SURVEY ON DEEP RESIDUAL NETWORK FOR IMAGE RECOGNITION

[1]Kiran Singla, [2]Prakash Mohod

[1]Student, [2]Assistant Professor
[1]Computer Science and Engineering, MTECH
G.H.R.I.E.T., Nagpur, India

Abstract : Deep convolutional neural networks have shown remarkable performance in image classification tasks in recent years. Due to the complexity and vanishing gradient problem, it normally takes a lot of time and more computational power to train deeper neural networks. Deep residual networks were introduced to tackle these problems, making the training process faster and attain more accuracy compared to their equivalent Deep neural networks. Since then more and more residual network variants and architectures have been proposed, and they form a residual networks family together. In this paper, we provide a comprehensive survey of the recent developments on deep residual networks. we summarize different residual network architectures and compare their performances on CIFAR-10 and CIFAR-100 dataset for image classification.

***IndexTerms* - Image classification, Residual network, Highway Network, Residual Networks of Residual Networks; PyramidNet, Pyramidal RoR, stochastic depth, wide residual network.**

I. Introduction

The emergence of deep convolutional neural networks has greatly contributed to advancements in solving complex tasks [6, 4, 5, 7, 10, 14, 15] in computer vision with significantly improved performance. Increasing the network depth is known to improve the model capabilities, which can be seen from AlexNet [6] with 8 layers, VGG [15] with 19 layers, and GoogleNet [14] with 22 layers. However, increasing the depth can be challenging for the learning process because of the vanishing/exploding gradient problem [2], which hamper convergence from the beginning. When deeper networks are able to start converging, a degradation problem has been exposed, with the network depth increasing, accuracy gets saturated and then degrades rapidly. Unexpectedly, such degradation is not caused by overfitting, and adding more layers to a suitably deep model leads to higher training error.

Deep residual networks [1] avoid this problem by using identity skip-connections, which help the gradient to flow back into many layers without vanishing. The identity skip-connections facilitate training of very deep networks up to thousands of layers that helped residual networks win five major image recognitions tasks in ILSVRC 2015 [24] and Microsoft COCO 2015 [16] competitions.

However, an obvious drawback of residual networks is that every percentage of improvement requires significantly increasing the number of layers, which linearly increases the computational and memory costs [1].

Very deep residual models also suffer vanishing gradients and overfitting problems; Thus, the performance of thousand layer ResNets is worse than hundred-layer ResNets. The Identity Mapping ResNets (Pre-ResNets) [12] simplified the residual networks training by BN-ReLU-conv order. Pre- ResNets can alleviate the vanishing gradients problem, so that the performance of thousand-layer Pre-ResNets can be further improved.

The Wide Residual Networks (WRN) [9] treated the vanishing gradients problem by decreasing depth and increasing the width of residual networks. Nevertheless, the exponentially increasing number of parameters brought by broader networks worsens the overfitting problem. As a result, dropout and drop-path methods are usually used to alleviate overfitting, and a method on ResNets Stochastic Depth residual networks (SD) [13], which can improve test accuracy and reduce training time. All kinds of residual networks are based on one basic hypothesis: By using shortcut connections, residual networks perform residual mapping fitted by stacked nonlinear layers, which is easier to be optimized than the original mapping [1].

A study supports that deep residual networks act like ensembles of relatively shallow networks [3]. This is achieved by showing the existence of exponential paths from the output layer to the input layer that gradient information can flow. Also, observations show that removing a layer from a residual network, during the test time, has a modest effect on its performance. Contrary to this, deleting a single layer in plain network architectures such as a VGG-network [15] damages the network by causing additional severe errors. Additionally, it shows that most of the gradient updates during optimization come from ensembles of relatively shallow depth. Moreover, residual networks do not resolve the vanishing gradient problem by preserving the gradient through the entire depth of the network. Instead, they avoid the problem by ensembling exponential networks of different length. This raises the importance of multiplicity that refers to the number of possible paths from the input layer to the output layer [3]. Inspired by these observations, multi-residual networks (Multi-ResNet) [11] which increase the multiplicity of the network, while keeping its depth fixed were proposed. It achieved better performance by increasing the number of residual functions in each residual block. The accuracy of a shallow multiresidual network is similar to a deep 110-layer residual network.

