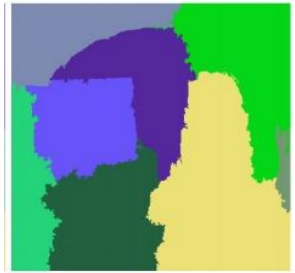
Faster R-CNN

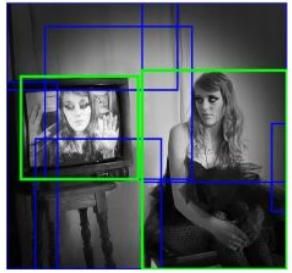
Towards Real-Time Object Detection with Region Proposal Networks

Jaeyoung Lee

Contents

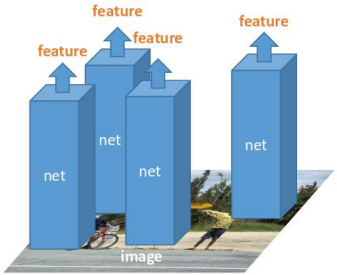
History (MS Research)



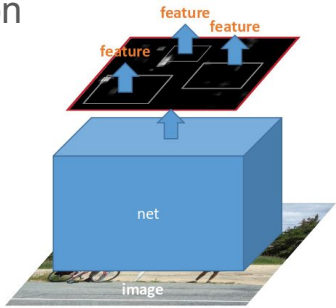


Region Proposal using Selective Search
(Traditional Segmentation)

- R-CNN (Region-CNN)
  - Region Proposal + CNN on ROI for Classification
- SPP-Net (Spatial Pyramid Pooling Nets)
  - Region Proposal + [Shared CNN on Full Image + FCs on ROI] for Classification
    - Remove redundant usage of CNN
- Fast R-CNN
  - Upgrade SPP-Net's Second Part
    - Refine SPP-Net
- Faster R-CNN
  - Replace Previous Region Proposal Process with CNN
    - Shared CNN between Region Proposal and Classification
    - Simple to Understand intuitively
  - Faster
  - More accurate

