$. Specifically, for the t-th iteration,

$$\begin{aligned} \mathbf{H}_t &= \text{concatenate}([h_1', h_2', \cdots, h_t']), \\ \mathcal{E}_t &= \text{softmax}(\frac{(\mathbf{H}_t \mathbf{W}^q)(\mathbf{H}_t \mathbf{W}^k)^\top}{\sqrt{d}})(\mathbf{H}_t \mathbf{W}^v)\mathbf{W}^o, \\ \mathcal{F}_t &= \text{ReLU}(\text{LayerNorm}(\mathcal{E}_t)[t]), \end{aligned} \tag{3}$$

where $h_t' \in \mathbb{R}^d$ is the glimpse feature h_t concatenated with the positional encoding, $\mathbf{W}^q, \mathbf{W}^k$, $\mathbf{W}^v, \mathbf{W}^o \in \mathbb{R}^{d \times d}$ are the learnable parameters, $\mathcal{F}_t \in \mathbb{R}^d$ is the global internal representation integrated during the t-th iterations, and the notation $\mathbf{X}[t]$ represents the t-th row of the matrix $\mathbf{X}$. Note that for an experiment setting with T iterations, $\mathcal{F}_T$ represents the final feature and will be decoded by the global classifier $f_g(\cdot; \theta_g)$ to predict the label.

3.3 Localization Network

We propose Localization Network to predict an affine transformation matrix $\mathbf{A}$ for the glimpse generation. More intuitively, $\mathbf{A}$ can represent a target region of the input image, where the parameter a^s is the ratio of the size of the glimpse-region to the input image, a^x and a^y denote the translation of the region origin. In MGNet, we let $a^s \in [a^s_{\min}, a^s_{\max}]$ so that the glimpse-region size is adaptive, where $a^s_{\min} = 0.2$ and $a^s_{\max} = 0.5$. Since we prevent the Glimpse Generator from sampling beyond the image range, the range of a^x and a^y should be within $[a^s - 1, 1 - a^s]$. In detail, given a t-th global internal representation $\mathcal{F}_t$, we produce the parameter of matrix $\mathbf{A}_{t+1}$ by

$$\begin{aligned} [a^s_{t+1}{}', a^x_{t+1}{}', a^y_{t+1}{}'] &= \Phi(\sigma(f_l(\mathcal{F}_t; \theta_l)); s), \\ a^s_{t+1} &= a^s_{t+1}{}' \cdot (a^s_{max} - a^s_{min}) + a^s_{min}, \\ [a^x_{t+1}, a^y_{t+1}] &= (2 \cdot [a^x_{t+1}{}', a^y_{t+1}{}'] - 1) \cdot (1 - a^s_{t+1}), \end{aligned} \tag{4}$$

where σ is sigmoid function, $f_l(\cdot; \theta_l)$ is a fully-connected layer, Φ is a gradient re-scaling operation and s is a gradient re-scaling factor. The gradient re-scaling operation

$$\Phi(x; s) = x; \quad \nabla_x \Phi(x; s) = s \tag{5}$$

is applied to tackle the gradient issue as we empirically find an exploding gradient problem in the Localization Network. The value of s is possibly around 0.01 to 0.02 in our setting. We show the hyper-parameter tuning in Supplementary Material Section 1.

3.4 Joint Classifiers Learning

Given dataset $D = \{x^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^N$ where $x^{(i)}$ denotes the i-th input image and $\mathbf{y}^{(i)}$ is the corresponding label, MGNet jointly learns the glimpse feature together with the global internal feature in an end-to-end fashion. To realistically demonstrate the model's potential, we train our model on pure Cross-Entropy (CE) loss and hence the total loss $\mathcal{L}$ can be given as

$$\mathcal{L} = \alpha \mathcal{L}_{glimpse} + (1 - \alpha) \mathcal{L}_{global}, \tag{6}$$

where $\mathcal{L}_{glimpse}$ is the glimpse classifier loss, $\mathcal{L}_{global}$ is the global classifier loss, and α is a hyper-parameter that balances the weighting between the losses.

