International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01850-6

Dynamic Context Removal: A General Training Strategy for Robust Models on Video Action Predictive Tasks

Xinyu Xu[1] · Yong-Lu Li[1] · Cewu Lu[1]

Received: 10 October 2022 / Accepted: 12 July 2023

Abstract

Predicting future actions is an essential feature of intelligent systems and embodied AI. However, compared to the traditional recognition tasks, the uncertainty of the future and the reasoning ability requirement make prediction tasks very challenging and far beyond solved. In this field, previous methods usually care more about the model architecture design but little attention has been put on how to train models with a proper learning policy. To this end, in this work, we propose a simple but effective training strategy, Dynamic Context Removal (DCR), which dynamically schedules the visibility of context in different training stages. It follows the human-like curriculum learning process, i.e., gradually removing the event context to increase the prediction difficulty till satisfying the final prediction target. Besides, we explore how to train *robust* models that give consistent predictions at different levels of observable context. Our learning scheme is *plug-and-play* and easy to integrate widely-used reasoning models including Transformer and LSTM, with advantages in both effectiveness and efficiency. We study two action prediction problems, i.e., Video Action Anticipation and Early Action Recognition. In extensive experiments, our method achieves state-of-the-art results on several widely-used benchmarks.

Keywords Dynamic Context Removal · Video Action Anticipation · Early Action Recognition · Robustness

1 Introduction

A comprehensive understanding of action sequences, e.g., `open the can before pouring water out`, is a basic ability of humans. We usually know how to take multiple action steps to achieve a final target and are easy to reason out the next action based on the past context. It puts new requirements on embodied AI as advanced intelligence should possess the ability to understand the action order and predict the next one. Thus, action prediction matters. It also serves as a support for many applications like autonomous driving (Alvarez et al., 2020; Rasouli et al., 2019) and human-robot interaction (Koppula & Saxena , 2015; Ryoo et al., 2015), where the predictions on pedestrians and users are essential.

With the rapid evolution of deep learning techniques, the comprehensive understanding and analysis of human action videos attract attention in edging research. In the traditional recognition field, modern video models (Carreira & Zisserman , 2017; Fan et al., 2021; Feichtenhofer et al., 2019; Lin et al., 2019; Simonyan and Zisserman , 2014; Tran et al., 2018, 2019; Wang et al., 2016, 2018) leverage spatiotemporal modeling to learn both spatial patterns and temporal logics and achieve significant progresses in many video *recognition* problems (Damen et al., 2021; Goyal et al., 2017; Kay et al., 2017). Besides, there is also a growing interest in action *prediction* problems (Damen et al., 2018, 2021; Kuehne et al., 2014; Li et al., 2018; Stein & McKenna , 2013). Similarly, they both expect systems to discriminate the existing actions in videos. Differently, the observed video segment given for systems shifts in action prediction problems, while action recognition systems have all the contents of actions from videos.

However, due to the temporal misalignment between visual observation and target action semantics, action prediction problems are much more challenging than action

Communicated by Yoichi Sato.

✉ Yong-Lu Li
yonglu_li@sjtu.edu.cn

✉ Cewu Lu
lucewu@sjtu.edu.cn

Xinyu Xu
xuxinyu2000@sjtu.edu.cn

[1] Shanghai Jiao Tong University, Shanghai 200240, China

Published online: 13 August 2023