Pedestrian Detection using R-CNN

CS676A: Computer Vision
Project Report

Advisor: Prof. Vinay P. Namboodiri

Deepak Kumar (12228) Mohit Singh Solanki (12419)
Group-17

April 15, 2016

Abstract

Pedestrian detection has been an important problem for decades, given its relevance to a number of applications in robotics, including driver assistance systems, road scene understanding or surveillance systems. Pedestrian detection is an essential and significant task in any intelligent video surveillance system, as it provides the fundamental information for semantic understanding of the video footages. It has an obvious extension to automotive applications due to the potential for improving safety systems.

1 Introduction

People are among the most important components of a machines environment, and endowing machines with the ability to interact with people is one of the most interesting and potentially useful challenges for modern engineering. Detecting and tracking people is thus an important area of research, and machine vision is bound to play a key role. Applications include robotics, entertainment, surveillance, care for the elderly and disabled, and content-based indexing. Just in the US, nearly 5,000 of the 35,000 annual traffic crash fatalities involve pedestrians, hence the considerable interest in building automated vision systems for detecting pedestrians. As such, it has served as a playground to explore different ideas for object detection.

The main paradigms for object detection Viola & Jones variants, HOG+SVM rigid templates, deformable part detectors (DPM), were all explored till 2012. SIFT and HOG are blockwise orientation histograms, a representation we could associate roughly with complex cells in V1, the first cortical area in the primate visual pathway. But we also know that recognition occurs several stages downstream, which suggests that there might be hierarchical, multi-stage processes for computing features that are even more informative for visual recognition. Fukushimas neocognitron, a biologicallyinspired 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. Later work showed that stochastic gradient descent via backpropagation was effective

for training convolutional neural networks (CNNs), a class of models that extend the neocognitron. CNNs saw heavy use in the 1990s, but then fell out of fashion with the rise of support vector machines. In 2012, Krizhevsky et al[1]. rekindled interest in CNNs by showing substantially higher image classification accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Their success resulted from training a large CNN on 1.2 million labeled images, together with a few twists on LeCuns CNN (e.g., max(x, 0) rectifying non-linearities and dropout regularization). Then came R-CNN by Ross Girshick et al.[3] who showed improved performance of object detection on PASCAL VOC dataset. Further modifications to R-CNN lead to the creation of Fast-RCNN[2] and Faster-RCNN[5] which further increased the mAP for object detection on PASCAL VOC dataset.

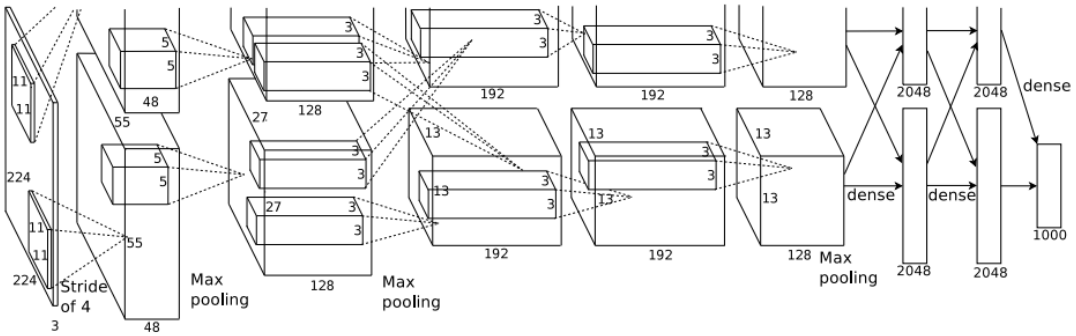


Figure 1: AlexNet Architecture

In this project we aim at modifying the existing R-CNN architecture to suit our pedestrian detection task. The original R-CNN was trained for 21 class detection of PASCAL VOC dataset.

2 Dataset

Multiple public pedestrian datasets have been collected over the years; INRIA, ETH, TUD-Brussels, Daimler (Daimler stereo), Caltech-USA, and KITTI are the most commonly used ones. They all have different characteristics, weaknesses, and strengths.

