



Residual Networks

Outline

by Alexander Kowsik

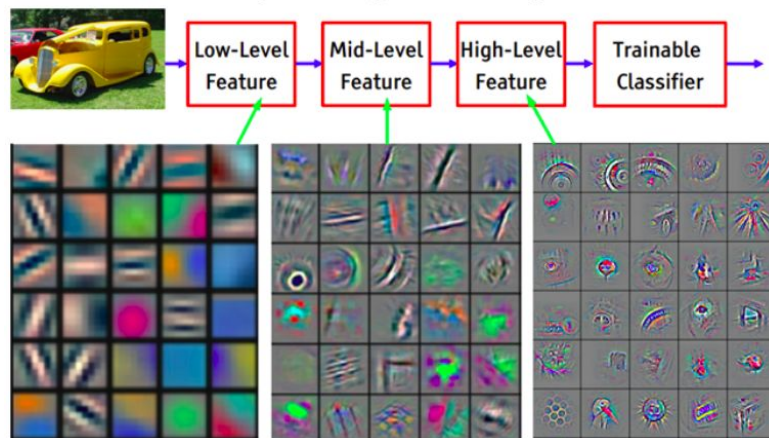
Outline

- **Introduction/Motivation**

- layers learn increasingly abstract features
- idea: more layers = better performance
- problem: vanishing/exploding gradients → at some point, networks not trainable anymore (25+ layer) → error rate increases
- solution: ResNets

- **Residual Networks**

- allows training for DEEP networks (up to 2000 layers)
- main idea: skip connections
- skip connections can be identity function
- rough idea: if layer hurts performance: “skip” it



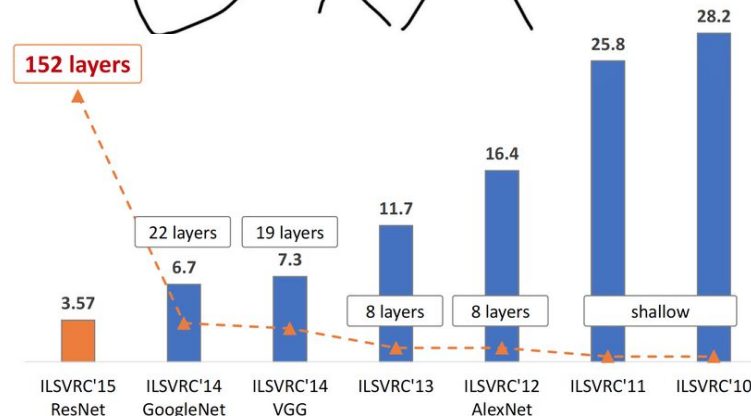
Outline

- Introduction/Motivation

- layers learn increasingly abstract features
- idea: more layers = better performance
- problem: vanishing/exploding gradients → at some point, networks not trainable anymore (25+ layer) → error rate increases
- solution: ResNets

- Residual Networks

- allows training for DEEP networks (up to 2000 layers)
- main idea: skip connections
- skip connections can be identity function
- rough idea: if layer hurts performance: “skip” it



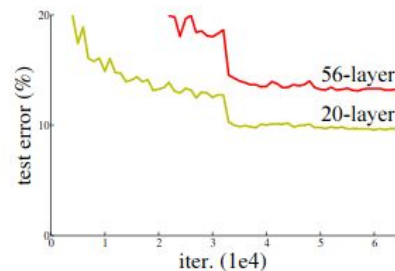
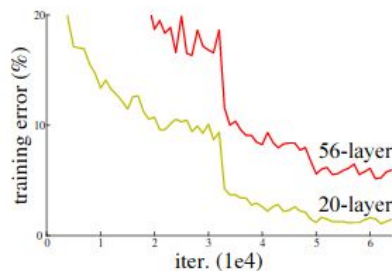
Outline

- Introduction/Motivation

- layers learn increasingly abstract features
- idea: more layers = better performance
- problem: vanishing/exploding gradients → at some point, networks not trainable anymore (25+ layer) → error rate increases
- solution: ResNets