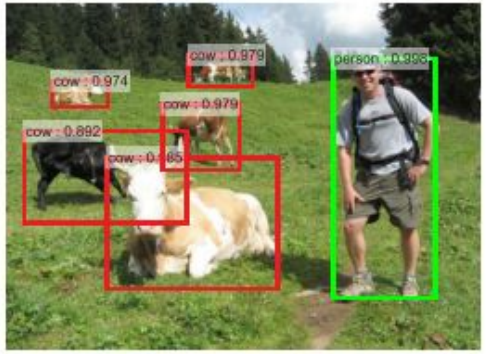


R-CNN
2000 nets on image regions



SPP-net
1 net on full image

Overall Architecture



1st Stage:
Contribution Point of Fast'er' R-CNN

Region Proposal Networks

Region Proposal

Input

Conv Net

Feature map

2nd Stage: Fast R-CNN

ROI Pooling

ROI

ROI

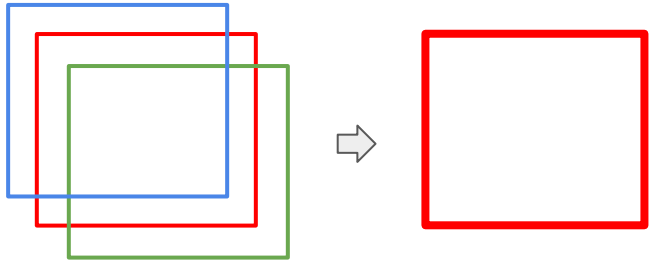
ROI

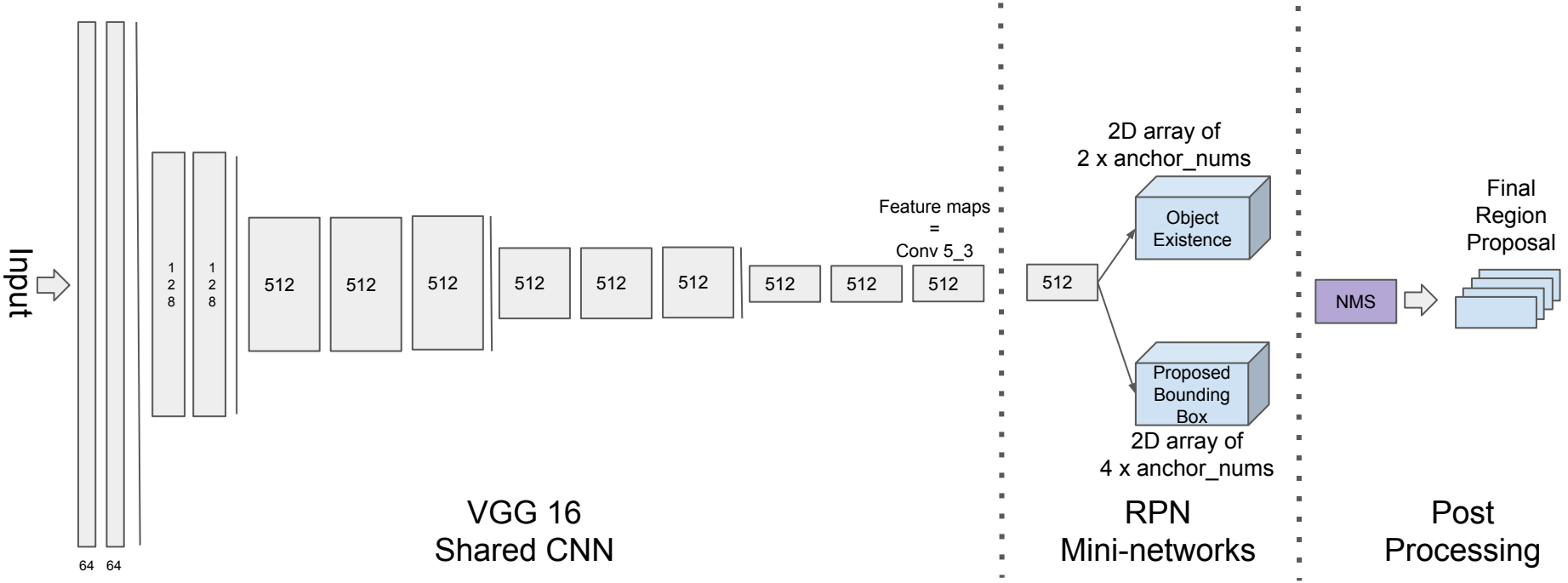
