Procedures from Instructional Video: Multimodal Procedural Knowledge Learning

Junyi Du
University of Southern California
junyidu@usc.edu

Abstract

Procedural knowledge is the knowledge required to perform certain tasks, and forms an important part of expertise. A major source of procedural knowledge is natural language instructions, and there are some prior efforts on interpreting readable instructions in natural languages into structured data that can be processed by machines. However, since we are experiencing a multi-modal world, natural language is not the only modal for keeping procedural knowledge, as videos are widely used as another form of instruction for professional teaching. Unfortunately, few attention has been paid to learning procedural knowledge in a multi-modal manner. In this paper, we mainly survey about the previous researches on procedural knowledge learning from vision modality, and works related to video understanding. Our goal is to explore the potential of video information for procedural knowledge learning, and try to address where are the improvements of the multi-modal methods come from.

1 Introduction

Procedural knowledge is the knowledge required to perform certain tasks, which extensively used in our daily life, and forms an important part of expertise. Continuous efforts have been dedicated to acquisition of structured procedural knowledge with symbolic meanings. One popular direction is learning procedures from natural language instructions or procedural texts, which contains descriptions for performing certain tasks in human natural language (Webster et al., 2012; Park and Motahari Nezhad, 2018; Du et al., 2019).

However, our experience of the world is multimodal: we see objects, hear sounds, feel texture, smell odors. Natural language is not the only modality that we perceive. When learning a new task, nowadays's people do not only follow some textual instructions of the task's procedures, but also refer to pictures and videos about the task as guidance. For example, there are tons of instructional video on Internet surrounding topics like "How to cook" or "How to repair". Unfortunately, few attention has been paid to learning procedural knowledge from multiple modalities.

In this paper, we mainly survey about how the modality of vision could be help for the learning of procedural knowledge. We study previous researches on learning procedural knowledge from multiple modalities. It's also our interest that in what extent the video can improve the learning of procedural knowledge. We try to address that what are the procedural knowledge can be easily learned from video while it's hard to learn from linguistic modality, to find out where are the improvements of the multi-modal methods come from.

2 Related Works

Instructional videos provide an intuitive way for human to learn how to complete a task, e.g. how to cook a fish, how to repair an A/C. A large number of previous works aim to endow machines with the ability to learn from instructional videos, which provide rich visual, audio and textual information of actions for finishing a task. Some of these works share similar interest with us, that aim at extracting a series of actions or procedural information from video:

2.1 Video dense caption

Video dense caption (Krishna et al., 2017) identify events in a video while simultaneously describing the detected events with natural language. They are actually doing scene captioning in the order of scenes in the given video.

However, while we are looking for learning procedural knowledge, video dense caption mainly

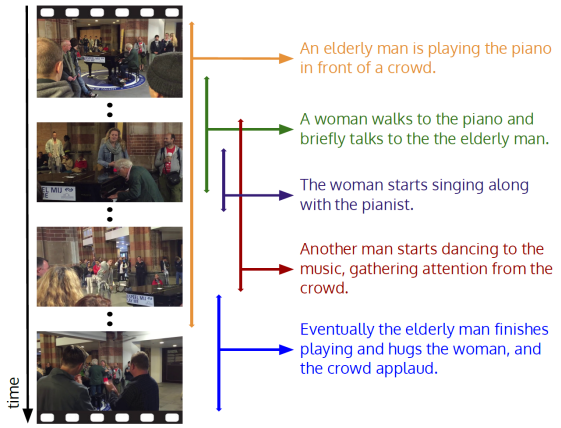

Figure 1: Example of video sample and its corresponding dense caption

focus on the events in the video. Their goal is to detect and extract these events, sentences in natural language describing actions, following the timeline of the video. And they only use vision modality, the video, as the input.

2.2 Visual + BERT

VideoBERT(Sun et al., 2019) is joint visual-linguistic model to learn high-level features. It shows visual features extracted from video segments can be injected into the prevailing pre-trained NLP model, BERT, as video tokens, then jointly fine-tune with text token in a masked training manner. There are similar ideas exist, like VisualBERT(Li et al., 2019), VL-BERT(Su et al., 2019), VIL-BERT(Lu et al., 2019) which also try to inject vision modality into the powerful BERT model to enable multi-modal representation learning. Among these competing models, VideoBERT should be highlighted since it is sharing our interest that it proposed a framework for aligning procedural steps in natural language with visual information in instructional video.

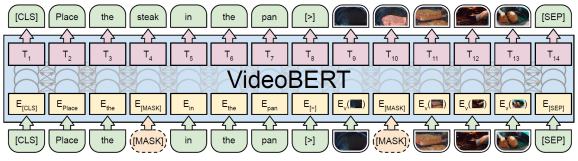

Figure 2: Illustration of VideoBERT in the context of a video and text masked token prediction

However, even it's sharing the same instructional video domain with us, VideoBERT focus on the pre-training strategy and representation. It's not extracting any structured procedural knowledge that we want. Then it comes to the question that how to exploit the abundant multi-modal representation to aid our research on learning of procedural knowledge.

There are also some interesting tasks for images or non-instructional videos that draw our attention:

2.3 Visual Question Answering

Visual Question Answering (VQA) is a widely explored problem in computer vision (Goyal et al., 2017). It's one of the most famous problem that crossing vision and linguistic modality. As the improvement of video modeling, some works start to focus on video question answering (Lei et al., 2018). In many video question answering tasks, models are required to have a deep understanding or even reasoning of information from the video, including procedural information.

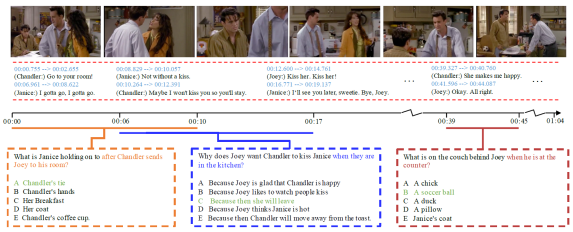

Figure 3: Examples of a video question answereing dataset (TVQA)

2.4 Action Anticipation

A problem with raising attention in video domain is to predict "what's next" for a given video. Liang et al. predict a pedestrian's future path jointly with future activities. Duarte et al. try to predict actions of human and investigate how the different cues contribute to of human actions. Although it's not linguistic modality related, the prediction of actions also require ability to modeling changes in future from an action level.

References

Junyi Du, He Jiang, Jiaming Shen, and Xiang Ren. 2019. Eliciting knowledge from experts: Automatic transcript parsing for cognitive task analysis. *CoRR*, abs/1906.11384.

Nuno Ferreira Duarte, Mirko Rakovic, Jovica Tasevski, Moreno Ignazio Coco, Aude Billard, and Jose Santos-Victor. 2018. Action anticipation: Reading

the intentions of humans and robots. *IEEE Robotics and Automation Letters*, 3(4):41324139.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. Tvqa: Localized, compositional video question answering. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander Hauptmann, and Li Fei-Fei. 2019. Peeking into the future: Predicting future person activities and locations in videos.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Hogun Park and Hamid Reza Motahari Nezhad. 2018. Learning procedures from text: Codifying how-to procedures in deep neural networks. In *Companion Proceedings of the The Web Conference 2018*, pages 351–358. International World Wide Web Conferences Steering Committee.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766.

Philip Webster, Victoria Uren, Ziqi Zhang, Fabio Ciravegna, and Andrea VARGA. 2012. Automatically extracting procedural knowledge from instructional texts using natural language processing.