RNP-VC

Aula 6: Detecção e Localização

Lucas Pereira, Rafael Teixeira, Lucas Assis, Anderson Soares
Instituto de Informática
Universidade Federal de Goiás (UFG)

Sumário

- No último episódio...
- Tarefas em visão computacional
- Detecção de 1 objeto
- Detecção de múltiplos objetos
- No próximo episódio...

No último episódio...

- Model zoo

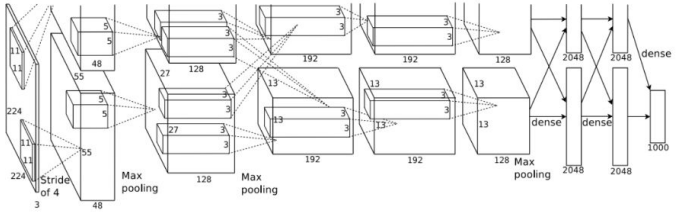
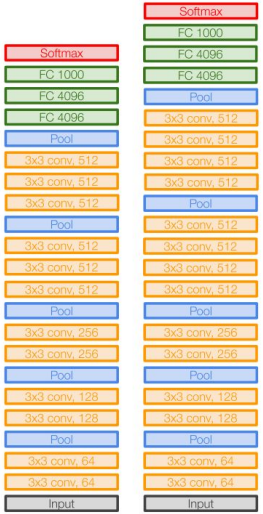
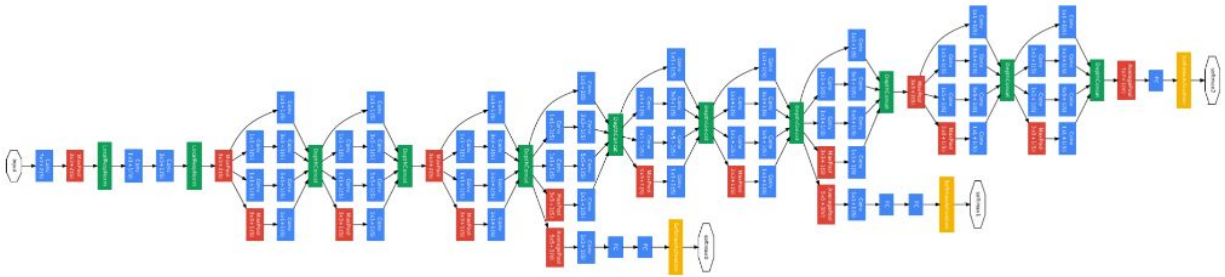
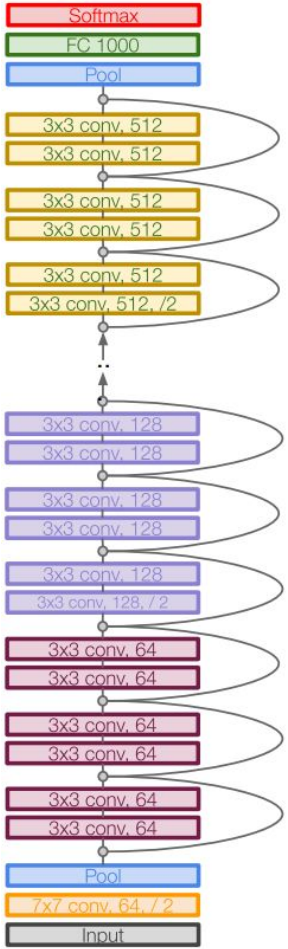


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.







No último episódio...

- Transfer Learning

	Datasets Similares	**Datasets Distintos**
Muitos dados disponíveis	Treine algumas camadas do modelo base e o classificador de saída	Treine um número maior de camadas (ou todas elas)
Poucos dados disponíveis	Treine apenas o classificador de saída	É... Houston, we have a problem

Tarefas em visão computacional

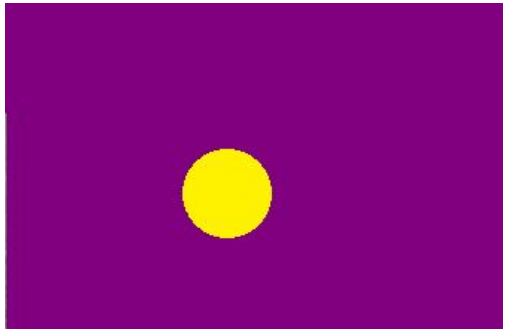
- Até agora: classificação

Tarefas em visão computacional

- Agora falta “só”...

Detecção e Localização

Segmentação

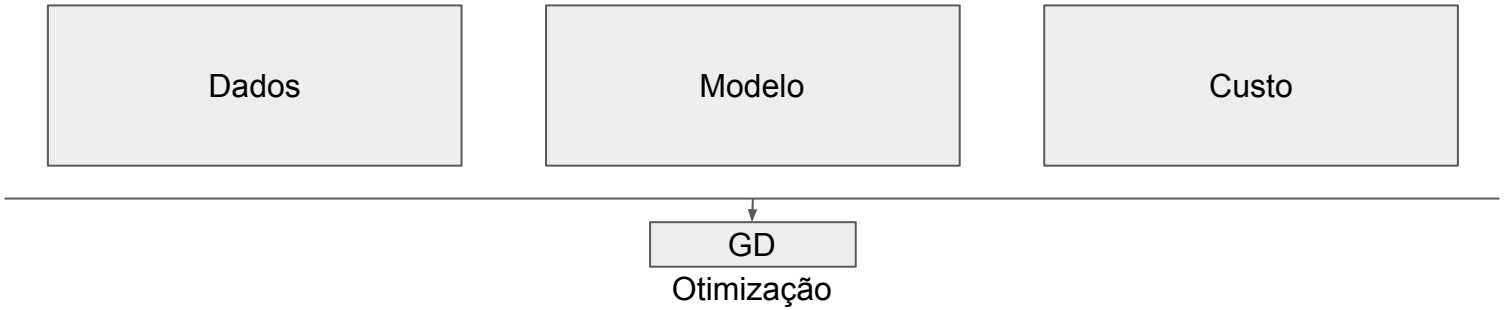


Tarefas em visão computacional

- Hoje: detecção e localização

Detecção de 1 objeto

- Estrutura de ML



Detecção de 1 objeto

- Como são os dados?

Detecção de 1 objeto

- Como são os dados?

Dados

Label:

$x_{min} = 35, y_{min} = 53, x_{max} = 52, y_{max} = 71$

ou

$x_{centro} = 43, y_{centro} = 62, w = 17, h = 18$

Detecção de 1 objeto

- Como são os dados?

Dados

Label:

x_{min} = ?, y_{min} = ?, x_{max} = ?, y_{max} = ?

Detecção de 1 objeto

- Como são os dados?

Dados

Label:

x_{min} = ?, y_{min} = ?, x_{max} = ?, y_{max} = ?

Detecção de 1 objeto

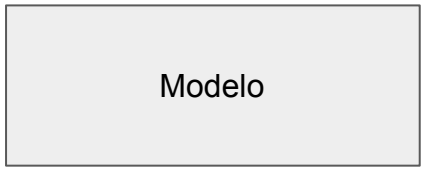
- Como são os dados?

