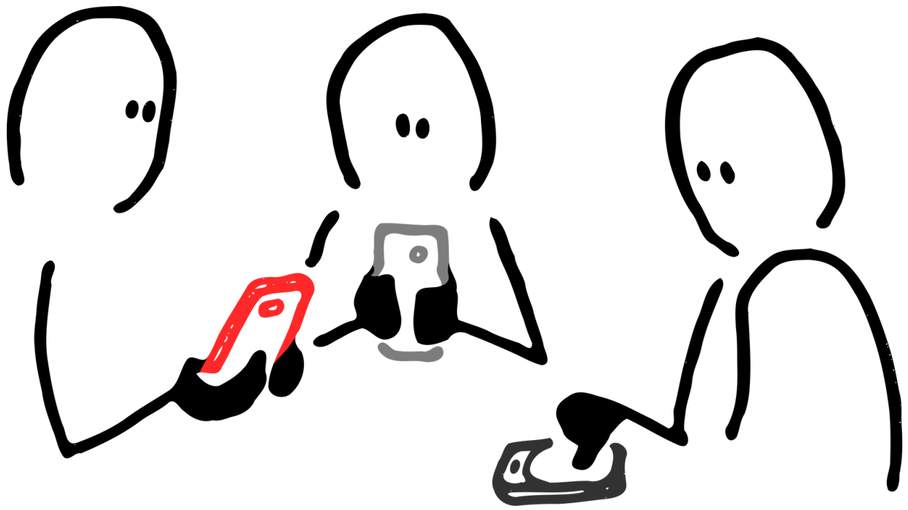
Deep Residual Learning for Image Recognition

Kaiming He , Xiangyu Zhang, Shaoqing Ren and Jian Sun

DADS6003: MACHINE LEARNING



Team Member

1. 6710422001 Sirima Pangpradang
2. 6710422002 Seriphap Siangnok
3. 6710422013 Voranitha Chaiaroon

CONTENT

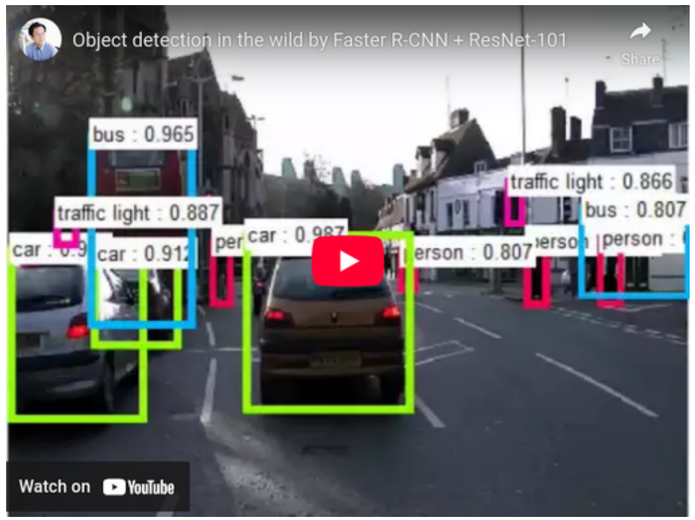
- Introduction
- Related Work
- Deep Residual Learning
- Experiments
- Conclusion

1. Introduction

Real-World Use

RESNET

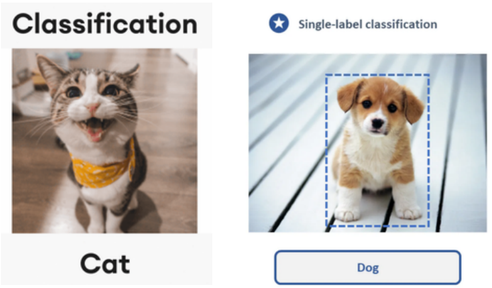
1 Self-driving Cars



Camera captures scene → Detects pedestrians, cars, traffic signs

Extracts visual features → Aids object detection

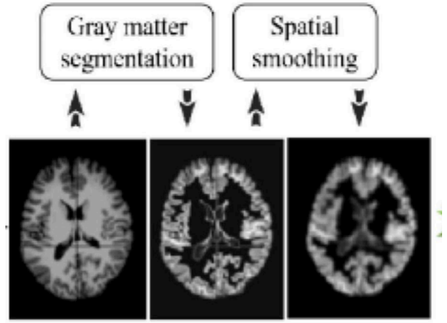
2 Smart Photo Search



Google Photos classifies images (cat, people, beach)

Categorizes images based on features

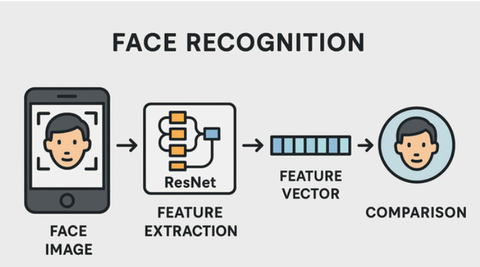
3 Medical Imaging



Analyzes X-ray/CT → Assesses disease risk

Learns from labeled medical images

4 Face Unlock



Smartphone scans face → Converts to vector → Compares with stored profile

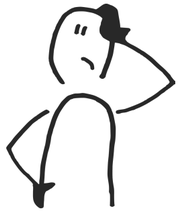
Identifies unique facial features

5 Warehouse System

Camera scans shelf → Counts inventory

Detects items → Notifies low stock

Introduction –
Why Residual Learning ?

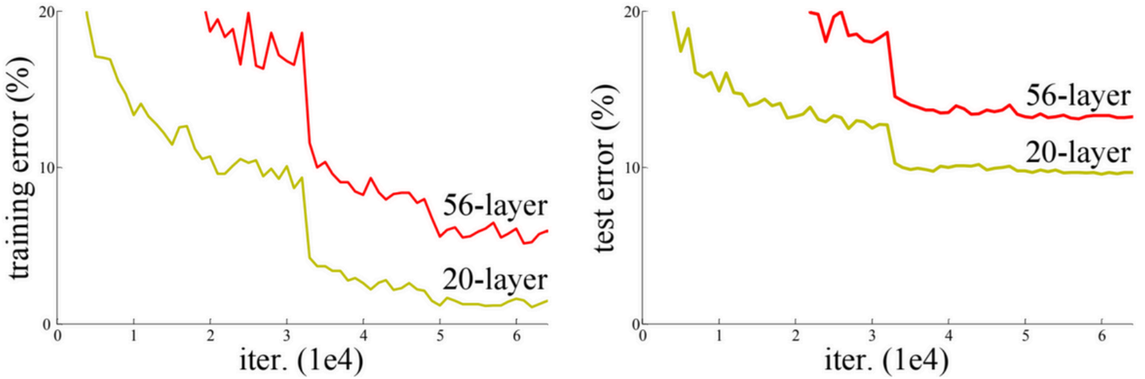


Deep Neural Networks Problem

In deep learning, it seems logical that the deeper a model is the smarter it should be, right?
But in reality, the opposite can happen —

The deeper the model, the worse the accuracy becomes!