Fully Connected Layer

Class

BBox

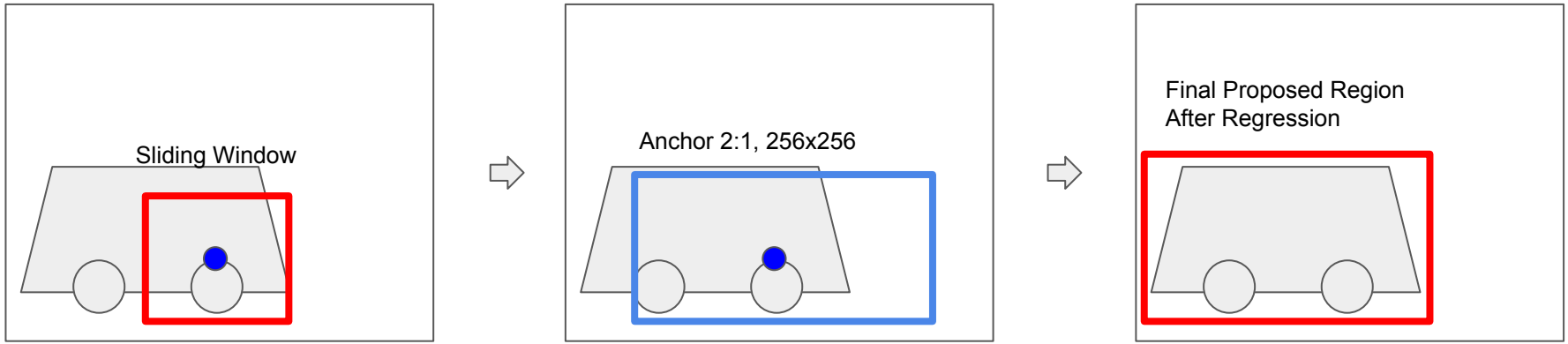
Region Proposal Networks





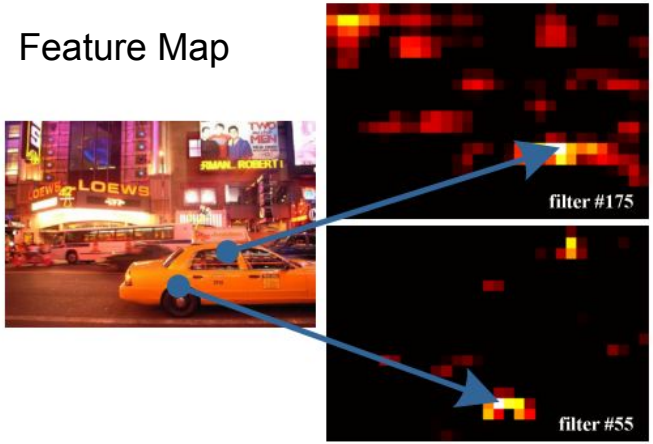
Region Proposal Process

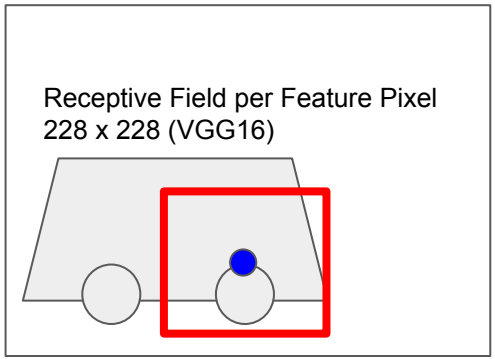
- Search Every Features of All Objects inside Sliding Window
- “Anchor” Rough Location and Bounding Box
- Regress Exact Location and Size



