Residual Attention Network for Image Classification

ECBM E4040 Deep Learning and Neural Networks Final Project

Yi-Pei Chan
Department of Statistics
Columbia University
New York, NY 10027
yc3700@columbia.edu

***Abstract*—In this work, we summarized the concepts and methods done in the original paper [1] and reviewed related influential literature. We further reproduced the model and evaluated the performance of the model on CIFAR-10 dataset, where we achieved 84.85% of test accuracy. At the end of the paper, we discussed possible reasons for discrepancies in results between our model and those of the original paper. We also provided promising future works that can potentially improve the model performance in the original paper.**

I. Introduction

Deep convolutional neural networks (CNN) have demonstrated a series of breakthroughs in the field of image classifications in recent years, even surpassing human-level accuracy in tasks such as the ImageNet Challenge [2,3]. The state-of-the-art architectures have become popular for their ability to learn feature representations well from raw data, without having to hand-design explicit features [4]. Integrating from low level features such as edges, color, gradient orientation to high level abstract features like complex shapes and objects [5], these neural networks classify in a comprehensive end-to-end fashion. Several works have suggested that the depth of the network is an important factor in achieving good classification performance, as deeper networks can learn more abstract features [6]. While deep networks may achieve lower classification error rates, they are harder to train with increasing depth mainly because of the following two reasons: Vanishing or exploding gradients [7] and harder optimization [8].

One effective solution for the problems of deep CNNs is to use Residual Networks [9]. In addition, residual networks are suggested to be powerful tools for image classification, as demonstrated in ILSVRC 2015 where the models achieved a top-5 error rate of 3.57% [9]. The greatest difference between the aforementioned two architectures is that Residual Networks possess shortcut connections. Unlike normal convolutional layers, these shortcut connections persist throughout the model and serve as a means in generating gradients with back propagation.

As there are abundance of works and researches on image classification, we would review past influential literature in the next section. We selected two major topics - residual networks, and visual attention- that were most relevant to the development of the original paper. The concepts and ideas from the past works were used in building our final model to evaluate results. The rest of the paper is organized as follows. In section 3 we will describe the CIFAR-10 dataset, and the framework we used for our implementations. In Section 4 we will discuss results and show how does our Residual Attention Network compared to its equivalent Convolutional Neural Network. Last, our conclusion will point out possible reasons for discrepancies between our results and those of the authors of the original paper, and we would propose few considerations that must be accounted when employing Residual Attention Network.

II. Related Work

A. Residual Networks

The residual network, also known as ResNet, was introduced by He et al. in 2015 [9]. The primary idea of the work was incorporating the residual connection into convolutional neural networks, which was an idea similar to its predecessor, Highway Network, and Inception with shortcut connections. The residual connections between blocks in the sequential layers provide a clear path for gradients to back propagate to early layers of the network and thus made training of deep networks more feasible by avoiding vanishing gradient problem or dead neurons. In addition, the residual blocks that allow additive interaction between input and output of two convolution layers make the learning process faster. The residual unit originally introduced by the authors in [9] is described as in equation 1.

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{1}$$

In [9], the authors as well adopted batch normalization, which was implemented right after each convolution and before activation, as it is believed to accelerate the training process by reducing internal covariate shift [11]. However, [9] did not use dropout, as they followed the idea that batch normalization regularizes the model and reduces the need for dropout [11]. Later, there are derivations and improved researches based on the idea of [9] with variations of the residual architecture. For example, He et al. in 2016 improved test error on classification on the CIFAR-10 dataset by using identity mappings as the skip connections and after-addition activation.

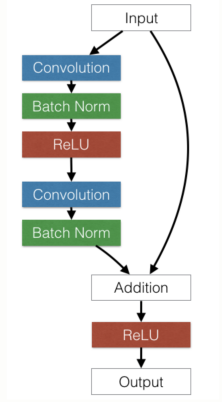

Fig. 1. ResNet Architecture [12]

B. Visual Attention

Attention mechanism was originally introduced with the aim of making neural networks focus on different parts of their input [13]. The derivation of the idea was successfully implemented in Natural Language Processing to do machine translation by Bahdanau et al.[14] at the beginning. For image classification applications, three common approaches: sequential process, region proposal and control gates, were used to implement top-down attention mechanism.

