Yasrab R. SRNET: A shallow skip connection based convolutional neural network design for resolving singularities. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 34(4): 924–938 July 2019. DOI 10.1007/s11390-019-1950-8

SRNET: A Shallow Skip Connection Based Convolutional Neural Network Design for Resolving Singularities

Robail Yasrab

Computer Vision Laboratory, School of Computer Science, University of Nottingham, Nottingham, NG8-1BB, U.K.

E-mail: robail.yasrab@nottingham.ac.uk

Received June 12, 2018; revised May 24, 2019.

Abstract Convolutional neural networks (CNNs) have shown tremendous progress and performance in recent years. Since emergence, CNNs have exhibited excellent performance in most of classification and segmentation tasks. Currently, the CNN family includes various architectures that dominate major vision-based recognition tasks. However, building a neural network (NN) by simply stacking convolution blocks inevitably limits its optimization ability and introduces overfitting and vanishing gradient problems. One of the key reasons for the aforementioned issues is network singularities, which have lately caused degenerating manifolds in the loss landscape. This situation leads to a slow learning process and lower performance. In this scenario, the skip connections turned out to be an essential unit of the CNN design to mitigate network singularities. The proposed idea of this research is to introduce skip connections in NN architecture to augment the information flow, mitigate singularities and improve performance. This research experimented with different levels of skip connections and proposed the placement strategy of these links for any CNN. To prove the proposed hypothesis, we designed an experimental CNN architecture, named as Shallow Wide ResNet or SRNet, as it uses wide residual network as a base network design. We have performed numerous experiments to assess the validity of the proposed idea. CIFAR-10 and CIFAR-100, two well-known datasets are used for training and testing CNNs. The final empirical results have shown a great many of promising outcomes in terms of performance, efficiency and reduction in network singularities issues.

Keywords convolutional neural network (CNN), wide residual network (WRN), dropout, skip connection, deep neural network (DNN)

1 Introduction

Convolutional neural networks (CNNs) offer the state-of-the-art performance in many areas of computer vision. Since the groundbreaking success of AlexNet[1] at the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC 2012)[2], very deep CNNs[1−10] have been designed to achieve better performance and improved results. Recently, very deep CNNs have become the de-facto standard in attaining better performance and results. However, with the exponential increase in the number of layers and the depth of these networks, several issues emerge, such as vanishing gradients and degradation. A number of methods were suggested to facilitate the deeper CNN training, including better optimizers[11], well-designed initialization strategies[12,13], skip connections[14,15], layer-wise training[16], and knowledge transfer[17,18]. Network singularity is one of the common issues in deep neural networks, which hinder the network to converge to its optimal accuracy. The three common singularities that occur in deep neural networks (DNNs) are known as elimination, overlap, and linear dependence singularities. There are different reasons behind the occurrence of these singularities. For example, a node elimination in DNN may lead to elimination singularities. Overlap singularities occur due to the collapse of nodes into each other[19], whereas linear dependence singularities arise due to linearly dependent nodes. Fig.1 shows the three types of singularities and critical reasons for their occurrence. These singularities influence the network performance in different ways. According to [20], the elimination and overlap singularities considerably slow down the network learning process, especially in shallow CNNs. Saxe *et al.*[21] stated that linear dependence sin-

Regular Paper

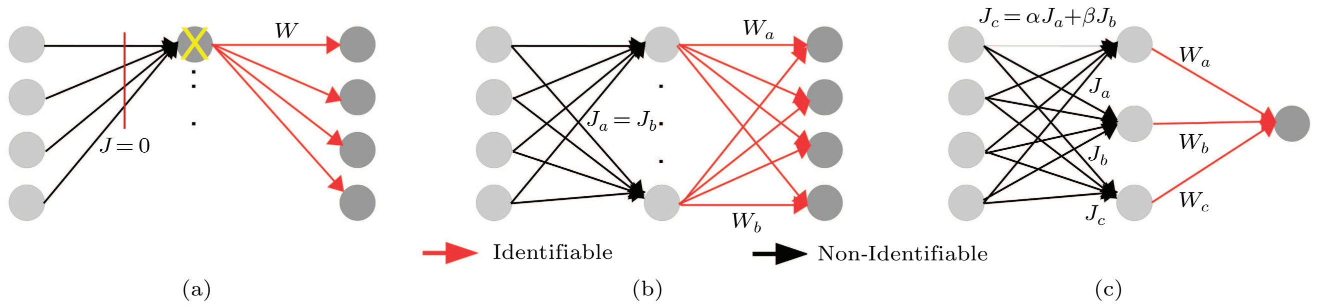

Fig.1. Pictorial depiction of (a) elimination, (b) overlap singularities and (c) linear dependence singularities.

gularities arose commonly in randomly initialized deep linear NN design and turned out to be more severe with an increase in network depth. The elimination and overlap singularities are related to non-linear NN, whereas linear dependence singularities arise only in linear NNs[19]. However, non-identifiability is one of the common aspects of all these singularities. According to [22], the "Hessian Loss" function turns out to be singular at these singularities; these are also known as higher-order saddles or degenerates.

Additional information flow inside the network is required to resolve the singularity-related issues. ResNet[23] has practically demonstrated that adding additional information flow to network architecture can offer better performance and mitigate the depth-related issues. According to Orhan and Pitkow[22], skip connections are described as extra links among different layers, carrying the information from top layers to bottom layers. Skip connections have offered a great deal of improvement in training very deep NNs[24−26] since their introduction. Besides offering diverse and additional information to final network layers for better classification, skip layers also provide a unique contribution to eliminate singularities. Residual networks of residual networks (RoR)[27,28] are another variant of ResNet which tries to resolve the diminishing feature problem. It adds level-wise shortcut connections to the original residual network, which improves its learning capacity. As a result, it exhibits excellent performance in the image classification task as compared with original ResNet architecture, although it lacks a diversity of visual features that are essential for final detection tasks.

To mitigate network singularities, this research proposed a network design that is intended to provide a novel network communication flow and offers a diversity of features for the final assessment task. The proposed network architecture uses a different method to transfer valuable details from top to bottom layers. The skip connections design and placement resolve the three types of singularities. These additional and diverse communication links are based on well-assessed feature flow architecture, trained and tested for performance improvement. We propose a general three-level skip connection architecture that can be replicated to any CNN architecture by position skip connections of three different levels of architecture. The proposed hypothesis is tested initially with ResNet that resulted in an improved performance with fewer network parameters when compared with early benchmark approaches. It also provides a shallow and wide CNN design that avoids feature diminishing problems. Experiments were performed on popular benchmark datasets: CIFAR-10 and CIFAR-100, and results were compared with those of similar approaches to evaluate the proposed method.

The rest of the paper is arranged as follows. Section 2 is about the literature review of early approaches that tried to improve network performance with different network designs. The proposed network design and methodology are presented in Section 3. Section 4 presents the optimization of the proposed network and Section 5 is about the experiment and analysis of results. The final section of this paper is about the conclusions.

2 Related Work

In the past few years CNNs of varying depth have been developed and it has been shown that deeper architectures in general offer a better performance. From early AlexNet (5-conv+3-fc)[1] to the VGG networks (16-conv+3-fc)[7] and GoogleNet (21-conv+1-fc)[29], both the depth and the accuracy kept growing. However, very deep CNNs face the critical issues of degradation and vanishing gradients[30]. In 2015 ResNets[23] came out as a savior to mitigate the degradation problem and achieved excellent results in a variety of classification jobs. Since then, many optimized models based on ResNets have been proposed, and those became part of the evolving residual-networks family. ResNet was a follow-up of earlier highway networks[26], and got a huge success by winning ImageNet and COCO 2015 competition and attaining state-of-the-art performance benchmarks. He *et al.*[25] proposed another innovative model named as Pre-ResNet that formulates a direct path for propagating the information through the whole network, which offers easy training and enhanced generalization. Later, Shen *et al.*[31] proposed a Pre-ResNet inspired model that involves weighted residuals for very deep networks (WResNet). It also removes ReLU from the highway and employs weighted residual functions to develop a direct-path. This mythology is also able to train more than 1 000 layers of residual networks and achieves good accuracy.

