Wide Residual Networks

Authors :
Sergey Zagoruyko
Nikos Komodakis

Accepted in:
British Machine Vision Conference(BMVC)
Year: 2016

Presented by:
Yasmeena Akhter, P19CS209

Paper link: http://www.bmva.org/bmvc/2016/papers/paper087/paper087.pdf

Motivation

CNN have seen a gradual increase in number of layers from AlexNet to InceptionNet

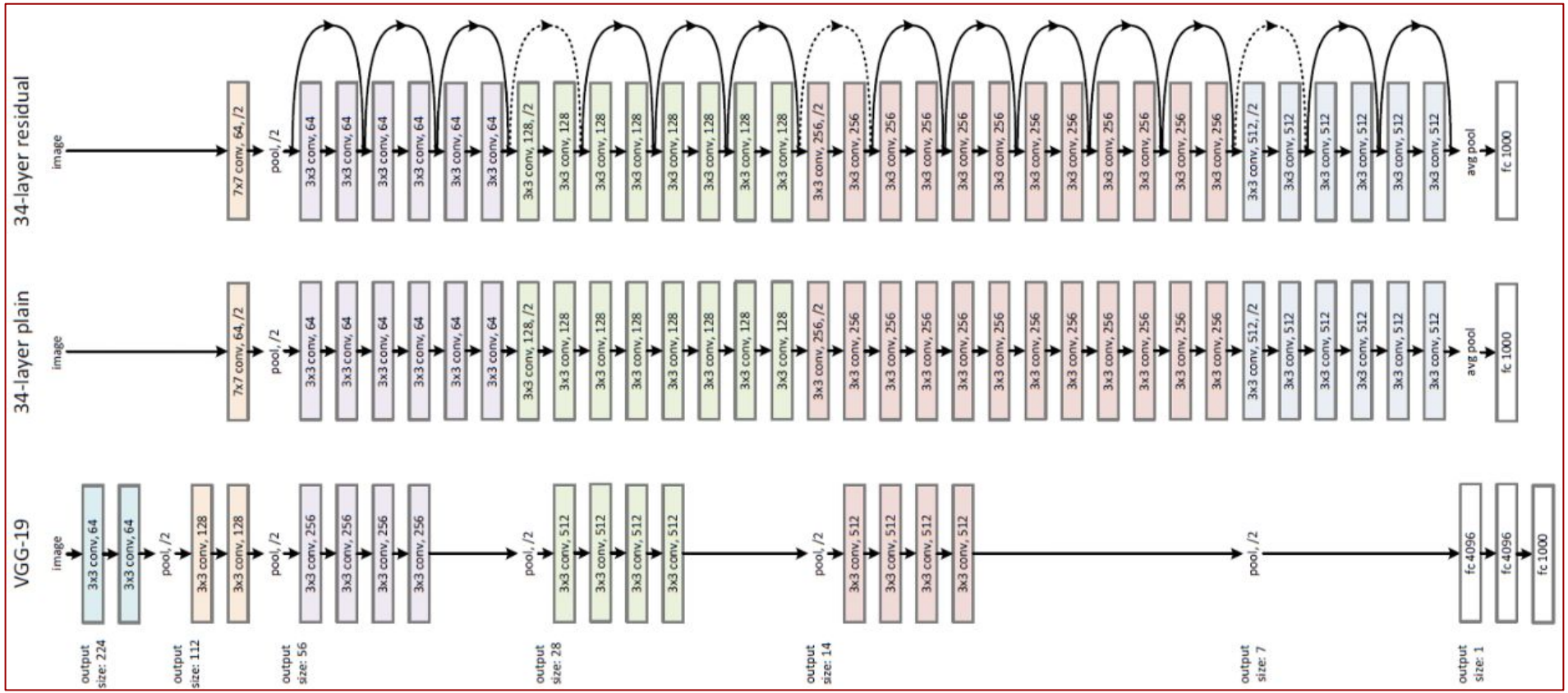
Simple stacking of layers didn't show much good performance because of vanishing/exploding gradient problem

Few techniques ,such as well-designed initialization strategies, better optimizers ,skip connections, knowledge transfer and layer-wise training were also used