Dados

imagem	classe	x_{min}	y_{min}	x_{max}	y_{max}
bola_0.jpg	bola	35	53	52	71
bandeira_0.jpg	bandeira	21	25	33	100
sem_objeto_0.jpg	background	nan	nan	nan	nan
…	…	…	…	…	…

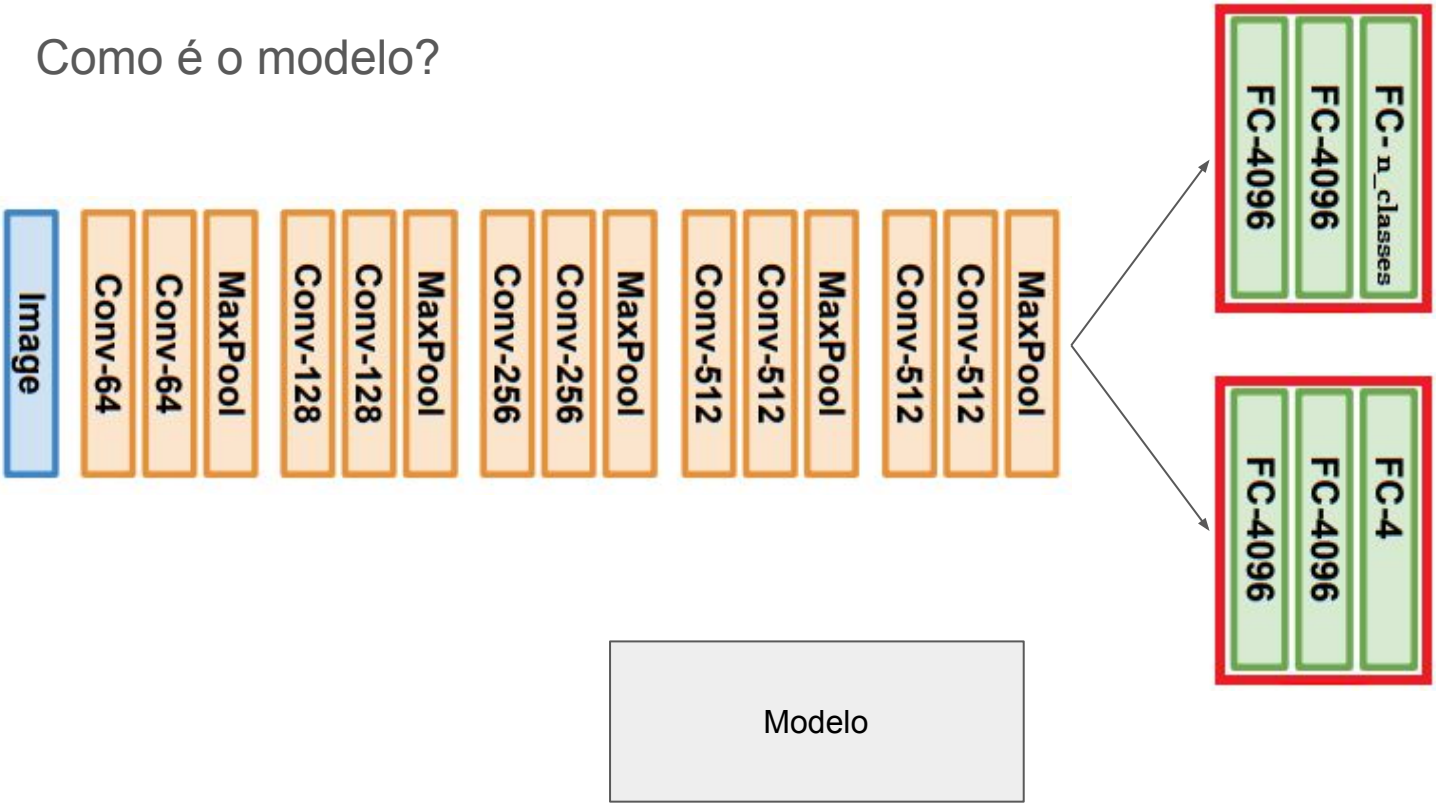
Detecção de 1 objeto

- Como é o modelo?



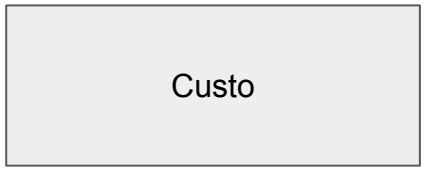
Detecção de 1 objeto

- Como é o modelo?



Detecção de 1 objeto

- Como é a função de custo?



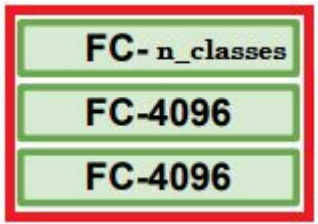
Detecção de 1 objeto

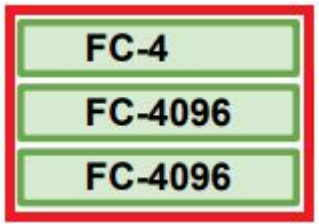
- Como é a função de custo?

Classificação:
- CE
- BCE
- Hinge
- ...

Regressão:
- L2
- L1
- ...

$$\mathcal{L}_{total} = \mathcal{L}_{class} + \lambda \mathcal{L}_{bbox}$$



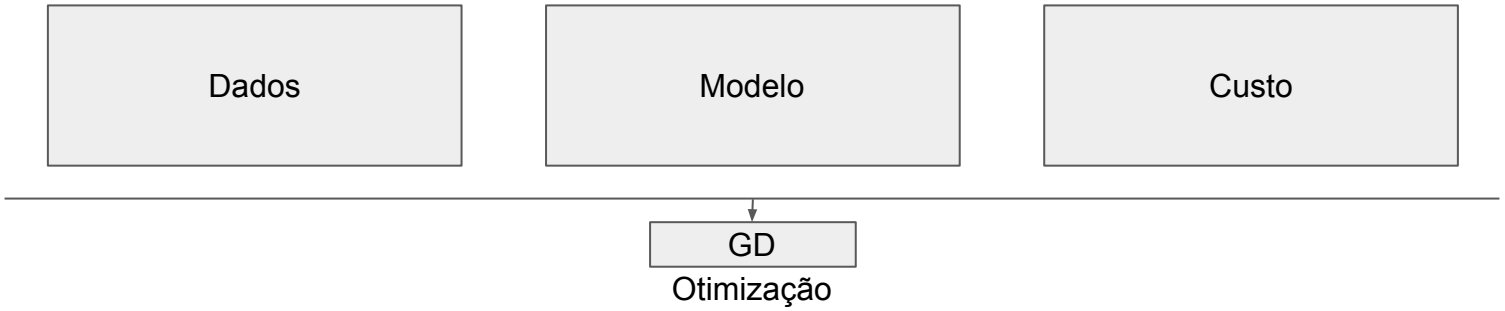


Custo

Detecção de 1 objeto

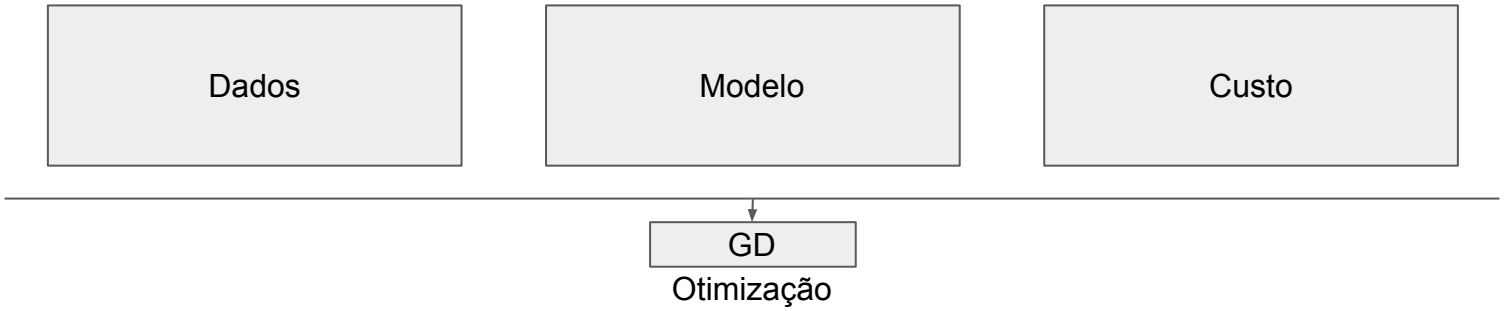
- Otimização

$$\mathcal{L}_{total} = \mathcal{L}_{class} + \lambda \mathcal{L}_{bbox}$$



Detecção de multiplos objetos

- Estrutura de ML



Detecção de múltiplos objetos

- Como são os dados?