INRIA is amongst the oldest and as such has comparatively few images. It benefits however from high quality annotations of pedestrians in diverse settings (city, beach, mountains, etc.), which is why it is commonly selected for training. ETH and TUD-Brussels are mid-sized video datasets. Daimler is not considered by all methods because it lacks colour channels. Daimler stereo, ETH, and KITTI provide stereo information. All datasets but INRIA are obtained from video, which makes INRIA dataset unique and it doesn't depend on the optical flow within the images.

INRIA dataset has both positive and negative datasets. But we use only the positive dataset for training. This is because, a lot of negative samples are generated in the process of region proposal.

3 Detection using R-CNN

The R-CNN object detection system consists of three modules.

1. Category Independent Region Proposal which define the set of candidate detections available to our detector.
2. Large Convolutional neural network that extracts a fixed-length feature vector from each region.
3. Set of class-specefic (Pedestrian/Background) linear SVMs.

3.1 Region Proposal

A variety of methods are available for generating category-independent region proposals like Objectness, Selective Search, Category-Independent object proposal, constrained parametric min-cuts (CPMC), multi-scale combinatorial grouping etc. Here we use selective search[4] for our purpose which gives region proposals at different scales.

Selective search works by over-segmenting the entire image. The underlying assumption is that pixels of the same object are highly correlated. Now the scale at which segmenting is done is increased and the segments are combined based on similarity measure. This gives us various segments at various scales. We treat each of these generated segments as a separate object and generate bounding box corresponding to these segments. This method generates ~2000 region proposals per image.

