Residual Neural Network

Why Deep Residual Learning?

- “Degradation of Accuracy” Problem
    - Accuracy gets saturated then decreases rapidly with the depth increase
    - The problem is not caused by overfitting
    - Adding more layers to a suitably deep model leads to higher training error
- Not all systems are similarly easy to optimize

Deep Residual Learning Network in Short

- Let the stacked layers fit a residual mapping
- Feedforward Neural Networks with “Shortcut connections”
    - Skipping one or more layers
    - These connections perform identity mapping
    - Their outputs are added to the outputs of the stacked layers
- Results from the experiments on the paper:
    - Deep residual nets are easy to optimize, but the counterpart “plain” nets (that simply stack layers) exhibit higher training error when the depth increases
    - Deep residual nets can easily enjoy accuracy gains from greatly increased depth, producing results substantially better than previous networks.

Related Works

- Residual Representations
  - VLAD and Fisher Vector are powerful shallow representations for image retrieval and classification.
  - Encoding residual vectors is shown to be more effective than encoding original vectors.
- Shortcut Connections
  - An early practice of training multi-layer perceptrons (MLPs) is to add a linear layer connected from the network input to the output. Few intermediate layers are directly connected to auxiliary classifiers for addressing vanishing/exploding gradients.
  - An “inception” layer is composed of a shortcut branch and a few deeper branches
  - “highway networks” present shortcut connections with gating functions [15].
  -

Residual Learning Reformulation

- **H(x)** is an underlying mapping to be fit by a few stacked layers.
- The hypothetical residual function **F(x):= H(x) − x**
- The original function becomes **F(x)+x**
- This reformulation is motivated by the counterintuitive phenomena about the degradation problem.
- The degradation problem suggests that the solvers might have difficulties in approximating identity mappings by multiple nonlinear layers.
- With the residual learning reformulation, if identity mappings are optimal, the solvers may simply drive the weights of the multiple nonlinear layers toward zero to approach identity mappings.

Residual Learning

- In real cases, it is unlikely that identity mappings are optimal, but reformulation may help to precondition the problem.
- If the optimal function is closer to an identity mapping than to a zero mapping, it should be easier for the solver to find the perturbations with reference to an identity mapping than to learn the function as a new one.
- The learned residual functions in general have small responses, suggesting that identity mappings provide reasonable preconditioning.

Identity Mapping by Shortcuts

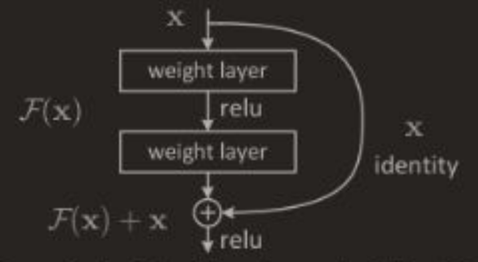

Figure 2. Residual learning: a building block.

- Residual learning to every few stacked layers.
- Building block formula: **y = F(x, {Wi})+ x. (1)**
  - x and y are the input and output vectors of the layers.
  - The function F(x, {Wi}) represents the residual mapping to be learned.
  - For the example in Fig. 2 that has two layers, F = W2σ(W1x) in which σ denotes activation function
  - The biases are omitted for simplifying notations.
  - The operation F + x is performed by a shortcut connection and element-wise addition.
  - We adopt the second nonlinearity after the addition (i.e., σ(y), see Fig. 2).
  - The shortcut connections in Equation (1) introduce neither extra parameter nor computation complexity.

Identity Mapping by Shortcuts (Continued)

- The dimensions of x and F must be equal in Eqn.(1). If this is not the case (e.g., when changing the input/output channels), we can perform a linear projection Ws by the shortcut connections to match the dimensions: **y = F(x, {Wi})+Wsx. (2)**
- Square matrix Ws can also be used in Eqn.(1). But the paper argued the identity mapping is sufficient for addressing the degradation problem and is economical, and thus Ws is only used when matching dimensions.
- The form of the residual function F is flexible. The function F which represents more than three layers are possible. If F has only a single layer, Eqn.(1) is similar to a linear layer: y = W1x+x with no observed advantages yet.
- Although the above notations are about fully-connected layers for simplicity, they are applicable to convolutional layers. The function F(x, {Wi}) can represent multiple convolutional layers. The element-wise addition is performed on two feature maps, channel by channel.

Read More

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). DOI:10.1109/cvpr.2016.90. (Alternate link: https://arxiv.org/pdf/1512.03385.pdf.)