Learning Tasks for Instructional Videos: A Position Paper

Cheston Tan & Keng-Teck Ma
Institute for Infocomm Research and A*STAR Artificial Intelligence Initiative
1 Fusionopolis Way, Singapore 138632
{cheston-tan@i2r.a-star.edu.sg, makt@scei.a-star.edu.sg}

Abstract

This abstract serves as a form of brainstorming for various tasks (and potential Challenges) associated with learning from instructional videos. The hope is that this will serve as a basis for further discussion among the community.

1. Introduction

One of the goals of the FIVER workshop is to create a common Challenge, and another goal is to set the community research agenda for the next few years. Here, we brainstorm possibilities of tasks relating to learning from instructional videos, rather than propose one single task or challenge. In this nascent field, where the core research problems are still not clear, it is important to keep in mind the full range of possibilities first, before deciding on specific tasks to narrow down to.

2. Tasks relating to learning from instructional videos

We present a set of tasks relating to learning from instructional videos. For each task, we will briefly describe the task, describe a potential usage/application, and explain the underlying research challenge.

By "task", we refer to the computer vision or machine learning task to be performed on instructional videos, not to the task being demonstrated or carried out within the instructional video.

Note that we will exclude existing tasks which are less specific to instructional videos and could be generically applied to other kind of videos. For example, we will not talk about retrieving a video given a keyword, nor classifying a video into action categories. We will list the tasks categorically, organized from the simplest set of inputs (a single video) to the most complex.

Note that henceforth, for brevity, we use the term "video" to mean "instructional video".

2.1. Input is a single video

Task: Produce a set of instructions from the video. This is perhaps the most obvious task. Example applications of this task include: producing a recipe from an un-annotated cooking video, or producing a set of instructions that were demonstrated or instructed to a robot. The research challenges for this basic task include: ignoring of irrelevant content within the video, such as visual (e.g. wiping forehead), auditory (e.g. coughing) and linguistic content (e.g. side remarks), and also determining the appropriate level of granularity for the instructions. This includes determining which steps are too obvious (e.g. "put the bowl onto the table") and within a step, what should be said or not said (e.g. "crack two eggs" versus "crack two eggs into the smaller bowl" [because the bigger bowl is needed for something else]).

Task: Align the visual content to instructions in some other modality. If the video comes with an auditory or textual set of instructions, the task is to determine which part of the video corresponds to which individual instruction. This would allow viewers of the video to arbitrarily navigate between instruction steps (e.g. skip easy steps like "crack two eggs into a bowl"). The challenges here are similar to the previous task, and this task is in some sense a simpler version, and the same underlying algorithms could be applied to solve both tasks.

2.2. Input is a collection of videos

Task: Produce a set of instructions from the videos. This task is similar to the above task for a single video. However, with multiple videos, there are additional challenges. There may be multiple variations of doing a task, seen across the different videos. For example, even for something as simple as making a cup of coffee, the order of adding coffee powder, sugar and creamer is not strictly fixed. Moreover, some videos may show a person adding sugar, then tasting the coffee, then adding sugar again. How to produce a concise representation that properly captures all these variations is challenging.

Task: Choose a single "best" video (or rank all the videos), according to various definitions of goodness, e.g. well-explained instructions, clear visual demonstration, conciseness, amount of detail, etc. This is extremely useful, for example when doing an online search of "how to […]", and many videos are returned, and the searcher wants to know which video is the "best" one to watch. The research challenge is the element of subjectivity in these measures of goodness, and the fact that some of these are fairly novel research areas in themselves (e.g. "clarity of visual demonstration").

2.3. Input is one instruction set and a collection of videos

Task: Return a list of videos that are consistent with, or relevant to the given instructions (or conversely, eliminate videos that are inconsistent or irrelevant). One example application is to search for recipes given a set of ingredients and set of cooking techniques that a viewer knows or does not know (e.g. exclude videos that require poaching an egg or the *sous vide* method of cooking).

Task: Return a novel video, either spliced together from clips in the collection, or generated from scratch (e.g. using a GAN). The potential application is something relatively novel: automatic creation of instructional videos, given just a set of instructions, and some database of videos to learn or splice from. This would save enormous time and effort to perform and edit instructional videos. If the instructional steps are not novel themselves, but rather their composition and ordering, then the research challenge here is to generate a video that is consistent and smooth.

3. Conclusion

This abstract serves as a form of brainstorming for various tasks (and potential Challenges) associated with learning from instructional videos. The hope is that this will serve as a basis for further discussion among the community.