Video You Only Look Once: Overall Temporal Convolutions for Action Recognition

Longlong Jing, Xiaodong Yang, Yingli Tian

PII: S1047-3203(18)30023-3
DOI: https://doi.org/10.1016/j.jvcir.2018.01.016
Reference: YJVCI 2128

To appear in: *J. Vis. Commun. Image R.*

Received Date: 25 July 2017
Accepted Date: 27 January 2018

Please cite this article as: L. Jing, X. Yang, Y. Tian, Video You Only Look Once: Overall Temporal Convolutions for Action Recognition, *J. Vis. Commun. Image R.* (2018), doi: https://doi.org/10.1016/j.jvcir.2018.01.016

Video You Only Look Once: Overall Temporal Convolutions for Action Recognition

Longlong Jing[a], Xiaodong Yang[b], Yingli Tian[a,c,*]

[a] The Graduate Center, City University of New York
[b] NVIDIA RESEARCH
[c] The City College of New York, City University of New York

Abstract

In this paper, we propose an efficient and straightforward approach, video you only look once (VideoYOLO), to capture the overall temporal dynamics from an entire video in a single process for action recognition. It remains an open question for action recognition on how to deal with the temporal dimension in videos. Existing methods subdivide a whole video into either individual frames or short clips and consequently have to process these fractions multiple times. A post process is then used to aggregate the partial dynamic cues to implicitly infer the whole temporal information. On the contrary, in VideoYOLO, we first generate a proxy video by selecting a subset of frames to roughly reserve the overall temporal dynamics presented in the original video. A 3D convolutional neural network (3D-CNN) is employed to learn the overall temporal characteristics from the proxy video and predict action category in a single process. Our proposed method is extremely fast. VideoYOLO-32 is able to process 36 videos per second that is 10 times and 7 times faster than prior 2D-CNN (Two-stream [1]) and 3D-CNN (C3D [2]) based models, respectively, while still achieves superior or comparable classification accuracies on the benchmark datasets, UCF101 and HMDB51.

Keywords: Video Understanding, Video Classification, Action Recognition, Convolutional Neural Network

1. Introduction

With more videos flourishing on the Internet, video-based applications such as automatic categorization, searching, indexing, and retrieval of videos have drawn significant attention from the multimedia community. As one of the fundamental tasks for video analytics, action recognition provides the crucial visual cues and is a field of increasing importance field in vision research. Unlike the image-based applications such as image classification, object

[*] Corresponding author

Email addresses: ljing@gradcenter.cuny.edu (Longlong Jing), xiaodongy@nvidia.com (Xiaodong Yang), ytian@ccny.cuny.edu (Yingli Tian)