



# Model Evaluation

## Metrics for object detection

Reporter: 01-XGH Time: 2018/12/09



# CONTENTS

**01** | Motivation

**02** | Different metrics

**03** | Important definitions

**04** | How to do it



# Motivation

the lack of consensus used by different works and implementations concerning the **evaluation metrics of the object detection problem**. Although on-line competitions use their own metrics to evaluate the task of object detection, just some of them offer reference code snippets to calculate the accuracy of the detected objects.

Researchers who want to evaluate their work using different datasets than those offered by the competitions, need to implement their own version of the metrics. Sometimes a wrong or different implementation can create different and biased results. Ideally, in order to have trustworthy benchmarking among different approaches, it is necessary to have a flexible implementation that can be used by everyone regardless the dataset used.

## Different Metrics

There are three criteria for evaluating the performance of detection algorithms: **detection speed** (Frames Per Second, FPS), **precision**, and **recall**. The most commonly used metric is Average Precision (AP), derived from precision and recall. AP is usually evaluated in a category specific manner, i.e., computed for each object category separately. In generic object detection, detectors are usually tested in terms of detecting a number of object categories. To compare performance over all object categories, the mean AP ( mAP) averaged over all object categories is adopted as the final measure of performance.

## Different competitions, different metrics

**PASCAL VOC Challenge:** current metrics used by the current PASCAL VOC object detection challenge are the **Precision x Recall curve** and **Average Precision**.

**COCO Detection Challenge:** 12 metrics used for characterizing the performance of an object detector on COCO.

**Google Open Images Dataset V4 Competition** also uses mean Average Precision (mAP) over the 500 classes to evaluate the object detection task.

**ImageNet Object Localization Challenge** defines an error for each image considering the class and the overlapping region between ground truth and detected boxes. The total error is computed as the average of all min errors among all test dataset images.



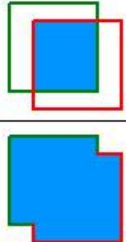
# Important definitions

## Intersection Over Union (IOU)

Intersection Over Union (IOU) is a measure based on Jaccard Index that evaluates the overlap between two bounding boxes. It requires a ground truth bounding box and a predicted bounding box. By applying the IOU we can tell if a detection is valid (True Positive) or not (False Positive).

IOU is given by the overlapping area between the predicted bounding box and the ground truth bounding box divided by the area of union between them:

$$\text{IOU} = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})}$$

$$IOU = \frac{\text{area of overlap}}{\text{area of union}} = \frac{\text{area of overlap}}{\text{area of union}}$$




## Important definitions

**True Positive (TP):** A correct detection. Detection with  $\text{IOU} \geq \text{threshold}$

**False Positive (FP):** A wrong detection. Detection with  $\text{IOU} < \text{threshold}$

**False Negative (FN):** A ground truth not detected

**True Negative (TN):** Does not apply. It would represent a corrected misdetection. In the object detection task there are many possible bounding boxes that should not be detected within an image. Thus, TN would be all possible bounding boxes that were correctly not detected (so many possible boxes within an image). That's why it is not used by the metrics.

*threshold*: depending on the metric, it is usually set to 50%, 75% or 95%.

# Important definitions

## Precision

Precision is the ability of a model to identify **only** the relevant objects. It is the percentage of correct positive predictions and is given by:

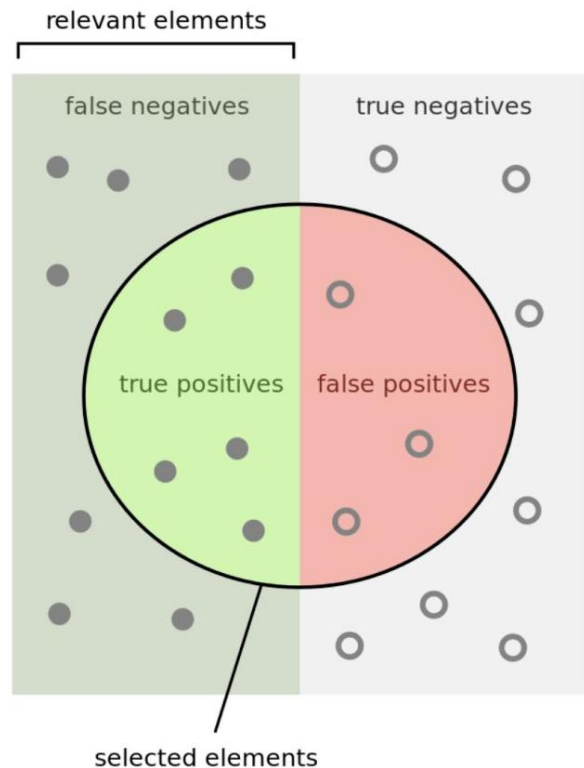
$$Precision = \frac{TP}{TP + FP} = \frac{TP}{\text{all detections}}$$

