RECURRENT LAYER ATTENTION NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Capturing long-range feature relations has been a central issue on convolutional neural networks (CNNs). To tackle this, attempts to integrate end-to-end trainable attention module on CNNs are widespread. Main goal of these works is to adjust feature maps considering spatial-channel correlation inside a convolution layer. In this paper, we focus on modeling relationships among layers and propose a novel structure, 'Recurrent Layer Attention network,' which stores the hierarchy of features into recurrent neural networks (RNNs) that concurrently propagating with CNN and adaptively scales feature volumes of all layers. We further introduce several structural derivatives for demonstrating the compatibility on recent attention modules and the expandability of proposed network. For semantic understanding on learned features, we also visualize intermediate layers and plot the curve of layer scaling coefficients(*i.e.*, layer attention). Recurrent Layer Attention network achieves significant performance enhancement requiring a slight increase on parameters in an image classification task with CIFAR and ImageNet-1K 2012 dataset and an object detection task with Microsoft COCO 2014 dataset.

1 INTRODUCTION

Concatenating all features in the order of layers in convolutional neural network (CNN) provides new interpretation, features form a sequence, consisting of features with small receptive fields to large receptive fields. Interestingly, recurrent neural network (RNN) is one of the representatives for modeling the sequential information. (Sutskever et al., 2014; Hochreiter & Schmidhuber, 1997). On the other hands, Recent attempts to utilize attention mechanism for empowering CNN a better representational power are prevalent (Hu et al., 2018b; Wang et al., 2017).

Motivated by intrinsic characteristics of CNN and RNN, and recent attention works on computer vision, we present the Recurrent Layer Attention Network (RLA network), which is differentiable and light-weight while improving the representational power of CNN by a slightly different way from other attention works in computer vision. The main goal of our work is applying global weights balance *among* layers by inheriting the feature hierarchy from previous CNN layers. We accomplish the goal by two main structural designs: employing our inter-layer attention mechanism to make the network re-adjust features, and utilizing RNN to memorize the feature hierarchy with concurrently propagating parallel with CNN.

We hypothesize that RLA network gains additional class discriminability through inheriting the informative feature hierarchy, such as repetitive appearances of important features or relevant features with different receptive fields. For example, our network raises the activation of neurons that is responsible to the whole body of zebra, using a history of features of relatively smaller receptive field. We demonstrate our hypothesis through the Grad-CAM visualization of intermediate features and corresponding layer attention value (*i.e.*, the importance of layer.)

We evaluate RLA network on image classification and object detection tasks using benchmark datasets: CIFAR, ImageNet-1K, and Microsoft COCO. On both tasks, RLA network gets comparable results with the state-of-the-arts, Squeeze-and-Excitation network (Hu et al., 2018b), and superior results than original ResNet architecture. Moreover, we suggest the compatibility of the RLA network by introducing the expansion of RLA network combining our inter-layer attention mechanism toward recent attention works (Hu et al., 2018b) on ablation study. Incorporating RLA network to recent attention works (Call these networks as *intra-layer* attention networks), further variations of RLA network can be recognized as the generalized attention network considering correlations among both inside and outside a layer.

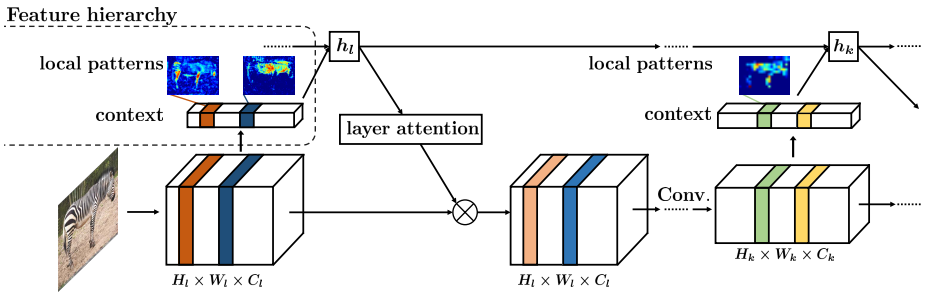

Figure 1: Concept diagram of Recurrent Layer Attention network (RLA network). Through concurrently propagating with CNN, RNN stores summarized feature and infers a scaling coefficient of feature volume considering a feature hierarchy. Recurrently applying this operation for every layer of CNN, RLA network applies overall weight balance of layers.

We summarize our contributions as follows:

- We propose two new concepts: the weight balancing of CNN features along layers (call it inter-layer attention), and the connection from shallow CNN layers to deep CNN layers through concurrently propagating RNN.
- We demonstrate the effectiveness of proposed RLA network for an image classification task and an object detection task on benchmark datasets. RLA network achieves similar or superior results compared with leading deep network while requiring small model parameters increase.
- Ablation studies investigate the effectiveness of two proposed concepts and show the compatibility towards existing intra-attention models, and the further expandability by tuning architectural designs. We show and discuss how RLA network is learned to interpret images by visualizing intermediate layers and plotting layer attention values.
- RLA network is easy to implement requiring basic operations of modern deep learning frameworks.

2 Related Work

Attention mechanism in computer vision. Attention mechanism can be interpreted as a methodology to bias the allocation of available neurons to the most informative components of input signal (Hu et al., 2018b). Recent main application of attention mechanism on computer vision is integrating end-to-end trainable attention module for deep CNNs. These attention modules are divided into two categories; spatial attention module, and channel-wise attention module. *Spatial attention* module learns spatial masks on feature maps for regulating the activations of neurons (Wang et al., 2017; Zhang et al., 2019a). *Channel-wise attention* module earns channel-wise distribution and then utilizes it to refine feature maps (Hu et al., 2018b;a). Several architectural designs both benefitting spatial and channel-wise attention are also stressed out (Woo et al., 2018; Zhang et al., 2019b; Chen et al., 2017; Hu et al., 2018a; Chen et al., 2018b).

