Evaluation Metrics Guide

Powered by **inmind.ai**

inmind.ai | contact@inmind.ai | +961 3 051 195 | Mkalles, Main Road, bldg. 757, Lebanon

Table of Contents

Evaluation Metrics Definitions

Generally, positive/negative defines whether we have a prediction or not, and True/False define whether our prediction is True/False

	Prediction	Reality	Definition
True Positive (TP)	Positive	Positive	Correctly detected cases
True Negative (TN)	Negative	Negative	Correctly not detected cases
False Positive (FP)	Positive	Negative	Wrongly detected cases
False Negative (FN)	Negative	Positive	Wrongly not detected cases

Accuracy: $\frac{Correct\ answers}{All\ Posibilities} = \frac{TP+TN}{TP+TN+FP+FN}$

- Measure of how correct our Data is.
- Best = 1 ----- Worst = 0

Precision: $\frac{Predicted\ Correctly}{All\ Predicted} = \frac{TP}{TP+FP}$

- Measure the fraction of relevant examples among all of the examples predicted that belong to a certain class.
- Best = 1 ----- Worst = 0

Recall: $\frac{Predicted\ Correctly}{All\ that\ should\ be\ Predicted} = \frac{TP}{TP+FN}$

- Measure the fraction of relevant examples which were predicted to belong to a class with respect to all examples that truly belong to a class.
- Best = 1 ----- Worst = 0

F-Score: $\frac{(1+\beta^2)\times Precision\times Recall}{(\beta^2\times Precision)+Recall}$

- Evaluate the machine learning model in one number by combining precision and recall
- A good f-score means that the model is effective in terms of how precisely it classifies the data and that it covers a good proportion of the cases that it should have classified correctly.
- Best = 1 ----- Worst = 0
    - If $\beta = 1$ neutral
    - If $\beta < 1$ focuses more on precision
    - If $\beta > 1$ focuses more on recall

- Before Continuing, we need to give an example:
    - Let us take a case where we have 95% of the population does not have cancer and 5% of the population has cancer.
    - If over 100 people we detect 100 persons that do not have cancer, then we will have a case where accuracy = 95%.
    - But in reality, we need this model to predict if a person has cancer or not so these results are <u>unacceptable</u> because we cannot find <u>any</u> persons that have cancer.
    - Another example is the case where 95% of data should be detected but we detect 100% of the data as True. In this case, we will have precision = 95% and Recall = 100% but the purpose of finding the 5% makes the model be very bad.
    - This is why a new metric will be introduced (balanced accuracy), in which we are taking into consideration how the dataset is distributed or balanced

---

True Negative Rate: $\frac{TN}{TN+FP}$

True Positive Rate: $\frac{TP}{TP+FN}$

Balanced Accuracy: $\frac{TNR+TPR}{2}$

- This parameter takes into consideration how balanced is the data.

