SSBNet: Improving Visual Recognition Efficiency by Adaptive Sampling Supplementary

Ho Man Kwan and Shenghui Song

The Hong Kong University of Science and Technology
hmkwan@connect.ust.hk eeshsong@ust.hk

In this supplementary, we provide some further results including 1.) performance with different saliency networks, 2.) comparison with another type of adaptive networks, 3.) detailed numerical results for the ablation study in the paper, and 4.) additional figures that visualize the sampling of the SSBNet.

1 Additional Results with Different Saliency Networks

Besides the 1×1 convolutional layer utilized for the saliency map estimation in the main paper, we also conducted additional experiments by increasing the kernel sizes of the convolutional layers from 1×1 to 3×3 and 5×5. The experiments were performed with ResNet-D-50/101/152 [2,3]. However, we did not notice any performance difference. This may be due to the fact that the saliency networks take the deep features as their inputs, which have large effective kernel sizes [6]. As a result, increasing the kernel sizes in the saliency networks did not lead to any performance improvement.

2 Comparison with Other Adaptive Networks

In the main paper, we referred to adaptive sampling as the methods that perform geometric samplings or transformations on the images or feature maps, where the samplings or transformations depend on the inputs. Examples of adaptive sampling methods include spatial transformer [4] and saliency sampler [7]. There are also approaches that adaptively skip the inference on individual units, such as pixels or blocks, to reduce the latency. For example, the stochastic sampling-interpolation network [10] samples pixels for inference, where the SBNet [8] selects blocks.

In table 1, we provided preliminary comparison results between SSBNet and the stochastic sampling-interpolation network [10] (referred to as SSIN for abbreviation), a recently proposed adaptive inference network. Following the same setting as the main paper, we trained SSBNet and SSIN based on ResNet-D-50 [3] on ImageNet [9]. For SSIN, instead of the standard 120 epochs of training, we trained the models with 200 epochs, following the original paper which used a longer training for the adaptive networks. We reported results of SSIN with loss weight λ of $\{0.001, 0.005, 0.010, 0.015\}$.

Table 1: Comparison to stochastic sampling-interpolation networks

	Model	FLOPS	Top-1(%)
120 Epochs	Baseline	4.3G	78.1
	SSB	3.0G	78.1
	SSIN, λ=0.001	3.9G	76.2
	SSIN, λ=0.005	3.4G	75.6
	SSIN, λ=0.010	3.2G	75.8
	SSIN, λ=0.015	3.0G	75.3
200 Epochs	SSIN, λ=0.001	4.1G	77.4
	SSIN, λ=0.005	3.7G	76.6
	SSIN, λ=0.010	3.3G	75.5
	SSIN, λ=0.015	3.0G	73.3

Baseline: ResNet-D-50 [3]
SSB: SSB-ResNet-D-50
SSINI: SSINI [10] + ResNet-D-50 [3]

In our experiments, SSB-ResNet-D-50 achieved better accuracy than SSIN. Notice that in the SSIN paper, the authors trained ResNet34 for their experiments, where we utilized ResNet-D-50, which is a deeper and improved version of ResNet [2]. The performance drop of SSIN reported here indicates that it may not fit the more complicated network and training scheme used in this paper. Compared to the baseline, the SSB-ResNet-D-50 has no drop in accuracy but with 30% less FLOPS.

3 Detailed Results of Ablation Studies

In Tables 2 and 3, more detailed results regarding the computation complexity, i.e., FLOPS, and accuracy of models with different sampling methods and sampling sizes are shown. All results are reported by the average of 3 runs of SSB-ResNet-D, which is based on ResNet-D [3]. As mentioned in the paper, four sampling methods are tested: 1) the proposed adaptive sampling in SSBNet, 2) the uniform sampling with the sampling mechanism in SSBNet (Equation 10 in the paper), 3) the uniform sampling with bilinear interpolation, and 4) the depthwise convolution for downsampling with bilinear interpolation for upsampling [5].

The results show that the models with adaptive sampling outperform the ones with uniform sampling on average, where the differences are larger with RandAugment [1]. With RandAugment, the SSB-ResNet-D-152 with adaptive sampling and sampling sizes of (16, 8, 4) outperformed all other networks that have similar complexity by at least 0.4% in accuracy; the SSB-ResNet-D-152 with adaptive sampling and sampling sizes of (12, 6, 3) outperformed the network with depthwise convolution and bilinear sampling in accuracy, but with 11% less computation.