Residual networks of Residual networks (RoR) [8] adds level-wise shortcut connections upon original residual networks to promote the learning capability of residual networks, and achieved state-of-the-art results on CIFAR-10 and CIFAR-100. Instead of sharply increasing the feature map dimension at units that perform downsampling, PyramidNet [19] gradually increase the feature map dimension at all units and has superior generalization ability.

There are many different deep residual network architectures proposed till now, our study summarize these architectures and compare their performances on CIFAR-10 and CIFAR-100 dataset for image classification. The remainder of the paper is organized as follows, Section II summarize different residual network architectures and Section III compare their performances on CIFAR-10 and CIFAR-100 dataset for image-classification, leading to conclusion.

II. FAMILY OF DEEP RESIDUAL NETWORKS

This section summarizes the different architectures modals of deep residual network.

1. HIGHWAY NETWORK

Inspired by Long Short Term Memory recurrent neural networks, highway networks [23] make use of a learned gating mechanism for regulating information flow. Due to this gating mechanism, a neural network can have paths along which information can flow across several layers without attenuation. We call such paths information highways, and such networks highway networks. Highway networks with hundreds of layers can be trained directly using stochastic gradient descent and with a variety of activation functions.

While the traditional plain neural architectures become increasingly difficult to train with increasing network depth (even with variance-preserving initialization), the optimization of highway networks is not hampered even as network depth increases to a hundred layers. It is also worth pointing out that the highway networks always converge significantly faster than the plain ones. It achieved competitive test error of 7.72 and 32.39 on CIFAR-10 and CIFAR-100 dataset respectively.

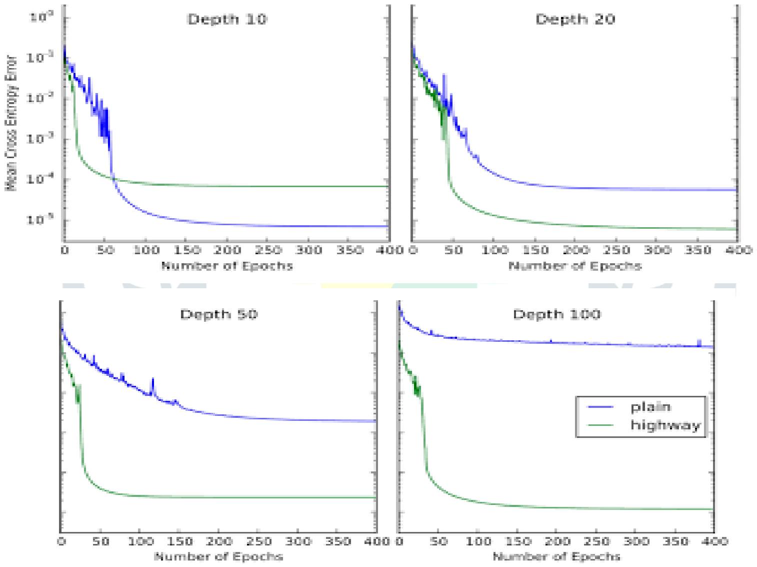


Figure 1. Comparison of optimization of plain networks and highway networks of various depths.

Training both plain networks and highway networks with the same architecture and varying depth, have shown interesting results. While for 10 layers plain network show very good performance, their performance significantly degrades as depth increases. Highway networks on the other hand do not seem to suffer from an increase in depth at all. The result of the 100 layer highway network is about 1 order of magnitude better than the 10 layer one, and is on par with the 10 layer plain network.

2. RESNET

ResNets [1] simplified Highway Networks using a simple skip connection mechanism to propagate information to deeper layers of networks. ResNets are simpler and more effective than highway Networks. Instead of hoping each few stacked layers directly fit a desired underlying mapping, it explicitly let these layers fit a residual mapping. This is done by using feedforward neural networks with “shortcut connections” (Fig 2). Shortcut connections are those skipping one or more layers. Mostly, the shortcut connections simply perform identity mapping, and their outputs are added to the outputs of the stacked layers (Fig. 2). Identity shortcut connections add neither extra parameter nor computational complexity. The entire network can still be trained end-to-end by SGD with backpropagation and can be easily implemented using common libraries without modifying the solvers.

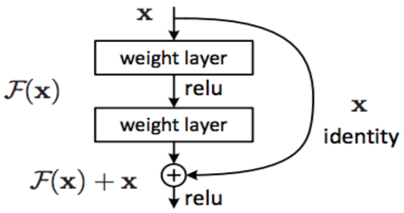

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2. Residual learning: a building block.

