Deformable Part Models are Convolutional Neural Networks

Xuewen Yang

August 21 2018

1. Object Localization Module

The Scale-Transferrable Detection Network (STDN) consists of a base network and two task specific prediction subnetworks. The role of the base network is to do feature extraction. The first subnet is used for object classification, and the second subnet is used for bounding box position regression.

1.1. Anchor Boxes.

The author associate a set of default anchor boxes with each feature map which is got by scale-transfer module. The scale of anchor boxes is the same as that of SSD [3]. Following DSSD, they use [1:6; 2:0; 3:0] aspect ratios at every prediction layer. Anchors are matched to any ground truth with intersection-over-union (IoU) higher than a threshold (0:5). The remaining anchors are treated as background. After the matching step, most of the default anchor boxes are negatives (no matched). They use hard negative mining so that the ratio between the negatives and positives is at most $3:1$.

1.2. Classification Subnet.

The role of the classification subnet is to predict the probability of each anchor belonging to a category. It contains a 1×1 convolution layer and two 3×3 convolution layers. Each convolution layer has a batchnorm layer [2] and a relu layer in front of it. The last convolution layer has K A filters, where K is the number of object classes and A is the number of anchors per spatial location. The classification loss is the softmax loss over multiple classes confidences.

1.3. Box Regression Subnet.

The purpose of this subnet is to regress the offset from each anchor box to the matched ground-truth object. The box regression subnet has the same structure as the classification subnet except that its last convolution layer has 4A filters. Smooth L1 loss is used for the localization loss and the bounding box loss is only used for positive samples

1.4. Training Settings.

Our detector is based on the MXNet [1] framework. All our models are trained with SGD solver on NVIDIA TITAN Xp GPU. We follow almost the same training strategy as SSD [3], including a random expansion data augmentation trick which is helpful for detecting small objects.

References

[1] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *ICCV*, 2015. 1

[2] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *CVPR*, 2015. 1

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 1