- Sequential process by its name operates in a sequential fashion [15,16,17], allowing end-to-end optimization using recurrent neural networks and long short-term memory that are able to capture different kinds of attention in a goal-driven way[1].
- Region proposal has been used in various image detection researches [18]. Unlike the image detection process, applying region proposal in image classification would require an additional region proposal stage added before feedforward classification, and unsupervised learning is often used to meet such requirement. For example, Xiao et.al in 2014 avoided using expensive annotations like bounding box or part information from end-to-end for fine grained image analysis [19].
- Control gates have been extensively used in long short-term memory to control the flow of information within the decision process. When applying attention in image classification, control gates for neurons use the top information to guide bottom-up feedforward process in updating information during training processes [15, 21].

Recent researches for soft attention have developed to make it possible to be trained end-to-end for convolutional network [22,23], and the bottom-up top-down feedforward structure has been successfully applied to human pose estimation[24]. The authors of the original paper were inspired by the idea and utilized a similar structure in their work. The framework is special in that it can in a single feedforward process mimic not only bottom-up feedforward process in producing low resolution feature maps with strong semantic information but also a top-down attention feedback producing dense features to inference on each pixel [1].

The stacked structure is a common approach of mixed attention mechanism which is believed to capture different types of attention and refine attention for complex images with its incremental nature; however, directly stacking Attention Modules would lead to performance drop. Therefore, the authors proposed attention residual learning mechanism, and stacked multiple Attention Modules in residual network.

III. Methods

A. Dataset

We reproduced the model based on the idea of [1]. Given limited time and computational resources, we were motivated to explore the adaptation and performance of our network on image classification tasks with the CIFAR-10 datasets.

The CIFAR-10 is labeled subsets of the 80 million tiny images dataset, which were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton [25]. It is a widely used dataset, consisting of 60,000 color images of size 32×32 divided into 10 classes, with 50,000 training images and 10,000 test images.

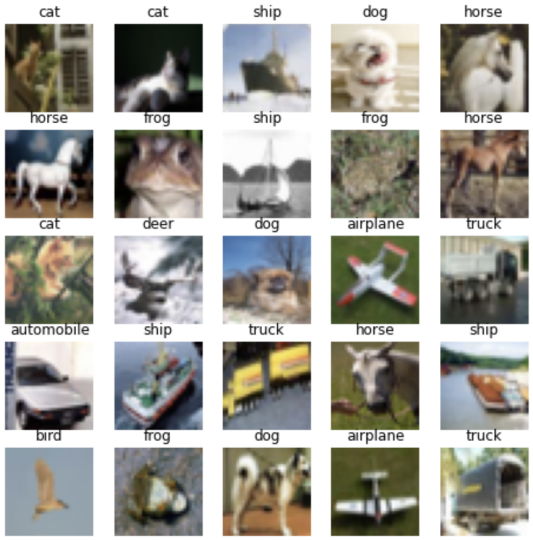

Fig. 2. Images randomly sampled from CIFAR-10

B. Final Model

The architecture of our final model resembles the idea of the original paper. We as well adopted the same hyperparameters as the original work, which are $\{p = 1, t = 2, r = 1\}$ [1]. The respective meanings for the three hyperparameters are as following,

- p: the number of pre-processing Residual Units before splitting into trunk branch and mask branch
- t: the number of Residual Units in trunk branch
- r: the number of Residual Units between adjacent pooling layer in the mask branch

The output loss calculated by the final layer were computed as a softmax function for a cross-entropy loss. The softmax function is defined as follow and it normalizes scores across all classes to sum to 1:

$$S(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{2}$$

The cross-entropy loss is defined as:

$$\text{Loss }_j = -\log\left(\text{Softmax}\left(\mathbf{x}^T\mathbf{w}_j\right)\right) \tag{3}$$

which is feasible for us to backpropagate the loss. The optimizer used in our model to minimize the softmax cross entropy loss was Adam and the learning rate was set to be 0.001.

