Residual Networks of Residual Networks: Multilevel Residual Networks

IEEE Transactions on Circuits and Systems for Video Technology, 2017

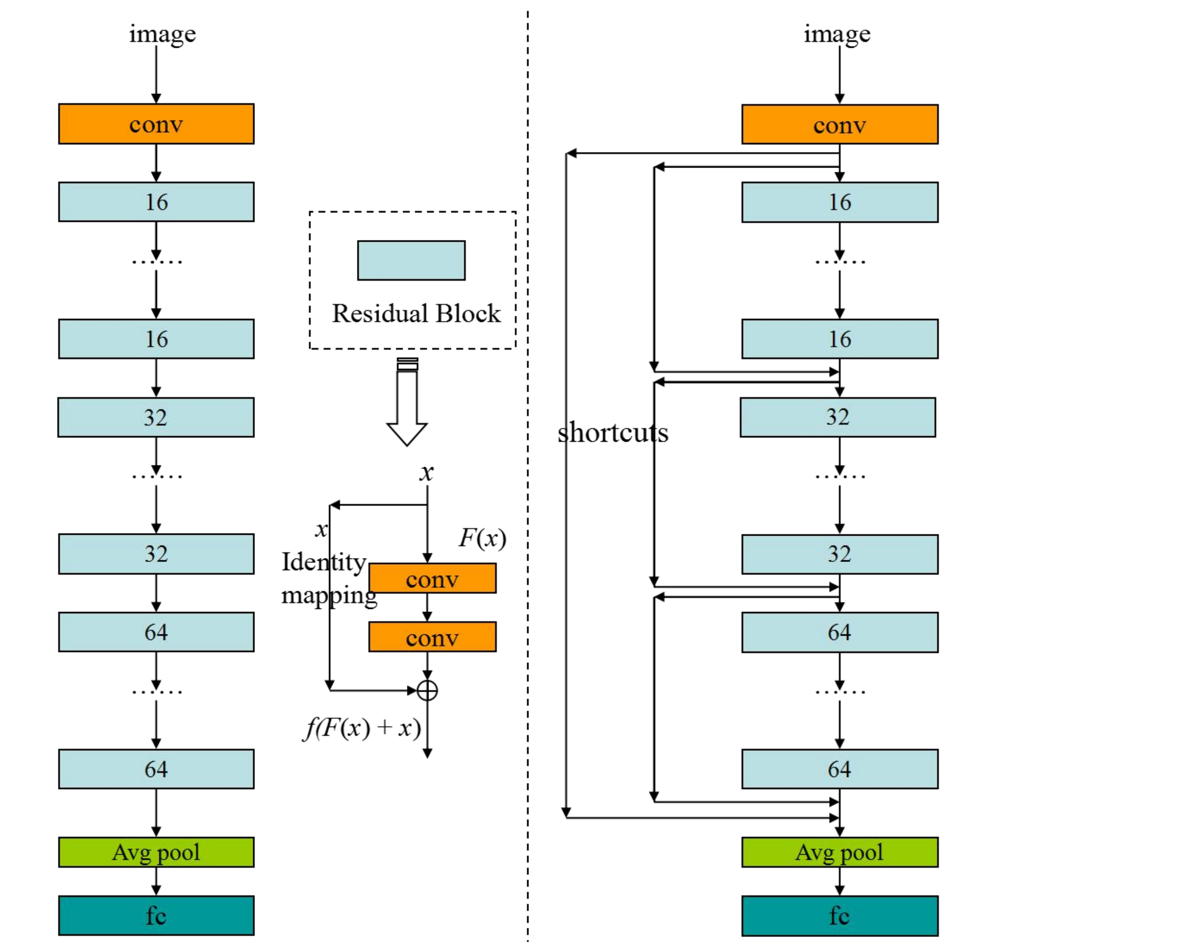
Ke Zhang*（张珂）*, Miao Sun, Tony Han, Xingfang Yuan, Liru Guo, Tao Liu

Department of Electronic and Communication Engineering, North China Electric Power University, Baoding, Hebei, China

ICIP 2017

Introduction

A residual-networks family with hundreds or even thousands of layers dominates major image recognition tasks, but building a network by simply stacking residual blocks inevitably limits its optimization ability. This paper proposes a novel residual-network architecture, Residual networks of Residual networks (RoR), to dig the optimization ability of residual networks. RoR substitutes optimizing residual mapping of residual mapping for optimizing original residual mapping. In particular, RoR adds level-wise shortcut connections upon original residual networks to promote the learning capability of residual networks. More importantly, RoR can be applied to various kinds of residual networks (ResNets, Pre-ResNets and WRN) and significantly boost their performance. Our experiments demonstrate the effectiveness and versatility of RoR, where it achieves the best performance in all residual-network-like structures. Our RoR-3-WRN58-4+SD models achieve new state-of-the-art results on CIFAR-10, CIFAR-100 and SVHN, with test errors 3.77%, 19.73% and 1.59%, respectively. RoR-3 models also achieve state-of-the-art results compared to ResNets on ImageNet data set.