Total Number of Ground Truths: Sum GTs = TP + FN

Total Number of Predictions: Sum Preds = TP + FP

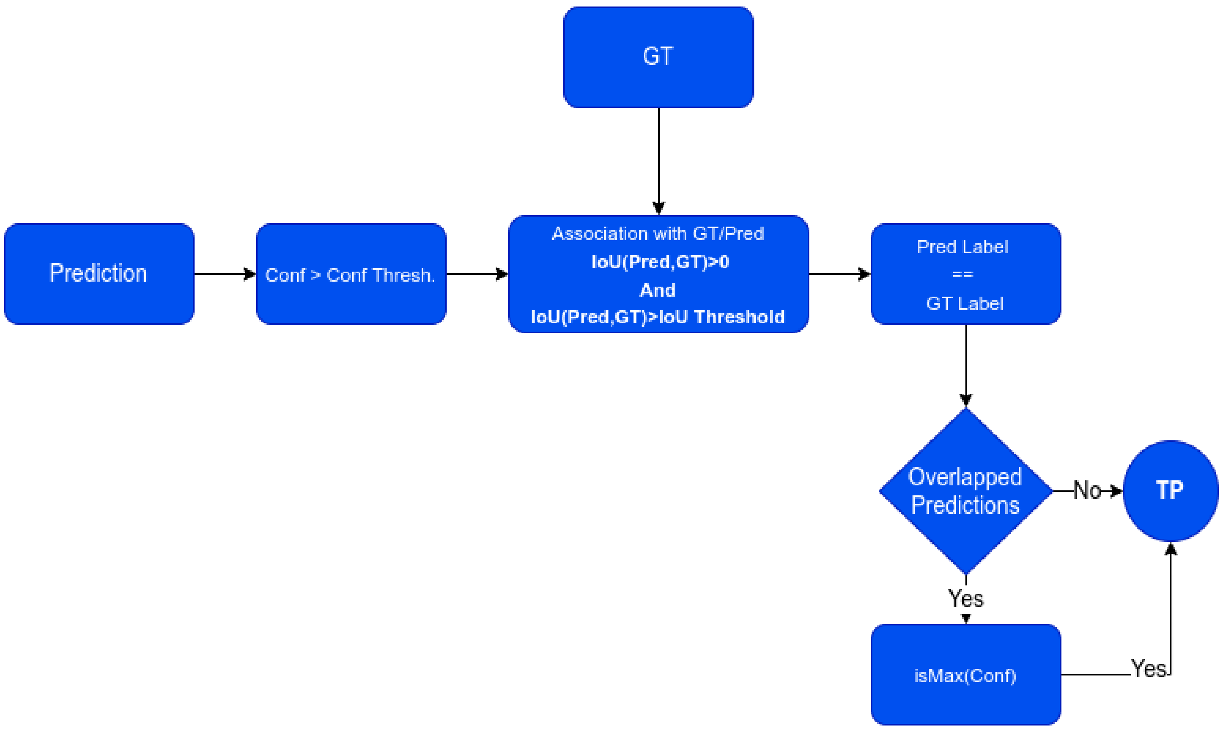
Evaluation Metrics Cases

As previously discusses, it exists 4 different linkage case as follows: TP, TN, FP, and FN. The evaluation approach is heavily inspired by Pascal VOC[2] and the open-source evaluation toolkit[1].

True Positive (TP)

To have a true positive association the following rules should be met:

a. Prediction confidence should be higher than the confidence threshold[1].
b. An association should be made with the ground truth[2]. IoU(Pred, GT)>0.
c. The association IoU should be higher than IoU Threshold[2].
d. The predicted label class should be the same as the ground truth label class.
In the case of multiple overlapping predictions, the prediction having the highest confidence score is associated with the GT[2].



TRUE NEGATIVE (TN)

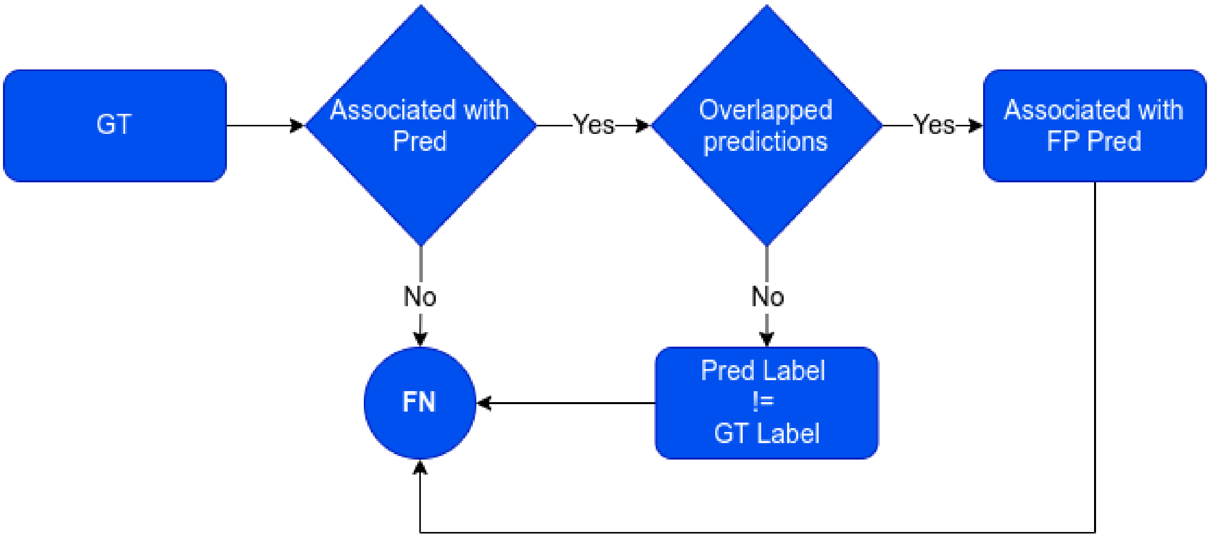
To have a true negative association the following rules should be met:

a. Image labeling should be empty indicating the absence of a detectable object.
b. The image prediction response should be also empty.