However, all these spatial and channel-wise attention modules, as we call *intra-layer* attention, only handles interaction *inside* a CNN layer. Therefore, currently suggested modules structurally lack to model relation among visual features captured in different CNN layers. Therefore, we suggest *inter-layer* attention mechanism. Through global weight balancing along CNN layers, our module models complex relationships among visual features of different receptive fields.

RNN in computer vision. Numbers of computer vision papers utilize RNN for various tasks containing visual question answering (Ren et al., 2015a; Malinowski et al., 2015), image captioning (Xu et al., 2015), multi-object classification (Chen et al., 2018a; Wang et al., 2016), and video description (Donahue et al., 2015). They share common ground on the way that exploiting RNN; extracting visual features from images using pre-trained CNNs, then employ RNN to model sequential procedures of each task. For example, obtained CNN features per subsequent time frame series are becoming inputs to belonging RNN units, and RNN conduct a prediction over video recognition and description tasks (Donahue et al., 2015). However, there are only few works to exploit RNN for directly enhancing

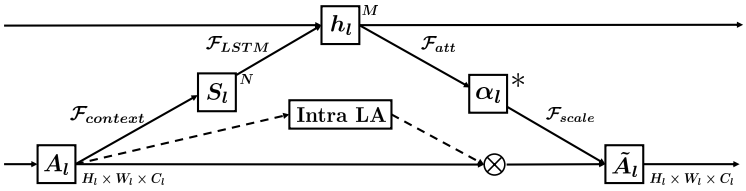

Figure 2: The structure of Recurrent Layer Attention network ($*$ = 1), RLA-vector ($* = C_l$), and IA+RLA (including a dotted line). $\otimes$ denotes scale operations belonging to intra-layer attention.

the performance of deep neural networks on vision tasks that do not require sequential decisions such as image classification or single object detection. One trial (Mnih et al., 2014) propose the deep network that employs RNN and region attending mechanism, but is non-differentiable.

In contrast to earlier works, we manipulate CNN and RNN to simultaneously affect to each other at every layer of CNN so that enhancing the representational power of CNN. To the best of our knowledge, our concept of conveying information from shallow CNN layer to deep CNN layer through the use of concurrently propagating memory units is the first attempt.

3 Recurrent Layer Attention Network

We start by defining several expressions to reduce unnecessary ambiguity. In our works, *local patterns* denote class-specific characteristics in images such as a head of a cat, which are collected as a response of the convolution operation. $\mathcal{G}$ denotes a convolution operation alone, or the arbitrary combinations of convolution operations and other mapping functions, such as residual block (He et al., 2016) or inception Module (Szegedy et al., 2016). *Features* are the results of convolution operation $\mathcal{G}$. Precisely, $\mathcal{A} = [A_1, A_2, \cdots, A_L]$ and $A_l = [A_l^1, A_l^2, \cdots, A_l^C]$ denote the set of feature volumes and maps in CNN, where $A_l^c \in \mathbb{R}^{H \times W}$ is the response of c-th convolution filter of the l-th CNN layer.

RLA network consists of three operations: summarizing features, storing summarized features in LSTM, and inferring the layer attention. Right after the convolution operation $\mathcal{G}$, RLA module "summarizes" the feature volume into a statistic, which is a bag of the local patterns in images. Then it "stores" the context in LSTM hidden units, and finally "infers" the layer attention of how much to concentrate on feature volume from $\mathcal{G}$. The concept figure describing this procedure is in Figure 1.

The intuition of utilzing RNN is derived from recognizing concatenated feature volumes along CNN layers as a sequence. Because each element of the feature sequence is a prior information implying which local patterns emerged, we hypothesize that LSTM hidden units which is carrying the hierarchy of features can provide meaningful information to CNN such as the importance of latter layer. Therefore, we utilize the attention mechanism (Xu et al., 2015; Bahdanau et al., 2014) to empower CNN reducing the activation of overlapped local patterns, or emphasizing the activation of important local patterns which highly contribute to the class discriminability.

Following subsections are organized as follows. The formulation of RLA network is described in Subsection 3.1, several architectural derivatives for showing the expandability and the compatibility toward intra-attention modules are introduced in Subsection 3.2, and the light-weightness of the RLA network is described in Subsection 3.3.

3.1 Formulation

We present our RLA network by describing one forward-passing procedure of l-th layer. The feature volume A_l is yielded by passing $\mathcal{G}^l$:

$$A_l = \mathcal{G}^l(A_{l-1}). \tag{1}$$

Summarizing Feature. Feature volume A_l is summarized as the context. Here, the context $S_l \in \mathbb{R}^N$ signifies a simplified representation of local patterns in the image, and is computed at every layer of CNN through:

$$S_l = \mathcal{F}_{context}(A_l), \tag{2}$$

where $\mathcal{F}_{context}$ is a combination of feature volume summarizing operation and consequent down-sampling (or up-sampling) operation. We investigate various types of feature summarizing operations for finding proper statistics that embeds the feature volume on Section 3.2. In order to feed S_l into LSTM cell, down-sampling and up-sampling techniques must be adaptively applied per layers, because RNN requires a fixed size of input to each cell. Note here, all the same feature summarizing operation is used on every layer.

Storing Context. The context captured in each earlier layer of CNN is recurrently inserted into LSTM hidden units $h_l \in \mathbb{R}^M$, and selectively embedded in h_l through the updating functions of LSTM with h_{l-1}:

$$h_l = \mathcal{F}_{LSTM}(h_{l-1}, S_l). \tag{3}$$

Here, $\mathcal{F}_{LSTM}$ is a LSTM cell updating operation described in (Hochreiter & Schmidhuber, 1997). The LSTM hidden states and cell states of the first stage, h_1 is initialized with zero.

