Supplementary Materials for "Deep Reinforced Attention Learning for Quality-Aware Visual Recognition"

Duo Li and Qifeng Chen

The Hong Kong University of Science and Technology
duo.li@connect.ust.hk cqf@ust.hk

A More Related Work

Non-local Attention Networks. Keeping track of non-local architectures [13] as a self-attention mechanism, A^2-Net [4] gathers global features with a second-order attention pooling and distributes the information to each local position in a two-step configuration. AA-Net [3] and Fully Attention model [11] further develop a two-dimensional relative self-attention mechanism to replace convolutions entirely. DANet [5] and CCNet [8] consider spatial and channel non-local modules simultaneously to strengthen contextual modeling for semantic segmentation. EMANet [9] is inspired by the EM algorithm and computes the attention map in an iterative fashion. Without exception, our proposed learning method could also be applied to these specialized attention modules for quality-aware image classification by inspecting their feature state and the corresponding non-local attention action in the same manner.

Reinforcement Learning for Network Engineering. Deep reinforcement learning has been adopted in the area of neural network slimming. For example, BlockDrop [15] utilizes a policy network to learn the optimal block dropping strategy for each image sample, simultaneously selecting minimal layer configurations for the inference route and preserving the desired prediction accuracy. SkipNet [14] uses a gating network to selectively skip redundant layers conditioned on the preceding activation and proposes a hybrid learning regime to optimize the discrete skipping decisions. RNP [10] leverages the Q-learning algorithm to assess the importance of feature maps and dynamically prunes the network based on the input images and the current feature maps to retain the recognition ability. N2N [1] applies a reinforcement learning model to learn the policy of channel selection, condensing a large teacher network into a small student one by removing redundant layers and shrinking the size of remaining layers. AMC [6] employs reinforcement learning to efficiently sample from the network architecture space, leading to highly compressed models while preserving their accuracy. In this regime, the automatically learned deep compression policy could outperform the conventional rule-based pipelines. In addition to these post-processing approaches, reinforcement learning has been applied to automate the design process of neural architectures, referred to as Neural Architecture Search (NAS). NASNet [16, 17] and MetaQNN [2] lead this trend of

utilizing reinforcement learning skills to search for a well-performing network architecture. Unlike these methods which concentrate on network compression and architecture search, we propose to measure and boost the quality of attention generation under the reinforcement learning framework. To the best of our knowledge, little progress with reinforcement learning has been made in the fundamental problem of handcrafted attention networks, which is of vital importance in the neural architecture design.

B Visualization Results

B.1 Channel Attention

As illustrated in Fig. 1, we compare the distributions of channel attention vectors in all building blocks of SE-ResNet-50 before and after applying our method. The most significant difference lies in the last stage (`conv5_x`), where the attention weights tend to be more diverse across different channels in our reinforced attention networks. Since each object category always exhibits preference to discriminative visual features in certain channels, our method facilitates better adaption and specialization of high-level features by improving the recalibration quality of attention modules. The enhanced representation learning ultimately boosts the visual recognition performance of the original attention networks, as validated by the quantitative results in the main paper.

B.2 Spatial Attention

Although our method is not directly designed to explore attended regions in the image space, the intermediate feature maps in the backbone network could be back-projected to the image space using Grad-CAM [12], merely in order to provide an intuition of the improved spatial attention with our method. Grad-CAM computes the importance of each location in the image space using gradients with respect to a specific class. We provide representative visualization results by applying Grad-CAM to the last convolutional layer of CBAM-ResNet-50 and our proposed reinforced attention version on the ImageNet validation set, as shown in the heat maps in Fig. 2 and 3. By observing and comparing salient regions where the network allocates more resources for correct prediction, it is clear that those regions covering the main objects enjoy higher quality of attention in our reinforced attention networks.

Specifically, we make the observations that the attended regions in our reinforced attention networks are guided to be more correct (such as the top four comparisons in Fig. 2) and more complete (such as the bottom four comparisons in Fig. 2 and the top four comparisons in Fig. 3), covering the objects of interest. The attended regions could be precise even with objects heavily occluded (see "hippopotamus, hippo, river horse, Hippopotamus amphibius" in Fig. 3) or partially observed (see "barometer" in Fig. 3). When the critical feature regions for recognition are disconnected (see "acoustic guitar" in Fig. 3) or occupy a small