Extremely deep residual nets are easy to optimize, but the counterpart "plain" nets (that simply stack layers) exhibit higher training error when the depth increases. The essential difference between residual and highway networks is that in the latter residual links are gated and weights of these gates are learned. It achieved competitive test error of 5.93 and 25.16 on CIFAR-10 and CIFAR-100 dataset respectively using 164 layer ResNet. This showed that, deep residual nets can easily enjoy accuracy gains from greatly increased depth, producing results substantially better than previous networks.

3. RESNET IN RESNET

Resnet in Resnet (RiR) [21] a deep dual stream architecture that generalizes ResNets and standard CNNs and is easily implemented with no computational overhead. Its architecture combines residual networks and standard convolutional networks in parallel residual and non-residual streams. RiR consistently improves performance over ResNets, outperforms architectures with similar amounts of augmentation on CIFAR-10. It achieved competitive test error of 5.01 and 22.90 on CIFAR-10 and CIFAR-100 dataset respectively.

4. Weighted Residual Networks (WResNet)

Inspired by Pre-ResNets, Shen et al. [18] proposed weighted residuals for very deep networks (WResNet), which removed the ReLU from highway and used weighted residual functions to create a direct-path. This method is also capable of 1000+ layers residual networks training and achieves good accuracy. It achieved competitive test error of 4.70 on CIFAR-10 dataset.

5. Convolutional Residual Memory Networks

The Convolutional Residual Memory(CRM) Networks [24] is a memory mechanism enhanced convolutional neural network architecture based on augmenting convolutional residual networks with a long short term memory mechanism. It is capable of selectively identifying features to remember throughout the layers of a CNN. It exploits the well-known property of LSTMs to both allow gradient information to be propagated backwards for many steps and remember features derived from inputs over many processing steps. Also allowing CNNs to be extended with a parallel network taking intermediate representations as input and subjecting them to alternative algorithmic manipulations.

A CRM Network interfaces a residual network and an LSTM. Though it achieved competitive test error of 4.65 and 20.35 on CIFAR-10 and CIFAR-100 dataset respectively, but the number of parameters in this model is too large.

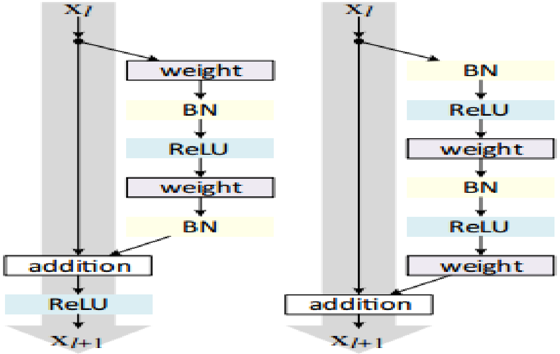
6. Deep Networks with Stochastic Depth

Long training time is a serious concern as networks become very deep. The forward and backward passes scale linearly with the depth of the network. Even on modern computers with multiple state-of-the-art GPUs, architectures like the 152-layer ResNet require several weeks to converge on dataset. The researcher is faced with an inherent dilemma: shorter networks have the advantage that information flows efficiently forward and backward, and can therefore be trained effectively and within a reasonable amount of time. However, they are not expressive enough to represent the complex concepts that are commonplace in computer vision applications. Very deep networks have much greater model complexity, but are very difficult to train in practice and require a lot of time and patience. It proposed a new theory where one would like to have a deep network during testing but a short network during training

Stochastic depth, a training procedure that enables the seemingly contradictory setup to train short networks and use deep networks at test time. It randomly skips layers entirely. It is achieved by introducing skip connections in the same fashion as ResNets, however the connection pattern is randomly altered for each minibatch. It starts with very deep networks but during training, for each mini-batch, it uses Bernoulli random variables to randomly drop a subset of layers and remove their corresponding transformation functions and bypass them with the identity function [13]. It achieved competitive test error rate of 5.23 and 24.58 on CIFAR-10 and CIFAR-100 dataset with 1.7M parameters.

7. Identity Mappings in Deep Residual Networks

[12] proposes a residual unit, which makes training easier and improves generalization.



(a) (b)

Figure 5. Left: (a) original Residual Unit in [1]; (b) Pre-resnet model.