Inferring Layer Attention. Considering the sequence of contexts previously fed into LSTM through the cascades of convolution operations and memory mechanism of LSTM until l-th layer, h_l infers the layer attention, $\alpha_l \in [0, 1]$, the scalar value for scaling the feature volume A_l:

$$\alpha_l = \mathcal{F}_{att}(h_l) = \sigma(W_2\delta(W_1 h_l)), \tag{4}$$

where $\mathcal{F}_{att}$ is composed of two fully-connected layers with $W_1 \in \mathbb{R}^{\frac{M}{r} \times M}$ and $W_2 \in \mathbb{R}^{1 \times \frac{M}{r}}$. δ and σ refer to the ReLU (Nair & Hinton, 2010) and sigmoid operations, respectively. Reduction ratio r denotes compressing ratio of the first fully-connected layer. Finally, the scaled feature volume $\tilde{A}_l$ is computed by element-wise multiplication $\mathcal{F}_{scale}$ between α_l and all elements of A_l:

$$\tilde{A}_l = \mathcal{F}_{scale}(\alpha_l, A_l) = \alpha_l \cdot A_l, \tag{5}$$

3.2 Instantiation

We introduce several derivatives of RLA network based on two key design variables, the attention structure and the context. All instances are experimented and discussed on Subsection 4.1.

Adoption of intra-layer attention mechanism. We show the compatibility of RLA network toward intra-layer attention mechanism. Adopting the intra-layer attention module, the scale operation for c-th channel of feature volume on l-th layer is given by:

$$\tilde{A}_l^c = \mathcal{F}_{scale}(\alpha_l, \mathcal{F}'_{scale}(A_l^c)) = \alpha_l \cdot \mathcal{F}'_{scale}(A_l^c), \tag{6}$$

where A_l^c denotes a c-th feature map of $A_l = [\tilde{A}_l^1, \tilde{A}_l^2, \cdots, \tilde{A}_l^C]$, and F'_{scale} denotes an arbitrary intra-layer attention mechanism which calibrate a feature volume inside the layer. Figure 2 depicts the integration of RLA with intra-layer attention module. We call this structure as Intra-layer Attention + RLA network (IA+RLA). To evaluate the performance of IA+RLA, we select Squeeze-and-Excitation block (Hu et al., 2018b) as IA module.

Distilling advantages from feature hierarchy. RLA network inherits two advantages: utilizing feature hierarchy to affect on subsequent features using RNN, and adaptively scaling features layer by layer. Because both two factors are latently contributing on guiding CNN to have better representational power, it's hard to discriminate effects of them. To evaluate different effectiveness of two factors, we additionally design a structure named RLA-vector. Between previously mentioned two concepts, RLA-vector only adopt the concept of exploiting feature hierarchy with concurrently propagating RNN, while not applying inter-layer attention mechanism. Utilizing the intra-layer attention mechanism in Squeeze-and-Excitation module (SE module) (Hu et al., 2018b), RLA-vector adjust feature volumes channel-wisely. In the other words, LSTM hidden units of RLA-vector predict a channel-sized vector, not a layer attention scalar. Through the comparison between Squeeze-and-Excitation block that does not inherits a feature hierarchy and RLA-vector, we can examine a pure impact on the usage of co-propagating RNN.

Context: simplified representation of feature volume. We focus on finding the context, a summarized feature volume, that assures preserving the information of local patterns during feature summarization procedure. Here, we concentrate on employing statistics that only summarize in

spatial ($H \times W$) dimensions of the feature volume, while not exploiting the channel-wise summary of feature volume. This intuition comes from the existence information of local patterns is less related to it's location.

First, we adopt global average pooled feature and global max pooled feature introduced in (Hu et al., 2018b; Woo et al., 2018) as contexts. Correspondingly, they summarize the spatial dimensions of the feature volume using average pooling and max pooling. We call these statistics as global average pooled features (GAP) and global max pooled features (GMP). The meaning of exploiting GAP for our works is considering the size or multiple appearances of local patterns in spatial dimensions of feature volume. On the other hand, GMP is used for catching the most salient part of local patterns.

Second, we additionally apply non-linear operations on GAP and GMP by attaching two fully connected layers, and call them as *GAP+MLP* and *GMP+MLP*. These statistics are for modeling the multiple appearance of similar local patterns at different channels. Seemingly, these statistics have same structure with SE module (Hu et al., 2018b). However, the goal for exploiting non-linearity is different. We use non-linearity for producing more informative context, while SE network is for predicting channel-wise scaling coefficients of the feature volume.

3.3 Model Complexity

Additional model parameters required for utilizing RLA module is given by:

$$4(M^2 + NM + M) + \frac{1}{r}\sum_{l=1}^{L}(M^2 + M), \tag{7}$$

where M is number of neurons in LSTM cell, N is a dimension of context, L is a number of layers in CNN, and r denotes the reduction ratio. As the formula implies, the additional model complexity caused by exploiting RLA module doesn't depend on the depth of backbone CNN. This is by the virtue of weight-sharing property from RNN. In contrast to ResNet-56 with SE module causes 1.51% parameters increase, ResNet-56 with our module (RLA) only increases 0.14% parameters compared to original ResNet-56.

4 Experiments and Analysis

In this section, we conduct ablation experiments for analyzing proposed instances and introduce experimental results on CIFAR and ImageNet-1K datasets for an image classification task, and Microsoft COCO datasets for an image detection task. We adopt ResNet (He et al., 2016) as a backbone CNN architecture for all experiments. Ensuring a fair comparison with our works, we re-implement the results of ResNet and Squeeze-and-Excitation network. Considering the readibility, implementation details over experiments are described on Appendix A and the parametric study for examining the effects of hyper-parameters (*i.e.*, M, N, r) are reported on Appendix B.