Detecção de múltiplos objetos

- Como são os dados?

Dados

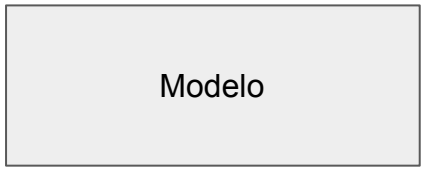
Detecção de múltiplos objetos

- Como são os dados?

Dados

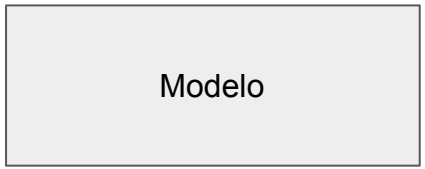
Detecção de múltiplos objetos

- Como é o modelo?



Detecção de múltiplos objetos

- Como é o modelo?
  - Modelos com 2 estágios (R-CNN, Fast R-CNN, Faster R-CNN)
  - Modelos com 1 estágio (YOLO, SSD, RetinaNet)



Detecção de múltiplos objetos

- R-CNN
  - R - Region Proposal (Selective Search)
    ✚
  - CNN - Redes convolucionais

arXiv:1311.2524v5 [cs.CV] 22 Oct 2014

Rich feature hierarchies for accurate object detection and semantic segmentation

Tech report (v5)

Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik
UC Berkeley
{rbg,jdonahue,trevor,malik}@eecs.berkeley.edu

Abstract

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012—achieving a mAP of 53.3%. Our approach combines two key insights: (1) one can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost. Since we combine region proposals with CNNs, we call our method R-CNN: *Regions with CNN features. We also compare R-CNN to OverFeat, a recently proposed sliding-window detector based on a similar CNN architecture. We find that R-CNN outperforms OverFeat by a large margin on the 200-class ILSVRC2013 detection dataset. Source code for the complete system is available at* http://www.cs.berkeley.edu/~rbg/rcnn.

1. Introduction

Features matter. The last decade of progress on various

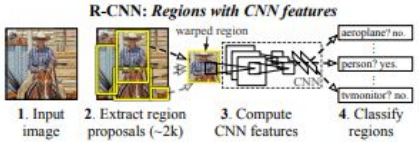


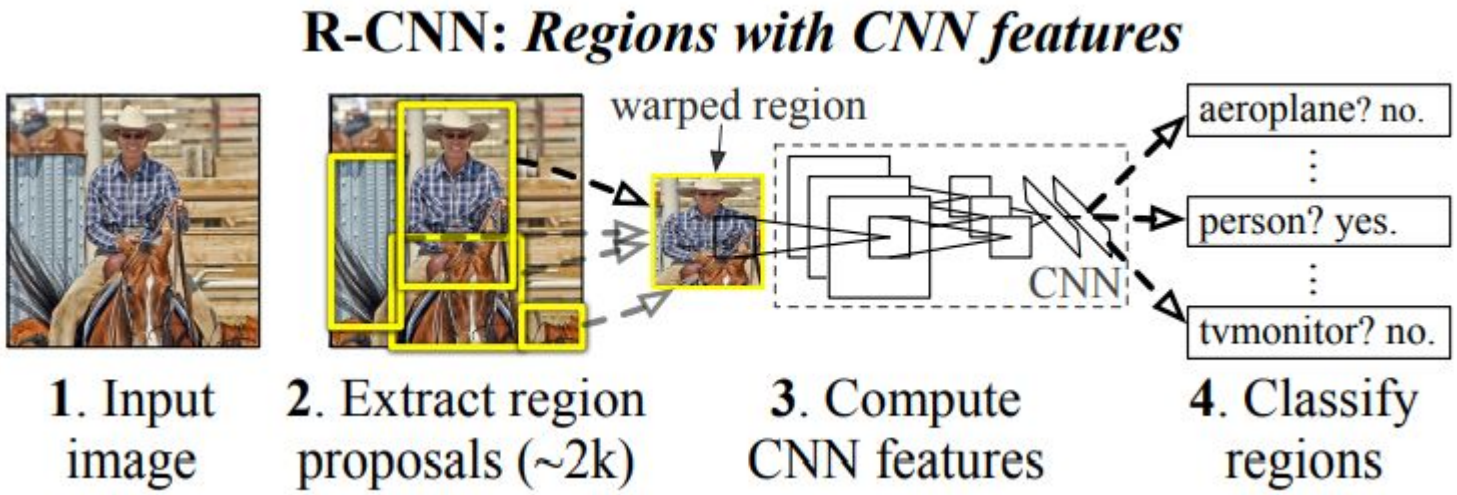
Figure 1: Object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs. R-CNN achieves a mean average precision (mAP) of **53.7% on PASCAL VOC 2010**. For comparison, [39] reports 35.1% mAP using the same region proposals, but with a spatial pyramid and bag-of-visual-words approach. The popular deformable part models perform at 33.4%. On the 200-class **ILSVRC2013 detection dataset, R-CNN's mAP is 31.4%**, a large improvement over OverFeat [34], which had the previous best result at 24.3%.

archical, multi-stage processes for computing features that are even more informative for visual recognition.

Fukushima's "neocognitron" [19], a biologically-inspired hierarchical and shift-invariant model for pattern recognition, was an early attempt at just such a process. The neocognitron, however, lacked a supervised training algorithm. Building on Rumelhart et al. [33], LeCun et al. [26] showed that stochastic gradient descent via back-

