Supplementary Material

1 Study on Hyper-parameters

We introduce two hyper-parameters in this work: gradient scaling factor s to tackle the exploding gradient problem in the localization network and α to balance the losses between the two classifiers. Figure 1 (a) shows that a suitable s works to improve training stability while (b) demonstrates that an appropriate balance between the global classifier and the glimpse classifier boosts accuracy.

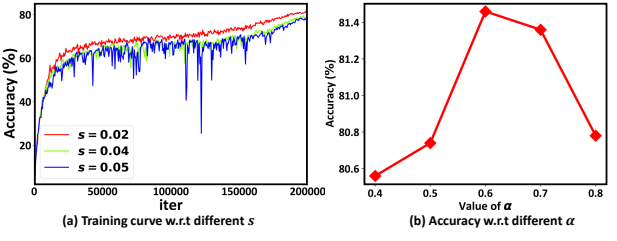

Figure 1: Study of different hyper-parameters with ResNet-18 on ImageNet100.

2 Experiments on Toy Datasets

We consider the same vanilla CNN backbone for FF-Nets and MGNet in the following experiments, which has four convolutional layers with 16, 32, 64, and 128 filters of kernel size 5, 5, 3, 3, respectively, followed by a global average pooling and a fully-connected layer.

Translated MNIST (T-MNIST). First, we consider the T-MNIST dataset to show the capability of MGNet to capture accurate task-relevant regions. The data are generated on-the-fly by placing each 28 $\times$ 28 pixels MNIST digit in a random location of a 112 $\times$ 112 blank patch. We let downsampling factor $M = 4$ for our MGNet, which means each glimpse will be 28 $\times$ 28 pixels. We also train an FF-Net as a baseline, which processes the full-resolution image at once.

Table 1 shows the comparison of the accuracy and computational cost of FF-Net and MGNet. MGNet stops at number of glimpses $T = 3$ as it matches the baseline accuracy, and since it computes on low-dimensionality glimpses, the computation amount can be reduced from 67.95 MFLOPs to 10.619 MFLOPs, which brings $\sim 7\times$ computational efficiency boost. Some samples of generated glimpses are shown in Figure 2 (a) to visualize the glimpse-regions. Note that we explicitly add an ℓ_2-norm to the size of the glimpse-region

Dataset	Structure		MFLOPs	Accuracy(%)
Translated MNIST	FF-Nets		67.950	99.48
	MGNet	1-glimpse	3.471	84.37
		2-glimpse	7.011	99.31
		3-glimpse	10.619	99.54
Gaussian Noise T-MNIST	FF-Nets		67.950	99.10
	MGNet	1-glimpse	3.471	63.19
		2-glimpse	7.011	97.88
		3-glimpse	10.619	98.87
		4-glimpse	14.296	99.02
		5-glimpse	18.042	99.07

Table 1: The comparison of the top-1 accuracy (%) and MFLOPs between FF-Net and MGNet. FF-Nets sweep the full 112×112 pixels input at once while our MGNet processes several glimpses of 28×28 pixels.

(i.e. $\left\|a^s - a^s_{\min}\right\|_2$) to enhance the ability of MGNet to capture precise location without supervised spatial guidance.

Gaussian Noise T-MNIST (GT-MNIST). Second, we consider the GT-MNIST dataset, which is generated by adding zero-mean Gaussian noise with a standard deviation of 0.3 to T-MNIST, followed by clipping to maintain the proper image data range. The experiment setup is the same as the previous one. As shown in Figure 2 (b), high-intensity downsampling hinders the capture of a task-relevant region because the first glimpse's spatial information may be ambiguous. However, by sequentially sampling more glimpses, MGNet matches the baseline with $\sim 4\times$ computational efficiency boost as shown in Table 1. This experiment further explores the ability of MGNet to integrate the information over multiple glimpses.

We further increase the noise intensity to study the behavior of MGNet. Interestingly, as shown in Figure 2 (c), if the first glimpse does not provide accurate spatial information, the glimpse-region adaptively grows larger to be more perceptive (the second example), and a search is performed to find the task-relevant region.