It investigates the propagation formulations behind the connection mechanisms of deep residual networks. It implies that identity shortcut connections and identity after-addition activation are essential for making information propagation smooth. It is called Pre-resnet model. It achieves a classification test error rate of 5.46% on CIFAR-10 dataset and 24.33 on CIFAR-100 using 164-layer model.

8. Deep Residual Networks with Exponential Linear Unit (ELU-ResNets)

[22] proposes the use of exponential linear unit instead of the combination of ReLU and Batch Normalization in Residual Networks. This not only speeds up learning in Residual Networks but also improves the accuracy as the depth increases.

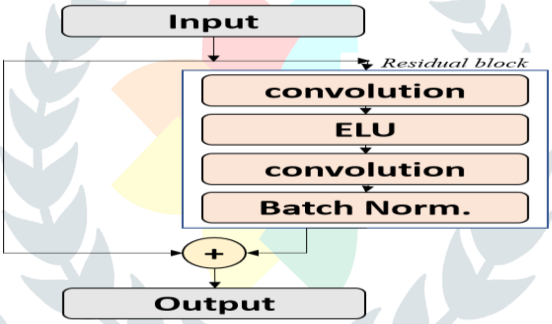


Figure 6. An ith Residual Block with Exponential Linear Unit (ELU) in Residual Networks

The ReLUs are non-negative and thus have mean activations larger than zero, whereas ELUs have negative values, which push the mean activations towards zero. ELUs saturate to a negative value when the input gets smaller. This decreases the forward propagated variation and information, which draws the mean activations to zero. Units with nonzero mean activations act as a bias for the next layer. If these units do not cancel each other out, then the learning causes a bias shift for units in the next layer. Therefore, ELUs decrease the bias shift as the mean activations are closer to zero. Less bias shift also speeds up learning by bringing standard gradient closer towards the unit natural gradient. It achieved competitive test error rate of 5.62 and 26.55 on CIFAR-10 and CIFAR-100 dataset using 110 layer model. Figure 6. Shows an ELU-Residual block.

9. Wide Residual Networks

In [9], Wide Residual Networks (WRN) treated the vanishing gradients problem by decreasing depth and increasing the width of residual networks. Widening consistently improves performance across residual networks of different depth, but increasing both depth and width helps until the number of parameters becomes too high and stronger regularization is needed; there doesn't seem to be a regularization effect from very high depth in residual networks as wide networks with the same number of parameters as thin ones can learn same or better representations. Furthermore, wide networks can successfully learn with a 2 or more times larger number of parameters than thin ones, which would require doubling the depth of thin networks, making them infeasibly expensive to train. WRN-40-4 achieved test error rate of 4.53 on CIFAR-10 dataset.

10. Multi-Residual Networks

Multi-residual networks (Multi-ResNet) [11] increase the multiplicity of the network, while keeping its depth fixed. This is achieved by increasing the number of residual functions in each residual block. Increasing the number of residual functions leads to a better performance than increasing the network depth. This leads to a lower error rate for the multi-residual network with the same number of convolutional layers as the deeper residual network. It supports the hypothesis that deep residual networks behave like ensembles, rather than a single extremely deep network in [3]. Multiresidual networks exploit multiple functions for the residual blocks which leads to networks that are wider, rather than deeper. A shallow multi-residual network of 14 layer depth is able to approximate the accuracy of a 110-layer residual network with 6.42% test error on CIFAR-10 dataset.

A model parallelism technique has been investigated to reduce the computational cost of multi residual networks. By splitting the computation of the multi-residual blocks among processors, the network is able to perform the computation faster.

11. Residual Networks of Residual Networks: Multilevel Residual Networks

In [8] RoR is proposed which adds level-wise shortcut connections upon original residual networks to promote the learning capability of residual networks. To dig the optimization ability of residual networks, RoR substitutes optimizing residual mapping of residual mapping for optimizing original residual mapping.

