STAT GR5242 Advanced Machine Learning

Improved Residual Network for Image Classification

Author: Jie Li (jl5246), Zhaoyang Wang (zw2551), Xiaofan Zhang (xz2735)

Abstract

Attention mechanism has been widely used to improve performance of deep learning models. Inspired by Fei Wang's paper, we introduce the attention concept to Residual Network named Residual Attention Network. However, Residual Attention Networks in the original paper are very deep and expensive to train, we design our own network architecture to reduce network size without lossing accuracy in order to improve model efficiency.

Extensive analyses are conducted on CIFAR-10 dataset to verfiy the effectiveness of model. our new architecture of Residual Attention Network reduces 40% model parameters and saves 75% training time than the original model, but only loss less than 1% on the model accuracy.

1. Introduction

Convolutional Neural Network (CNN) is the most popular neural netowrk model being used for image classification problem which can help break images and extracts the high-level features. However, instead of compressing an entire image into a static representation, the attention serves to select a focused location and enhances different representations of objects at that location. Therefore, in practice the model with attention modules can be very robust to noisy images with complex background.

The original paper introduced "Residual Attention Network", a convolutional neural network stacking Attention Modules which generate attention-aware features. In our task, we propose a new architecture of Residual Attention Network different from original paper for image classification. First, we construct and train a simple Residual Attention Network on CIFAR-10 dataset to compare Nesterov SGD and Adam optimizer. Then, we improve our model architecture by applying different regularization methods. Lastly, we visualize the model performance in training, validation accuracy and training time by Tensorbaord to compare our model with the original model from paper.

2. Method

2.1. Residual Learning

Degradation problem is exposed with the deeper neural network due to vanishing gradients and curse of dimensionality. The solution is to skip the training of few layers performing indentity mapping, which can be expressed in a general form:

$$y_l = x_l + F(x_l, W_l)$$

where F is a residual function. With skip connection, it is easy to propagate larger gradients to initial layers and these layers also could learn as fast as the final layers. The Figure 1 shows a single Residual Unit structure. In our task, we also apply pre-activation (ReLU and Batch Normalization) into Residual Unit, which could be much easier to train and generalize better than the original ResNet [11].

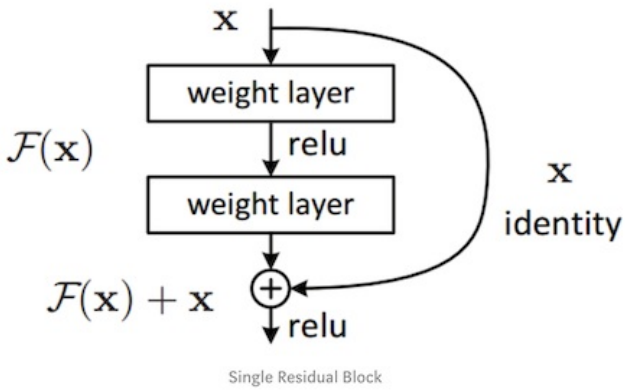


Figure 1. Single Residual Unit [6]

2.2. Attention Module

Deeper neural networks are difficult to train, so we apply Residual Network structure for image classification in order to alleviate optimization difficulty. One of the more unique developments in image classification architecture is the use of visual attention to determine patches of an image to focus on, and then classifying based on the area of focus [10].

Inspired by Fei Wang's paper, our attention is constructed by two branches: trunk branch and soft mask branch. Trunk branch $T(x)$ is built by two pre-activation Residual Units, which performs a global feature extractor to collect information globally. Soft mask branch $M(x)$ is considered as a particular feature selector which performs as a control gate to softly weight outputs from trunk branch. It applys bottom-up top-down fully convolutional structure as Figure 3: first down-sampling to quickly learn global information of the whole image, and then global information is expanded by up sampling to help guide input features back to original feature map.

To avoid degradation, we add a skip connection (identity mapping) into the Attention Module. Thus, the expression of Attention Module is:

$$H_{i,c}(x) = [1 + M_{i,c}(x, \theta)] * T_{i,c}(x, \phi)$$

where i ranges over all spatial positions and c is the index of the channel, θ and ϕ are parameters of soft mask branch and trunk branch respectively.

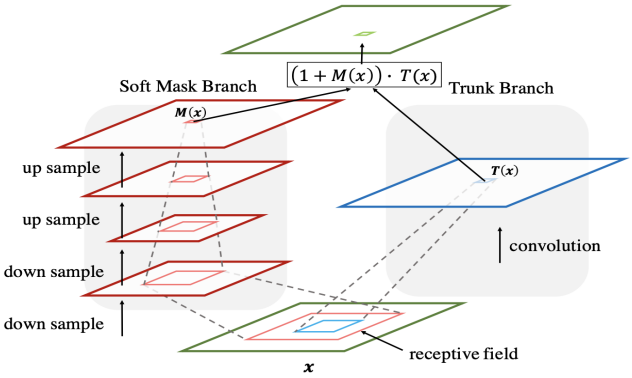


Figure 2. Attention Module [5]

2.3. Optimization Algorithm

In our task, we choose Adaptive Moment Estimation (Adam), a method that computes adaptive learning rates for each parameter. It uses estimations of first and second moments of gradient to scale learning rate individually for each weight of the neural network [1].