This is because of the ***'vanishing gradient'*** problem and the ***difficulty*** in training very deep models.

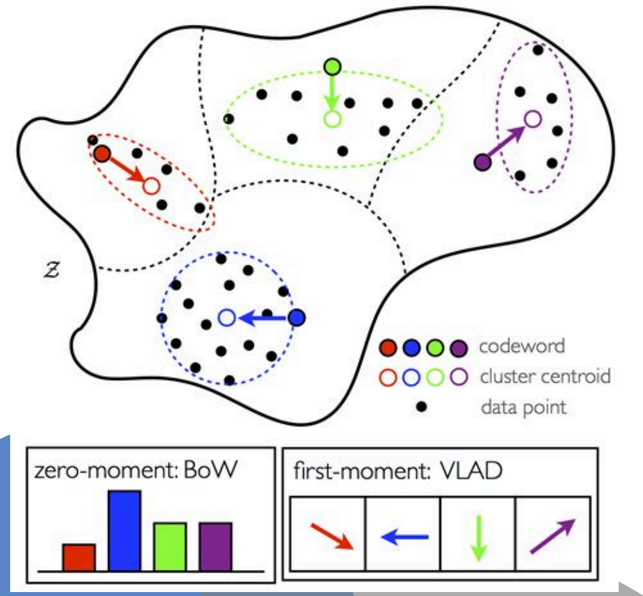


Picture: Graph compare error between network 20 and 56 yalers

Related Work – Residual Representations

1.Residual Representations Concept

- Think of it like learning **'what's missing'** from the original.
- Instead of learning the final answer, the model learns the **'difference'.**
- Helps improve accuracy and reduce computation, especially in image tasks.



🧠 ***In Image Recognition:***

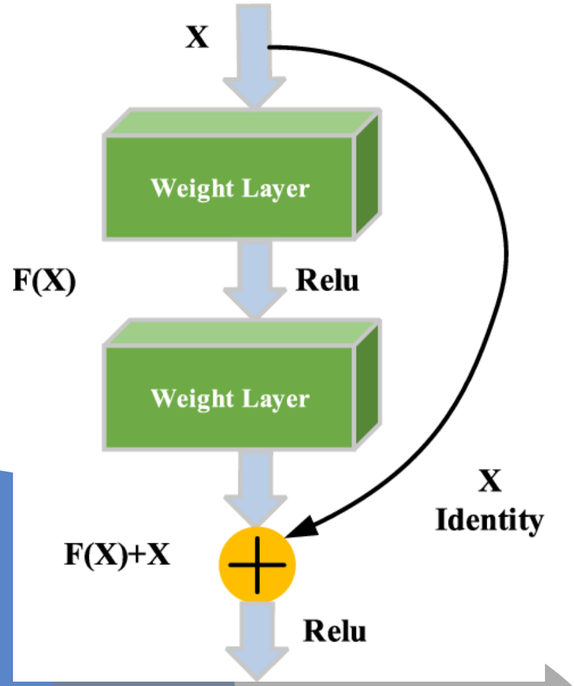
- Models (like VLAD or Fisher Vector) don't memorize the whole image.
- They remember the parts that make the image unique (the residual).
- This improves accuracy and efficiency.

VLAD (Vector of Locally Aggregated Descriptors)

Related Work – Shortcut Connections

2.Shortcut Connections concept

- A direct path from early layers to deeper ones (shortcut = identity mapping).
- Helps retain past knowledge while learning new features.



- Learn new info + keep the old
- Combine: Output = F(x) + x
- Pass to next layer → deeper learning without forgetting

"Shortcut connections are like memory highways in AI – they help ***the model remember what it already knows and build on top without starting over***"

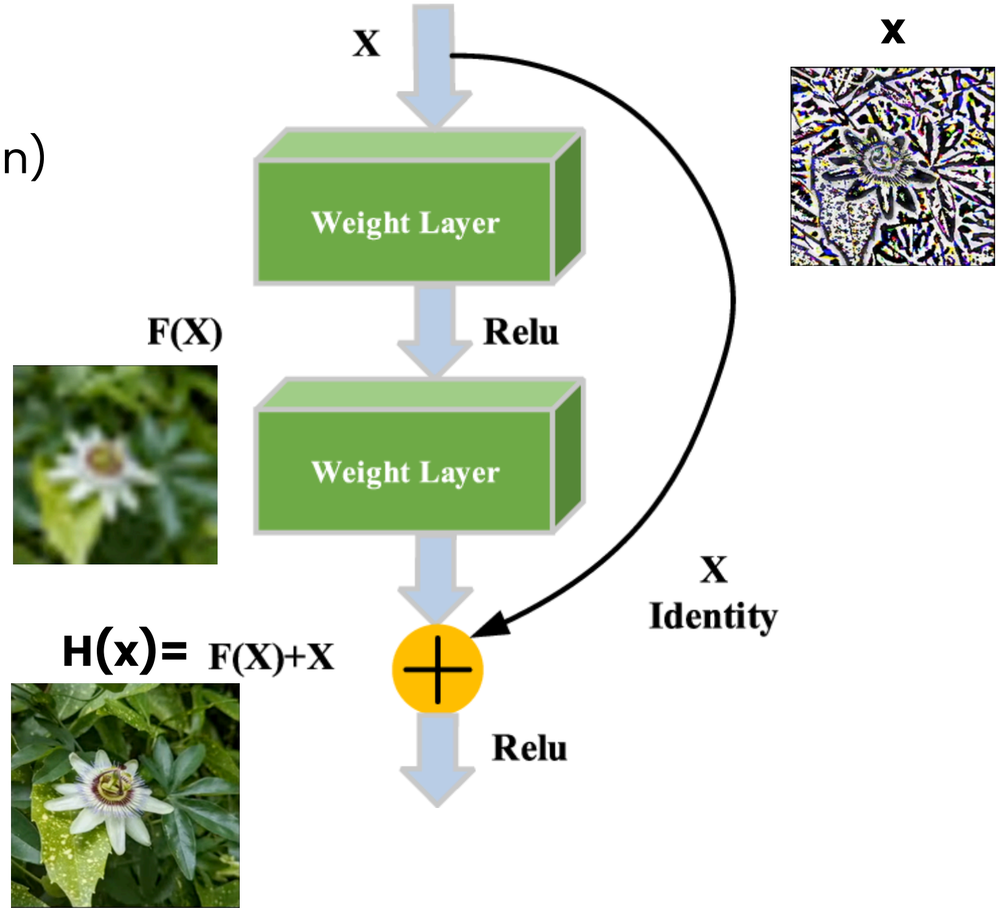
3. Deep Residual Learning

3.1 Residual Learning

📌 Instead of learning everything from scratch, the model learns only what needs to be added.

