Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

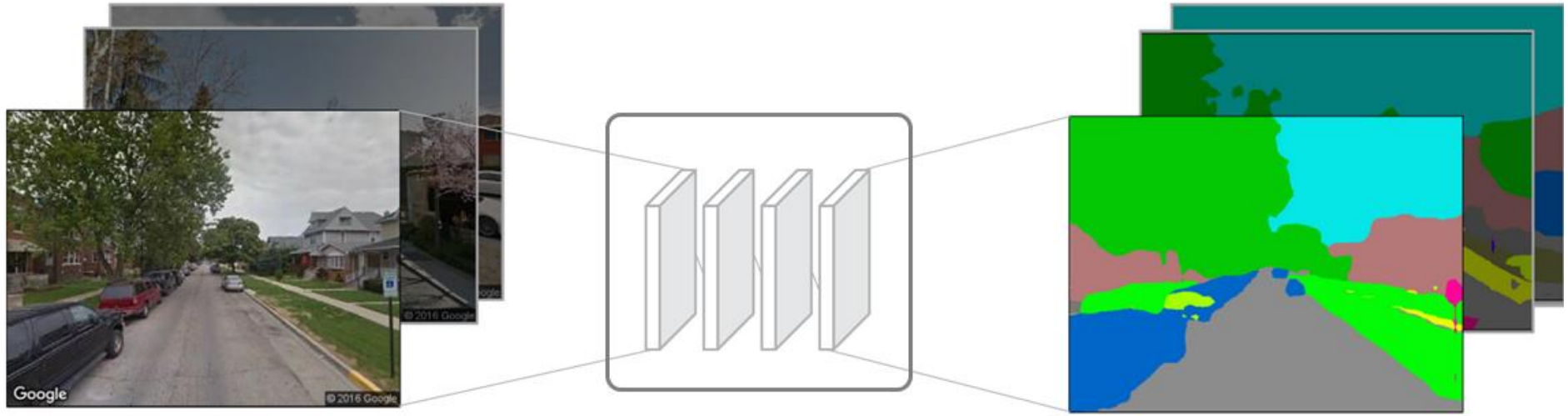
Sistemas Urbanos Inteligentes

Aprendizaje Multitarea (Multitask Learning)

Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación

¿Qué utilidad nos pueden entregar las CNN en contextos urbanos?

¿Cómo explicar la percepción?

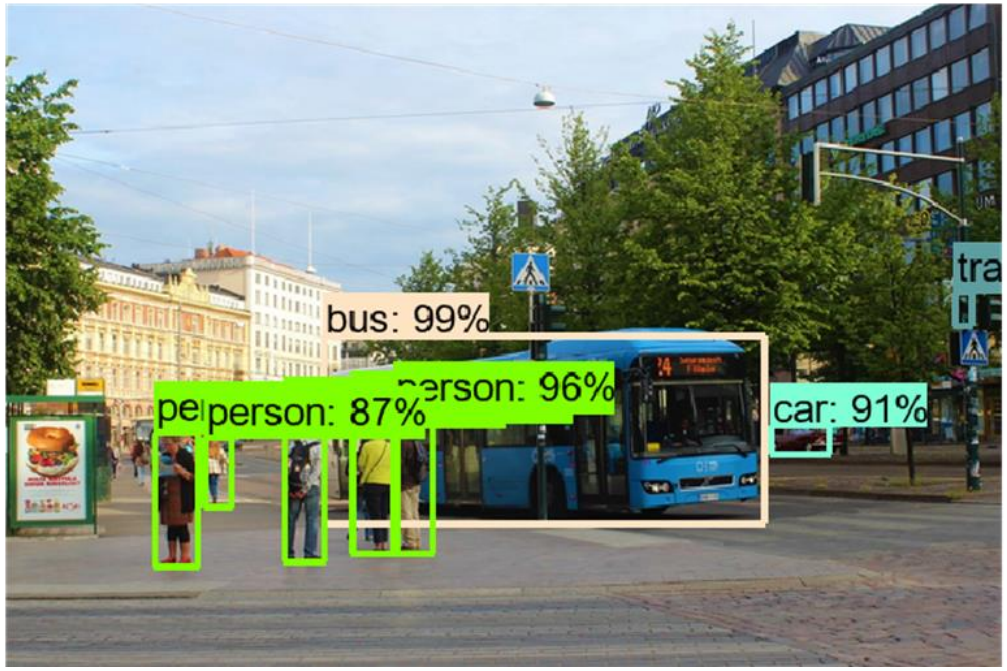


Segmentación semántica de imágenes

Rossetti, T., Lobel, H., Rocco, V., Hurtubia, R. (2019). Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. Landscape & Urban Planning, 181, 169-178

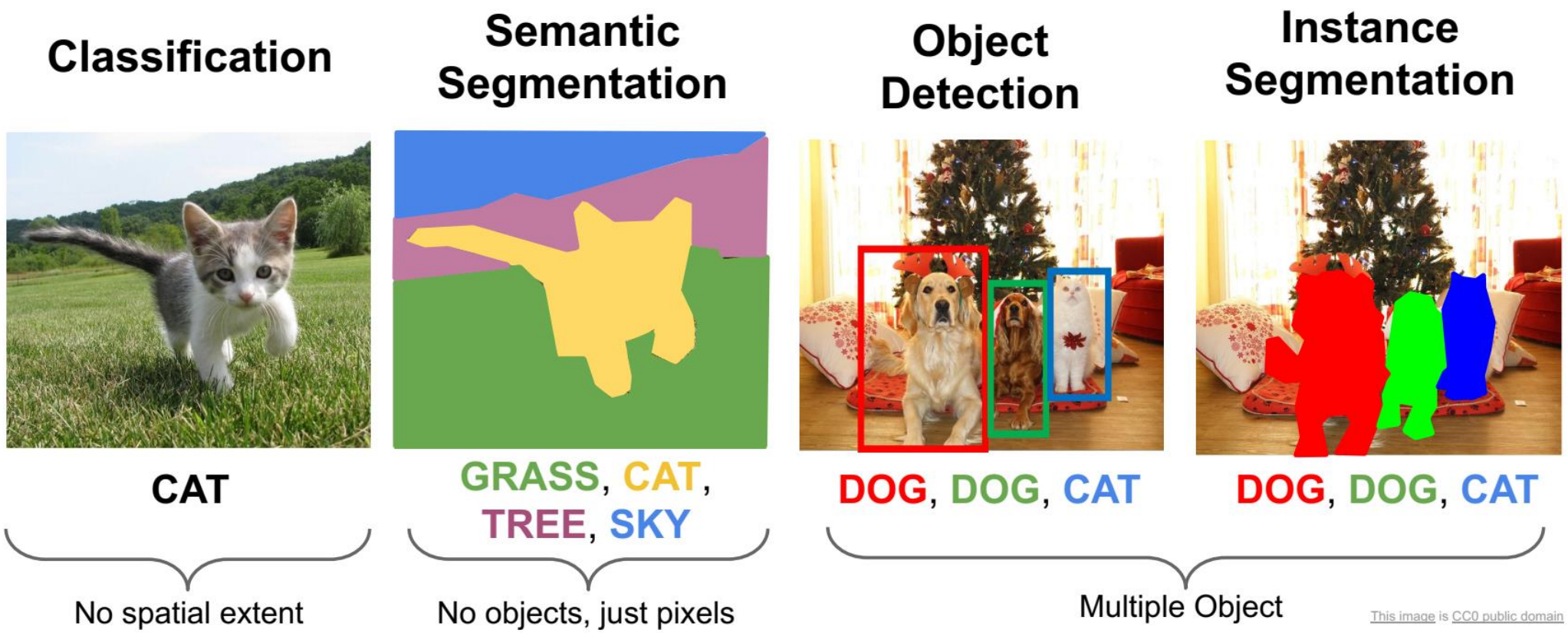
¿Cómo mejorar la caracterización de las imágenes?

