Faster R-CNN

Summary

1. Given the invention of fast R-CNN, the computational bottleneck of obj detection network is the region proposal component.

2. Propose Region Proposal Network that shares the conv. features with the detection network, thus enabling cost-free proposals.

Faster RCNN model

Region Proposal Network

1. Given a image, a ConvNet takes it as input and produces feature maps.
2. Given each location in the feature map, a $n \times n$ spatial window centered at the location is mapped to a lower-dimensional feature with a $n \times n$ conv. layer.
3. This feature is mapped into 2 separated FCN, to predict the bb regression parameters and objectness score, which together are referred to as region proposals.
4. At each location in the feature map, we predict multiple region proposals, each associated with an anchor.
5. An anchor is a pre-defined rectangular boxes with a specific scales and aspect ratios.

Classification

1. Given the region proposals and the feature maps produced by step 1 above, fast R-CNN is used to assign obj category label to the region proposals.

Benefits of the anchor-based detection

Translation-invariant anchors

1. "If one translates an obj. in an img., the proposal should translate accordingly."
2. The method is translation-invariant in terms of the anchor and the fun that computes the region proposals relative to the anchors.

Reduced model size

1. Each anchor has 6 conv. output layers (4 for bb regression and 2 for objectness score), so the no. of parameters scales linearly in the no. of conv. output layers.

Novel scheme for addressing multiple scales and ratios

1. Their method is based on "pyramid of anchors".
2. Only relies on images and feature maps of a single scale
3. Allows for sharing the bulk of the feature b/t the region proposal components and the object classifier without extra cost for addressing scales.
4. Prev. methods to address scale either :
   - Use input imgs of different scales and the feature maps are computed separately for each scale.
   - Use sliding window of multiple scale on the feature maps.

Loss functions for RPN

1. $$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*)$$

2. i is the idx of an anchor in a mini-batch.
3. p_i is the predicted probability of an anchor i being an object.
4. p_i^* is the GT label, and is 1 if the anchor corresponds to an obj and is 0 otherwise.
5. L_{cls} is the log-loss.
6. t_i are the predicted bb parameters.
7. $L_{reg}(t_i, t_i^*) = R(t_i - t_i^*)$ is the Robust L1 loss.