

# How Local is the Local Diversity? Reinforcing Sequential Determinantal Point Processes with Dynamic Ground Sets for Supervised Video Summarization

Yandong Li<sup>1</sup>0000000320051334, Liqiang Wang<sup>1</sup>0000000212654656, Tianbao Yang<sup>2</sup>0000000278585438, and Boqing Gong<sup>3</sup>0000000339155977

<sup>1</sup> University of Central Florida, Orlando, FL

<sup>2</sup> University of Iowa, Iowa City, Iowa

<sup>3</sup> Tencent AI Lab, Seattle, WA

**Abstract.** The large volume of video content and high viewing frequency demand automatic video summarization algorithms, of which a key property is the capability of modeling diversity. If videos are lengthy like hours-long egocentric videos, it is necessary to track the temporal structures of the videos and enforce local diversity. The local diversity refers to that the shots selected from a short time duration are diverse but visually similar shots are allowed to co-exist in the summary if they appear far apart in the video. In this paper, we propose a novel probabilistic model, built upon SeqDPP, to dynamically control the time span of a video segment upon which the local diversity is imposed. In particular, we enable SeqDPP to learn to automatically infer *how local the local diversity is supposed to be* from the input video. The resulting model is extremely involved to train by the hallmark maximum likelihood estimation (MLE), which further suffers from the exposure bias and non-differentiable evaluation metrics. To tackle these problems, we instead devise a reinforcement learning algorithm for training the proposed model. Extensive experiments verify the advantages of our model and the new learning algorithm over MLE-based methods.

## 1 Introduction

The Internet age has come to such a new phase that high-definition videos are both ubiquitous and dominant in the IP traffic featured by the boom of video sharing websites, online movies and television shows, and the emerging live video streaming services. Some statistics indicate that about 300 hours of video are uploaded to YouTube per minute and more than 500 million hours of video are watched on YouTube daily. Such a large volume of video content and high viewing frequency demand automatic video summarization algorithms. By distilling important events from the original video and condensing them to a short video clip (or a story board, text description, etc.), video summarization has a great potential in many real-world applications.

Video summarization has been one of the basic research areas in the fields of computer vision and multimedia for decades [2]. A variety of techniques have been proposed for different scenarios of video summarization. In general, a good video summary is supposed to describe main events [3–5] happened in the video and meanwhile remove the video shots that are redundant [6, 7] and/or unimportant [8, 9].



