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

Solving the moving average process above, we get expression of unbiased estimator $\hat{m}_t$, $\hat{v}_t$ and formula for updating parameter:

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

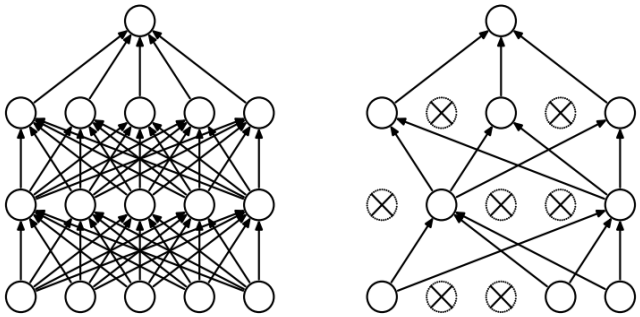
$$w_{t+1} = w_t - \eta \frac{\hat{m}_t}{\sqrt{v_t} + \varepsilon}$$

where m and v are moving averages, $\hat{m}, \hat{v}$ are unbiased estimator of m and v, g is gradient on current mini-batch, η is step size, and β_i are hyper-parameters of the algorithm.

Adam works well in practice and compares favorably to other adaptive learning-method algorithms as it converges very fast and the learning speed of the model is quiet fast and efficient. Also it rectifies every problem that is faced in other optimization techniques such as vanishing Learning rate, slow convergence or high variance in the parameter updates which leads to fluctuating loss function.

2.4. Dropout

To solve overfitting issue, we introduce dropout method in our Residual Attention Network. It refers to drop out units in a neural network. By dropping a unit out, we temporarily remove it from the network, along with all its incoming and outgoing connections [8].



(a) Standard Neural Net (b) After applying dropout.

Figure 3. Dropout [8]

2.5. Batch Normalization

Batch normalization is used for normalizing input data for each layer. It is a technique for improving speed, performance, and stability of model. For each feature, batch normalization computes the mean and variance of that feature in mini-batch. Then it subtracts the mean and variance and divides the feature by its mini-batch standard deviation.

3. Experiments and Results

Instead of reproducing the same models from original paper, we start with our simple base model, and then improve model structures by adding regu algorithms we just talked above. Also, we describe our model structure in detail and visualize the model performance in tensorboard.

3.1. Residual Attention Network Architecture

As Table 1, we provide our architecture of Residual Attention Network, and compare with the model structure from the original paper. We remove one Attention Module and skip connection, but add different regularization layers into model. Therefore, we reduce 4.7M parameters from paper model, which is about 40% model reduction.

Layer	Output Size	Our Model	Paper Model
Conv2D	32x32x32	(5x5 stride=1)	(5x5 stride=1)
Batch Norm	16x16x32	x1	Not Applied
Activation	16x16x32	ReLU	Not Applied
Max pooling	16x16x32	(2x2 stride=2)	(2x2 stride=2)
Residual Unit	16x16x128	x1	x1
Attention	16x16x128	x1	x1
Residual Unit	8x8x256	x1	x1
Attention	8x8x256	x1	x2
Residual Unit	4x4x1024	x3	x3
Batch Norm	4x4x1024	x1	Not Applied
Activation	4x4x1024	ReLU	Not Applied
AvgPooling2D	1x1x1024	(4x4)	(4x4)
Flatten	1x1x1024	x1	x1
Dropout	1x1x1024	x1	Not Applied
Dense	10	Softmax, L2 Norm	Softmax
Params (10^6)		6.8	11.5

Table 1. Model Structure

3.2. Implementation

The CIFAR-10 dataset consists of 50,000 training and 10,000 test images with 32 x 32 RGB images, representing 10 different labels. We implement data augmentation technique such that rotate, flip or shift images with the per-pixel RGB mean value subtracted.

We adopt the idea from paper using nesterov SGD with a momentum of 0.9. On the other hand, we also attempt Adam optimizer with a mini-batch size of 128 and 0.001 weight decay to compare the difference. Additionally, in order to make model converge faster, we create early stop mechanism and learning rate scheduler that shrink learning late for every 20 epochs.

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

3.3. Results

From figure 4, we find that the model with Adam increases 0.25 accuracy on both training and validation sets. And, the SGD case is early stoped at 48 epoch and 15 of these epochs have no improvement. Therefore, we believe the Adam optimizer is a better choice for our task.