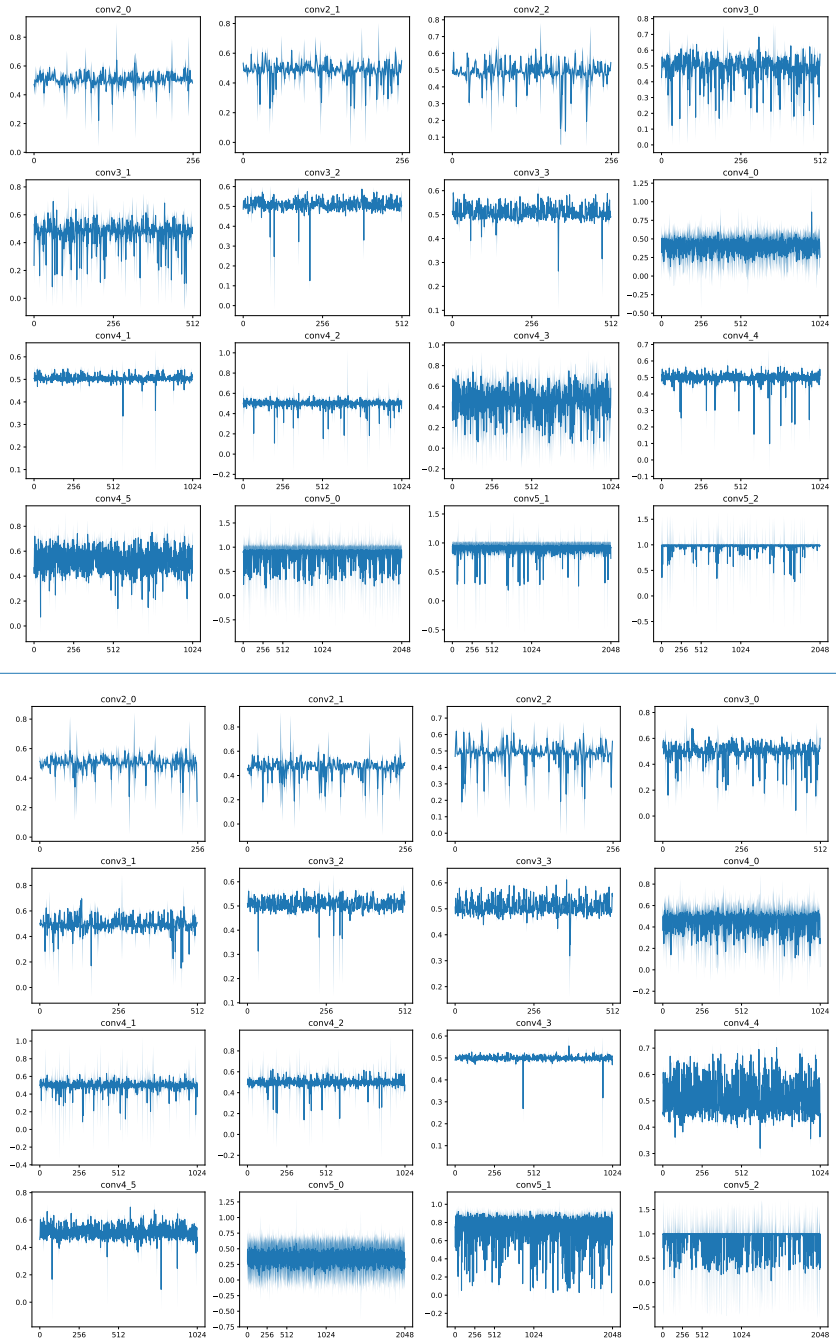


Fig. 1: Distributions of channel-attention vectors on the ImageNet validation set with SE-ResNet-50 before (*top*) and after (*bottom*) applying DREAL. The x-axis represents channel index and the y-axis represents magnitude.