4 More results on Visualization

To further illustrate the capability of SSBNet to focus on different locations of the feature maps at different layers, we provide additional visualization results in Figures 1-10. All figures are sampled from the outputs of SSB-ResNet-RS-152 with input size of 224×224. The input images are from ImageNet dataset [9]. Each figure provides the samples from one of the selected layers, and each row in the figures shows the results with different inputs. The figures are annotated with their index in the network, e.g. Layer 3-5 indicates the fifth layer in the third group of the building layers.

It can be observed that SSBNet is able to sample different positions at different layers. For example, in Figure 1, the network samples the feature maps uniformly; in Figure 2, it samples more heavier toward the objects; in Figure 5, the network zooms out from the feature maps, which increases the receptive field of the convolutional layers with respect to the input feature maps. We also noticed that the network weights different objects. In Figure 2b, it focuses on both the people and the dogs in the third feature maps, while in Figure 7b, it weights the dogs much heavier than the people.

Table 2: Comparison between uniform and adaptive sampling on ImageNet, with different sampling size

Model	Input size	Params	FLOPS	Top-1(%)
U-50	224 × 224	25.6M	2.61G	77.70
A-50	224 × 224	25.6M	2.61G	77.75
U-101	224 × 224	44.6M	3.35G	78.29
A-101	224 × 224	44.6M	3.35G	78.46
U-152	224 × 224	60.2M	4.13G	78.56
A-152	224 × 224	60.3M	4.14G	78.64

U: Uniform (12, 6, 3)
A: Adaptive (12, 6, 3)

Model	Input size	Params	FLOPS	Top-1(%)
U-50	224 × 224	25.6M	2.95G	77.96
A-50	224 × 224	25.6M	2.95G	78.09
U-101	224 × 224	44.6M	4.25G	78.68
A-101	224 × 224	44.6M	4.25G	78.88
U-152	224 × 224	60.2M	5.60G	79.05
A-152	224 × 224	60.3M	5.61G	79.19

U: Uniform (16, 8, 4)
A: Adaptive (16, 8, 4)

Model	Input size	Params	FLOPS	Top-1(%)
U-50	224 × 224	25.6M	3.38G	78.01
A-50	224 × 224	25.6M	3.38G	78.12
U-101	224 × 224	44.6M	5.40G	79.01
A-101	224 × 224	44.6M	5.40G	78.99
U-152	224 × 224	60.2M	7.49G	79.42
A-152	224 × 224	60.3M	7.49G	79.45

U: Uniform (20, 10, 5)
A: Adaptive (20, 10, 5)

Model	Input size	Params	FLOPS	Top-1(%)
B-50	224 × 224	25.6M	2.95G	77.62
D-50	224 × 224	25.9M	2.74G	77.95
B-101	224 × 224	44.6M	4.25G	78.41
D-101	224 × 224	45.3M	3.69G	78.42
B-152	224 × 224	60.2M	5.61G	78.86
D-152	224 × 224	61.4M	4.65G	78.81

B: Blinear (16, 8, 4)
D: DConv + Bilinear (14, 7, 4)

Table 3: Comparison between uniform and adaptive sampling on ImageNet, with RandAugment [1]

Model	Input size	Params	FLOPS	Top-1(%)
U-50	224 × 224	25.6M	2.61G	77.97
A-50	224 × 224	25.6M	2.61G	77.96
U-101	224 × 224	44.6M	3.35G	78.91
A-101	224 × 224	44.6M	3.35G	79.13
U-152	224 × 224	60.2M	4.13G	79.31
A-152	224 × 224	60.3M	4.14G	79.65

U: Uniform (12, 6, 3)
A: Adaptive (12, 6, 3)

Model	Input size	Params	FLOPS	Top-1(%)
U-50	224 × 224	25.6M	2.95G	78.07
A-50	224 × 224	25.6M	2.95G	78.22
U-101	224 × 224	44.6M	4.25G	79.27
A-101	224 × 224	44.6M	4.25G	79.50
U-152	224 × 224	60.2M	5.60G	79.71
A-152	224 × 224	60.3M	5.61G	80.12

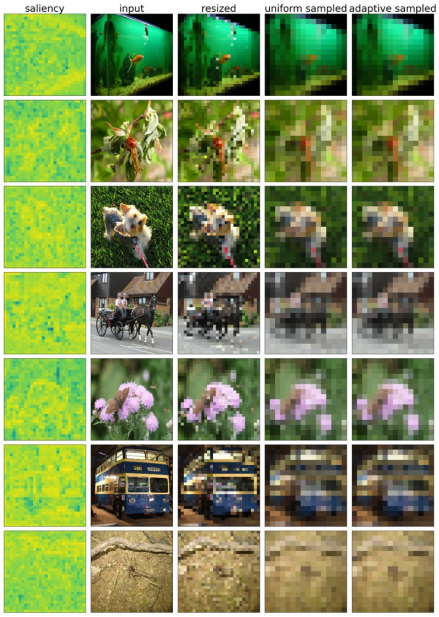
U: Uniform (16, 8, 5)
A: Adaptive (16, 8, 4)

