Active Multi-View Recognition with Deep Reinforcement Learning

Dinesh Narapureddy
Carnegie Mellon University
dnarapur@andrew.cmu.edu

George Tan
Carnegie Mellon University
georget1@andrew.cmu.edu

1. Introduction

Multi-view recognition is a more realistic task for object detection than single-image recognition. Traditional methods use single-view images for detection, but this ignores the other dimensions of the objects. In a real life application, agents can rotate or move objects for a multi-view image sequence of the objects which results in better learning and recognition. There has been sparse literature addressing this task and comes under the umbrella of active vision. Active multi-view recognition uses the current information to make a approximate guess of the next view to better understand a image. This mimics human behavior for various recognition tasks. For example, in case of a occluded object or viewing a object from unrecognizable viewpoints human plan an action hypothesizing the object and the next best view to recognize it. Inspired from human behavior for active vision we propose a reinforcement learning based method to recognize object categories.

1.1. Related works

Recent papers introduced CNNs for generalized Multi-view recognition by learning from images that cover the full sphere of viewpoints over an object [1] [4]. [5] have proposed to combine multiple view features using a view based fully connected network and show the state-art-the art object recognition. Next best view selection for 3D object recognition has been proposed by [5]. Taking advantage of the multi-image sequences, the agent can learn the best trajectory with pairs of images to achieve faster recognition [2]. Although most of the methods have addressed the problem with different approaches, they introduce prior knowledge and bias in how to search for the next best view. We believe reinforcement learning as the best method to learn the next best view selection. Recent improvements in deep reinforcement learning [3] show improvements to human level performance on tasks.

2. Proposal

We propose a deep reinforcement learning method to learn the next best view. We will use a Deep Q-Network [3] to learn the best actions, i.e. which direction to move for the next best view, for a given image. The method will be evaluated on the ModelNet40 dataset [5], which are 3D CAD models from the 40 categories at different viewpoints. Recognition accuracy and number of actions will be measured. Although accuracy may be the most important metric, it would be best if the least number of actions are taken while searching for the next best view.

References

[1] D. Jayaraman and K. Grauman. Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion. *CoRR*, abs/1605.00164, 2016. 1

[2] E. Johns, S. Leutenegger, and A. J. Davison. Pairwise decomposition of image sequences for active multi-view recognition. *CoRR*, abs/1605.08359, 2016. 1

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013. 1

[4] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. *CoRR*, abs/1505.00880, 2015. 1

[5] Z. Wu, S. Song, A. Khosla, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. *CoRR*, abs/1406.5670, 2014. 1