## Recall

Recall is the ability of a model to find all the relevant cases (all ground truth bounding boxes). It is the percentage of true positive detected among all relevant ground truths and is given by:

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{\text{all ground truths}}$$

# Important definitions



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

## Precision x Recall curve

The Precision x Recall curve is a good way to evaluate the performance of an object detector as the confidence is changed by plotting a curve for each object class. **An object detector of a particular class is considered good if its precision stays high as recall increases**, which means that if you vary the confidence threshold, the precision and recall will still be high. Another way to identify a good object detector is to look for a detector that can identify only relevant objects (0 False Positives = high precision), finding all ground truth objects (0 False Negatives = high recall).

A poor object detector needs to increase the number of detected objects (increasing False Positives = lower precision) in order to retrieve all ground truth objects (high recall). That's why the Precision x Recall curve usually starts with high precision values, decreasing as recall increases. This kind of curve is used by the PASCAL VOC 2012 challenge.

## Average Precision

Another way to compare the performance of object detectors is to calculate the area under the curve (AUC) of the Precision x Recall curve. As AP curves are often zigzag curves going up and down, comparing different curves (different detectors) in the same plot usually is not an easy task - because the curves tend to cross each other much frequently. That's why Average Precision (AP), a numerical metric, can also help us compare different detectors. In practice AP is the precision averaged across all recall values between 0 and 1.

From 2010 on, the method of computing AP by the PASCAL VOC challenge has changed. Currently, **the interpolation performed by PASCAL VOC challenge uses all data points, rather than interpolating only 11 equally spaced points as stated in their paper.**

## 11-point interpolation

The 11-point interpolation tries to summarize the shape of the Precision x Recall curve by averaging the precision at a set of eleven equally spaced recall levels  $[0, 0.1, 0.2, \dots, 1]$ :

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, \dots, 1\}} \rho_{\text{interp}}(r)$$

With

$$\rho_{\text{interp}}(r) = \max_{\tilde{r}: \tilde{r} \geq r} \rho(\tilde{r})$$

where  $\rho(\tilde{r})$  is the measured precision at recall  $\tilde{r}$

Instead of using the precision observed at each point, the AP is obtained by interpolating the precision only at the 11 levels  $r$  taking the **maximum precision whose recall value is greater than  $r$** .

## Interpolating all points

Instead of interpolating only in the 11 equally spaced points, you could interpolate through all points in such way that:

