Introduction to Residual Neural Network (ResNet)

Lu Lu

Jan 25, 2019 @Crunch Seminar

Overview

- Background
- Residual neural network
- Variants of residual blocks
- Some analysis

Why Deep Neural Networks?

- Shallow NNs (single hidden layer)
  - [1]Universal approximation theorem (uniformly)
  - [2]$\epsilon^{-d/n}$ neurons can approximate any C^n-function on a compact set in $\mathbb{R}^d$ with error ϵ
- Deep NNs: better than shallow NNs (of comparable size)
  - [3]Exists a function expressible by a 3-layer NN, which cannot be approximated by any 2-layer network (unless exponentially large)
  - [4]$\frac{\text{size}_{\text{deep}}}{\text{size}_{\text{shallow}}} \approx \epsilon^d$

---

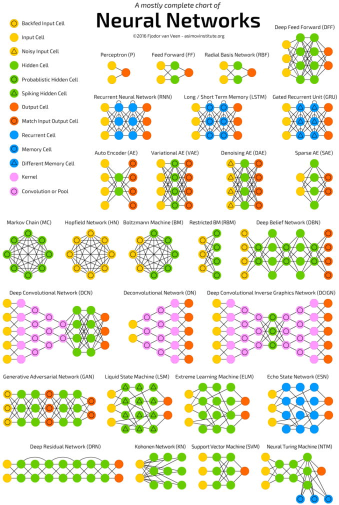
[1]Cybenko, Math. Control Signals Syst., 1989; Hornik et al., Neural Netw., 1989.

[2]Mhaskar, Neural Comput., 1996.

[3]Eldan & Shamir, COLT, 2016.

[4]Mhaskar & Poggio, Anal. Appl., 2016; Mhaskar et al., AAAI, 2017; Poggio et al., IJAC, 2017.

“In theory, theory and practice are the same. In practice, they are not.” — Albert Einstein?



Overview

- ~~Background~~
- Residual neural network
- Variants of residual blocks
- Some analysis

Residual Neural Network[5]

- Deep networks are hard to train: vanishing gradients
- Core idea: "identity shortcut connection" that skips one or more layers
- Widely used: simple & powerful

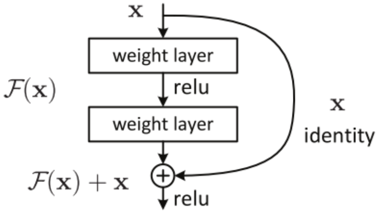
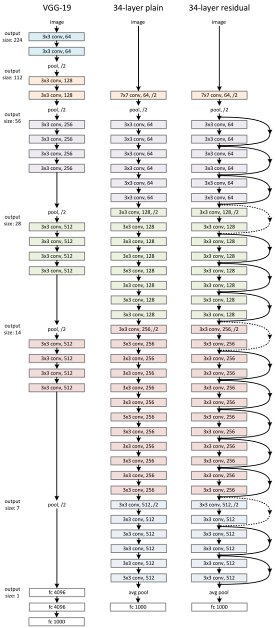


Figure 1: A residual block

[5] He et al., CVPR, 2016.

Residual Neural Network



Residual Neural Network

- $\mathcal{H}(\mathbf{x}) = \mathcal{F}(\mathbf{x}) + \mathbf{x} \Rightarrow \mathcal{F}(\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \mathbf{x}$ (residual)
- *Hypotheses*: the residual may be an easier function to fit

$\mathcal{F}$?

- If $\mathcal{F}$ has two layers, $\mathcal{F}(\mathbf{x}) = W_2\sigma(W_1\mathbf{x})$
- If $\mathcal{F}$ has one layers, $\mathcal{F}(\mathbf{x}) = W_1\mathbf{x}$,
  $\mathcal{H}(\mathbf{x}) = W_1\mathbf{x} + \mathbf{x} = (W_1 + 1)\mathbf{x}$. No advantage!

