Residual Attention Neural Network

E4040.2019Fall.ZZZZ.report
Jie Li jl5246, Xiaofan Zhang xz2735
Columbia University

Abstract

In this paper, we re-implement "Residual Attention Network", a convolutional neural network that adopts mixed attention mechanism into very deep structure for image classification task [4]. This particular model can generate attention-aware feature and is robust to noisy images. We construct 56-layer and 92-layer Residual Attention Networks to classify small size images and solve low accuracy problem by tuning parameters. To evaluate the effectiveness of models, performance analyses are conducted on CIFAR-10 and CIFAR-100 datasets.

Our Residual Attention Network with 56-layer achieves 3.26% and 5.38% error on CIFAR-10 and CIFAR-100 dataset. We also show Naive Attention Learning leads 3.75% performance drop than Attention Residual Learning. The model with 92-layer only reaches 66.81% accuracy.

1. Introduction

The image classification task is about how machines can understand of visual information. In recent years, researchers focus on training feedforward convolutional neural network with "deeper" structures, but we are looking for a new mechanism beyond the feature maps. "Rather than compress an entire image into a static representation, the attention allows for salient features to dynamically come to the forefront as needed" [5].

In this paper, we re-implement Residual Attention Network, a convolutional network that adopts mixed attention mechanism on CIFAR-10 and CIFAR-100 datasets. We train our models on google cloud environment with GPU to improve computational capacity. However, we are unable to train deeper models or use ImageNet dataset due to limited resources. We also point out the errors on the original paper corresponding to parameters of optimizer which triggers low performance.

2. Summary of the Original Paper

2.1 Work of the Original Paper

The original paper introduced Residual Attention Network that constructed with stacking attention modules and Residual units. The new model structure combines characteristic of both visual attention and Residual learning that could be robust to images with complex background, noise and extended to hundreds of layers.

The paper firstly showed model effectiveness between Residual Learning (ResNet) and Attention Residual Network (ARL). The researchers also conducted experiments to validate the model performance of both naïve stacking Attention Modules (NAL) and Attention Residual Network (ARL), and of different mask structures. They showed how ARL enjoyed noise resistant property.

Second, they constructed and trained several ARLs stacking different number of Attention Modules such as Attention-56, Attention-92, Attention-128 and Attention-164 to compare results on CIFAR-10, CIFAR-100 and ImageNet in major factors such as test accuracy and number of parameters.

2.2 Key Results of the Original Paper

Based on the original paper, researchers compared with Naive Attention Learning (NAL) and Attention Residual Learning (ARL), showing the performance drop.

1. Dot production with mask range from zero to one repeatedly will degrade the value of features in deep layers [4].
2. Soft mask is constructed as identical mapping of Residual Unit to enhance good features and suppress noises from trunk features [4].

Network	ARL (err %)	NAL (err %)
Attention-56	5.52	5.89
Attention-92	4.99	5.35

Table 1: Test error (%) on CIFAR-10 using ARL and NAL [4]

Their Attention-56 network achieves 5.52% test error on CIFAR-10, and Attention-92 network achieves 4.99% test error on CIFAR-10 and 21.71% on CIFAR-100 as the Table 2:

Network	CIFAR10	CIFAR100	Params (10^6)
Attention56	5.52		
Attention92	4.99	21.71	1.9
ResNet164	5.46	24.33	1.7

Table 2: Test error (%) on CIFAR-10 and CIFAR-100 [4]

They also displayed the test errors on CIFAR-10 using different mask structures. 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 3: Test error (%) on CIFAR-10 using different mask structures [4]

Residual Attention Network also outperforms Residual Network on noisy images, as shown in the following table. The result indicates the soft mask is resistant to wrong labels.

Noise Level	ResNet-164 (err %)	Attention-92(err %)
10%	5.93	5.15
20%	6.61	5.79
50%	8.35	7.27
70%	17.21	15.75

Table 4: Test error (%) on CIFAR-10 with label noises [4]

3. Methodology

The Residual Attention Network is stacked by multiple Attention modules beyond the modern convolutional neural network to perform image classification. Researchers borrowed the structure of pre-activation Residual Unit, ResNeXt and Inception to construct Attention Module. They introduced concept of trunk branch and soft mask branch to compose Attention.

3.1. Residual connections

Residual connections also called skip connection performs identity mapping and their results are added to the output of the stacked layers [3]. A residual layer can be described as follows:

$$H(x) = x + F(x)$$

, where $H(x)$ is referred as a residual model and $F(x)$ indicates the features generated by deep convolutional networks. It is used to allow gradients to flow through a network directly without passing through non-linear activation [7], so that avoid degradation problem when model gets deeper.

