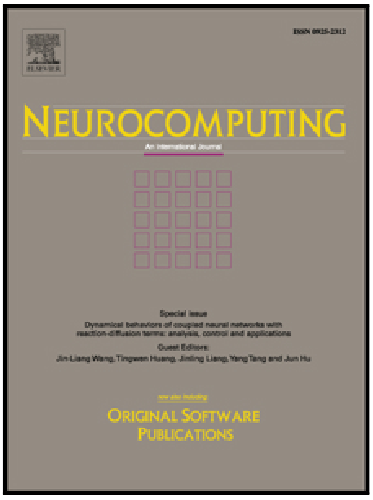
Communicated by Dr. Yu Jiang


ISSN 0925-2312
ELSEVIER
NEUROCOMPUTING
An International Journal
Special issue
Dynamical behaviors of coupled neural networks with reaction-diffusion terms: analysis, control and applications
Guest Editors:
Jin-Liang Wang, Tingwen Huang, Jinling Liang, Yang Tang and Jun Hu
now also including
ORIGINAL SOFTWARE PUBLICATIONS

The Forgettable-Watcher Model for Video Question Answering

Wenqing Chu, Hongyang Xue, Zhou Zhao, Deng Cai, Chengwei Yao

PII: S0925-2312(18)30807-5
DOI: 10.1016/j.neucom.2018.06.069
Reference: NEUCOM 19747

To appear in: *Neurocomputing*

Received date: 4 February 2018
Revised date: 6 May 2018
Accepted date: 25 June 2018

Please cite this article as: Wenqing Chu, Hongyang Xue, Zhou Zhao, Deng Cai, Chengwei Yao, The Forgettable-Watcher Model for Video Question Answering, *Neurocomputing* (2018), doi: 10.1016/j.neucom.2018.06.069

The Forgettable-Watcher Model for Video Question Answering

Wenqing Chu[a1], Hongyang Xue[a1], Zhou Zhao[b], Deng Cai[a], Chengwei Yao[b*]

[a]*The State Key Lab of CAD & CG, College of Computer Science and Technology Zhejiang University, China*
[b]*College of Computer Science, Zhejiang University, China*

Abstract

A number of visual question answering approaches have been proposed recently, aiming at understanding the visual scenes by answering the natural language questions. While the image question answering has drawn significant attention, video question answering is largely unexplored. Video-QA is different from Image-QA since the information and the events are scattered among multiple frames. In order to better utilize the temporal structure of the videos and the phrasal structures of the answers, we propose two mechanisms: the re-watching and the re-reading mechanisms and combine them into the forgettable-watcher model. Then we propose a TGIF-QA dataset for video question answering with the help of automatic question generation. Finally, we evaluate the models on our dataset. The experimental results show the effectiveness of our proposed models.

Keywords: Video analysis, Video question answering, Attention model

1. Introduction

Understanding the visual scenes is one of the ultimate goals in computer vision. A lot of intermediate and low-level tasks, such as object detection [1, 2], recognition [3, 4], segmentation [5, 6] and tracking [7, 8] have been studied towards this goal. One of the high-level tasks towards scene understanding is the visual question answering [9]. This task aims at understanding the scenes by answering the questions about the visual data. It also has a wide application, from aiding the visually-impaired, analyzing surveillance data to domestic robots [10, 11, 12].

The visual data we are facing everyday are mostly dynamic videos. However, most of the current visual question answering works only focus on images [13, 14, 15, 16, 17, 18]. The images are static and contain far less information than the videos. The task of image-based question answering cannot fit into real-world applications well since it ignores the temporal coherence of the scenes.

Existing video-related question answering works usually make use of additional information. The Movie-QA dataset [19] contains multiple sources of information: plots, subtitles, video clips, scripts and DVS transcriptions. These extra information is hard to retrieve in the real world, making it difficult to extend these approaches to general videos.

Unlike the previous works, we consider the more ubiquitous task of video question answering with only the visual data and the natural language questions. In our task, only the videos, the questions and the corresponding answer choices are presented. We first introduce a dataset collected on our own. Collecting a dataset is not an easy task. In image-based question answering (Visual-QA) [9], most current collection methods require humans to generate the question-answer pairs [9, 20]. This requires a significant amount of human labor. In addition, the video data has a temporal dimension compared with the image, which implies that the labor of the human annotators is multiplied. To avoid the significant increase of human labor, our solution is to employ the question generation approaches [21] to generate question-answer pairs directly from the texts accompanying the videos. Now the collection becomes collecting videos with descriptions. This inspires us to utilize the existing video description datasets. The TGIF (Tumblr GIF) dataset [22] is a large-scale video description dataset. The

*Corresponding author
Email addresses: wqchu16@gmail.com (Wenqing Chu[a]), hyxue@outlook.com (Hongyang Xue[a]), zhaozhou@zju.edu.cn (Zhou Zhao[b]), dengcai@cad.zju.edu.cn (Deng Cai[a]), yaochw@zju.edu.cn (Chengwei Yao[b])

[1] The first two authors contributed equally.