Segmentación + detección de objetos

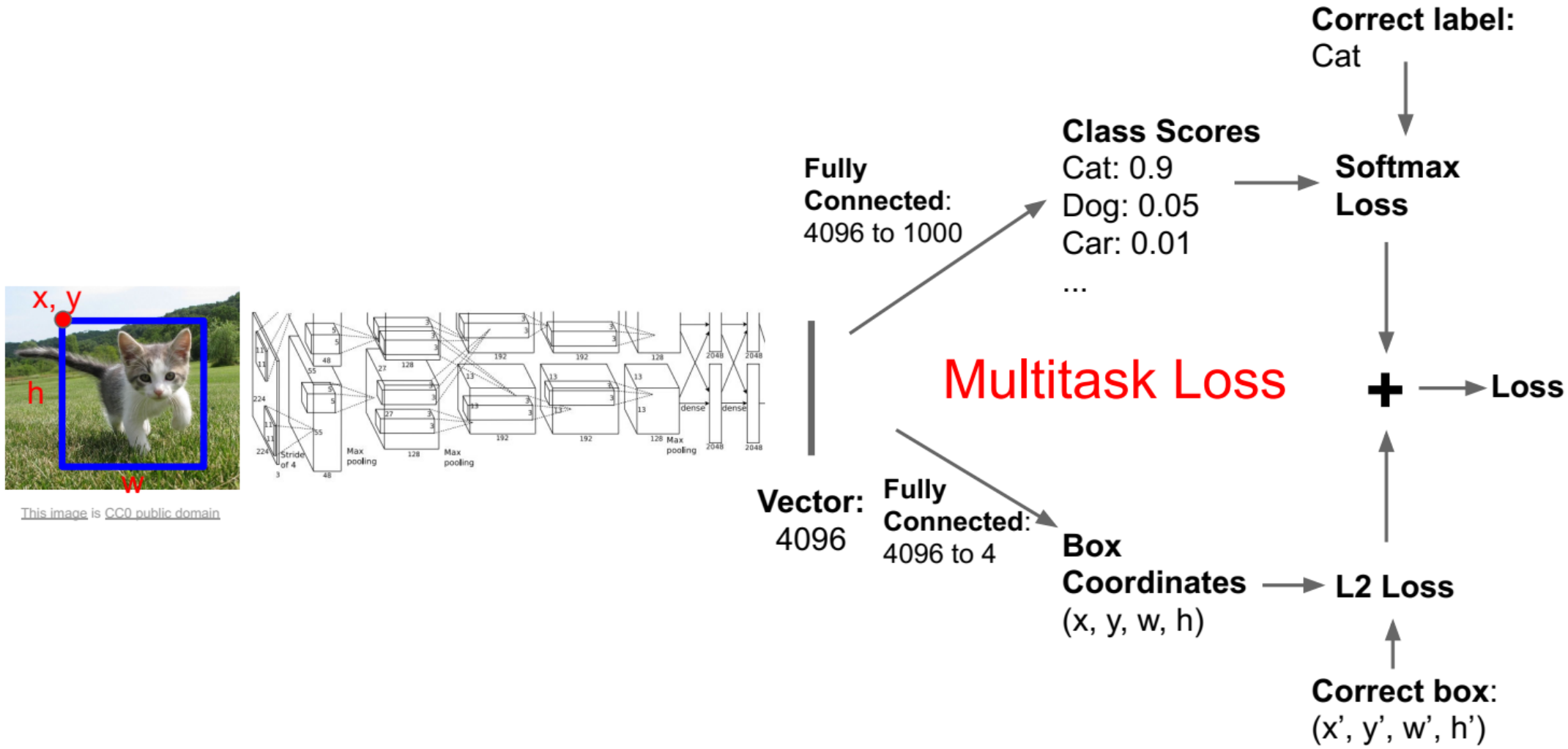


Ramírez, T., Rossetti, T., Lobel, H., Hurtubia, R. (2021). Measuring heterogeneous perception of urban space with massive data and machine learning: an application to safety. Landscape & Urban Planning, 208, 104002.

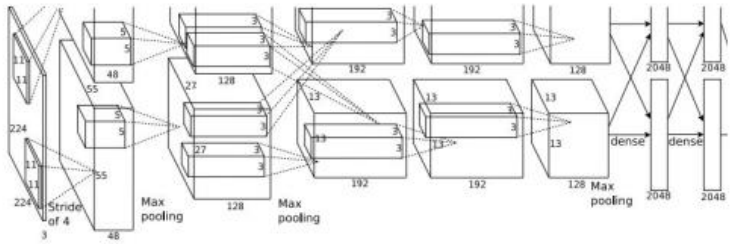
Necesitamos subir en la escala de problemas de visión por computador, hasta la detección de objetos



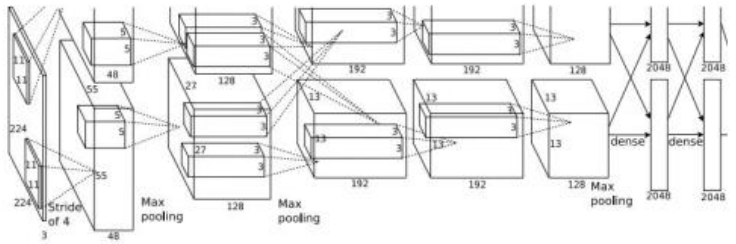
Para esto, necesitamos realizar dos predicciones de manera simultánea



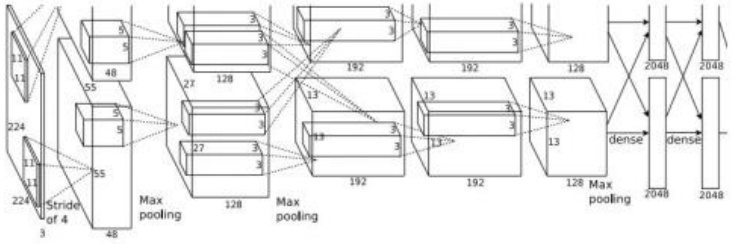
Para esto, necesitamos realizar dos predicciones de manera simultánea



CAT: (x, y, w, h)



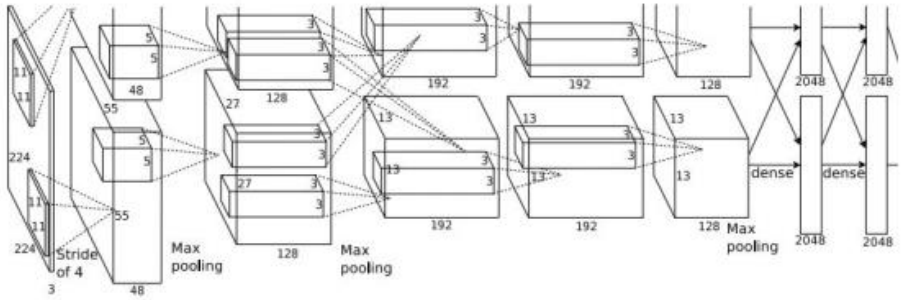
DOG: (x, y, w, h)
DOG: (x, y, w, h)
CAT: (x, y, w, h)



DUCK: (x, y, w, h)
DUCK: (x, y, w, h)
....

¿Qué problema nos genera este esquema de detección?

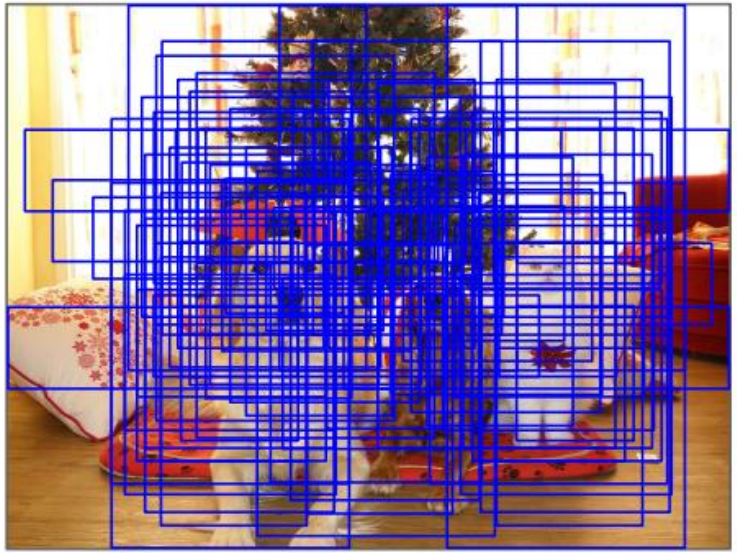
Podemos utilizar una ventana deslizante
para analizar todas las posiciones

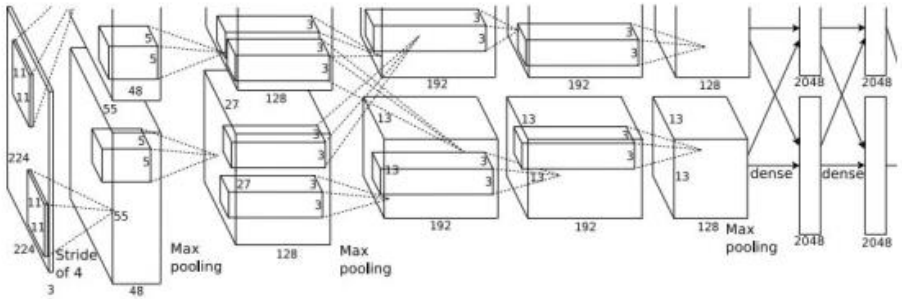


Dog? NO
Cat? NO
Background? YES

¿Qué problema nos genera este esquema de detección?

