R-CNN

Summary

1. Argue that prev learning approaches to obj detection did not work well because:
   - localization as regression did not work as well, as demonstrated by concurrent works.
   - Sliding-window detector:
     - to maintain high spatial resolution, these CNNs are not deep.
     - units high up in the network have large receptive fields and large strides, which makes precise localization hard.
2. They follow the "recognition using regions" paradigm.
3. At test time, their approach:
   - generate 2000 category-independent region proposals for the ipt img.
   - use affine image warping to compute a fixed-size CNN input from each region proposals.
   - extract a fixed-length feature vector from each warped region proposal.
   - classify each region with category-specific linear SVMs.
4. An additional simple bb regression method significantly reduce mislocalizations, which are the dominant error mode.
5. Given all scores for the regions computed by the SVM, they apply greedy non-maximum suppression for each class independently.
6. The NMS rejects a region if it has a IoU with a higher scoring selected region larger than a learned threshold.

Shared feature computations improves run-time efficiency

1. Two properties of their approaches are helpful for comp. efficiency:
   - The time spent on computing region proposals and CNN features are shared across all classes.
   - The features computed by the CNN are low-dimensional.

An interesting technique to visualize what the network has learned

1. Pick a single unit
2. Compute the unit's activations on a large set of held-out region proposals
3. Sort the proposals from highest to lowest activation
4. Perform NMS
5. Display the top-scoring regions.
6. They found units that fire on dog faces, etc.

Mini-batch selection

1. In each SGD iteration, they sample 32 positive windows (over all classes) and 96 background windows to construct a mini-batch of size 128.
2. They bias the sampling towards positive window because they are extremely rare compared to background.

Other interesting bits

1. Compared to the Deformable Part-based Model (DPM), sig. more of their errors result from poor localization, rather than confusion with background or object classes.
2. They argue that this indicates the CNN features are more discriminative than HOG features.
3. They tried 3 different img trans. methods and found that warp performs the best.

Qns

1. They mention that "loose localization likely results from our use of bottom-up regression proposals and the positional invariance learned from pre-training our CNN for whole-image cls".
What do they mean by this?

2.