IV. Results and Discussion

A. Model Results

The training accuracy and test accuracy achieved by our model for classifying images from CIFAR-10 dataset is around 98.40% and 84.85% respectively after 100 epochs. The test accuracy was lower than that of the original model which achieved test error rates as low as 3.9% [1].

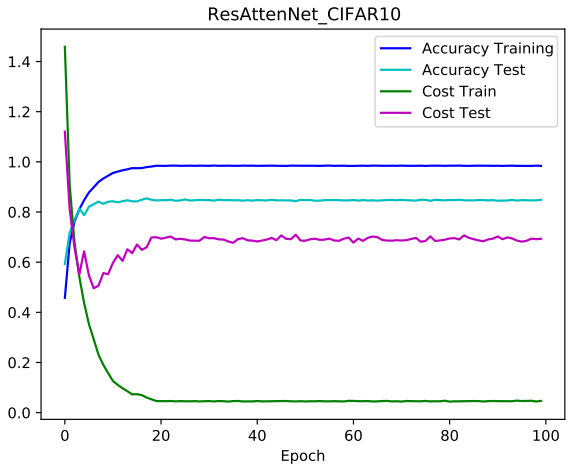

Fig. 3. Accuracies and Costs from our reproduced model

B. Discussion

This work was limited by computation power and time to test for different architectures. There are several further efforts that can be made to improve the results.

- The original paper test several combination of values and achieved the best testing error by using $\{p = 2, t = 4, r = 3\}$. We may try out different combinations to achieve better potential results when the computation resource permits.
- Since each residual block consists of multiple layers of convolutional layer, it is worthwhile testing to know what is the optimal amount of convolutional layers to put whithin one residual block.
- As the architecture is still new, there is no pre-trained model to serve as a starting point. An equally important question may be whether the initial convolution layer is not extracting the features of interest and whether the residual connections actually amplified the negative effects of poor initialization.

There are also some promising ideas to explore based on the paper. From the original work, the authors used maxpooling for down sample and used patching for up sample. The module is illustrated in Fig.4. It is very likely to improve performance by using autoencoder or variational autoencoder which turns the down sampling and up sampling into trainable processes and hopefully a more flexible network. However, this approach may be computational expensive for the increase in the number of tuning parameters.

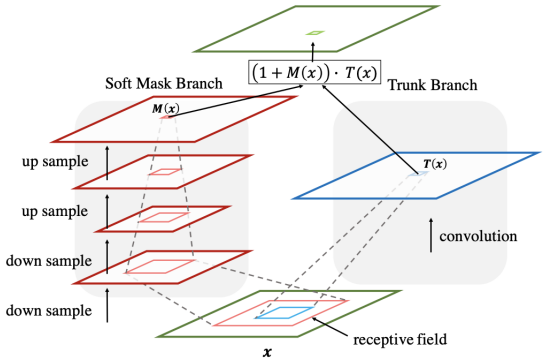

Fig. 4. The bottom-up top-down structure [1]

Another promising future work is to use different residual modules, such as the ResNeXt, which had better performance than the original ResNet and is a homogeneous, multi-branch architecture that requires a few hyperparameters [26].

References

[1] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. CoRR, abs/1704.06904, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. arXiv:1502.01852, 2015.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[4] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1989

[5] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. ECCV, 2014.

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

[7] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In International conference on artificial intelligence and statistics, pages 249–256, 2010.

[8] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015.
[10] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. CoRR, abs/1603.05027, 2016.
[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167, 2015.
[12] http://torch.ch/blog/2016/02/04/resnets.html
[13] A. Graves. Generating sequences with recurrent neural networks. In Arxiv preprint arXiv:1308.0850, 2013.
[14] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
[15] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In NIPS, 2014.
[16] 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.
[17] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In ICML, 2015.
[18] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Simul- ´ taneous detection and segmentation. In ECCV, 2014
[19] 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. CoRR, abs/1411.6447, 2014.
[20] 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.
[21] M. F. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber. Deep networks with internal selective attention through feedback connections. In NIPS, 2014.
[22]] 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.
[23] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In NIPS, 2015.
[24] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. arXiv preprint arXiv:1603.06937, 2016.
[25] Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.
[26] S. Xie, R. B. Girshick, P. Dollar, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. CoRR, abs/1611.05431, 2016.