Symbols in the Image:
H(x) = The final output the model aims to learn (Target function)
F(x) = The model learns to add or improve (Residual)
x = The original input or existing knowledge (Identity/Input)

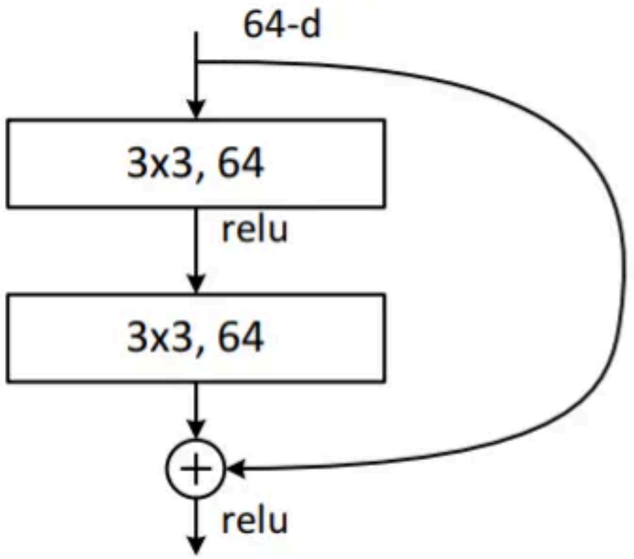
$$H(x) = F(x) + x$$



3.2. Identity Mapping by Shortcuts

Consider a building block defined as:

$$y = F(x, \{W_i\}) + x$$

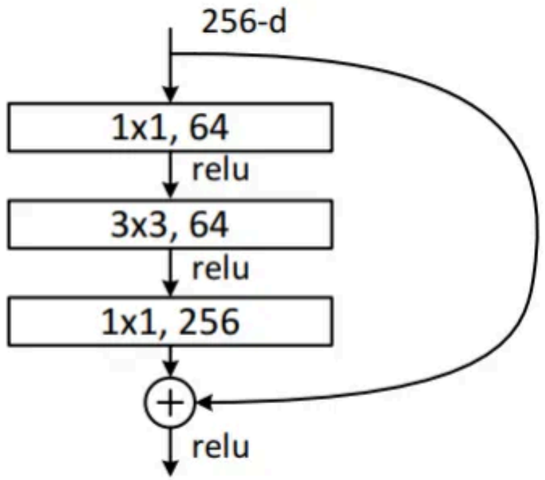


x : input vector
y : output vector
$F(x, \{W_i\})$: the residual mapping to be learned
(includes convolutions, batch normalization, and ReLU activation)

3.2. Identity Mapping by Shortcuts

If this is 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:

$$y = F(x, \{W_i\}) + W_s x.$$



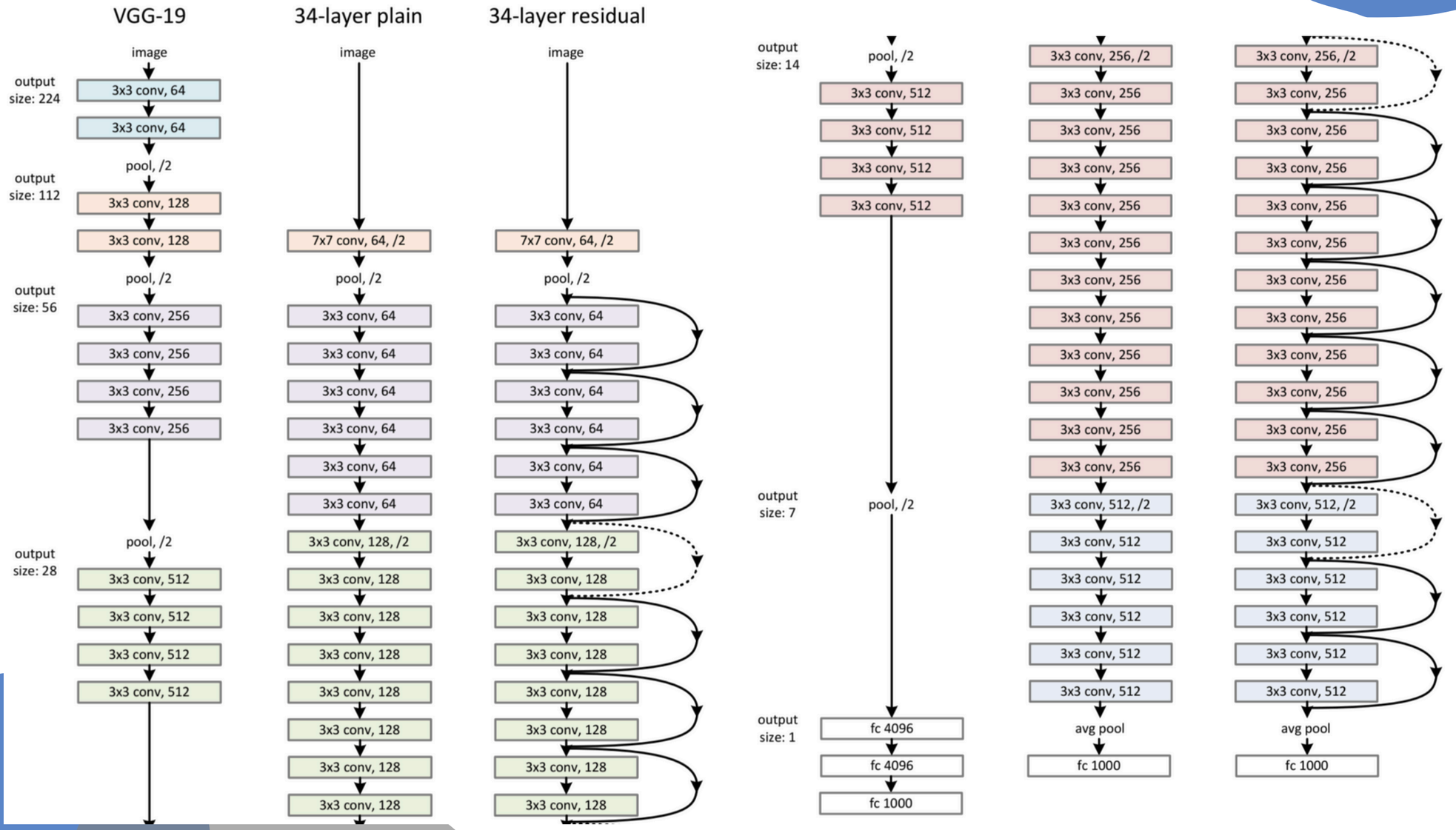
x : input vector

y : output vector

$F(x, \{W_i\})$: the residual mapping to be learned

W_s : A learnable projection matrix
(the shortcut connections to match the dimensions)