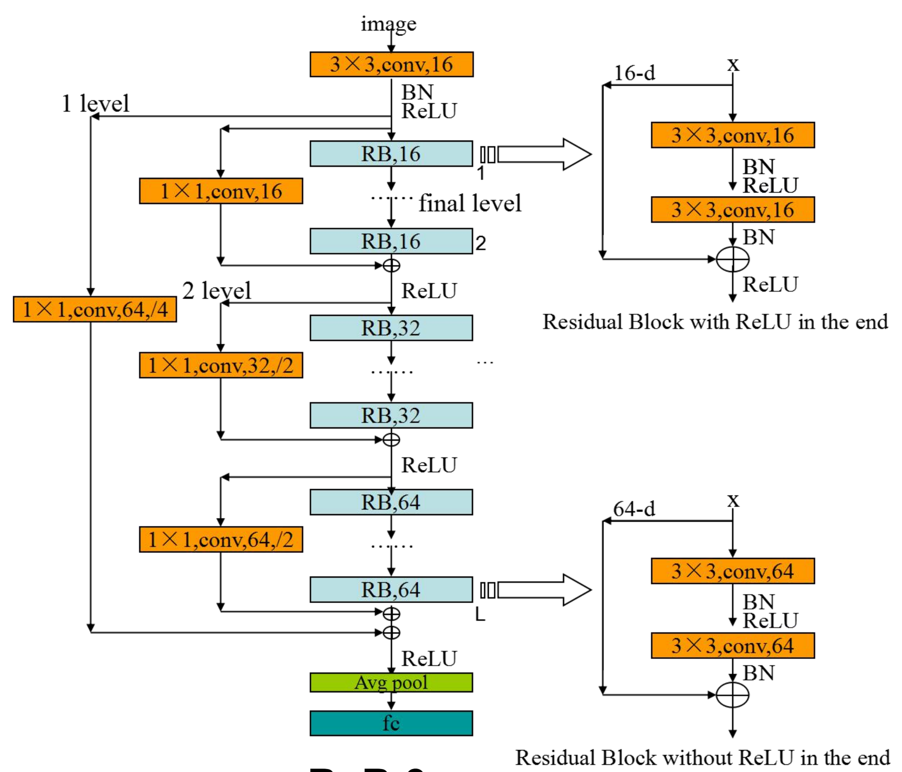


Residual Networks **RoR**

Methods

Architectures of RoR

RoR is based on a hypothesis: To dig the optimization ability of residual networks, we can optimize the residual mapping of residual mapping. So we add shortcuts level by level to construct RoR based on residual networks.



RoR-3

Optimization of RoR

- **Shortcut level number of RoR**

It is important to choose a suitable number of RoR levels for a satisfying performance. The more shortcut levels chosen, the more branches and parameters are added. The overfitting problem will be exacerbated, and the performance may decrease. However, RoR improvements will be less obvious if the number of levels is too small. So we must find a suitable number to keep the balance**. So we chose *m*=3.**