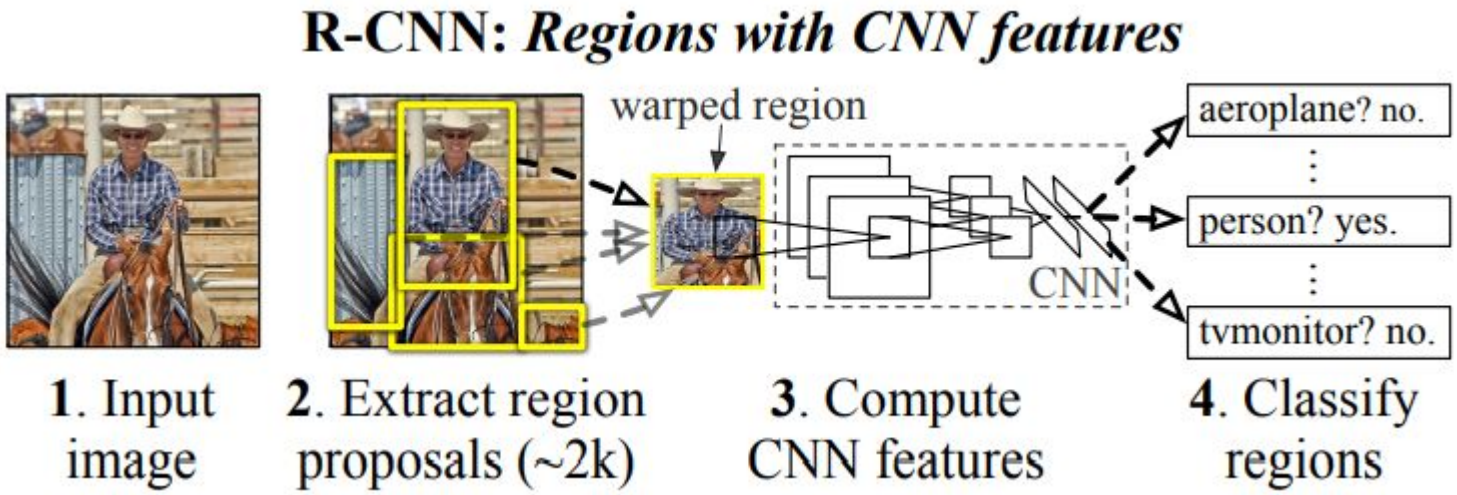
Detecção de múltiplos objetos

- R-CNN



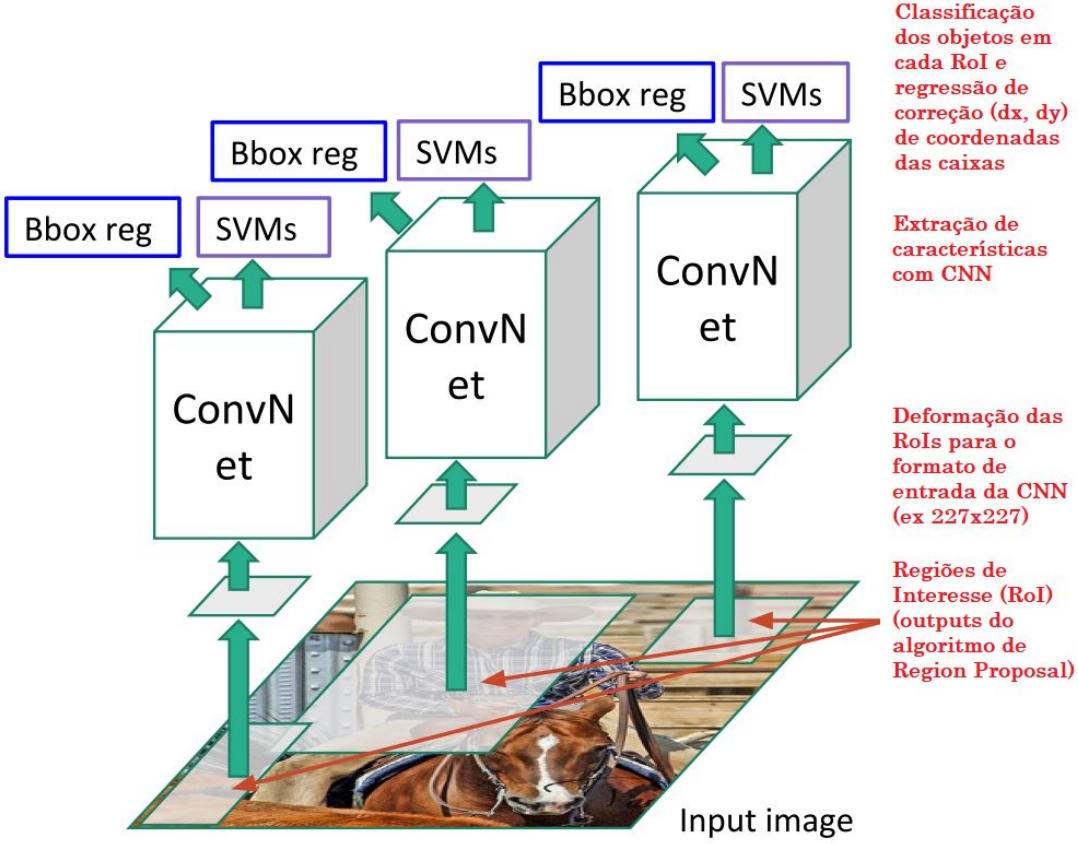
Detecção de múltiplos objetos

- R-CNN



Detecção de múltiplos objetos

- R-CNN



Detecção de múltiplos objetos

- Fast R-CNN

arXiv:1504.08083v2 [cs.CV] 27 Sep 2015

Fast R-CNN

Ross Girshick
Microsoft Research
rbg@microsoft.com

Abstract

This paper proposes a Fast Region-based Convolutional Network method (Fast R-CNN) *for object detection. Fast R-CNN builds on previous work to efficiently classify object proposals using deep convolutional networks. Compared to previous work, Fast R-CNN employs several innovations to improve training and testing speed while also increasing detection accuracy. Fast R-CNN trains the very deep VGG16 network 9× faster than R-CNN, is 213× faster at test-time, and achieves a higher mAP on PASCAL VOC 2012. Compared to SPPnet, Fast R-CNN trains VGG16 3× faster, tests 10× faster, and is more accurate. Fast R-CNN is implemented in Python and C++ (using Caffe) and is available under the open-source MIT License at https://github.com/rbgirshick/fast-rcnn.*

1. Introduction

Recently, deep ConvNets [14, 16] have significantly improved image classification [14] and object detection [9, 19] accuracy. Compared to image classification, object detection is a more challenging task that requires more complex methods to solve. Due to this complexity, current approaches (*e.g.*, [9, 11, 19, 25]) train models in multi-stage

while achieving top accuracy on PASCAL VOC 2012 [7] with a mAP of 66% (vs. 62% for R-CNN).[1]

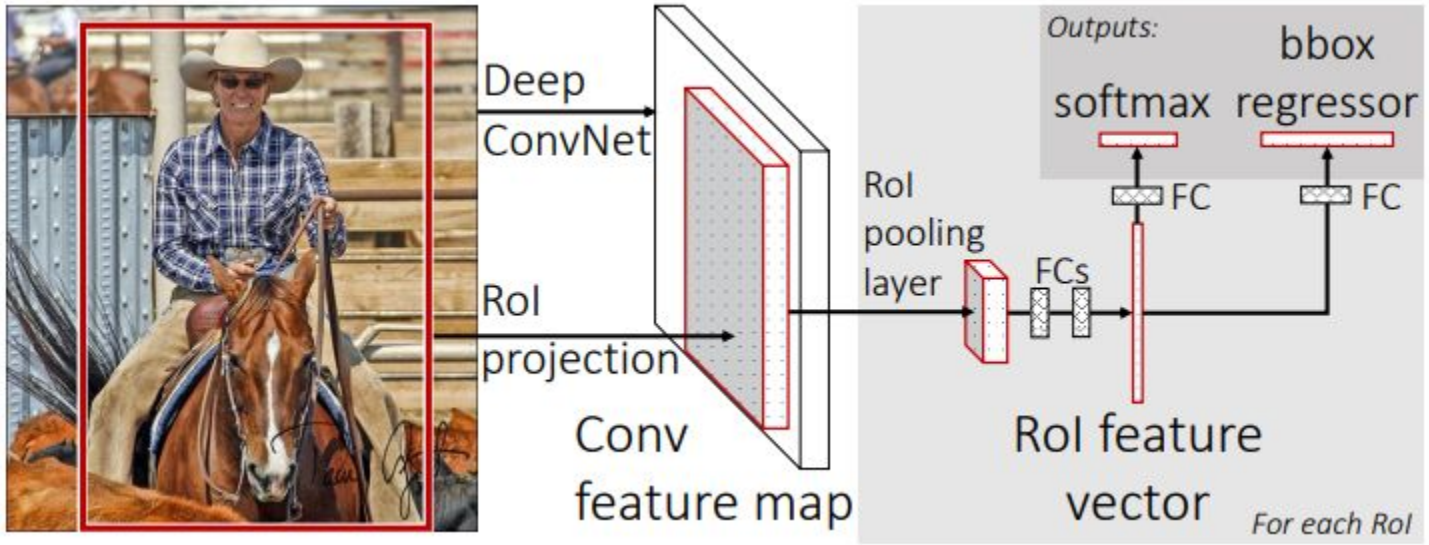
1.1. R-CNN and SPPnet

The Region-based Convolutional Network method (R-CNN) [9] achieves excellent object detection accuracy by using a deep ConvNet to classify object proposals. R-CNN, however, has notable drawbacks:

1. **Training is a multi-stage pipeline.** R-CNN first fine-tunes a ConvNet on object proposals using log loss. Then, it fits SVMs to ConvNet features. These SVMs act as object detectors, replacing the softmax classifier learnt by fine-tuning. In the third training stage, bounding-box regressors are learned.
2. **Training is expensive in space and time.** For SVM and bounding-box regressor training, features are extracted from each object proposal in each image and written to disk. With very deep networks, such as VGG16, this process takes 2.5 GPU-days for the 5k images of the VOC07 trainval set. These features require hundreds of gigabytes of storage.
3. **Object detection is slow.** At test-time, features are extracted from each object proposal in each test image. Detection with VGG16 takes 47s / image (on a GPU).