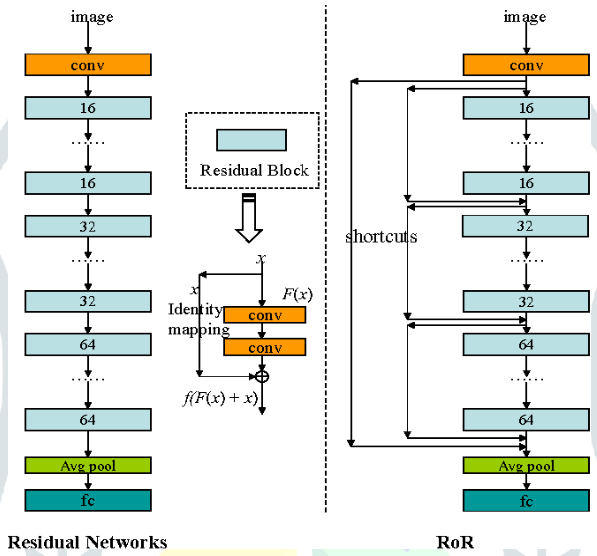


Figure 7. The image on the left is an original residual network, which contains a series of residual blocks, and each residual block has one shortcut connection. The number (16, 32, or 64) on each residual block is the number of output feature map. The image on the right is residual networks of residual networks architecture with three shortcut levels. RoR is constructed by adding identity shortcuts level by level based on original residual networks.

RoR achieved better performance than ResNets by using the same number of layers on different data sets. RoR is not only suitable for original ResNets, but also fits in nicely with other residual networks. Any residual network can be improved by RoR. Hence, RoR has a good prospect of successful application on various image recognition tasks.

12. Deep Pyramidal Residual Networks

Generally, deep neural network architectures are stacks consisting of a large number of convolutional layers, and they perform downsampling along the spatial dimension via pooling to reduce memory usage. Concurrently, the feature map dimension (i.e., the number of channels) is sharply increased at downsampling locations, which is essential to ensure effective performance because it increases the diversity of high-level attributes. This also applies to residual networks and is very closely related to their performance.

In [19] instead of sharply increasing the feature map dimension at units that perform downsampling, it gradually increases the feature map dimension at all units to involve as many locations as possible. This design has proven to be an effective means of improving generalization ability.

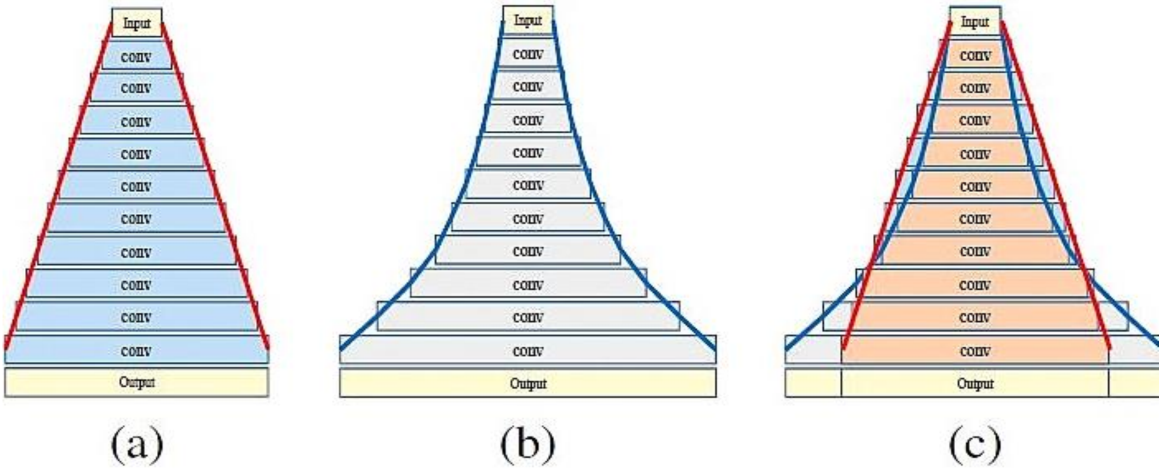


Figure 8. Visual illustrations of (a) additive PyramidNet, (b) multiplicative PyramidNet, and (c) a comparison of (a) and (b).

In two ways pyramidNet can be implemented, additive and multiplicative PyramidNets. The main difference between them is that the feature map dimension of an additive network gradually increases linearly, whereas the dimension of a multiplicative network increases geometrically. That is, the dimension slowly increases in input-side layers and sharply increases in output-side layers.