4.1 Ablation Study

The main goal of ablation study is demonstrating the expandability of RLA network. Key concepts illustrated in this part are as follows: the compatibility of RLA toward intra-layer attention mechanism, the impacts of inheriting feature hierarchy using RNN for empowering CNN to have better representational power, the effectiveness of introduced statistics for context. We adopt ResNet-56 as backbone structure for the ablation study.

Adopting Channel Attention. Table 1 shows every instances outperform ResNet-56 on CIFAR-10 and CIFAR-100 dataset. Interestingly, we find out a IA+RLA, the integrated structure of intra-layer attention and our approach, works worse than SE and RLA alone, but still better than ResNet. Note here, these results might have different tendency belonging to the backbone structure and dataset. The reason for that is two designs are on the complementary positions. Specifically, intra-layer attention mechanism adjusts features with similar size of receptive fields, while inter-layer attention mechanism focus on calibrating features with different size of receptive fields. We leave further experiments and analysis on this observation as future works.

Table 1: Ablation experimental results on instances showing the number of parameters and Top-1 errors for CIFAR-10/100 datasets. All ablation experiments utilize ResNet-56 as backbone architecture.

	Instance	Params	CIFAR-10	CIFAR-100
Backbone	–	0.862M	6.98	29.19
Baseline	RLA	0.864M	6.42	27.15
Channel attention	SE	0.870M	6.30	28.08
	IA+RLA	0.870M	6.42	28.49
	RLA-vector	0.876M	6.28	27.33
Context	GMP	0.864M	6.79	28.87
	GAP+MLP	0.871M	6.50	28.23
	GMP+MLP	0.871M	6.18	28.73

Distillation of Feature Hierarchy. Supporting a key design concept of RLA network, the exploitation of concurrently propagating RNN, we observe RLA-vector significantly works well compared to SE network. These experimental results claims that utilizing the feature hierarchy through RNN alone also remarkably aids CNN to interpret an image better. This result is interesting since it suggests the potential that the proposed concept of inheriting feature hierarchy through RNN can be solely applied to other neural network architectural designs.

Utilizing Various Context. Various feature summary statistics for the context differently affects on performance. Similar to the results reported in (Hu et al., 2018b; Woo et al., 2018), GMP reports higher error than GAP. Interestingly, the results on GAP+MLP and GMP+MLP are different depending on the dataset. Through this observation, we could recognize considering multiple appearances of local patterns on different channels in feature volume has negligible impact. We confirm GAP is the most appropriate context considering the performance increase and the model complexity.

4.2 IMAGE CLASSIFICATION

CIFAR (Krizhevsky & Hinton, 2009) dataset is consist of 50K training images and 10K validation images. We train networks for 160 epochs with the learning rate is initialized with 0.1 and divided by 10 when reached 80 and 120 epochs. The training and validation curve over CIFAR dataset are depicted in Figure 3. Through the optimization schedule, we observe that RLA network achieves lower training/validation error. Table 2 (left) summarizes experimental results. RLA distinctly surpasses an original ResNets and most of them also outperform the state-of-the-art channel-wise attention, SE network. At the view of model complexity, the requiring increase of parameters when applying RLA module to ResNet-56 is about 11 times smaller than applying SE module.

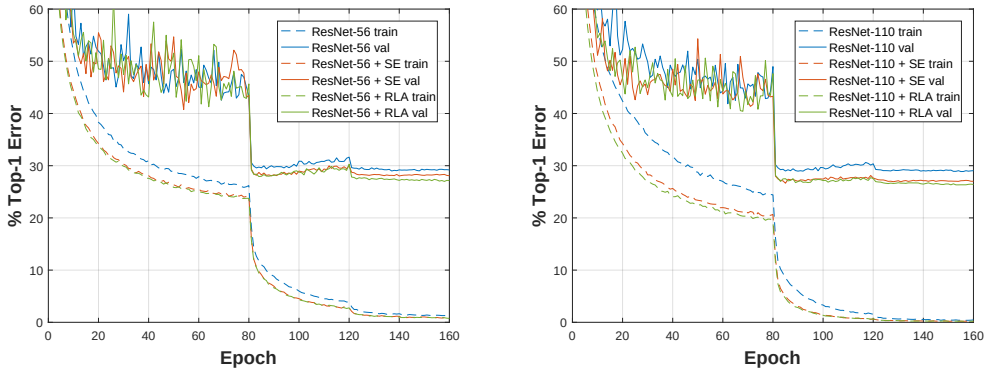

Figure 3: Training and validation error curve on the CIFAR-100 dataset of the ResNet-56 basebone(left) and ResNet-110 backbone(right).

Table 2: Image classification results on CIFAR-10/100 dataset(left), ImageNet-1K dataset(top-right). Object detection results on Microsoft COCO dataset(bottom-right). Correspondingly, Top-1 Errors for CIFAR, Top-1/5 Errors for ImageNet-1K, and mAP@.5 and mAP@[.5, .95] for Microsoft COCO are reported. We highlight cases that achieves the highest score on each metric.

	Params	△Params	CIFAR-10/100
ResNet-20	0.275M	–	8.75 / 32.21
SE-20	0.278M	0.86%	7.96 / **31.08**
RLA-20	0.276M	**0.30%**	**7.48** / 31.77
ResNet-32	0.471M	–	7.51 / 31.90
SE-32	0.475M	0.84%	7.32 / 30.34
RLA-32	0.471M	**0.20%**	**6.83 / 29.65**
ResNet-44	0.667M	–	7.17 / 31.13
SE-44	0.672M	0.83%	**6.64** / 29.54
RLA-44	0.668M	**0.17%**	6.68 / **28.75**
ResNet-56	0.858M	–	6.97 / 29.19
SE-56	0.871M	1.51%	**6.30** / 28.08
RLA-56	0.859M	**0.14%**	6.40 / **27.15**
ResNet-110	1.735M	–	6.61 / 28.95
SE-110	1.755M	1.15%	6.04 / 27.01
RLA-110	1.737M	**0.11%**	**5.92 / 26.40**

