MAC5921 – Deep Learning

Aula 08 – 12/09/2023

Nina S. T. Hirata

Classification

Dog

Cat

Classification + Localization

Dog

Cat

Object Detection

Cat, Dog

Semantic Segmentation

Grass, Cat

Instance Segmentation

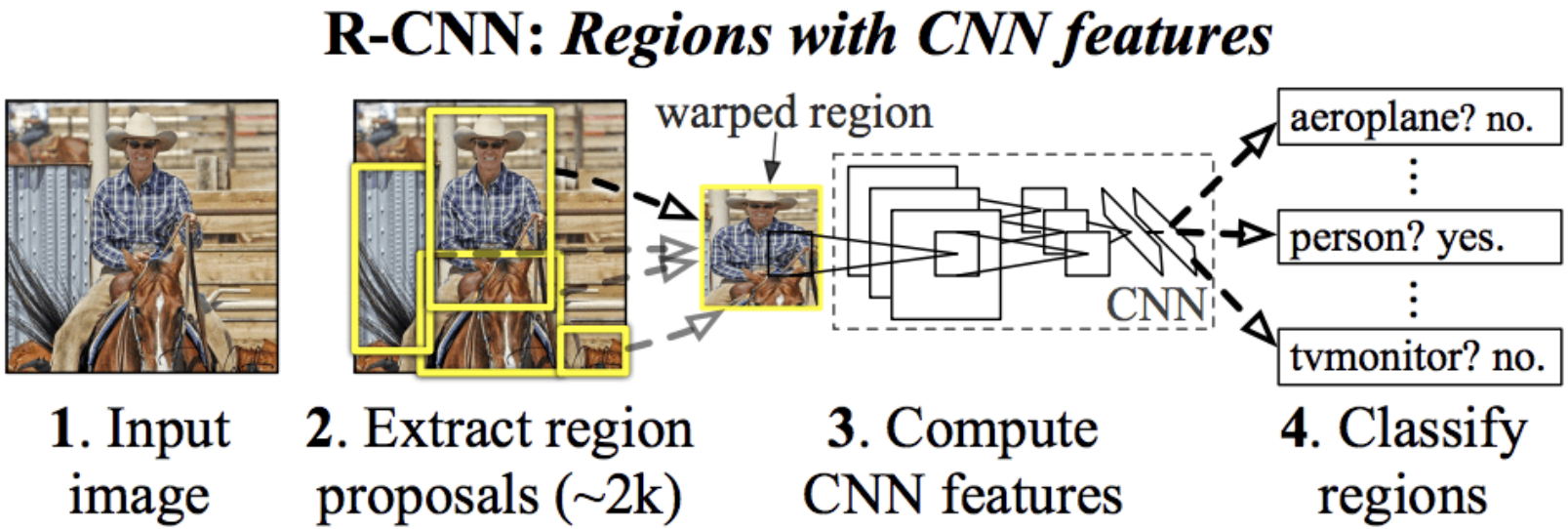
Cat, Cat, Cat, Cat, Cat

Object detection

Modelos de 2 estágios: família R-CNN

Modelos de 1 estágio: YOLO, SSD, ...