3.3. Network Architectures



3.4. Implementation

1. Data Preprocessing

- Image Resizing
- Image Cropping
- Color Augmentation
- Per-pixel Mean Subtraction

3.4. Implementation

2. Model Training

- Batch Normalization (BN)
- Weight Initialization
- Training Procedure

3. Model Testing

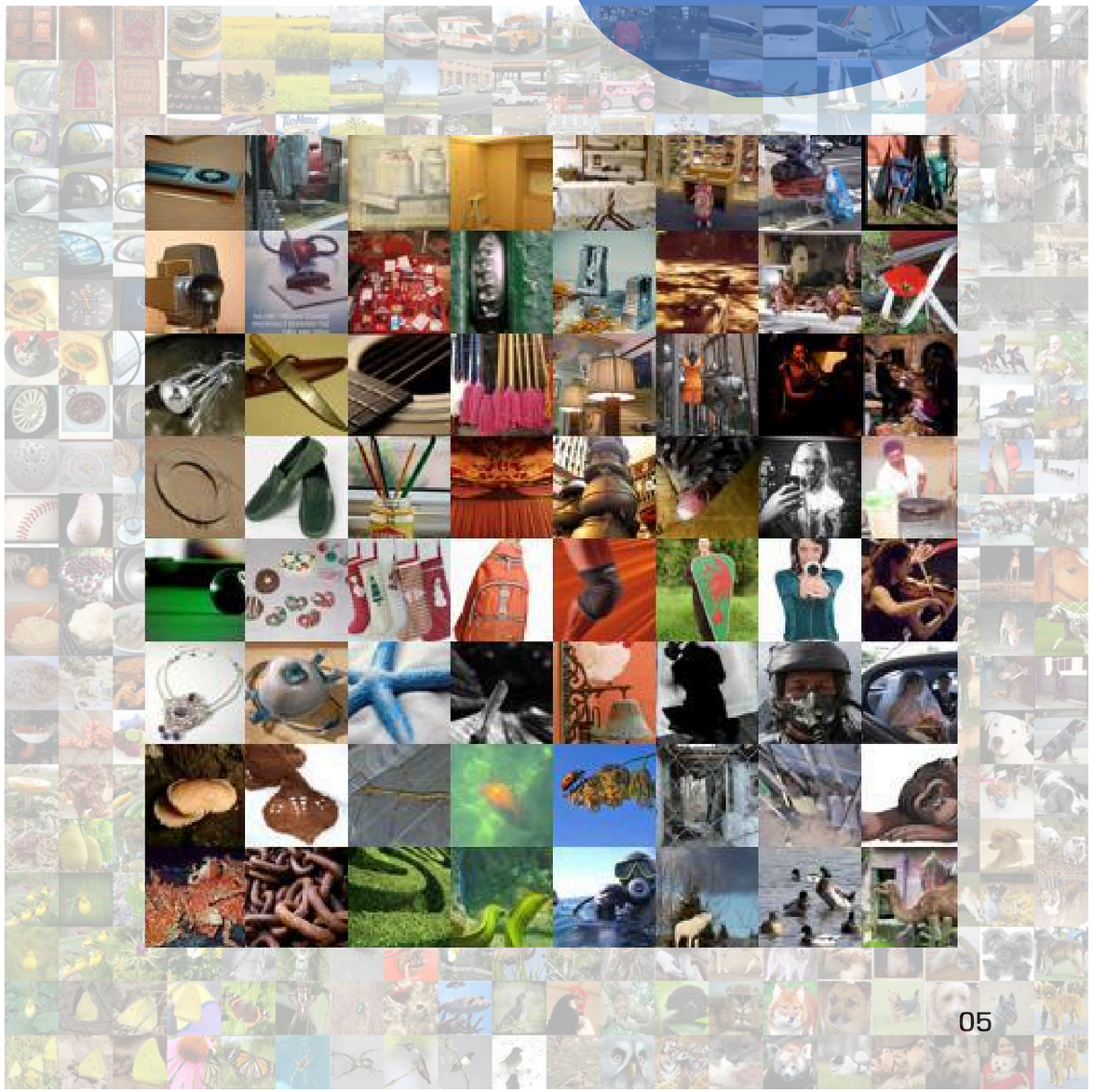
- Standard 10-crop testing for comparisons.
- Best results : fully-convolutional form and average the scores at multiple scales.

4.Experiments

4.1.ImageNet Classification

Evaluate with the ImageNet2012 classification dataset, that consists of 1000classes

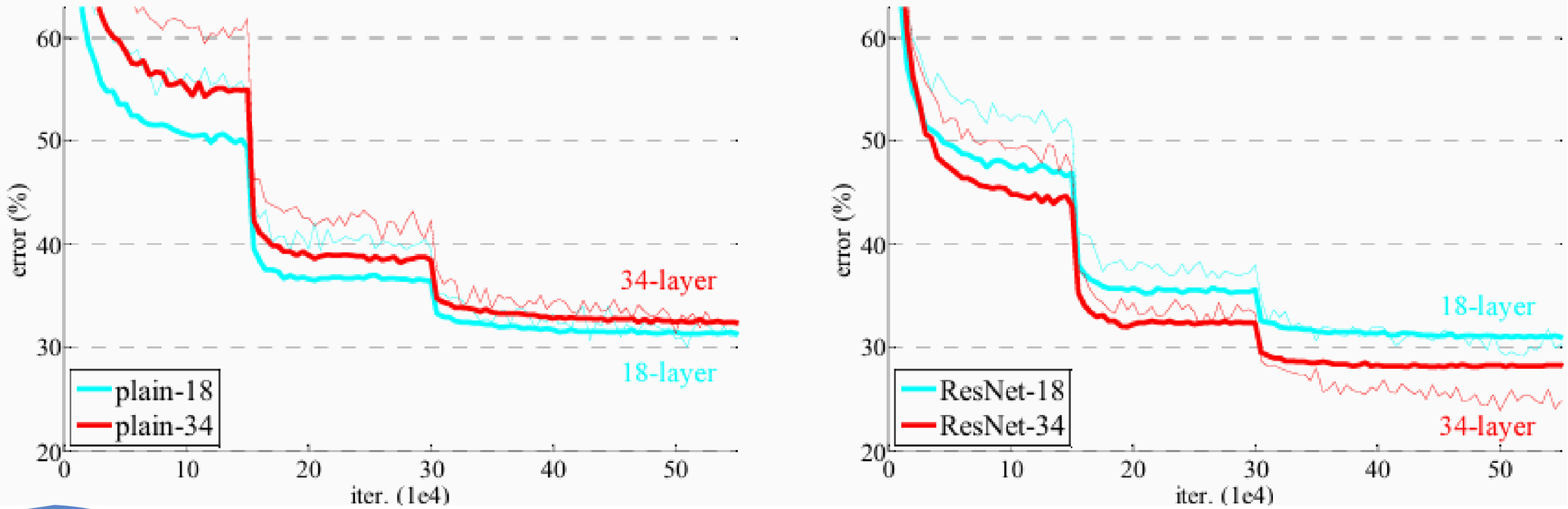
- Trained: 1.28million images
- Validated: 50k images
- Tested 100k images



