EVERY HUMAN MAKES MISTAKES: EXPLORING THE SENSITIVITY OF DEEP-LEARNED OBJECT DETECTORS TO HUMAN ANNOTATION NOISE

1) BACKGROUND

- Recent success in object detection depends on meticulously annotated, large scale datasets.
- On Amazon's Mechanical Turk, precisely annotating a single object takes an average of 88 seconds [1], highlighting the high cost of the annotation process.
- **Current Research:** efforts to reduce annotation costs primarily focus on noise correction before or during training.
- **Gap:** analysis of effect of specific types of human annotation noise on detector performance.

2) RESEARCH QUESTION

- How sensitive are deep-learned object detectors to the four types of human annotation noise?

3) 4 TYPES OF HUMAN ANNOTATION NOISE

(A) Missing Annotations

(B) Extra Annotations

(C) Inaccurate Bounding Boxes

(D) Wrong Classification Labels

9) CONTACT INFO

Author: Laurens Michielsen - L.L.Michielsen@student.tudelft.nl
Responsible Professor: Dr. Jan van Gemert
Supervisor: Dr. Osman S. Kayhan

4) METHODOLOGY

- For each noise type:
  1. Train and evaluate model on noise-free train and test sets
  2. For i in range(5):
     a. Generate additional 10% of noise
     b. Train model on corrupted train set
     c. Evaluate model on noise-free test set
  3. Aggregate the results over the runs
- Experiments conducted:
  - 3 times for YOLOv8 with PASCAL dataset
  - Once for YOLOv8 with VisDrone dataset
  - Once for YOLOv8 with Brain-Tumor dataset
  - Once for Faster R-CNN with PASCAL dataset

6) CONCLUSIONS

- Annotation noise in smaller datasets harms performance more than in larger datasets.
- Noise in the annotations of smaller objects harms performance more than larger objects.
- YOLOv8 shows resilience to low levels of missing annotations and inaccurate bounding boxes, but is sensitive to all levels of wrong classification labels.
- Extra annotations have a regularizing effect on YOLOv8.
- Faster R-CNN is more sensitive to all noise-types compared to YOLOv8, except for inaccurate bounding boxes, where performance is similarly.

7) LIMITATIONS

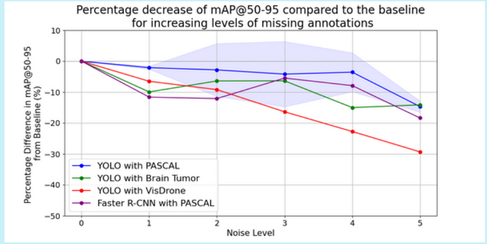
- YOLOv8 with VisDrone and Brain-Tumor and Faster R-CNN with PASCAL are each run only once.
- Hyperparameter tuning was not performed.
- Hyperparameters were equal across all noise types and noise levels.

8) REFERENCES

[1] Su, H., Deng, J., Fei-Fei, L.: Crowdsourcing annotations for visual object detection. In: AAAI Human Computation Workshop (2012)

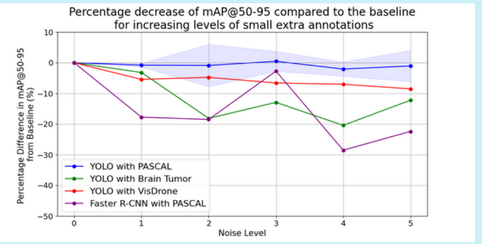
5) EXPERIMENTAL RESULTS

Missing Annotations:



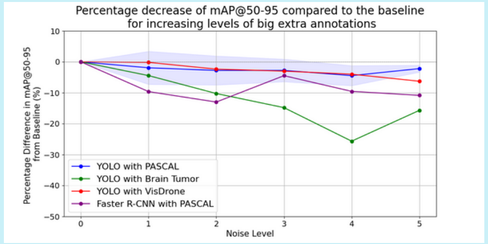
YOLOv8 is more resilient to missing annotations compared to Faster R-CNN. YOLOv8 maintains performance with greater variability at lower noise levels. Missing annotations of smaller objects and missing annotations in smaller datasets have a more detrimental effect on performance compared to their larger counterparts.

Small Extra Annotations:



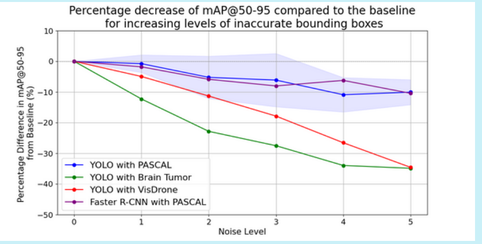
Small extra annotations have a regularizing effect on YOLOv8. Small extra annotations impact performance more in the presence of a significant amount of small objects. Faster R-CNN is sensitive to small extra annotations. Small extra annotations in smaller datasets cause a pronounced performance decline.

Big Extra Annotations:



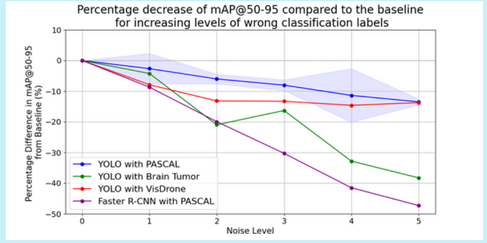
Big extra annotations have a regularizing effect on YOLOv8. Faster R-CNN is sensitive to big extra annotations, but less compared to small extra annotations. Big extra annotations in smaller harm performance more compared to larger datasets.

Inaccurate bounding boxes:



YOLOv8 and Faster R-CNN are equally robust to to inaccurate bounding boxes, with moderate sensitivity at low levels and severe sensitivity at higher levels. Inaccurate bounding boxes for small objects and smaller datasets harm performance more signifcantly compared to larger objects and datasets.

Wrong Classification Labels:



YOLOv8 is sensitive to all levels of wrong classification labels. Faster R-CNN is significantly more sensitive to wrong classification labels compared to YOLOv8. Wrong classification labels of smaller objects and smaller datasets harm detector performance more compared to their larger counterparts.