Residual Attention Network for Image Classification

Fei Wang[1], Mengqing Jiang[2], Chen Qian[1], Shuo Yang[3], Cheng Li[1],
Honggang Zhang[4], Xiaogang Wang[3], Xiaoou Tang[3]
[1]SenseTime Group Limited, [2]Tsinghua University,
[3]The Chinese University of Hong Kong, [4]Beijing University of Posts and Telecommunications
[1]{wangfei, qianchen, chengli}@sensetime.com, [2]jmq14@mails.tsinghua.edu.cn
[3]{ys014, xtang}@ie.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk, [4]zhhg@bupt.edu.cn

Abstract

In this work, we propose "Residual Attention Network", a convolutional neural network using attention mechanism which can incorporate with state-of-art feed forward network architecture in an end-to-end training fashion. Our Residual Attention Network is built by stacking Attention Modules which generate attention-aware features. The attention-aware features from different modules change adaptively as layers going deeper. Inside each Attention Module, bottom-up top-down feedforward structure is used to unfold the feedforward and feedback attention process into a single feedforward process. Importantly, we propose attention residual learning to train very deep Residual Attention Networks which can be easily scaled up to hundreds of layers.

*Extensive analyses are conducted on CIFAR-10 and CIFAR-100 datasets to verify the effectiveness of every module mentioned above. Our Residual Attention Network achieves state-of-the-art object recognition performance on three benchmark datasets including CIFAR-10 (3.90% error), CIFAR-100 (20.45% error) and ImageNet (4.8% single model and single crop, top-5 error). Note that, our method achieves **0.6%** top-1 accuracy improvement with **46%** trunk depth and **69%** forward FLOPs comparing to ResNet-200. The experiment also demonstrates that our network is robust against noisy labels.*

1. Introduction

Not only a friendly face but also red color will draw our attention. The mixed nature of attention has been studied extensively in the previous literatures [34, 16, 23, 40]. Attention not only serves to select a focused location but also enhances different representations of objects at that location. Previous works formulate attention drift as a sequential process to capture different attended aspects. However, as far as we know, no attention mechanism has been applied to feedforward network structure to achieve state-of-art results in image classification task. Recent advances of image classification focus on training feedforward convolutional neural networks using "very deep" structure [27, 33, 10].

Inspired by the attention mechanism and recent advances in the deep neural network, we propose Residual Attention Network, a convolutional network that adopts mixed attention mechanism in "very deep" structure. The Residual Attention Network is composed of multiple Attention Modules which generate attention-aware features. The attention-aware features from different modules change adaptively as layers going deeper.

Apart from more discriminative feature representation brought by the attention mechanism, our model also exhibits following appealing properties:
(1) Increasing Attention Modules lead to consistent performance improvement, as different types of attention are captured extensively. Fig.1 shows an example of different types of attentions for a hot air balloon image. The sky attention mask diminishes background responses while the balloon instance mask highlighting the bottom part of the balloon.
(2) It is able to incorporate with state-of-the-art deep network structures in an end-to-end training fashion. Specifically, the depth of our network can be easily extended to hundreds of layers. Our Residual Attention Network outperforms state-of-the-art residual networks on CIFAR-10, CIFAR-100 and challenging ImageNet [5] image classification dataset with significant reduction of computation (**69%** forward FLOPs).

All of the aforementioned properties, which are challenging to achieve with previous approaches, are made possible with following contributions:
(1) *Stacked network structure*: Our Residual Attention Network is constructed by stacking multiple Attention Modules. The stacked structure is the basic application of mixed attention mechanism. Thus, different types of attention are able to be captured in different Attention Modules.