3.2. Residual Inception Unit

The inception layer is computed by three convolution branches such as a 1 x 1 convolution branch, a 3 x 3 convolution branch, and a double 1 x 1 convolution branch [6]. The three effective types of convolution filters allow the module to focus on different scaled features of the inputs.

3.3. Attention Module

The Attention Module consists of two branches: trunk branch $T(x)$ and soft mask branch $M(x)$.

Trunk branch performs as a global feature processor that connects with a number of Residual Inception Units. Each trunk branch has own mask branch to guide feature learning.

Soft Mask Branch applies bottom-up top-down structure in order to quickly collect global information and combine the information with original feature maps [4]. Also, they added two consecutive 1x1 convolution layers and sigmoid activation function to normalize the output, which is used to control gates for neurons of trunk branch [4]. It also introduced skip connection into the structure to capture information from different scale. Thus, the output of attention module looks like:

$$H(x) = [1 + M(x)] * F(x) \qquad (1)$$

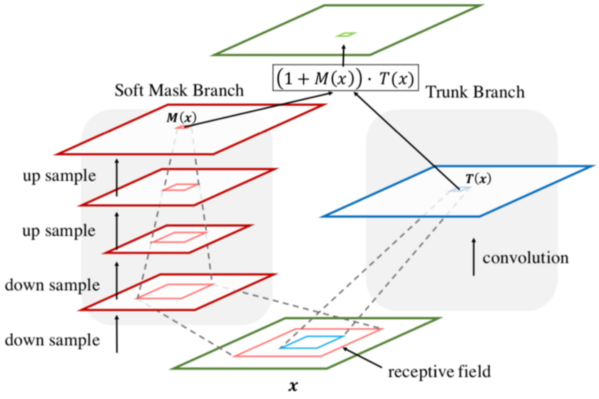


Figure 1: Attention Module with trunk branch and mask branch

4. Implementation

In this section, we describe and explain our model implementation about the structure of Residual Attention Network, dataset and training procedure. We implement our work on Google Cloud environment with one GPU (NVIDIA Tesla P100), use library TensorFlow (2.0) and Tensorboard (2.0.2).

4.1. Deep Learning Network

The Table 5 is showing the overview structure in detail of Residual Attention Network: 56-layer and 92-layer, stacking multiple Attention modules on the modern convolutional neural network.

Layer	Attention -56	Attention -92	Output Size
Conv2D	5x5x32, stride=2		32x32x32
Max pooling	2x2 stride=2		16x16x32
Residual Unit	(32,32,128)		16x16x128
Attention	x1	x1	16x16x128
Residual Unit	(64,64,256)		8x8x256
Attention	x1	x2	8x8x256
Residual Unit	(128,128,512)		4x4x512
Attention	x1	x3	4x4x512
Residual Unit	(256,256,1024) x3		4x4x1024
Avg pooling	4x4 stride=1		1x1x1024
Dropout			1x1x1024
FC, Softmax			10
Params	28M	65M	
Depth	56	92	

Table 5: Residual Attention Network architecture details for CIFAR-10 and CIFAR-100

4.2. Datasets

The CIFAR-10 and CIFAR-100 datasets consist of 50,000 training set and 10,000 test set with 32 x 32 RGB images, representing 10/100 different image labels. We apply the data augmentation technique that generate image rotation, shifting, and horizontal flip, with the per-pixel RGB mean value subtracted.

4.3. Training Procedure

In original paper, researchers used Nesterov SGD with a momentum of 0.9, but does not work well in our task. We change to use Adam optimizer with a mini-batch size of 128 and 0.001 weight decay. Additionally, we also create the early stop mechanism and learning rate scheduler that set initial learning rate to 1e-4 and reduce by 0.1 factor for every 20 epochs.

To improve model performance, three different regularization methods are applied such as batch normalization, dropout and L2 norm to avoid overfitting and improve speed, performance, and stability of our models.