$$\sum_{r=0}^1 (r_{n+1} - r_n) \rho_{interp}(r_{n+1})$$

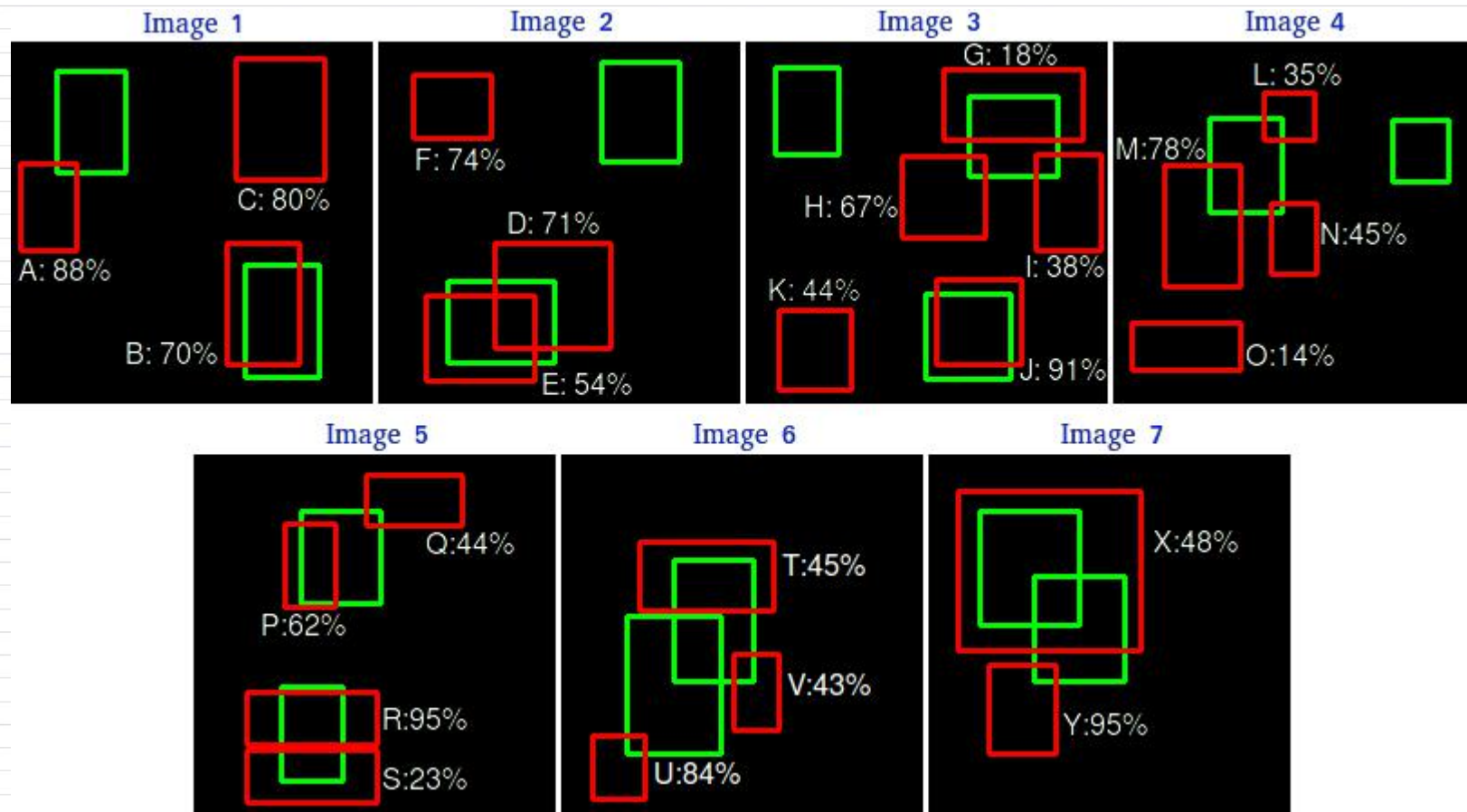
With

$$\rho_{interp}(r_{n+1}) = \max_{\tilde{r}: \tilde{r} \geq r_{n+1}} \rho(\tilde{r})$$

where  $\rho(\tilde{r})$  is the measured precision at recall  $\tilde{r}$

In this case, instead of using the precision observed at only few points, the AP is now obtained by interpolating the precision at **each level**,  $r$  taking the **maximum precision whose recall value is greater or equal than  $r+1$** . This way we calculate the estimated area under the curve.

## An example





# An example

The table shows the bounding boxes with their corresponding confidences. The last column identifies the detections as TP or FP. In this example a TP is considered if IOU  $\geq$  30%, otherwise it is a FP. By looking at the images above we can roughly tell if the detections are TP or FP.

In some images there are more than one detection overlapping a ground truth (Images 2, 3, 4, 5, 6 and 7). For those cases the detection with the highest IOU is taken, discarding the other detections. This rule is applied by the PASCAL VOC 2012 metric: "e.g. 5 detections (TP) of a single object is counted as 1 correct detection and 4 false detections".

Images	Detections	Confidences	TP or FP
Image 1	A	88%	FP
Image 1	B	70%	TP
Image 1	C	80%	FP
Image 2	D	71%	FP
Image 2	E	54%	TP
Image 2	F	74%	FP
Image 3	G	18%	TP
Image 3	H	67%	FP
Image 3	I	38%	FP
Image 3	J	91%	TP
Image 3	K	44%	FP
Image 4	L	35%	FP
Image 4	M	78%	FP
Image 4	N	45%	FP
Image 4	O	14%	FP
Image 5	P	62%	TP
Image 5	Q	44%	FP
Image 5	R	95%	TP
Image 5	S	23%	FP
Image 6	T	45%	FP
Image 6	U	84%	FP
Image 6	V	43%	FP
Image 7	X	48%	TP
Image 7	Y	95%	FP