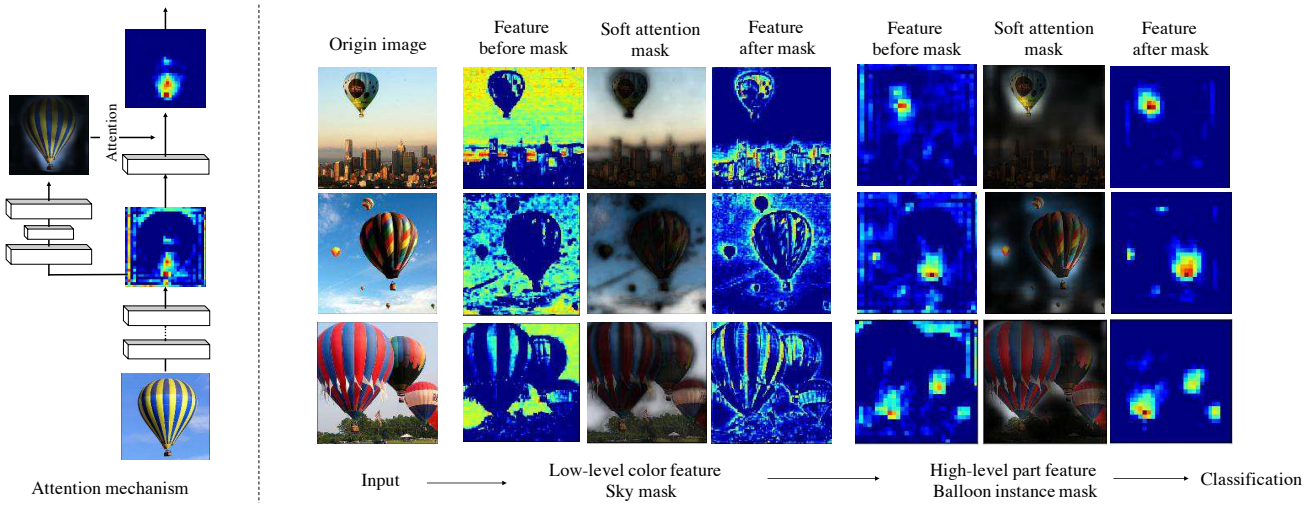

Figure 1: **Left:** an example shows the interaction between features and attention masks. **Right:** example images illustrating that different features have different corresponding attention masks in our network. The sky mask diminishes low-level background blue color features. The balloon instance mask highlights high-level balloon bottom part features.

(2) *Attention Residual Learning*: Stacking Attention Modules directly would lead to the obvious performance drop. Therefore, we propose attention residual learning mechanism to optimize very deep Residual Attention Network with hundreds of layers.

(3) *Bottom-up top-down feedforward attention*: Bottom-up top-down feedforward structure has been successfully applied to human pose estimation [24] and image segmentation [22, 25, 1]. We use such structure as part of Attention Module to add soft weights on features. This structure can mimic bottom-up fast feedforward process and top-down attention feedback in a single feedforward process which allows us to develop an end-to-end trainable network with top-down attention. The bottom-up top-down structure in our work differs from stacked hourglass network [24] in its intention of guiding feature learning.

2. Related Work

Evidence from human perception process [23] shows the importance of attention mechanism, which uses top information to guide bottom-up feedforward process. Recently, tentative efforts have been made towards applying attention into deep neural network. Deep Boltzmann Machine (DBM) [21] contains top-down attention by its reconstruction process in the training stage. Attention mechanism has also been widely applied to recurrent neural networks (RNN) and long short term memory (LSTM) [13] to tackle sequential decision tasks [25, 29, 21, 18]. Top information is gathered sequentially and decides where to attend for the next feature learning steps.

Residual learning [10] is proposed to learn residual of identity mapping. This technique greatly increases the depth of feedforward neuron network. Similar to our work, [25, 29, 21, 18] use residual learning with attention mechanism to benefit from residual learning. Two information sources (query and query context) are captured using attention mechanism to assist each other in their work. While in our work, a single information source (image) is split into two different ones and combined repeatedly. And residual learning is applied to alleviate the problem brought by repeated splitting and combining.

In image classification, top-down attention mechanism has been applied using different methods: sequential process, region proposal and control gates. Sequential process [23, 12, 37, 7] models image classification as a sequential decision. Thus attention can be applied similarly with above. This formulation allows end-to-end optimization using RNN and LSTM and can capture different kinds of attention in a goal-driven way.

Region proposal [26, 4, 8, 38] has been successfully adopted in image detection task. In image classification, an additional region proposal stage is added before feedforward classification. The proposed regions contain top information and are used for feature learning in the second stage. Unlike image detection whose region proposals rely on large amount of supervision, e.g. the ground truth bounding boxes or detailed segmentation masks [6], unsupervised learning [35] is usually used to generate region proposals for image classification.

Control gates have been extensively used in LSTM. In image classification with attention, control gates for neu-

rones are updated with top information and have influence on the feedforward process during training [2, 30]. However, a new process, reinforcement learning [30] or optimization [2] is involved during the training step. Highway Network [29] extends control gate to solve gradient degradation problem for deep convolutional neural network.

However, recent advances of image classification focus on training feedforward convolutional neural networks using "very deep" structure [27, 33, 10]. The feedforward convolutional network mimics the bottom-up paths of human cortex. Various approaches have been proposed to further improve the discriminative ability of deep convolutional neural network. VGG [27], Inception [33] and residual learning [10] are proposed to train very deep neural networks. Stochastic depth [14], Batch Normalization [15] and Dropout [28] exploit regularization for convergence and avoiding overfitting and degradation.