Podemos utilizar una ventana deslizante
para analizar todas las posiciones

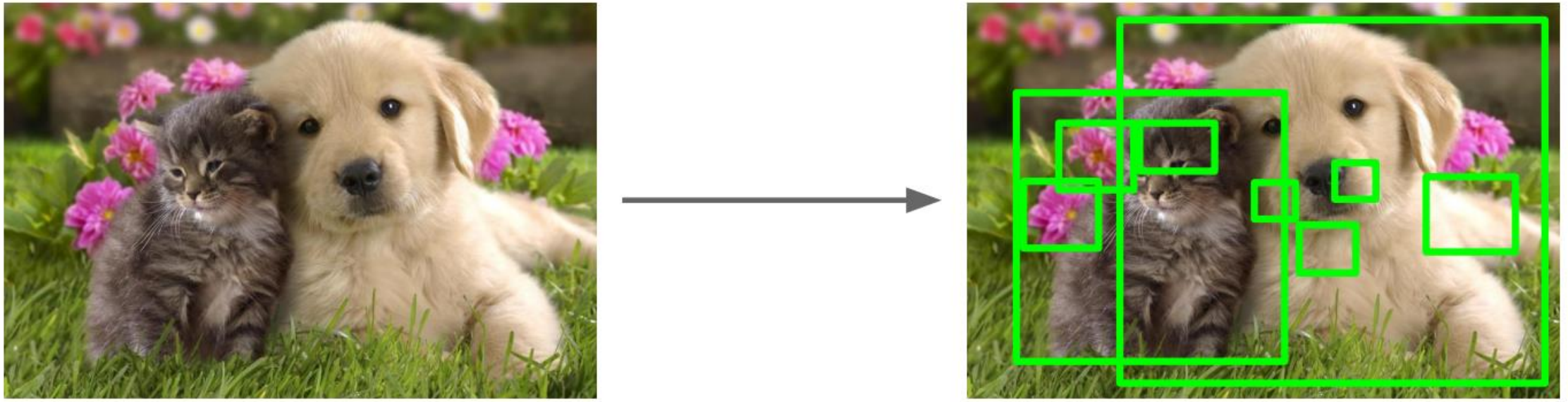




Dog? NO
Cat? YES
Background? NO

Necesitamos generar menos ventanas candidatas

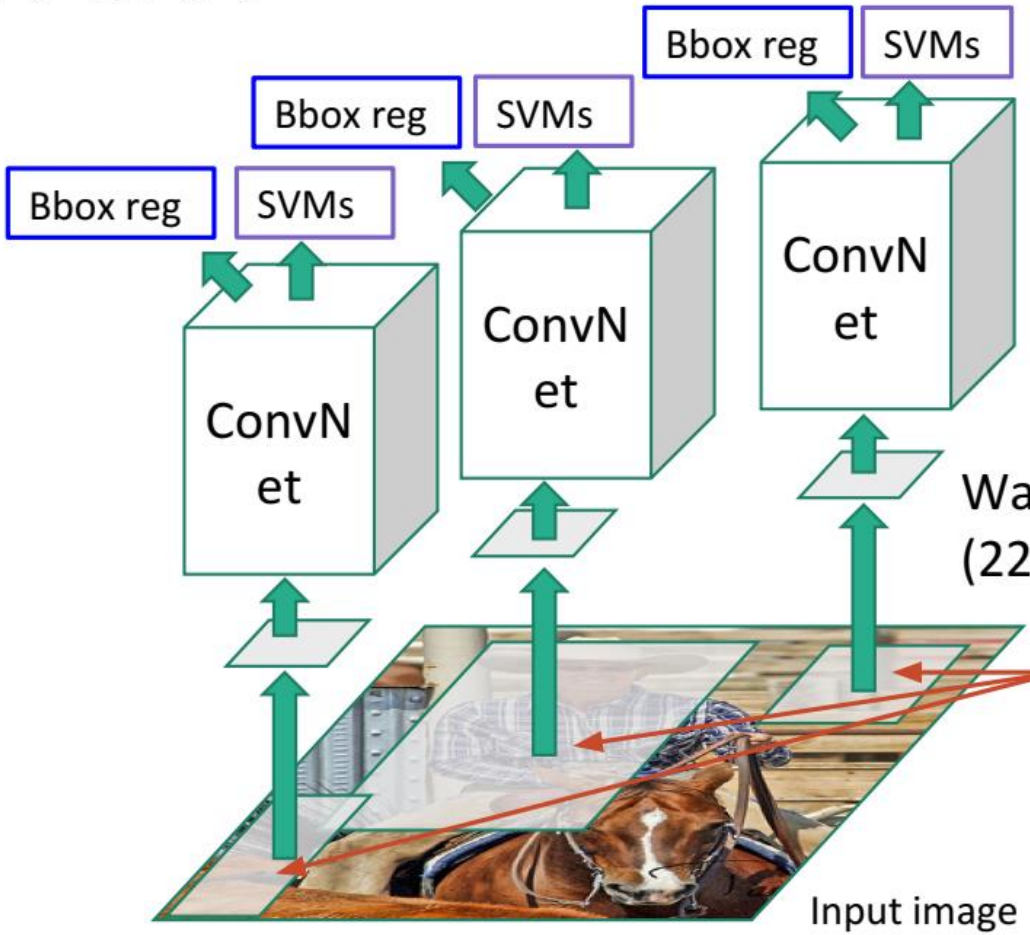
Region Proposals entregan rápidamente ventanas donde podrían haber objetos



Alexe et al, "Measuring the objectness of image windows", TPAMI 2012
Uijlings et al, "Selective Search for Object Recognition", IJCV 2013
Cheng et al, "BING: Binarized normed gradients for objectness estimation at 300fps", CVPR 2014
Zitnick and Dollar, "Edge boxes: Locating object proposals from edges", ECCV 2014

R-CNN

Predict “corrections” to the RoI: 4 numbers: (dx, dy, dw, dh)



Classify regions with SVMs

Forward each region through ConvNet (ImageNet-pretranied)

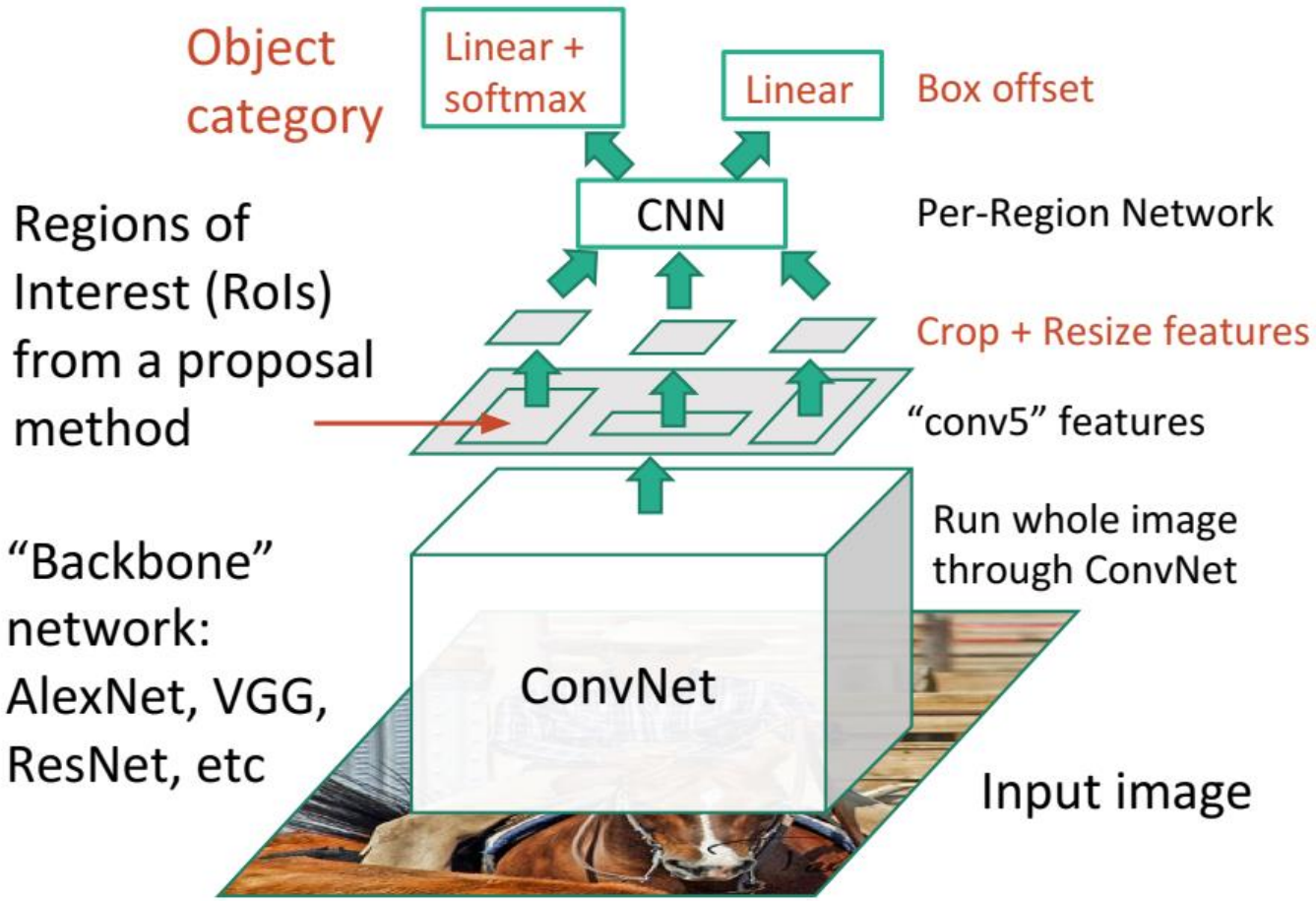
Warped image regions (224x224 pixels)

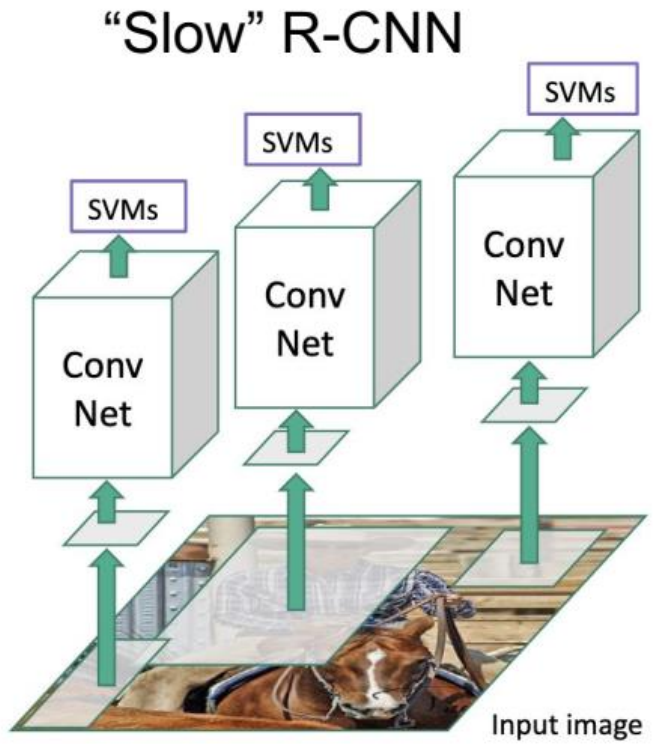
Regions of Interest (RoI) from a proposal method (~2k)