R-CNN



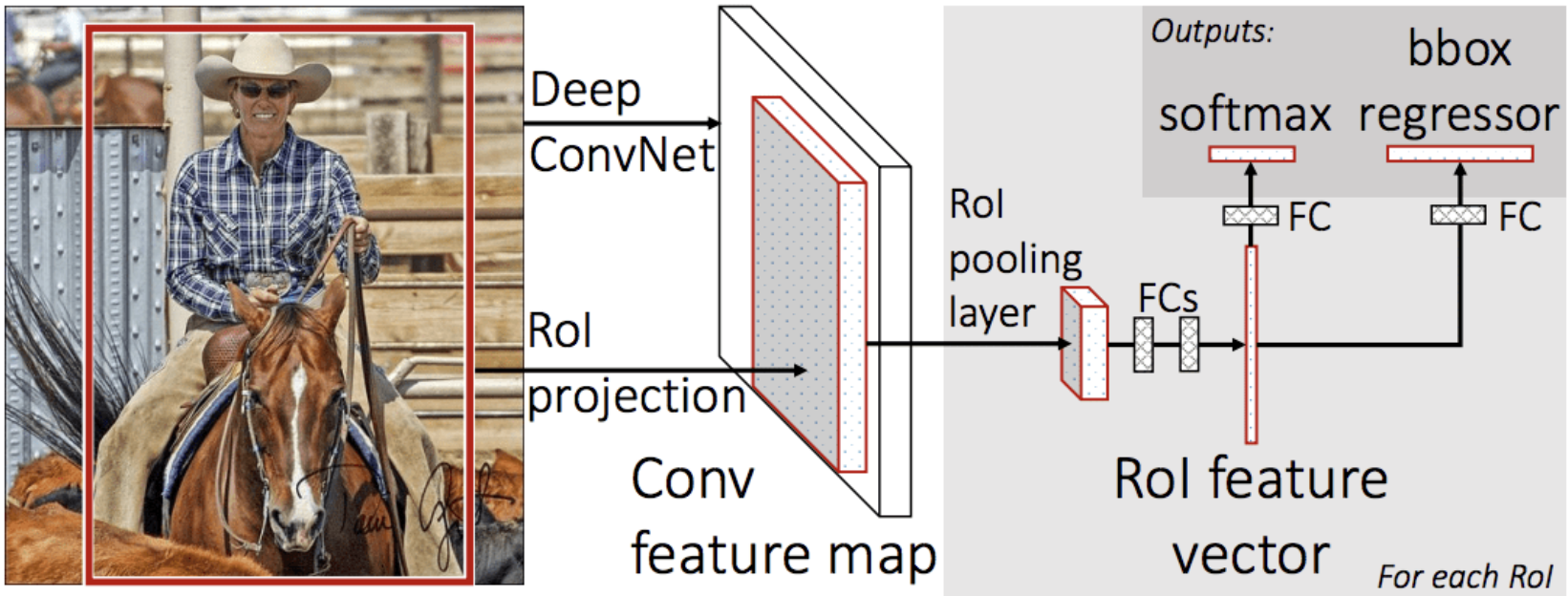
Rich feature hierarchies for accurate object detection and semantic segmentation

https://arxiv.org/abs/1311.2524

Cada region proposal precisa passar pela CNN de classificação

Isso é lento, se pensar que para cada imagem são consideradas cerca de 2000 region proposals

Fast R-CNN



Fast R-CNN

https://arxiv.org/abs/1504.08083

Passa uma imagem pela CNN apenas uma vez

O retângulo no mapa de features correspondente às region proposals são pooled para um tamanho padrão

Introduz-se saída adicional à rede: regressão relativa às coordenadas dos BBs

Target

u target class, v target BB

Saídas

Classificação: softmax scores para K classes

$$p = (p_0, p_1, \ldots, p_K)$$

Regressão: bounding box coordinates, $k = 0, 1, \ldots, K$

$$t^k = (t_x^k, t_y^k, t_w^k, t_h^k)$$

u target class, v target BB
p predicted class scores, t^u predicted BB for class u

Loss:

$$\mathcal{L}(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geqslant 1] L_{loc}(t^u, v)$$

$$L_{cls}(p, u) = -\log p_u$$

$$L_{loc}(t^u, v) = \sum_{i \in \{x,y,w,h\}} \text{smooth}_{L_1}(t_i^u - v_i)$$

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5\,x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}$$

Sempre há detalhes ...