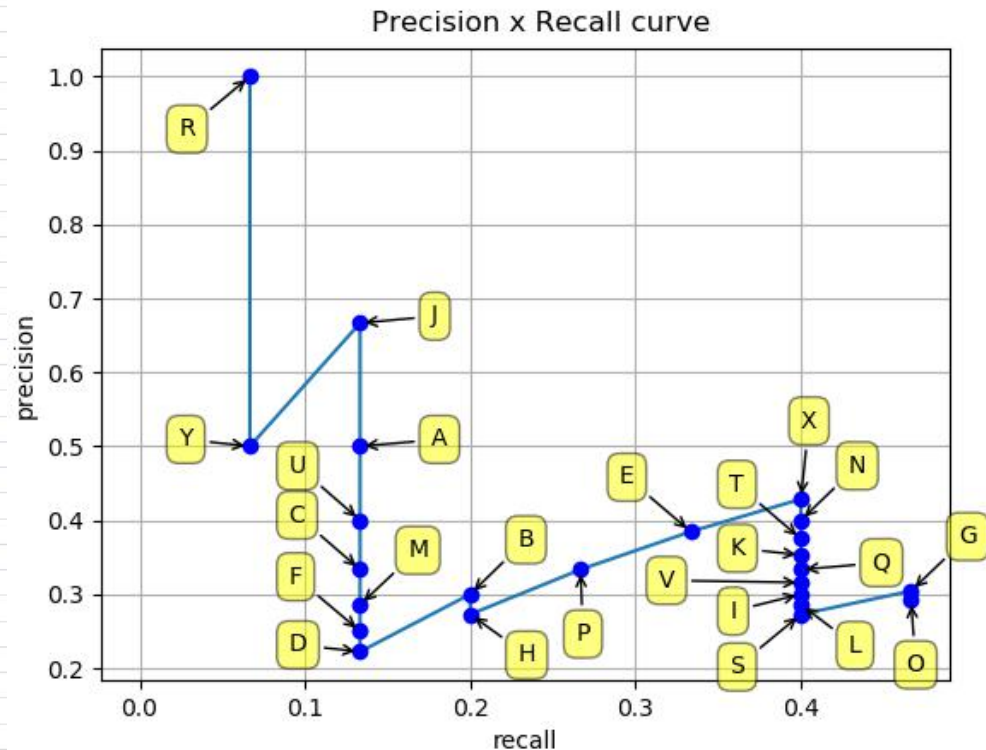
# An example

The Precision x Recall curve is plotted by calculating the precision and recall values of the accumulated TP or FP detections. For this, first we need to order the detections by their confidences, then we calculate the precision and recall for each accumulated detection as shown in the table right:

Images	Detections	Confidences	TP	FP	Acc TP	Acc FP	Precision	Recall
Image 5	R	95%	1	0	1	0	1	0.0666
Image 7	Y	95%	0	1	1	1	0.5	0.0666
Image 3	J	91%	1	0	2	1	0.6666	0.1333
Image 1	A	88%	0	1	2	2	0.5	0.1333
Image 6	U	84%	0	1	2	3	0.4	0.1333
Image 1	C	80%	0	1	2	4	0.3333	0.1333
Image 4	M	78%	0	1	2	5	0.2857	0.1333
Image 2	F	74%	0	1	2	6	0.25	0.1333
Image 2	D	71%	0	1	2	7	0.2222	0.1333
Image 1	B	70%	1	0	3	7	0.3	0.2
Image 3	H	67%	0	1	3	8	0.2727	0.2
Image 5	P	62%	1	0	4	8	0.3333	0.2666
Image 2	E	54%	1	0	5	8	0.3846	0.3333
Image 7	X	48%	1	0	6	8	0.4285	0.4
Image 4	N	45%	0	1	6	9	0.4	0.4
Image 6	T	45%	0	1	6	10	0.375	0.4
Image 3	K	44%	0	1	6	11	0.3529	0.4
Image 5	Q	44%	0	1	6	12	0.3333	0.4
Image 6	V	43%	0	1	6	13	0.3157	0.4
Image 3	I	38%	0	1	6	14	0.3	0.4
Image 4	L	35%	0	1	6	15	0.2857	0.4
Image 5	S	23%	0	1	6	16	0.2727	0.4
Image 3	G	18%	1	0	7	16	0.3043	0.4666
Image 4	O	14%	0	1	7	17	0.2916	0.4666