Girshick et al, “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014.
Figure copyright Ross Girshick, 2015;

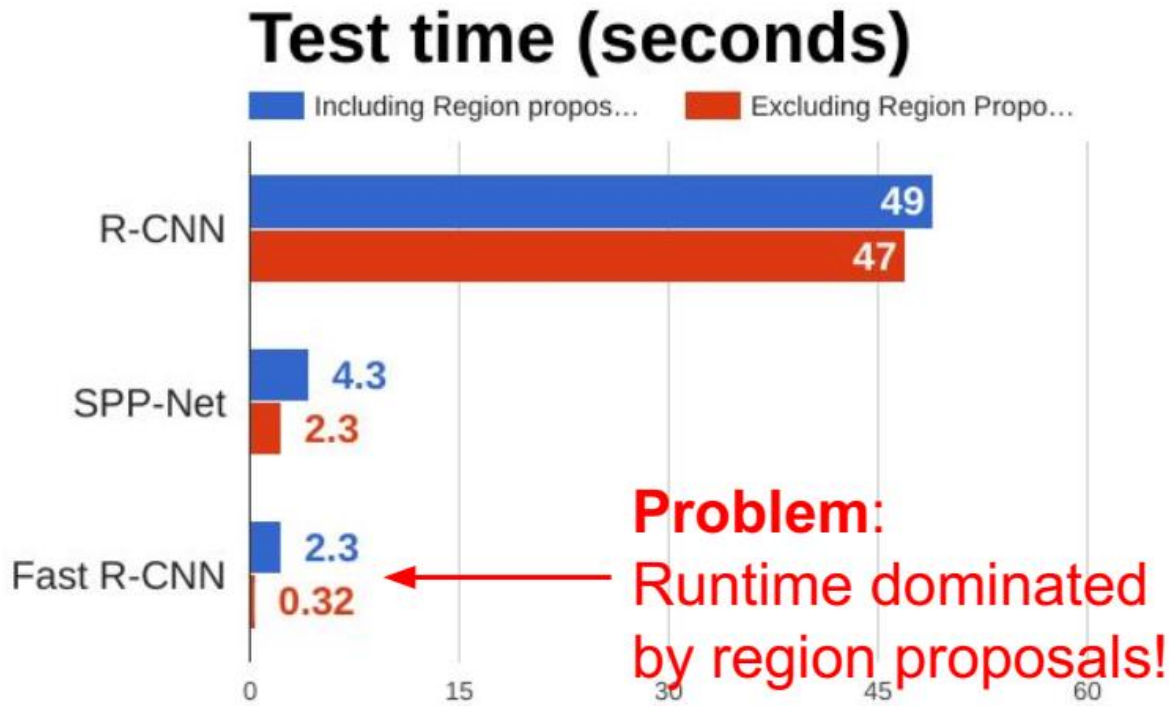
¿Cuál es el punto débil de este esquema?

Fast R-CNN





Girshick, "Fast R-CNN", ICCV 2015. Figure copyright Ross Girshick, 2015;



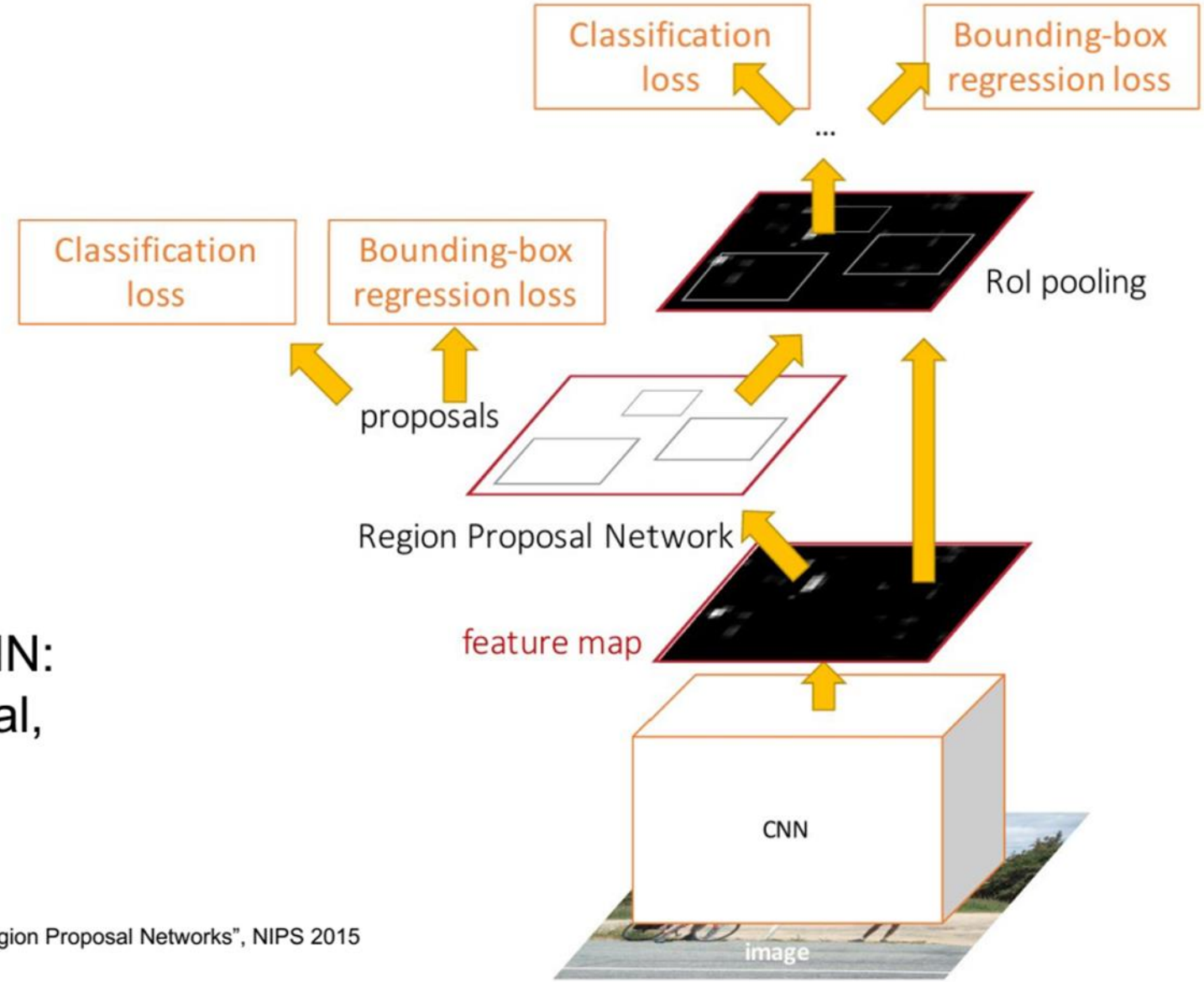
Girshick et al, “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014.
He et al, “Spatial pyramid pooling in deep convolutional networks for visual recognition”, ECCV 2014
Girshick, “Fast R-CNN”, ICCV 2015

Faster R-CNN:

Make CNN do proposals!

Insert **Region Proposal Network (RPN)** to predict proposals from features

Otherwise same as Fast R-CNN: Crop features for each proposal, classify each one



Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick;

Region Proposal Network

In practice use K different anchor boxes of different size / scale at each point

Input Image
(e.g. 3 x 640 x 480)

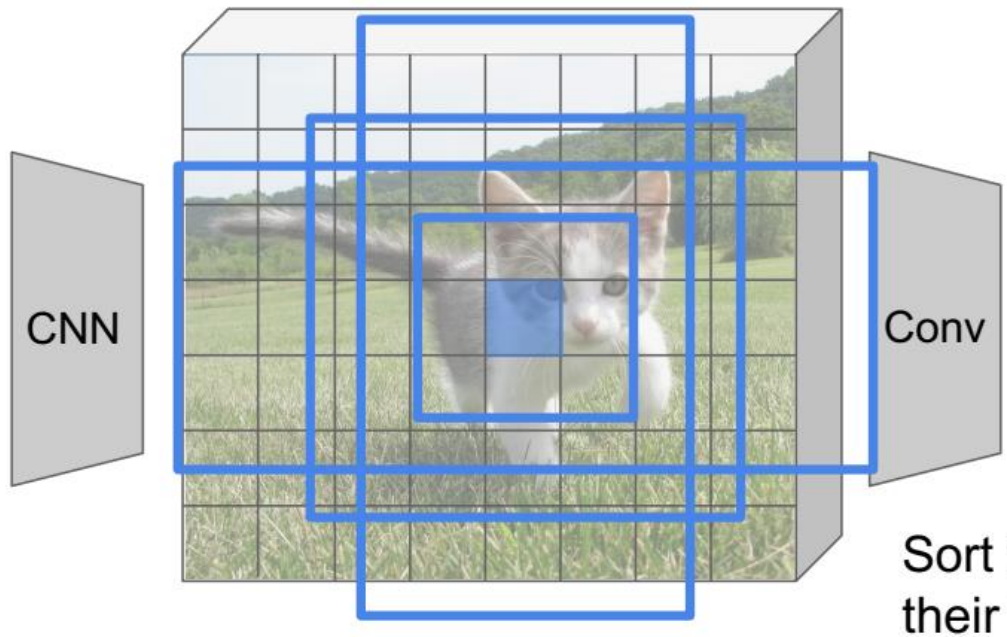


Image features
(e.g. 512 x 20 x 15)

