Hyomin Seo

Title : Understanding Rectifiers

<u>Explain how the main technique from ``Delving Deep into Rectifiers'' may alleviate a core issue in the training scheme for ``Going Deeper with Convolutions.''</u>

Core issues in 'Going Deeper with Convolutions'

Improving the performance of deep neural networks = increasing size and depth.

1. Increased size leads to larger number of parameters, making the network inclined to overfitting, creating major bottleneck between labeled examples and limited training set.
2. Increased size leads to dramatically increased use of computational resources ; waste of computation. (for 2 linked convolutional layers, increase in the number of their fitlers results in quadratic increase of computation.)

Definition of Rectifiers :

Rectified activation units : Rectified LInear activation function (Relu) :
Piecewise linear function that will output the input directly if it is positive, otherwise it will output zero.
Instead of increasing the dimension and depth of the system, applicating PReLU (Parametric Rectified Linear Unit) can improve the model fitting with nearly zero extra computation cost and small overfitting risk - alleviating the core issues of increasing dimensions and depth in order to increase the performance of the system.

<u>Explain the core effect of pre-activation from ``Identity Mappings in Deep Residual Networks'' compared to the original residual formulation.</u>

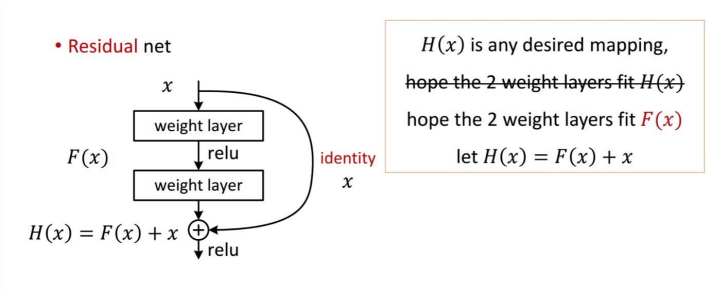


Illustration on Residual set.

Residual network
Optimizing F(x), the residual mapping will result in getting desired H(x).

Problem : Rapid degradation
If the depth of the network deepens, the accuracy gets saturated and degrades rapidly.

Continued

Residual unit : $y_l = h(x_l) + F(x_l, W_l)$

$x_{l+1} = f(y_l)$

↳ identity mapping

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

$\Rightarrow x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i)$

↳ Sum. residual function

$\frac{\partial \varepsilon}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L} \frac{\partial x_L}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L}\left(1 + \frac{\partial}{\partial x_l}\sum_{i=1}^{L-1} F(x_i, W_i)\right)$ ε as Loss

direct propagation — propagation through multiple weight layers.

Question : Does f identity mapping yields the best performance?
the paper proves that setting f as identity map composed of ReLU and BN without intercepting shortcut path.

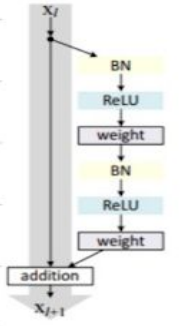
Full pre activation

$\rightarrow y_{l+1} = f(y_l) + F(f(y_l), W_{l+1})$

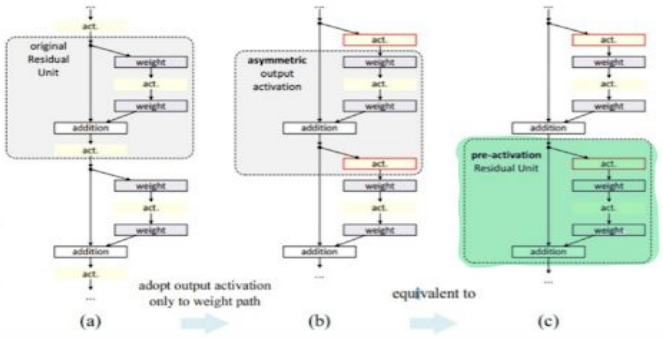
impose activation only on F by introducing $\hat{f}$

$\rightarrow y_{l+1} = f(y_l) + F(\hat{f}(y_l), W_{l+1})$

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



) full pre-activation



The resulting performance of system with full pre activation shows that the accuracy on training set is lower than that of the original model, the overall accuracy on the test was evidently higher.

- optimizing was achieved easily on full pre activation model
- Regularization is done by BN in full pre activation model.