Model	Input size	Params	FLOPS	Top-1(%)
U-50	224 × 224	25.6M	3.38G	78.37
A-50	224 × 224	25.6M	3.38G	78.36
U-101	224 × 224	44.6M	5.40G	79.41
A-101	224 × 224	44.6M	5.40G	79.88
U-152	224 × 224	60.2M	7.49G	80.19
A-152	224 × 224	60.3M	7.49G	80.28

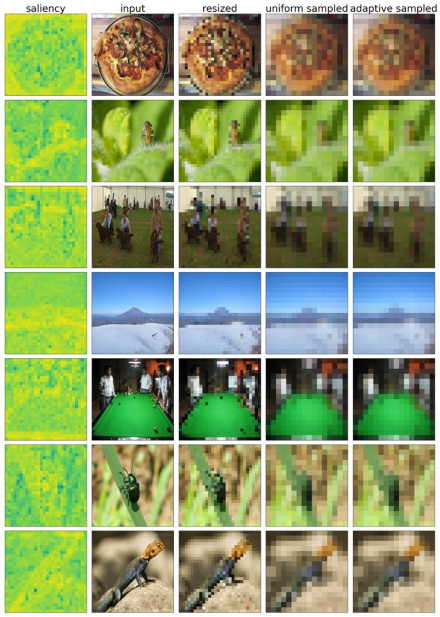
U: Uniform (20, 10, 5)
A: Adaptive (20, 10, 5)

Model	Input size	Params	FLOPS	Top-1(%)
B-50	224 × 224	25.6M	2.95G	77.77
D-50	224 × 224	25.9M	2.74G	78.06
B-101	224 × 224	44.6M	4.25G	79.02
D-101	224 × 224	45.3M	3.69G	79.15
B-152	224 × 224	60.2M	5.61G	79.58
D-152	224 × 224	61.4M	4.65G	79.58

B: Blinear (16, 8, 4)
D: DConv + Bilinear (14, 7, 4)



