國立高雄科技大學
National Kaohsiung University of Science and Technology

Region of Interest (RoI) Pooling and Align

Speaker: Shih-Shinh Huang

February 21, 2021

Outline

- Introduction
  - Background
  - RoI Feature Extraction
- RoI Pooling
  - Overview
  - Pooling Steps
- RoI Align
  - Overview
  - Align Steps
  - Bilinear Interpolation

Introduction

- Background
  - Detecting and segmenting objects from images are important tasks in computer vision area.
    - Faster R-CNN: well-known method for detection

      Shaoqing Ren, *et. al.*, "Faster R-CNN Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. on PAMI*, 2017.

    - Mask R-CNN: well-known method for segmentation

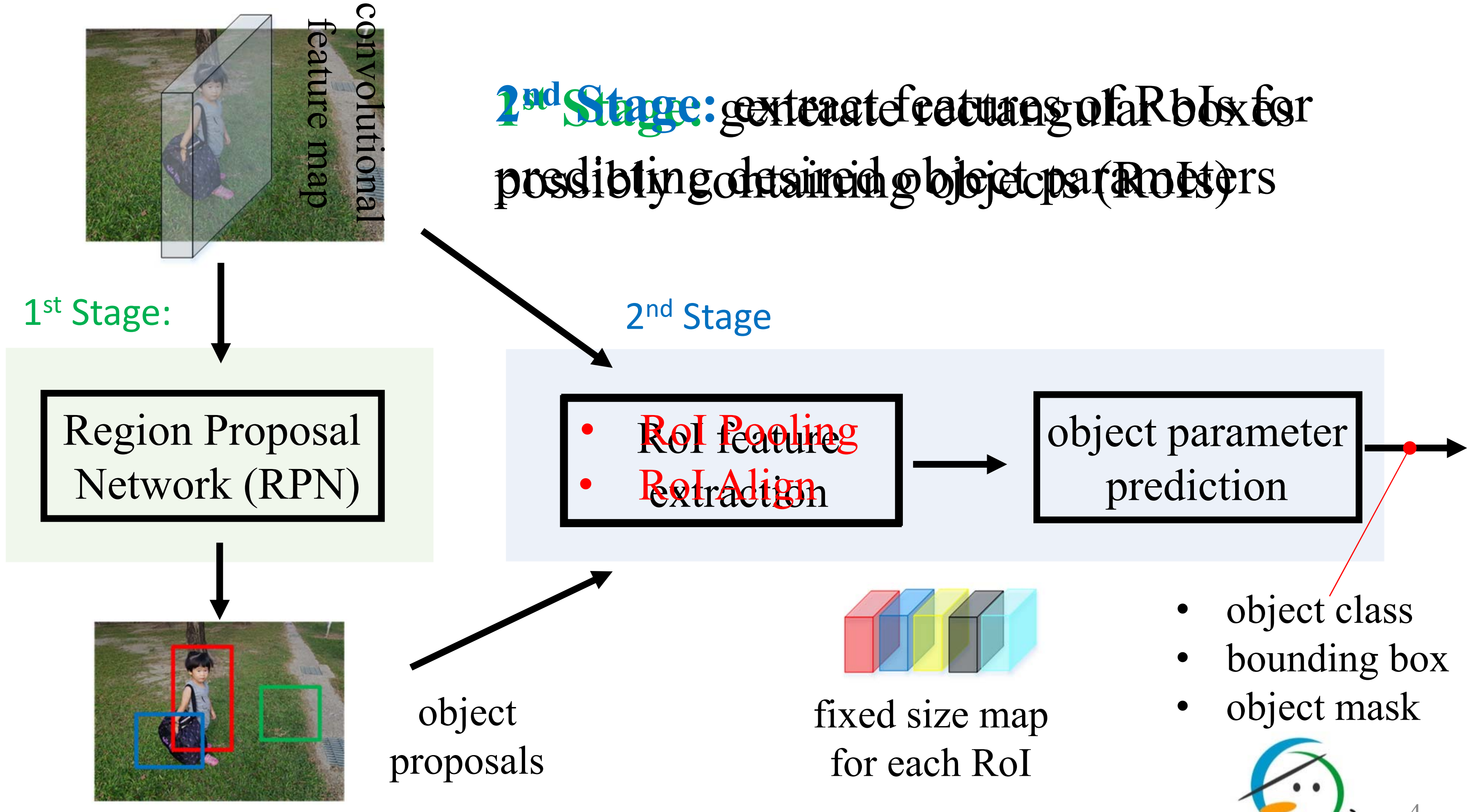
      Kaiming He, *et. al.*, "Mask R-CNN," *IEEE ICCV*, 2017