As shown in Figure 2, the glimpse classifier can be regarded as an auxiliary loss and will be dropped after training, so neither extra memory nor computation power is required during inference. We show the effect of glimpse classifier in Supplementary Material Section 1.

Global Classifier. During the t-th iteration, the global classifier takes $\mathcal{F}_t$ as input to make a global prediction, and it is trained by averaging all t-th prediction loss $\mathcal{L}_{global-t}$:

$$\mathcal{L}_{global-t} = \mathbb{E}_{x,y \sim D}[\mathcal{H}(y, f_g(\mathcal{F}_t; \theta_g))], \quad \mathcal{L}_{global} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_{global-t}, \tag{7}$$

where $\mathcal{H}$ denotes the CE loss and T is the number of glimpses.

Glimpse Classifier. Similarly, the glimpse classifier takes h_t as input and jointly learns by averaging all t-th glimpse loss $\mathcal{L}_{glimpse-t}$:

$$\mathcal{L}_{glimpse-t} = \mathbb{E}_{x,y \sim D}[\mathcal{H}(y, f_p(h_t; \theta_p))], \quad \mathcal{L}_{glimpse} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_{glimpse-t}. \tag{8}$$

4 Experimental Results

4.1 ImageNet100

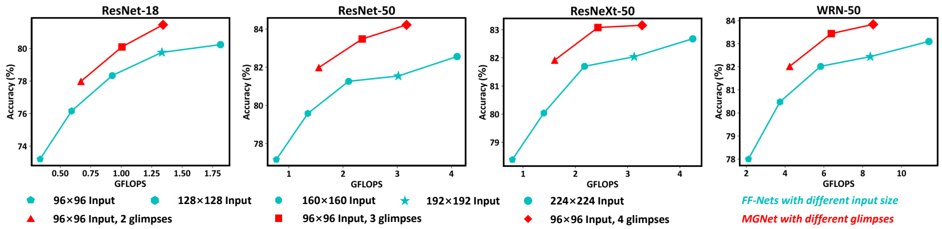

Figure 3: Top-1 accuracy (%) comparison between FF-Nets and MGNet in terms of computational cost on ImageNet100. MGNet is trained once and exit on the different number of glimpses to show the accuracy of early-exit. FF-Nets with different input sizes are trained separately to explore the trade-off between the accuracy and computation of one-pass strategy. The results show that given the same model capacity, MGNet consistently outperforms FF-Nets among various backbones while having fewer computation.

In this section, we evaluate MGNet on ImageNet100, which is the first 100 classes of ImageNet [6]. We demonstrate some experiments on toy datasets in Supplementary Material Section 2 to better understand how the RDA mechanism works.

Network		GFLOPs	Latency (ms)	Accuracy (%)
ResNet-18	FF-Net	1.815	87.7	80.24
	MGNet	**1.343**	**59.4**	**81.46**
ResNet-50	FF-Net	4.104	240.1	82.56
	MGNet	**3.172**	**167.6**	**84.22**
ResNeXt-50	FF-Net	4.246	313.1	82.68
	MGNet	**3.276**	**198.3**	**83.16**
WRN-50	FF-Net	11.413	486.6	83.10
	MGNet	**8.542**	**369.0**	**83.84**

Table 1: GFLOPs and inference latency on ImageNet100.

We implement FF-Net as a special case of MGNet with the number of glimpses $T = 1$ and downsampling factor $M = 1$, which means the Glimpse Generator performs identity transformation without downsampling. As our comparison does not depend on backbone architecture, we evaluate it with ResNet-18 [18], ResNet-50 [18], ResNeXt-50 [64], and WRN-50 [66] backbones. To ensure the models' convergence to sufficiently demonstrate their capability, we train both FF-Nets and MGNet in 400 epochs with SGD. The peak learning rate is set to be 0.1 using a one-cycle scheduler [54]. For data augmentation, we train models with Auto Augmentation [5]. For MGNet, we set total glimpses $T = 4$ and downsampling factor $M = 7/3$, which still requires less computation than baseline. The hyper-parameter α is set to be 0.6, s is 0.02 for ResNet-18, and 0.01 otherwise.

