Efficient Recognition and Classification of Objects of Interest in Video Streams

Author : Jatin Aggarwal
Type of work: Research seminar

Motivation

In this paper, we are driving a state-of-the-art method of object detection. We explore optimization techniques for the classification of objects efficiently. For this, we are reviewing different works that categorically define action recognition. As we know 3D convolution neural network model is computationally expensive in terms of memory and run-time. We want to find more efficient ways to process the videos for downstream tasks. Therefore, our focus is a 2D convolution model which is memory efficient with good performance. Based on the model comparisons and knowledge of these works, we distill the technique in one state-of-the-art model. We explore the application for scene representation learning in videos from different data sets.

Task Description

In this work, we are conducting a literature study to find different approaches to efficient video processing and comparing them based on a same data set. These papers are based on different data sets with literature criteria explained in bluets point to understand the task.

- Find out the research literature for object recognition and classification.
- Use only 2D neural network study for further consideration, no 3D convolution model.
- Categorizing and comparison of action recognition techniques.

Relevant literature

Our focus of this study is based on the following literature.

- AdaFocus V2: "End-to-End Training of Spatial Dynamic Networks for Video Recognition" First, the authors proposed a differentiable interpolation-based operation for selecting patches. That allows the gradient back-propagation throughout the whole model. Then explained three tailored training techniques. Which addresses the optimization problems during end- to-end training. Sth-Sth V2 data set is used.
- VidConv: "A modernized 2D ConvNet for Efficient Video Recognition"[2] The main objective of this paper they explains how 2D Convolution neural network is inferior to 3D ConvNet is not true in all cases concerning action recognition. Furthermore, this study

selects a minimal design of the 2D ConvNet and proves that its performance is still very competitive while preserving efficiency. Also use Sth-Sth V2 data set.

- GabriellaV2: "Towards better generalization in surveillance videos for Action Detection"[3]. This paper focused on a system based on tracklet generation using an object detector with a tracker. This study shows how to resolve overlapping actor problems with a good explanation of tracklet action classification and post-processing units. The data set in this study was AVA-kinetics.

- "Exploiting Instance-based Mixed Sampling via Auxiliary Source Domain Supervision for Domain-adaptive Action Detection"[4]. This study proposed DA-AIM, a novel algorithm tai- lored for unsupervised domain adaptive action detection. DA-AIM considers the inherent characteristics of action detection and mixes 3D video clips. AVA-kinetics data set is used here.

- "A practitioner's guide to improve the logistics of spatiotemporal deep neural networks for animal behavior analysis"[5]. This publication explores a suite of optimization techniques for representative neural network architecture. The I3D model was trained to perform action classification on freely behaving mice in a home cage setup. This use HMDB51 data set, to find the suggested result that simple optimizations in data loading protocols and network specification yield significant reductions with model the run-time and system's overall accuracy without any scarify.

References

1 Y. Wang et al., "AdaFocus V2: End-to-End Training of Spatial Dynamic Networks for Video Recognition," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20030-20040, doi: 10.1109/CVPR52688.2022.01943.

2 Ishandave, zaccy, akash k, sarah.shiraz@knights.ucf.edu, yogesh, shah@crcv.ucf.edu., "GabriellaV2 Towards better generalization in surveillance videos for Action Detection" The WACV workshop in IEEE computer vision foundation 2022.

3 Chuong H. Nguyen, Su Huynh, Vinh Nguyen, Ngoc Nguyen CyberCore AI, Ho Chi Minh, Viet Nam., VidConv: A modernized 2D ConvNet for Efficient Video Recognition., https://arxiv.org/abs/2207.03782v1 2022

4 Yifan Lu, Gurkirt Singh, Suman Saha, Luc Van Gool.,Exploiting Instance-based Mixed Sampling via Auxiliary Source Domain Supervision for Domain-adaptive Action Detection., https://arxiv.org/pdf/2209.15439.pdf.,2022

5 Lakshmi Narasimhan Govindarajan Rohit Kakodkar Thomas Serre lakshmi govindarajan, rohit kakodkar, thomas serre Brown University, Providence, RI, USA A practitioner's guide to improve the logistics of spatiotemporal deep neural networks for animal behavior analysis.https://serre-lab.clps.brown.edu/wp-content/upload/2022/09/VAIB22$_{C}ameraReady.pdf.$, 2022