國立高雄科技大學
National Kaohsiung University of Science and Technology

Introduction

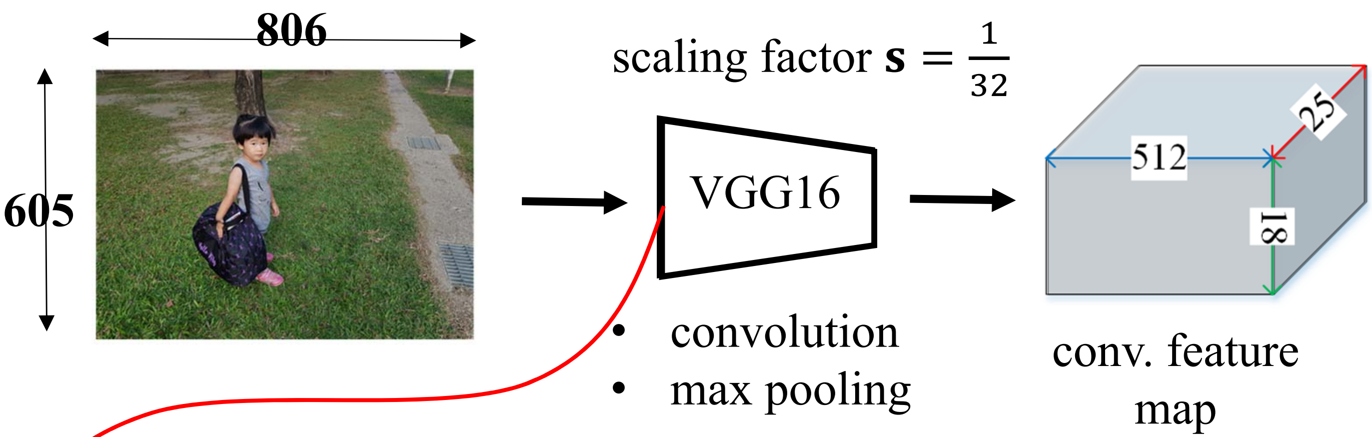
1st Stage: generate rectangular boxes possibly containing objects (RoIs)

2nd Stage: extract features of RoIs for predicting desired object parameters



Introduction

- RoI Feature Extraction: input
  - Feature Map: feature of the input image obtained from a deep convolutional neural network.



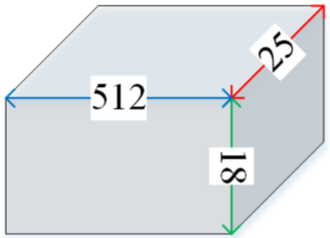
$$\left\lfloor \frac{806}{32} \right\rfloor = 25$$

$$\left\lfloor \frac{605}{32} \right\rfloor = 18$$

K. Simonyan et. al. "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ICLR*, 2015.

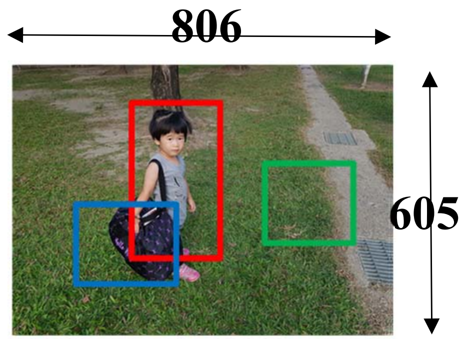
Introduction

- RoI Feature Extraction: input
  - RoI List: object proposals from RPN



conv. feature map

Region Proposal Network (RPN)



1st RoI: (224, 89), (439, 426)

top-left corner

2nd RoI: (537, 217), (732, 383)

bottom-right corner

3rd RoI: (134, 312), (355, 489)

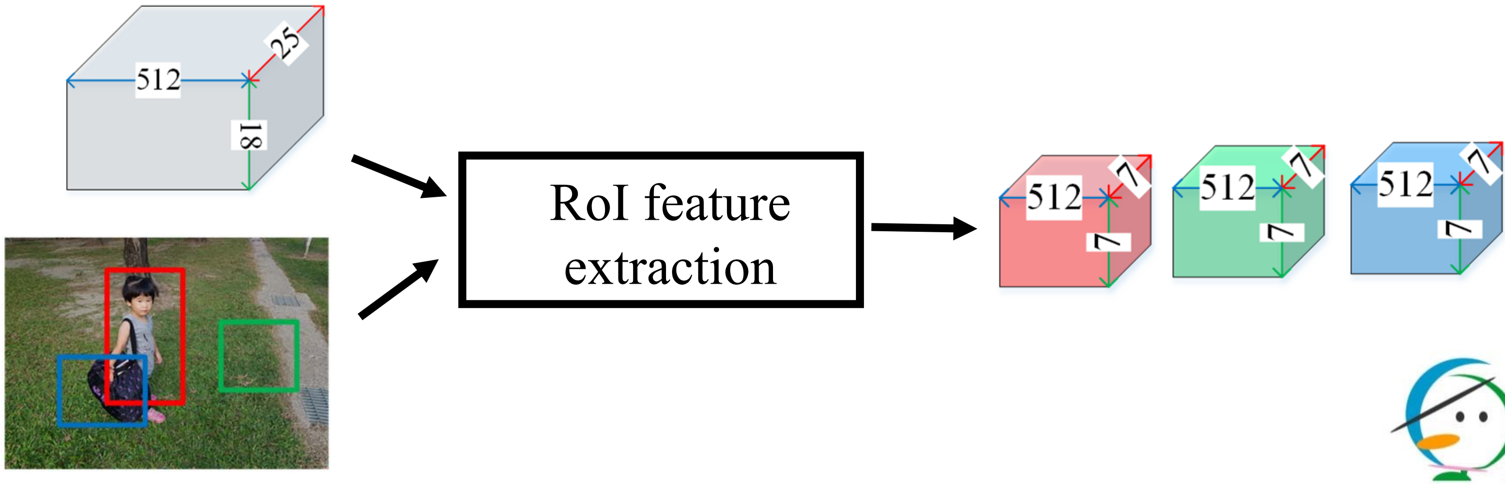
Quarter Unit: Region Proposal Network (PRN)

https://youtu.be/0tBhRfEzUWs

Shaoqing Ren, *et. al.*, "Faster R-CNN Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. on PAMI*, 2017.