The discussion on shallow vs deep networks has been ongoing for a long time in machine learning field[32,33]. That debate outlines that comparatively shallow networks require exponentially more components as compared with deeper networks. The residual networks researchers tried to formulate a thin-architecture in order to increase its depth and have fewer parameters. Afterwards, introducing a bottleneck block leads to thinner ResNet blocks. [34] assesses that the residual block with identity mapping architecture leads to training a very deep network, although it weakens residual networks, as it undermines the gradient flows through the network. This whole situation leads to avoiding the network from learning anything during training. Ultimately, it is possible that only a few blocks are learning some useful representations, or many blocks are sharing very few details with a small contribution to the final goal. This problem is known as diminishing feature reuse outlined by [26]. Zagoruyko and Komodakis[34] also tried to resolve the issues by building wide residual networks. As a result, it proved that widening of ResNet blocks offers a great deal of efficiency regarding enhancing the performance of ResNet besides increasing their depth. It has also shown that wider residual networks provide a significant improvement over [25], with 50 times fewer layers and twice faster. Later, Huang *et al.*[24] attempted to resolve similar diminishing feature problems by randomly disabling connections amongst residual blocks during training. The method randomly drops a subset of layers and bypasses them using an identity mapping intended for every mini-batch. This technique could be taken as a particular case of dropout[35], where every residual block has an identity scalar weight on which the dropout method is applied. A popular CNN design DenseNet[36] makes use of a densely connected path to concatenate the input features with the output features, allowing every micro-block to obtain the raw information from entire previous micro-blocks.

The proposed idea of skip connections worked well for ResNet and was later introduced in most NN architectures; however, the most appropriate explanation of skip connections is offered by Orhan and Pitkow[22]. Skip connections turned out to be the solution for overlap, elimination and linear singularities. All these singularities cause features degradation and loss of information. Though the study of [19] provides an excellent mathematical model for orthogonal and other skip connectivity, it lacks in terms of skips connection placement and different levels of connectivity. There are many earlier studies performed to assess the influence of singularities on networks[19,20]. These studies have evaluated the impact of elimination and overlap singularities on gradient-based learning. Saxe *et al.*[21] studied the linear dependence singularity problem and assessed that the real reason behind such a type of singularities is an increase in depth.

Zhang *et al.*[27] used a similar idea of additional connections to network design. All kinds of residual networks are based on one primary hypothesis: by using the shortcut connections, residual networks perform residual mapping fitted by stacked nonlinear layers; as a result, it becomes easier to optimize the network than the original mapping[23]. Another popular skip connection based architecture named as Residual Networks of Residual Networks (RoR)[27] adds level-wise shortcut connections upon original residual networks to offer better learning capability. PyramidNet[37] offers an innovative idea that instead of sharply increasing the feature maps dimension at units, it gradually increases the feature map dimensions. This network also offers superior generalization ability.

3 Methodology

This section outlines the detailed overview of the architecture of the proposed WRN-ROR, together with types of proposed residual blocks. Experiments are performed to support and confirm the proposed network architectures that further improve the performance of the existing CNN networks.

3.1 Resolving Singularities

As mentioned earlier, there are three different kinds of singularities: elimination, overlap and linear-dependence singularities[22]. The primary reason for the occurrence of the first kind of elimination singularities is the obliteration of a hidden unit, e.g., when the unit's incoming (or outgoing) weights turn out to be zero. As a result, the outgoing (or incoming) links of the network unit are non-identifiable. The overlap singularities happened due to the permutation symmetry of the NNs-hidden units. It occurs at a particular network layer where two units become identical as a result. In this situation, the outgoing links of the network block are no longer recognizable individually. Linear dependence singularities happen due to the linear dependency of the hidden units of a network. Therefore, the outgoing links of these units are no longer individually identifiable.

How do additional communication-links/skip-connections help to eliminate these singularities and other depth-related issues? According to Orhan and Piktow[22] adding skip connections between end-to-end network blocks helps to eliminate singularities. Skip connection offers the capability to keep these units active even when their adjustable incoming or outgoing links turned out to be zero. Additional links help to eradicate the overlap singularities via breaking the permutation symmetry of the neural network (NN) hidden units at a particular layer. Therefore, in case of identical unit weights, the units do not break down, as their discrete skip connections still disambiguate them. Skip connections also help to eliminate the linear dependence, through accumulating linearly independent inputs to the network units. Fig.2 shows that additional skip connections are helpful to resolve all three types of singularities.

Inspired by methodologies mentioned above, the proposed network design makes use of different strategies to resolve many singularities and other depth-related issues for the CNN training. In this scenario, the novel design of blocks and the additional skip connection among layers will offer a great deal of support for better results. Section 3 and Section 4 will explain these aspects in a more detailed way.

3.2 Network Architecture

A normal identity mapping based residual block can be shown as:

$$x_i = \mathcal{F}(x_i, W_i),$$
$$x_{i+1} = x_i + \mathcal{F}(x_i, W_i), \tag{1}$$

where x_i is the input and x_{i+1} is the output of the network. $\mathcal{F}$ denotes the residual function and W_i are block parameters. ResNet is composed of a sequence of residual blocks stacked on each other. Normally, ResNet is built on two kinds of residual blocks, namely, 1) two (3×3) consecutive CONV layers with ReLU and BN units, and 2) a well-known ResNet bottleneck architecture, where one (3×3) CONV layer is sandwich among two (1×1) CONV layers (bottleneck layers). The later architecture with a $(1 \times 1)(3 \times 3)(1 \times 1)$ block structure is proved to be more efficient[25]; therefore this architecture is adopted for our further experiments.

According to [34], there are three different ways to augment the power of residual blocks, 1) more CONV layers per block, 2) more feature planes per block, and 3) bigger filter sizes. First two ideas are enticing, as adding extra layers and filters leads to greater performance boost[38]. However, some researchers[7,39] proved that a smaller filter size (3×3) is more useful as compared with larger ones. Therefore, the large size filters are not used in the proposed network. We prefer the

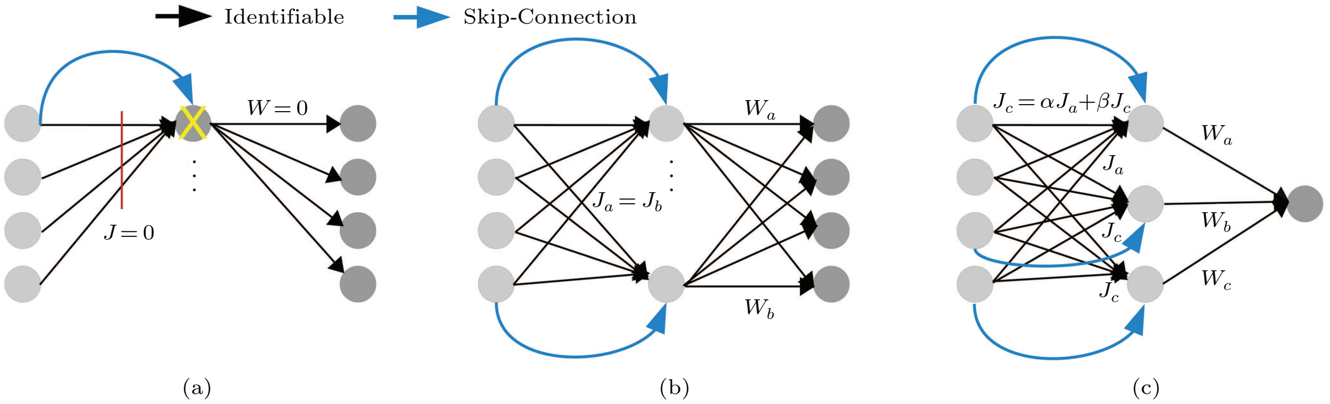

Fig. 2. Resolving three different types of singularities by skip connections. (a) Elimination. (b) Overlap singularities. (c) Linear dependence singularities.