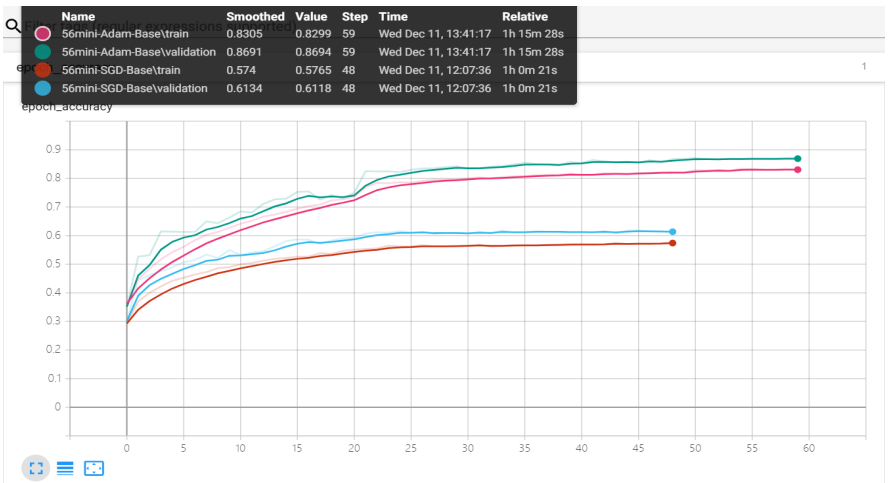


Figure 4. Training accuracy using Adam and SGD optimizor

Then, we realize that base Residual Attention Network has 0.039 overfitting between training and validation accuracy, so we invite our regularization methods to handle overfitting issue. For following image, it reduces a little overfitting, but obviously improves the model accuracy by 0.024.

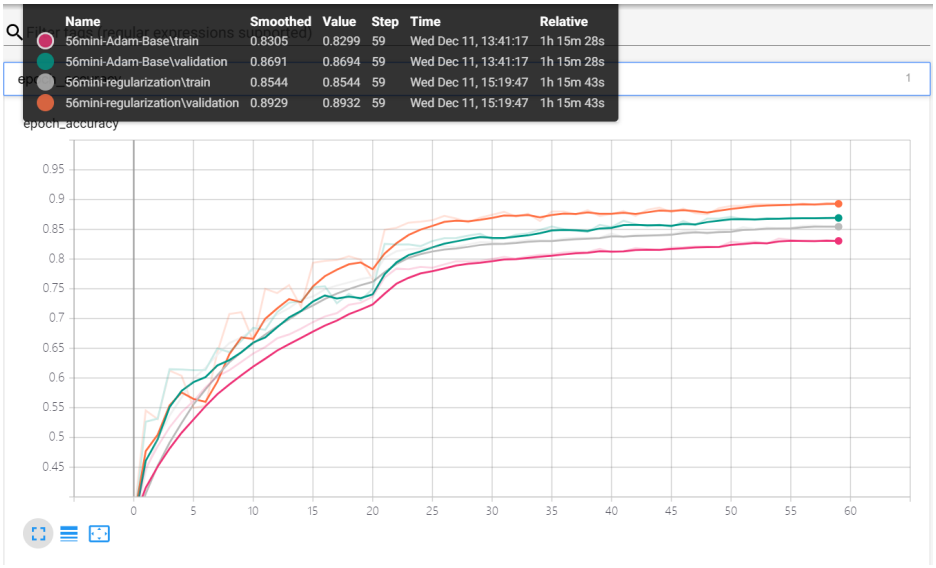


Figure 5. Training accuracy using regularization

Finally, we compare our model with original model from paper. From figure 6, we see that paper model reaches 0.90 on validation accuracy, but spending almost 5 hours. On the other hand, our model only takes 1 hour to early converge at 0.893 on validation accuracy. In short, our new architecture of Residual Attention Network reduces 40% model parameters and saves 75% training time on CIFAR-10 than the original model, but only loss less than 1% on the model accuracy.