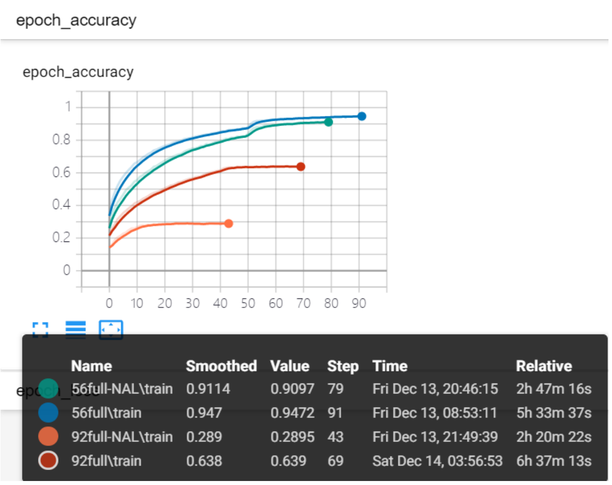
5. Results

5.1. Project Results

First of all, we show that the Residual Attention Network performance drops by 3.75% on CIFAR-10 dataset using Naive Residual Learning (NAL), compared with Attention Residual Learning (ARL) as Table 6 shows:

Network	ARL (error)	NAL (error)
Attention-56	5.28%	9.03%
Attention-92	36.2%	40.05%

Table 6: Error (%) using ARL and NAL on CIFAR-10



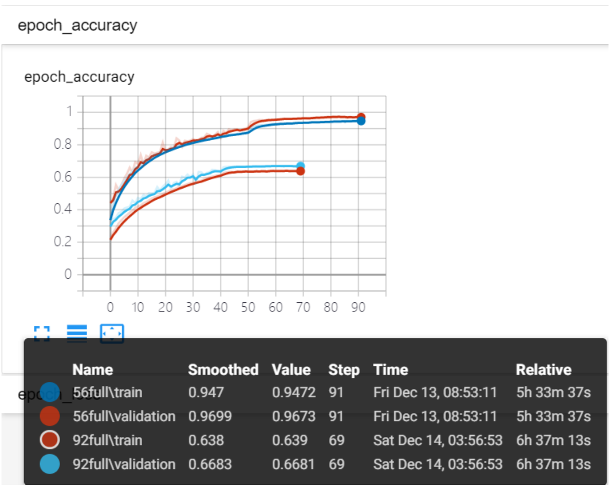
	Name	Smoothed	Value	Step	Time	Relative
●	56full-NAL\train	0.9114	0.9097	79	Fri Dec 13, 20:46:15	2h 47m 16s
●	56full\train	0.947	0.9472	91	Fri Dec 13, 08:53:11	5h 33m 37s
●	92full-NAL\train	0.289	0.2895	43	Fri Dec 13, 21:49:39	2h 20m 22s
●	92full\train	0.638	0.639	69	Sat Dec 14, 03:56:53	6h 37m 13s

Figure 2: Training accuracy and running time details for Attention-56 and Attention-92 using ARL and NAL on CIFAR-10

Next, we evaluate the Attention-56 and Attention-92 Networks on CIFAR-10 by validation error, test error and training time. For the small images (32 x 32), Attention-56 works very well and achieves 3.27% train error and 5.28% validation error spending 337 minutes on training. Attention-92, stacking more Attention Modules only ends up at 66% accuracy.

Network	Train Error	Val Error	Train Time
Attention-56	3.27%	5.28%	337 min
Attention-92	36.1%	33.19%	403 min

Table 7: Error (%) on Attention-56 and Attention-92 on CIFAR-10



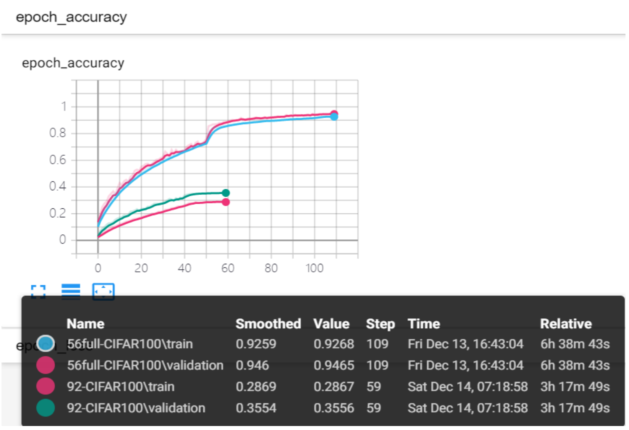
	Name	Smoothed	Value	Step	Time	Relative
●	56full\train	0.947	0.9472	91	Fri Dec 13, 08:53:11	5h 33m 37s
●	56full\validation	0.9699	0.9673	91	Fri Dec 13, 08:53:11	5h 33m 37s
●	92full\train	0.638	0.639	69	Sat Dec 14, 03:56:53	6h 37m 13s
●	92full\validation	0.6683	0.6681	69	Sat Dec 14, 03:56:53	6h 37m 13s

Figure 3: Training accuracy and running time details for Attention-56 and Attention-92 on CIFAR-10

Finally, we also both models on CIFAR-100. Attention-56 works very well and achieves 7.32% train error and 5.40% validation error spending 337 minutes on training. Attention-92, stacking more Attention Modules only ends up at 36% accuracy as Table 8.