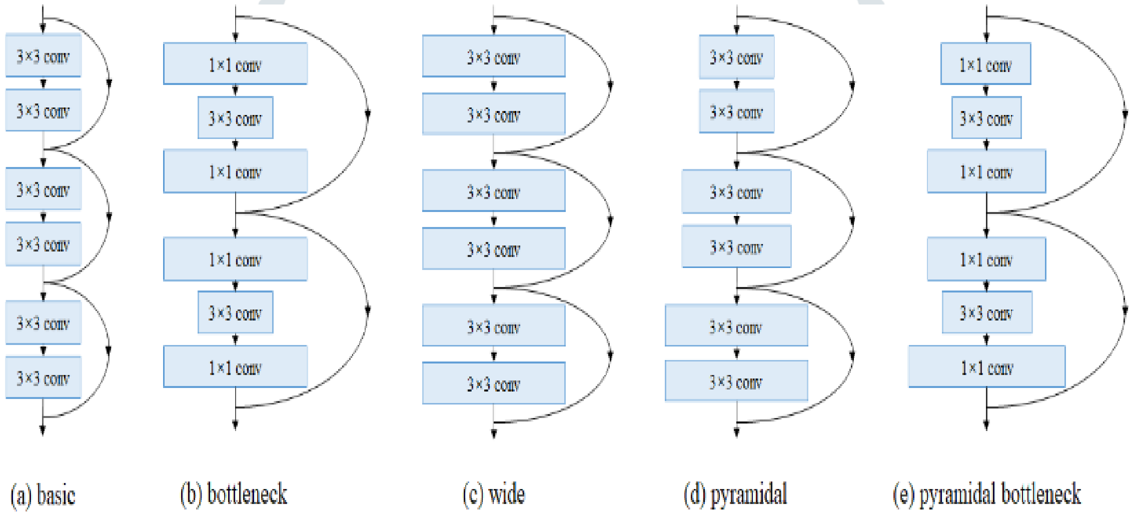


Figure 9. Schematic illustration of (a) basic residual units [1], (b) bottleneck residual units, (c) wide residual units [9], (d) pyramidal residual units [19], and (e) pyramidal bottleneck residual units [19].

13. Pyramidal RoR for Image Classification

Sharply increasing the number of feature map channels in RoR makes the characteristic information transmission in the network incoherent, which losses a certain of information related to classification prediction, and limits the classification performance. To effectively solve the above problem Pyramidal RoR network model is proposed by analyzing the performance characteristics of RoR and combining with the PyramidNet [19].

Based on RoR, the Pyramidal RoR network model [20] with channels gradually increasing is designed. It contains multi-level shortcuts. So that different layers of information can be transmitted to each other, and increasing the feature dimension gradually makes information transmission more smoothly.

Drop-path is used to avoid over-fitting and save training time. Pyramidal RoR can give more control over bias shift and vanishing gradients and get excellent image classification performance.

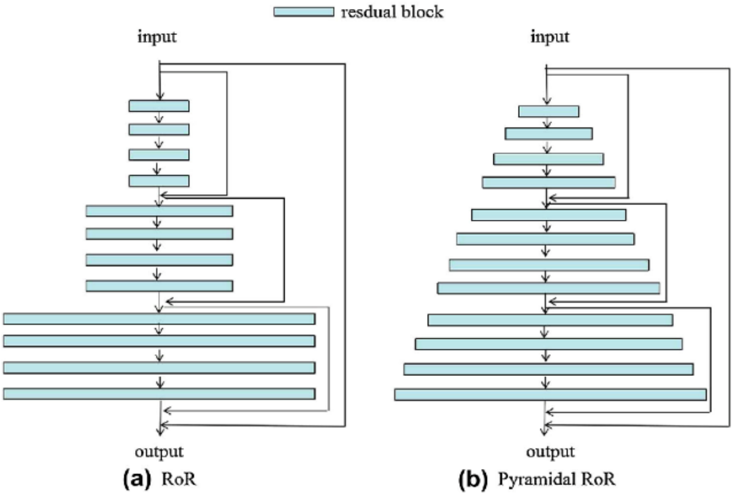


Figure 10. Pyramidal RoR architecture

III. COMPARISION ON CIFAR-10 AND CIFAR-100 DATASET

Table below compares the deep residual network test error rate on CIFAR-10/100 dataset for image classification.

