CNN Object Detection At HD And UHD At The Edge

- David Packwood
- 17/09/2018

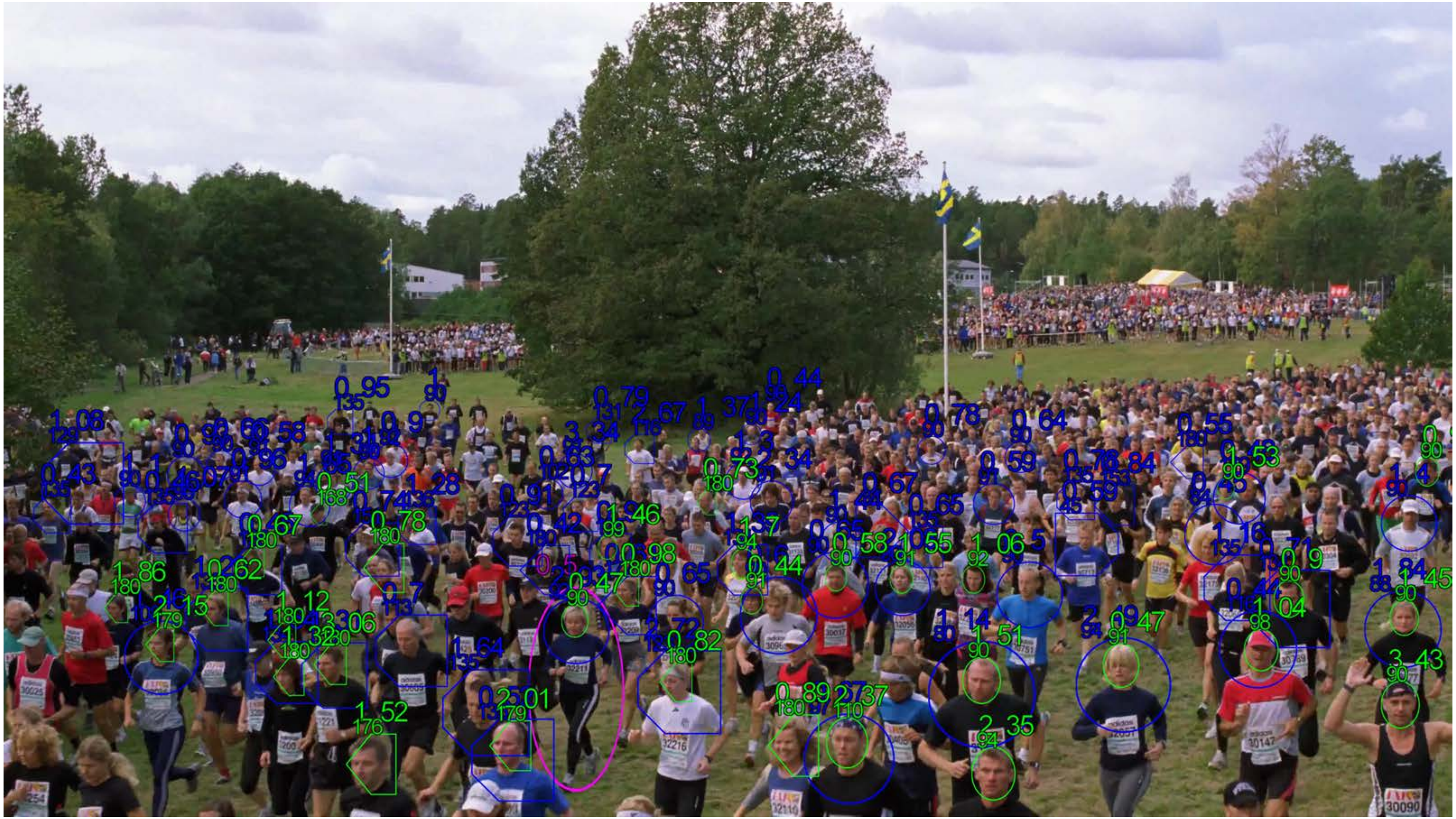
Motivation

Resolution = View Distance

Resolution = View Distance

- Cameras at the edge have HD+ resolution sensors today.
- CNN Object Detection systems commonly published have an input window of size in the range 300-600 pixels.
- Any Object Detection system will have a minimum size object it can possibly detect.
- The exact size of objects in pixel terms is dependent on the given optical system in question (lens, sensor etc.)
- But roughly speaking downscaling an image by half you can expect your view distance to halve.
- For this reason it is desirable to do object detection at the native resolution of the optical system.