FALSE NEGATIVE (FN)

False-negative association is the result of one of the following cases:

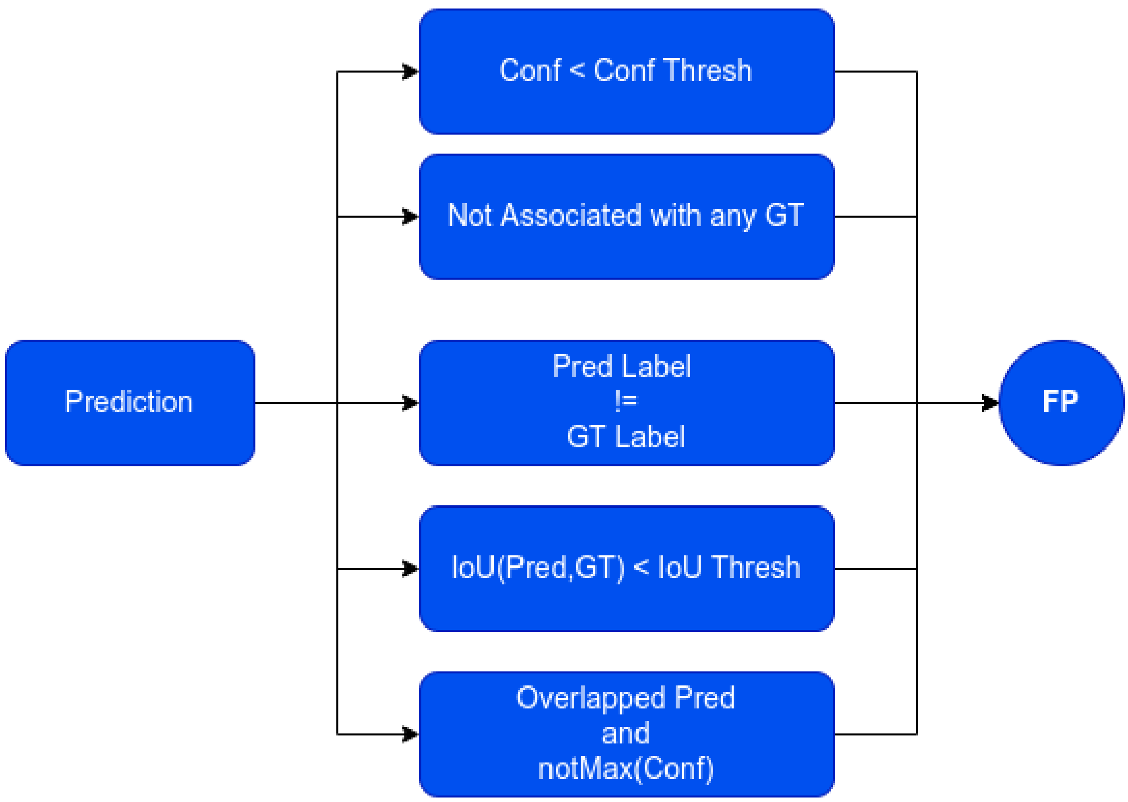
1. No prediction is associated with ground truth.
2. Prediction associated with a ground truth have different labels class[3].
3. Ground truth is associated with a false positive prediction[3]. (false positive case can be checked in the FP section)