Anchor is an object?
K x 20 x 15

Box transforms
4K x 20 x 15

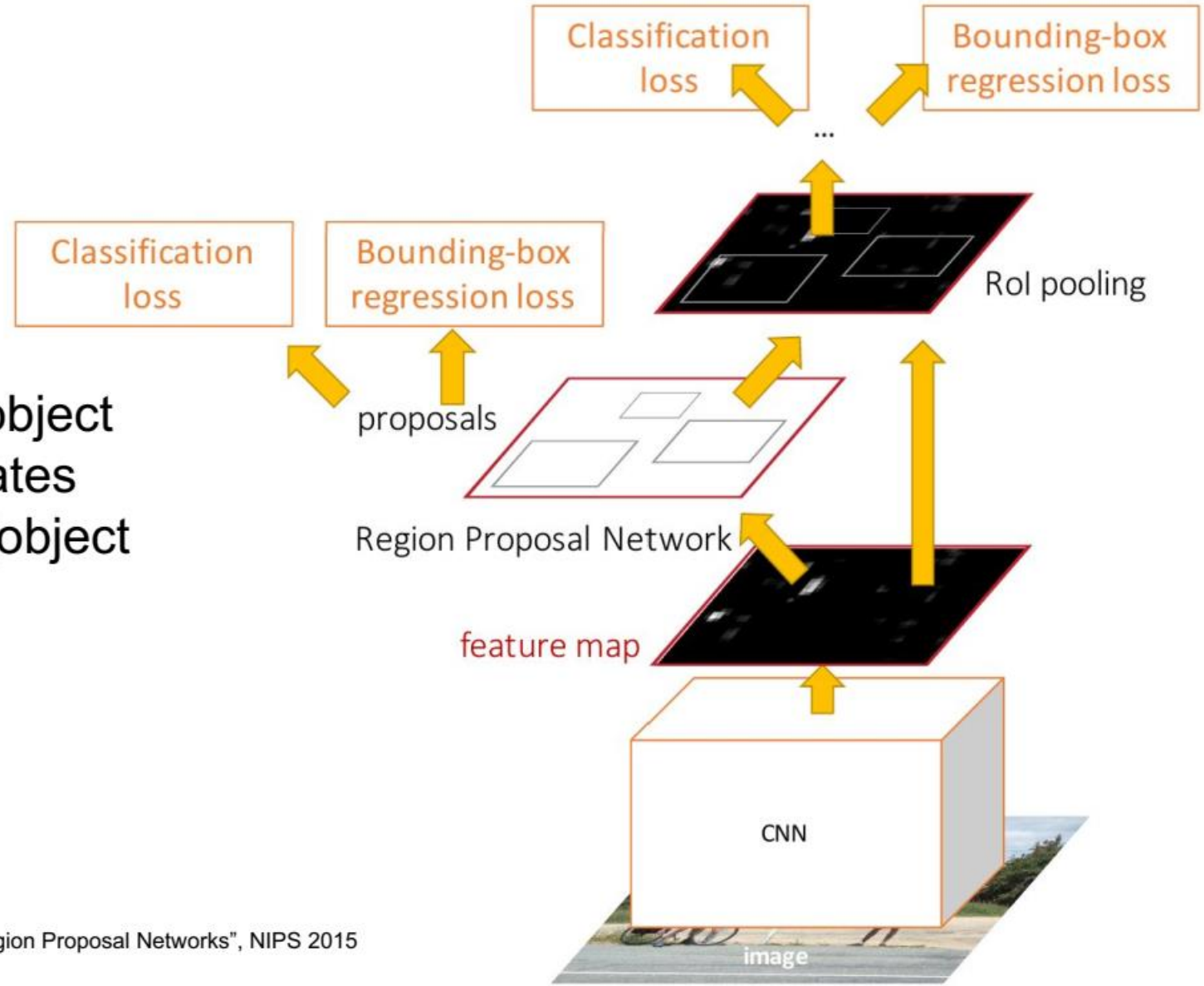
Sort the K*20*15 boxes by their “objectness” score, take top ~300 as our proposals

Faster R-CNN:

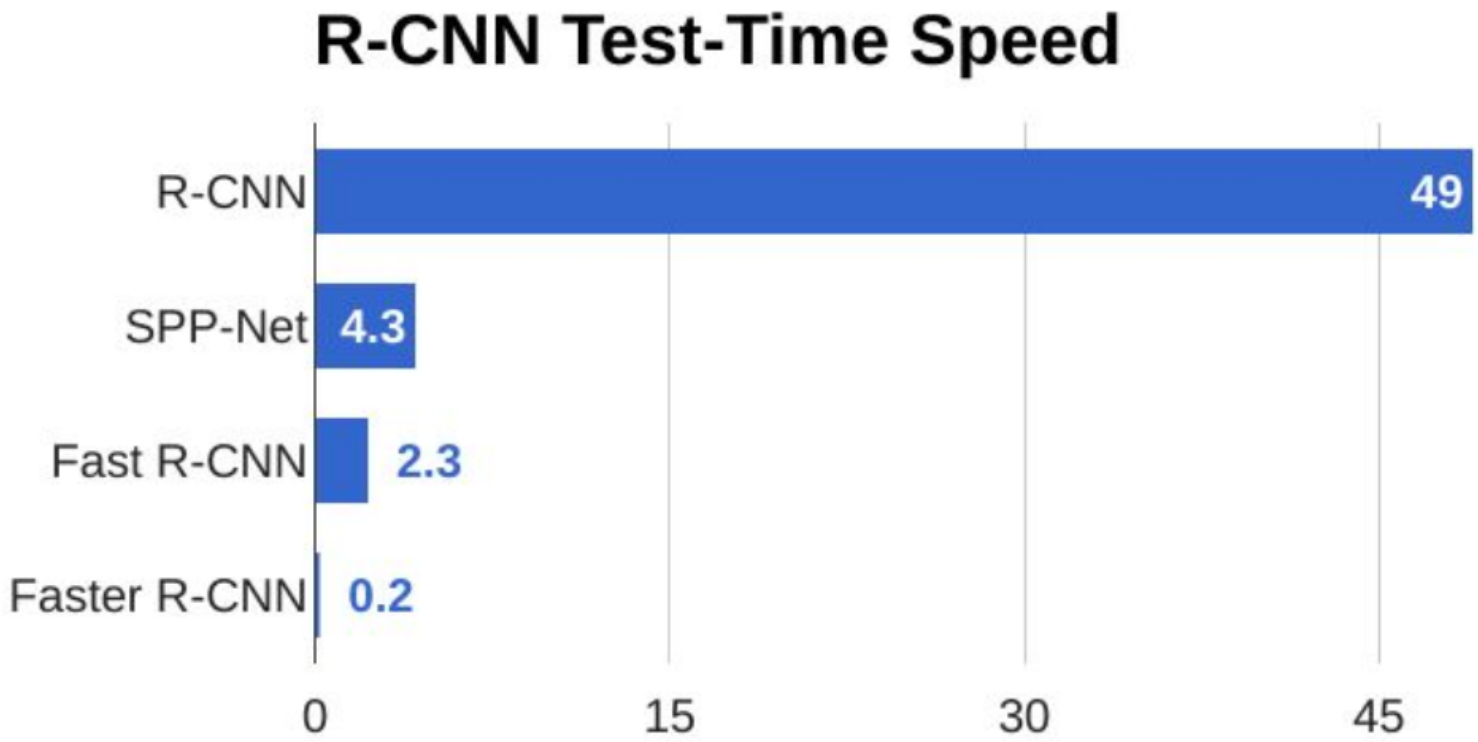
Make CNN do proposals!