Soft attention developed in recent work [3, 17] can be trained end-to-end for convolutional network. Our Residual Attention Network incorporates the soft attention in fast developing feedforward network structure in an innovative way. Recent proposed spatial transformer module [17] achieves state-of-the-art results on house number recognition task. A deep network module capturing top information is used to generate affine transformation. The affine transformation is applied to the input image to get attended region and then feed to another deep network module. The whole process can be trained end-to-end by using differentiable network layer which performs spatial transformation. Attention to scale [3] uses soft attention as a scale selection mechanism and gets state-of-the-art results in image segmentation task.

The design of soft attention structure in our Residual Attention Network is inspired by recent development of localization oriented task, *i.e*. segmentation [22, 25, 1] and human pose estimation [24]. These tasks motivate researchers to explore structure with fined-grained feature maps. The frameworks tend to cascade a bottom-up and a top-down structure. The bottom-up feedforward structure produces low resolution feature maps with strong semantic information. After that, a top-down network produces dense features to inference on each pixel. Skip connection [22] is employed between bottom and top feature maps and achieved state-of-the-art result on image segmentation. The recent stacked hourglass network [24] fuses information from multiple scales to predict human pose, and benefits from encoding both global and local information.

3. Residual Attention Network

Our Residual Attention Network is constructed by stacking multiple Attention Modules. Each Attention Module is divided into two branches: mask branch and trunk branch. The trunk branch performs feature processing and can be adapted to any state-of-the-art network structures. In this work, we use pre-activation Residual Unit [11], ResNeXt [36] and Inception [32] as our Residual Attention Networks basic unit to construct Attention Module. Given trunk branch output $T(x)$ with input x, the mask branch uses bottom-up top-down structure [22, 25, 1, 24] to learn same size mask $M(x)$ that softly weight output features $T(x)$. The bottom-up top-down structure mimics the fast feedforward and feedback attention process. The output mask is used as control gates for neurons of trunk branch similar to Highway Network [29]. The output of Attention Module H is:

$$H_{i,c}(x) = M_{i,c}(x) * T_{i,c}(x) \tag{1}$$

where i ranges over all spatial positions and $c \in \{1, ..., C\}$ is the index of the channel. The whole structure can be trained end-to-end.

In Attention Modules, the attention mask can not only serve as a feature selector during forward inference, but also as a gradient update filter during back propagation. In the soft mask branch, the gradient of mask for input feature is:

$$\frac{\partial M(x,\theta)T(x,\phi)}{\partial \phi} = M(x,\theta)\frac{\partial T(x,\phi)}{\partial \phi} \tag{2}$$

where the θ are the mask branch parameters and the ϕ are the trunk branch parameters. This property makes Attention Modules robust to noisy labels. Mask branches can prevent wrong gradients (from noisy labels) to update trunk parameters. Experiment in Sec.4.1 shows the robustness of our Residual Attention Network against noisy labels.

Instead of stacking Attention Modules in our design, a simple approach would be using a single network branch to generate soft weight mask, similar to spatial transformer layer [17]. However, these methods have several drawbacks on challenging datasets such as ImageNet. First, images with clutter background, complex scenes, and large appearance variations need to be modeled by different types of attentions. In this case, features from different layers need to be modeled by different attention masks. Using a single mask branch would require exponential number of channels to capture all combinations of different factors. Second, a single Attention Module only modify the features once. If the modification fails on some parts of the image, the following network modules do not get a second chance.

The Residual Attention Network alleviates above problems. In Attention Module, each trunk branch has its own mask branch to learn attention that is specialized for its features. As shown in Fig.1, in hot air balloon images, blue color features from bottom layer have corresponding sky mask to eliminate background, while part features from top layer are refined by balloon instance mask. Besides, the incremental nature of stacked network structure can gradually refine attention for complex images.