Introduction

- RoI Feature Extraction: output
  - RoI feature maps: each one is the feature map within a RoI and with the fixed size $k \times k \times c$
    - k: pre-defined size ($k = 7$ in original paper)
    - c: num. of channels of input conv. feature map



RoI Pooling

- Overview
  - RoI pooling is firstly proposed in Fast R-CNN for object detection.

    Ross Girshick, "Fast R-CNN," *IEEE ICCV*, 2015

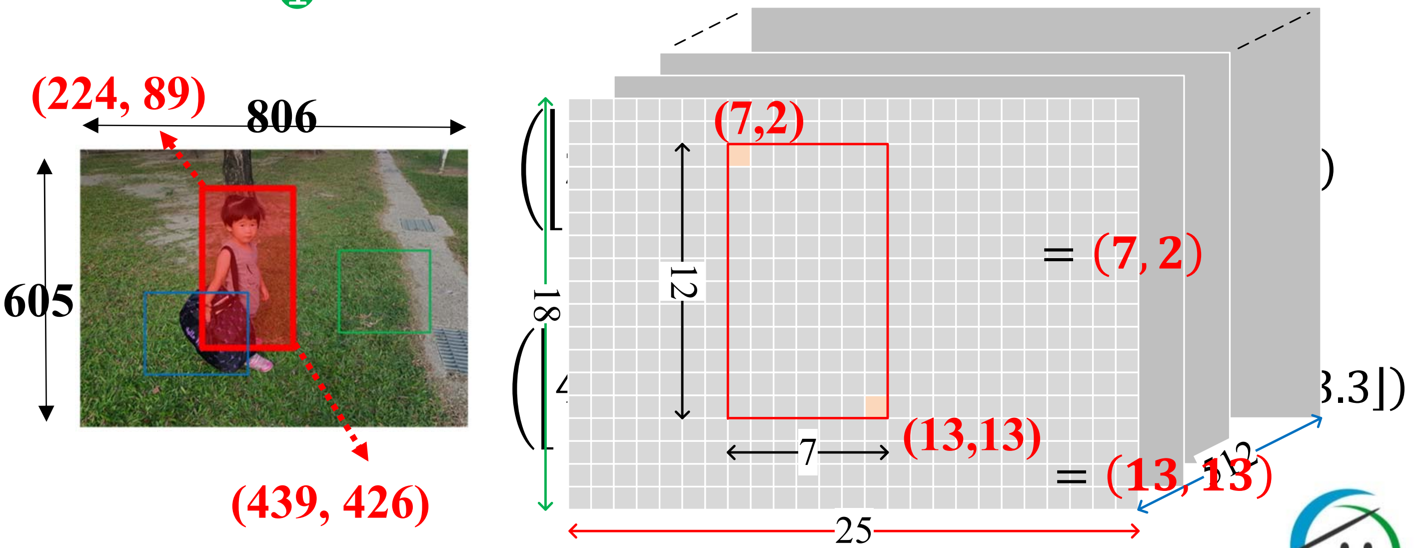
  - RoI pooling uses max pooling operation for computing the RoI feature map.
    - make the conv. feature map be reused for all RoIs
    - make the architecture be trained end-to-end.

RoI Pooling

- Overview: three steps
  - Step 1 (RoI mapping): map RoI bounding box from image to conv. feature map by quantization
  - Step 2 (RoI division): divide the mapped RoI into $k \times k$ grids by quantization
  - Step 3 (max pooling): find the maximum of all points in a grid

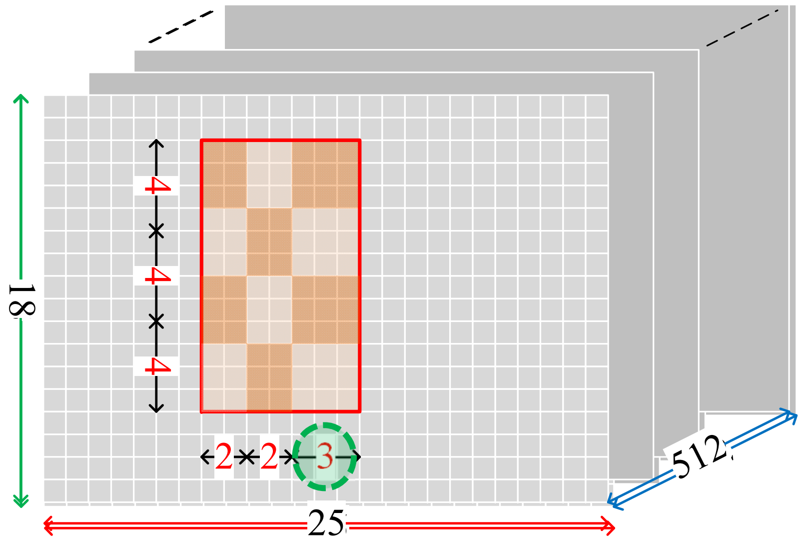
RoI Pooling

- Pooling Steps: RoI mapping
  - multiple the coordinates of RoI corners by the scaling factor *s*
  - quantize the resulting coordinates by floor operator ❶