(x, y) é top-left coordinate do BB

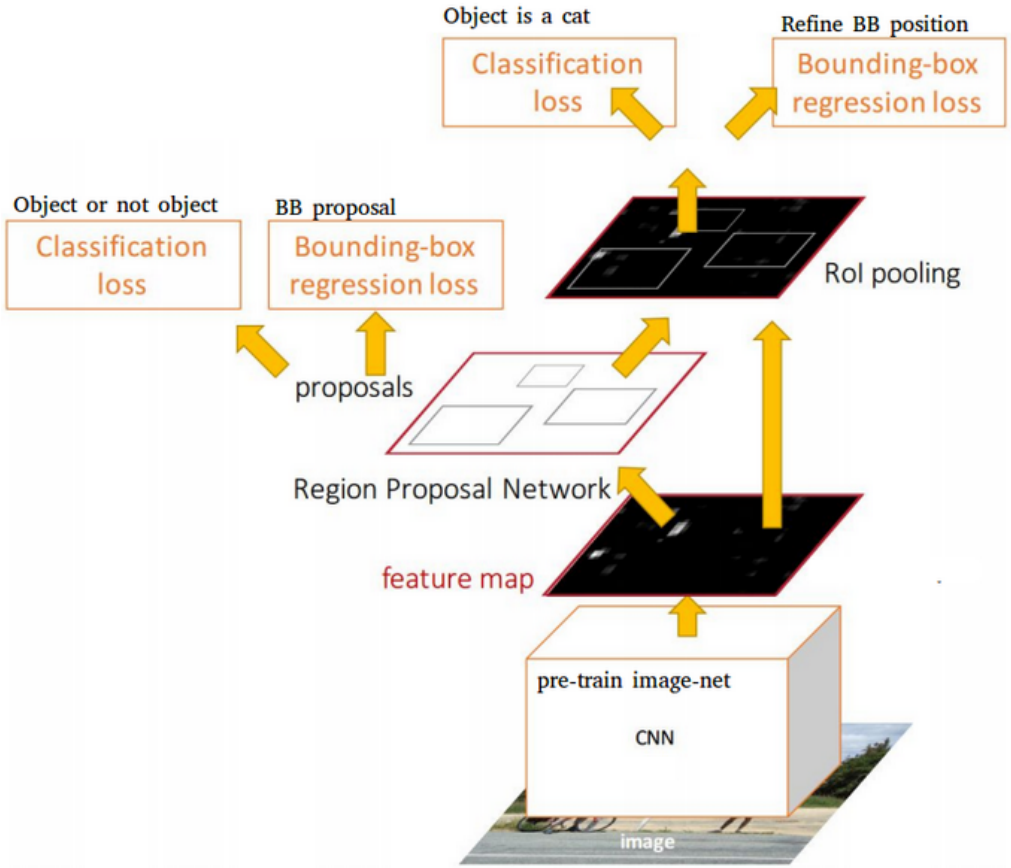
Mas em outros trabalhos, é o centro

E em outros, os quatro valores do BB são top-left e right-bottom coordinates

L_2 loss pode levar a exploding gradient

Normaliza ground-truth BB variables v_i para ter média 0 e variância 1