TABLE 1

TEST ERROR RATE (%) COMPARISON OF DEEP RESIDUAL NETWORKS ON CIFAR-10 AND CIFAR-100 DATASET

Model (Parameters)	**CIFAR-10**	**CIFAR-100**
Highway Network	7.72	32.29
ResNet-164(2.5M)	5.93	25.16
Pre-ResNet-164(2.5M)	5.46	24.33
Pre-ResNet-1001(10.2M)	4.62	22.71
ELU-ResNets-110 (1.7M)	5.62	26.55
ResNet-110+SD(1.7M)	5.23	24.58
ResNet in ResNet (10.3M)	5.01	22.90
WResNet-d (19.3M)	4.70	-
WRN-28-10 (36.5M)	4.17	20.50
CRMN(>40M)	4.65	20.35
RoR-110+SD (1.7M)	5.08	23.48
RoR-WRN-56-4（13.3M）	3.77	19.73
multi-resnet（145M）	3.73	19.60
PyramidNet (28.3M)	3.77	18.29
Pyramid RoR+SD (depth=110, α=48) (1.7M)	4.35	21.41
Pyramid RoR+SD (depth=110, α=270) (28.3M)	3.33	16.82

We observe that the test error gradually decreases with each new model. Although multi-resnet, PyramidNet and CRMN achieve competitive test errors, the number of parameters in these models is too large. Residual networks have gradually improved performance with each model.

Stochastic depth method has been implemented with many other residual networks models to decrease computational time and save training time. Pyramidal RoR models with only 1.7M parameters can outperform other deep models having large number of parameters like CRMN, RiR, WRestnet-d.

CONCLUSION

In this paper, we provide a comprehensive survey of different deep residual network architectures and compare their performances on CIFAR-10 and CIFAR-100 dataset for image classification. This survey shows the different residual models that have a good prospect of successful application on various image recognition tasks and computer vision tasks.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition". *arXiv preprint arXiv:1512.03385, 2015*.

[2] Y. Bengio, P. Simard, and P. Frasconi. "Learning long-term dependencies with gradient descent is difficult". *IEEE transactions on neural networks, 5(2):157–166, 1994.*

[3] Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks". *In Advances in Neural Information Processing Systems,* pages 550–558, 2016.

[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,* 1(4):541–551, 1989

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition". *In ICML*, 2014.

[6] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks". *In NIPS*, 2012.

[7] M. D. Zeiler and R. Fergus. "Visualizing and understanding convolutional neural networks". *In ECCV,* 2014.

[8] Ke Zhang, Miao Sun, Tony X. Han, Xingfang Yuan, Liru Guo, Tao Liu; "Residual Networks of Residual Networks: Multilevel Residual Networks". *arXiv:1608.02908v2 [cs.CV]*.

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

[10] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: "Integrated recognition, localization and detection using convolutional networks". *arXiv preprint arXiv:1312.6229*, 2013.

[11] M. Abdi, S. Nahavandi;" Multi-Residual Networks: Improving the Speed and Accuracy of Residual Networks". *arXiv:1609.05672v4*, 2016.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mapping in deep residual networks," *arXiv preprint arXiv:1603.05027*, 2016.

[13] G. Huang, Y. Sun, Z. Liu, and K. Weinberger, "Deep networks with stochastic depth," *arXiv preprint arXiv:1605.09382*, 2016.

[14] Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions". *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pages 1–9, 2015.

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

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," *In European Conference on Computer Vision, pages 740–755. Springer*, 2014.

[17] Chu, D. Yang, R. Tadinada; "Visualizing Residual Networks", *arXiv:1701.02362v1.*

[18] F. Shen, and G. Zeng, "Weighted residuals for very deep networks," *arXiv preprint arXiv:1605.08831*, 2016.

[19] Han D, Kim J, Kim J, "Deep pyramidal residual networks," in *Proc. CVPR.*, 2017.

[20] Ke Zhang, Liru Guo, Ce Gao, Zhenbing Zhao, "Pyramidal RoR for Image Classification", *arXiv:1710.00307 [cs.CV]*.

[21] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[22] Shah, S. Shinde, E. Kadam, and H. Shah, "Deep residual networks with exponential linear unit," *arXiv preprint arXiv:1604.04112*, 2016.

[23] R.K. Srivastava, K. Gre, J. Schmidhuber, "Highway networks," *arXiv:1505.00387v2, 2015.*

[24] Joel Moniz, Christopher Pal, "Convolutional Residual Memory Networks," *arXiv:1606.05262, 2016.*