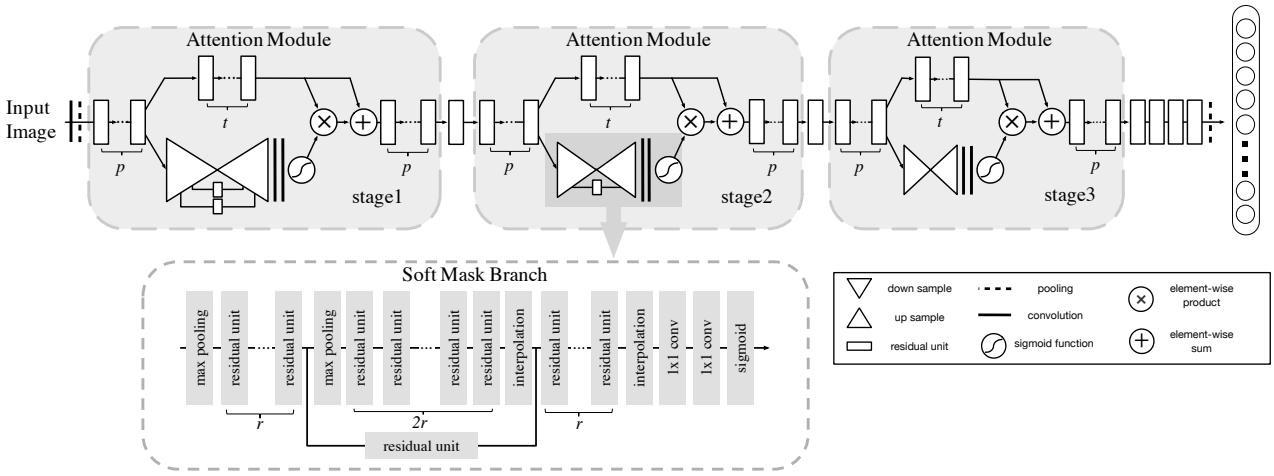

Figure 2: Example architecture of the proposed network for ImageNet. We use three hyper-parameters for the design of Attention Module: p, t and r. The hyper-parameter p denotes the number of pre-processing Residual Units before splitting into trunk branch and mask branch. t denotes the number of Residual Units in trunk branch. r denotes the number of Residual Units between adjacent pooling layer in the mask branch. In our experiments, we use the following hyper-parameters setting: $\{p = 1, t = 2, r = 1\}$. The number of channels in the soft mask Residual Unit and corresponding trunk branches is the same.

3.1. Attention Residual Learning

However, naive stacking Attention Modules leads to the obvious performance drop. First, dot production with mask range from zero to one repeatedly will degrade the value of features in deep layers. Second, soft mask can potentially break good property of trunk branch, for example, the identical mapping of Residual Unit.

We propose attention residual learning to ease the above problems. Similar to ideas in residual learning, if soft mask unit can be constructed as identical mapping, the performances should be no worse than its counterpart without attention. Thus we modify output H of Attention Module as

$$H_{i,c}(x) = (1 + M_{i,c}(x)) * F_{i,c}(x) \tag{3}$$

$M(x)$ ranges from $[0, 1]$, with $M(x)$ approximating 0, $H(x)$ will approximate original features $F(x)$. We call this method attention residual learning.

Our stacked attention residual learning is different from residual learning. In the origin ResNet, residual learning is formulated as $H_{i,c}(x) = x + F_{i,c}(x)$, where $F_{i,c}(x)$ approximates the residual function. In our formulation, $F_{i,c}(x)$ indicates the features generated by deep convolutional networks. The key lies on our mask branches $M(x)$. They work as feature selectors which enhance good features and suppress noises from trunk features.

In addition, stacking Attention Modules backs up attention residual learning by its incremental nature. Attention residual learning can keep good properties of original features, but also gives them the ability to bypass soft mask branch and forward to top layers to weaken mask branch's feature selection ability. Stacked Attention Modules can gradually refine the feature maps. As show in Fig.1, features become much clearer as depth going deeper. By using attention residual learning, increasing depth of the proposed Residual Attention Network can improve performance consistently. As shown in the experiment section, the depth of Residual Attention Network is increased up to 452 whose performance surpasses ResNet-1001 by a large margin on CIFAR dataset.

3.2. Soft Mask Branch

Following previous attention mechanism idea in DBN [21], our mask branch contains fast feed-forward sweep and top-down feedback steps. The former operation quickly collects global information of the whole image, the latter operation combines global information with original feature maps. In convolutional neural network, the two steps unfold into bottom-up top-down fully convolutional structure.

From input, max pooling are performed several times to increase the receptive field rapidly after a small number of Residual Units. After reaching the lowest resolution, the global information is then expanded by a symmetrical top-down architecture to guide input features in each position. Linear interpolation up sample the output after some Residual Units. The number of bilinear interpolation is the same as max pooling to keep the output size the same as the input feature map. Then a sigmoid layer normalizes the output

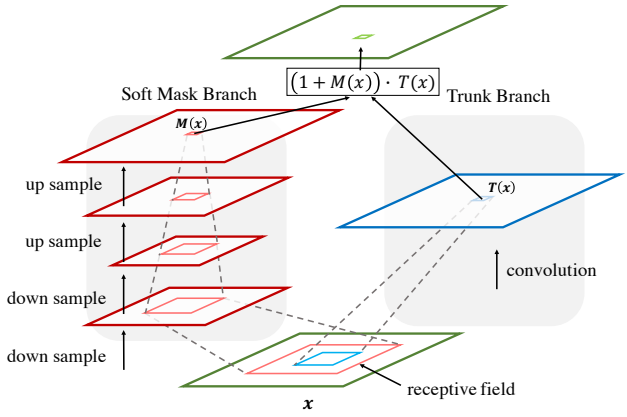

Figure 3: The receptive field comparison between mask branch and trunk branch.

range to $[0, 1]$ after two consecutive 1×1 convolution layers. We also added skip connections between bottom-up and top-down parts to capture information from different scales. The full module is illustrated in Fig.2.