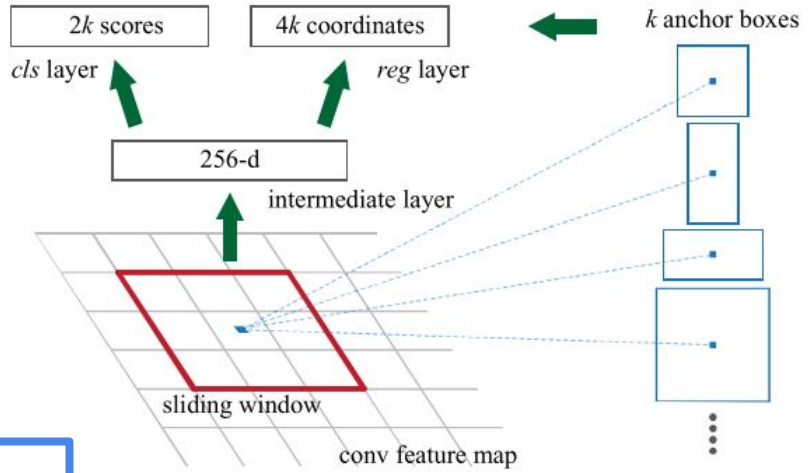
Anchoring on Feature Maps

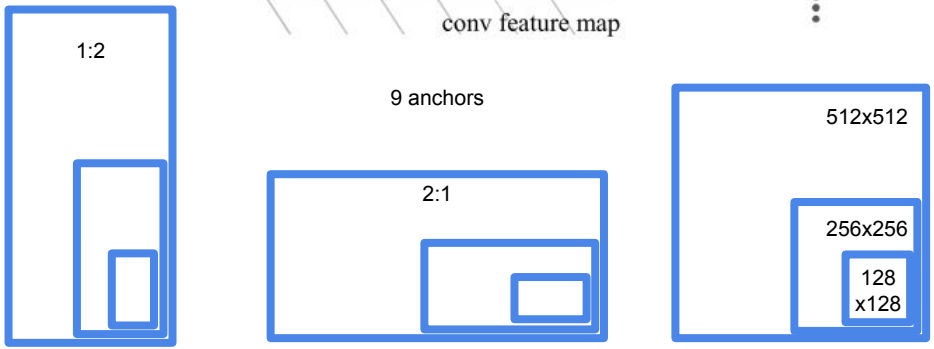
- Feature Map





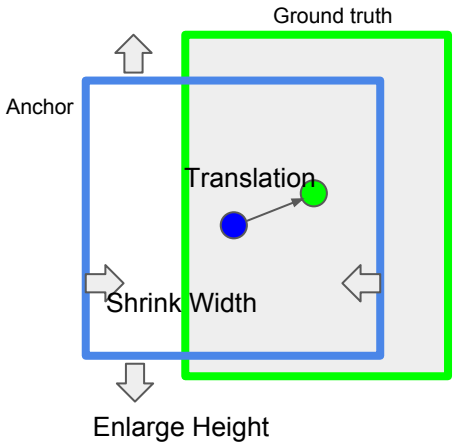
- Mimic Sliding Window





Bounding Box Regression

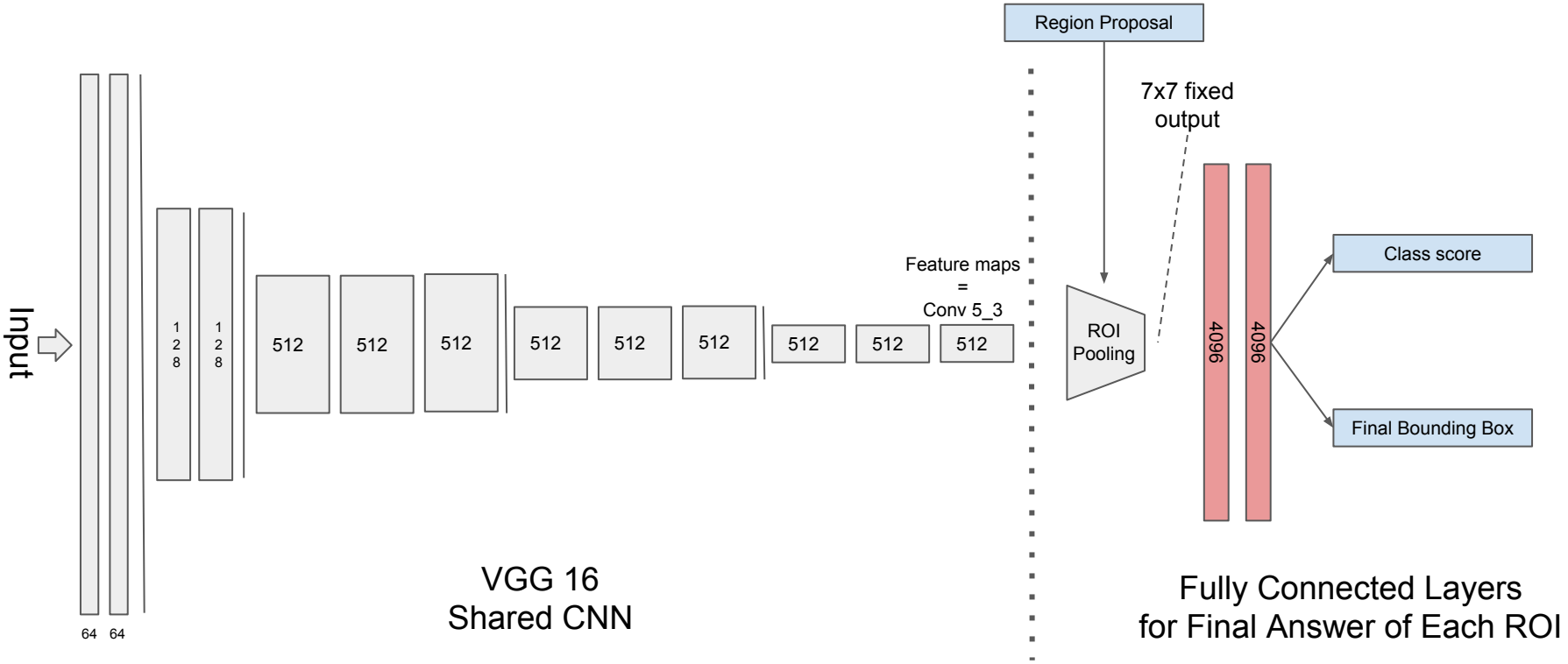
- Guess from Features, Not all Object's Region
- C.F.
  - BBox Reg in 2nd Stage (Final Classification & Localization) is used Since R-CNN
- Regression with Baselines - 9 Anchor Boxes with various scales, ratios