We present a fair comparison in terms of the number of parameters, backbone architecture, training settings, and computational cost. The following experiments show the potential of MGNet to simultaneously reduce computation, improve adversarial robustness, enhance general robustness and be more interpretable in real-world datasets. Visualization of success and failure cases are shown in Supplementary Material Section 3.

Network			GFLOPs	Average	Noise			Blur				Weather				Digital			
					Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	JPEG
ResNet-18	FF-Nets		1.8146	46.21	36	37	32	28	35	42	41	41	47	**62**	71	52	60	59	52
	MGNet	1-glimpse	0.3342	50.56	44	41	39	**38**	**48**	48	43	37	47	50	69	56	66	70	**63**
		2-glimpse	0.6695	52.77	**45**	**43**	**40**	**38**	**48**	**49**	47	41	**51**	57	73	**58**	68	**71**	**63**
		3-glimpse	1.0058	**53.23**	**45**	**43**	**40**	**38**	47	**49**	**49**	**43**	**51**	59	**74**	**58**	**69**	**71**	62
ResNet-50	FF-Nets		4.1042	53.24	46	46	42	37	43	49	49	**47**	54	**64**	76	60	64	64	59
	MGNet	1-glimpse	0.7677	55.03	50	46	46	43	**53**	50	47	42	53	53	73	60	70	**73**	66
		2-glimpse	1.5523	**57.36**	**51**	**48**	**47**	**43**	**53**	**52**	**54**	**47**	**56**	60	**77**	**61**	**72**	**73**	**67**
ResNeXt-50	FF-Nets		4.2455	53.01	47	47	42	37	42	47	47	**48**	54	**63**	**76**	**59**	64	62	58
	MGNet	1-glimpse	0.7937	55.57	**51**	49	47	**43**	**52**	50	46	45	56	54	74	**59**	70	**73**	**65**
		2-glimpse	1.6042	**57.13**	**51**	**50**	**49**	42	**52**	**51**	**52**	**48**	**58**	59	**76**	**59**	**71**	**73**	**65**
WRN-50	FF-Nets		11.413	54.75	48	49	45	39	46	49	49	**49**	55	**64**	**77**	61	65	66	60
	MGNet	1-glimpse	2.1101	56.76	53	50	49	45	55	51	48	44	55	56	74	60	70	74	67
		2-glimpse	4.2372	**59.02**	**54**	**51**	**50**	**46**	**55**	**52**	**55**	**49**	**58**	62	**77**	**62**	**73**	**75**	**68**

Table 2: Top-1 accuracy (%) evaluation of MGNet and FF-Nets on ImageNet100-C.

4.1.1 Early-Exit

Early-exit allows a model to be trained once and specialized for efficient deployment, addressing the challenge of efficient inference across resource-constrained devices such as edge-devices [62]. MGNet is designed to process multi-glimpse sequentially; hence it can naturally early-exit without adding any overhead.

Table 1 shows that *given the same model capacity*, MGNet with four 96 × 96 glimpses always outperforms FF-Nets with standard 224 × 224 inputs while holding less computation. For the latency in the practical usage, we are testing on Intel Xeon E5-2650 without GPU. Additionally, since the input is smaller for each forward pass, MGNet requires noticeably less memory (e.g., reduce by 26.4% in ResNet-18). Therefore, the acceleration is more prominent when the memory resources are limited. We further demonstrate the early exits' accuracy of the same MGNet and train FF-Nets individually with various input sizes to explore the trade-off between these two manners' computational cost and performance. We observe that RDA mechanisms can consistently outperform the one-pass manner among various backbones. As shown in Figure 3, with the same backbone ResNet-50, MGNet with four 96 × 96 glimpses outperforms FF-Net with a full 224 × 224 input by 1.66% accuracy, while the computation is only about 77.28% of the latter. For ResNeXt-50, MGNet with two 96 × 96 glimpses matches the performance of FF-Net with 192 × 192 input while requiring only 51.36% computation. **This experiment shows that an image classifier can be more efficient and effective by including RDA mechanisms.**

4.1.2 Common Corruptions

The models we train on clean data are directly evaluated on the common corruptions benchmark [20] (reduced to 100 classes) ImageNet100-C, which consists of 15 different corruption types generated algorithmically from noise, blur, weather, and digital categories. Each corruption type has five severity levels, so the total number of corruption types is 75.