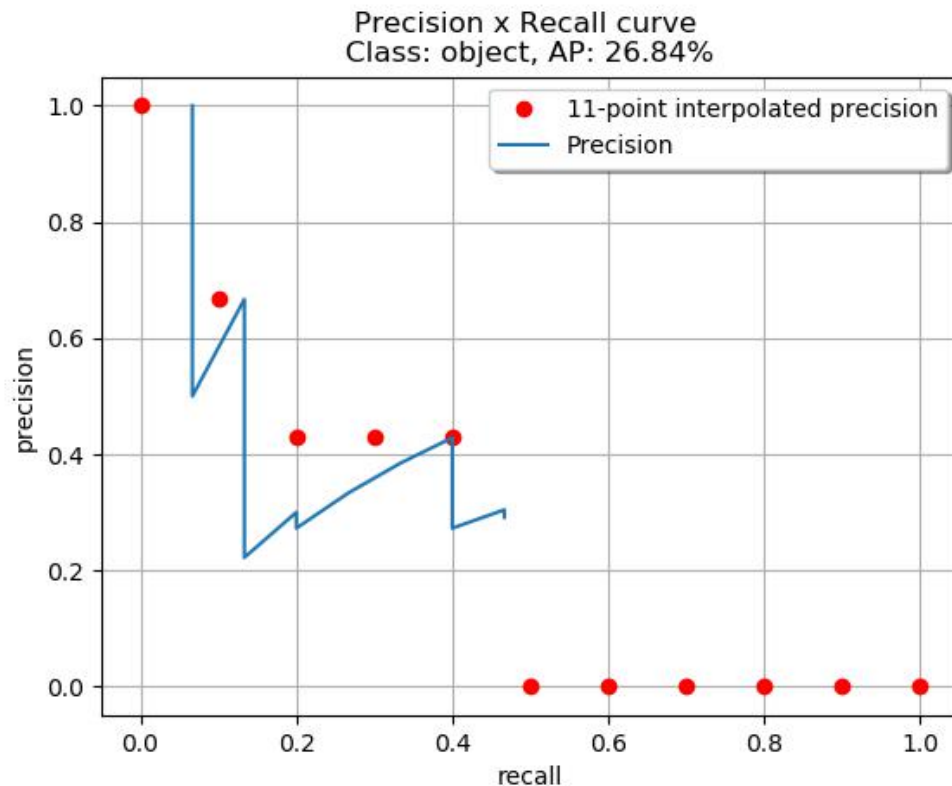
## An example

The Precision x Recall curve is plotted by calculating the precision and recall values of the accumulated TP or FP detections. For this, first we need to order the detections by their confidences, then we calculate the precision and recall for each accumulated detection as shown in the table right:



## Calculating the 11-point interpolation

The idea of the 11-point interpolated average precision is to average the precisions at a set of 11 recall levels (0,0.1,...,1). The interpolated precision values are obtained by taking the maximum precision whose recall value is greater than its current recall value as follows:



## An example

By applying the 11-point interpolation, we have:

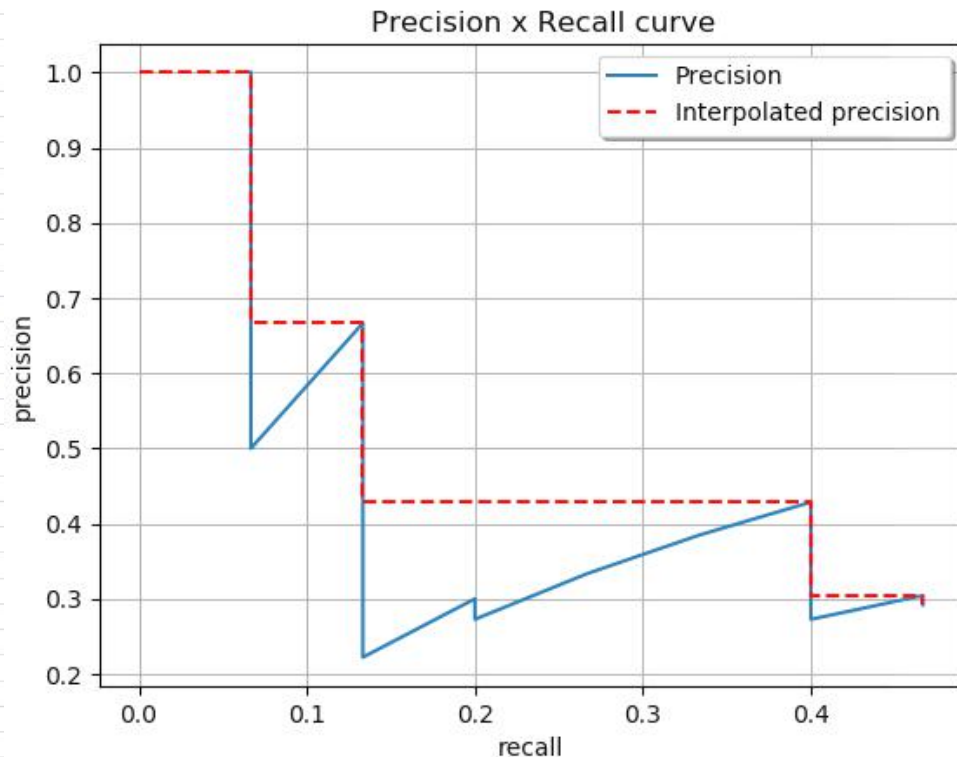
$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, \dots, 1\}} \rho_{\text{interp}}(r)$$