$$t_x = (G_x - P_x)/P_w$$
$$t_y = (G_y - P_y)/P_h$$
$$t_w = \log(G_w/P_w)$$
$$t_h = \log(G_h/P_h).$$

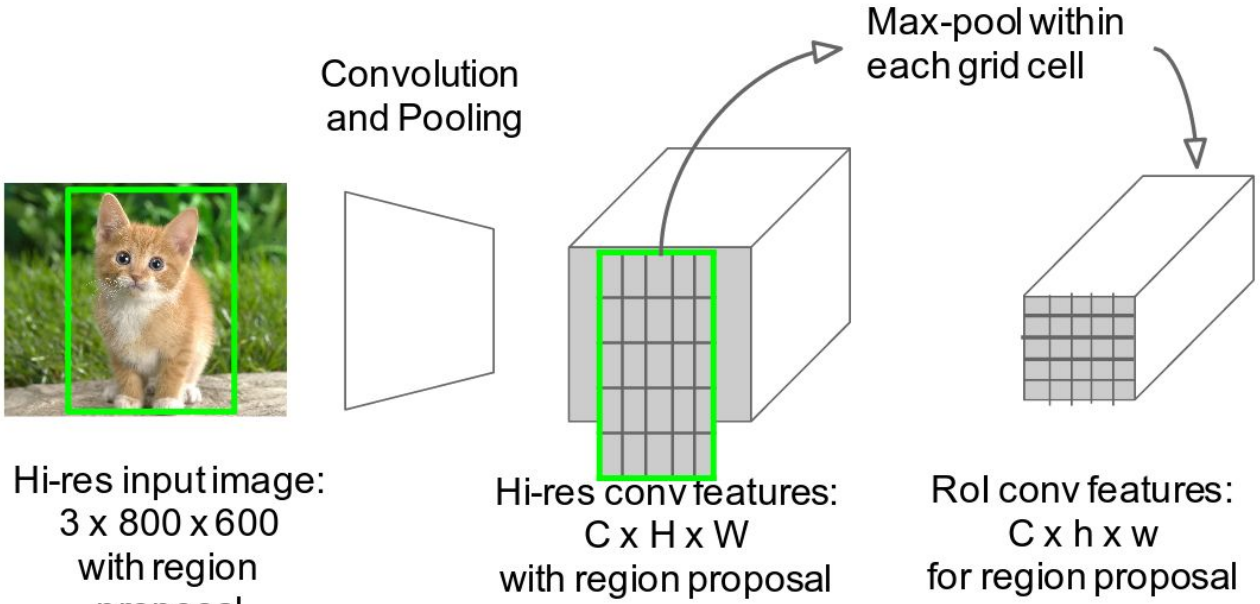
P: Predicted
G: Ground Truth

Classification & Final Localization