4.1.ImageNet Classification

Result

- Deep ResNets can be trained without difficulties
- Deeper ResNet have lower training error, and also lower test error



4.2.CIFAR-10 and Analysis

Evaluate with the CIFAR-10 dataset, that consists of 60,000 small 32×32 color images categorized into 10 different classes, including animals and vehicles

Sample of CIFAR-10 dataset (Class and Name)

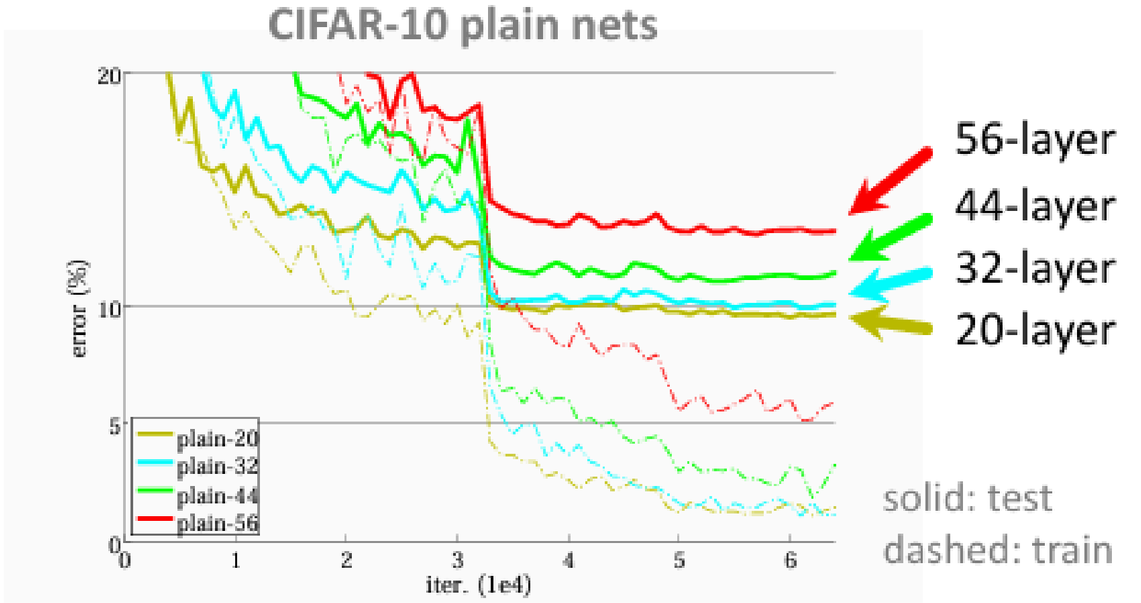
First 25 images of the dataset

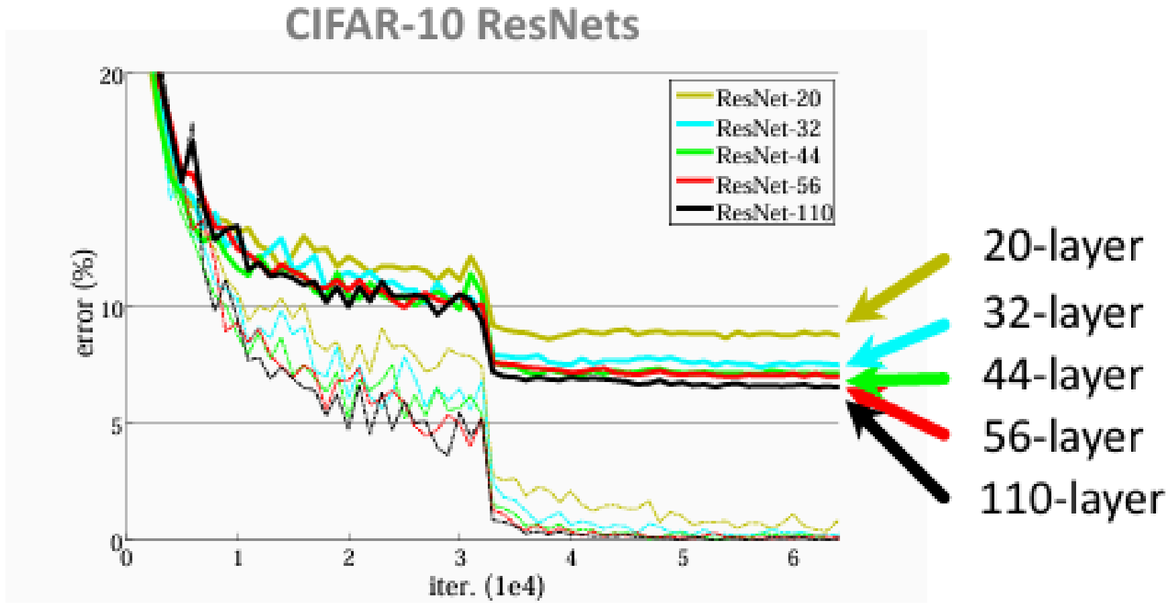
First 25 images with class names

4.2.CIFAR-10 and Analysis

Result

- Deep ResNets can be trained without difficulties
- Deeper ResNet have lower training error, and also lower test error