wide network design proposed by Zagoruyko *et al.*[34] The proposed idea was to make use of wide CONV layers in comparison to a deeper network. The resultant network was wider and shallower having an improved performance. Therefore, there is no need to stack hundreds of ResNet blocks to get top results anymore. Table 1 shows the wide residual network (WRN)[34] architecture, where each ResNet block is repeated "N" times. The idea is to get a wider network, a deep CNN design. According to most recent results, WRN offers state-of-the-art performances; therefore, the proposed network will employ WRN as base architecture.

Table 1. Structure of the Resent Blocks

ResNet-Block	Output-Size	Block Type=RB(1,3,1)
Conv-1	32×32	$\|3 \times 3, 16\|$
Conv-2	32×32	$\|1 \times 1, 16 \times k\|$
		$\|3 \times 3, 16 \times k\| \times N$
		$\|1 \times 1, 16 \times k\|$
Conv-3	16×16	$\|1 \times 1, 32 \times k\|$
		$\|3 \times 3, 32 \times k\| \times N$
		$\|1 \times 1, 32 \times k\|$
Conv-4	8×8	$\|1 \times 1, 64 \times k\|$
		$\|3 \times 3, 64 \times k\| \times N$
		$\|1 \times 1, 64 \times k\|$
Average-pool	1×1	$\|8 \times 8\|$

Note: N is the times a unit can be replicated and can be set at the start of network training.

To overcome the degradation problem in residual network, we have offered a solution with residual connections among entire ResNet blocks. The idea is inspired by [27], which is formulated to further enhance the performance of residual networks by adding level-wise shortcut connections among the original ResNet blocks. That leads to improve the capability of the ResNet for different classification tasks. According to Zhang *et al.*[27], "if the residual mapping is easier to learn, the residual mapping of the residual mapping should be easier to learn". The proposed architecture is shown in Fig.3, where multi-level shortcut connections are added to the ResNet blocks and different levels of the network. The figure shows different levels of skip connections from top to bottom of the network. It shows that the network is interlined with a plenty of communication links connecting the whole architecture top-to-bottom. In this image, we can see different connections with different colour ranges. Each type of connections is designed and applied for some specific flow of information. Each skip connection is based on offering a unique type of information to the specified level of the network to help the system to produce efficient results. The details of connections mentioned above are presented in Section 3. The "level 1" shortcuts are offering block-to-block connection. The "level 2" connections are connecting each level of WRN-ResNet, while the "level 3" connections are linking the top to the bottom levels of the proposed architecture. Through adding different levels of shortcut connections, the higher level residual blocks can transfer information to the underlying residual blocks. This kind of setup plays a significant role in suppressing the vanishing gradient.

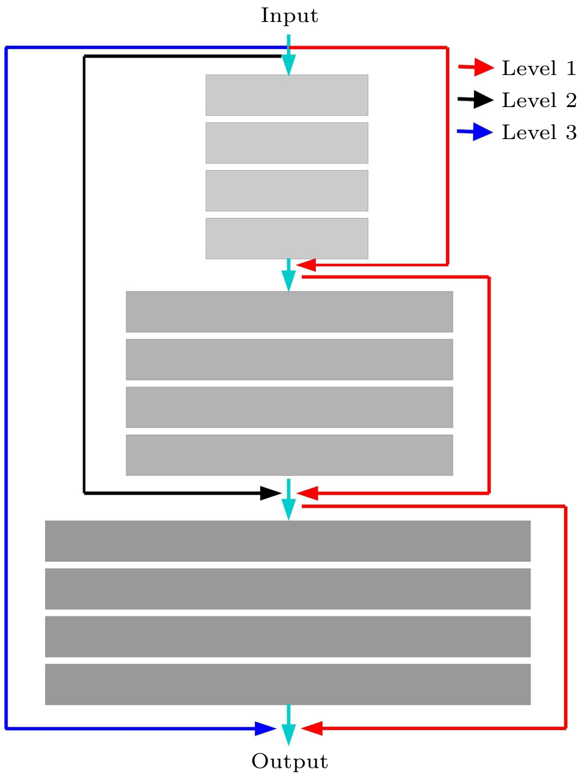

Fig.3. Basic architecture of network.

3.3 Building Block

This subsection describes the structures of different residual blocks and outlines the most suitable for SR-NET classification network. To improve the network performance and capability, there is a need to ameliorate the ResNet basic building block. Since the introduction of residual networks, there are many transformations that have occurred in the fundamental architecture, for example, the pre-activation ResNets[25] which tried to resolve the backward gradient flow problem. However, there is still some room for improvements. In this scenario, a number of experiments are performed on different building blocks to empirically assess the performance. Meanwhile, different block types, activation functions, and batch normalization (BN)[40] layers are assessed to find the most optimal combination.

To define convolutions in a residual block, let RB(K) denote the residual block model, where K denotes the kernel size of the conv units in the model. For instance, RB(3, 3) means a ResNet block with two conv layers with a filter size 3×3. According to Network-in-Network (NiN) architecture, proposed by Lin *et al.*[5], the 1×1 conv is less expensive as compared with the 3×3 conv block. Inspired by this idea the proposed ResNet block is a NiN[5] style block with 3×3 and 1×1 conv layers that bind together in a set. The proposed block architecture will be applied in the WRN mode, as WRN is similar to ResNet, which is composed of ResNet blocks (Conv, Relu, BN layers and Shortcut connections) as a core structure of networks. The residual block is the core of the proposed WRN-ROR architecture which directly determines the performance of the classification network. Table 1 shows the proposed blocks architecture for classification task experiments. The proposed block and connection architecture could be applied to any ResNet family architecture. The flexibility of design enables wide and easy applications to any NN architecture. Furthermore, experiments are performed to assess the performance empirically; details are presented in Section 5.

3.4 Shortcut Connection

According to Zhang *et al.*[27] "by digging the optimization ability of ResNet, the residual mapping of the residual network can be optimized". In this research, the author suggests that the optimization ability of ResNet could be enhanced by adding extra shortcut connection among different levels and blocks of the residual network. Zhang *et al.*[27] stated that additional connections transfer the neural network learning problem to that learning the residual mapping of residual mapping, which is simpler and easier to learn than the original ResNet. Another critical reason for incorporating additional communication links among residual units, blocks, and levels is to create several direct paths for information propagation. Due to this additional information flow among residual architectures, the classification performance will be ultimately improved. In this way, the upper block of the residual network will propagate information to lower blocks. By channelling additional information flow into the network from top to down, we can resolve the vanishing gradient problem. (2) is an extension of (1), and here it generalizes the hyper-residual skip connection architecture.