Example



CNN Object Detection

What Is Tractable At The Edge?

What Hardware Is Available?

In the interest of impartiality…

NVIDIA Jetson TX2

- 1.5 TOPs Single Precision
- £400 for the module

Google Edge TPU

- (rumoured) 4 TOPs INT8

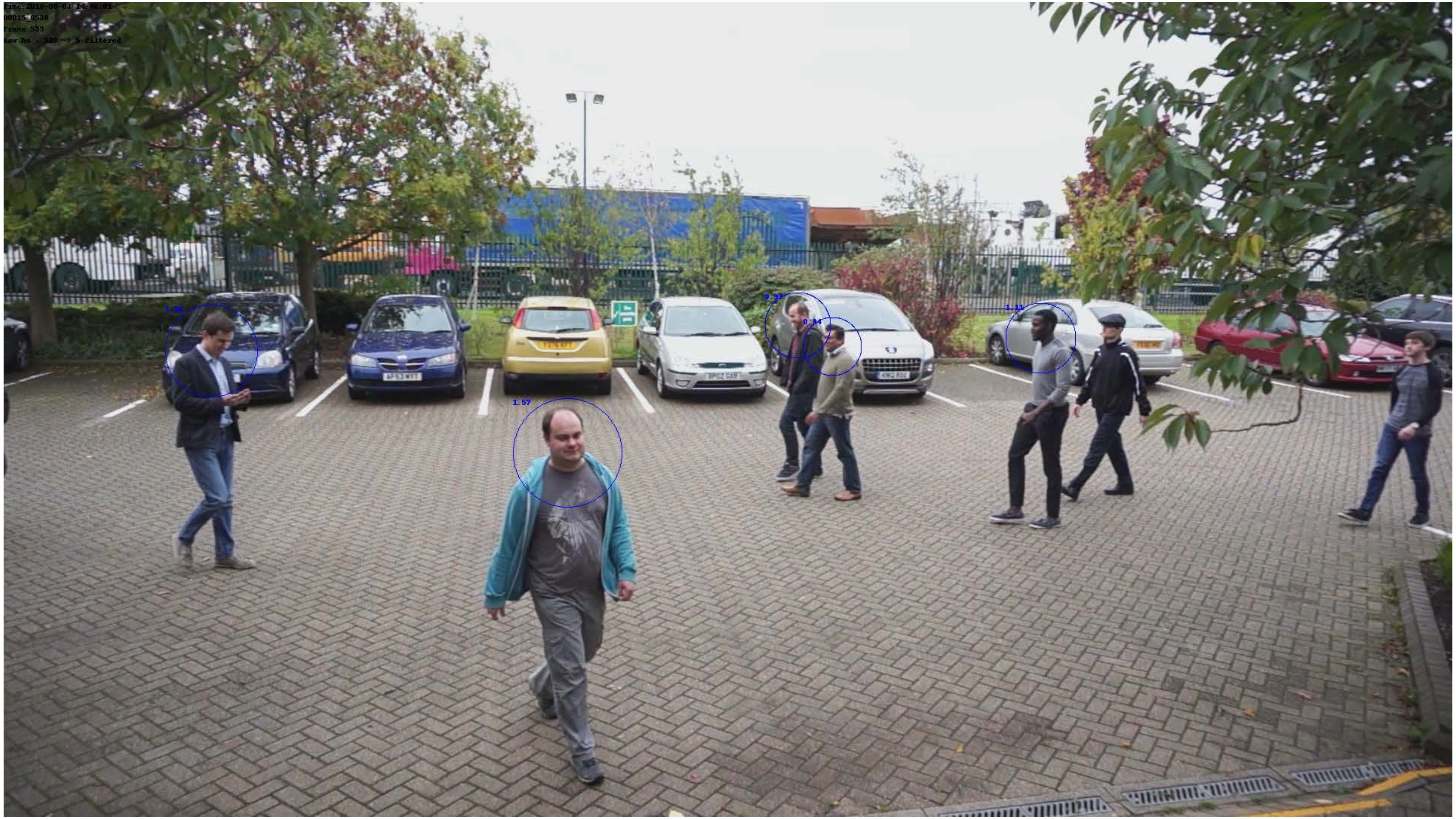
Some Thoughts

- The number of TOPs available at the edge is O(1)
- These TOPs are not cheap (could be desirable that they are shared between multiple sensors).