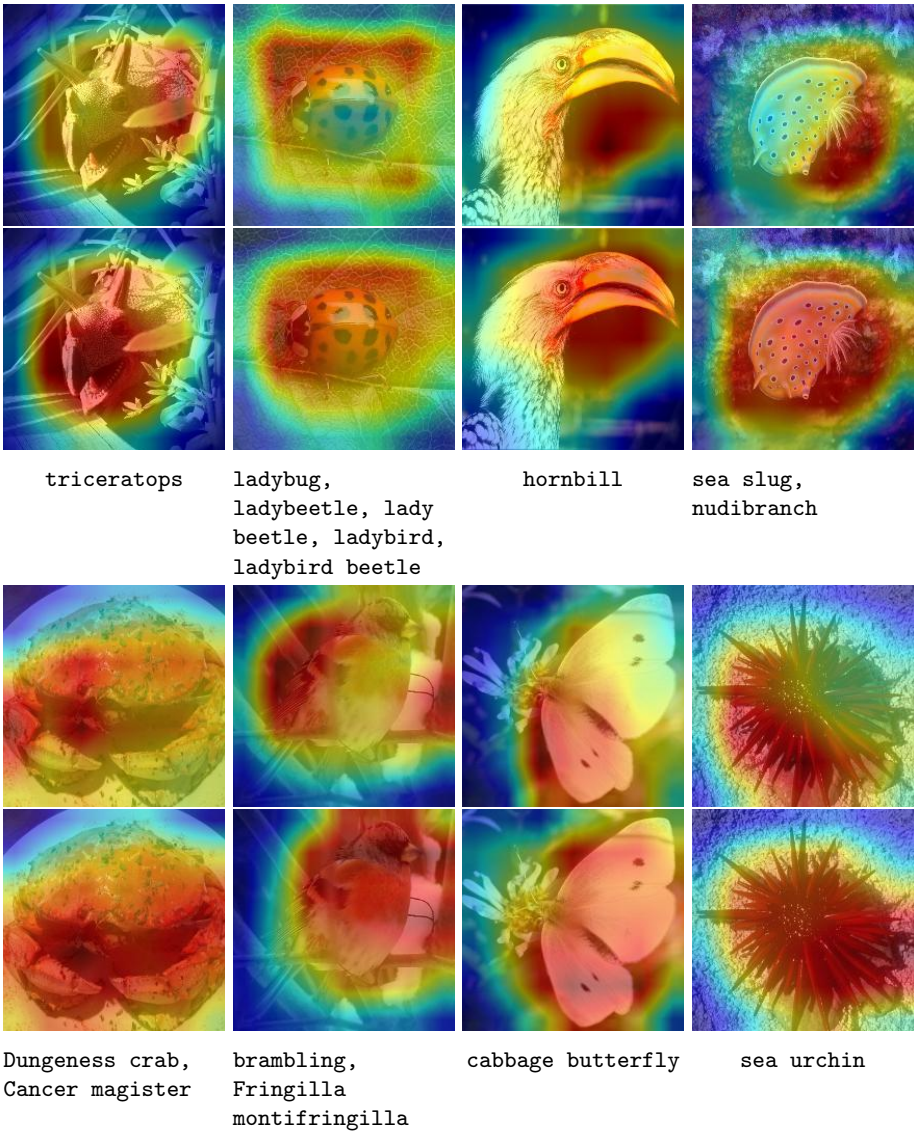
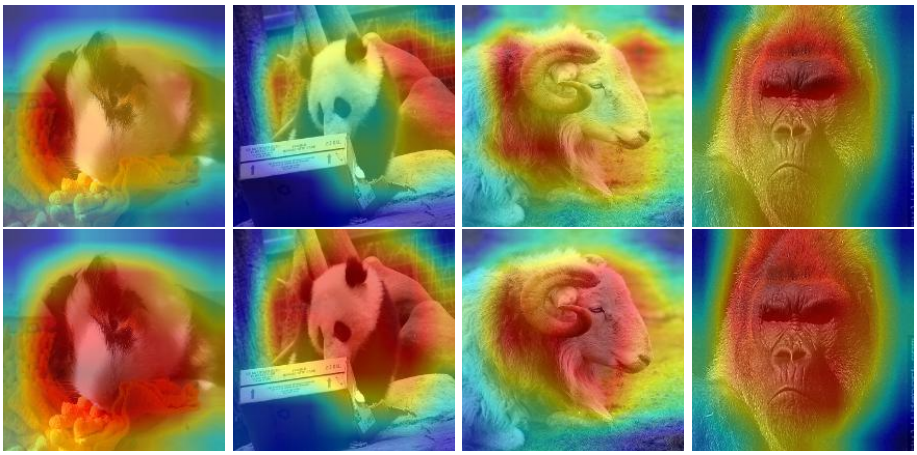
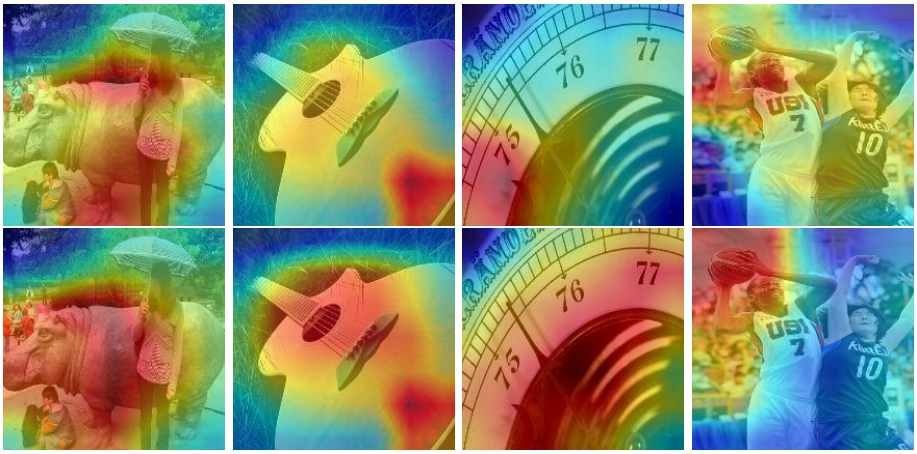