- **Identity Mapping Types of RoR**

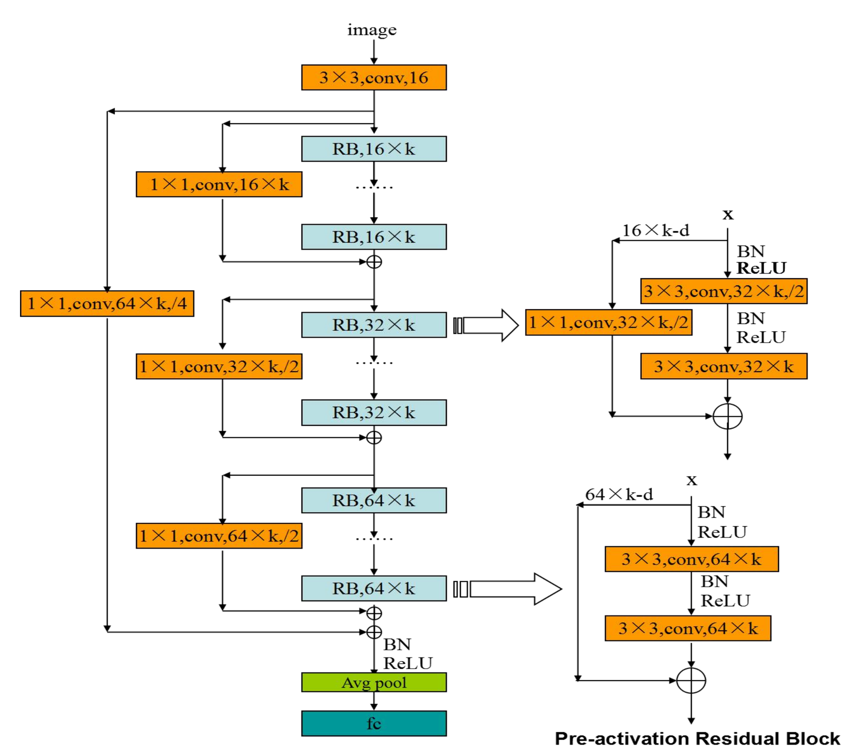
We all used **Type A** in the final shortcut level, and **Type B** in the other shortcut levels.

- **Maximum Epoch Number of RoR**

In this paper, we choose **500** as the maximum epoch number for RoR and Stochastic Depth method.

- **Drop Path by Stochastic Depth**

Overfitting can be a critical problem for the CIFAR-100 data set. Adding extra shortcuts to the original ResNets can cause the overfitting problems to be even more severe. So in this paper we use **the stochastic depth droppath method** in our RoR except for the ImageNet data set, and it can significantly alleviate overfitting, especially on the CIFAR-100 data set.



Pre-RoR-3 or RoR-3-WRN

Results

CIFAR-10 and CIFAR-100 Classification by RoR

CIFAR-10 500 Epoch	ResNets	ResNets+SD	RoR-3	RoR-3+SD
110-layer	5.43	5.63	5.08	5.04
164-layer	5.07	5.06	4.86	4.90

CIFAR-100 500 Epoch	ResNets	ResNets+SD	RoR-2	RoR-2+SD	RoR-3	RoR-3+SD
110-layer	26.80	23.83	27.19	23.60	26.64	23.48
164-layer	25.85	23.29	-	-	27.45	22.47

Versatility of RoR for other residual networks

In this paper, we constructed the RoR architecture based on other two residual networks: Pre-ResNets and WRN.

500 Epoch	Pre-ResNets	Pre-RoR-3	Pre-ResNets+SD	Pre-RoR-3+SD
164-layer CIFAR-10	5.04	5.02	4.67	4.51
164-layer CIFAR-100	25.54	25.33	22.49	21.94

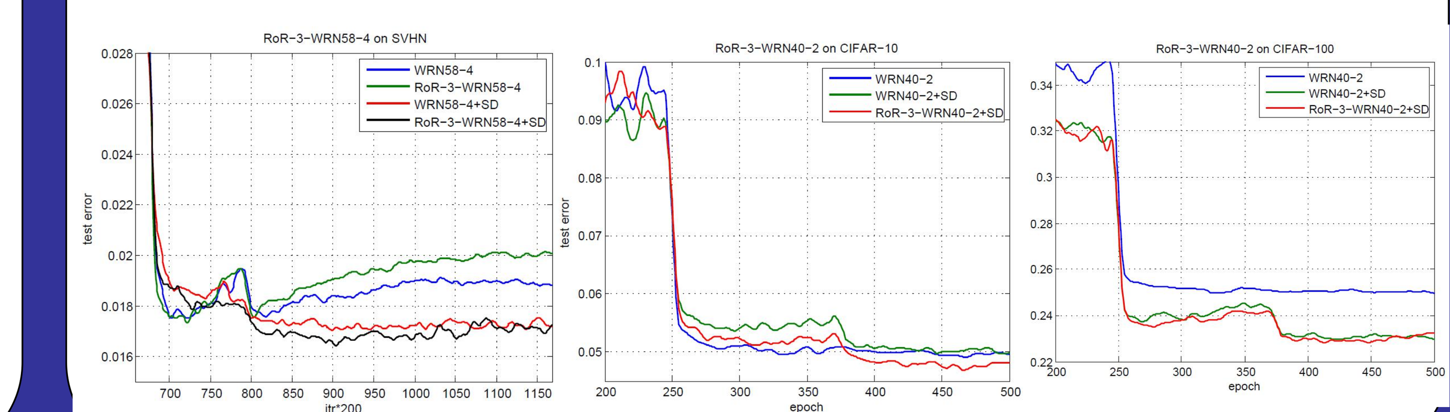
500 Epoch	WRN40-2	RoR-3-WRN40-2	WRN40-2 +SD	RoR-3-WRN40-2+SD
CIFAR-10	4.81	5.01	4.80	4.59
CIFAR-100	24.70	25.19	22.87	22.48

