Rich feature hierarchies for accurate object detection and semantic segmentation

Link

arXiv

Summary

- This paper uses deep CNN as feature extractor to localize and segment objects. They also showed that supervised pre-training on a large dataset (ILSVRC) followed by fine-tuning on a target smaller dataset (PASCAL) yields significant performance boost on the target dataset.
- Their object detection system consists of three modules.
    1. Generate a category independent region proposals using selective search.
    2. Use a large CNN to extract fixed length feature vector from each region.
    3. Use class specific linear SVM to classify each feature vector.
- Two properties the detection efficient compared to other approaches at that time. First is the features extracted by CNN are shared across all categories. And second is feature vectors computed by CNN (4096-dimensional) were low-dimensional compared to others.
- In each SGD iteration they sample 32 positive windows and 96 negative windows. The sampling was biased towards the positive windows because there are lot more background windows compared to actual objects in original image.
- When a region contains both object and background, they use intersection over union (IOU) to label the region. If the overlap is below an IOU threshold, it is defines as negative for the object. The threshold used was 0.3, which was obtained by a grid search. Using a higher or lower threshold resulted in decreased performance.
- When no fine-tuning is used feature from second fully connected layer fc_7 performs worse than features from the first fully connected layer fc_6. It suggests that the $fc_7 7$ learns too much domain-specific features for ILSVRC dataset. Even if both fully connected layers are removed, features from the convolution layer only ($pool_5$), produces quite good results.
- Fine-tuning improves performance significantly. The boost for fine-tuning for fc_6 and fc_7 is larger than only using $pool_5$ features.
- Fine-tuning from VGG net substantially outperforms fine-tuning from AlexNet, increasing mean average precision from 58.5% to 66%.
- They also use bounding box regression to reduce localization error. For this they train a linear regression model to predict new detection window for each class given $pool_5$ features for a given selective search region proposal. This increases mean average precision by 3-4%.