RoI Pooling

- Pooling Steps: RoI division
  - divide the width/height of the mapped RoI by k
  - quantize the width/height by floor operator
    ❷



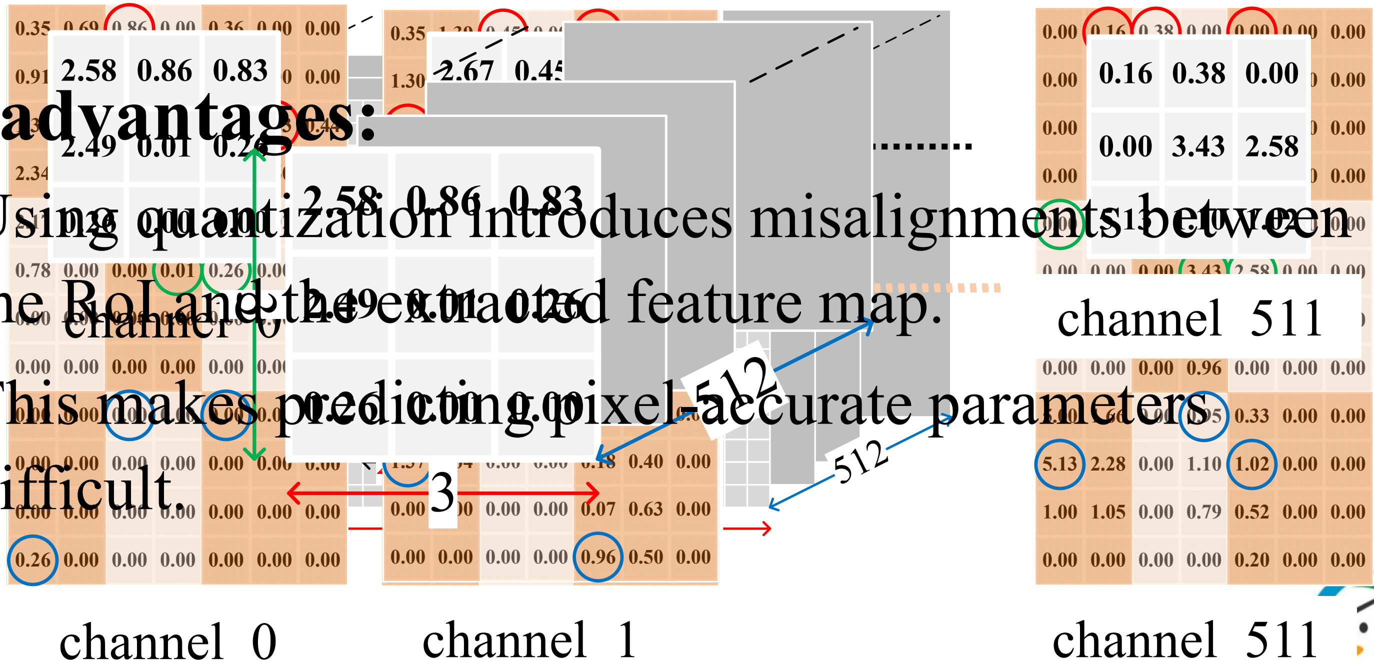
$(k = 3)$

grid width: $\left\lfloor \frac{7}{3} \right\rfloor = \lfloor 2.33 \rfloor = \mathbf{2}$

grid height: $\left\lfloor \frac{12}{3} \right\rfloor = \lfloor 4.00 \rfloor = \mathbf{4}$

RoI Pooling

- Pooling Steps: max pooling
  - take the maximum of points in each grid
  - aggregate the results in all channels



Disadvantages:

- Using quantization introduces misalignments between the RoI and the extracted feature map.
- This makes predicting pixel-accurate parameters difficult.

RoI Align

- Overview
  - RoI align was proposed and used for instance segmentation in Mask R-CNN.

Kaiming He, *et. al.*, “Mask R-CNN,” *IEEE ICCV*, 2017

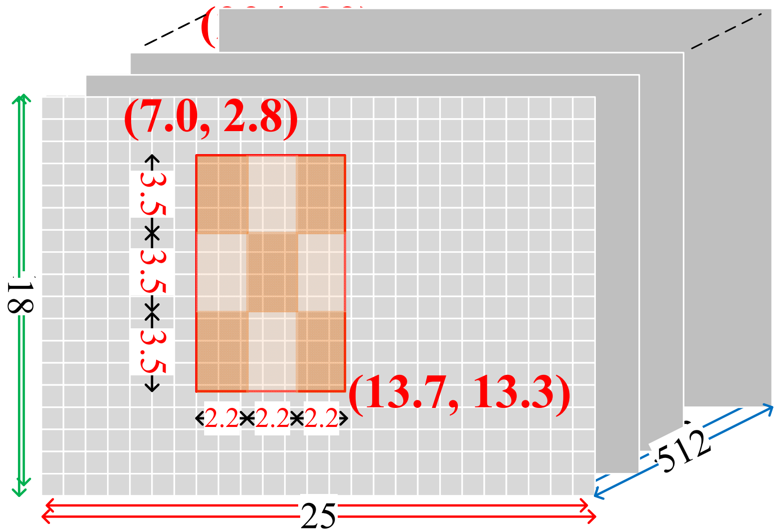
  - Objective: perform data pooling more accurate.
    - perform RoI mapping and division without quantization.
    - interpolate the values of sampling points for data pooling.

