Review: End-to-End Object Detection with Transformers

Sanatan Mishra
October 18, 2025

Goals

- Object detection: predict a set of labels and bounding boxes for each relevant object
- Simplify previous object detection systems by removing the need for postprocessing to clean up model output

Innovation: DETR Model

- Problem statement: see object detection as a simple set detection model
- Model architecture: convolutional neural network (CNN) backbone → encoder-decoder transformer model → 3-layer feed-forward network
  - Uses parallel sequence generation as autoregression is too costly
- Loss function: sum of pairwise matching costs for each boundary box predicted
  - Minimizes error in set prediction, consistent with the problem statement

Results/Conclusion

- DETR has similar performance to Faster R-CNN on most subsets of the COCO dataset without requiring as much postprocessing or other domain knowledge
- DETR is significantly better than Faster R-CNN on large objects due to the self-attention
- All parts of the DETR model and loss function necessary
- DETR lacks strong class-specialization, allowing for generalization to out-of-distribution images
- Limitations:
    - DETR not as good as Faster R-CNN on small objects
    - Possibly compute-intensivity (not mentioned in the paper but an important consideration for drone applications)