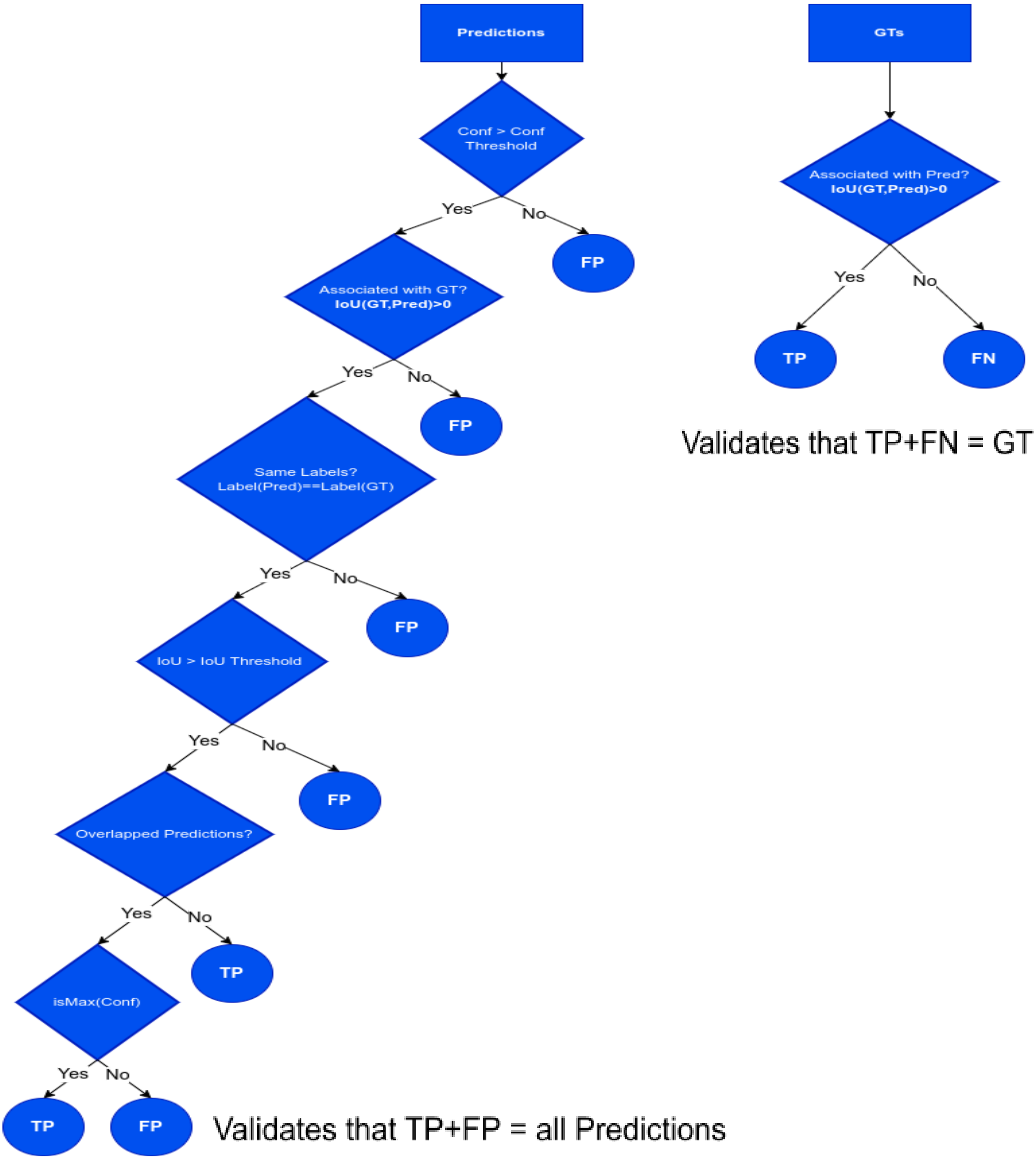
False Positive (FP)

False-positive association is the result of **one** of the following scenarios:

1. Prediction confidence is less than the confidence threshold[1].
2. Prediction is not associated with any ground truth (case of prediction only). **IoU(Pred,Any(GT)) < 0.**
3. In the case of association with ground truth, the predicted label class is different from the ground truth label class.
4. In the case of association with ground truth, the intersection over union score does not meet the threshold. **IoU(Pred, GT)< IoU Threshold.**
5. In the case of multiple predictions overlapping the same ground truth, the prediction does not meet the maximum prediction confidence between overlapping predictions. **isMax(Conf(~Current_Pred), All(~Pred)) == False.**



Evaluation Process Flow Chart



Validates that TP+FN = GT

Validates that TP+FP = all Predictions

References

[1] Padilla, R., Passos, W. L., Dias, T. L., Netto, S. L., & da Silva, E. A. (2021). A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics, 10(3), 279.

[2]Hoiem, D., Divvala, S. K., & Hays, J. H. (2009). Pascal VOC 2008 challenge. World Literature Today, 24.

[3] K. Koech, On Object Detection Metrics With Worked Example. *Medium* (2022), (available at https://towardsdatascience.com/on-object-detection-metrics-with-worked-example-216f173ed31e).