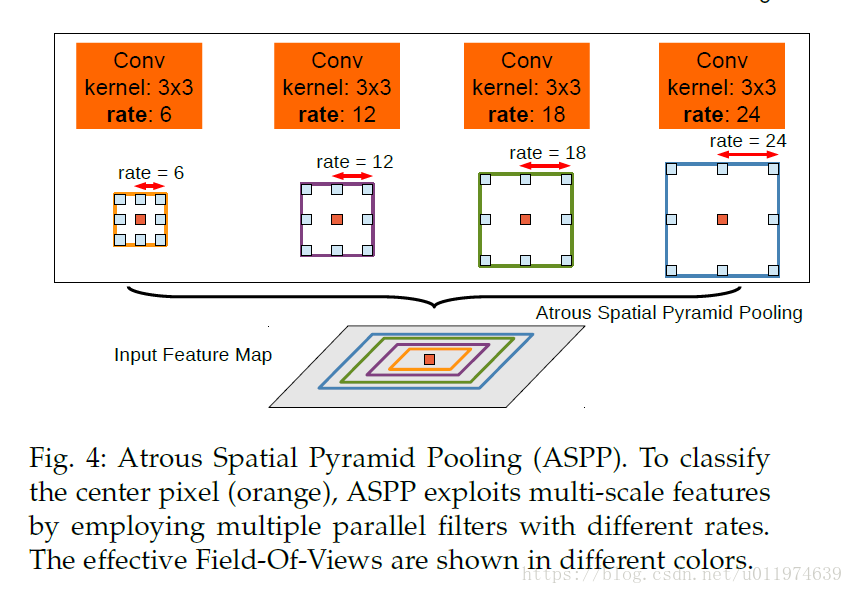
# Other



The atrous spatial pyramid pooling. To classify the center pixel, ASPP exploits multi-scale features by employing multiple parallel filters with different rates. The effective of Field-of-view are shown in different colors.

1. Learning based temporal action localization methods require vast amounts of training data. However, such largescale video datasets, which are expected to capture the dynamics of every action category, are not only very expensive to acquire but are also not practical simply because there exists an uncountable number of action classes. This poses a critical restriction to the current methods when the training samples are few and rare (e.g. when the target action classes are not present in the current publicly available datasets). To address this challenge, we conceptualize a new example-based action detection problem where only a few examples are provided
2. To this end, in this paper, we consider a one(few)-shot learning scenario of action localization; given one(or a few) example of new action classes, typically one per class,
3. Since all the network components are differentiable, our localization network can be trained in an end-to-end fashion Metric learning methods are also employed by many one-shot learning algorithms that generate good results. Deep siamese networks [22] train a convolutional network to embed examples so that samples in the same class  
   are close while samples in different classes are far away. [43, 36, 39] refine this idea by introducing recurrence and attention mechanisms.
4. The one-shot problem setup requires that the classes during testing must not be present during training. Thumos14 contains 20 classes from UCF-101, so only the other 81  
   classes in UCF-101 are used during training the encoder. We denote the two splits of UCF-101 as UCF-101-81 and UCF-101-20. After training the encoder, we use a small part of Thumos14 validation set (contain 6 classes) to train the fully connected network with the features extracted by the video encoder, and use the remaining 14 classes in the  
   test set to test our one-shot localization network. The two splits of Thumos14 validation set and test set are denoted as Thumos-val-6 and Thumos-test-14. Similarly for activity net, we split the 100 classes into 80-20 splits. Our one-shot action localization network is trained on videos containing only the 80 classes in the training set, denoted by ActivityNet-train-80, and is tested on the other 20 classes in the validation set, denoted by ActivityNet-val-20.
5. We see from Table 1 and 2 that, although there is still a performance gap between one-shot and fully supervised action detection, our method significantly outperforms the state-of-the-art method when tested in one-shot setup2 方法比较垃圾的时候该如何措辞
6. Our network is trained under the meta-learning framework so that it can quickly adapt to new classes with just one(few) training samples. 说明自己方法只要少量样本
7. This usually requires either careful feature engineering, or a significant number of  
   samples. This is far from what we desire: ideally, robots should be able to learn  
   from very few demonstrations of any given task, and instantly generalize to new  
   situations of the same task, without requiring task-specific engineering