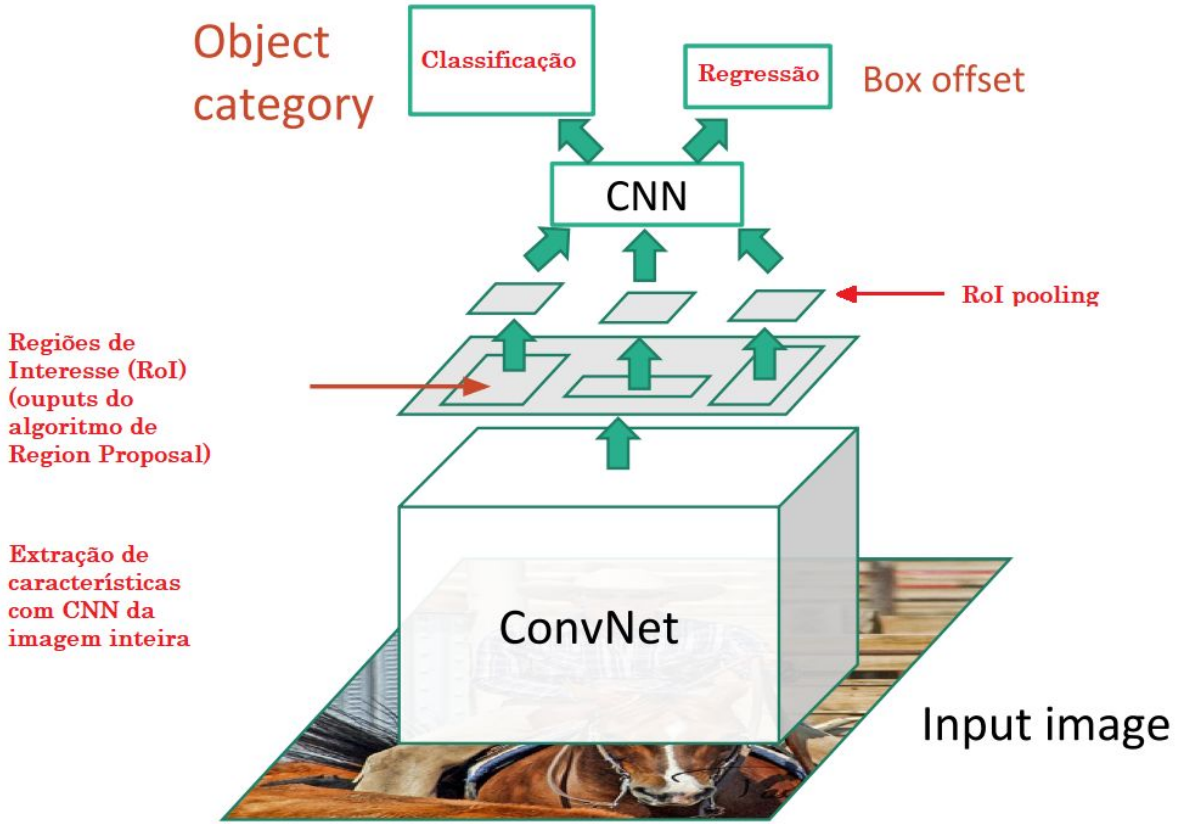
Detecção de múltiplos objetos

- Fast R-CNN



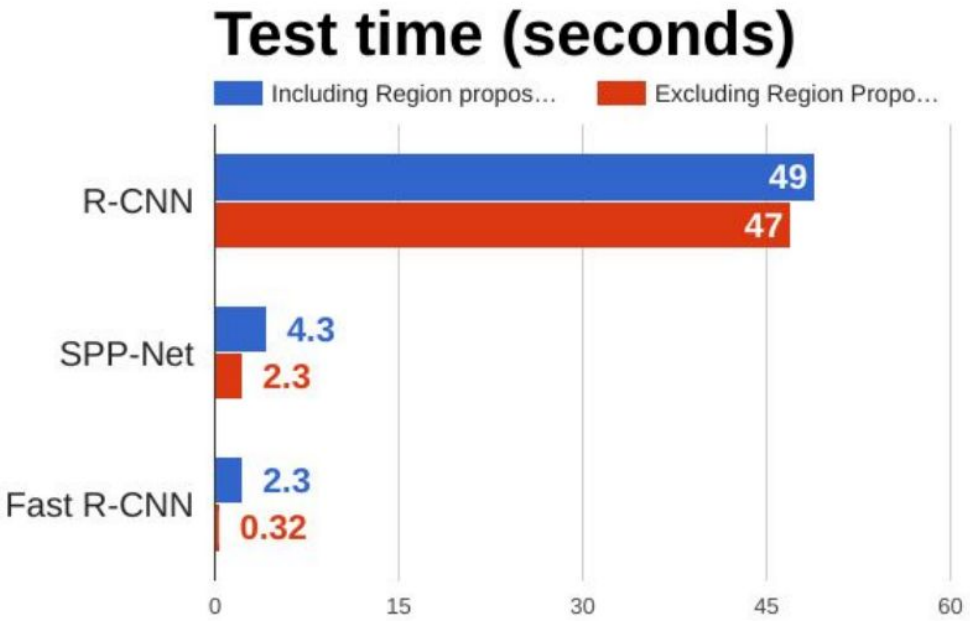
Detecção de múltiplos objetos

- Fast R-CNN



Detecção de múltiplos objetos

- R-CNN vs Fast R-CNN



Detecção de múltiplos objetos

- Faster R-CNN

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

Shaoqing Ren* **Kaiming He** **Ross Girshick** **Jian Sun**
Microsoft Research
{v-shren, kahe, rbg, jiansun}@microsoft.com