Table 2 shows that MGNet yields a substantial improvement in general robustness compared to FF-Nets. For example, MGNet with ResNet-18 backbone with three glimpses increases the average accuracy by 6.56% compared to FF-Nets, while the computational cost is merely 55% of the latter. On average, MGNet with two glimpses outperforms FF-Nets by 4.76% with only 36.9% computational cost. **The progress of MGNet perceiving from a rough overview to detailed parts makes it more robust, even with a single glimpse.**

4.1.3 Adversarial Robustness

Recent work show that deep neural networks can be simply fooled by adversarial examples [13, 57]. In this section, we compare the adversarial robustness between FF-Nets and MGNet without adversarial training [41].

FGSM [13] is one of the most popular methods to generate adversarial examples during a single iteration,

$$x + \varepsilon \cdot \mathrm{sgn}(\nabla_x L(\theta, x, y)), \tag{9}$$

where x is an input image, y is the label, θ denotes the parameters, L is the loss function, sgn returns the sign, and ε is the attack step size. PGD [41] is an iterative variant of FGSM,

$$x^{k+1} = \Pi_{x+\mathcal{S}}\; x^k + \varepsilon \cdot \mathrm{sgn}(\nabla_x L(\theta, x, y)), \tag{10}$$

where k is the iteration index and $\mathcal{S}$ denotes the set of perturbations that formalizes the manipulative power of the adversary. In the following experiments, we consider the PGD attacks with $4/255$ ℓ_∞-bounded and step size $\varepsilon = 1/255$ on different numbers of steps.

As shown in Figure 4, with the same strength of the PGD attacks, the adversarial robustness of MGNet significantly outperforms FF-Nets. For example, with four attack steps, the top-1 accuracy of FF-Net with ResNet-50 drastically drops to 7.6%, while MGNet still maintains 44.2%. Even with 300 attack steps, the accuracy of MGNet still maintains 10.86% while FF-Nets drops to 0.96% with only 20 attack steps. The result is consistent across various backbones. We infer that the increment of robustness may come from the ensemble, but MGNet even requires less computational cost than a single pass of FF-Nets. Note that we intend to show the intrinsic feature of MGNet against adversarial attacks rather than propose a defense method. Besides, the one-stage end-to-end trainable property allows MGNet to be combined with various adversarial defense methods to achieve higher adversarial robustness.

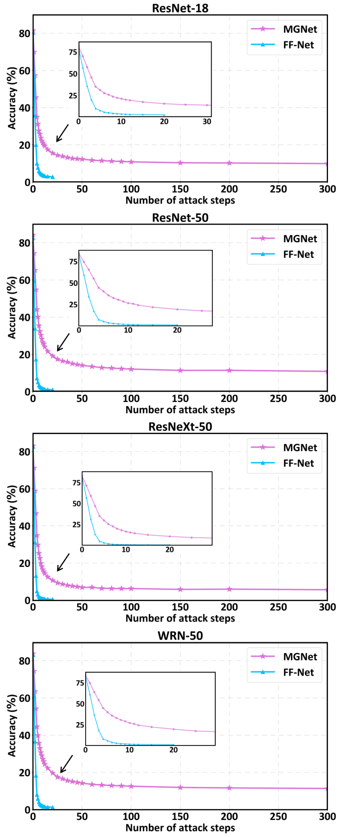

Figure 4: The top-1 accuracy performance comparison over different number of PGD attack without adversarial training on ImageNet100.

4.2 Tiny ImageNet

We evaluate MGNet on Tiny ImageNet [34] to explore the performance on images with lower resolution. Tiny ImageNet is a subset of ImageNet. It includes 200 distinct categories, and each contains 500 training images, 50 validation images, and 50 test images. All the images are resized to 64 × 64 pixels, where the original size is 224 × 224 pixels on ImageNet.