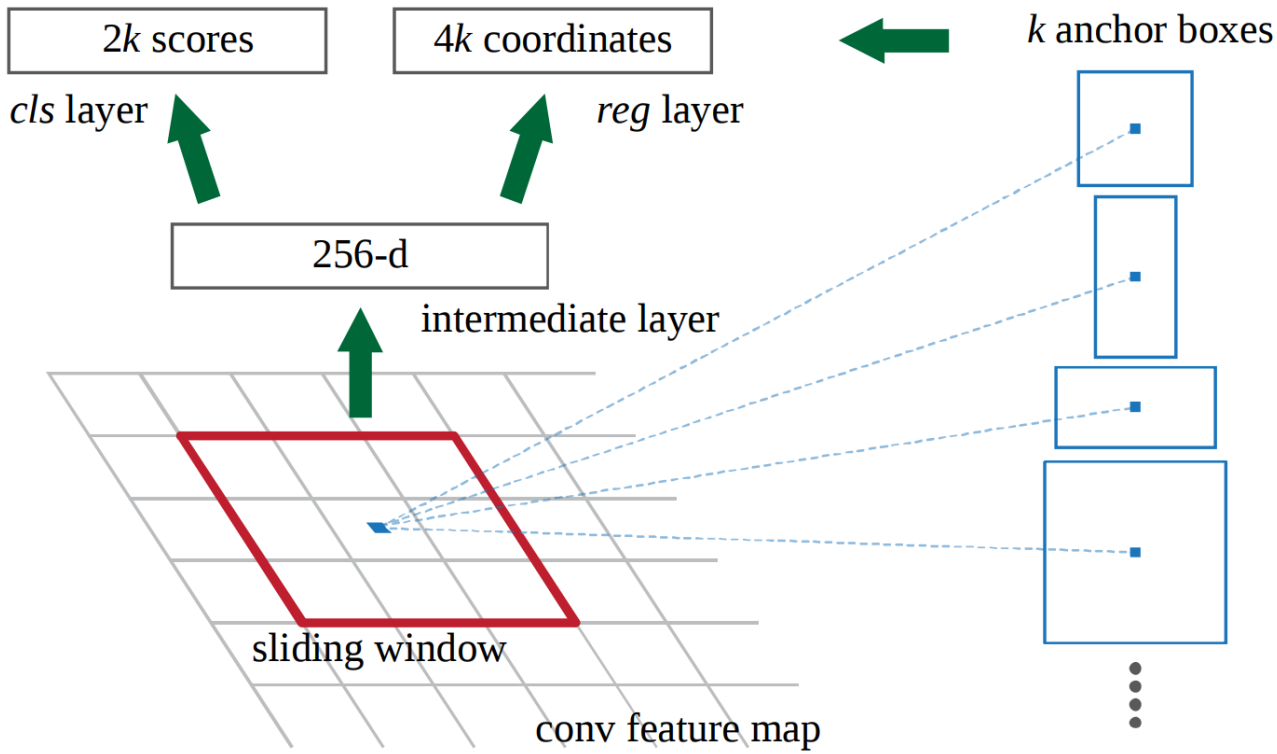
Faster R-CNN = RPN + Fast R-CNN



Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

https://arxiv.org/abs/1506.01497

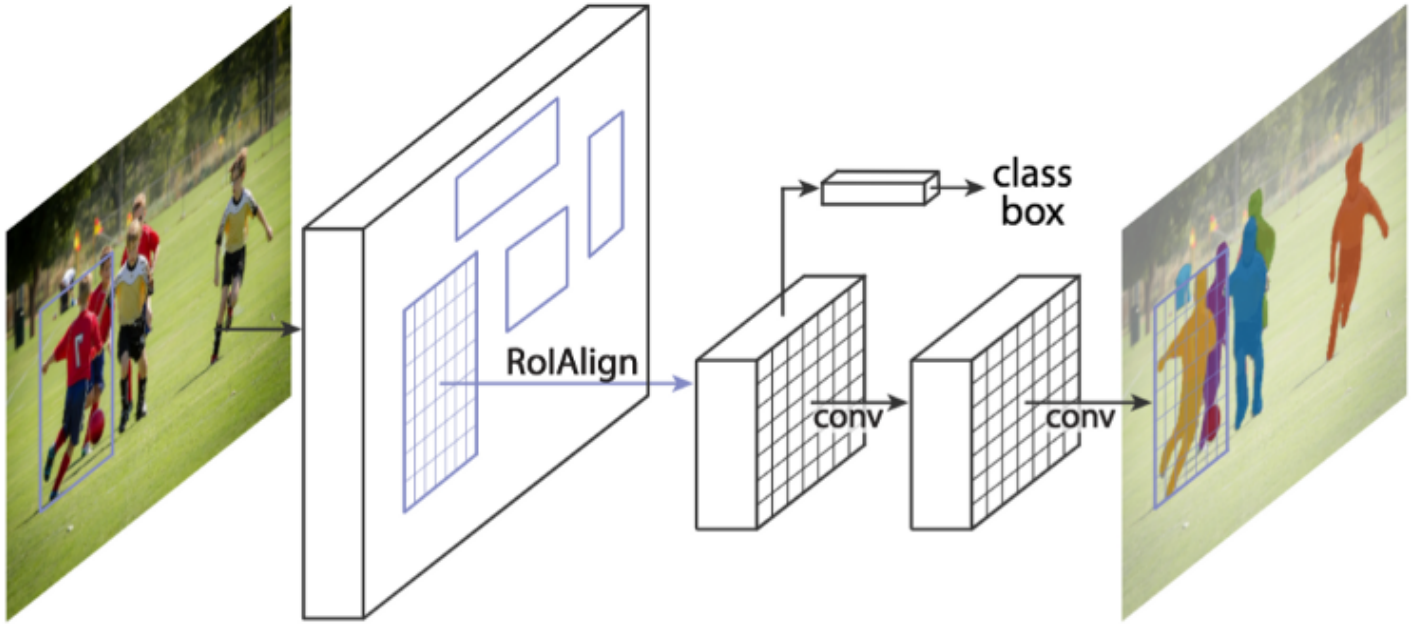
RPN – Region Proposal Network



Quais bounding box enviar para a parte de classificação?