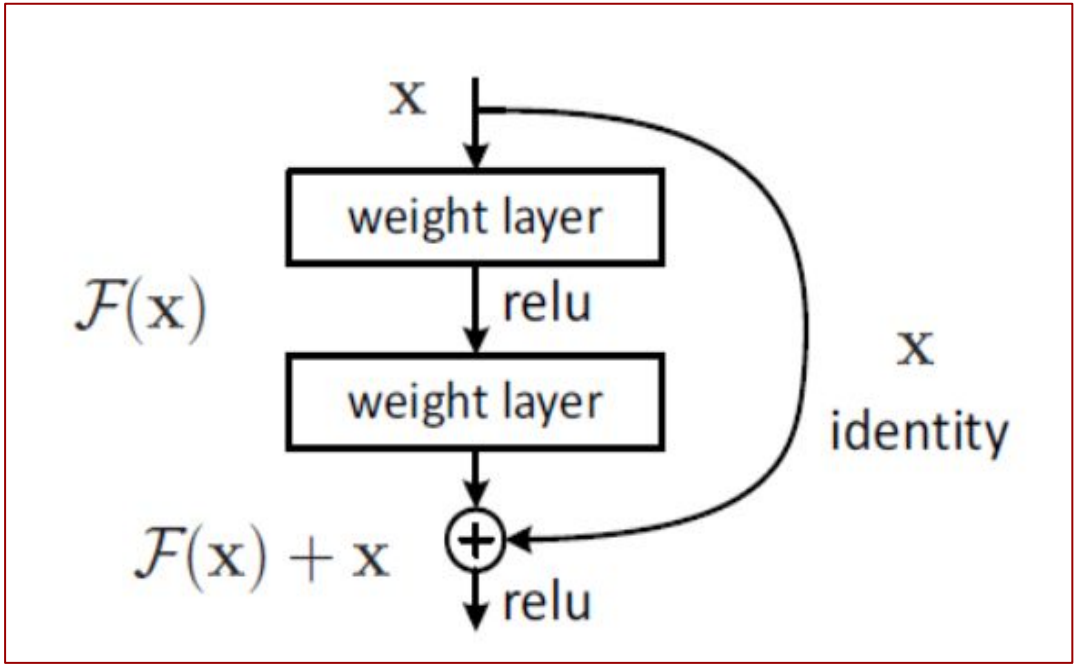
Residual networks encouraged scaling up the number of layers,by showing speedy convergence of deep nets

ResNets model are based on the order of activations inside the ResNet block and its depth [1]

ResNet Architecture



Simple ResNet Block



Problems With ResNet

Circuit Complexity Theory: suggests that shallow circuits can require exponentially more components than deeper circuits.

To overcome this, authors have used tried to make them thin as possible. Also introduced «bottleneck» block resulting in thinning of ResNet block

Diminishing Feature Reuse: As gradient flows through the network there is nothing to force it to go through residual block weights and it can avoid learning anything during training

- **So one or more residual blocks learn no useful representations**

Approach

Approach is based on the motivation of the mentioned observations

To address these issues authors have build their approach on the top of the work done in [3]

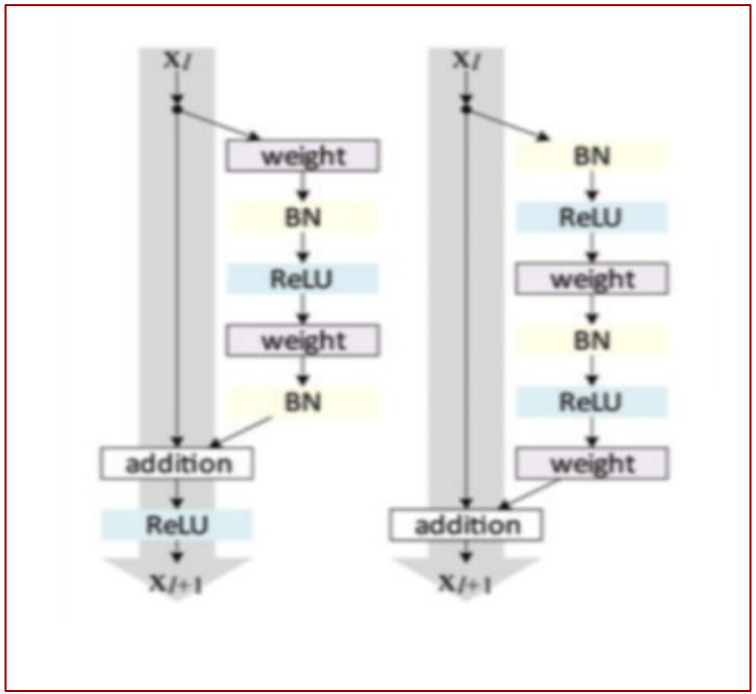
Approach tries to answer the question of how wide deep residual networks should be and address the problem of training

Authors have presented wider deep residual networks that significantly improve over baseline, having 50 times less layers and being more than 2 times faster

Resulting network architectures are termed as wide residual networks (WRNs)

[3] Identity mappings in deep residual networks.