Abstract

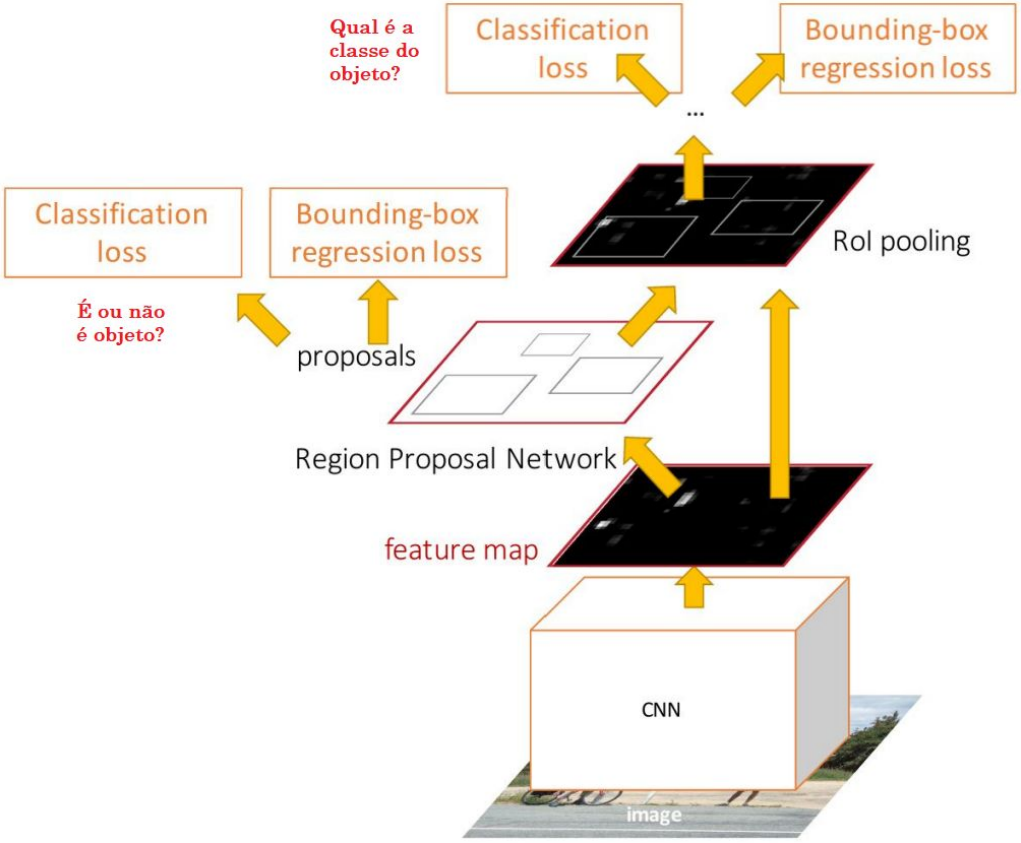
State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet [7] and Fast R-CNN [5] have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, we introduce a *Region Proposal Network* (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully-convolutional network that simultaneously predicts object bounds and objectness scores at each position. RPNs are trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. With a simple alternating optimization, RPN and Fast R-CNN can be trained to share convolutional features. For the very deep VGG-16 model [19], our detection system has a frame rate of 5fps (*including all steps*) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007 (73.2% mAP) and 2012 (70.4% mAP) using 300 proposals per image. Code is available at `https://github.com/ShaoqingRen/faster_rcnn`.

1 Introduction

Recent advances in object detection are driven by the success of region proposal methods (*e.g.*, [22]) and region-based convolutional neural networks (R-CNNs) [6]. Although region-based CNNs were computationally expensive as originally developed in [6], their cost has been drastically reduced thanks to sharing convolutions across proposals [7, 5]. The latest incarnation, Fast R-CNN [5], achieves near real-time rates using very deep networks [19], *when ignoring the time spent on region proposals*. Now, proposals are the computational bottleneck in state-of-the-art detection systems.