$$\begin{cases} x_1 = \mathcal{F}(x_0, W_0), \\ x_2 = \mathcal{F}(x_1, W_1) + x_1, \\ x_{i+1} = \mathcal{F}(x_i, W_i) + x_i + \\ \qquad \dfrac{1}{i-1} \sum\limits_{l+1}^{i-1} R_l x_l . i = 2, ..., I-1, \end{cases} \tag{2}$$

where x_1 shows fully connected feed forward NN network architecture with no skip connections. In the next step, we introduce identity skip connection to x_2, among adjacent layers of NN architecture. There is one thing to be noticed that there is no skip connection from the input layer. Later, x_{i+1} shows an architecture of hyper-residual skip connection architecture, which allows communication flow among all layers of the network. Additional communication connection adds few percents to the final accuracy of the network. However, which shortcut connection is the most appropriate? Which one does not add additional load to the network? And which one offers better efficiency? In this scenario, ResNets and Pre-activation ResNets[23,25] were assessed with a number of shortcuts, e.g., projection shortcut, identity-mapping shortcut. The experimental results shown by He *et al.*[25] outlined that the identity-mapping shortcut is the most appropriate shortcut connections. The key reason is that identity-mapping does not involve any parameters, which leads to lower computational requirements and also reduces the chances of overfitting in contrast with other types of shortcuts. Moreover, it offers the better capability to pass the gradient through, which ultimately stabilizes the training phase. In the proposed SRNet, only identity mapping is not employed, as the dimensions of the features differ among individual residual units as well as different levels. Therefore, we use a mixed approach for residual blocks and level-wise blocks. Using only the projection shortcut or a zero-padded shortcut for all the residual units blocks is not a suitable solution, as [25] outlined that a projection shortcut can hamper information propagation and ultimately leads to some optimization problems, particularly in case of very deep architecture. According to Han *et al.*[37] the zero-padded shortcut does not lead to the overfitting problem as no extra trainable parameters exist. Fig.4 shows the different types and levels of shortcut connections in the proposed CNN.

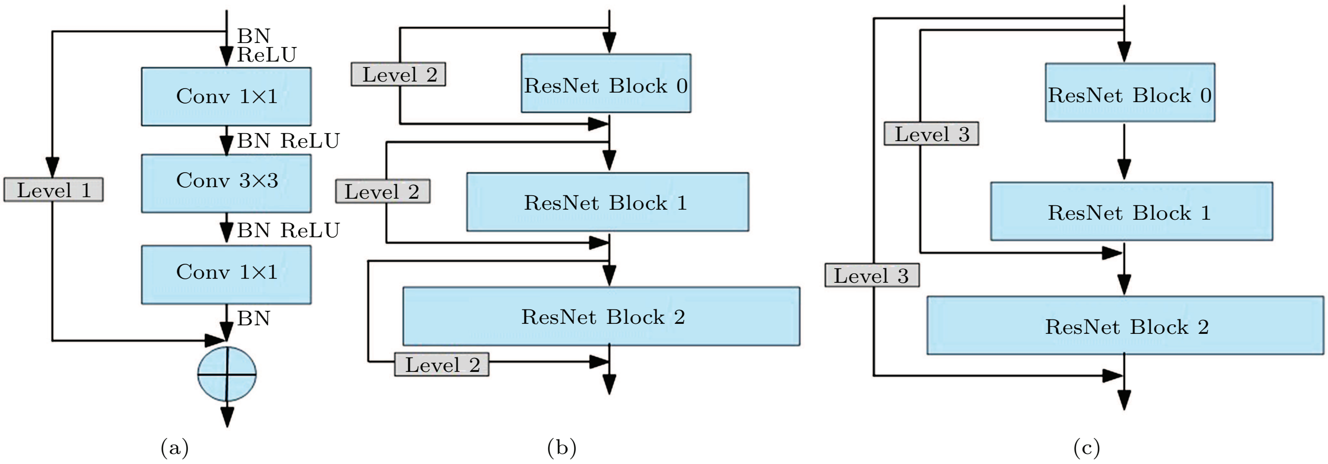

Fig.4. Structure of the shortcut connection.

3.5 ReLU and BN Unit's Placements

ReLU activation function[41] is one of the critical parts of a residual network for nonlinearity. According to Han *et al.*[37], the performance of the neural network extensively depends on the number and placement of the ReLUs in architecture. In the case of original ResNet, it is seen that using ReLUs after the addition of weight layers adversely influences the performance of the network. As shown in (3), ReLU is used over the addition of the ResNet block, and it has the function of filtering the non-negative values, where $\boldsymbol{F}_{(i,l)}$ signifies the i-th residual function of the i-th residual unit and x_i^f holds the outputs of residual network. It is empirically proven that by simply removing ReLUs after every addition in the ResNet Unit that leads to a significant improvement in the performance①. Later, He *et al.*[25] proposed Pre-Activation ResNet that holds BNs and ReLUs before the convolution layer and eradicates the ReLUs after the addition unit, and this structure leads to a great deal of performance boost. In (4) the ReLUs are removed after addition, to create identity path. As a result, the network offers a great deal of performance even at the depth larger than 1 000 layers.

$$x_i^f = ReLU(\boldsymbol{F}_{(i,l)}(x_{i-1}^l) + (x_{i-1}^l)), \quad (3)$$

$$x_i^f = (\boldsymbol{F}_{i,l}(x_{i-1}^l) + (x_{i-1}^l)). \quad (4)$$

According to Han *et al.*[37], the extensive use of ReLUs in the block leads to diminished performance. [37] has empirically showed that removing the first ReLU from the block improves the network performance.

The principal objective of the BN layer in the neural network is to normalize the activations, which offers fast convergence and improved performance. According to Han *et al.*[37], BN layers can be employed to maximize the capability of a single ResNet block. The authors also proved empirically that the use of the BN layer at the end of each block leads to better performance improvement. Therefore, in the proposed network block, the BN layers are added in each ResNet unit before and after convolution layers, in order to get better performance.

Ultimately, the final version of the proposed network's ResNet unit is employed by considering all the aforementioned rules. Fig.5 shows the different alternatives of ResNet units used in earlier experiments. The final version of proposed CNN's resent unit is Unit-D, where the surrounding ReLUs are removed from addition layers. In addition, the BN unit is being placed on both of the aforementioned places.

4 Optimization of Proposed Network

This section will outline some of the vital optimization aspects of the proposed network for experiments.

4.1 Shortcut Connection's Levels

In pursuit of an optimum number of shortcut connection's levels for the proposed CNN architecture, we have assessed many designs. Let L be the number of connection levels. The original residual network connection level is $L = 1$. For the proposed SRNet CNN we have set $L = 3$. According to Zhang *et al.*[27] increasing L leads to worsening the performance. Therefore, it is assessed that $L = 3$ is the best standard

① http://torch.ch/blog/2016/02/04/resnets.html, May 2019.