Architecture used



Left: Resnet(Original) Right: ResNet with Identity Mapping

$$x_{l+1} = x_l + F(x_l, W_l)$$

Above equation represents the Residual block with identity mapping

where x_{l+1} and x_l are input and output of the l^{th} unit in the network, F is a residual function and W_l are parameters of the block

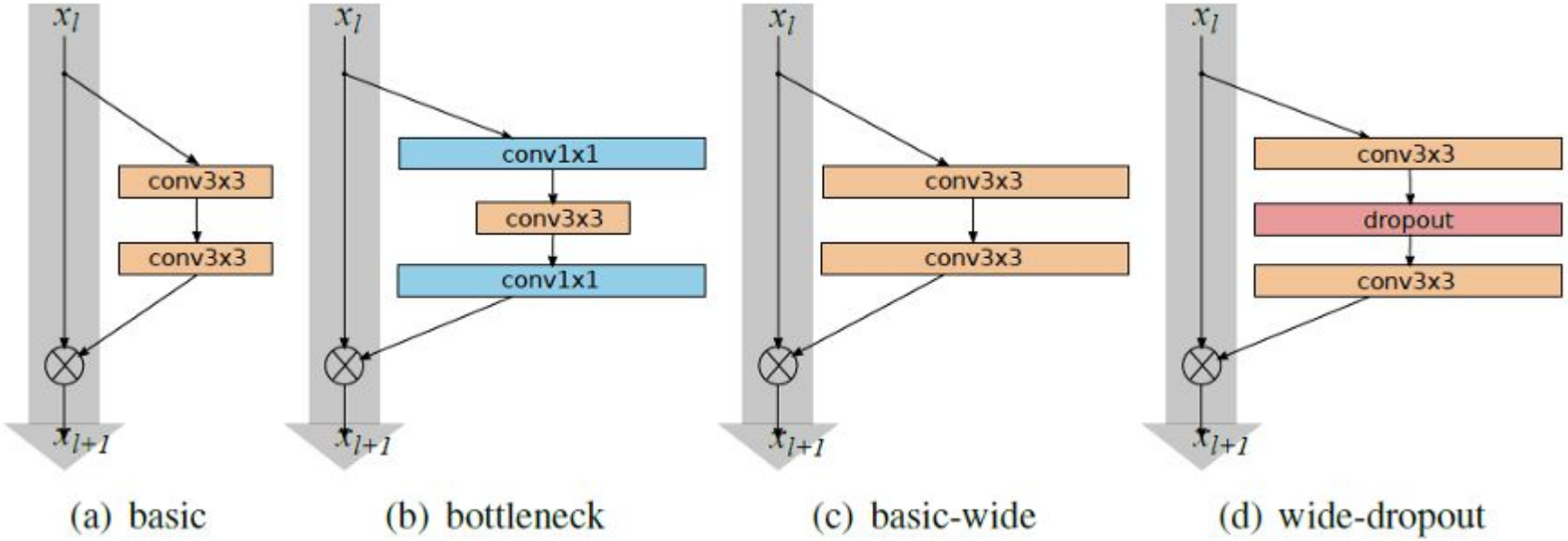
Same architecture is used as given in the baseline[with an exception of bottleneck block [3]

Filter of size (3x3) are used for experiments

Two factors are introduced by authors as deepening factor l and widening factor k,

l is the number of convolutions in a block and k multiplies the number of features in convolutional layers

Figure 1: Various residual blocks used in the paper. Batch normalization and ReLU precede each convolution (omitted for clarity)



(a) basic (b) bottleneck (c) basic-wide (d) wide-dropout

Dropout in residual blocks

Authors have used the dropout as regularization into each residual block between convolution and after ReLU to perturb batch normalization in the next residual block and prevent it from overfitting

Dataset used

- CIFAR
- SVHN
- COCO
- ImageNet

Highlights of the paper:

Widening consistently improves performance across residual networks of different depth

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 unfeasibly expensive to train

How WRNs are different from DRNs

- No bottleneck blocks are used
- Number of convolutions between skip connections is varied/different
- Regularization layers are introduced in residual blocks
- Reduced the number of layers and training time as well

Note: For further explanation on number and type of convolutions used refer to experiments section in the paper

Overall contribution of the paper

Paper presented a detailed experimental study of residual network architectures that thoroughly examines several important aspects of ResNet block structure

Paper proposed a novel widened architecture for ResNet blocks that allowed residual networks with significantly improved performance

Paper proposed a new way of utilizing dropout within deep residual networks so as to properly regularize them and prevent overfitting during training

Lastly, Paper has shown that the proposed ResNet architectures achieved state-of-the-art results on several datasets, dramatically improving accuracy and speed of residual networks.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015

[2] Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks." *arXiv preprint arXiv:1605.07146* (2016).

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. CoRR, abs/1603.05027, 2016.

Thank you