We select downsampling factor $M = 2$ and total glimpses $T = 3$ for MGNet to make an appropriate comparison with FF-Nets. In this setting, MGNet will receive three 32×32 pixels glimpses while FF-Nets, as usual, will receive a 64×64 pixels image. We first compare our baseline implementation with [56]. Next, same as [63], we remove the max-pooling layer followed by the first convolutional layer as we will reduce the input image size further to 32×32 pixels. Note that these networks are initially designed for 224×224 pixels images. We use the notation † to mark modified networks that we select as the backbones to compare FF-Nets and MGNet.

Network		GFLOPs	Accuracy (%)
ResNet-18	[56]	0.1497	52.40
	Ours	0.1497	53.97
ResNet-18†	FF-Net	0.5657	57.14
	MGNet	0.4301	57.72
ResNet-34	[56]	0.3009	53.20
	Ours	0.3009	55.08
ResNet-34†	FF-Net	1.1705	58.71
	MGNet	0.8837	58.38

† No max-pooling layer followed by the first convolutional layer.

Table 3: GFLOPs and accuracy (%) evaluation on Tiny ImageNet.

As shown in Table 3, the feedforward baselines of our implementation are slightly higher than [56] baselines. It can benefit from our learning-rate scheduler choice and the larger training epochs that ensure the models are fully converged. In these experiments, we show the potential of RDA mechanism to reduce computation while maintaining accuracy in smaller image scales. For example, using ResNet-18† as the backbone, FF-Net and MGNet achieve a comparable accuracy while the latter requires only 76% FLOPs. This improvement may not be so significant at larger image scales. Nevertheless, we claim these results are reasonable because the smaller the image is, the less redundant computing is spent on unimportant regions.

5 Conclusion

In this paper, we explore the capability of a recurrent downsampled attention mechanism based model for image classification. MGNet achieves comparable predictive performance on ImageNet100 while holding several benefits: 1) requires less computation amount; 2) can early-exit on-the-fly; 3) is intrinsically more robust against adversarial attacks and common corruptions; and 4) explicitly informs more spatial information. Furthermore, we can directly train MGNet in an end-to-end manner from scratch.

Although we intuitively propose to train MGNet by gradient re-scaling, it harms the convergence speed, and such that we cannot afford to explore MGNet on ImageNet dataset. Future work can focus on tackling this problem or improving MGNet submodules.

Beyond that, there is no apparent limitation for MGNet to be combined with recent work such as pruning, quantization, knowledge distillation, and adversarial defense methods to achieve more promising performance. We hope that this work will spur the related research direction that focuses on the exploration of recurrent downsampled attention mechanism to improve vision models further.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Int. Conf. Learn. Represent. (ICLR)*, 2015.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Int. Conf. Learn. Represent. (ICLR)*, 2015.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 40(4):834–848, 2018.

[4] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Towards the limit of network quantization. In *Int. Conf. Learn. Represent. (ICLR)*, 2017.

[5] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 248–255, 2009.

[7] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. pages 100–105, 2015.

[8] S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, infer, repeat: Fast scene understanding with generative models. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pages 3225–3233, 2016.

[9] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4476–4484, 2017.

[10] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 70, pages 1243–1252, 2017.

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Int. Conf. Learn. Represent. (ICLR)*, 2019.

[12] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 580–587, 2014.

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Int. Conf. Learn. Represent. (ICLR)*, 2015.

[14] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 37, pages 1462–1471, 2015.

[15] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[16] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1229–1238, 2016.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Int. Conf. Comput. Vis. (ICCV)*, pages 1026–1034, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 770–778, 2016.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 770–778, 2016.

[20] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Int. Conf. Learn. Represent. (ICLR)*, 2019.

[21] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *CoRR*, abs/2006.16241, 2020.

[22] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *Int. Conf. Learn. Represent. (ICLR)*, 2020.

[23] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[24] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *CoRR*, abs/1810.04020, 2018.

[25] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[26] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 42(8):2011–2023, 2020.