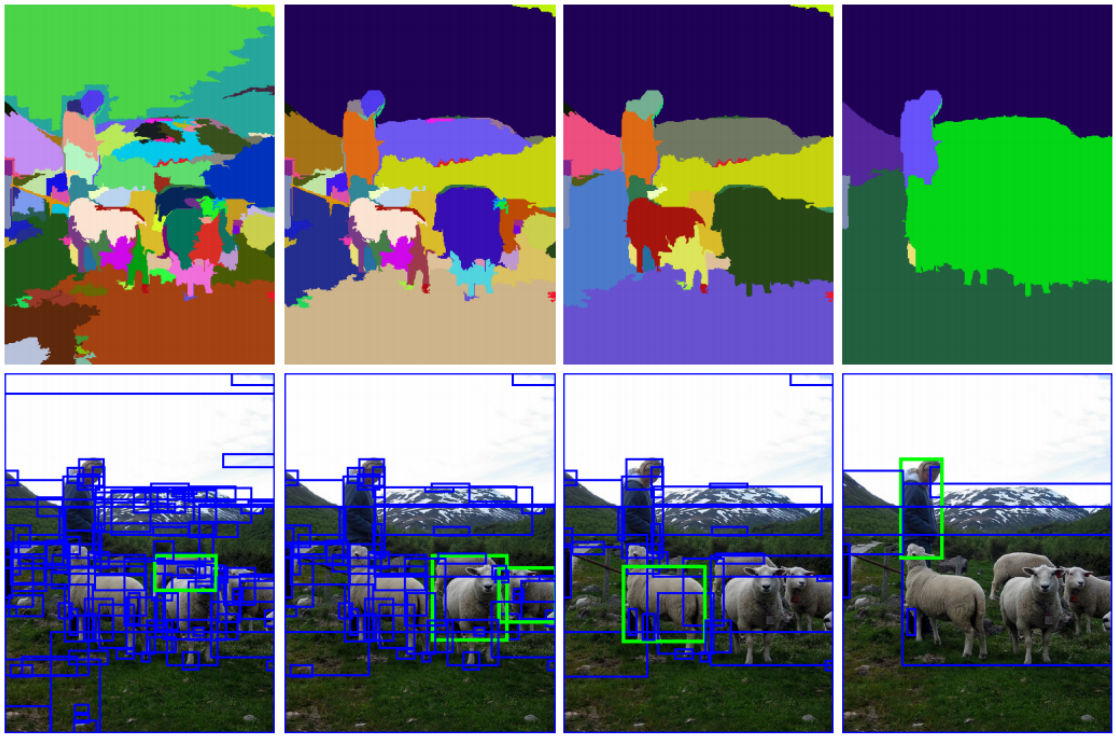


Figure 2: Selective Search

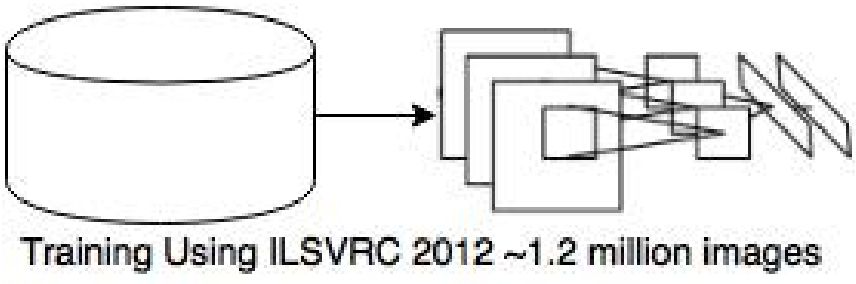
3.2 Feature Extraction

A 4096-dimensional feature vector is extracted from each region proposal using the Caffe implementation of the CNN described by Krizhevsky et al[1]. Features are computed by forward propagating a mean-subtracted 227 x 227 RGB image through five convolutional layers and two fully connected layers. In order to compute features for a region proposal, the image data in that region is first converted into a form that is compatible with the CNN (its architecture requires inputs of a fixed 227 x 227 pixel size). Of the many possible transformations of our arbitrary-shaped regions, the simplest is chosen. Regardless of the size or aspect ratio of the candidate region, all pixels are warped in a tight bounding box around it to the required size. Prior to warping, the tight bounding box is dialated so that at the warped size there are exactly p pixels of warped image context around the original box (p = 16).

Figure 3: Image warping and dialation

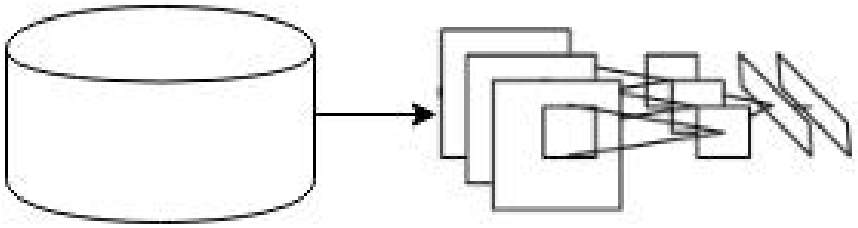
4 Training

We use a pre-trained caffe model trained on ILSVRC 2012 dataset. This is to avoid random initialization of RCNN weights and to avoid over-fitting of data. This model is trained for 1000 classes of ILSVRC dataset.



4.1 Fine-Tuning

To adapt our CNN to the new task (detection) and the new domain (warped INRIA windows), stochastic gradient descent (SGD) training of the CNN parameters is continued using only warped region proposals from INRIA dataset. Aside from replacing the CNNs ImageNet-specific 1000-way classification layer with a randomly initialized 2-way classification layer (for the pedestrian plus background), the CNN architecture is unchanged. All region proposals with ≥ 0.5 IoU overlap with a ground-truth box are treated as positives for that boxs class and the rest as negatives. SGD is started at a learning rate of 0.001 ($1/10^{th}$ of the initial pre-training rate), which allows fine-tuning to make progress while not clobbering the initialization. In each SGD iteration, 32 positive windows (over pedestrian) and 96 background windows are uniformly sampled to construct a mini-batch of size 128. The sampling is biased towards positive windows because they are extremely rare compared to background.



Fine Tune on INRIA dataset

4.2 Pedestrian Classifier

For our purpose its clear that an image region tightly enclosing a person should be a positive example. Similarly, its clear that a background region, which has nothing to do with person, should be a negative example. Less clear is how to label a region that partially overlaps a person. This issue is resolved with an IoU overlap threshold, below which regions are defined as negatives. 0.3 is chosen as the threshold as mentioned in the actual paper[3]. Authors claim that selecting this threshold carefully is important. Setting it to 0.5, as in, decreased mAP by 5 points. Similarly, setting it to 0 decreased mAP by 4 points. Positive examples are defined simply to be the ground-truth bounding boxes. Once features are extracted and training labels are applied, a linear SVM is optimised. Since the training data is too large to fit in memory, the standard hard negative mining method is adapted. Hard negative mining converges quickly and in practice mAP stops increasing after only a single pass over all images.