The bottom-up top-down structure has been applied to image segmentation and human pose estimation. However, the difference between our structure and the previous one lies in its intention. Our mask branch aims at improving trunk branch features rather than solving a complex problem directly. Experiment in Sec.4.1 is conducted to verify above arguments.

3.3. Spatial Attention and Channel Attention

In our work, attention provided by mask branch changes adaptably with trunk branch features. However, constrains to attention can still be added to mask branch by changing normalization step in activation function before soft mask output. We use three types of activation functions corresponding to mixed attention, channel attention and spatial attention. Mixed attention f_1 without additional restriction use simple sigmoid for each channel and spatial position. Channel attention f_2 performs $L2$ normalization within all channels for each spatial position to remove spatial information. Spatial attention f_3 performs normalization within feature map from each channel and then sigmoid to get soft mask related to spatial information only.

$$f_1(x_{i,c}) = \frac{1}{1 + exp(-x_{i,c})} \tag{4}$$

$$f_2(x_{i,c}) = \frac{x_{i,c}}{\|x_i\|} \tag{5}$$

$$f_3(x_{i,c}) = \frac{1}{1 + exp(-(x_{i,c} - \text{mean}_c)/\text{std}_c)} \tag{6}$$

Where i ranges over all spatial positions and c ranges over all channels. mean_c and std_c denotes the mean value and standard deviation of feature map from c-th channel. x_i denotes the feature vector at the ith spatial position.

Activation Function	Attention Type	Top-1 err. (%)
$f_1(x)$	Mixed Attention	**5.52**
$f_2(x)$	Channel Attention	6.24
$f_3(x)$	Spatial Attention	6.33

Table 1: The test error (%) on CIFAR-10 of Attention-56 network with different activation functions.

Layer	Output Size	Attention-56	Attention-92
Conv1	112×112	7×7, 64, stride 2	
Max pooling	56×56	3×3 stride 2	
Residual Unit	56×56	$\begin{pmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{pmatrix} \times 1$	
Attention Module	56×56	Attention ×1	Attention ×1
Residual Unit	28×28	$\begin{pmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{pmatrix} \times 1$	
Attention Module	28×28	Attention ×1	Attention ×2
Residual Unit	14×14	$\begin{pmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{pmatrix} \times 1$	
Attention Module	14×14	Attention ×1	Attention ×3
Residual Unit	7×7	$\begin{pmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{pmatrix} \times 3$	
Average pooling	1×1	7×7 stride 1	
FC,Softmax	1000		
params×10^6		31.9	51.3
FLOPs×10^9		6.2	10.4
Trunk depth		56	92

Table 2: Residual Attention Network architecture details for ImageNet. Attention structure is described in Fig. 2. We make the size of the smallest output map in each mask branch 7×7 to be consistent with the smallest trunk output map size. Thus 3,2,1 max-pooling layers are used in mask branch with input size 56×56, 28×28, 14×14 respectively. The Attention Module is built by pre-activation Residual Unit [11] with the number of channels in each stage is the same as ResNet [10].

The experiment results are shown in Table 1, the mixed attention has the best performance. Previous works normally focus on only one type of attention, for example scale attention [3] or spatial attention [17], which puts additional constrain on soft mask by weight sharing or normalization. However, as supported by our experiments, making attention change adaptively with features without additional constraint leads to the best performance.

4. Experiments

In this section, we evaluate the performance of proposed Residual Attention Network on a series of benchmark datasets including CIFAR-10, CIFAR-100 [19], and ImageNet [5]. Our experiments contain two parts. In the first part, we analyze the effectiveness of each component in the Residual Attention Network including attention residual

learning mechanism and different architectures of soft mask branch in the Attention Module. After that, we explore the noise resistance property. Given limited computation resources, we choose CIFAR-10 and CIFAR-100 dataset to conduct these experiments. Finally, we compare our network with state-of-the-art results in CIFAR dataset. In the second part, we replace the Residual Unit with Inception Module and ResNeXt to demonstrate our Residual Attention Network surpasses origin networks both in parameter efficiency and final performance. We also compare image classification performance with state-of-the-art ResNet and Inception on ImageNet dataset.