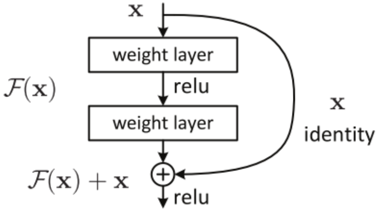
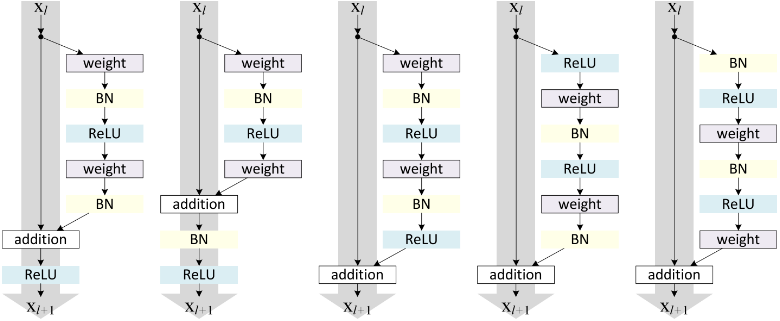


Figure 2: A residual block

Overview

- ~~Background~~
- ~~Residual neural network~~
- Variants of residual blocks
- Some analysis

Variants of Residual Blocks[6]



(a) original (b) BN after addition (c) ReLU before addition (d) ReLU-only pre-activation (e) **full pre-activation**

[6] He et al., ECCV, 2016.

Variants of Residual Blocks

Accuracy can be gained more efficiently by increasing the cardinality than by going deeper or wider.

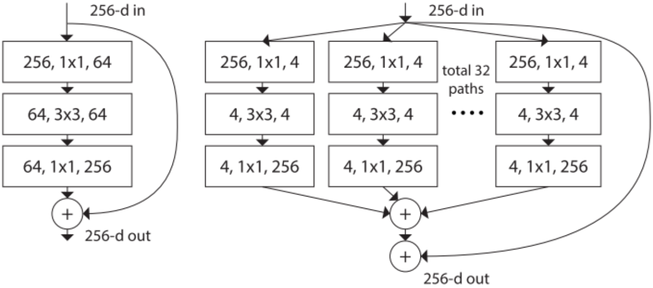


Figure 3: ResNeXt[7]: split-transform-merge

[7] Xie et al., CVPR, 2017.

Variants of Residual Blocks

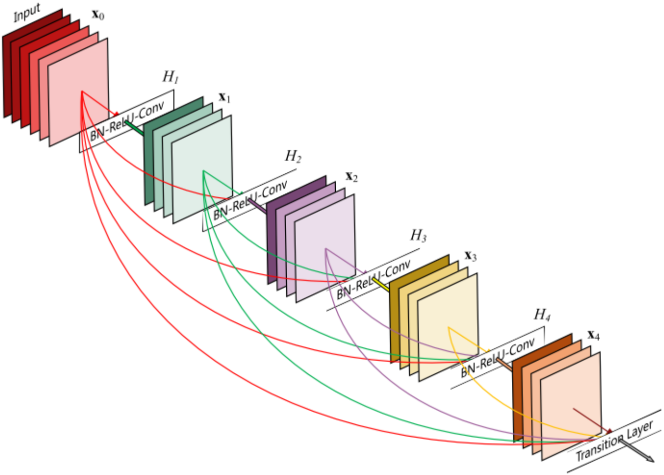


Figure 4: DenseNet[8]

[8] Huang et al., CVPR, 2017.

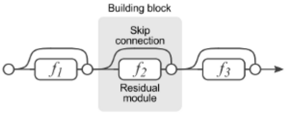
Overview

- ~~Background~~
- ~~Residual neural network~~
- ~~Variants of residual blocks~~
- Some analysis

Unraveled View[9]

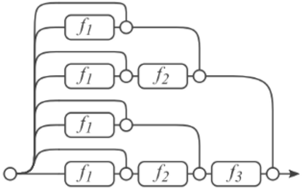
$$
\begin{aligned}
y_3 &= y_2 + f_3(y2) \\
&= [y_1 + f_2(y_1)] + f_3(y_1 + f_2(y_1)) \\
&= [y_0 + f_1(y_0) + f_2(y_0 + f_1(y_0))] + f_3(y_0 + f_1(y_0) + f_2(y_0 + f_1(y_0)))
\end{aligned}
$$