Scaling Up Object Detection to Full HD

- Consider a start of the art CNN object detection system (YOLOv3)
- Some simple arithmetic shows that naively scaling YOLOv3 to receive as input a full HD frame requires roughly 0.4 TOPs per frame.
- If we are willing to downgrade on detection performance, a tiny version of Yolo (TinyYOLO) is available, at full HD frame 0.08 TOPs, so at 30 fps 2.4 TOPs, this seems just about tractable.
- But is this a smart thing to do?

Deep Networks On The Whole Frame?



Region Proposals

Region Proposals Are Already A Thing

- The Region Proposal idea is to pass the whole image through a CNN and generate a feature map for the images.
- RCNN Object Detection systems (especially Faster RCNN) etc. create region proposals from this feature map and then a more generic classifier decides what is in the region of interest.
- A commonly used CNN to generate the feature map is VGG-16 (0.4 TOPs for a full HD image)
- Then the classifier should be applied.

What Classes Do We Really Need?

Are current benchmarks enough?

Giraffes (COCO)?

Non copyright image of a giraffe.
(Please use your imagination)

What are current benchmarks testing?

- Detection performance on a wide variety of classes.
- A lot of variation of pose and setting within a class.
- Multiple objects of interest per “small” image.
- No temporal information.

Super Lightweight Region Proposals

A very simple idea

Type	Output Channels	Size	Output
RGB	3	-	1920x1080
Conv	16	3x3	1918x1078
Conv	16	3x3	1916x1076
Pool	16	2x2	958x538
Conv	32	3x3	954x534
Conv	32	3x3	950x530
Pool	32	3x3	475x265
Conv	64	3x3	467x257
Conv	16	16x16	

- 0.04 TOPs (1.2 TOPs at 30fps)
- 300K Free Parameters (cacheable)
- After a region is suggested we can put it through an “expert” second stage system to verify the output.
- We can also exploit temporal information to track objects between frames.

Some Ideas On Implementation

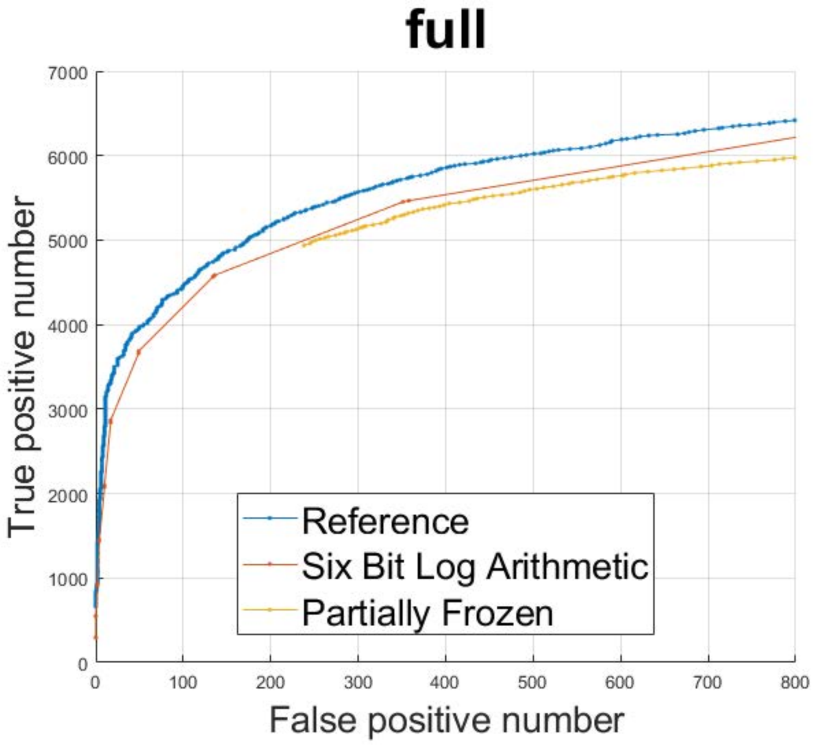
We May Also Save OPs By Implementation

When is an OP not an OP?

A couple of ideas:

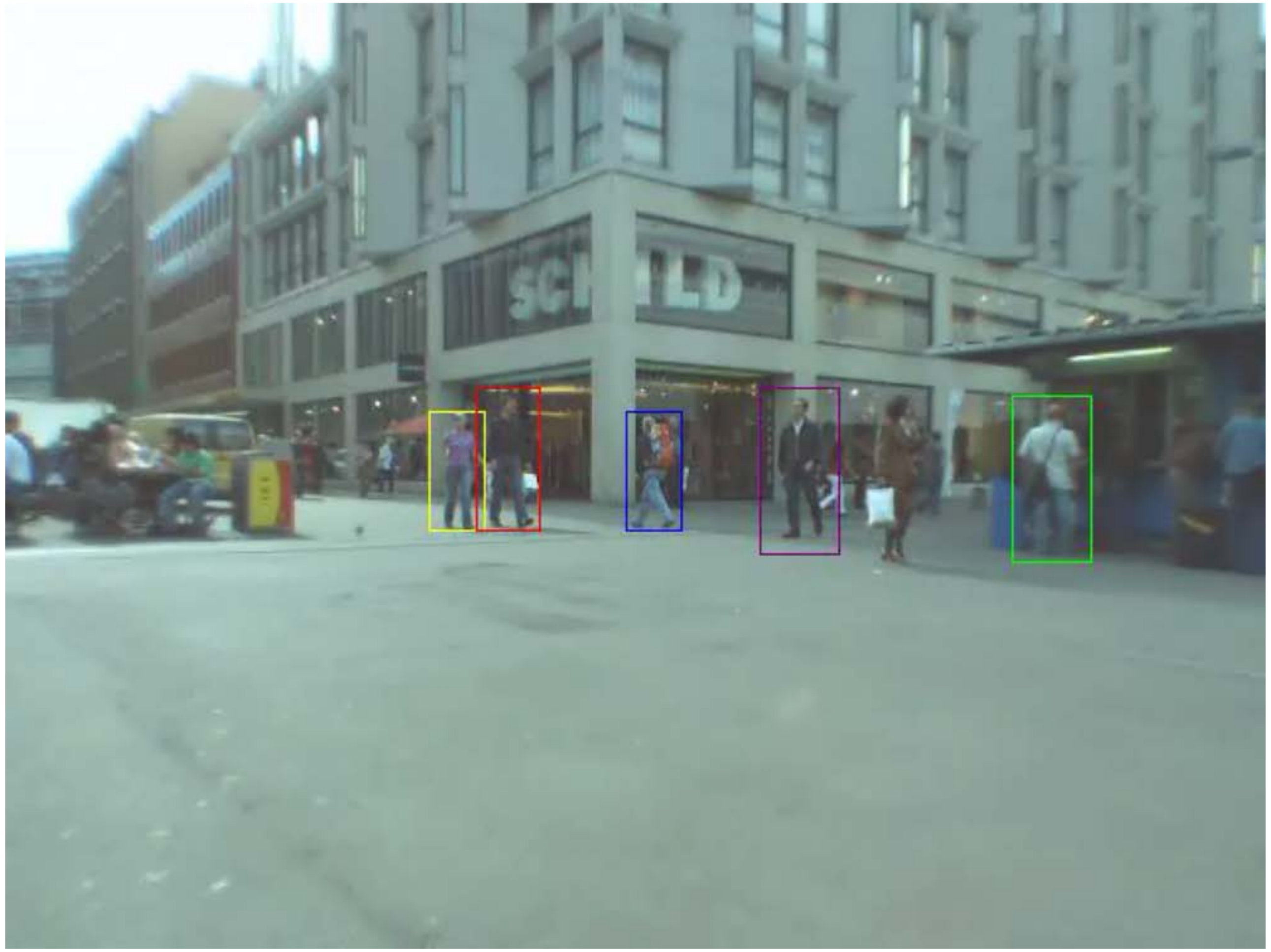
- Implement the CNN by a hardwired circuit (rather than programmable weights)
- Implement the CNN with arithmetic in the logarithmic domain.

Results



Some Videos

Pedestrian Detections



Video from:
A Mobile Vision System for Robust Multi-Person Tracking, A. Ess & B. Leibe & K. Schindler & L. van Gool, IEEE Conference on Computer Vision and Pattern Recognition (CVPR 08)

Features

Closing Thoughts

Closing Thoughts

- Native resolution object detection is desirable.
- Existing benchmarks emphasize detection performance at all costs.
- Performance per compute cost is a desirable metric as a starting point.
- A more practical benchmark would be reduced class set at higher resolution.
- Deep CNNs on sparse/high res images are very expensive.
- Temporal information is part of the answer.
- Resolution is only increasing. Can compute keep up?

Thank You

Danke

Merci

谢谢

ありがとう

Gracias

Kiitos

감사합니다

धन्यवाद

תודה

arm