Learning and Verification of Task Structure in Instructional Videos

Medhini Narasimhan[1,2], Licheng Yu[2], Sean Bell[2], Ning Zhang[2], Trevor Darrell[1]

[1]UC Berkeley, [2]Meta AI

https://medhini.github.io/task_structure

Abstract

Given the enormous number of instructional videos available online, learning a diverse array of multi-step task models from videos is an appealing goal. We introduce a new pre-trained video model, VideoTaskformer, focused on representing the semantics and structure of instructional videos. We pre-train VideoTaskformer using a simple and effective objective: predicting weakly supervised textual labels for steps that are randomly masked out from an instructional video (masked step modeling). Compared to prior work which learns step representations locally, our approach involves learning them globally, leveraging video of the entire surrounding task as context. From these learned representations, we can verify if an unseen video correctly executes a given task, as well as forecast which steps are likely to be taken after a given step. We introduce two new benchmarks for detecting mistakes in instructional videos, to verify if there is an anomalous step and if steps are executed in the right order. We also introduce a long-term forecasting benchmark, where the goal is to predict long-range future steps from a given step. Our method outperforms previous baselines on these tasks, and we believe the tasks will be a valuable way for the community to measure the quality of step representations. Additionally, we evaluate VideoTaskformer on 3 existing benchmarks—procedural activity recognition, step classification, and step forecasting—and demonstrate on each that our method outperforms existing baselines and achieves new state-of-the-art performance.

1. Introduction

Picture this, you're trying to build a bookshelf by watching a YouTube video with several intricate steps. You're annoyed by the need to repeatedly hit pause on the video and you're unsure if you have gotten all the steps right so far. Fortunately, you have an interactive assistant that can guide you through the task at your own pace, verifying each step as you perform it and interrupting you if you make a mistake. A composite task such as "*making a bookshelf*" involves multiple fine-grained activities such as "*drilling holes*" and "*adding support blocks*." Accurately categorizing these activities requires not only recognizing the individual steps that compose the task but also understanding the task structure, which includes the temporal ordering of the steps and multiple plausible ways of executing a step (e.g., one can beat eggs with a fork or a whisk). An ideal interactive assistant has both a high-level understanding of a broad range of tasks, as well as a low-level understanding of the intricate steps in the tasks, their temporal ordering, and the multiple ways of performing them.

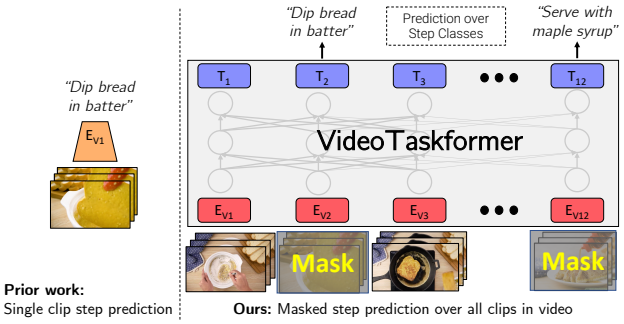

Figure 1: Prior work [13, 12] learns step representations from single short video clips, independent of the task, thus lacking knowledge of task structure. Our model, VideoTaskformer, learns step representations for masked video steps through the global context of all surrounding steps in the video, making our learned representations aware of task semantics and structure.

As seen in Fig. 1, prior work [12, 13] models step representations of a single step independent of the overall task context. This might not be the best strategy, given that steps for a task are related, and the way a step is situated in an overall task may contain important information about the step. To address this, we pre-train our model with a masked modeling objective that encourages the step representations to capture the *global context* of the entire video. Prior work lacks a benchmark for detecting mistakes in videos, which

*Work done while an intern at Meta AI. Correspondence to medhini@berkeley.edu