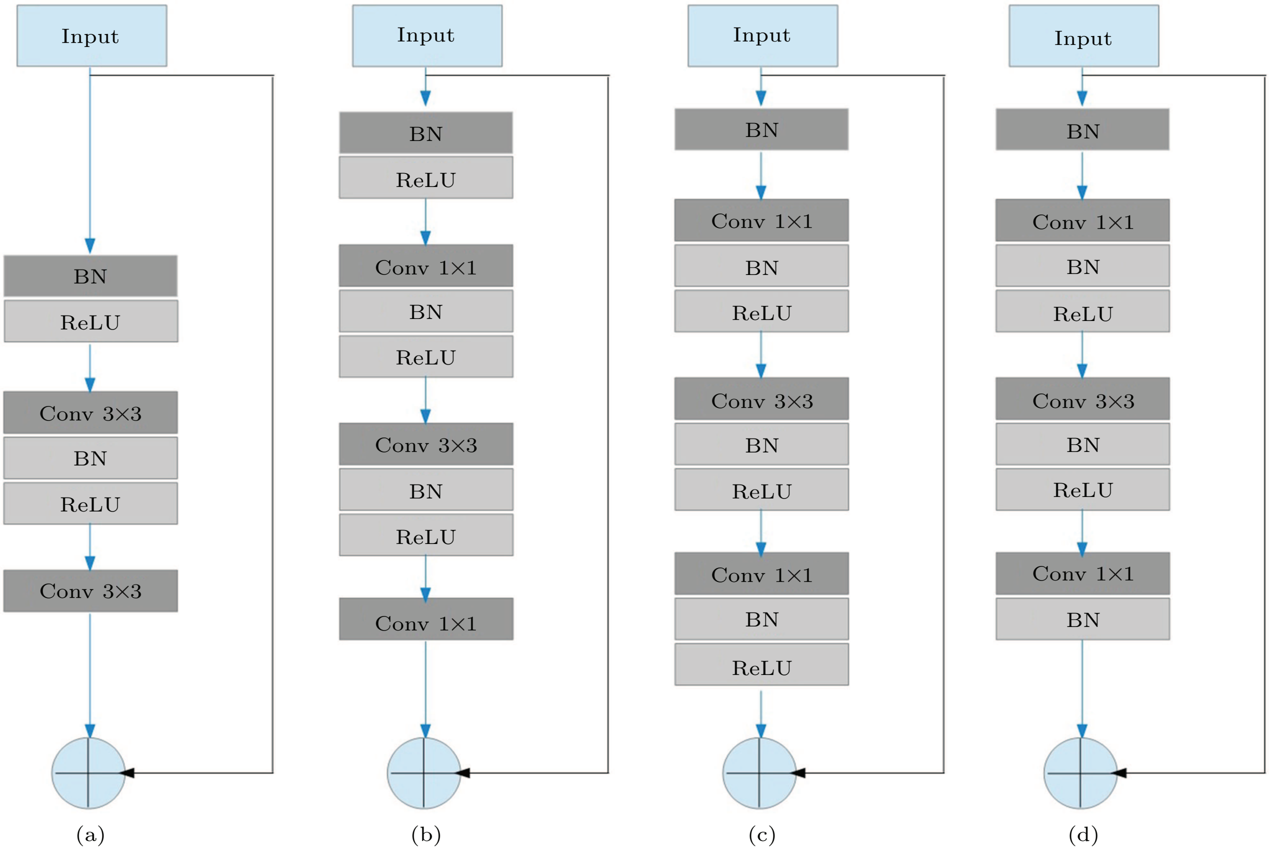

Fig.5. Structure of the resent blocks.

for adding level-based connections. Choosing the right number of levels is very important for getting a satisfactory performance. Adding additional shortcut connections leads to an additional number of parameters that ultimately result in the overfitting problem and diminish the network performance. Therefore, keeping $L = 3$ offers a reasonable amount of parameters to take advantage of additional connections. Fig.4 shows the possible levels of connections proposed for SRNet. The mentioned three different types of skips offer a diverse information flow and transform any given architecture to a wide CNN framework. As aforementioned, due to the parallel processing nature of current GPUs, wider CNN architectures are more feasible for training besides a deeper architecture.

4.2 Mapping of Proposed Network

To transfer the information from top to bottom in the proposed network, we need to map the network. In this situation, we have empirically analyzed different alternatives. According to He *et al.*[23] there are four different types of mapping or shortcut connections. 1) For identity mapping shortcut, no additional parameters are required. 2) A projection shortcut connection (1×1 Conv) uses a 1×1 convolution layer with “Xavier” weight initialization methodology. We have analyzed with different weight initialization methodologies and found that Xavier is extremely useful, especially for convolutional layers, as it offers a uniform distribution over the interval. 3) Non-linearity shortcut connections further improve the performance of B-Type connections. It has additional BN and ReLU units to normalize the distribution of features coming out of a convolution layer and these features might be negative which will be truncated by a non-linearity like ReLU. 4) Pooling shortcut connection is a bit computational expensive shortcut connection, though it offers a wide diversity of feature maps. Feeding network with diverse information is one of crucial base-lines of the proposed research. Therefore, it is considered to be significant to introduce a diversity of data even in shortcut connections.

The type-1 shortcut is the simplest category of connections that introduces neither extra parameters nor computational complexity. From computational complexity and additional parameters point of view type-1 and type-2 mappings are better than type-3 and type-4. Although type-4 adds additional parameters, it also offers more information that is necessary to process and get better accuracy. In the case of the proposed CNN, we have used a combination of above-mentioned connection. Level 1 used the type-2 shortcut connections. The level-2 used the type-3 connection, and level-3 used the type-4 connection. Fig.6 shows different alternatives for mapping for the proposed network.

4.3 Epochs

He *et al.*[23,25] proposed to use 164 epochs to train and optimize the residual network. Zhang *et al.*[27] showed empirically that 500 epochs could bring significant improvements in performance. We have set goals to check network performance after every 10 epochs, until it begins converging. We also set a dynamic learning rate that goes down by factor of 10 after every 20 epochs. It is proven to be very suitable for network performance and learning capability.

4.4 Dropout

SRNet network widens the CNN architecture and adds additional training parameters (due to additional level based shortcut connections). Therefore, the problem of overfitting is critical. In the case of training CIFAR-100 dataset, this problem needs to be addressed. This kind of architecture leads to a serious overfitting problem. The most famous methods to mit-

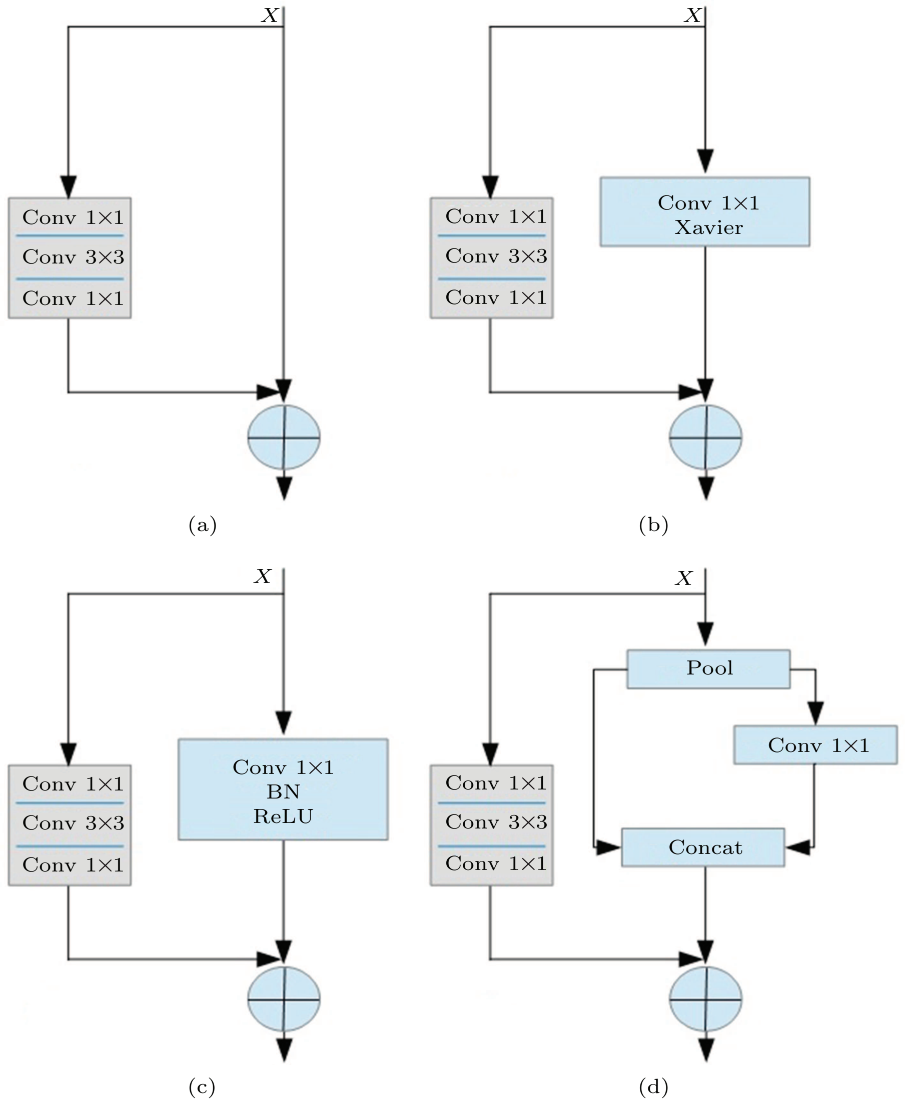

Fig.6. Mapping of the ResNet connections.

