Reaction Report

RotationNet: Joint Object Categorization and Pose Estimation Using Multiviews from Unsupervised Viewpoints

Rushikesh Dudhat

What I like about this paper: In this paper, the authors developed a Multiview approach for object recognition where a partial set of images is can be used to make inferences about the object and viewpoint from which the image is taken. The property of RotationNets to take input images sequentially has a lot of potential application in a real-world scenario like CCTV cameras where the likelihood of the object can be updated with every new frame. Rotation nets treat viewpoints as latent variables and learn them in an unsupervised manner. Thus this eliminates the requirement for pose normalization which is often susceptible to noise and errors. In the training process of RotationNets, it is assumed that the images are captured from different predefined viewpoints. Here only the category labels are given as input. The parameters to optimize are the viewpoints and R(Neural nets). An 'incorrect view' class is added which shows the likelihood that the given image does not correspond to a viewpoint. For the correct viewpoint, the incorrect category prediction will be close to 1. This addition of the incorrect class label is one of the most significant aspects of the paper since it helps in generating cost functions for both class and pose estimations. Initially, the parameters of the RotationNet are kept constant and the viewpoint is estimated in forward propagation and then errors are backpropagated keeping the viewpoints constant. This process is done iteratively for all the images. RotationNet (with VGG-M architecture) significantly outperformed existing methods with both the ModelNet40 and ModelNet10 datasets.

What I don't like about this paper: It is not clear as to how the RotationNets identify the basis axis and achieve inter-class and intra-class pose alignment. Also, the assumption that the probability of the incorrect class should be one if the image does not correspond to the particular viewpoint is not intuitive. It is possible that for an incorrect viewpoint the image still shows a high probability of being in the correct class. And dragging down this probability during the training may cause overfit. An experiment could have also been done on real-world data like CCTV footage to see how accurately it classifies the objects. It requires the viewpoint of test images to be from a predefined set of viewpoints. However, this could be a limitation in a real-world application. For example, the images might be coming at irregular degree intervals (5°, 15°, etc.)

Future directions: There should be some consideration for the objects which are symmetric about some axis. Most of the real-world objects are symmetric about one axis or another. If such information can be embedded while training the data sets, then it is possible that we get the same predictions quickly. For example, cars are symmetric around the vertical plane passing between the headlights. Thus the symmetric views around this axis can be avoided like +/- 30 degrees, +/- 60 degrees, and +/- 90 degrees. In this case, effectively there will only be 7 views instead of 12 i.e taking views in a semi-circle instead of a circle. There should ideally be no change in the performance of the RotationNet and the training time can be effectively reduced. Additionally, real-world images will not come from predefined viewpoints. This can be solved by adding a probability term that measures the probability of an image corresponding to a particular viewpoint.