Choosing an evaluation metric

We wish to measure the performance of the inference engine in predicting if a person is wearing a mask and safety helmet. Evaluating the engine on the above task requires **evaluation of two operations:**

Classification of bounding box

In the dataset the distribution of classes is non-uniform . So a simple accuracy-based metric will introduce biases. It is also important to assess the risk of misclassifications. So, it is necessary to **vary the model score threshold** that determines what is counted as a model-predicted positive detection of the class while evaluating classification. The metric should also not be biased towards any particular class.

Regression of bounding box parameters

Model object detections are determined to be true or false depending upon the IoU threshold which summarises how well overlap with ground truth boxes.It is necessary to use **different IOU(intersection over union) thresholds** to find the number of positive predictions in order to evaluate the performance of the regressor in localising the object.

Proposed metric:

mAP(mean average precision) averaged over 24 classes and 10 IOU thresholds.

Calculating the mAP score

No of object classes:

- The object detector classifies the bounding boxes into **three classes.(head, nonhead, background).**
- The mask classification has a total of **4 classes.(yes, no, invisible, wrong).**
- The helmet classification has **2 classes. (yes, no).**

So each detected bounding box is mapped to one of 24(3x4x2) classes.

No of IOU Threshold values: 10 values from 0.5 to 0.95

CALCULATION : First the average precision for a specific class and IOU threshold value is computed by generating the **precision-recall curve** from the model's detection output, by varying the model score threshold. To do this unambiguously, **the AP score is defined as the mean precision of 11 equally spaced recall values.** This AP score is then averaged over all 24 classes and 10 IOU values.