Non-maximum supression

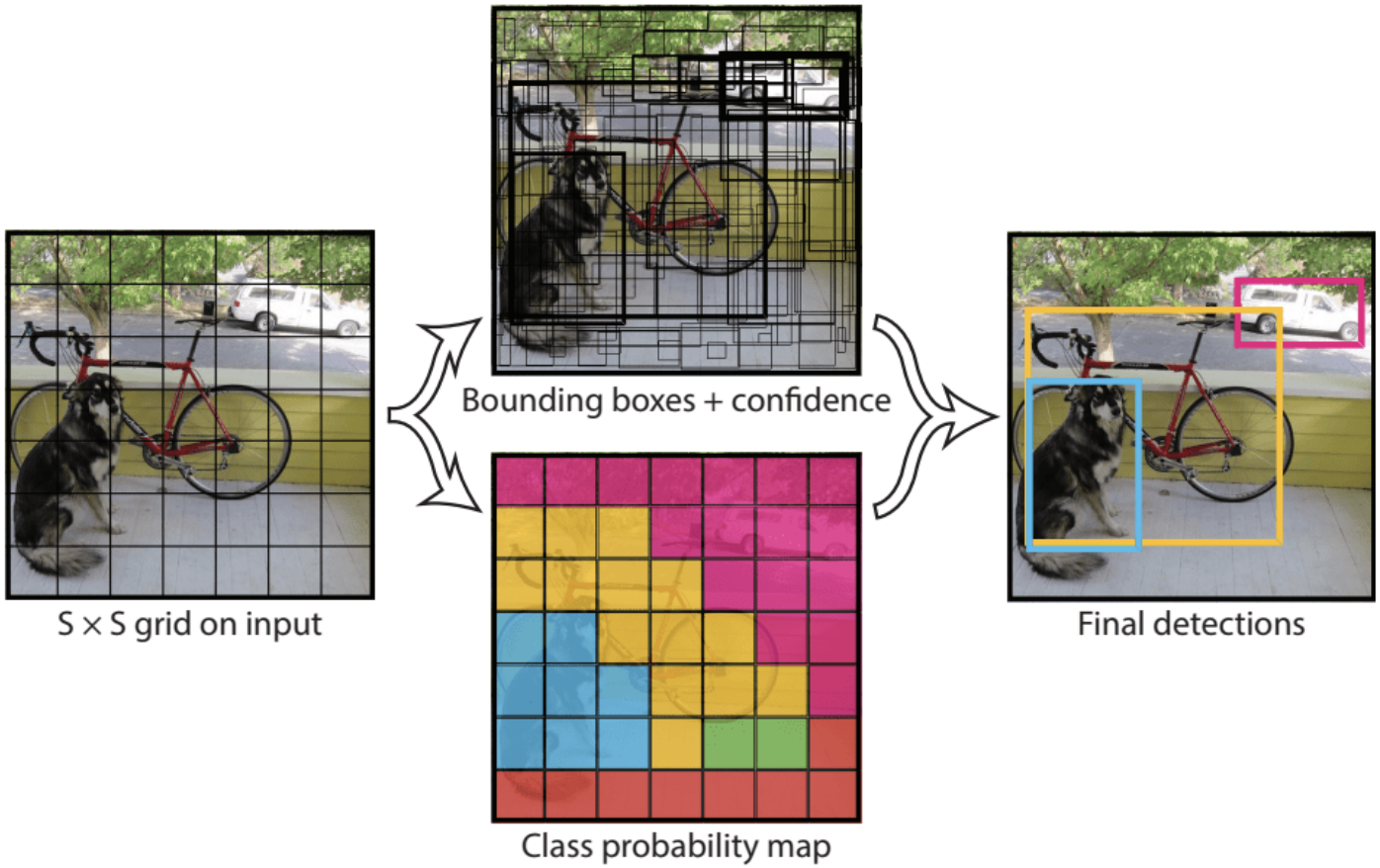
Mask RCNN



Mask R-CNN

https://arxiv.org/abs/1703.06870

YOLO – You Only Look Once



You Only Look Once: Unified, Real-Time Object Detection

https://arxiv.org/abs/1506.02640

Treinamento:

"Basta" minimizar a loss

Predição: o que são exatamente os resultados preditos e como comparar predições de diferentes modelos ??

mAP (mean average precision) https://github.com/rafaelpadilla/Object-Detection-Metrics