RoI Align

- Overview: four steps
  - Step 1 (RoI mapping): multiply the scaling factor to map RoI to conv. feature map
  - Step 2 (RoI division): divide the width/height of mapped RoI by k to have $k \times k$ grids.
  - Step 3 (Interpolation): interpolate the values of all sampling points (each grid has $s \times s$ points)
  - Step 4 (max pooling): find the maximum of all $s \times s$ sampling points in a grid.

RoI Align

- Align Steps
  - RoI mapping: multiple the scaling factor
  - RoI division: divide width/height by k



$(k = 3)$

$$\left(\left\lfloor 224 \times \frac{1}{32} \right\rfloor, \left\lfloor 89 \times \frac{1}{32} \right\rfloor\right) = (7.0, 2.8)$$

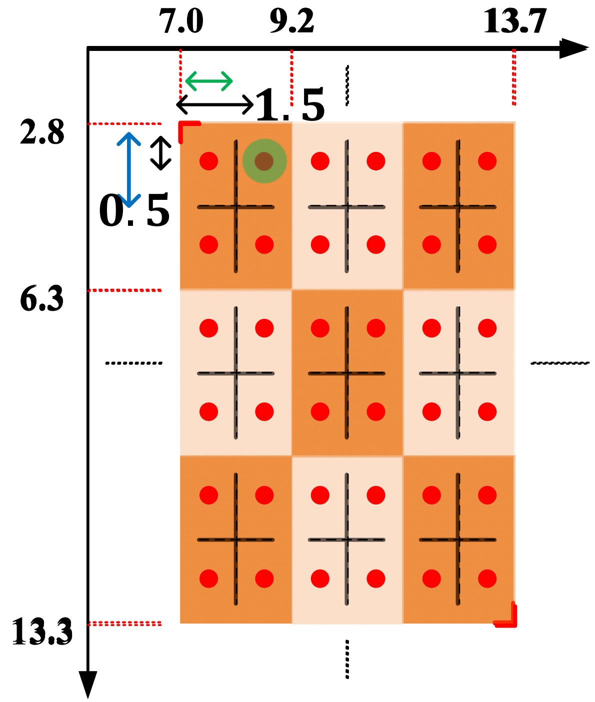
grid width: $6.7 \div 3 = \mathbf{2.2}$

$$\left(\left\lfloor 439 \times \frac{1}{32} \right\rfloor, \left\lfloor 426 \times \frac{1}{32} \right\rfloor\right) = (13.7, 13.3)$$

grid height: $10.5 \div 3 = \mathbf{3.5}$

RoI Align

- Align Steps: interpolation
  - divide each grid into $s \times s$ cells ($s = 2$)
  - take the centroids of cells as sampling points