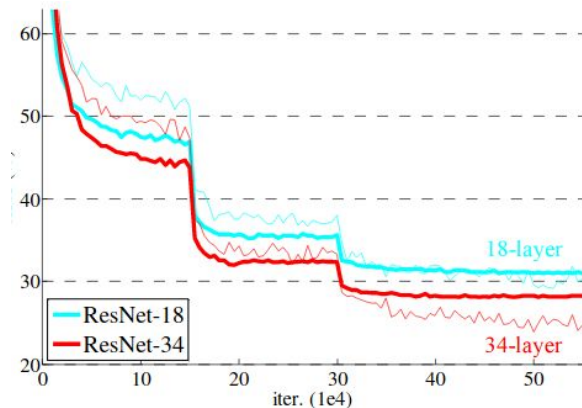
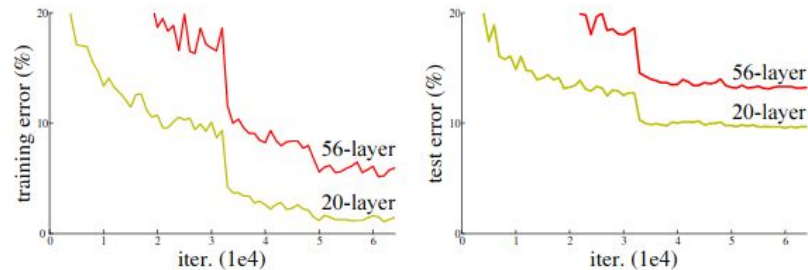
- Residual Networks

- allows training for DEEP networks (up to 2000 layers)
- main idea: skip connections
- skip connections can be identity function
- rough idea: if layer hurts performance: “skip” it



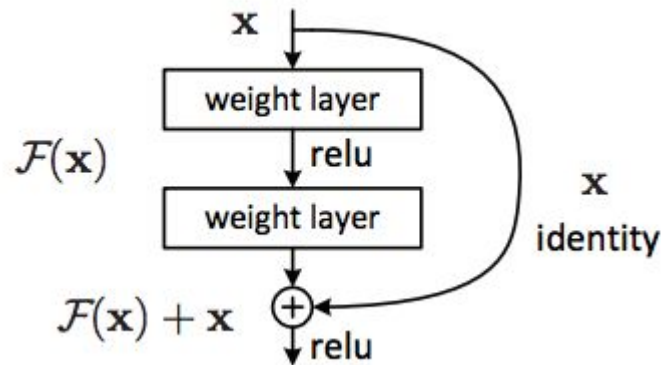
Outline

- Introduction/Motivation
 - layers learn increasingly abstract features
 - idea: more layers = better performance
 - problem: vanishing/exploding gradients → at some point, networks not trainable anymore (25+ layer) → error rate increases
 - solution: ResNets
- Residual Networks
 - allows training for DEEP networks (up to 2000 layers)
 - main idea: skip connections
 - skip connections can be identity function
 - rough idea: if layer hurts performance: “skip” it