4.1. CIFAR and Analysis

Implementation. The CIFAR-10 and CIFAR-100 datasets consist of $60,000$ 32×32 color images of 10 and 100 classes respectively, with $50,000$ training images and $10,000$ test images. The broadly applied state-of-the-art network structure ResNet is used as baseline method. To conduct fair comparison, we keep most of the settings same as ResNet paper [10]. The image is padded by 4 pixels on each side, filled with 0 value resulting in 40×40 image. A 32×32 crop is randomly sampled from an image or its horizontal flip, with the per-pixel RGB mean value subtracted. We adopt the same weight initialization method following previous study [9] and train Residual Attention Network using nesterov SGD with a mini-batch size of 64. We use a weight decay of 0.0001 with a momentum of 0.9 and set the initial learning rate to 0.1. The learning rate is divided by 10 at 64k and 96k iterations. We terminate training at 160k iterations.

The overall network architecture and the hyper parameters setting are described in Fig.2. The network consists of 3 stages and similar to ResNet [10], equal number of Attention Modules are stacked in each stage. Additionally, we add two Residual Units at each stage. The number of weighted layers in trunk branch is $36m$+20 where m is the number of Attention Module in one stage. We use original 32×32 image for testing.

Attention Residual Learning. In this experiment, we evaluate the effectiveness of attention residual learning mechanism. Since the notion of attention residual learning (ARL) is new, no suitable previous methods are comparable therefore we use "naive attention learning" (NAL) as baseline. Specifically, "naive attention learning" uses Attention Module where features are directly dot product by soft mask without attention residual learning. We set the number of Attention Module in each stage $m = \{1, 2, 3, 4\}$. For Attention Module, this leads to Attention-56 (named by trunk layer depth), Attention-92, Attention-128 and Attention-164 respectively.

We train these networks using different mechanisms and

Network	ARL (Top-1 err. %)	NAL (Top-1 err.%)
Attention-56	**5.52**	5.89
Attention-92	**4.99**	5.35
Attention-128	**4.44**	5.57
Attention-164	**4.31**	7.18

Table 3: Classification error (%) on CIAFR-10.

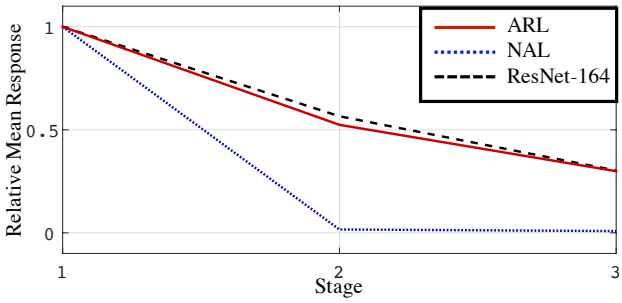

Figure 4: The mean absolute response of output features in each stage.

summarize the results in the Table 3. As shown in Table 3, the networks trained using attention residual learning technique consistently outperform the networks trained with baseline method which proves the effectiveness of our method. The performance increases with the number of Attention Module when applying attention residual learning. In contrast, the performance of networks trained with "naive attention learning" method suffers obvious degradation with increased number of Attention Module.

To understand the benefit of attention residual learning, we calculate mean absolute response value of output layers for each stage. We use Attention-164 to conduct this experiment. As shown in the Fig. 4, the response generated by the network trained using naive attention learning quickly vanishes in the stage 2 after four Attention Modules compared with network trained using attention residual learning. The Attention Module is designed to suppress noise while keeping useful information by applying dot product between feature and soft mask. However, repeated dot product will lead to severe degradation of both useful and useless information in this process. The attention residual learning can relieve signal attenuation using identical mapping, which enhances the feature contrast. Therefore, it gains benefits from noise reduction without significant information loss, which makes optimization much easier while improving the discrimination of represented features. In the rest of the experiments, we apply this technique to train our networks.

Comparison of different mask structures. We conduct experiments to validate the effectiveness of encoder-decoder structure by comparing with local convolutions without any down sampling or up sampling. The local convolutions soft mask consists of three Residual Units us-

ing the same number of FLOPs. The Attention-56 is used to construct Attention-Encoder-Decoder-56 and Attention-Local-Conv-56 respectively. Results are shown in Table 4. The Attention-Encoder-Decoder-56 network achieves lower test error 5.52% compared with Attention-Local-Conv-56 network 6.48% with a considerable margin 0.94%. The result suggests that the soft attention optimization process will benefit from multi-scale information.

Mask Type	Attention Type	Top-1 err. (%)
Local Convolutions	Local Attention	6.48
Encoder and Decoder	Mixed Attention	**5.52**

Table 4: Test error (%) on CIFAR-10 using different mask structures.