	Params	△Params	Top-1/5%
ResNet-18	1.169M	–	29.61 / 10.47
SE-18	1.178M	**0.76%**	**28.81 / 9.83**
RLA-18	1.190M	1.83%	29.35 / 10.31
ResNet-50	2.555M	–	24.84 / 7.55
SE-50	2.808M	**9.90%**	**23.64 / 7.03**
RLA-50	2.897M	13.35%	23.83 / 7.08
ResNet-101	4.454M	–	22.99 / 6.51
SE-101	4.933M	10.72%	**21.93 / 6.14**
RLA-101	4.824M	**8.29%**	22.35 / 6.37

	mAP@.5	mAP@[.5, .95]
ResNet-50	47.4	27.7
RLA-50	**48.2**	**27.8**
ResNet-101	50.3	30.5
RLA-101	**51.2**	**30.7**

ImageNet-1K 2012 (Deng et al., 2009) dataset is comprised of around 1.28M training images and 50K validation images. We train networks for 100 epochs on ImageNet-1K 2012 dataset with the learning rate is set to 0.1 and divided by 10 every 30 epochs. Data augmentation and optimization details follows the methods that ResNet (He et al., 2016) adopts. More implementation details are discussed on Appendix A. Table 2 (top-right) describes the experimental results. We notice that RLA network outperforms ResNet, but ranks below the SE network on performance. We interpret these results come from structural limitation of RLA module, when applied to backbone CNNs with large channels. Giving an example of ResNet-50 as a backbone architecture, RLA scale the network only using 16 scale coefficients (total numbers of residual blocks), while SE module adjusts the network using 15104 scale coefficients (total numbers of adjusting channels over all residual blocks). As mentioned in ablation study, we expect the integration of intra-layer mechanism and RLA network produces better results.

4.3 Object detection

Microsoft COCO 2014 (Lin et al., 2014) dataset contains 83K training images and 41K validation images. We adopt Faster R-CNN (Ren et al., 2015b) as a detection method and replace the baseline network from ResNet to RLA network. We train networks for 490K iterations with ImageNet-1K 2012 pre-trained ResNet and RLA networks. More implementation details are discussed on Appendix A. Table 2 (bottom-right) summarizes the experimental results. Here, we observe the improvements from original ResNet baseline, verifying a generalization performance on the object detection task.

5 Visualization

Grad-CAM visualziation. We hypothesized that RLA network gains additional class discriminability by considering repetitive appearances of unimportant features or relevant features with different receptive fields. For demonstrating above hypothesis and semantically understanding the way how RLA network learned, we applied Grad-CAM visualization (Selvaraju et al., 2017) over intermediate CNN layers. This visualization technique is of interest to our work since it provides a measure of importance of each pixel in a feature map towards the overall decision of the CNN (Chattopadhay et al., 2018).

Grad-CAM visualization is generally applied for the last CNN layer because the last CNN layer is normally recognized as the most semantically salient layer. Instead, we apply the Grad-CAM technique for the layers which are located directly before scale operation with layer attention

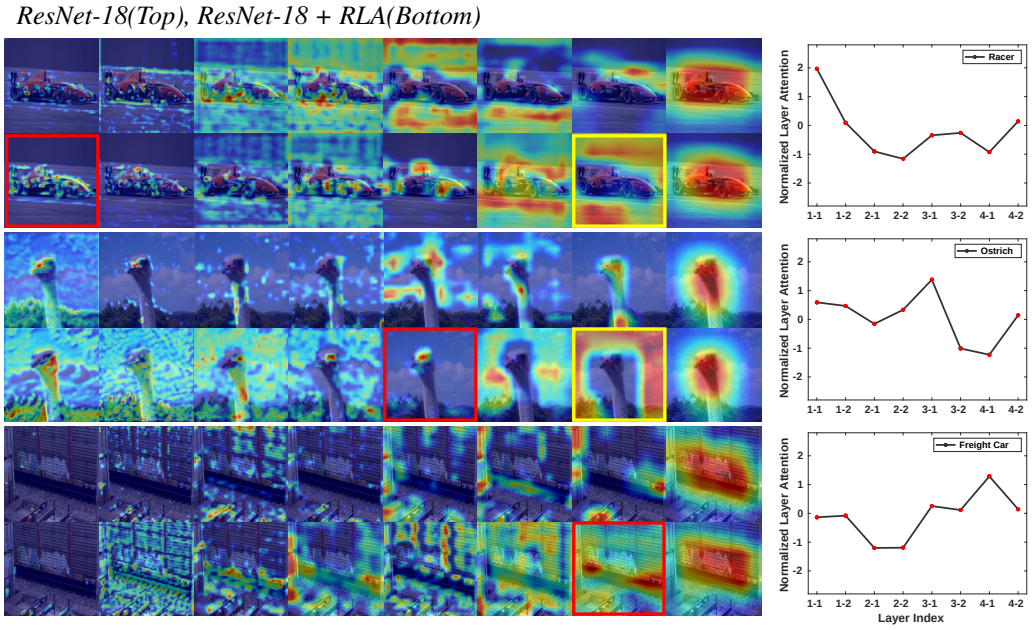

Figure 4: Grad-CAM visualization on the intermediate layers of RLA network(left) and normalized layer attention curve(right). Layer attention is layer-wisely normalized by substracting sample mean and dividing by sample standard deviation. The layer that achieve the highest layer attention value marked with red border, and a representative of layers that captures class-agnostic local patterns marked with yellow border.

values (*i.e.*, every last layer of each residual block). Figure 4 depicts the Grad-CAM visualization results on intermediate layers (left) and normalized layer attention curve (right). Interpreting with corresponding layer attention value, we could gain several insight about how RLA network induces their model parameters to be learned.

