Notes - Deep Residual Learning

Friday, January 4, 2019 12:37 AM

What is the focus of the research paper?
To ease the training of the networks that are substantially deeper than those used previously.

What is the specialty about this approach?
This approach reformulates the layers as residual functions with reference to the layer inputs, instead of learning unreferenced functions.

What is the advantage of deep networks?
They naturally integrate low/mid/high-level features and classifiers in an end-to-end multi-layer fashion and levels of features can be enriched by the number of stacked layers (depth). Depth is important aspect and it proved to be an important factor to increase the performance. Because of this the winning models in ImageNet used the deep networks in their approaches.

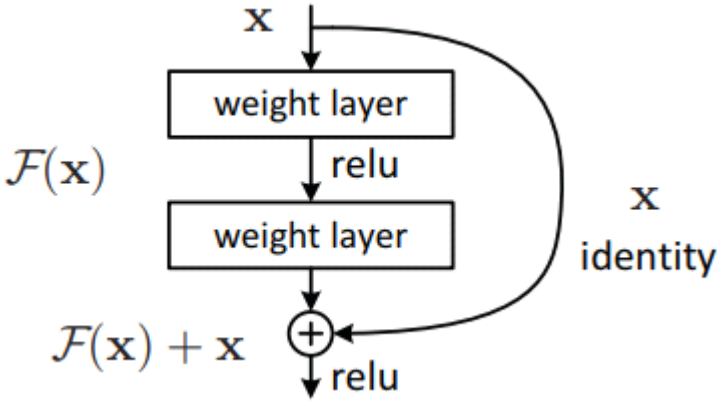
Can we improve the performance just by stacking the layers?
If the layers are stacked, it will raise to the problem of vanishing / exploding gradients which can hamper the convergence from the beginning.

How the vanishing / exploding gradients problem were addressed before?
This problem has be solved mostly by using the normalized initialization and intermediate normalization layers which enable networks with tens of layers to start converging for SGD with back propagation.

What is a degradation problem and relation with the depth of the network?
With the network depth increasing, accuracy gets saturated and then degrades rapidly. This degradation is not caused by overfitting and adding more layers to a suitable deep model also leads to higher training errors.

What is the approach used by the research paper?
Instead of making each few stacked layers directly fit a desired underlying mapping, these layers are explicitly made to fit a residual mapping.

What is the mathematical explanation for the approach?
If the desired underlying mapping is H(x), then the stacked non-linear layers fit another mapping of F(x) : = H(x) - x. So the original mapping is recast into F(x)+x.



What is meant by the shortcut connections and how they will be useful?
The formulation of F(x)+x can be done by feed forward neural networks with shortcut connections.

These shortcut connections skip one or more layers.
In the deep residual mapping approach the shortcut connections perform identity mapping and their outputs are added to the outputs of the stacked layers. These identity shortcut connections won't add extra parameters or computational complexity.