Noisy Label Robustness. In this experiment, we show our Residual Attention Network enjoys noise resistant property on CIFAR-10 dataset following the setting of paper [31]. The confusion matrix Q in our experiment is set as follows:

$$Q = \begin{pmatrix} r & \frac{1-r}{9} & \cdots & \frac{1-r}{9} \\ \frac{1-r}{9} & r & \cdots & \frac{1-r}{9} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-r}{9} & \frac{1-r}{9} & \cdots & r \end{pmatrix}_{10\times 10} \tag{7}$$

where r denotes the clean label ratio for the whole dataset.

We compare ResNet-164 network with Attention-92 network under different noise levels. The Table 5 shows the results. The test error of Attention-92 network is significantly lower than ResNet-164 network with the same noise level. In addition, when we increase the ratio of noise, test error of Attenion-92 declines slowly compared with ResNet-164 network. These results suggest that our Residual Attention Network can perform well even trained with high level noise data. When the label is noisy, the corresponding mask can prevent gradient caused by label error to update trunk branch parameters in the network. In this way, only the trunk branch is learning the wrong supervision information and soft mask branch masks the wrong label.

Comparisons with state-of-the-art methods. We compare our Residual Attention Network with state-of-the-art methods including ResNet [11] and Wide ResNet [39] on

Noise Level	ResNet-164 err. (%)	Attention-92 err. (%)
10%	5.93	5.15
30%	6.61	5.79
50%	8.35	7.27
70%	17.21	15.75

Table 5: Test error (%) on CIFAR-10 with label noises.

Network	params$\times 10^6$	CIFAR-10	CIFAR-100
ResNet-164 [11]	1.7	5.46	24.33
ResNet-1001 [11]	10.3	4.64	22.71
WRN-16-8 [39]	11.0	4.81	22.07
WRN-28-10 [39]	36.5	4.17	20.50
Attention-92	1.9	4.99	21.71
Attention-236	5.1	4.14	21.16
Attention-452†	8.6	**3.90**	**20.45**

Table 6: Comparisons with state-of-the-art methods on CIFAR-10/100. †: the Attention-452 consists of Attention Module with hyper-parameters setting: $\{p = 2,\ t = 4,\ r = 3\}$ and 6 Attention Modules per stage.

CIFAR-10 and CIFAR-100 datasets. The results are shown in Table 6. Our Attention-452 outperforms all the baseline methods on CIFAR-10 and CIFAR-100 datasets. Note that Attention-92 network achieves 4.99% test error on CIFAR-10 and 21.71% test error on CIFAR-100 compared with 5.46% and 24.33% test error on CIFAR-10 and CIFAR-100 for ResNet-164 network under similar parameter size. In addition, Attention-236 outperforms ResNet-1001 using only half of the parameters. It suggests that our Attention Module and attention residual learning scheme can effectively reduce the number of parameters in the network while improving the classification performance.

4.2. ImageNet Classification

In this section, we conduct experiments using ImageNet LSVRC 2012 dataset [5], which contains $1,000$ classes with 1.2 million training images, $50,000$ validation images, and $100,000$ test images. The evaluation is measured on the non-blacklist images of the ImageNet LSVRC 2012 validation set. We use Attention-56 and Attention-92 to conduct the experiments. The network structures and hyper parameters can be found in the Table 2.

Implementation. Our implementation generally follows the practice in the previous study [20]. We apply scale and aspect ratio augmentation [33] to the original image. A 224×224 crop is randomly sampled from an augment image or its horizontal flip, with the per-pixel RGB scale to $[0, 1]$ and mean value subtracted and standard variance divided. We adopt standard color augmentation [20]. The network is trained using SGD with a momentum of 0.9. We set initial learning rate to 0.1. The learning rate is divided by 10 at 200k, 400k, 500k iterations. We terminate training at 530k iterations.

Mask Influence. In this experiment, we explore the efficiency of proposed Residual Attention Network. We compare Attention-56 with ResNet-152 [10]. The ResNet-152 has 50 trunk Residual Units and 60.2×10^6 parameters com-

Network	params$\times 10^6$	FLOPs$\times 10^9$	Test Size	Top-1 err. (%)	Top-5 err. (%)
ResNet-152 [10]	60.2	11.3	224×224	22.16	6.16
Attention-56	31.9	6.3	224×224	**21.76**	**5.9**
ResNeXt-101 [36]	44.5	7.8	224×224	21.2	5.6
AttentionNeXt-56	31.9	6.3	224×224	**21.2**	**5.6**
Inception-ResNet-v1 [32]	-	-	299×299	21.3	5.5
AttentionInception-56	31.9	6.3	299×299	**20.36**	**5.29**
ResNet-200 [11]	64.7	15.0	320×320	20.1	4.8
Inception-ResNet-v2	-	-	299×299	19.9	4.9
Attention-92	51.3	10.4	320×320	**19.5**	**4.8**

Table 7: Single crop validation error on ImageNet.

pared with 18 trunk Residual Units and 31.9×10^6 parameters in Attention-56. We evaluate our model using single crop scheme on the ImageNet validation set and show results in Table 7. The Attention-56 network outperforms ResNet-152 by a large margin with a 0.4% reduction on top-1 error and a 0.26% reduction on top-5 error. More importantly, Attention-56 network achieves better performance with only 52% parameters and 56% FLOPs compared with ResNet-152, which suggests that the proposed attention mechanism can significantly improve network performance while reducing the model complexity.