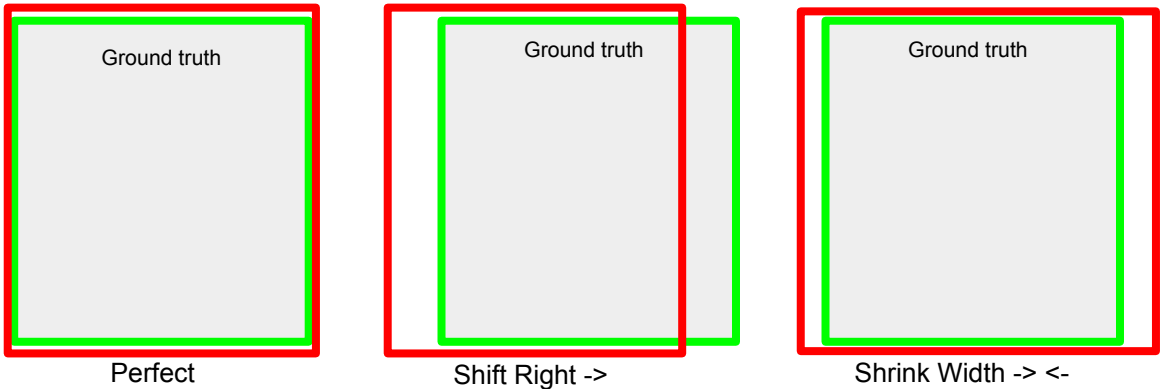


ROI Pooling

- Crop ROI on Feature maps
- Warp ROI into Fixed 7x7 grid



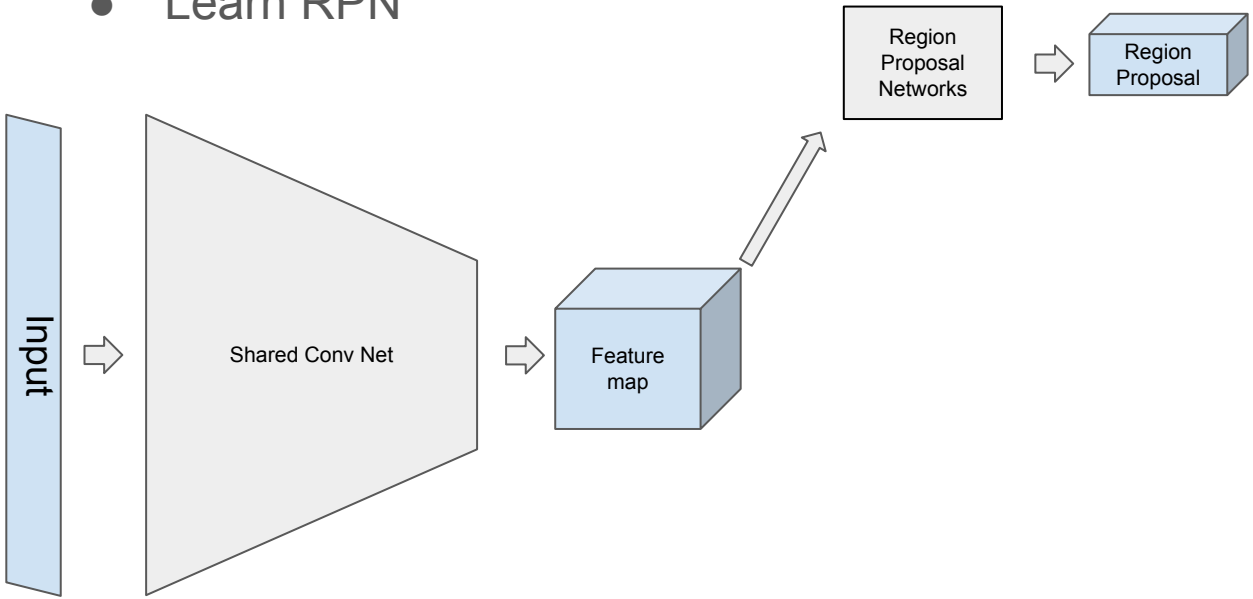
Bounding Box Regression Again...