Fig. 2: Grad-CAM visualization results for spatial attention. Evaluated images are selected from eight categories with the ground truth labels annotated at the bottom of each pair. The top one in each pair corresponds to results from the baseline CBAM-ResNet50, while the bottom one represents the result improved using our DREAL method.



guinea pig, Cavia cobaya | giant panda, panda, panda bear, coon bear, Ailuropoda melanoleuca | ram, tup | gorilla, Gorilla gorilla



hippopotamus, hippo, river horse, Hippopotamus amphibius | acoustic guitar | barometer | basketball

Fig. 3: Grad-CAM visualization results for spatial attention. Evaluated images are selected from eight categories with the ground truth labels annotated at the bottom of each pair. The top one in each pair corresponds to results from the baseline CBAM-ResNet-50, while the bottom one represents results improved using our DREAL method.

portion in the whole image (see "basketball" in Fig. 3), our proposed reinforced attention network can still localize them successfully, taking advantage of the quality-aware guidance from the extra critic network for image classification.

Furthermore, our method could bring about substantial performance improvement on the person re-identification benchmarks. Analogously, the success of reinforced spatial attention model in this specific task may also arise from removing its attention from misleading regions of the pedestrian image to effectively alleviate the negative effects of occlusion and cluttered backgrounds. For DukeMTMC-reID, a more complicated scenario usually with multiple objects in the scene, we choose two images of the same person from the gallery and query sets respectively and observe that our method helps the attention map to always focus on the same informative parts of this person, such as the bag, pants and so on, while the baseline attention map fails to locate these salient regions more often.

B.3 Critique and Reward

We track the predicted critic value and the reward for each attention module in the SE-ResNet-50 during the entire training period, as demonstrated in Fig. 4. The variation of expected critiques closely follows their corresponding actual rewards, speaking for the effectiveness of the regression loss $\mathcal{L}_r$ in updating the critic network $\boldsymbol{\phi}$ to make precise predictions. The values of critique and reward rise up gradually as the optimization goes (they may drop before the first learning rate decay arrives since the model predictions can be noisy and not reliable enough even with the aid of attention modules at this very early period), speaking for the effectiveness of the quality loss $\mathcal{L}_q$ in updating the attention module $\boldsymbol{\theta}$ to yield high-quality attention maps in favor of the final recognition performance. The visual analysis of critique and reward further discloses the learning dynamics of our proposed DREAL method and justifies the principle of our design.

As an additional note, the first line of Fig. 4 shows a trend that the critic network doesn't converge quite well in the shallower layers when compared to the deeper ones where the critic and reward values are much closer. This observation somewhat reveals that our DREAL method shows more effectiveness in the deeper layers. We conduct ablation studies by separately applying DREAL to each stage of the top-performing SRM-ResNet-101 and summarize the top-1 error on ImageNet in Table 1. Though there is a performance improvement regarding each stage compared to the baseline, applying DREAL to all stages leads to the best result as what we show in the main paper. But in some cases where computational budget becomes the key consideration, we may remove our method from some shallow layers as a trade-off. Furthermore, we note that the attention modules themselves are more important in deeper layers, such as `conv4_x` and `conv5_x`, as claimed in the SENet paper [7] and our visual analysis.

Table 1: Top-1 error on the ImageNet validation set when applying DREAL to different stages of SRM-ResNet-101.