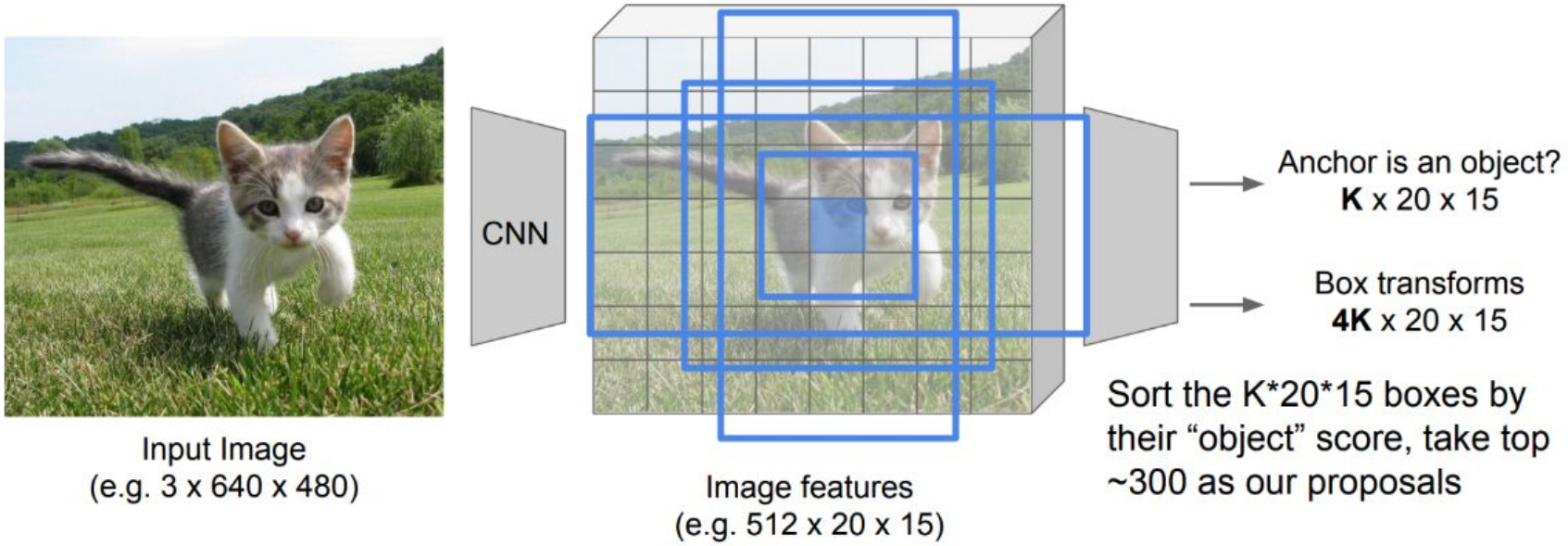
Detecção de múltiplos objetos

- Faster R-CNN



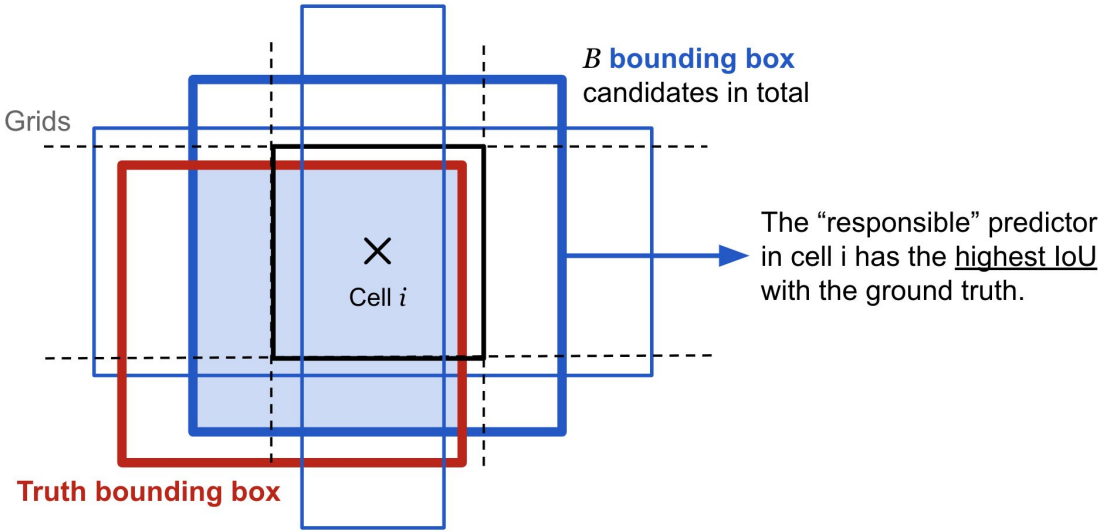
Detecção de múltiplos objetos

- Faster R-CNN



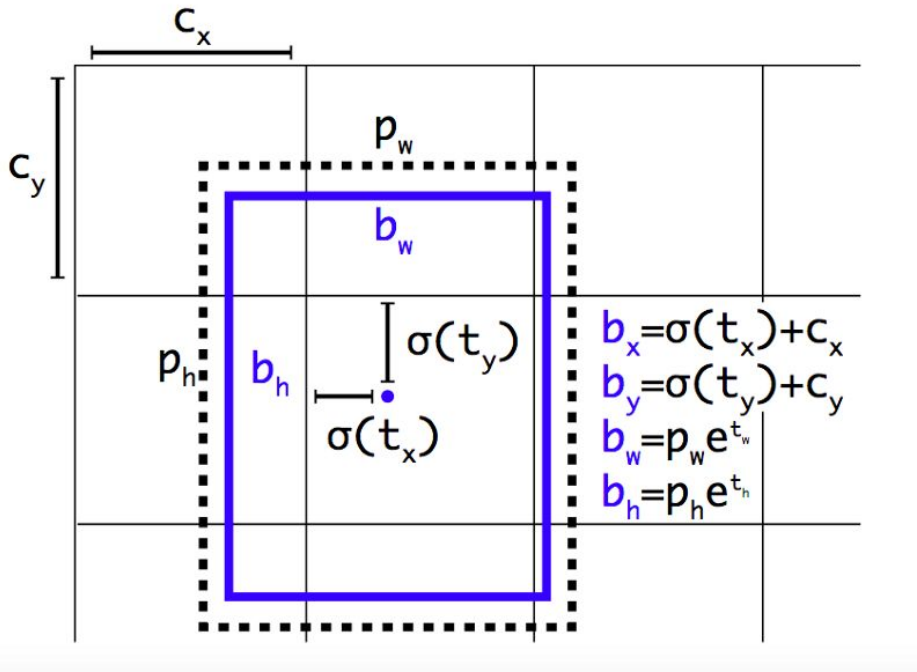
Detecção de múltiplos objetos

- Faster R-CNN



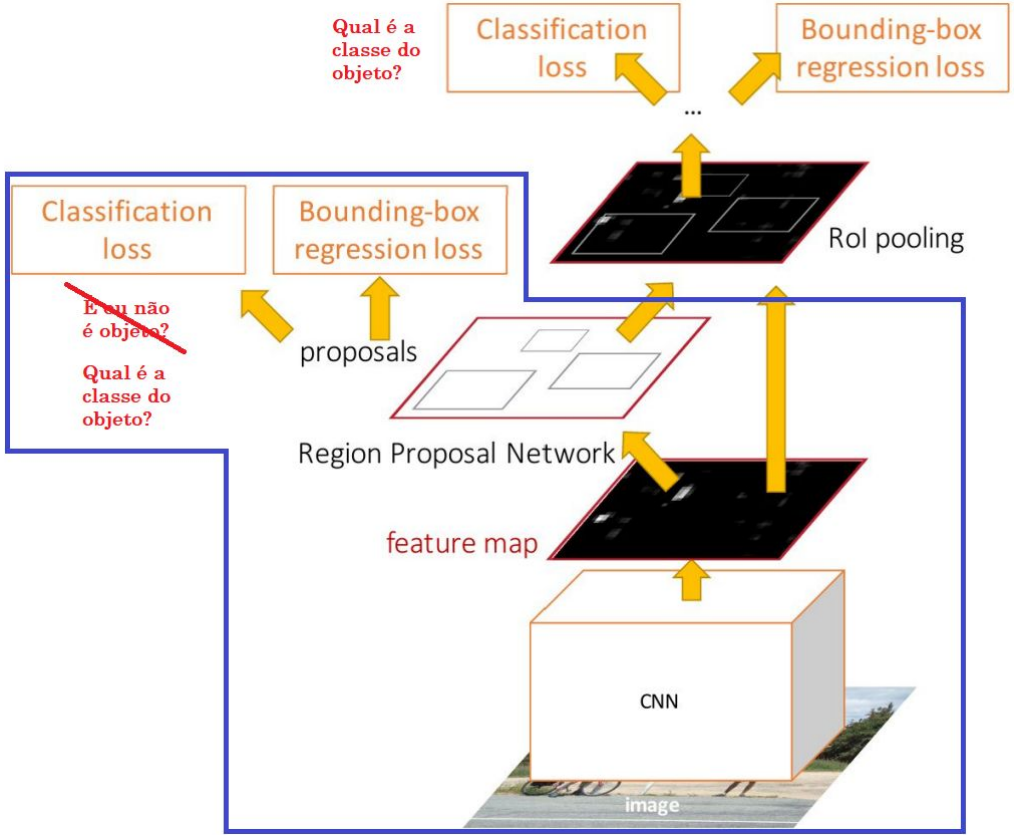
Detecção de múltiplos objetos

- Faster R-CNN



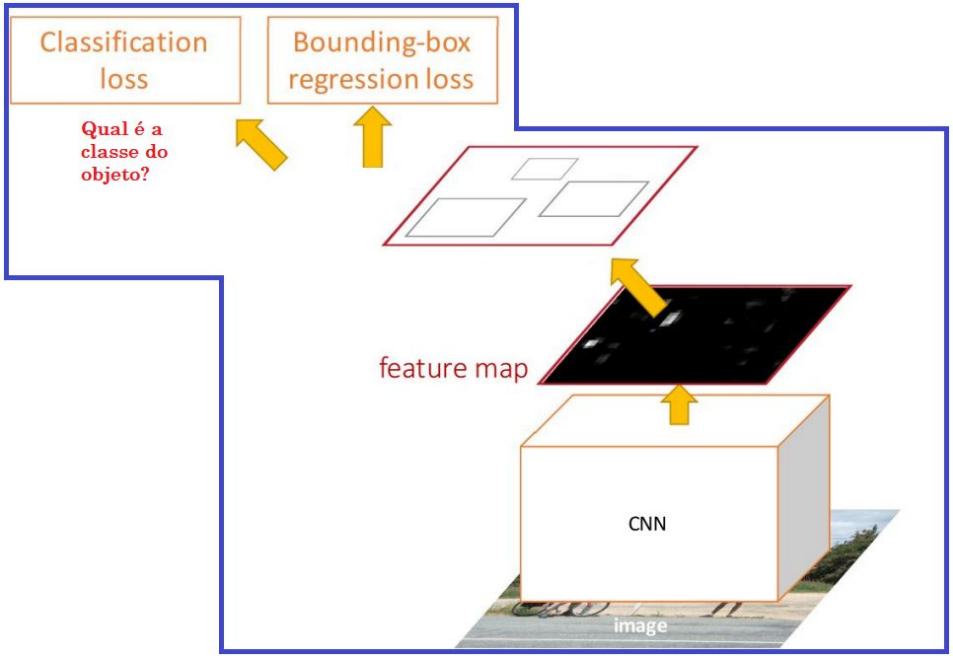
Detecção de múltiplos objetos

- E se...