Perfect | Shift Right -> | Shrink Width -> <-

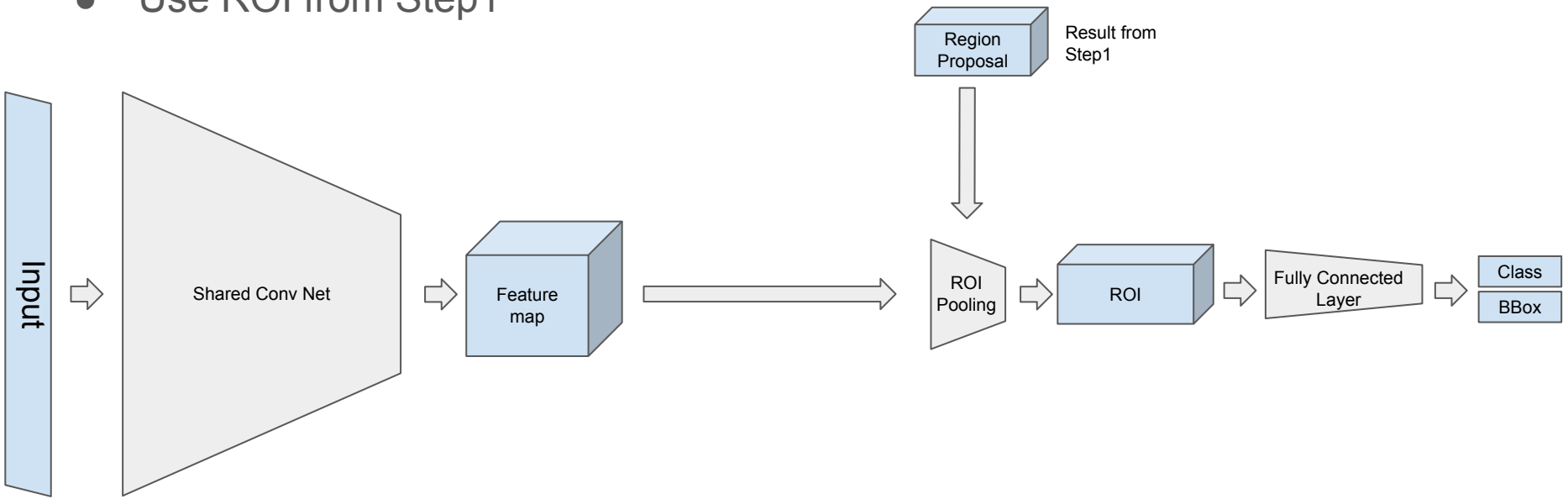
Learning: Step 1

- Initialize Pre-trained Shared Conv Nets
- Finetune Shared Conv Nets
- Learn RPN