$$AP = \frac{1}{11} (1 + 0.6666 + 0.4285 + 0.4285 + 0.4285 + 0 + 0 + 0 + 0 + 0 + 0)$$

$$AP = 26.84\%$$

## Calculating the interpolation performed in all points

By interpolating all points, the Average Precision (AP) can be interpreted as an approximated AUC of the Precision x Recall curve. The intention is to reduce the impact of the wiggles in the curve. By applying the equations presented before, we can obtain the areas as it will be demonstrated here. We could also visually have the interpolated precision points by looking at the recalls starting from the highest (0.4666) to 0 (looking at the plot from right to left) and, as we decrease the recall, we collect the precision values that are the highest as shown in the image



## Calculating the interpolation performed in all points

Looking at the plot above, we can divide the AUC into 4 areas (A1, A2, A3 and A4).

Calculating the total area, we have the AP:

$$AP = A1 + A2 + A3 + A4$$

$$A1 = (0.0666 - 0) \times 1 = \mathbf{0.0666}$$

$$A2 = (0.1333 - 0.0666) \times 0.6666 = \mathbf{0.04446222}$$

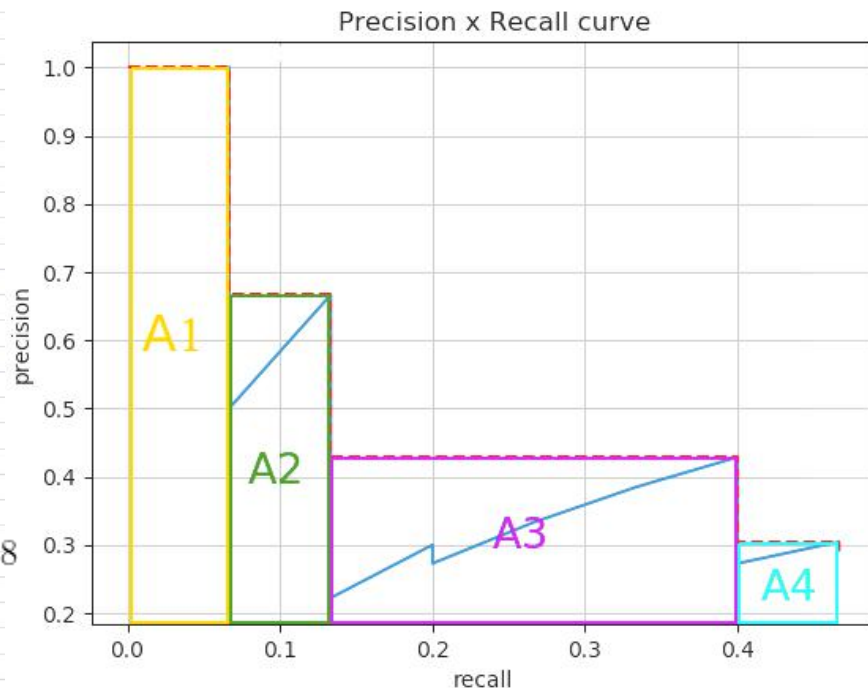
$$A3 = (0.4 - 0.1333) \times 0.4285 = \mathbf{0.11428095}$$

$$A4 = (0.4666 - 0.4) \times 0.3043 = \mathbf{0.02026638}$$

$$AP = 0.0666 + 0.04446222 + 0.11428095 + 0.02026638$$

$$AP = 0.24560955$$

$$AP = \mathbf{24.56\%}$$







**How to do it ?**



# Game Over

THANK YOU!

最后还是来点呱唧呱唧啊！