A Perceptual Prediction Framework for Self Supervised Event Segmentation

Sathyanarayanan N. Aakur
University of South Florida
Tampa, FL, USA
saakur@mail.usf.edu

Sudeep Sarkar
University of South Florida
Tampa, FL, USA
sarkar@usf.edu

Abstract

Temporal segmentation of long videos is an important problem, that has largely been tackled through supervised learning, often requiring large amounts of annotated training data. In this paper, we tackle the problem of self-supervised temporal segmentation that alleviates the need for any supervision in the form of labels (full supervision) or temporal ordering (weak supervision). We introduce a self-supervised, predictive learning framework that draws inspiration from cognitive psychology to segment long, visually complex videos into constituent events. Learning involves only a single pass through the training data. We also introduce a new adaptive learning paradigm that helps reduce the effect of catastrophic forgetting in recurrent neural networks. Extensive experiments on three publicly available datasets - Breakfast Actions, 50 Salads, and INRIA Instructional Videos datasets show the efficacy of the proposed approach. We show that the proposed approach outperforms weakly-supervised and unsupervised baselines by up to 24% and achieves competitive segmentation results compared to fully supervised baselines with only a single pass through the training data. Finally, we show that the proposed self-supervised learning paradigm learns highly discriminating features to improve action recognition.

1. Introduction

Video data can be seen as a continuous, dynamic stream of visual cues encoded in terms of coherent, stable structures called "*events*". Computer vision research has largely focused on the problem of recognizing and describing these events in terms of either labeled actions [19, 18, 19, 2, 1] or sentences (captioning) [1, 36, 35, 13, 5, 39]. Such approaches assume that the video is already segmented into atomic, stable units sharing a semantic structure such as "*throw ball*" or "*pour water*". However, the problem of temporally localizing events in untrimmed video has not been explored to the same extent as activity recognition or captioning. In this work, we aim to tackle the problem of temporal segmentation of untrimmed videos into its constituent events in a self-supervised manner, without the need for training data annotations.

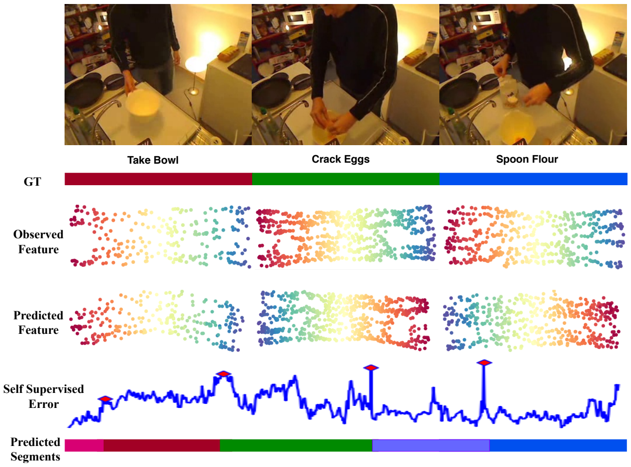

Figure 1: **Proposed Approach**: Given an unsegmented input video, we encode it into a higher level feature. We predict the feature for next time instant. A self-supervised signal based on the difference between the predicted and the observed feature gives rise to a possible event boundary.

To segment a video into its constituent *events*, we must first define the term *event*. Drawing from cognitive psychology [42], we define an event to be a "*segment of time at a given location that is perceived by an observer to have a beginning and an end*". Event segmentation is the process of identifying these beginnings and endings and their relations. Based on the level of distinction, the granularity of these events can be variable. For example, *throw ball* and *hit ball* can be events that constitute a larger, overarching event *play baseball*. Hence, each event can be characterized by a stable, internal representation that can be used to anticipate future visual features within the same event with high correlation, with increasing levels of error as the current transitions into the next. Self-supervised learning paradigms of "*predict, observe and learn*" can then be used to provide