Jointly train with 4 losses:

1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates



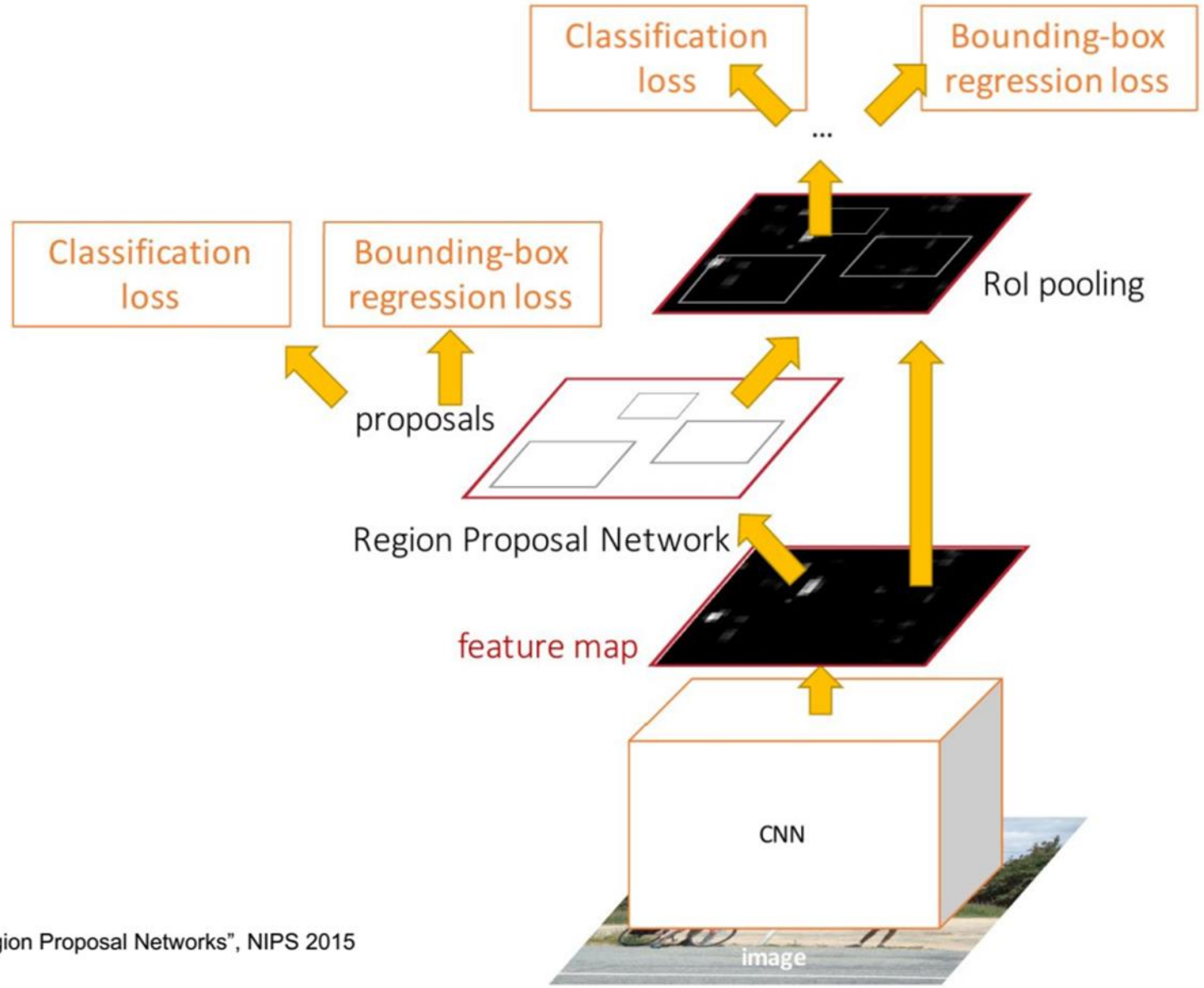
Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015


R-CNN Test-Time Speed
R-CNN
SPP-Net
Fast R-CNN
Faster R-CNN
49
4.3
2.3
0.2
0
15
30
45

Faster R-CNN:

Make CNN do proposals!

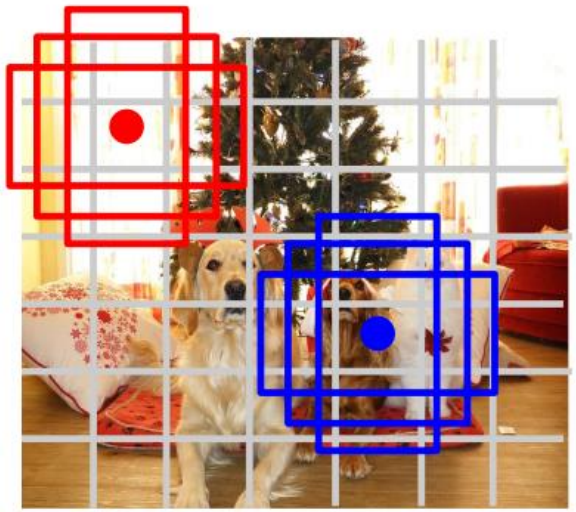
¿Necesitamos realmente la segunda subred?



Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

Single-Stage Object Detectors: YOLO / SSD / RetinaNet

Input image
3 x H x W



Divide image into grid
7 x 7

Image a set of **base boxes** centered at each grid cell
Here B = 3

Within each grid cell:

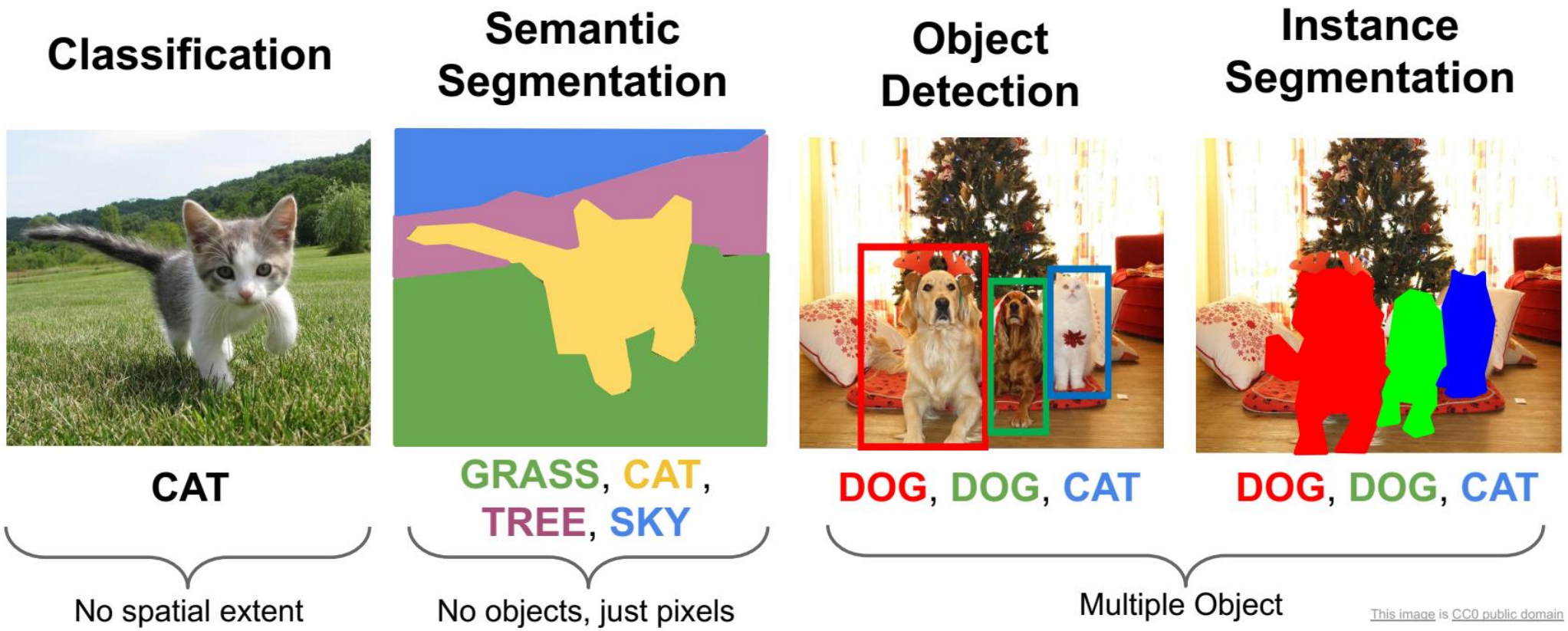
- Regress from each of the B base boxes to a final box with 5 numbers: (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)
- Looks a lot like RPN, but category-specific!

Output:
7 x 7 x (5 * B + C)