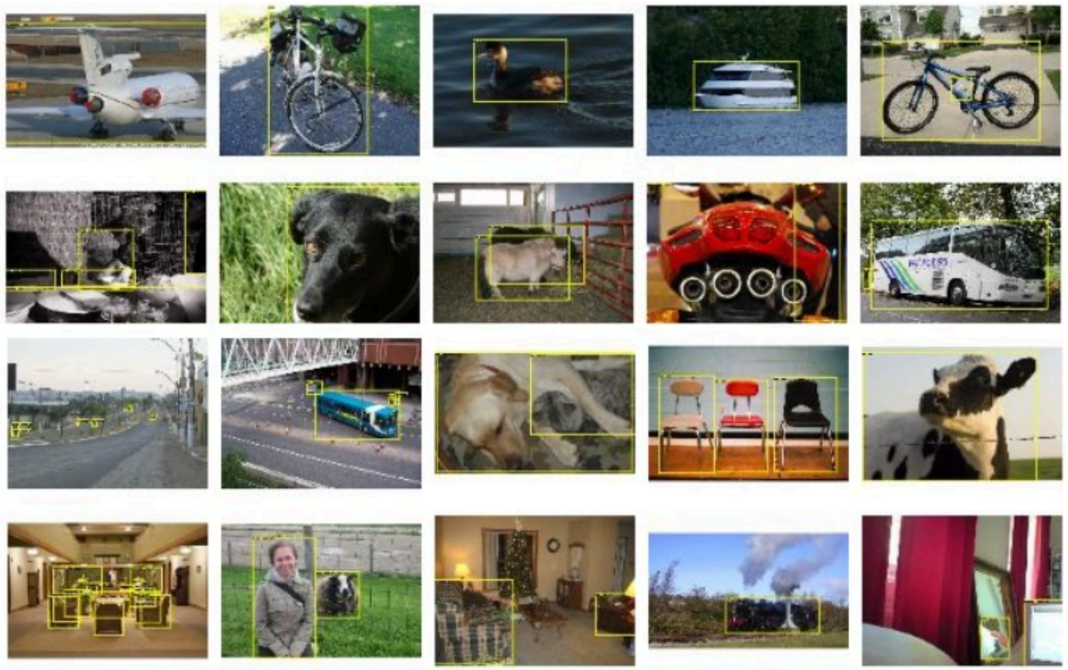




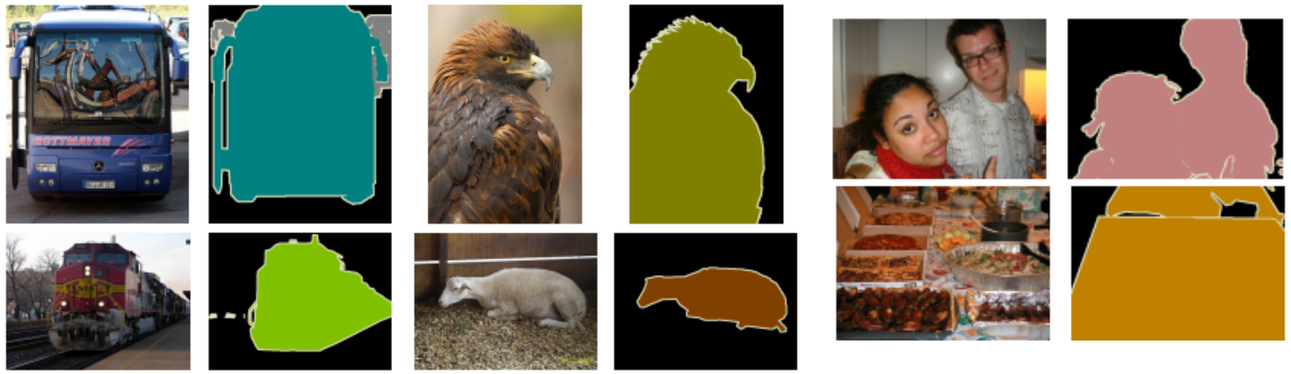
4.3.Object Detectionon PASCAL and MS COCO

PASCAL VOC and MS COCO datasets contain real images along with their corresponding bounding boxes and segmentation masks for object detection and segmentation tasks.

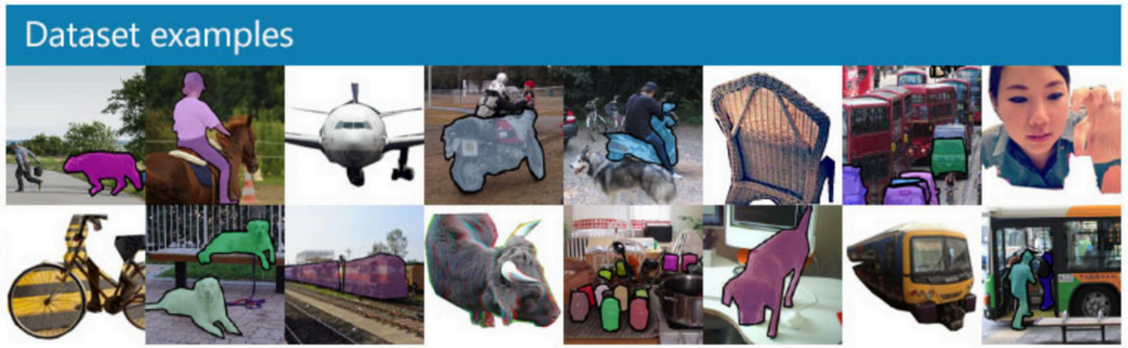
bounding boxes



segmentation masks



bounding boxes and segmentation masks



4.3.Object Detectionon PASCAL and MS COCO

Comparison of Models

- **Faster R-CNN** is used as the base detection model for both **VGG-16** and **ResNet-101**

PASCAL VOC dataset

- VOC 2007: ResNet-101 achieves **76.4%** mAP, outperforming VGG-16 (**73.2%** mAP).
- VOC 2012: ResNet-101 scores **73.8%**, while VGG-16 gets **70.4%**.