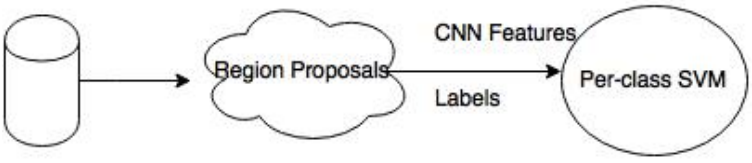


Figure 4: Linear SVM Training

5 Test Time Detection

At test time, selective search is run on the test image to extract around 2000 region proposals (selective searchs fast mode is used in all experiments). Each proposal is warped and forward propagated through the CNN in order to read off features from the desired layer. Then, for pedestrian class, each extracted feature vector is scored using the SVM trained for that. Given all scored regions in an image, a greedy non-maximum suppression is applied that rejects a region if it has an intersection-over-union (IoU) overlap with a higher scoring selected region larger than a learned threshold.

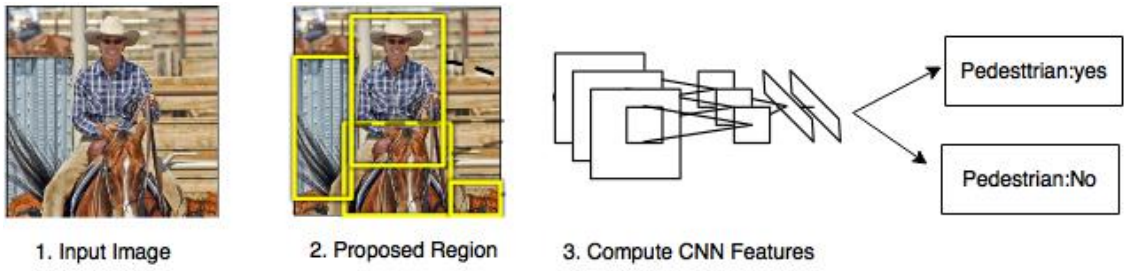


Figure 5: Detection at test time

6 Results

We changed the original CNN architecture to suit to our needs by changing the number of classes. The original paper had 21 classes. In our case we just had two classes, background and pedestrian. We also modified the code so that the algorithm works with the given annotations of INRIA dataset.

6.1 Performance

Method Used	Performance
Discriminately Trained DPM (based on HOG)	88%
Our R-CNN based approach	88.2%

- Trained number of region of interest = around 1.2 million
- Number of training Images = 614
- Total time taken for training = 4.7 hours
- Total no. of test images = 288

Note: Even though the number of training images is less. The number of image proposals from each image is ~2000. This makes the training set large and sufficient to train the CNN.

6.2 Hardware Specs

1. OS - Ubuntu 14.04 (On Amazon Web Service)
2. Memory - 15GB
3. GPU - 1x NVIDIA GRID (Kepler G104) + 8 x hardware hyperthreads from Intel Xeon E5-2670

6.3 Detection Results

Figure 6: Sample results on test images

7 Future Work

1. As Caltech dataset is huge and has continuous frames, can test the accuracies obtained on this dataset. Also this happens to be the most widely used dataset for pedestrian detection currently and acts like a benchmark for comparison.
2. Test the technique on other category-independent region proposal methods like BING, MCG, CPMC etc.
3. Test on classifiers other than Linear SVM like clustering, Nearest Neighbours etc.

Acknowledgements

We would like to thank Prof. Vinay P. Namboodiri for his guidance throughout the project. Special thanks to the course TAs Adarsh Chauhan, Yeshi Dolma and Samrath Patidar for clarifications on the aspects of CNN about which we had very little idea. We would like to express the deepest appreciation to Siddhant Manocha, Satvik Gupta and Kundan Kumar, our batch-mates for helping us understand the R-CNN pipeline better. All the help provided by every other student enrolled in the course is appreciated.

References

[1] Geoffrey E. Hinton Alex Krizhevsky, Ilya Sutskever. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

[2] Ross Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.

[4] Gevers2 J.R.R. Uijlings, van de Sande and A.W.M. Smeulders2. Selective search for object recognition. *ICCV*, 2011.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.