Depth and Width Analysis

The performance can be improved by increasing depth or width.

Depth	CIFAR-10 Pre-RoR-3+SD	CIFAR-100 Pre-RoR-3+SD
110-layer	4.63	23.05
164-layer	4.51	21.94
218-layer	4.51	21.43
1202-layer	4.49	20.64

Depth with Width	CIFAR-10	CIFAR-100
RoR-3-WRN40-2	4.59	22.48
RoR-3-WRN40-4	4.09	22.11
RoR-3-WRN58-2	4.23	21.50
RoR-3-WRN58-4	3.77	19.73



Comparisons with State-of-the-art Results

CIFRA-10, CIFAR-100 and SVHN

Method (#Parameters)	CIFAR-10	CIFAR-100	SVHN
NIN [5]	8.81	35.68	2.35
FitNet [8]	8.39	35.04	2.42
DSN [9]	7.97	34.57	1.92
All-CNN [10]	7.25	33.71	-
Highway [28]	7.72	32.39	-
ELU [22]	6.55	24.28	-
FractalNet (30M) [29]	4.59	22.85	1.87
ResNets-164 (2.5M) [12] [13]	5.93	25.16	-
FitResNet, LSUV [26]	5.84	27.66	-
Pre-ResNets-164 (2.5M) [13]	5.46	24.33	-
Pre-ResNets-1001 (10.2M) [13]	4.62	22.71	-
ELU-ResNets-110 (1.7M) [31]	5.62	26.55	-
PELU-ResNets-110 (1.7M) [24]	5.37	25.04	-
ResNets-110+SD (1.7M) [15]	5.23	24.58	1.75
ResNet in ResNet (10.3M) [30]	5.01	22.90	-
SwapOut (7.4M) [32]	4.76	22.72	-
WResNet-d(19.3M) [33]	4.70	-	-
RoR-3-164 (2.5M)	4.86	22.47(+SD)	-
Pre-RoR-3-164+SD (2.5M)	4.51	21.94	-
RoR-3-WRN40-2+SD (2.2M)	4.59	22.48	-
Pre-RoR-3-1202+SD (19.4M)	4.49	20.64	-
RoR-3-WRN40-4+SD (8.9M)	4.09	20.1 1	-
RoR-3-WRN58-4+SD (13.3M)	**3.77**	**19.73**	**1.59**

ImageNet

Method	Top-1 Error	Top-5 Error
ResNets-18 [38]	28.22	9.42
RoR-3-18	**27.84**	**9.22**
ResNets-34 [12]	24.52	7.46
ResNets-34 [38]	24.76	7.35
RoR-3-34	**24.47**	**7.13**
ResNets-101 [12]	21.75	6.05
ResNets-101 [38]	21.08	5.35
RoR-3-101	**20.89**	**5.24**
ResNets-152 [12]	21.43	5.71
ResNets-152 [38]	20.69	5.21
RoR-3-152	**20.55**	**5.14**

During training on ImageNet, we noticed that RoR is slower than ResNets. **So instead of training RoR from scratch, we used the pretrained ResNets models.** The weights from pretrained ResNets models remained unchanged, but the new added weights were initialized. 10 epochs for fine-tuning RoR. **SD was not used here** because SD made RoR difficult to converge on ImageNet.

Conclusions

This paper proposes a new Residual networks of Residual networks architecture (RoR), which was proved capable of obtaining a new state-of-the-art performance on CIFAR-10, CIFAR-100, SVHN and ImageNet for image classification. Through empirical studies, this work not only significantly advanced the image classification performance, but can also provided an effective complement to the residual-networks family in the future. In other words, any residual network can be improved by RoR. Hence, RoR has a good prospect of successful application on various image recognition tasks.