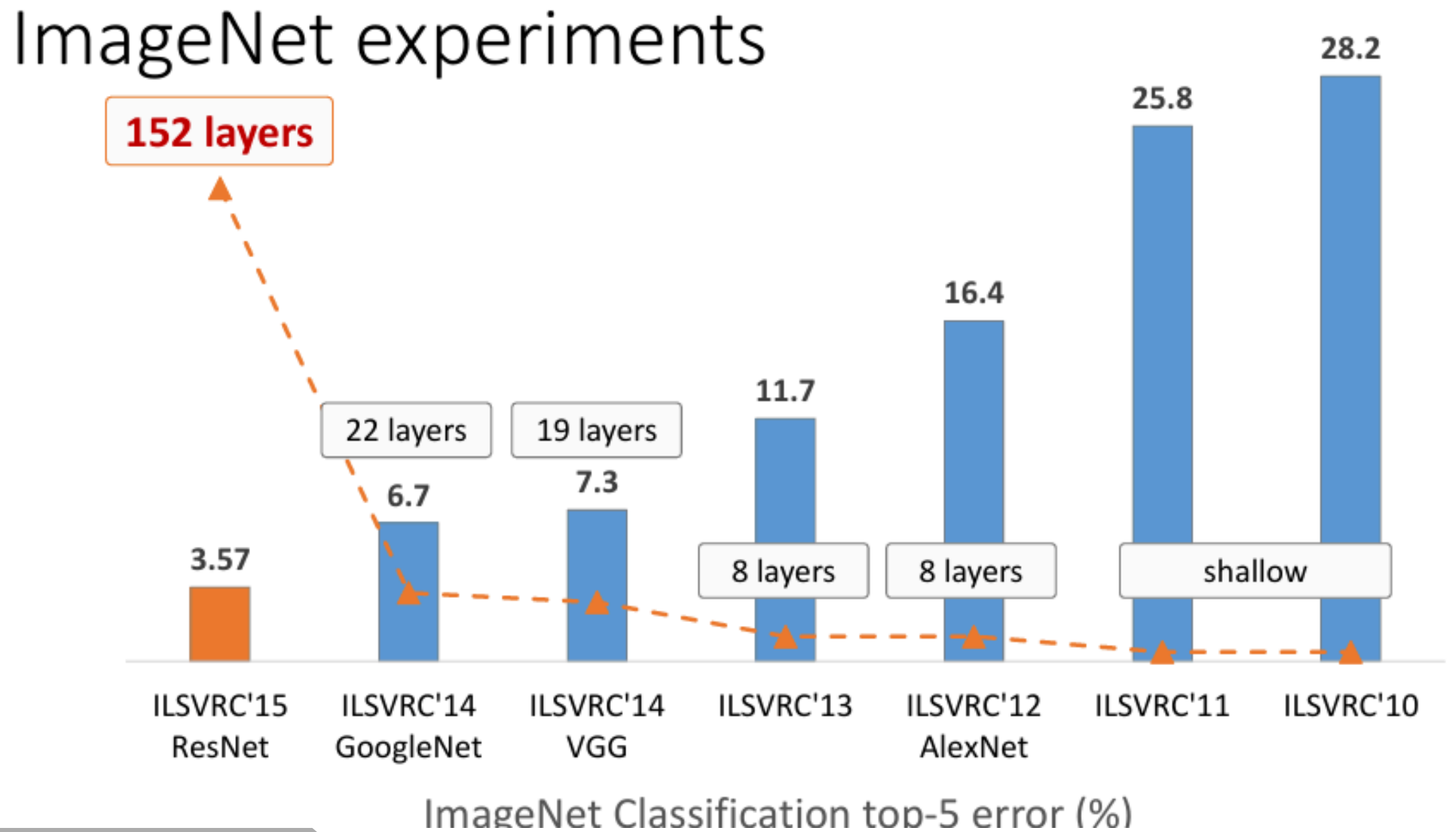
MS COCO dataset

- ResNet-101 boosts mAP@[.5, .95] by 6.0%, a **28% relative improvement** over VGG-16.

5.Conclusion

5.Conclusion

Deeper is still better – Increasing network depth improves performance, as ResNets mitigate optimization challenges, leading to better feature learning.



5.Conclusion

Features matter – Well-learned features from deep networks transfer effectively across tasks, enhancing detection and segmentation.

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

ResNet-101

task	2nd-place winner	MSRA	margin (relative)
ImageNet Localization (top-5 error)	12.0	9.0	**27%**
ImageNet Detection (mAP@.5)	53.6	62.1	**16%**
COCO Detection (mAP@.5:.95)	33.5	37.3	**11%**
COCO Segmentation (mAP@.5:.95)	25.1	28.2	**12%**

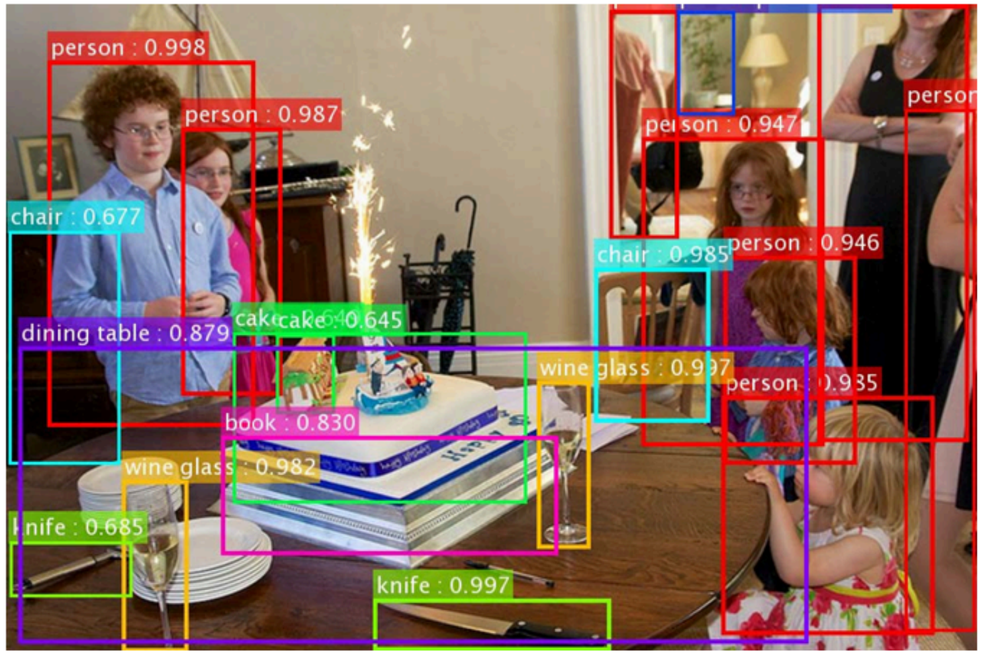
absolute 8.5% better!