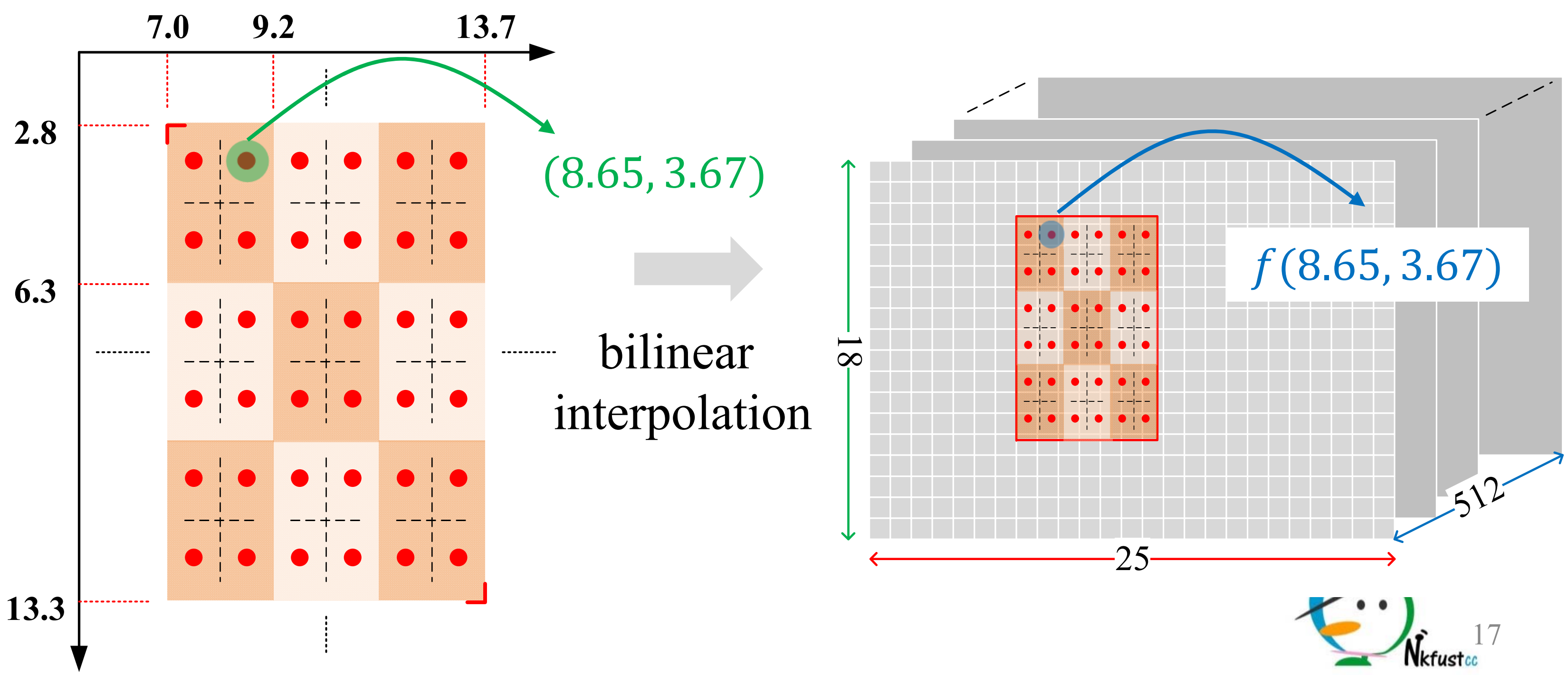


$$x: \overbrace{7.0}^{\text{staring coordinate}} + \overbrace{\left(\frac{2.2}{s = 2}\right) \times (1 + 0.5)}^{\text{displacement}} = \mathbf{8.65}$$

$$y: 2.8 + \left(\frac{3.5}{s = 2}\right) \times (0 + 0.5) = \mathbf{3.67}$$

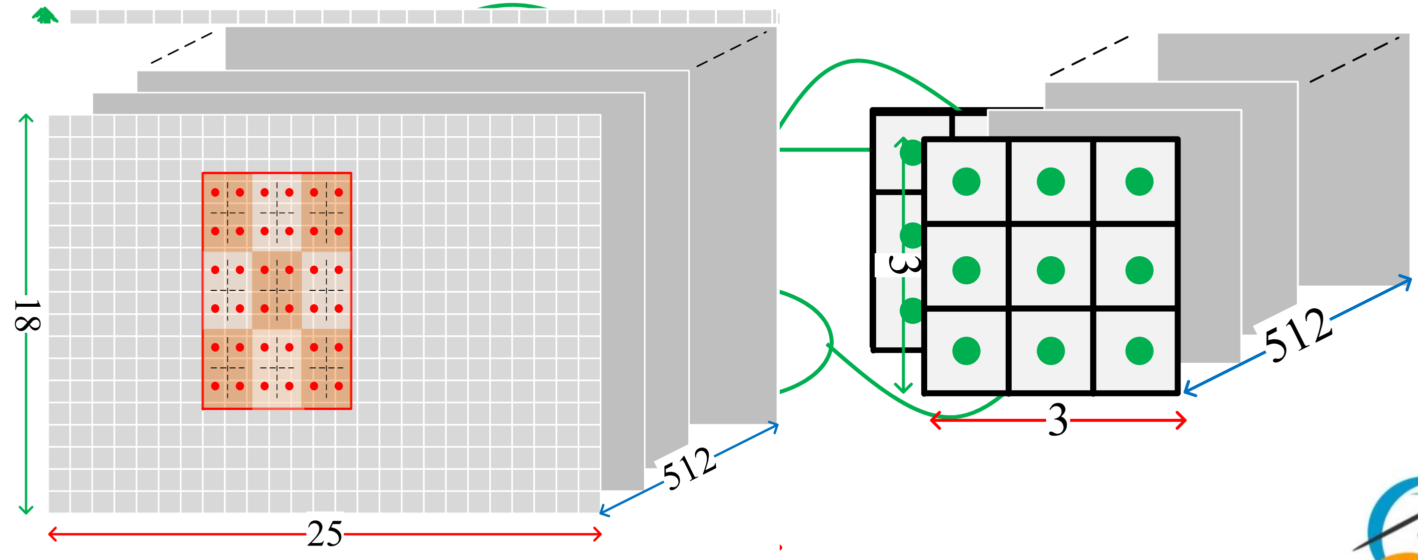
RoI Align

- Align Steps: interpolation
  - interpolate the feature value of every sampling point by bilinear interpolation.



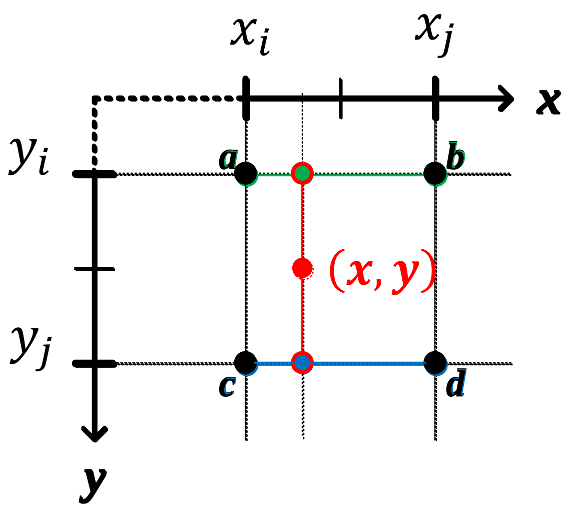
RoI Align

- Align Steps: max pooling
  - take the maximum of feature values of sampling points in each grid
  - aggregate the results in all channels



