Center for Research in Computer Vision

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Aisha Urooj Khan

M.S., Lahore University of Management Sciences, 2013
B.Eng., NED University of Engineering & Technology, 2009

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

14 November 2022, 11:00 A.M.
Trevor Colbourn Hall 351

DISSERTATION COMMITTEE

Professor Mubarak Shah, *Chair*, shah@crcv.ucf.edu
Professor Niels Lobo, *Co-Chair*, nlobo@ucf.edu
Professor Qian Lou, qian.lou@ucf.edu
Professor Teng Zhang, teng.zhang@ucf.edu

DISSERTATION RESEARCH IMPACT

Given an image or a video, visual question answering (VQA) deals with a challenging task of answering a question related to the contents of the visual input. VQA has practical applications such as assisting people with visual impairments and helping radiologists for early diagnosis of fatal diseases. While VQA systems are increasingly finding real-world applications, there is a compelling need to provide these systems capabilities to explain their decisions. Such capabilities are imperative to improve a system's reliability and trustworthiness. *Attention* is a mechanism used by VQA methods to link the text (question and answer) to specific visual regions, also referred to as grounding. Thus, VQA grounding is a means of verifying that correct visual content is being inspected to assess how the answer was determined. With its significance for critical applications, VQA also serves as a foundation of further research areas such as embodied AI, language-guided navigation, and visual dialogue.

This dissertation makes several contributions in the field of VQA and grounding by presenting: 1) new algorithm for multimodal question answering that processes each input modality individually and collectively as needed allowing to overcome language biases, hence improving performance for truly vision-based questions; 2) a mechanism to measure the reliability of VQA methods by verifying that the correct visual information is being inspected; 3) techniques to improve the interpretability of such methods in two types of neural networks: CNNs and transformers; 4) an efficient approach to learn a compact video representation, i.e., its underlying spatiotemporal scene-graph and utilize it to solve video question answering.

SELECTED PUBLICATIONS (h-index: 5, total citation: 150)

1. **Analysis of Hand Segmentation in the Wild**, A Urooj and A Borji, in *IEEE CVPR* 2018.
2. **Segmenting Sky Pixels in Images,** C L Place*, A Urooj*, and A Borji, in *IEEE WACV* 2019.
3. **MMFT-BERT: MultiModal Fusion Transformer with BERT Encodings for Visual Question Answering,** A Urooj, A Mazaheri, N Lobo, and M Shah, in *EMNLP* 2020.
4. **Found a Reason for me? Weakly-supervised Grounded Visual Question Answering using Capsules,** A Urooj, H Kuehne, K Duarte, C Gan, N Lobo and M Shah, in *IEEE CVPR*, 2021.
5. **Weakly Supervised Grounding for VQA in Vision-Language Transformers**, A Urooj, H Kuehne, C Gan, N Lobo and M Shah, in *ECCV*, 2022. *Oral.*
6. **Learning to Predict Situation Hyper-Graphs for Video Question Answering,** A Urooj, H Kuehne, B Wu, C Gan, N Lobo, and M Shah, *submitted*.

DISSERTATION

VISUAL QUESTION ANSWERING: TRADE-OFFS BETWEEN TASK ACCURACY AND EXPLAINABILITY

Understanding data from multiple modalities such as vision, language, and sound requires extensive research. Visual question answering (VQA) is one such task which requires answering questions about complex situations in images or videos. In this dissertation, we aim to increase the capability of AI systems by developing algorithms that extract knowledge from copious amounts of visual (images, videos) and text (question, subtitles, captions) inputs with the focus on question-answering.

First, we explore the VQA task in a multi-modal setup of video and subtitles. We propose to solve VQA while ensuring individual and combined processing of multiple input modalities. Our approach named MMFT-BERT benefits from processing multimodal data (video and text) using a process which adopts the BERT encodings individually and uses a novel transformer-based fusion method to fuse them together. We show that MMFT-BERT can improve VQA accuracy particularly for vision-only questions.

Despite high VQA accuracies, these systems are struggling to ground text in the visual content. Therefore, in the second part of this dissertation, we focus on a weakly-supervised grounding setting: the grounding of relevant visual entities by training on the VQA task alone without any object-level supervision. To tackle this problem, we propose a visual capsule module with a query-based selection mechanism of capsule features, that allows the model to focus on relevant regions based on the textual cues about visual information in the question. We show that integrating the proposed capsule module in VQA systems significantly improves their performance on grounding.

Up to this point, our approach solves grounding in CNN-based VQA methods. Recently, visual-language transformers have shown tremendous performance on visual question answering (VQA) and grounding. However, most of those systems still rely on pre-trained object detectors during training, which limits their applicability. To mitigate this limitation, in the third part of this dissertation, we combine the ideas of attention and capsules in a capsule-transformer architecture. Our approach leverages capsules by transforming each visual token into a capsule representation in the visual encoder; it then uses activations from language self-attention layers as a text-guided selection module to mask those capsules before they are forwarded to the next layer. We show that the integration of capsules continues to benefit the grounding ability in transformers as well.

Finally, now that we have studied grounding in images, in the fourth part of this dissertation, we return to our original goal, video question answering. We propose to solve the VQA task in videos and develop a method which identifies the underlying activities and scene structure throughout the video, hence empowering the proposed network by learning strong spatiotemporal representations required to correctly answer the complex reasoning-based questions.

Aisha Urooj Khan

1986	Born in Karachi, Pakistan
2005-2009	B.Eng., NED UET, Karachi, Pakistan
2011-2013	M.S., Lahore University of Management Sciences, Pakistan
2020	Research Scientist Intern, MIT-IBM Watson AI Lab
2021	Research Scientist Intern, MIT-IBM Watson AI Lab
2016-2022	Ph.D., University of Central Florida, Orlando, Florida

Selected Awards

2022	ECCV Travel Award
2019	NeurIPS Travel Award
2018	Grace Hopper Celebration Travel Award

Selected Talks

2022	*Weak Grounding in Vision-Language Transformers*. **ECCV Workshop on Compositional and Multimodal Perception**.
2022	*Weakly Supervised VQA Grounding*, **Google AI Research**, Mountain View, CA.
2021	*VQA and Answer Grounding*, **Meta AI Research** (FAIR), Montreal, CA.