igate this problem are DropPath[24,42] and Dropout[35]. The DropPaths method averts co-adaptation of parallel paths through randomly dropping the path. According to He *et al.*[25], the network is not able to converge to an efficient solution by dropping an identity mapping path randomly. Dropping an identity mapping negatively impacts training. Therefore, we have used the Dropout methodology for the sake of fixing the overfitting issues. Though, it is assessed that Dropout is less effective while used in convolutional layers[34]. Therefore, dropout layers are carefully planted among convolution units to resolve the overfitting problems.

5 Experiment and Analysis

This section will present the empirical analysis of the proposed CNN on benchmark datasets CIFAR-10 and CIFAR-100.

5.1 Implementation

CIFAR-10 and CIFAR-100 are used for the proposed CNN analysis. CIFAR-10 is a well-known dataset containing 10 classes of objects, and CIFAR-100 includes 100 object classes. There are 5 000 images in each dataset used for training classification and 10 000 images for each dataset for testing classification. The standard data augmentation is used for proposed experiments. The proposed CNN is designed to train in Caffe. Caffe Berkeley Vision library[43] is used to develop and train the whole network. Caffe offers a great deal of freedom to write network layers and train the network according to the proposed requirements. The proposed network was trained on single NVIDIA Tesla K40c GPU with 12 G memory.

5.2 Training Setup

The network is trained using the backpropagation[44] by Stochastic Gradient Descent (SGD) with Nesterov momentum on CIFAR-10 and CIFAR-100 datasets. We have used a mini batch of 120 images. The network is initialized using the initialization method of He *et al.*[13] The learning rate is set to 0.1 for CIFAR-10 dataset and 0.5 for CIFAR-100 dataset. The learning rate decay factor is set to α, and decay is performed at 150 and 225 epochs. The network is initialized by MSRA[13]. The weight decay parameters are set to 0.000 1. The momentum is set to 0.9 and dampening is set to 0. We have experimented with different values of momentum such as [0.5, 0.9, 0.95, 0.99], and later we fix it to 0.9 to efficiently dampen the velocity and reduce the kinetic energy of our SGD function. It is analysed during our experiments that momentum 0.9 offers a more stable final loss.

For network initialization, we have used the proposed method of network initialization of He *et al.*[13] These parameters are set after initial experiments to check that CNN is converging during training. It is observed that early stop approach is useful to stop CNN in early stages of training. It is noted that a higher number of epochs can lead to having an over-fitted model. Therefore, we have designed a network that will be trained for a given number of epochs, while we will decrease the learning rate by the factor of 10 after every 50 epochs. This approach helped to train and coverage the network more efficiently. Setting batch size during training is also a critical aspect of CNN learning process. Higher batch size leads to a buffer overflow problem that is a vital issue in network training. Therefore, according to our GPUs availability, we set the batch size to 120 images to avoid such issues. Images were shuffled to ensure the diversity in each batch. The cross-entropy loss was used to sum up the entire losses of the mini-batch pixels.

5.3 CIFAR-10

We have used the mean squared error (MSE) loss for assessment of our network outcomes. (5) shows the proposed MSE loss calculation method, and calculates the mean $\left(\frac{1}{m}\sum_{i=1}^{m}\right)$ of the squares of the errors $(Y_i - \hat{Y}_i)^2$, where m is the total number of given classes, Y is the inference result of our proposed model, and $\hat{Y}$ is the ground truth. We transform it into test error for our network error assessment. (5) is given below:

$$\begin{aligned}&\textsf{Test error}: \ J_{\text{test}}(\theta)\\&= \frac{1}{2m_{\text{test}}}\sum_{i=1}^{m_{\text{test}}}(h_{\theta}((x_{\text{test}}^{(i)}) - y_{\text{test}}^{(i)})^2),\end{aligned} \tag{5}$$

where $J_{\text{test}}(\theta)$ is the test error calculated through taking squares of the errors of network test output $x_{\text{test}}^{(i)}$ and ground truth $y_{\text{test}}^{(i)}$. Proposed CNN had a test error-rate of 4.137 on CIFAR-10, which is much better to original Pre Residual Network with 110 layers. On the other hand, the numbers of parameters are not much higher compared with other wide models like VGG-16[7]. The results have shown that through increasing the number of skip connection levels, the network performance is improved, though we have achieved this performance on the expense of a bigger network size. It is observed that adding the three-level skip connection design to network offers a great deal of performance boost compared with two-level skip connection designs. The two-level connection architecture offers an error of 5.643 while adding one more level we achieved a lower error of 4.137.

5.4 CIFAR-100

Table 2 shows the results of CIFAR-100 experiments. In proposed experiments, we have used SGD along with the batch size of 128. The network is trained and validation (test set) checks are performed after every 10 epochs, to assess the performance of network training. The learning rate is set to 0.1. Different from CIFAR-10 dataset we have used the Dropout layer for CIFAR-100 experiments. The Dropout layer settled among the $(1\times3\times1)$ residual unit, after the (3×3) convolution layer. After many trials, the dropout rate is set to 0.3. In the absence of dropout layer the problem of overfitting can be escalated, as additional branches of skip links incorporate extra trainable parameters. Table 2 shows that ResNet-110 and Pre-ResNet-110 without using Dropout or Spectral Dropout (SD) offered the test error of 25.16% and 24.33% respectively. However, our proposed ResNet block with just 16 layers and additional skip connections offers better performance. It is also noticed that the proposed network provided better training time. The time is substantially lowered as compared with traditional approaches, as shown in Table 3. Regarding lowering error rates, we have observed a similar trend in CIFAR-100 dataset training. Similar to CIFAR-10 results we have a better network accuracy rate and lowered test error rate by using three different levels of skip connections. The two-level skip connection architecture offered a 23.07 error rate while adding one more level. There is a better accuracy achieved with a lower test error rate of 22.19. Therefore, we beat most of benchmark results in terms of the test error as well as the network size.