We select three target class having local patterns of different visual characteristics; racer, ostrich, freight car. Here, different visual characteristics signify different receptive fields of seemingly salient local patterns for discriminating target class. Sorting in ascending order, we recognize racer class has local patterns of smallest receptive fields such as a wing, wheels, ostrich class has local pattern of medium receptive field, head. Third, freight car class seems to have no specific local pattern or parts by itself. As results, we observe following interesting facts.

First, RLA network learns to enhance the layer attention value on layers that catch more seemingly salient features for each target class. The layers marked with red border denote layers that achieve the highest layer attention value. The first layer that captures simple pattern or small parts of racer class, the fifth layer that focus to catch head part of ostrich class, the seventh layer that captures global information or relations among locally ambiguous features for freight car class belongs that category. Through these observation, we hypothesize that RLA network emphasize the layer that catches class-specific local patterns for each target class.

Second, deep layers of RLA network tend to learn class-agnostic features after acquiring class-specific or semantically salient features in previous layers. The layers marked with yellow border denote a representative that less focus on class-specific local patterns in the images. Giving an example of ostrich class, the last layer of RLA network tend to learn class-agnostic local patterns, background. More visualizations on different target classes are on Appendix E. Comparing with original ResNet keeps allocating more class-specific neurons until deep layers, these observations are quite interesting. We additionally note that distinctly different highlighting pattern between close layers is because the residual architecture learns the ‘residual’ while keeping previously obtained features in earlier layers.

Furthermore, This observation accords with the investigation about the class selectivity (Morcos et al., 2018) on Gather-Excite network (Hu et al., 2018a); the intra-attention mechanism applied CNN has lower class selectivity in deeper layers compared to original backbone. We understand this finding is

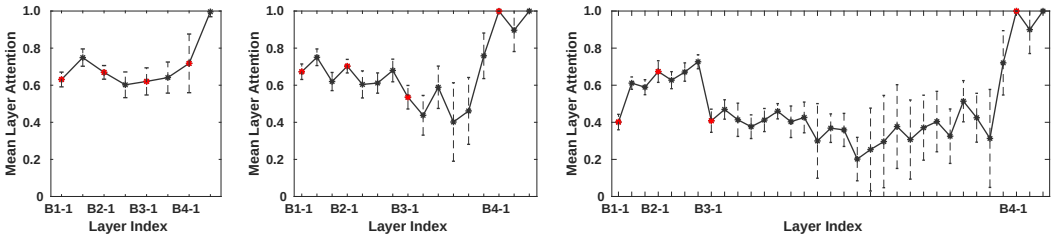

Figure 5: Mean layer attention curve over layers of RLA-18 (left), RLA-50 (middle), and RLA-101 (right). Data points signify mean layer attention value calculated on ImageNet 1K validation dataset. Vertical range on each point denotes an ± 1 standard deviation. We mark the first layer of each bottleneck block as red for visibility.

identically applied to our inter-attention mechanism (*i.e.*, RLA network) and our results support the results of previous work.

Mean layer attention curve. Figure 5 shows a mean layer attention curve along RLA network layers. Investigation on layer attention values over RLA network enable deeper understanding on the residual architecture, and traditional debates on CNN.

First, layer attention values are relatively high in shallow layers and rather lower values in middle layers. We recognize this observation could be another way for further understanding residual architecture. The intuition behind the residual learning is to let the layers learn the perturbations with reference to an identity function, and original paper (He et al., 2016) supports their intuition via displaying the standard deviation of feature responses along layers. Here, feature responses are defined as the outcomes (batch-normalized) of residual blocks. Comparing with that, we notice that our plot provides similar information but more straightforward. Through the observation on layer attention values decrease along layers, we could reason the intuition of the residual learning as latter layers learn less important layers.

Second, the variance of layer attention values increase along the layers except the last residual block, and layers in last residual block of RLA network has the biggest layer attention values that close to 1. We interpret traditional debates on CNN, shallow layers in CNN learn simple or "low-level" features while deep layers in CNN learn powerful, semantic, or "high-level" features (Donahue et al., 2014; Zeiler & Fergus, 2014) supports these observation. Supported by previous investigation, we think it's natural that RLA network learns large layer attention close to the latest CNN layers which are semantically salient, and layer attention values drastically vary on deep layers that contains abstract features.

6 Conclusion

In this paper, we first propose inter-layer attention mechansim to enhance the representational power of CNN. We structurized our mechanism as 'Recurrent Layer Attention network' by utilizing two new concepts: the weight balancing of CNN features along layers and the link from shallow CNN layers to deep CNN layers via RNN for directly conveying the feature hierarchy. We introduce structural derivatives of RLA network: 'IA+RLA' for proving an applicability of our work toward recent intra-layer attention mechanism, and 'RLA-vector' for distilling the impacts of proposed two new concepts. We also precisely select statistics for the context by focusing local patterns preserved in feature summarization procedure. We evaluate RLA network using CIFAR and ImageNet-1k 2012 datasets for an image classification task, and also verify it's generalization ability toward an object detection task via experiments on Microsoft COCO dataset. For demonstrating our hypothesis that RLA network gains additional class discriminability and semantically understanding how RLA network induce their model parameters to be learned, we visualize RLA network utilizing Grad-CAM visualization, plot the layer attention value curve, and report several interesting findings.

For future works, we are planning to integrate our inter-layer attention mechanism to intra-layer attention mechanism with heavier experiments first, and to utilize the concept of making any kinds of arbitrary connection from earlier layer to latter layer through RNN in other domains.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, 2018.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5659–5667, 2017.

Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Order-free rnn with visual attention for multi-label classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018a.

Xinlei Chen and Abhinav Gupta. An implementation of faster rcnn with study for region sampling. *arXiv preprint arXiv:1702.02138*, 2017.

Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. a^2-nets: Double attention networks. 2018b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pp. 647–655, 2014.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 9401–9411, 2018a.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018b.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pp. 1–9, 2015.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pp. 2204–2212, 2014.

Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*, 2018.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pp. 2953–2961, 2015a.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015b.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2017.

Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2285–2294, 2016.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.

Jianwei Yang, Jiasen Lu, Dhruv Batra, and Devi Parikh. A faster pytorch implementation of faster r-cnn. *https://github.com/jwyang/faster-rcnn.pytorch*, 2017.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019a.

Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention. *CoRR*, abs/1904.02998, 2019b. URL `http://arxiv.org/abs/1904.02998`.

A IMPLEMENTATION

Image recognition task. Data augmentation and optimization details follow He et al. (2016). We exploit scale augmentation, random-size cropping, and random horizontal flipping, not using the color augmentation. For optimization, we use nesterov SGD with momentum 0.9 with weight decay of 0.0001, and initialize weights by the strategy described in He et al. (2015).

Object detection task. For fast implementation on object detection tasks, we utilize ImageNet-1K pretrained ResNet, SE networt, and RLA network. Implementation details follows the works investigating the fast implementation of faster R-CNN (Yang et al., 2017; Chen & Gupta, 2017).

B PARAMETRIC STUDY

We present hyperparameters for designs of RLA such as a reduction ratio r of layer attention, a dimension of context N, the number of neurons M in LSTM cell. Since hyper-parameters are highly related to model complexity, it's a crucial issue to select them. In this section, we explain the standard to select hyper-parameters and demonstrate effects on the performance of the network.

Dimension of Context. Because feature volumes over the layers of CNN have different number of channels, summarized feature volume must be downsampled or upsampled for being inserted into LSTM. However, these operations cause an unintended lack or lost of information. Therefore, we select a dimension of context as a minimum number of downsampling or upsampling to be executed.

Size of LSTM Units. In our works, the size of LSTM hidden units, M, highly affects model and computational complexity. Following other works to find proper RNN hidden units size by rule-of-thumb or conducting experiments (Wang et al., 2016; Donahue et al., 2015), we perform experiments for finding the trade-off among the performance and complexities. Table 3 shows experimental results on varying $M \in [2, 4, 8, 16]$. Surprisingly, the performance of RLA does not drop as M decrease, rather increase. We deduce the reason of this observation is because the distributions of contexts are clearly separable, only small numbers of LSTM hidden units are enough for reducing considerable bias.

Reduction Ratio. Neither the reduction ratio, r, has golden rule as hidden units. Accordingly, we conduct experiments on the reduction ratio $r \in [2, 4, 8, 16]$, and found out there exists only slight differences on top-1%-error contrast to experimental results with varying hidden units. Therefore, we compress model complexity of RLA network by choosing small M which also affects model performance, and further regulate the complexity using reduction ratio.

C DETAILS ON EXPERIMENTS

Exploiting big M on RLA-vector. Following interesting tendency that smaller M induce performance increase, we conduct an ablation experiments with $M = 4$. However, we found that RLA-vector with small M yields prominent performance decrease. We interpret this as, reduced model capacity by decreasing M can not scale channel-sized vector of each layer in CNN. For the fair comparison with SE network, we set $M = 32$ considering that the average size of global average pooled feature volumes also equals to 32 in SE-ResNet-56.

hyperparameter selection in ImageNet-1K experiments. We observe the tendency on performance by changing hyperparameter in parametric study. However, This tendency does not always fit in other dataset too. Considering the data diversity of ImageNet-1K 2012 datasets, we exploit commonly used RNN hidden units size for ResNet-56, $M = 512$, as exploited in previous CNN-RNN paper (Wang et al., 2016). We note that further consideration on differing LSTM hidden units size M possibly reports better performance even with smaller parameters, as shown in the parametric study on CIFAR dataset. About reduction ratio, we apply $r = 8$ for RLA-18/50 and $r = 16$ for RLA-101.

D STRUCTURE OF RESNET-50 + RLA

We describe the location of connection between CNN and RNN in RLA-50 network. When exploiting vanila CNN, applying RLA to backbone CNN is straightforward, making connections with concur-

rently propagating RNN just after every convolution layer of the network. However, it's confusing to apply RLA when using residual architecture He et al. (2016). Table 4 shows the structure of backbone CNN using ResNet-50 and ResNet-50 + RLA. Note here, RNN that concurrently propagating with CNN is not depicted for ResNet-50 + RLA. In short, RLA network stores the context and infers the layer attention for scaling features *per each residual block*.

E GRAD-CAM VISUALIZATION ON VARIOUS OBJECT CLASS

For aiding semantic understanding of how RLA network induces their model paramters to be learned, we provide Grad-CAM visualization Selvaraju et al. (2017) results of the intermediate feature maps over the CNN layers. We could find out similar observations debated on the main article on other target class; the tendency to enhance the layer attention value of layers such that catches the most semantically important visual features, and another tendency to learn non class-specific features in deep layers. Figure 6 depicts visualization results on the target class of junco, achidna, killer whales, leonberg, tiger cat, sleeping bags.

Table 3: Comparison on model performance with different size of LSTM hidden units (M) and reduction ratios (r). All parametric experiments exploit a ResNet-56 as backbone architecture, and the context of "GAP" on CIFAR-10 and CIFAR-100 datasets.

LSTM units size (M)	CIFAR-10/100
2	6.58 / 27.47
4	6.42 / 27.15
8	6.35 / 27.92
16	6.54 / 27.72

Reduction ratio (r)	CIFAR-10/100
2	6.32 / 28.13
4	6.58 / 28.06
8	6.49 / 28.84
16	6.54 / 28.12

Table 4: The structure of ResNet50 + RLA, where M denotes the size of LSTM hidden units.