[27] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[28] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pages 4107–4115, 2016.

[29] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pages 2017–2025, 2015.

[30] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. *CoRR*, abs/2103.03206, 2021.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017.

[32] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. A survey of deep learning applications to autonomous vehicle control. *CoRR*, abs/1912.10773, 2019.

[33] Hugo Larochelle and Geoffrey E. Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pages 1243–1251, 2010.

[34] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[35] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[36] Renjie Liao, Yujia Li, Yang Song, Shenlong Wang, William L. Hamilton, David Duvenaud, Raquel Urtasun, and Richard S. Zemel. Efficient graph generation with graph recurrent attention networks. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pages 4257–4267, 2019.

[37] S. P. Liversedge and J. Findlay. Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, 4:6–14, 2000.

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3431–3440, 2015.

[39] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, and Xiaoou Tang. Face model compression by distilling knowledge from neurons. In *AAAI Conf. Artif. Intell. (AAAI)*, pages 3560–3566, 2016.

[40] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet V2: practical guidelines for efficient CNN architecture design. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 11218, pages 122–138, 2018.

[41] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Int. Conf. Learn. Represent. (ICLR)*, 2018.

[42] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *CoRR*, abs/1406.6247, 2014.

[43] Aamir Mustafa, Salman H. Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. In *Int. Conf. Comput. Vis. (ICCV)*, pages 3384–3393, 2019.

[44] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 427–436, 2015.

[45] Yael Niv, Reka Daniel, Andra Geana, Samuel J Gershman, Yuan Chang Leong, Angela Radulescu, and Robert C Wilson. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21):8145–8157, 2015.

[46] Bruno A Olshausen. 20 years of learning about vision: Questions answered, questions unanswered, and questions not yet asked. In *20 Years of Computational Neuroscience*, pages 243–270. 2013.

[47] Yao Qin, Nicholas Frosst, Colin Raffel, Garrison W. Cottrell, and Geoffrey E. Hinton. Deflecting adversarial attacks. *CoRR*, abs/2002.07405, 2020.

[48] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 9908 of *Lecture Notes in Computer Science*, pages 525–542, 2016.

[49] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 779–788, 2016.

[50] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pages 91–99, 2015.

[51] Ronald A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7:17 – 42, 2000.

[52] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *Int. Conf. Learn. Represent. (ICLR)*, 2015.

[53] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4510–4520, 2018.

[54] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019.

[55] Suraj Srinivas and R. Venkatesh Babu. Data-free parameter pruning for deep neural networks. In *Brit. Mach. Vis. Conf. (BMVC)*, pages 31.1–31.12, 2015.

[56] Lei Sun. Resnet on tiny imagenet. *Submitted on*, 14, 2016.

[57] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Int. Conf. Learn. Represent. (ICLR)*, 2014.

[58] Karen Ullrich, Edward Meeds, and Max Welling. Soft weight-sharing for neural network compression. In *Int. Conf. Learn. Represent. (ICLR)*, 2017.

[59] David C Van Essen and Charles H Anderson. Information processing strategies and pathways in the primate visual system. *An introduction to neural and electronic networks*, 2:45–76, 1995.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pages 5998–6008, 2017.

[61] Qingzhong Wang and Antoni B. Chan. CNN+CNN: convolutional decoders for image captioning. *CoRR*, abs/1805.09019, 2018.

[62] Xiaofei Wang, Yiwen Han, Victor C. M. Leung, Dusit Niyato, Xueqiang Yan, and Xu Chen. Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Commun. Surv. Tutorials*, 22(2):869–904, 2020.

[63] Jiayu Wu, Qixiang Zhang, and Guoxi Xu. Tiny imagenet challenge. *Technical Report*, 2017.

[64] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5987–5995, 2017.

[65] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *Int. J. Autom. Comput.*, 17(2):151–178, 2020.

[66] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Brit. Mach. Vis. Conf. (BMVC)*, 2016.

[67] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *CoRR*, abs/1805.08318, 2018.

[68] Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, and Pushmeet Kohli. Towards robust image classification using sequential attention models. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 9480–9489, 2020.