5.5 Results and Analysis

The key focus of this research was to show that additional information flow in a CNN is contributing to the final detection task. The depth of the proposed CNN was kept to a constant amount of parameters. Zagoruyko and Komodakis[34] stated the significant reasons for keeping CNN width between 16-28 offers steady improvements in the network performance. The network depth is kept shallow to speed up the network training. It is assessed that a widening factor is more efficient as compared with a deepening factor. It helps to spread jobs to multiple GPUs that are designed to perform efficiently in parallel computing. After experimenting with different depths and width factors, we have discovered the current experimental parameters offer efficient performance and constant improvements in network performance. This analysis provided a rea-

Table 2. Test Error on CIFAR-10 and CIFAR-100 with Different ResNet Networks

Block Type	Depth	Parameter (M)	CIFAR-10	CIFAR-100
Network in Network[5]	-	-	8.810	35.67
Highway Network[26]	-	-	7.720	32.39
Stochatic Depth Network[24]	110	1.7	5.230	25.58
Original Residual Network (100)[23]	110	1.7	6.430	25.16
Pre-activation ResNet[25]	110	1.7	6.370	24.33
Wide Residual Network-16[34]	16-8	11.0	4.270	22.07
RoR + WRN-58-4[27]	58—4	-	3.770	25.19
SRNet (Lev1+Lev2)	16—4	3.76	5.643	23.07
SRNet (Lev1+Lev2+Lev3)	16—4	4.58	4.137	22.19

Table 3. Analysis of Proposed Network

CNN	Memory Access (MB)	Comp.	Add.	Div.	Exp.	Activation	Param.
ResNet-18	26.68	151.55 K	305.16 K	139.27 K	5	607.24 K	895.11 K
WRN-16	393.18	704.51 K	721.00 K	475.24 K	100	2.48 M	2.77 M
WRN-16-RoR	541.34	1.92 M	3.42 M	1.92 M	20	9.57 M	4.11 M
SRNet (Lev1+Lev2)	507.25	1.39 M	3.29 M	1.39 M	20	8.54 M	3.77 M
SRNet (Lev1+Lev2+Lev3)	571.73	1.92 M	3.77 M	1.92 M	20	10.16 M	4.58 M

Note: Comp.: computation; Add.: addition; Div.: division; Exp.: exponential; Param.: parameter.

lization that the depth and the width of network equally play a significant role in network learning capabilities. Choosing a suitable depth or width makes a fundamental difference in final network performance. Increasing the depth until the network performance saturates and later gradually augmenting the network width can offer prolific results. The final results have shown that embedding additional communication links among different levels of a convolution network offers a great deal of information flow inside the network. This additional information flow leads to better performance and improved results. It is established that when an image is passed through extensive neural layers, it sometimes losses all critical details necessary for the detection of essential objects in the image. The additional branches from top layers to middle and bottom layers are proved to be useful in the regeneration of critical features that are significant in final detection layers. This research has demonstrated that besides going too deep, we can make use of a typical wide CNN with additional communication layers to get improved results as compared with benchmark counterparts. This research also shows that adding more width to the network can lead to better performance at the expense of additional trainable parameters.

The use of skip connections at different depths of the network helps us to eradicate network singularities along with maintaining similar performance. The proposed research makes use of specially designed skip connections which are placed top, middles and bottom of a wide CNN architecture. The wide network design is preferred over deep network design due to its high performance over deep design. A wide CNN with specialized skip connections leads us to deal with elimination, overlap and linear dependence singularities. The design and mechanism of fetching and feeding additional information into the network helps the overall architecture to tackle overcoming elimination, overlap and linear dependence singularities. As a result, the network is populated with enough information to deal with any feature loss that ultimately leads to the aforementioned singularities. The proposed design is a general idea that can be applied to any network suffering from training related complications (overfitting, singularities, etc.). The key aspect of this architecture is that it does not add additional load to network resources regarding memory and processing parameters. While designing and placing skip connections, it is ensured that these connections are not memory or processing-extensive. This paper has presented the results on CIFAR-10 and CIFAR-100 datasets. The proposed CNN is altered for CIFAR-100 experiments as it requires additional parameters. Therefore dropout layers are introduced to save the network from overfitting. Table 3 offers an interesting view of proposed architecture along with earlier similar designs. It provides the description of networks regarding memory accesses, arithmetic operations and the number of parameters. This contrast demonstrates a great deal of insight into the effective network design of SRNet. It offers better architecture regarding all above mentioned CNN parameters. Fig.7 offers a graphical view of the training process of the SRNet, where the green curve shows the network accuracy that converges after approximately 50 000 iterations. The blue line shows the network loss that gradually decreases and starts converging approximately after 50 000. The yellow curve shows the network test loss. It shows similar convergence as training loss. This graph collectively shows network training and testing process with very natural convergence of loss and accuracy. These curves show a normal network training behavior, where training accuracy and loss start from a higher point finally converge to a constant level.

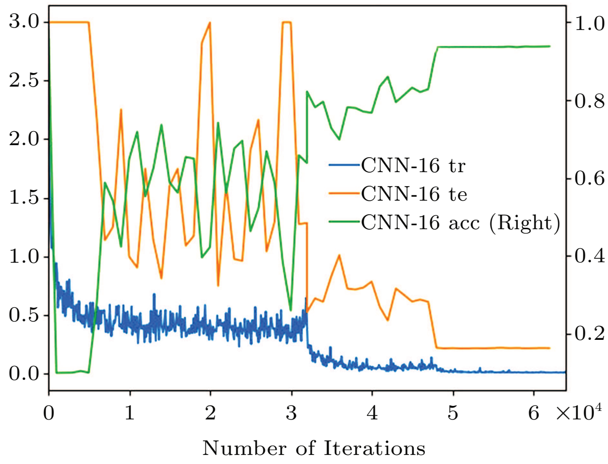

Fig.7. Training (tr), testing (te) and accuracy (acc) curves on the CIFAR-10 dataset.

The key aspect that differentiates the proposed research from earlier studies is that it offers a practical design for application of wide CNN with additional skips. Any CNN architecture can be transformed and improved by adding other levels of skip connections. The proposed types and levels of additional communication links suggested in this research are experimented for different datasets and architectures. It is assured that the addition of these skip connections offers better accuracy and helps to resolve singularities. The idea of this research is not merely removing singularities but also improving network performance and efficiency.

The idea of adding additional communication channels offers a great deal of performance boost to overall network architecture, though it comes with some additional resource cost. Through the addition of skip connection among different levels of the network, there is an increase in trainable parameters. This increase depends on the levels of skip connections we are using. The use of three-level skip connections has added only 2.58% increase into overall trainable parameters and storage space.

6 Conclusions

The fundamental idea of the proposed novel convolutional network architecture is to enhance and update CNN blocks structure and add additional communication links in the network. The objective of the research is to introduce different level skip connections to mitigate network singularities. As, our experiments and empirical results showed that through lessening singularities we can contribute a big part in the network efficiency and performance. The use of skip connection leads to many benefits like improving network information flow, better top-to-bottom connectivity, resolving vanishing gradients, feature renewal, and most importantly extenuating singularities. The resultant CNN network offered better information flow and fix for vanishing gradient. The new network blocks equipped with different levels of connections based architecture lead to significant improvement in the information flow inside the network, which ultimately offers better classification performance. The proposed architecture was trained and tested on well-known benchmark datasets CIFAR-10 and CIFAR-100. The results showed that our proposed network outperformed earlier similar approaches. Furthermore, these reformed residual units could be utilized in any CNN architecture to enhance the capability and performance of the network. And the proposed intriguing finding will facilitate further advancements in deep neural network design and architecture.

7 Future Work

Incorporating skip connections in CNN design offer capabilities to deal with different singularities. On the other hand, skip connections introduce some additional parameters in the network. These days the size and speed of the network is also a critical aspect that needs considerable attention. The future of work of our research is to propose Binary CNN architecture with added skip connection to deal with singularities above. It is already proven by Rastegari *et al.*[45] that the idea of binarizing network weight and parameters offers a great deal of resource efficient CNN architecture. Currently, [46–48] proved that binary CNN is a feasible idea and can be used for a wide range of real-world applications. We intend to design and deploy a binary CNN architecture with skip connections in a real-time object recognition environment.

Acknowledgement(s) I would like to acknowledge Tony Pridmore, Michael Pound, Khan Faraz, Mohammadreza Soltaninejad and John Atanbori of Computer Vision Laboratory, School of Computer Science, University of Nottingham, for insightful discussions.