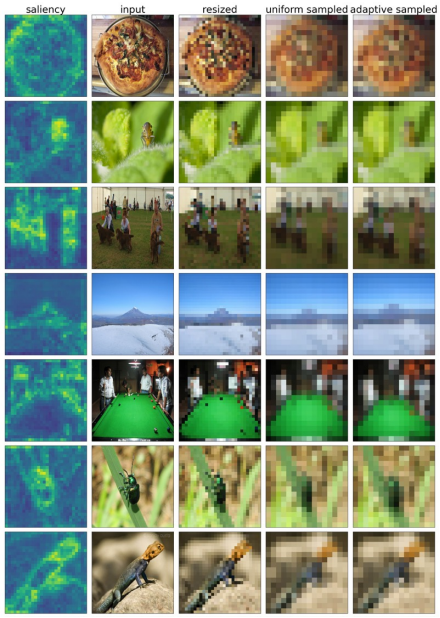
(a) Example 1



(b) Example 2

Fig. 1: Layer 2-2

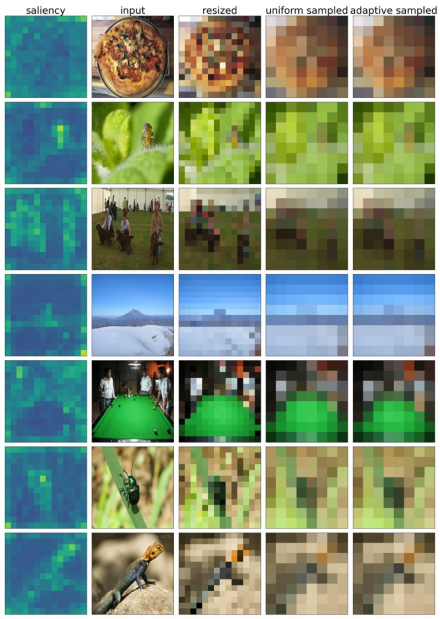
(a) Example 1



(b) Example 2

Fig. 2: Layer 2-7

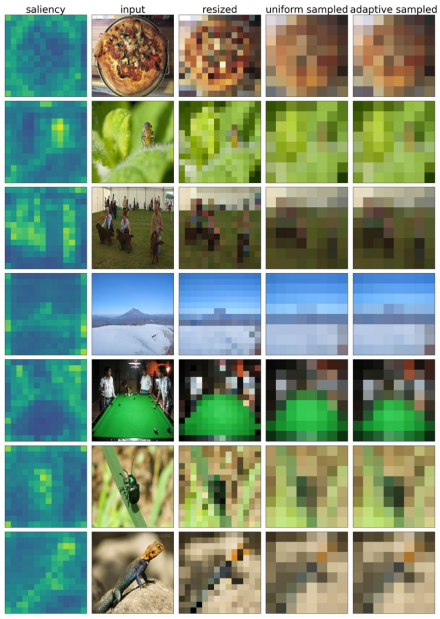
(a) Example 1



(b) Example 2

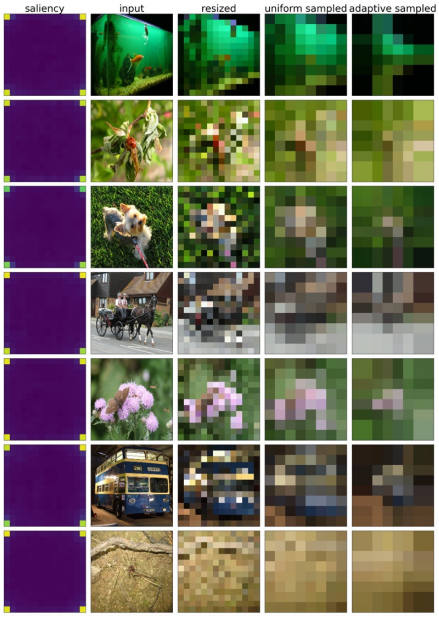
Fig. 3: Layer 3-5

(a) Example 1

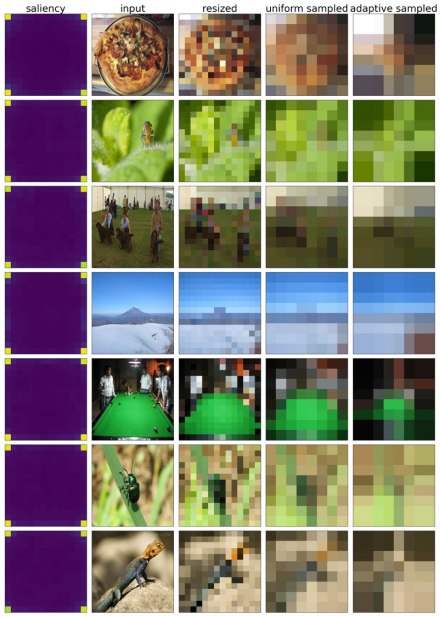


(b) Example 2

Fig. 4: Layer 3-10



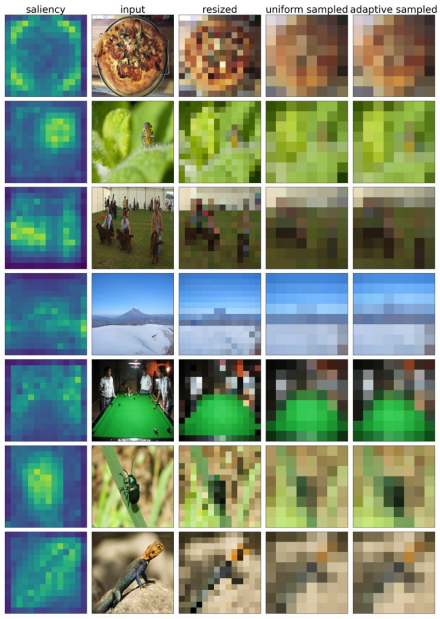
(a) Example 1



(b) Example 2

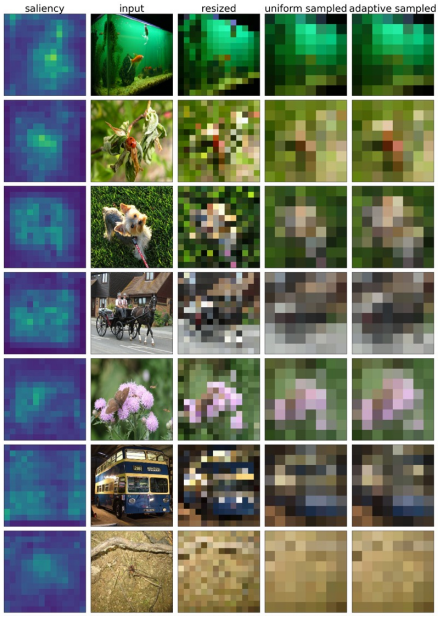
Fig. 5: Layer 3-15

(a) Example 1

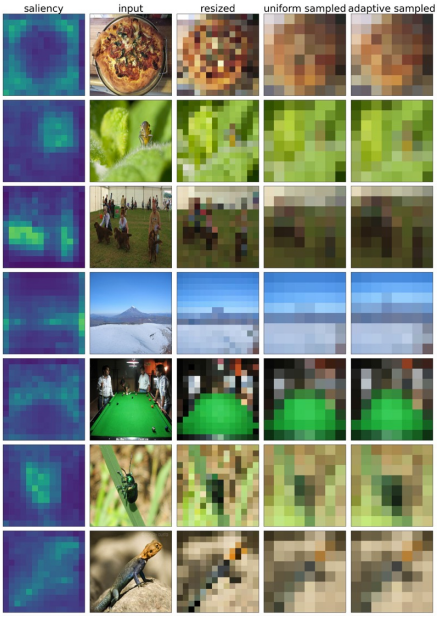


(b) Example 2

Fig. 6: Layer 3-20



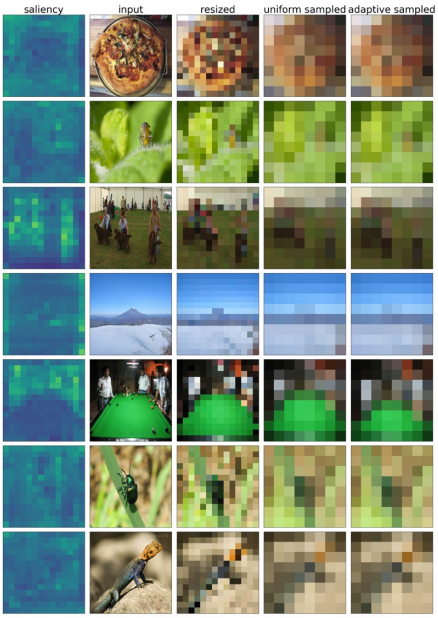
(a) Example 1



(b) Example 2

Fig. 7: Layer 3-25

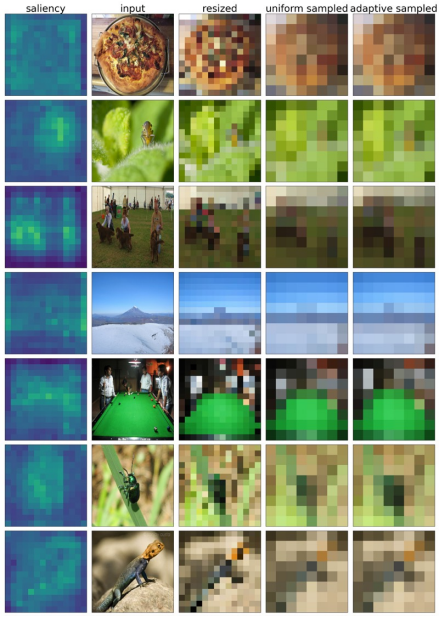
(a) Example 1



(b) Example 2

Fig. 8: Layer 3-30

(a) Example 1



(b) Example 2

Fig. 9: Layer 3-35

(a) Example 1

(b) Example 2

Fig. 10: Layer 4-3

References

1. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: Randaugment: Practical automated data augmentation with a reduced search space. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 18613–18624. Curran Associates, Inc. (2020) 2, 4
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 1, 2
3. He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 1, 2
4. Jaderberg, M., Simonyan, K., Zisserman, A., kavukcuoglu, k.: Spatial transformer networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015) 1
5. Li, D., Zhou, A., Yao, A.: Hbonet: Harmonious bottleneck on two orthogonal dimensions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019) 2
6. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016) 1
7. Recasens, A., Kellnhofer, P., Stent, S., Matusik, W., Torralba, A.: Learning to zoom: a saliency-based sampling layer for neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 51–66 (2018) 1
8. Ren, M., Pokrovsky, A., Yang, B., Urtasun, R.: Sbnet: Sparse blocks network for fast inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 1
9. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y 1, 3
10. Xie, Z., Zhang, Z., Zhu, X., Huang, G., Lin, S.: Spatially adaptive inference with stochastic feature sampling and interpolation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 531–548 (2020) 1, 2