FINAL ORAL EXAMINATION

OF

Amir Mazaheri
M.Sc., University of Central Florida, 2016
B.S., Sharif University of Technology, 2013

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)

25 March, 2020, 10:00 A.M.
https://meet.ucf.edu/ctran/FQRPC6KK

DISSERTATION COMMITTEE
Professor Mubarak Shah, Chair, shah@crcv.ucf.edu
Professor Peter Hancock, peter.hancock@ucf.edu
Professor Lotzi Bölöni, lboloni@cs.ucf.edu
Professor Fei Liu, feiliu@cs.ucf.edu
DISSEMINATION RESEARCH IMPACT

Video data is explosively growing due to ubiquitous video acquisition devices. Recent years witness a surge of substantial video volumes from surveillance, health care, and personal mobile phones, to name a few. Meanwhile, thanks to social media and on-demand video streaming industry, videos and corresponding textual descriptions are being produced and stored every moment, continuously. This amount of video + text “big data” makes today the best time ever for Computer Vision and Machine Learning (ML) to introduce and solve tasks related to a common understanding of videos and text. Furthermore, a practical analysis of such amount of video and text data is impossible for humans. Text and video joint-understanding approaches are solutions to various real-world use-cases, like social media analysis, video search engines, etc. These practical solutions change the way that we utilize the available data and have an impact on the industry, social security, and the future research tracks.

This dissertation makes contributions to the above tasks by proposing: (1) A novel framework for multi-concept video retrieval model which utilizes inter and intra-shots correlations (2) A Spatio-temporal attention model to solve a novel form of Visual Question Answering (3) A new research problem, Visual Text Correction, to detect and correct inaccuracies in video descriptions (4) A Generative model to generate videos from natural language sentences in wild datasets by constructing a latent path.

SELECTED PUBLICATIONS


DISSENYATION
VIDEO CONTENT UNDERSTANDING USING TEXT

The joint understanding of videos and text is the backbone of many real-world use-cases that is needed to exploit the knowledge captured in today's massive video data. However, understanding of text and videos could be very challenging, since they have fundamentally different characteristics. In this dissertation, we propose methods that are essential to the design of many practical solutions for real-world problems in the context of a joint understanding of video and text.

First, we address the problem of Video Retrieval and propose a novel approach, employing a latent ranking SVM, which integrates the advantages of various recent works in text and image retrieval, such as choosing ranking over structured prediction and modeling interdependencies between querying concepts and so on. Our model captures the positive and negative correlation among the shots, and we train the whole system using a ranking-SVM objective function. We also introduce a simple and effective technique to make our model robust to outliers, since the concept labels of shots in the training are scarce and noisy.

While the video retrieval is about finding videos related to a query (set of conceptual words), it does not fully exploit knowledge contained in both video and text. Next, we study a novel form of Visual Question Answering (VQA), Video-Fill-In-the-Blank (VFIB). Given a video and a description sentence with one missing word, VFIB problem is to find the missing word automatically. We propose a framework to solve the FIB problem, which consists of text and visual encoding. We employ a novel text encoding module, "lr/rl LSTMs," that first encodes the left and right sentence fragments separately and also combines each fragment with an external memory corresponding to the opposite fragment. For the visual encoding, we employ spatial and temporal attention models to select discriminative visual representations to find the missing word. Our method won second place in Large Scale Movie Description Challenge-Fill In the Blank (LSMDC-FIB).

Related to the above, we introduce a new problem called Visual Text Correction (VTC), i.e., finding and replacing an inaccurate word in the textual description of a video. We propose a deep network that can simultaneously detect an inaccuracy in a sentence, and fix it by replacing the inaccurate word(s). Our method leverages the short and long term dependencies between words of a video caption and also semantic interdependence of video and the words in the caption. Our proposed formulation can solve the VTC problem employing a novel network in two steps: (1) inaccuracy detection, and (2) correct word prediction. Our experiments and performance analysis demonstrate that the proposed method provides excellent results and also highlights the challenges in solving the VTC problem. To the best of our knowledge, this work is the first of its kind for the Text Correction Task.

So far, we have processed videos and text (sentences or set of concept words) to solve video retrieval, video fill in the blanks, and video text correction. Next, we propose to synthesize the content of a video directly using sentences. We tackle the complicated problem of Video Generation by regressing the first and last frames' latent representations and employing a context-aware interpolation method to build up the latent representations of in-between frames. We propose a stacking "upPooling" block to sequentially generate RGB frames out of each latent vector and progressively increase the resolution. Moreover, our proposed Discriminator encodes videos based on single and multiple frames. We provide quantitative and qualitative results to support our arguments and show the superiority of our method over well-known baselines like RNNs.
Amir Mazaheri

1990  Born in Tehran, Iran
2008  B.S., Sharif University of Technology, Tehran, Iran
2016  M.Sc., University of Central Florida, Orlando, FL
2018  Research Scientist Intern, Netflix, Los Gatos, CA
2013-2020  Ph.D., University of Central Florida, Orlando, FL