Deep Residual Learning for Image Recognition

Summary written by Sanika Phatak
April, 24, 2020

Summary: This paper introduces residual learning framework (ResNet) to ease the training of deep networks. Deeper networks have degradation problem as the depth of layers increases which indicates that all the systems are not easy to optimize. Residual networks tackle this issue by reformulating the layers as learning residual functions with reference to layer inputs which are easier to optimize and can gain accuracy with increased depth. Residual networks are also called shortcut connections as they skip over layers in the residual block by using residual mapping which allows considerable depth in networks with very less increase in computation time.

Related work: ResNet adopts from the fact that residual representations help the solvers converge faster than normal representations as shown in multigrid method [1]. Highway networks [5],[6] present shortcut connections with gating functions and the idea of ResNet is concurrent to this idea.

Approach: ResNet introduces residual mapping in deep networks. With $H(x)$, the underlying mapping fit by a few layers, stacked layers can be assumed to map a residual function $H(x) - x$ making the original function $F(x) + x$. The degradation problem suggests that solvers have difficulty mapping non linear layers and with optimal residual mapping, its suggested that the learning is easier due to reasonable preconditioning. The identity mapping building block is represented by $y = F(x, W_i) + x$ where x and y are inputs and outputs respectively and function F is the residual mapping learned which can be represented by $F = W_2\sigma(W_1x)$ for a two layer block with σ being the activation function ReLU [3]. If the dimensions of x and F are not equal, linear projections (W_s) are performed to match the dimensions and are represented by $y = F(x, W_i) + W_sx$. ResNet is tested against a baseline plain network with equal number of layers and both the architectures based on VGG net [4]. The residual shortcuts introduced in the plain layer for the residual network and identity mapping is used when outputs are of same dimension. If the output are of different dimensions, two methods are tested: (A) padding of zeros to identity shortcut to increase dimensions and (B) Projection shortcut. Batch Normalization [2] is done after each convolution layer to avoid vanishing gradients problem and SGD (Stochastic Gradient Descent) solver is used. Deep bottleneck residual architectures have 3 layers with 1x1, 3x3 and 1x1 convolutions and identity shortcuts.

Experiments and results: The two architectures are evaluated on ImageNet 2012 dataset. The networks are evaluated for 18-layer and 34-layer nets and training errors are compared and it is observed that the plain network with 34 layers has higher training error than the 18 layer plain network but the 34 layer ResNet performs better than the 18-layer ResNet by 2.8%. This shows that the degradation problem which occurs in the deep plain net probably due to exponentially low convergence rates is taken care of the the ResNet structure. The 18-layer ResNet and plain net are comparable accurate but ResNet converges faster. Three different shortcut methods are evaluated, (A) zero padding shortcuts used for increasing dimensions, (B) projection shortcuts used for increasing dimensions and others are identity shortcuts and (C) all the shortcuts are projections. (B) performs better than (A) as the zero padded dimensions in (A) have no residual learning and (C) performs marginally better due to extra parameters introduced but all these methods perform considerably better than plain nets which indications projection shortcuts are not essential for degradation problem. For the bottleneck residual blocks, layers up to 152 are tested which increase the accuracy with depth but still have lesser complexity (11.3 billion FLOPS) as compared to a shallower VGG-16/19 net (15.3/19.6 billion FLOPS.

Strengths: The main strength of ResNet is that it drastically improves the learning of deep networks without increasing any parameters or introducing complexities. ResNet enables to have very deep networks (almost 8 times deeper than popular networks like VGG), which enhance the accuracy without increasing the computational requirements.

Weaknesses: The paper mainly demonstrates how residual mapping reduces boosts the learning process and accuracy but does not give an explanation about why deeper networks have optimization difficulties causing degradation problem and how ResNet solves this issue.

Reflections: ResNet is a powerful tool to increase performance and accuracy of very deep networks. It is a simple and elegant method to tackle the degradation problem in deep networks. Deeper networks improve recognition as well as localization and ResNet can be a very important tool to get very accurate models for object recognition and classification in real world. There is a lot of scope of further research in how ResNet improves performance.

References

[1] W. Briggs and S. McCormick. A multigrid tutorial. 2000.

[2] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift.

[3] V. Nair and G. Hinton. Rectified linear units improve restricted boltzmann machines. *ICML*, 2010.

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

[5] R. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv*, 1505(00387), 2015.

[6] R. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. *arXiv*, 1507(06228), 2015.