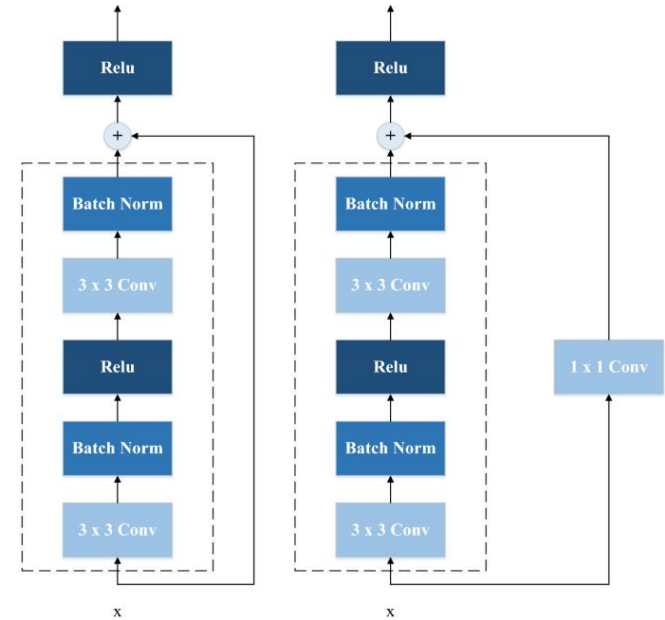
Outline

- Introduction/Motivation
 - layers learn increasingly abstract features
 - idea: more layers = better performance
 - problem: vanishing/exploding gradients → at some point, networks not trainable anymore (25+ layer) → error rate increases
 - solution: ResNets
- Residual Networks
 - allows training for DEEP networks (up to 2000 layers)
 - main idea: skip connections
 - skip connections can be identity function
 - rough idea: if layer hurts performance: “skip” it



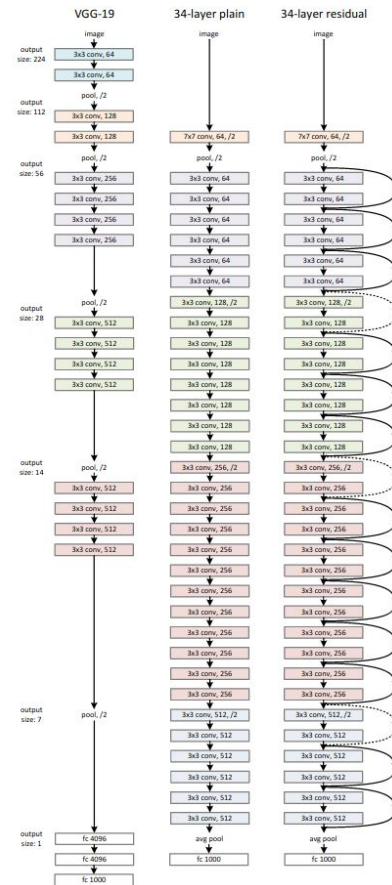
Outline

- **Structure/Variants**
 - structure in detail (ReLU before/after skip connection, ...)
 - skip distances
 - **dimension differences** + solutions: 1x1 convolutions, padding, pooling, ...
 - different network architectures (HighwayNets, DenseNets, ...)
- **ResNets Internals**
 - Forward propagation
 - Backprop
 - Training of ResNets: Gradient Descent, ...
 - advantages/disadvantages/performances
- **Summary/Conclusion**



Outline

- Structure/Variants
 - structure in detail (ReLU before/after skip connection, ...)
 - skip distances
 - **dimension differences** + solutions: 1x1 convolutions, padding, pooling, ...
 - different network architectures (HighwayNets, DenseNets, ...)
- ResNets Internals
 - Forward propagation
 - Backprop
 - Training of ResNets: Gradient Descent, ...
 - advantages/disadvantages/performances
- Summary/Conclusion





Outline

- Structure/Variants
 - structure in detail (ReLU before/after skip connection, ...)
 - skip distances
 - **dimension differences** + solutions: 1x1 convolutions, padding, pooling, ...
 - different network architectures (HighwayNets, DenseNets, ...)
- ResNets Internals
 - Forward propagation
 - Backprop
 - Training of ResNets: Gradient Descent, ...
 - advantages/disadvantages/performances
- Summary/Conclusion

During **backpropagation** learning for the normal path

$$\Delta w^{\ell-1,\ell} := -\eta \frac{\partial E^\ell}{\partial w^{\ell-1,\ell}} = -\eta a^{\ell-1} \cdot \delta^\ell$$

and for the skip paths (nearly identical)

$$\Delta w^{\ell-2,\ell} := -\eta \frac{\partial E^\ell}{\partial w^{\ell-2,\ell}} = -\eta a^{\ell-2} \cdot \delta^\ell.$$

Structure of the site

- site will be built with **pandoc**
- video at the top (**kdenlive?**)
 - slides with powerpoint
 - presentation of the topic
- rest of site: topic in detail
 - written elaboration
 - supportive graphics and diagrams (maybe animations or interactive elements if possible)
 - goal: something similar to *distill.pub* articles

Hello World

Second headline

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.



alt text