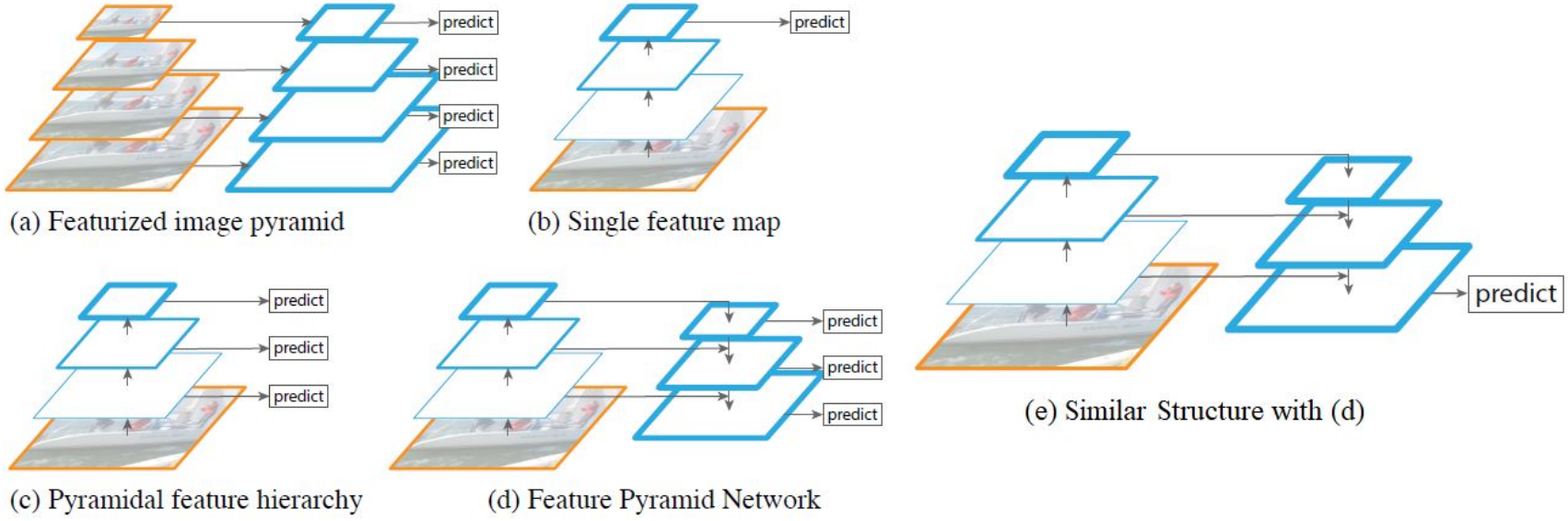
Detecção de múltiplos objetos

- YOLO/SSD/RetinaNet



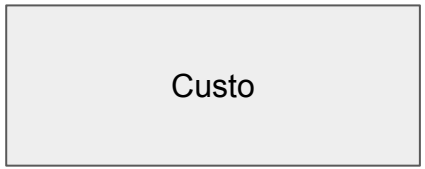
Detecção de múltiplos objetos

- Feature Pyramid Netwroks



Detecção de múltiplos objetos

- Como é a função de custo?

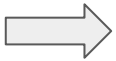


Detecção de múltiplos objetos

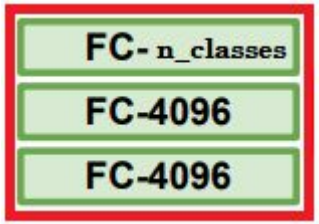
- Como é a função de custo?

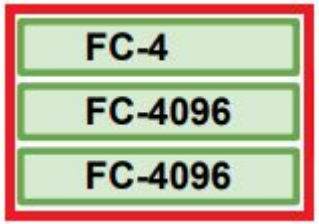
Classificação:
- CE
- BCE
- Hinge
- ...

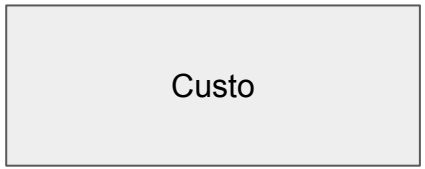
Regressão:
- L2
- L1
- ...



$$\mathcal{L}_{total} = \mathcal{L}_{class} + \lambda \mathcal{L}_{bbox}$$







Para cada âncora!

Detecção de múltiplos objetos

- Focal Loss

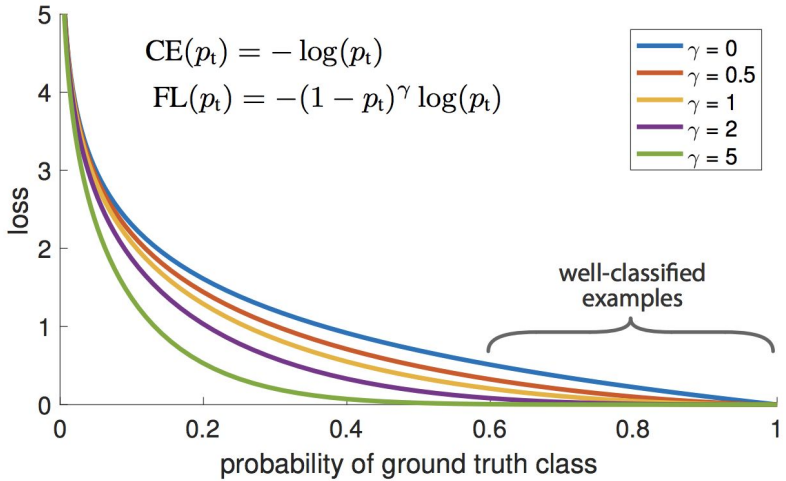


Figure 1. We propose a novel loss we term the *Focal Loss* that adds a factor $(1 - p_t)^{\gamma}$ to the standard cross entropy criterion. Setting $\gamma > 0$ reduces the relative loss for well-classified examples ($p_t > .5$), putting more focus on hard, misclassified examples. As our experiments will demonstrate, the proposed focal loss enables training highly accurate dense object detectors in the presence of vast numbers of easy background examples.

Detecção de múltiplos objetos

- Object Detection in 20 years: A Survey (artigo de revisão bibliográfica deste ano!)

arXiv:1905.05055v2 [cs.CV] 16 May 2019

Object Detection in 20 Years: A Survey

Zhengxia Zou, Zhenwei Shi, *Member, IEEE*, Yuhong Guo, and Jieping Ye, *Senior Member, IEEE*

Abstract—Object detection, as of one the most fundamental and challenging problems in computer vision, has received great attention in recent years. Its development in the past two decades can be regarded as an epitome of computer vision history. If we think of today's object detection as a technical aesthetics under the power of deep learning, then turning back the clock 20 years we would witness the wisdom of cold weapon era. This paper extensively reviews 400+ papers of object detection in the light of its technical evolution, spanning over a quarter-century's time (from the 1990s to 2019). A number of topics have been covered in this paper, including the milestone detectors in history, detection datasets, metrics, fundamental building blocks of the detection system, speed up techniques, and the recent state of the art detection methods. This paper also reviews some important detection applications, such as pedestrian detection, face detection, text detection, etc, and makes an in-deep analysis of their challenges as well as technical improvements in recent years.

Index Terms—Object detection, Computer vision, Deep learning, Convolutional neural networks, Technical evolution.

1 Introduction

Object detection is an important computer vision task that deals with detecting instances of visual objects of a certain class (such as humans, animals, or cars) in digital images. The objective of object detection is to develop computational models and techniques that provide one of the most basic pieces of information needed by computer vision applications: *What objects are where?*

As one of the fundamental problems of computer vision, object detection forms the basis of many other computer vision tasks, such as instance segmentation [1–4], image captioning [5–7], object tracking [8], etc. From the application point of view, object detection can be grouped into two research topics "general object detection" and "detection applications", where the former one aims to explore the methods of detecting different types of objects under a unified framework to simulate the human vision and cognition, and the later one refers to the detection under specific application scenarios, such as pedestrian detection, face detection, text detection, etc. In recent years, the rapid development of deep learning techniques [9] has brought new blood into object detection, leading to remarkable breakthroughs and pushing it forward to a research hot-spot with unprecedented attention. Object detection has now been widely used in many real-world applications, such

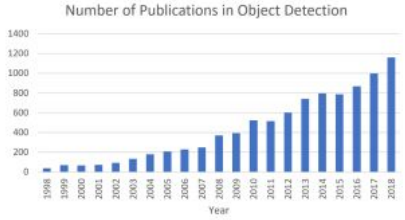


Fig. 1. The increasing number of publications in object detection from 1998 to 2018. (Data from Google scholar advanced search: *allintitle: "object detection" AND "detecting objects"*.)

decades.

- **Difference from other related reviews**

A number of reviews of general object detection have been published in recent years [24–28]. The main difference between this paper and the above reviews are summarized as follows:

No próximo episódio...

- Segmentação semântica