1 Metrics

Evaluating our machine learning algorithm is an essential part of any project. The way we choose our metrics influences how the performance of machine learning algorithms is measured and compared. They influence how to weight the importance of different characteristics in the results and finally, the ultimate choice of which algorithm to choose.

1.1 Precision, Recall & F1 score

Precision measures how accurate is your predictions. i.e. the percentage of your predictions are correct.

Recall measures how good you find all the positives. For example, we can find 80% of the possible positive cases in our top K predictions.

Their definisions are:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

where

- TP = True positive
- TN = True negative
- FP = False positive
- FN = False negative

1.2 Intersection over Union

Intersection over Union (IoU) measures the overlap between 2 boundaries. We use that to measure how much our predicted boundary overlaps with the ground truth (the real object boundary). In some datasets, we predefine an IoU threshold (say 0.5) in classifying whether the prediction is a true positive or a false positive.

1.3 Our metrics

According to [2] we use the following metrics in order to quantify our results:

frame-AP measures the area under the precision-recallcurve of the detections for each frame (similar to the PASCAL VOC detection challenge [1]). A detection is correct if the intersection-over-union with theground truth at that frame is greater than and the ac-tion label is correctly predicted.

video-AP measures the area under the precision-recallcurve of the action tubes predictions. A tube is correctif the mean per frame intersection-over-union with theground truth across the frames of the video is greaterthan and the action label is correctly predicted.

AUC measures the area under the ROC curve, a metricpreviously used on this task. An action tube is correctunder the same conditions as invideo-AP. Following [3] , the ROC curve is plotted until a false positive rateof 0.6, while keeping the top-3 detections per class andper video. Consequently, the best possible AUC scoreis 60%.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[2] Georgia Gkioxari and Jitendra Malik. Finding action tubes. *CoRR*, abs/1411.6031, 2014.

[3] Yicong Tian, Rahul Sukthankar, and Mubarak Shah. Spatiotemporal deformable part models for action detection. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 2642–2649, Washington, DC, USA, 2013. IEEE Computer Society.