Review 7 - 1

Unsupervised Learning from Narrated Instruction Videos

Summary

This paper, presents a method to learn the main steps of a task from a set of videos with textual description in an unsupervised way. The main difficulties in this task are from high variability across languages and video. To address these challenges, this paper develops an unsupervised learning approach that takes advantage of the complementarity of the visual signal in the video and the corresponding natural language narration to resolve their ambiguities. Firstly, this paper formulates the problem as two clustering tasks, one in text and the other in video. The model learns the relationships among main logical objects and uses them to conduct the text clusters for different mains step of the tasks. Given the output of the text clustering that identified the important steps forming a task, the model uses a discriminative clustering approach with global ordering constraints to learn the video clusters. The main idea behind discriminative clustering is to find a clustering of the data that can be easily recovered by a linear classifier through the minimization of an appropriate cost function over the assignment matrix. In the experiment section, this paper demonstrates the proposed approach has been tested on a new annotated dataset of challenging real-world instruction videos containing complex person-object interactions in a variety of indoor and outdoor scenes.

Main Contributions:

1. Develop a new unsupervised learning approach that takes advantage of the complementary nature of the input video and the associated narration. The method solves two clustering problems, one in text and one in video, applied one after each other and linked by joint constraints to obtain a single coherent sequence of steps in both modalities.
2. Collect and annotate a new challenging dataset of real-world instruction videos. The dataset contains about 800000 frames for five different tasks.
3. Conduct experiments and demonstrate that unsupervised method can discover the main steps to achieve the task and locate the intervals in the videos.
4. Provide a new chance for large scale learning from instruction videos on the Internet.

Review 7 - 2

Predicting Motivations of Actions by Leveraging Text

Summary

The key problem this paper aims to tackle is predicting why a person has performed an action in images instead of simply recognizing actions. To study this problem, this paper first assembles an image dataset of people (about 10, 000 people) and annotated them with their actions, motivations, and scene. Then this paper demonstrates visual features alone may not be sufficient prediction the intention behind the actions very well. To combine the textual knowledge with image features, this paper then extracts common-sense from text. The idea is to create a factor graph over several concepts (actions, motivations, and scenes). The unary potentials come from visual classifiers, and the potentials for the relationships between concepts can be estimated from large amounts of text. Hence, the model converts this problem into a "fill in the blanks" problem and leverages pre-trained language model to infer the motivation with highest probability along with the given concept.

Main Contributions:

1. Propose a new problem of predicting the motivations behind actions.
2. Provide a dataset with images, people annotation, actions and motivations.
3. Provide a basic framework to learn the motivations from pre-trained text models.
4. Experimentally demonstrate only using images features failed to tackle the problem very well.

Cons:

1. This model can learn feature from image-text embedding feature space.
2. Doesn't provide quantified results.