Stage	`conv2_x`	`conv3_x`	`conv4_x`	`conv5_x`	None (**baseline**)	Full (**ours default**)
Top-1 Err.(%)	21.132	20.946	20.858	20.794	21.404	20.474

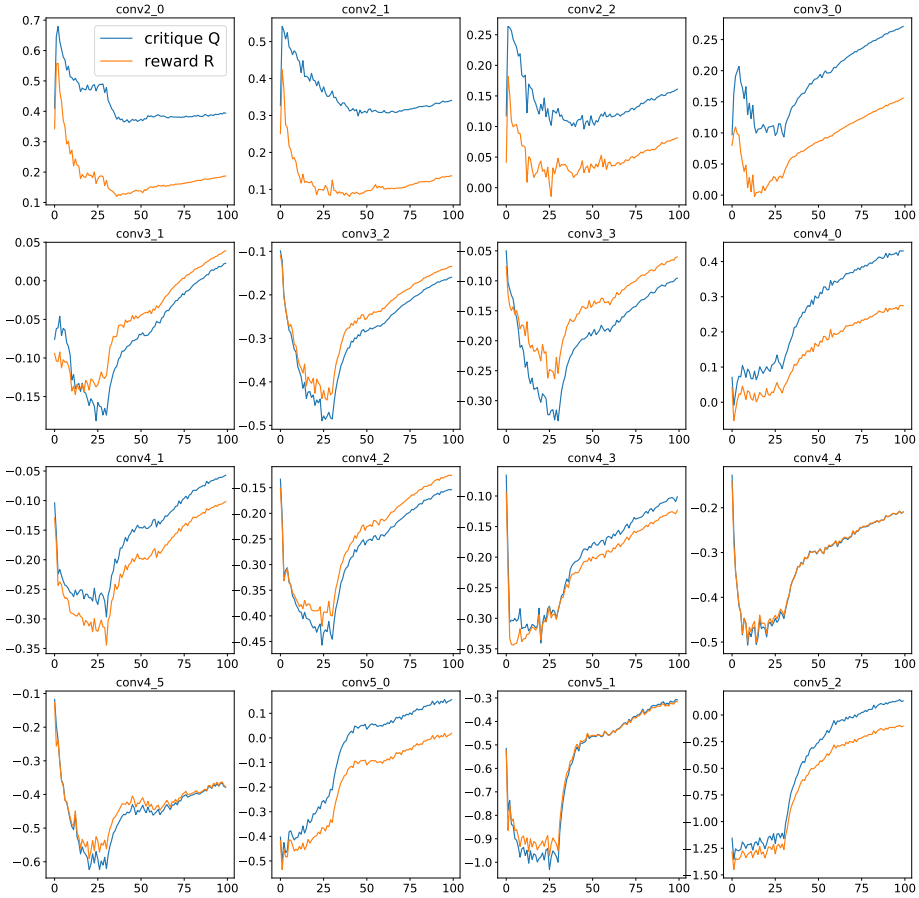


Fig. 4: The critique Q and reward R in each attention module of the SE-ResNet-50 during all the 100 training epochs. The x-axis represents the training epoch and the y-axis represents the value.

References

1. Ashok, A., Rhinehart, N., Beainy, F., Kitani, K.M.: N2N learning: Network to network compression via policy gradient reinforcement learning. In: ICLR (2018)
2. Baker, B., Gupta, O., Naik, N., Raskar, R.: Designing neural network architectures using reinforcement learning. In: ICLR (2017)
3. Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V.: Attention augmented convolutional networks. In: ICCV (2019)
4. Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: Aˆ2-nets: Double attention networks. In: NeurIPS (2018)
5. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR (2019)
6. He, Y., Lin, J., Liu, Z., Wang, H., Li, L.J., Han, S.: AMC: Automl for model compression and acceleration on mobile devices. In: ECCV (2018)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
8. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNet: Criss-cross attention for semantic segmentation. In: ICCV (2019)
9. Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H.: Expectation-maximization attention networks for semantic segmentation. In: ICCV (2019)
10. Lin, J., Rao, Y., Lu, J., Zhou, J.: Runtime neural pruning. In: NIPS (2017)
11. Parmar, N., Ramachandran, P., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. In: NeurIPS (2019)
12. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
13. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
14. Wang, X., Yu, F., Dou, Z.Y., Darrell, T., Gonzalez, J.E.: SkipNet: Learning dynamic routing in convolutional networks. In: ECCV (2018)
15. Wu, Z., Nagarajan, T., Kumar, A., Rennie, S., Davis, L.S., Grauman, K., Feris, R.: BlockDrop: Dynamic inference paths in residual networks. In: CVPR (2018)
16. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. In: ICLR (2016)
17. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: CVPR (2018)