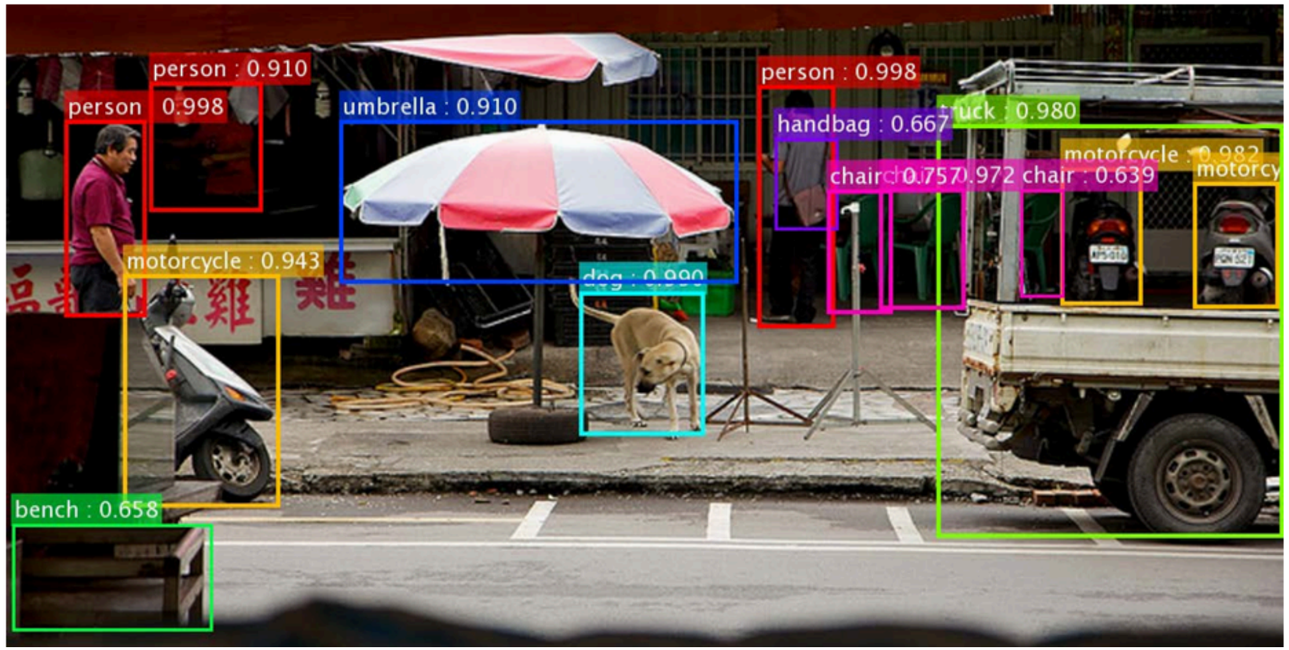
Mean Average Precision (mAP)

results are all based on ResNet-101 features are well transferrable

5.Conclusion

Faster R-CNN is just amazing – Faster R-CNN combined with ResNet achieves superior object detection accuracy and efficiency.

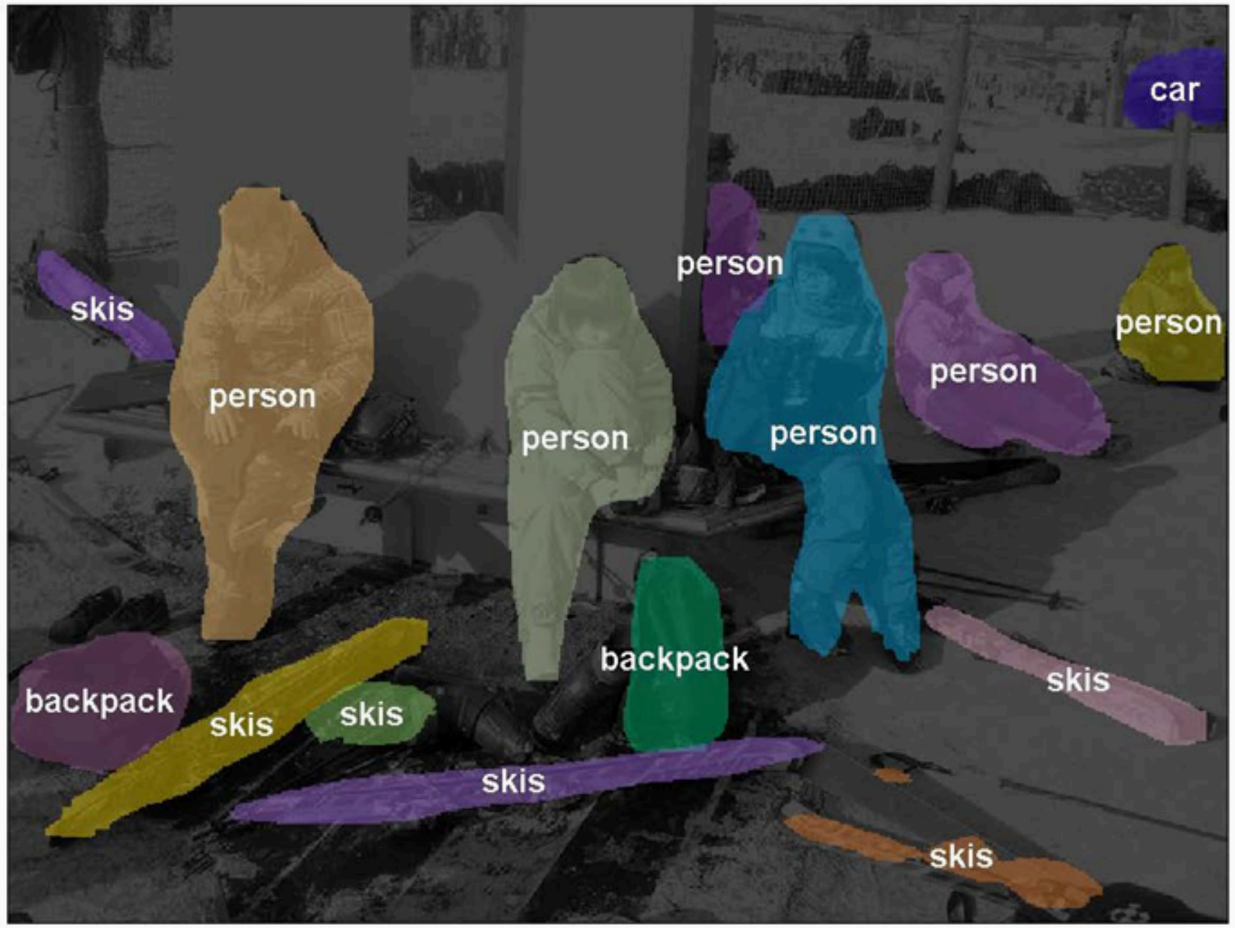




5.Conclusion

Faster R-CNN is just amazing – continue

input



THANK YOU

CML 2016 Tutorial on Deep Residual Networks