David Park

CS 4395.001

Summary of *End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding*

In *End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding* by Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, Shiliang Pu, Fei Wu and affiliated with Zhejiang University, Shanghai Institute for Advanced Study of Zhejiang University, Shanghai AI Laboratory, Hikvision Research Institute, Northwestern Polytechnical University, the authors explore the idea of one-shot natural language spatial video grounding. This idea involves identifying and localizing objects in a video based on a single natural language sentence that describes the object's appearance and location.

There has been prior work done on natural language video grounding. In general there are three main branches, temporal grounding, spatio-temporal grounding, and spatial grounding, which the paper focuses on. Both deep neural networks and weakly-supervised learning, using multiple instances learning, have been used for this. Multiple methods for one-shot learning have been researched. For example, one method proposes a meta-learning-based approach to perform one-shot action localization by capturing task-specific prior knowledge. However, one-shot natural language spatial video grounding follows the setting from a fully convolutional neural network architecture to solve the segmentation tasks.

To the best of their knowledge, the authors are the initiators within the idea of one-shot natural language spatial video grounding. Some existing approaches for similar tasks have limitations, such as requiring multiple training examples, which makes the task more difficult. One-shot natural language spatial video grounding uses an information tree to capture the relationships between objects and their attributes in both the video and the sentence, thus, overcoming the limitations that similar approaches have.

The authors evaluated their work with two video grounding benchmarks, *VidSTG* and *VID-sentence*. These benchmarks are widely used video grounding benchmarks based on large datasets. The authors computed the Intersection over Union (IoU) for the predicted spatial bounding box and the ground-truth per frame. The prediction is considered as accurate if the IoU average of all the frames exceeds a certain threshold of alpha which the authors have set to 0.4, 0.5, and 0.6 during training. The baselines used for evaluation are an extension of fully supervised models like OMRN, and other widely known methods like STPR. IT-OS consistently achieved the best performance on two benchmarks and multiple experimental settings .They then compare one-shot IT-OS with fully supervised methods when training multiple baselines and IT-OS with all labels on the VID-sentence dataset. This comparison between IT-OS and the fully supervised OMRN is less than 4%.

One of the authors, Fei Wu, has had over 17,000 citations throughout his academic career. I think this author's work is important because being able to identify and localize objects in a video with limited natural language descriptions will most likely translate well out of a testing environment and into a "real-world' scenario. Natural

language sentences describing a frame or video may be incomplete or sparse, thus, having a need for one-shot natural language spatial video grounding.