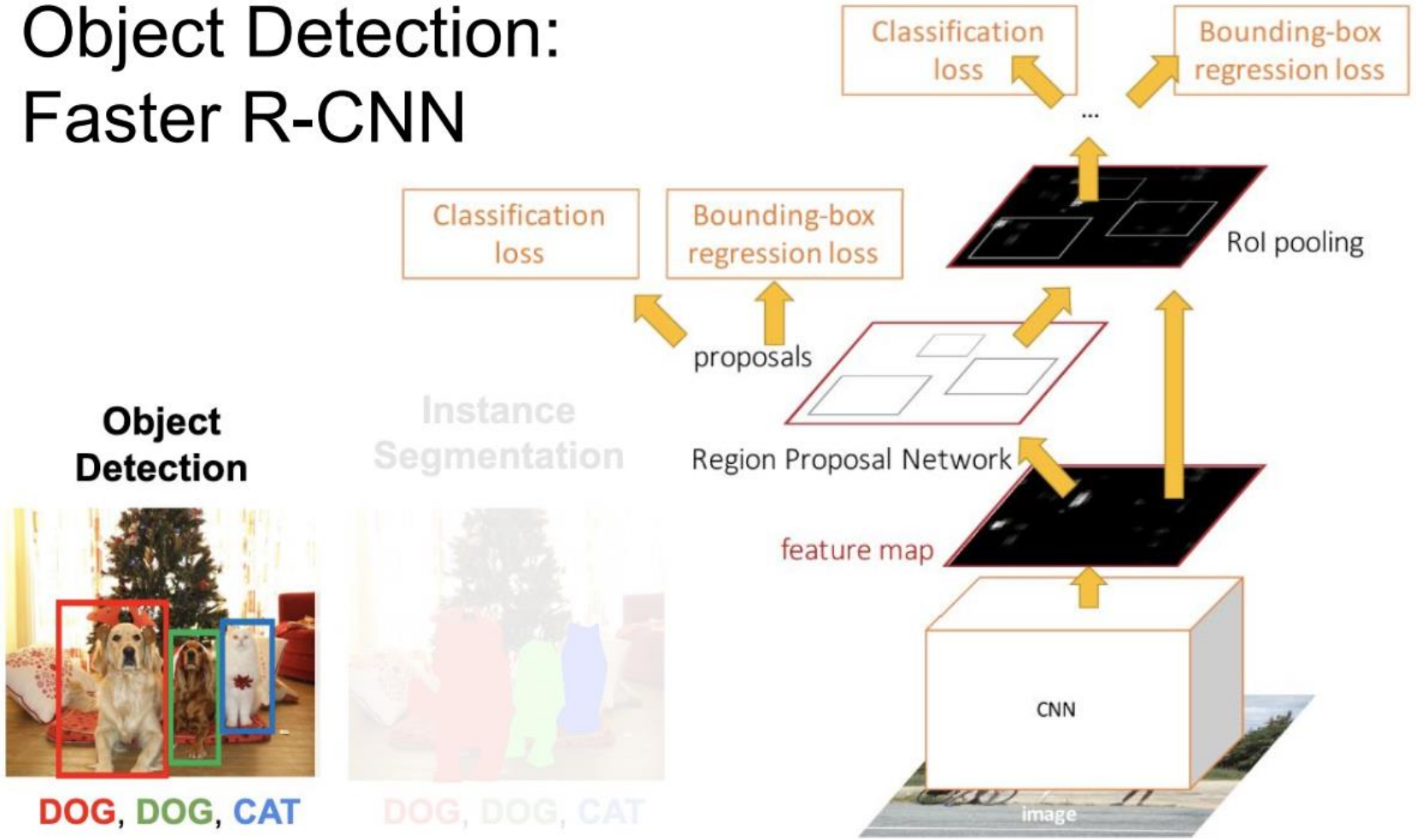
Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al, "SSD: Single-Shot MultiBox Detector", ECCV 2016
Lin et al, "Focal Loss for Dense Object Detection", ICCV 2017

https://youtu.be/YDFf-TqJOFE

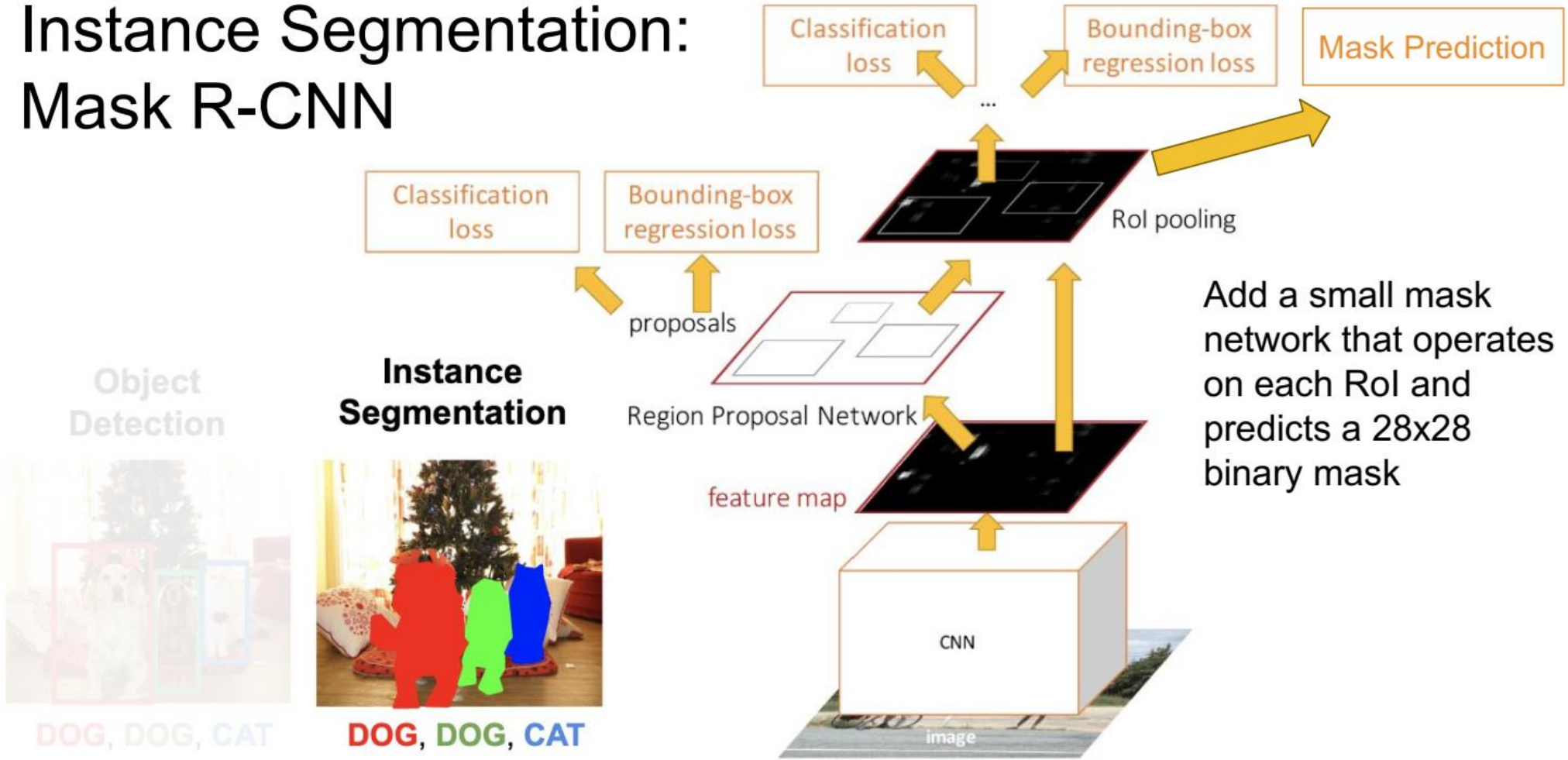
Para lograr la segmentación a nivel de instancia, utilizamos las mismas ideas



Object Detection: Faster R-CNN



Instance Segmentation: Mask R-CNN

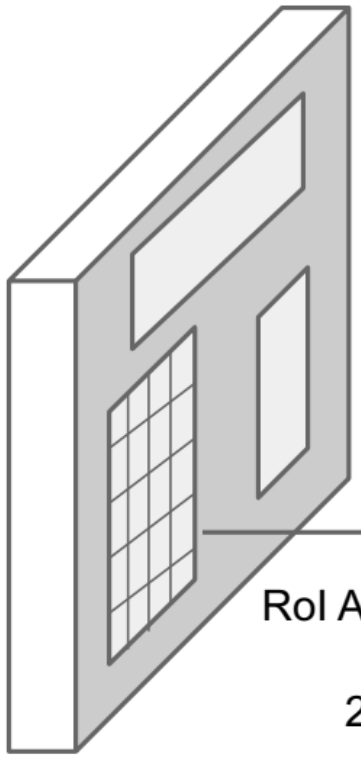


Add a small mask network that operates on each RoI and predicts a 28x28 binary mask