RoI Align

- Bilinear Interpolation: overview
  - It is a way to interpolate point value on a 2D grid
  - It is performed using linear interpolation twice, respectively, in horizontal and vertical directions



$$f(x_i, y_i) = a \rightarrow \boldsymbol{f(x, y_i)} \leftarrow f(x_j, y_i) = b$$

$$\downarrow$$

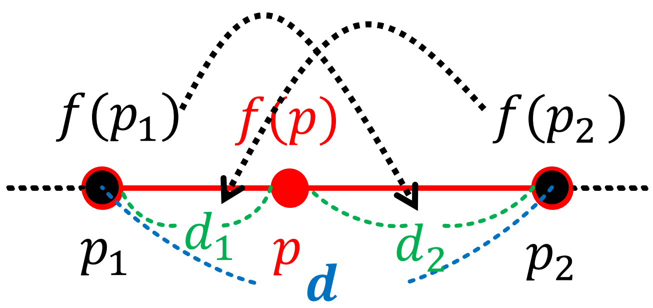
$$\boldsymbol{f(x, y)} = ?$$

$$\uparrow$$

$$f(x_i, y_j) = c \rightarrow \boldsymbol{f(x, y_j)} \leftarrow f(x_j, y_j) = d$$

RoI Align

- Bilinear Interpolation: overview
  - Linear interpolation: interpolate the value $f(p)$ of a point p between p_1 and p_2
    - $f(p)$ is the weighted sum of $f(p_1)$ and $f(p_2)$
    - w_1 and w_2 are inversely proportional to d_1 and d_2

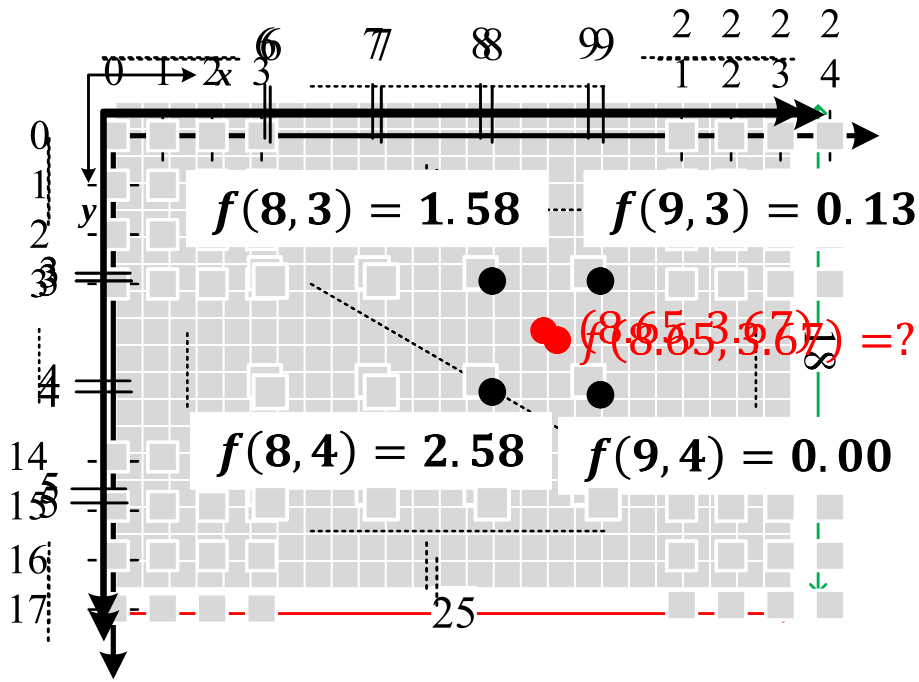


$$f(p) = \frac{d_2}{d} \times f(p_1) + \frac{d_1}{d} \times f(p_2)$$

- $w_1 + w_2 = 1.0$
- $w_1 \propto \frac{1}{d_1}$
- $w_2 \propto \frac{1}{d_2}$

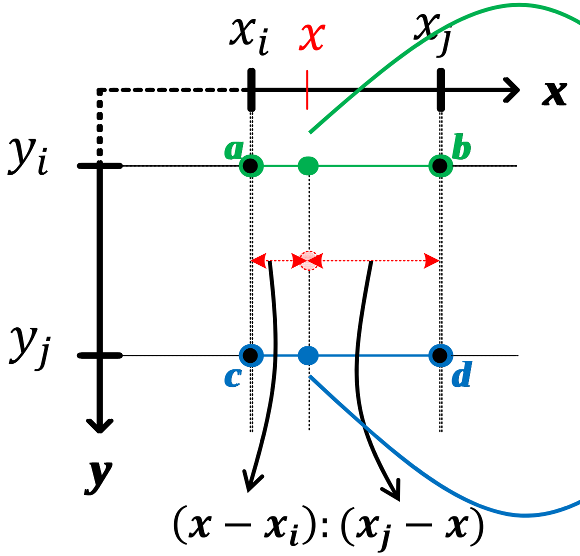
RoI Align

- Bilinear Interpolation
  - consider the conv. feature map as a square lattice of integral points
  - take the four closest integral points around the point p for interpolation



RoI Align

- Bilinear Interpolation
  - horizontal interpolation: apply linear interpolation in horizontal direction

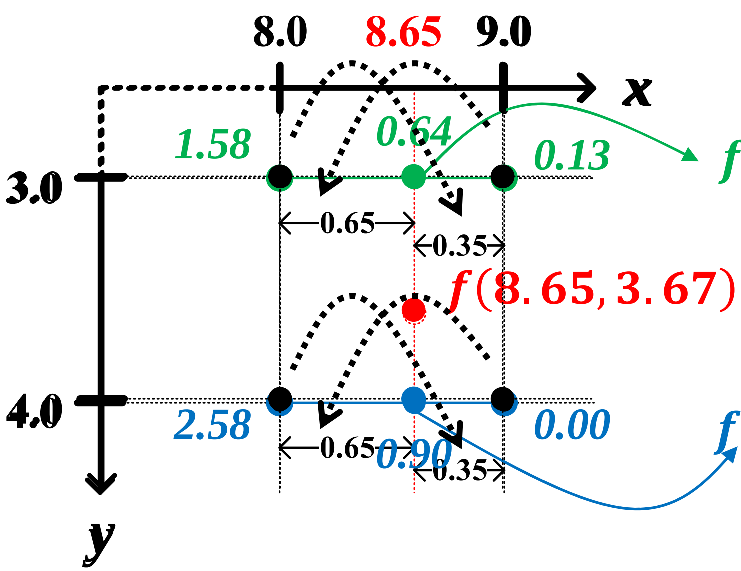


$$f(x, y_i) = \frac{(x_j - x)}{(x_j - x_i)} \times a + \frac{(x - x_i)}{(x_j - x_i)} \times b$$

$$f(x, y_j) = \frac{(x_j - x)}{(x_j - x_i)} \times c + \frac{(x - x_i)}{(x_j - x_i)} \times d$$

RoI Align

- Bilinear Interpolation:
  - horizontal interpolation: apply linear interpolation in horizontal direction



$$f(8.65, 3.0) = \frac{0.35}{1.0} \times 1.58 + \frac{0.65}{1.0} \times 0.13$$

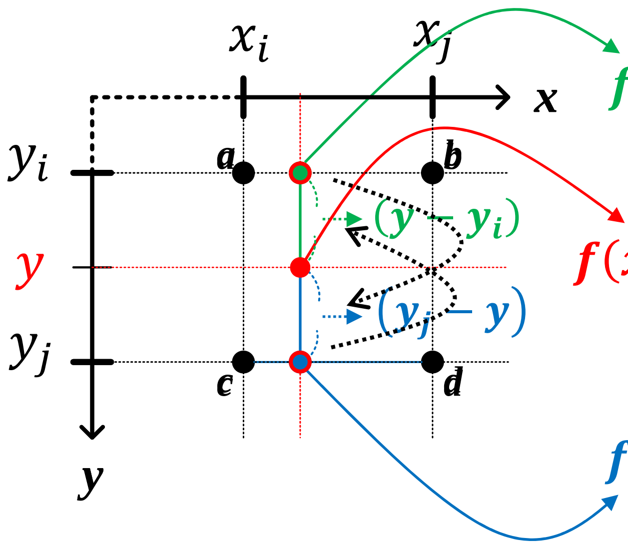
$$= 0.64$$

$$f(8.65, 4.0) = \frac{0.35}{1.0} \times 2.58 + \frac{0.65}{1.0} \times 0.00$$

$$= 0.90$$

RoI Align

- Bilinear Interpolation
  - vertical interpolation: apply linear interpolation in vertical direction



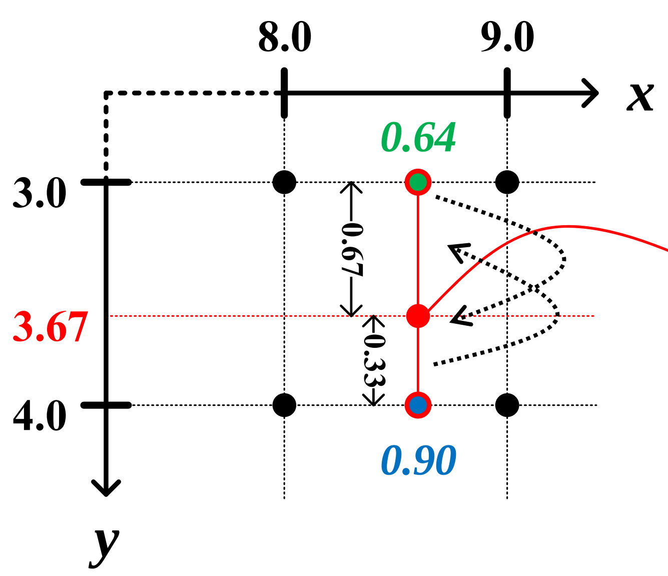
$$f(x, y_i) = \frac{(x_j - x)}{(x_j - x_i)} \times a + \frac{(x - x_i)}{(x_j - x_i)} \times b$$

$$f(x, y) = \frac{(y_j - y)}{(y_j - y_i)} \times f(x, y_i) + \frac{(y - y_i)}{(y_j - y_i)} \times f(x, y_j)$$

$$f(x, y_j) = \frac{(x_j - x)}{(x_j - x_i)} \times c + \frac{(x - x_i)}{(x_j - x_i)} \times d$$

RoI Align

- Bilinear Interpolation
  - vertical interpolation: apply linear interpolation in vertical direction



$f(8.65, 3.0) = 0.64$

$$f(8.65, 3.67) = \frac{0.33}{1.0} \times 0.64 + \frac{0.67}{1.0} \times 0.90$$

$f(8.65, 4.0) = 0.90$

$$= 0.81$$


thank you
© 1996, 2002 SANRIO CO., LTD.