Implementing a Positive Control

6/11/24

Introduction

Establishing a positive control is necessary to validate that my deep learning model is performing as expected. This validation requires comparing evaluation metrics generated by my model against established benchmarks from the literature using an identical task and dataset. A significant discrepancy would indicate potential model dysfunction. The challenge is finding a paper to which I can compare my model.

I did find this paper: 'Pedestrian Segmentation from Complex Background Based on Predefined Pose Fields and Probabilistic Relaxation'. The paper compares an image segmentation method to CNN-based methods. They use the Penn Fudan dataset - an image database of pedestrians around a campus/urban environment. Each of the 170 samples contains at least one labelled pedestrian (one mask, one bounding box).

In section 4.2.1 of the paper's 'Quantitative Evaluation', they compare their method notably to a Mask R-CNN - this model architecture is almost identical to that of my model. For the sake of completeness, I will give an overview here:

Mask R-CNN Structure:

Firstly, the network uses a CNN module to extract feature maps using numerous kernels.

These feature maps are passed through to the Region Proposal Network (RPN) module that considers pixel k-many 'windows' at evenly spaced 'anchors' over each filter map. This network learns to select windows that most likely contain an object of interest and 'proposes' them to downstream Fully-Connected layers following ROI pooling (to normalise dimensionality anchor windows differing in size and aspect ratio). In this way the network only pays attention to promising regions within the samples, without the need for a selective-search algorithm which is computationally less efficient. Up to here, the Mask R-CNN is identical to the Faster R-CNN architecture.

The discrepancy lies in the output heads. A Faster R-CNN network has two output heads: one for classification (kx2 outputs for each ROI), and one for bounding-box regression (kx4 outputs for each ROI). The Mask R-CNN has an additional output head that outputs the object mask for a given input sample.

As this is the only discrepancy, I believe its performance in the paper is suitable as a reference for the positive control.

Goal:

Validate my model's performance (Average Precision and Recall (AP and AR)) relative to the findings in a sufficiently similar implementation example from the literature.

Hypothesis:

If my Faster R-CNN implementation is functioning correctly, it should achieve detection metrics (AP and AR) comparable to the published Mask R-CNN benchmarks on the Penn-Fudan dataset.

Rationale:

1. The core detection architecture is identical
2. The dataset and task (pedestrian detection) are standardized
3. The evaluation protocols for AP and AR are consistent
4. The segmentation head in Mask R-CNN does not affect detection metrics

Experimental Plan:

1. Train Faster R-CNN on Penn-Fudan dataset using:
    - ResNet-50 backbone
    - Standard detection heads
    - Default training parameters
2. Evaluate using COCO metrics:
    - Average Precision (AP)
    - Average Recall (AR)
3. Compare against published benchmarks:

- Mask R-CNN: AP = 79.25%, AR = 92.63%
- Other architectures (for context):

- Yolact++: AP = 92.20%, AR = 94.02%
- DeepLabv3: AP = 78.06%, AR = 92.83%

Reference Selection

The paper "Pedestrian Segmentation from Complex Background Based on Predefined Pose Fields and Probabilistic Relaxation" (Caisse Amisse, Jijón-Palma and António, 2021) provides suitable benchmark metrics for comparison. They evaluate multiple CNN-based architectures on the Penn-Fudan dataset, which contains 170 images of pedestrians in urban environments with pixel-level annotations (masks and bounding boxes).

Architectural Comparison

Base Architecture Similarity

The paper benchmarks Mask R-CNN, which shares the same fundamental detection architecture as my Faster R-CNN implementation:

1. Backbone: ResNet-50 feature extractor
2. Region Proposal Network (RPN)
3. ROI Pooling layer
4. Classification and bounding box regression heads

Key Differences

The main architectural difference is that Mask R-CNN includes an additional segmentation head for mask prediction, while my Faster R-CNN implementation focuses solely on detection. This difference could potentially impact detection performance through:

1. Multi-task Learning Effects:
    - The additional mask supervision might help the shared layers learn better feature representations
    - The model must balance detection and segmentation objectives, which could affect optimization
2. Parameter Updates:
    - Gradients from the mask head flow back through the shared layers
    - This could influence how the detection-related parameters are updated during training

However, these implementations still serve as valid reference points because:

1. The core detection architecture remains identical
2. The published metrics provide a reasonable expected performance range

Multi-Study Validation

Two independent studies support the comparison Faster R-CNN and Mask R-CNN metrics in the positive control:

1. - Pedestrian Detection Reference Study (Caisse Amisse, Jijón-Palma and António, 2021)
   - Mask R-CNN: AP = 79.25%, AR = 92.63%
   - Similar task (except for mask generation), making it a good reference study for the positive control
   - The frozen COCO ResNet50 backbone is identical to the one in my model (they also use transfer learning)
2. Vehicle Detection Study (Tahir, Shahbaz Khan and Owais Tariq, 2021)

   - Faster R-CNN: AP = 76.3%, AR = 76%
   - Mask R-CNN: AP = 74.3%, AR = 74.35%
   - Shows consistent relative performance between the two architectures

These studies demonstrate that:

1. Faster R-CNN and Mask R-CNN may achieve comparable metrics
2. My implementation's performance (AP = 87%, AR = 92%) is similar to that of the Mask R-CNN with a ResNet50 backbone

Results

My implementation achieved:

- AP @ IoU 0.50:0.95 = 87% (8% > Mask R-CNN in Caisse Amisse, Jijón-Palma and António, 2021)
- AR @ IoU 0.50:0.95 = 92% (equivalent to Mask R-CNN AR in Caisse Amisse, Jijón-Palma and António, 2021)

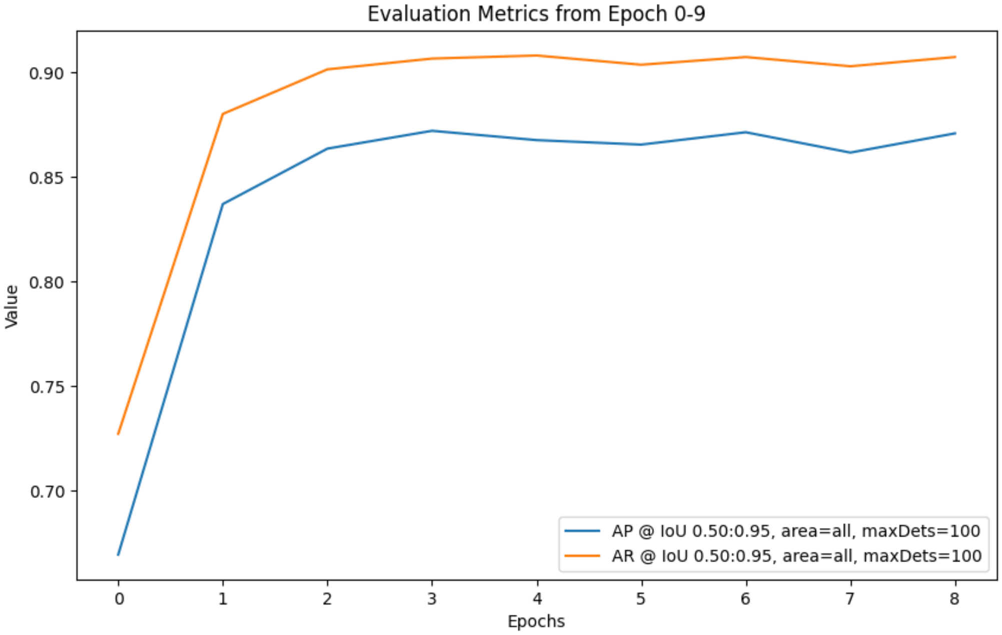

Figure 1: mAP and mAR

These metrics fall well within the expected range established by the literature benchmarks, validating that my implementation is functioning correctly. The higher discrepancy in AP could be attributed to the additional mask-predictor head in the reference study.

Conclusion

The positive control demonstrates that my Faster R-CNN implementation:

1. Achieves performance consistent with published benchmarks
2. Shows no evidence of implementation errors or dysfunction
3. Can be confidently applied to new detection tasks

- Caisse Amisse, Jijón-Palma, M.E. and António, J. (2021). PEDESTRIAN SEGMENTATION FROM COMPLEX BACKGROUND BASED ON PREDEFINED POSE FIELDS AND PROBABILISTIC RELAXATION. Boletim de Ciências Geodésicas, [online] 27(3). doi:https://doi.org/10.1590/s1982-21702021000300017.

- Tahir, H., Shahbaz Khan, M. and Owais Tariq, M. (2021). Performance Analysis and Comparison of Faster R-CNN, Mask R-CNN and ResNet50 for the Detection and Counting of Vehicles. 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). doi:https://doi.org/10.1109/icccis51004.2021.9397079.