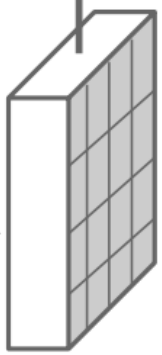
He et al, "Mask R-CNN", ICCV 2017

Mask R-CNN

CNN
+RPN

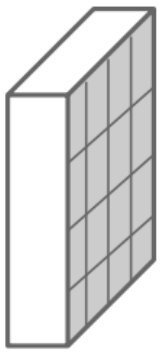


RoI Align



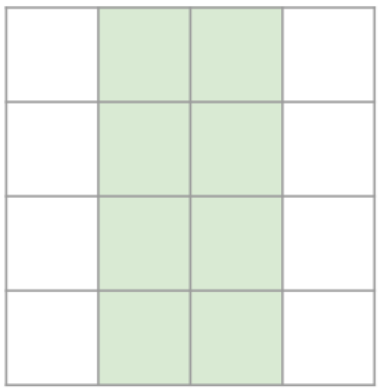
256 x 14 x 14

Conv



256 x 14 x 14

Conv

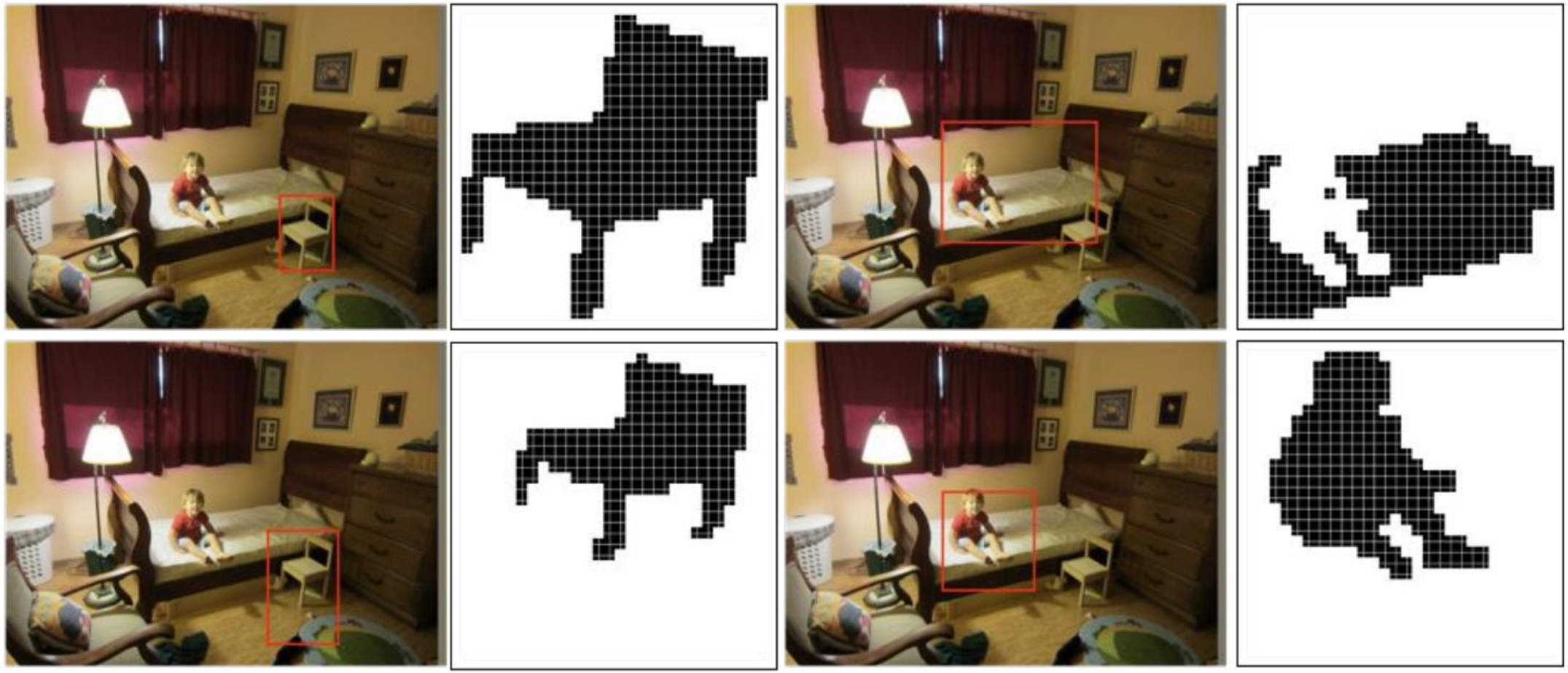


Classification Scores: C
Box coordinates (per class): 4 * C

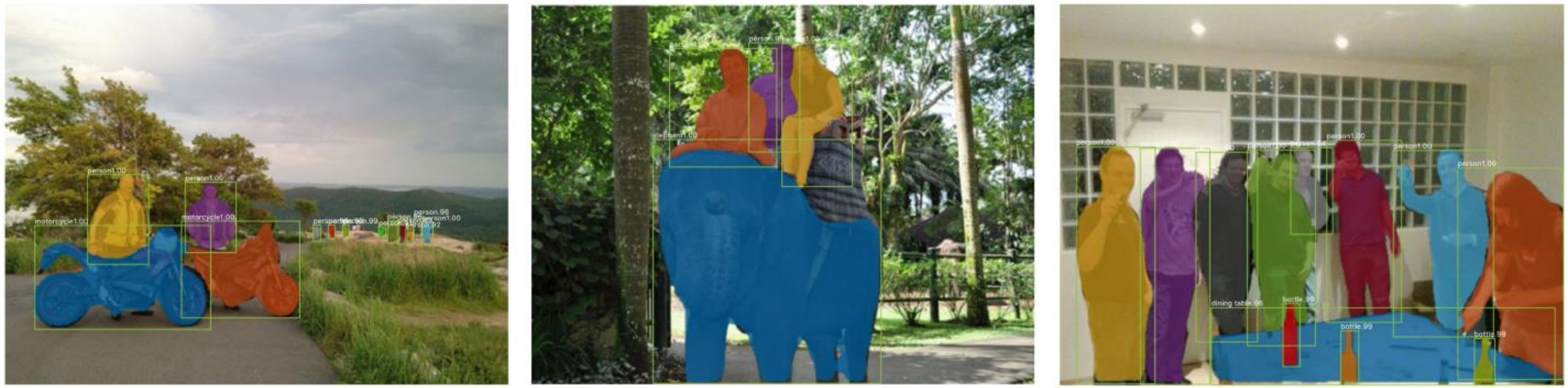
Predict a mask for
each of C classes

C x 28 x 28

He et al, “Mask R-CNN”, ICCV 2017



He et al, “Mask R-CNN”, ICCV 2017

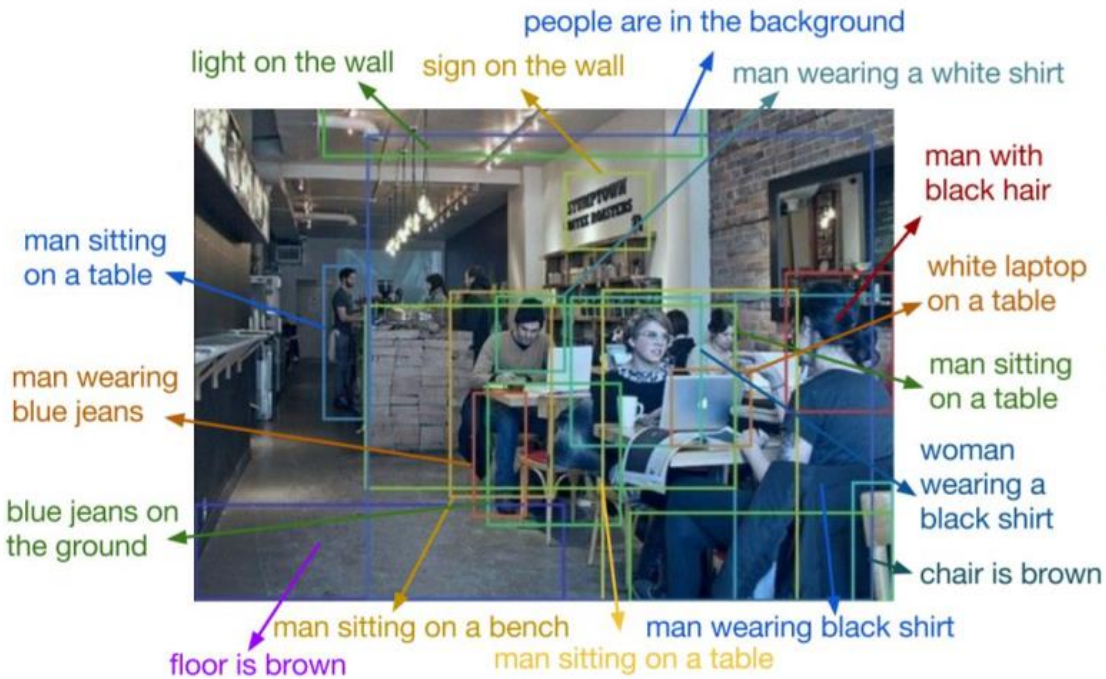


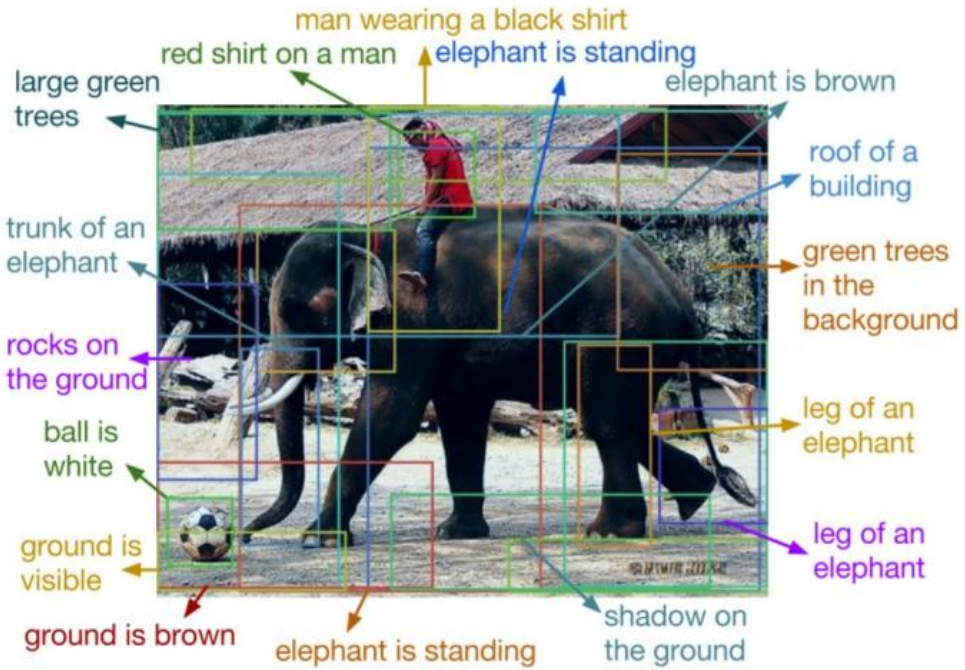
He et al, “Mask R-CNN”, ICCV 2017

He et al, “Mask R-CNN”, ICCV 2017

https://youtu.be/OOT3UIXZztE

Object Detection + Captioning = Dense Captioning





Johnson, Karpathy, and Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning", CVPR 2016

https://youtu.be/25jyl67-poQ

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Sistemas Urbanos Inteligentes

Aprendizaje Multitarea (Multitask Learning)

Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación