Object detection on large-scale egocentric video dataset

Alexandra Maha Abbas
MSc in Big Data

UNIVERSITY of STIRLING

1 Introduction & Problem

This dissertation project tackles an object detection challenge on the **egocentric EPIC-KITCENS Dataset**. Egocentric vision is a subfield of **Computer vision** that focuses on images and video recorded by **wearable cameras** that gives access to a **unique aspect of people's interaction with objects** and other people, their attention and intention.

Google GLASS Microsoft HoloLens magic leap

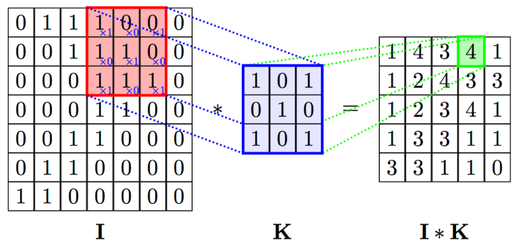
2 EPIC-KITCHENS Dataset

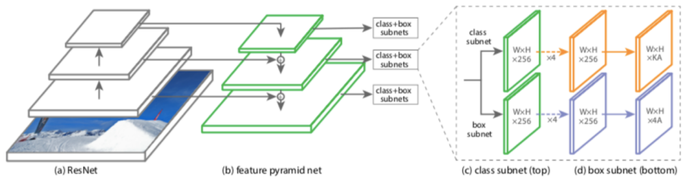
The EPIC-KITCHENS Dataset is a **large-scale egocentric video dataset, capturing various non-scripted activities in native kitchen environments**, which has been recently published (April 2018) by D. Damen *et al*. [1]. It features **55 hours of video** (11.5 million frames) that is labelled for **39.6 thousand action segments and 424.2 thousand object bounding boxes**. Primary aim of the research was to capture natural multi-tasking and parallel-goal interactions.

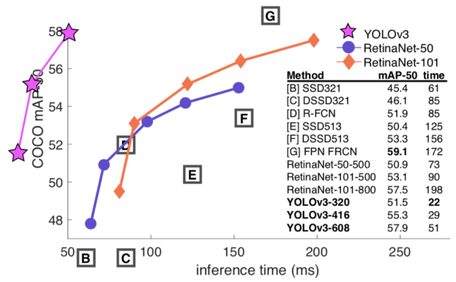
3 Object detection network of choice: RetinaNet

Convolutional Neural Networks (CNNs) are powerful state-of-the-art **image processing methods** widely used in Computer vision for **image recognition, object detection and segmentation**. It was **AlexNet** [2] by A. Krizhevsky, I. Sutskever, and H. Geoffrey E. that popularized CNNs in Computer vision research.

RetinaNet [3] is a **one stage unified fully convolutional neural network** consisting of a **ResNet backbone** combined with a **Feature Pyramid Network (FPN)**, a classification subnet and a box regression subnet. It is a **highly accurate** object detector with **comparable inference** time to other one stage networks.

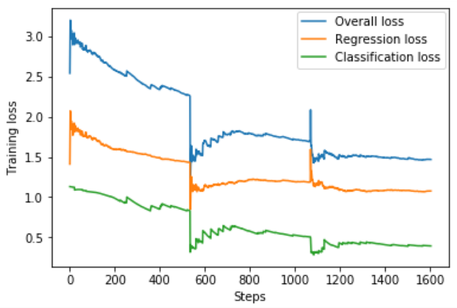






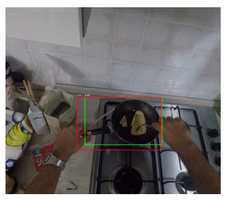
4 Implementation

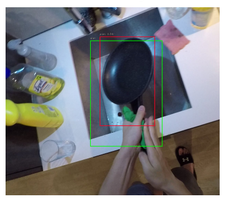
Due to time and computing power constraints only a part of the data was used. I reduced the number of classes from 352 to one and decided to make a **'pan' detector**. RetinaNet was implemented using the excellent **Keras RetinaNet framework**. Initial weights were loaded from a **RetinaNet-50 pretrained on MS COCO**. While training, the **backbone of the network was frozen**. It was trained through **3 epochs** with **batch size 4** and **535 steps per epoch**.



5 Results

RetinaNet was **able to outperform the baseline results** in object class 'pan'.





	IoU threshold	Faster R-CNN (mAP)	RetinaNet (mAP)
Seen kitchens	0.5	0.6760	**0.7412**
Unseen kitchens		0.6288	**0.7273**

6 References

[1] D. Damen et al., "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset," no. April, pp. 1–12, 2018.
[2] A. Krizhevsky, I. Sutskever, and H. Geoffrey E., "ImageNet Classification with Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst. 2012.
[3] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," Proc. IEEE Int. Conf. Comput. Vis., vol. 2017–October, pp. 2999–3007, 2017.

For further information contact: *Alexandra Maha Abbas* *ama00151@students.stir.ac.uk*

BE THE DIFFERENCE