Network	Train Error	Val Error	Train Time
Attention-56	7.32%	5.40%	402 min
Attention-92	71.33%	64.4%	200 min

Table 8: Error (%) for Attention-56 and Attention-92 on CIFAR-100



	Name	Smoothed	Value	Step	Time	Relative
●	56full-CIFAR100\train	0.9259	0.9268	109	Fri Dec 13, 16:43:04	6h 38m 43s
●	56full-CIFAR100\validation	0.946	0.9465	109	Fri Dec 13, 16:43:04	6h 38m 43s
●	92-CIFAR100\train	0.2869	0.2867	59	Sat Dec 14, 07:18:58	3h 17m 49s
●	92-CIFAR100\validation	0.3554	0.3556	59	Sat Dec 14, 07:18:58	3h 17m 49s

Figure 4: Training accuracy and running time details for Attention-56 and Attention-92 on CIFAR-100

5.2. Comparison of Results

First thing first, we compare with Nesterov SGD and Adam optimizer on Attention-56 networks on CIFAR-10. In our experiment, we get very bad results as paper setting, however, the Attention-56 network with Adam optimizer reaches 5.28%, closed to the paper result 5.52%, so we decide to use Adam optimizer for our network models training.

Second, the paper shows the networks trained using Attention Residual Learning (ARL) outperforms Naive Attention Learning (NAL). In our experiment, we are succeed to show 3.75% performance difference between Naive Residual Learning (NAL) and Attention Residual Learning (ARL) on CIFAR-10 dataset.

Third, we reproduce Attention-92 architecture followed by the structure provided by original paper, but does not work well on both CIFAR-10 and CIFAR-100. The model on the paper reaches 4.99% and 21.71% error, but ours is larger than 30%. More importantly, Attention-92 they built only has 1.9 million parameters, even smaller than Attention-56. We will talk about more concern in the next discussion section.

5.3. Discussion of Insights Gained

Comparing with original paper, we list three major aspects regarding model structures and training procedure which might cause the different results in our experiment, and also help for future improvements.

The Residual Attention Network architecture from original paper is designed for ImageNet dataset, a set of 224 x 224 RGB images. Because of the down-sampling mechanism in Attention Module, it causes the dimension issue if apply on the small size image datasets as CIFAR. Therefore, we believe that Attention-92 Network is only good for the large size image classification. That might be the reason we end up high error rate.

Our training procedure setting is slightly different than original paper, which also might lead to a biased result. Due to limitation of computation resource, we set 150 epochs, small learning rate, and increase the batch size from 64 to 128 in order to speed up the training process, but it might lead models not completely converge.

Lastly, the paper mentions skip connections in the soft mask branch, but they did not put details how it works and constructs. In our experiment, we design simple skips connections {2, 1, 0} as showing in paper diagrams. For future improvements, we would like to research on skip connection setting. We might also apply Attention modules in different networks such as AlexNet and VGG in small image classification. Additionally, we may introduce attention on further computer vision applications such as object detection and localization.

6. Conclusion

In our paper, we re-implement the Residual Attention Network which stacks multiple Attention Modules on modern convolutional neural network. The major benefits of our work are two parts. We learn the mechanism behind Attention Module, which softly weight learn features from bottom-up top-down structure. More importantly, we are experienced in building deeper networks from scratch to solve image classification problems, and applying different regularization methods such as L2 norm, dropout, and batch normalization to achieve 5.28% error on CIFAR-10 images.

7. References

[1] Our Github

[2] V. Fung, An Overview of ResNet and its Variants, 2017.7.15, https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

[3] K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recoginition, arXiv: 1512.03385, 2015.12, 10

[4] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, Xiao. Wang and X. Tang, Residual Attention Network for Image Classification, arXiv:1704.06904, 2017.4.23

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, arXiv: 1502.03044, 2016.4.19

[6] T. Edirisooriya, Residual Attention Networks for Image Classification

[7] What are "residual connections" In RNN?

8. Appendix

8.1 Individual student contributions in fractions - table

Full name	Jie Li	Xiaofan Zhang
UNI	jl5246	xz2735
Fraction of (useful) total contribution	1/2	1/2
What I did 1	Building Model Structure	Building Model Structure
What I did 2	Training	Tensorboard
What I did 3	Github, Code comment	Paper