Different Basic Units. In this experiment, we show Residual Attention Network can generalize well using different basic unit. We apply three popular basic units: Residual Unit, ResNeXt [36], and Inception [32] to construct our Residual Attention Networks. To keep the number of parameters and FLOPs in the same scale, we simplify the Inception. Results are shown in Table 7.

When the basic unit is ResNeXt, the AttentionNeXt-56 network performance is the same as ResNeXt-101 while the parameters and FLOPs are significantly fewer than ResNeXt-101. For Inception, The AttentionIncepiton-56 outperforms Inception-ResNet-v1 [32] by a margin with a 0.94% reduction on top-1 error and a 0.21% reduction on top-5 error. The results show that our method can be applied on different network structures.

Comparisons with State-of-the-art Methods. We compare our Attention-92 evaluated using single crop on the ILSVRC 2012 validation set with state-of-the-art algorithms. Table 7 shows the results. Our Attention-92 outperforms ResNet-200 with a large margin. The reduction on top-1 error is 0.6%. Note that the ResNet-200 network contains 32% more parameters than Attention-92. The computational complexity of Attention-92 shown in the Table 7 suggests that our network reduces nearly half training time comparing with ResNet-200 by adding attention mechanism and reducing trunk depth. Above results suggest that our model enjoys high efficiency and good performance.

5. Discussion

We propose a Residual Attention Network which stacks multiple Attention Modules. The benefits of our network are in two folds: it can capture mixed attention and is an extensible convolutional neural network. The first benefit lies in that different Attention Modules capture different types of attention to guide feature learning. Our experiments on the forms of activation function also validate this point: free form mixed attention will have better performance than constrained (including single) attention. The second benefit comes from encoding top-down attention mechanism into bottom-up top-down feedforward convolutional structure in each Attention Module. Thus, the basic Attention Modules can be combined to form larger network structure. Moreover, residual attention learning allows training very deep Residual Attention Network. The performance of our model surpasses state-of-the-art image classification methods, *i.e.* ResNet on CIFAR-10 (3.90% error), CIFAR-100 (20.67% error), and challenging ImageNet dataset (0.6% top-1 accuracy improvement) with only 46% trunk depth and 69% forward FLOPs (comparing with ResNet-200). In the future, we will exploit different applications of deep Residual Attention Network such as detection and segmentation to better explore mixed attention mechanism for specific tasks.

References

[1] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*, 2015. 2, 3

[2] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, 2015. 3

[3] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. *arXiv preprint arXiv:1511.03339*, 2015. 3, 5

[4] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015. 2

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE. 1, 5, 7

[6] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *CVPR*, 2014. 2

[7] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, 2015. 2

[8] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 6

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2, 3, 5, 6, 7, 8

[11] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*, 2016. 3, 5, 7, 8

[12] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. *arXiv preprint arXiv:1511.05284*, 2015. 2

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997. 2

[14] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger. Deep networks with stochastic depth. *arXiv preprint arXiv:1603.09382*, 2016. 3

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 3

[16] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2001. 1

[17] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 3, 5

[18] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*, pages 361–369, 2016. 2

[19] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009. 5

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 7

[21] H. Larochelle and G. E. Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *NIPS*, 2010. 2, 4

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2, 3

[23] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014. 1, 2

[24] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *arXiv preprint arXiv:1603.06937*, 2016. 2, 3

[25] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015. 2, 3

[26] A. Shrivastava and A. Gupta. Contextual priming and feedback for faster r-cnn. In *ECCV*, 2016. 2

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 1, 3

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014. 3

[29] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *NIPS*, 2015. 2, 3

[30] M. F. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber. Deep networks with internal selective attention through feedback connections. In *NIPS*, 2014. 3

[31] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014. 7

[32] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. 3, 8

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 1, 3, 7

[34] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognitiona gentle way. In *International Workshop on Biologically Motivated Computer Vision*, pages 472–479. Springer, 2002. 1

[35] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015. 2

[36] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016. 3, 8

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[38] S. Yang, P. Luo, C. C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *ICCV*, 2015. 2

[39] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 7

[40] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan. Diversified visual attention networks for fine-grained object classification. *arXiv preprint arXiv:1606.08572*, 2016. 1