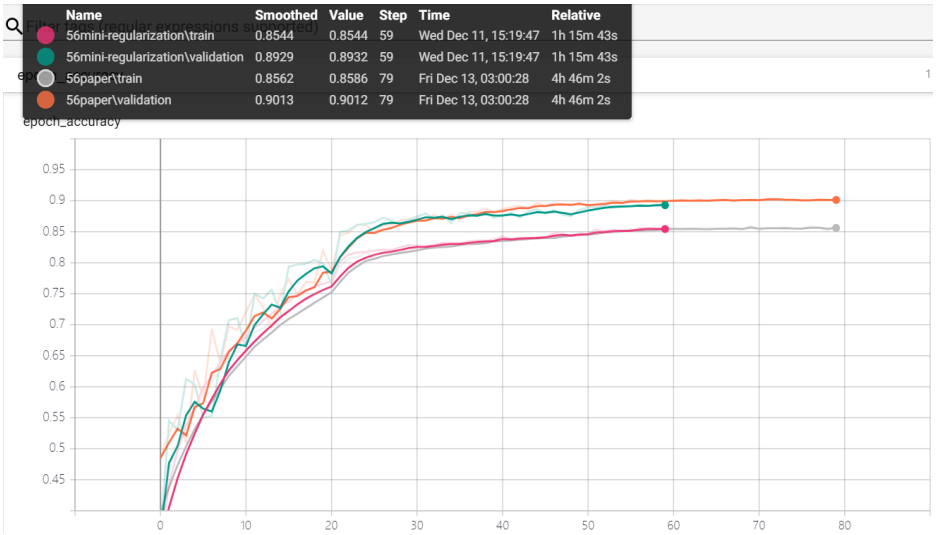


Figure 6. Comparing different model capacity

We visualize the histogram plot for the last layer's weights of Residual Attention Network.

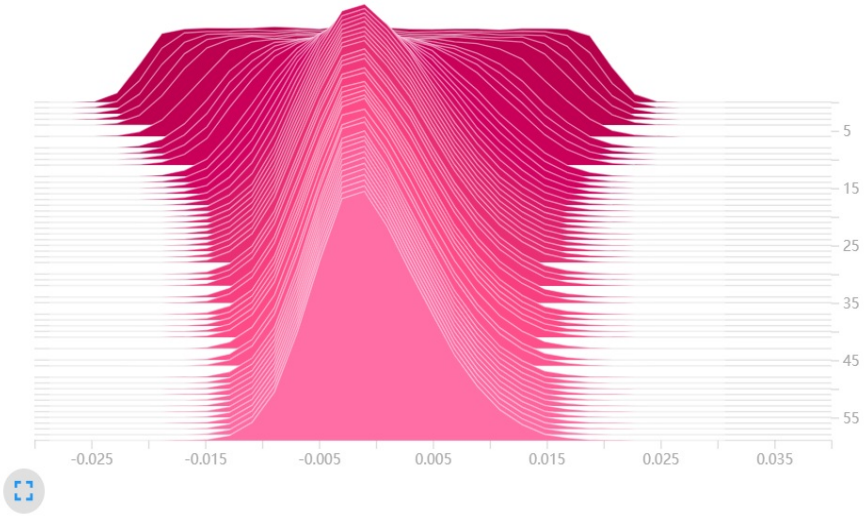


Figure 7. Histogram plot of last layer's weights

4. Discussion

A Residual Attention Network is stacking multiple Attention Modules on the Residual Network (ResNet). The benefits of our network is that capture mixed attention to collect global information from lower resolution.

In our experiments, the simple Attention Module work well in the small image (32x32) classification problem because it quick down-sampling to pixel level and mapping information back to original feature maps. Therefore, it reduces the model size, but not loss many accuracy. On the other hand, if we work on large images (224x224) dataset, stacking multiple Attention Modules and deeper sampling will be a good option to make network complex and powerful, but it requires powerful computation as well.

Also, we believe the Attention module is a extensible mechanism for other convolutional neural network. In this paper, we only apply Attention Module on the ResNet, but stacking on different base models such as VGG, Alexnet, MobileNet and LSTM might solve different tasks: object detection and NLP.

5. Conclusion

To summarize our work, due to recent Residual Attention Network is getting deeper, so it is very high cost on training models. We propose a new architecture of Residual Attention Network for image classification to reduce model size and improve model efficiency without accuracy loss. For training Residual Attention Network, we find that Adam optimizer is better and faster than Nesterov SGD. Rather than stacking multiple Attention Modules, adding batch normalization and dropout layers not only prevent overfitting issue, also reduce model size in order to improve training efficency.

Reference

[1] A.Walia, Types of Optimization Algorithms used in Neural Networks and Ways to Optimize Gradient Descent, 2017.6.10

[2] B.Fitzgerald , Applying attention to the CIFAR-10 dataset, 2017.1.8

[3] D.Kingma, J.Ba, Adam: A Method for Stochastic Optimization, arXiv: 1412.6980, 2017.1.30

[4] D.Vatterott, Attention in a Convolutional Neural Net, 2016.9.20

[5] 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

[6] K.He, X.Zhang, S.Ren and J.Sun, Deep Residual Learning for Image Recoginition,arXiv: 1512.03385, 2015. 12, 10

[7] 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

[8] N.Srivastava, G.Hinton, A.Krizhevsky, I.Sutskever
R.Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research 15 (2014) 1929-1958

[9] S.Sahoo, Residual blocks — Building blocks of ResNet, 2018.11.27

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

[11] V.Bushaev, Adam — latest trends in deep learning optimization, 2018.10.22