Output size	ResNet-50	ResNet-50 + RLA
112×112	conv, 1×1, 64, stride 2	
	max-pool, 3×3, stride 2	
56×56	$\begin{bmatrix} conv, & 1\times1, & 64 \\ conv, & 3\times3, & 64 \\ conv, & 1\times1, & 256 \end{bmatrix} \times 3$	$\begin{bmatrix} conv, & 1\times1, & 64 \\ conv, & 3\times3, & 64 \\ conv, & 1\times1, & 256 \\ fc, & [M,1] & \end{bmatrix} \times 3$
28×28	$\begin{bmatrix} conv, & 1\times1, & 128 \\ conv, & 3\times3, & 128 \\ conv, & 1\times1, & 512 \end{bmatrix} \times 4$	$\begin{bmatrix} conv, & 1\times1, & 128 \\ conv, & 3\times3, & 128 \\ conv, & 1\times1, & 512 \\ fc, & [M,1] & \end{bmatrix} \times 4$
14×14	$\begin{bmatrix} conv, & 1\times1, & 256 \\ conv, & 3\times3, & 256 \\ conv, & 1\times1, & 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} conv, & 1\times1, & 256 \\ conv, & 3\times3, & 256 \\ conv, & 1\times1, & 1024 \\ fc, & [M,1] & \end{bmatrix} \times 6$
7×7	$\begin{bmatrix} conv, & 1\times1, & 512 \\ conv, & 3\times3, & 512 \\ conv, & 1\times1, & 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} conv, & 1\times1, & 512 \\ conv, & 3\times3, & 512 \\ conv, & 1\times1, & 2048 \\ fc, & [M,1] & \end{bmatrix} \times 3$
1×1	global average-pool, 1000-d, fc, softmax	

ResNet-18(Top), ResNet-18 + RLA(Bottom)

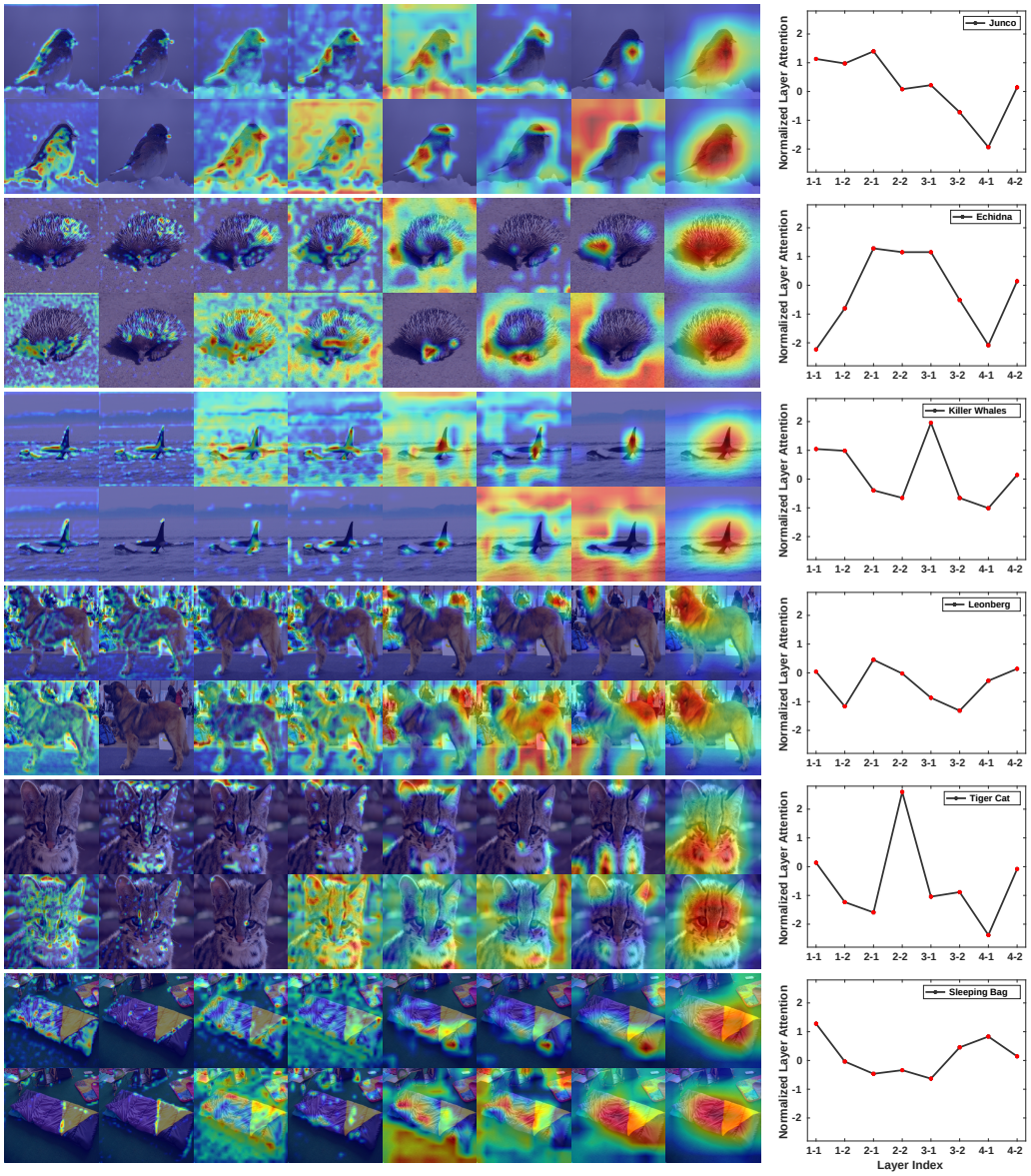

Figure 6: Grad-CAM visualization on the intermediate layers of RLA network. Target class of Junco, achidna, killer whales, leonberg, tiger cat, sleeping bags(top to bottom) are depicted.