2^n paths connecting input to output layers



(a) Conventional 3-block residual network

=

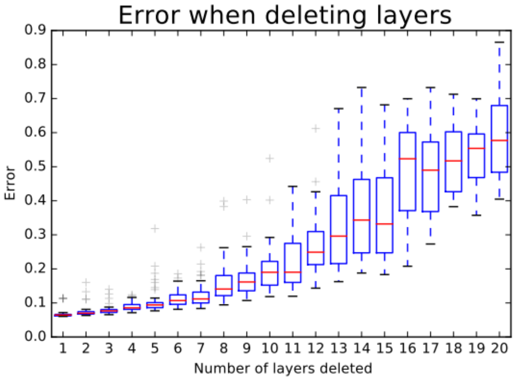


(b) Unraveled view of (a)

[9] Veit et al., NIPS, 2016.

Ensemble-like Behavior[10]

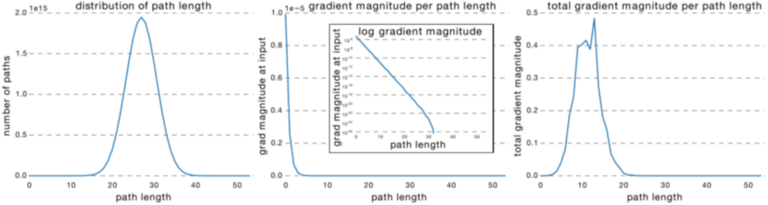
- Lesion study: randomly deleting several modules
- Paths do not strongly depend on each other



[10] Veit et al., NIPS, 2016.

Vanishing Gradients?[11]

- The effective paths are relatively shallow
- Only the short paths contribute gradients
- ResNet does not resolve vanishing gradients ~~by preserving gradient flow throughout the entire network~~. Rather, they enable very deep networks by *shortening the effective paths*.



(a) (b) (c)

[11] Veit et al., NIPS, 2016.

Universal Approximation

Recall[12]:

- To approximate any continuous function $[0,1]^d \to \mathbb{R}$ by ReLU NN: minimal width is $d+1$

ResNet with one hidden neuron:

$$\mathcal{H}(\mathbf{x}) = \mathbf{W}_2 \mathrm{ReLU}(\mathbf{W}_1 \cdot \mathbf{x} + b) + \mathbf{x}$$

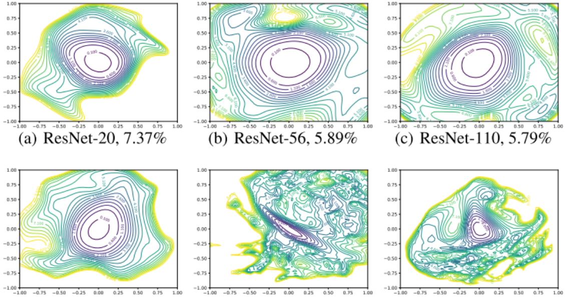
- Identity map (d dim) + one hidden neuron = $d+1$ units

ResNet with one neuron per hidden layer: universal approximation (in L^1) for any Lebesgue-integrable function as the depth $\to \infty$.[13]

[12] Hanin et al., arXiv, 2017.
[13] Lin & Jegelka, NIPS, 2018.

Why Easier to Train?

- 2D visualization of the loss surface by "filter normalization" method[14]
- BoostResNet[15]: a training algorithm (non-differentiable), training error decays exponentially with depth



(a) ResNet-20, 7.37% (b) ResNet-56, 5.89% (c) ResNet-110, 5.79%

(d) ResNet-20-NS, 8.18% (e) ResNet-56-NS, 13.31% (f) ResNet-110-NS, 16.44%

[14] Li et al., NIPS, 2018.

[15] Huang et al., ICML, 2018.

Overview

- ~~Background~~
- ~~Residual neural network~~
- ~~Variants of residual blocks~~
- ~~Some analysis~~

Cerebral Cortex

- Cajal Ramon (the father & the mother of modern neuroscience)
- Pyramidal cells (1888)