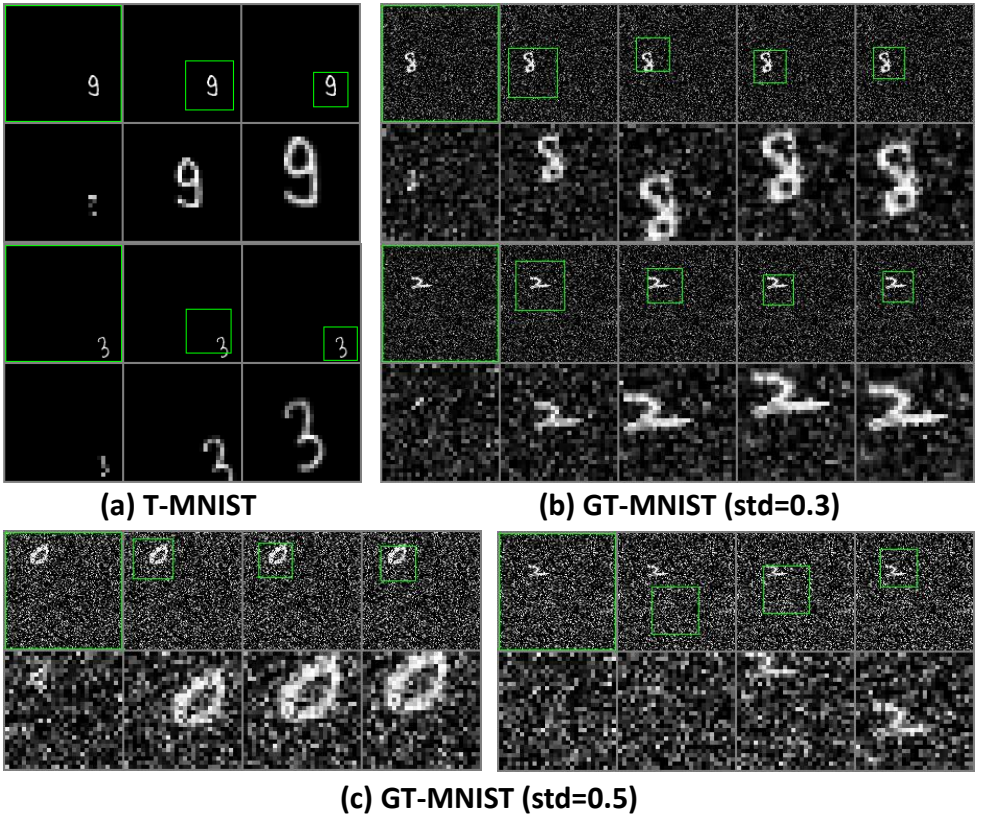

Figure 2: Visualization of the glimpses series generated by MGNet on various datasets. For each series, the first row shows the original image, each with a green box represent the glimpse-region, while the second row shows the generated glimpses (upsampled for better visualization) sampling from the glimpse-region.

3 Visualization of glimpses

Since the glimpse-region is dynamically obtained to capture the task-relevant region fields well, the real-world scene's visualization will illustrate this recurrent attention procedure. We visualize the predicted glimpse-region results taken from ImageNet100 validation dataset in Figure 3 which includes three common cases: 1) Figure 3 (a) shows the examples that the model has correct prediction from the first glimpse while still seeking a precise glimpse-region. This is because the Glimpse Classifier guides every glimpse, making the model more interpretable that explicitly provides more meaningful task-relevant regions; 2) Figure 3 (b) shows the general failure cases. We find that MGNet sometimes fails due to the over-complicated and confusing scene or the tiny object size; 3) Figure 3 (c) shows the progress that MGNet first predicts wrong but later corrects itself. Note that in some cases, a more precise glimpse-region is found later, showing the model integrates the information well during the iteration. In other cases, the model looks at a similar region but changes the prediction. We infer this improvement comes from the inherent ensemble feature.

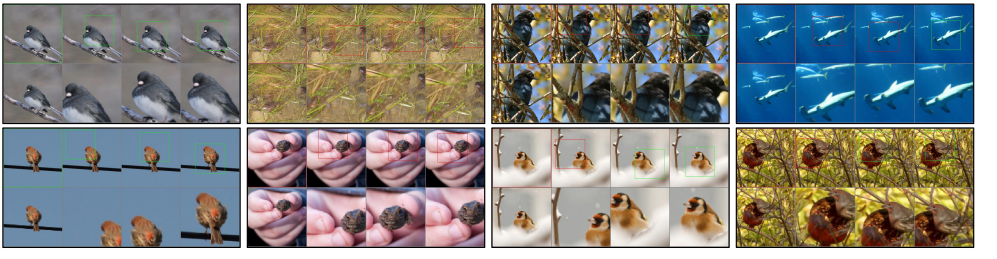

Figure 3: A visualization of the glimpses series generated by MGNet on ImageNet100 validation set. A green box denotes that the model has a correct prediction at this glimpse and the red vice versa.