References

[1] Krizhevsky A, Ilya S, Geoffrey E H. ImageNet classification with deep convolutional neural networks. In *Proc. the 26th Annual Conference on Neural Information Processing Systems*, December 2012, pp.1106-1114.

[2] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Berg A C. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, 115(3): 211-252.

[3] LeCun Y, Yoshua B, Geoffrey E H. Deep learning. *Nature*, 2015, 521(7553): 436-444.

[4] Zou W Y, Wang X, Sun M, Lin Y. Generic object detection with dense neural patterns and regionlets. arXiv:1404.4316, 2014. https://arxiv.org/abs/1404.4316, July 2018.

[5] Lin M, Chen Q, Yan S. Network in network. arXiv: 1312.4400, 2013. https://arxiv.org/abs/1312.4400, July 2018.

[6] Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. OverFeat: Integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229, 2013. https://arxiv.org/abs/1312.6229, July 2018.

[7] Simonyan K. Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014. https://arxiv.org/abs/1409.1556, July 2018.

[8] Yasrab R. ECRU: An encoder-decoder based convolution neural network (CNN) for road-scene understanding. *Journal of Imaging*, 2018, 4(10): Article No. 116.

[9] Yasrab R, Gu N, Zhang X. SCNet: A simplified encoder-decoder CNN for semantic segmentation. In *Proc. the 5th International Conference on Computer Science and Network Technology*, December 2016, pp.785-789.

[10] Yasrab R, Gu N, Zhang X. An encoder-decoder based convolution neural network (CNN) for future advanced driver assistance system (ADAS). *Applied Sciences*, 2017, 7(4): Article No. 312.

[11] Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. In *Proc. the 30th International Conference on Machine Learning*, June 2013, pp.1139-1147.

[12] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In *Proc. the 13th International Conference on Artificial Intelligence and Statistics*, May 2010, pp.249-256.

[13] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proc. the 2015 IEEE International Conference on Computer Vision*, December 2015, pp.1026-1034.

[14] Lee C Y, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. In *Proc. the 18th International Conference on Artificial Intelligence and Statistics*, May 2015, pp.562-570.

[15] Raiko T, Valpola H, LeCun Y. Deep learning made easier by linear transformations in perceptrons. In *Proc. the 15th International Conference on Artificial Intelligence and Statistics*, April 2012, pp.924-932.

[16] Schmidhuber J. Learning complex, extended sequences using the principle of history compression. *Neural Computation*, 1992, 4(2): 234-242.

[17] Chen T, Goodfellow I, Shlens J. Net2net: Accelerating learning via knowledge transfer. arXiv:1511.05641, 2015. https://arxiv.org/abs/1511.05641, November 2018.

[18] Romero A, Ballas N, Kahou S E, Chassang A, Gatta C, Bengio Y. FitNets: Hints for thin deep nets. arXiv: 1412.6550, 2014. https://arxiv.org/abs/1412.6550, July 2018.

[19] Wei H, Zhang J, Cousseau F, Ozeki T, Amari S. Dynamics of learning near singularities in layered networks. *Neural Computation*, 2008, 20(3): 813-843.

[20] Amari S I, Park H, Ozeki T. Singularities affect dynamics of learning in neuromanifolds. *Neural Computation*, 2006, 18(5), 1007-1065.

[21] Saxe A M, McClelland J L, Ganguli S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120, 2013. https://arxiv.org/abs/1312.6120, August 2018.

[22] Orhan A E, Pitkow X. Skip connections eliminate singularities. arXiv:1701.09175, 2017. https://arxiv.org/abs/1701.09175, September 2018.

[23] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.770-778.

[24] Huang G, Sun Y, Liu Z, Sedra D, Weinberger K Q. Deep networks with stochastic depth. In *Proc. the 14th European Conference on Computer Vision*, October 2016, pp.646-661.

[25] He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In *Proc. the 14th European Conference on Computer Vision*, October 2016, pp.630-645.

[26] Srivastava R K, Greff K, Schmidhuber J. Highway networks. arXiv:1505.00387, 2015. https://arxiv.org/abs/1505.00387, June 2018.

[27] Zhang K, Sun M, Han X, Yuan X, Guo L, Liu T. Residual networks of residual networks: Multilevel residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, 28(6): 1303-1314.

[28] Zhang K, Guo L, Gao C, Zhao Z. Pyramidal RoR for image classification. arXiv:1710.00307, 2017. https://arxiv.org/abs/1710.00307, May 2018.

[29] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In *Proc. the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp.1-9.

[30] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 1994, 5(2): 157-166.

[31] Shen F, Gan R, Zeng G. Weighted residuals for very deep networks. In *Proc. the 3rd International Conference on Systems and Informatics*, November 2016, pp.936-941.

[32] Bengio Y, LeCun Y. Scaling learning algorithms towards AI. In *Large-Scale Kernel Machines*, Bottou L, Chapelle O, DeCoste D, Weston J (eds.), MIT Press, 2017.

[33] Larochelle H, Erhan D, Courville A, Bergstra J, Bengio Y. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proc. the 24th International Conference on Machine Learning*, June 2007, pp.473-480.

[34] Zagoruyko S, Komodakis N. Wide residual networks. arXiv:1605.07146, 2016. https://arxiv.org/abs/1605.07146, January 2019.

[35] Srivastava N, Hinton G E, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014, 15(1): 1929-1958.

[36] Huang G, Liu Z, Weinberger K Q, Maaten L. Densely connected convolutional networks. arXiv:1608.06993, 2016. https://arxiv.org/abs/1608.06993, September 2018.

[37] Han D, Kim J, Kim J. Deep pyramidal residual networks. arXiv:1610.02915, 2016. https://arxiv.org/abs/1610.02915, July 2018.

[38] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.5987-5995.

[39] Szegedy C, Loffe S, Vanhoucke V, Alemi A A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proc. the 31st AAAI Conference on Artificial Intelligence*, February 2017, pp.4278-4284.

[40] Loffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. the 32nd International Conference on Machine Learning*, July 2015, pp.448-456.

[41] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines. In *Proc. the 27th International Conference on Machine Learning*, June 2010, pp.807-814.

[42] Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580, 2012. https://arxiv.org/abs/1207.0580, July 2018.

[43] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In *Proc. the 22nd ACM International Conference on Multimedia*, November 2014, pp.675-678.

[44] LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W, Jackel L D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1989, 1(4): 541-551.

[45] Rastegari M, Ordonez V, Redmon J, Farhadi A. XNOR-Net: ImageNet classification using binary convolutional neural networks. In *Proc. the 14th European Conference on Computer Vision*, October 2016, pp.525-542.

[46] Sheen S, Lyu J. Median binary-connect method and a binary convolutional neural network for word recognition. arXiv:1811.02784v1, 2018. https://arxiv.org/abs/1811.02784v1, December 2018.

[47] Lin X, Zhao C, Pan W. Towards accurate binary convolutional neural network. In *Proc. the 2017 Annual Conference on Neural Information Processing Systems*, December 2017, pp.344-352.

[48] Juefei-Xu F, Boddeti V N, Savvides M. Local binary convolutional neural networks. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.4284-4293.

Robail Yasrab received his Ph.D. degree in computer vision from the School of Computer Science, University of Science and Technology of China, Hefei, in 2017. He is currently a research fellow at the Computer Vision Laboratory, School of Computer Science, University of Nottingham, Nottingham, United Kingdom. His research interests include artificial intelligence, computer vision, deep leaning, plant phenotyping, and particularly the application and adaptation of modern machine learning techniques to real-world problems.