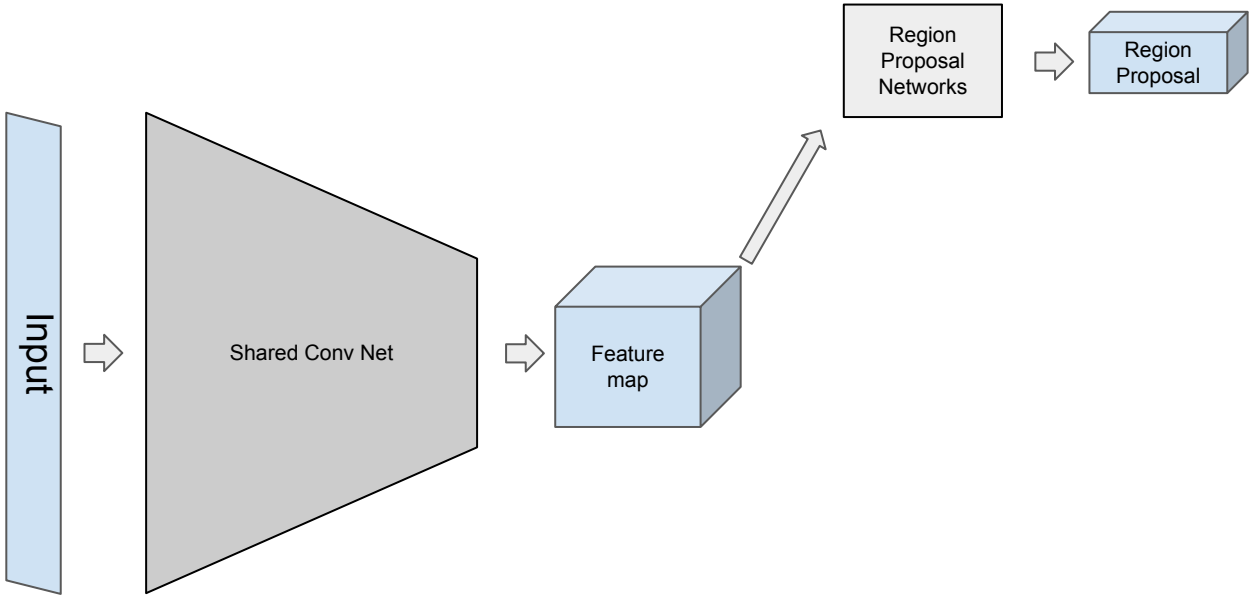
Learning: Step 2

- Finetune Shared Conv Nets
- Learn Fully-Connected-Layers
- Use ROI from Step1



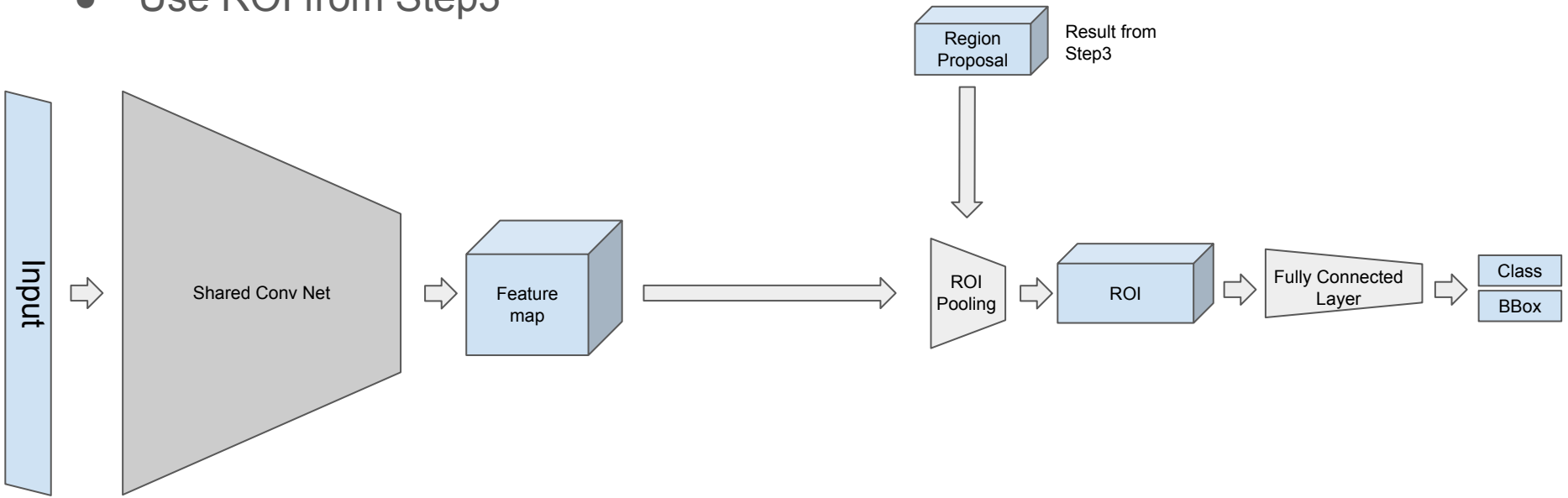
Learning: Step 3

- Fix Shared Conv Nets
- Learn RPN “Again”



Learning: Step 4

- Fix Shared Conv Nets
- Learn Fully-Connected-Layers “Again”
- Use ROI from Step3



Discussions

- How about Semantic Segmentation?
  - Pro
    - Intuitively Simple to Implementation
      - ROI Pooling?
    - Fine-grained localization for Pose estimation
  - Cons
    - Cost