Center for Research in Computer Vision

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Madeline Chantry Schiappa

B.S., University of Connecticut, 2010
M.S., University of Central Florida, 2018

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

27 June 2023, 1:00 P.M.
HEC 356

DISSERTATION COMMITTEE

Dr. Yogesh Singh Rawat, *Chair*, *yogesh@crcv.ucf.edu*
Dr. Mubarak Shah, *shah@crcv.ucf.edu*
Dr. Shibu Yooseph, *shibu.yooseph@crcv.ucf.edu*
Dr. Joseph Schmidt, *joseph.schmidt@ucf.edu*

DISSERTATION RESEARCH IMPACT

The primary objective of this research has been to explore and develop effective methodologies for assessing the robustness of computer vision models. Through rigorous experimentation and analysis, we have uncovered valuable insights into the strengths and weaknesses of various computer vision, visual-language, and video-language deep learning models. By investigating the vulnerabilities and limitations of these models, this work aims to contribute to the development of more reliable and trustworthy systems. The knowledge gained from this work can guide future investigations, inspire further research, and facilitate the responsible integration of computer vision technologies into society

SELECTED PUBLICATIONS

Schiappa, Madeline C., and Yogesh S. Rawat. "**SVGraph: Learning Semantic Graphs from Instructional Videos.**". IEEE Eighth International Conference on Multimedia Big Data (BigMM). IEEE, 2022

Schiappa, Madeline Chantry, Shruti Vyas, Hamid Palangi, Yogesh Rawat, and Vibhav Vineet. "**Robustness Analysis of Video-Language Models Against Visual and Language Perturbations.**" Advances in Neural Information Processing Systems Datasets and Benchmarks Track 35 (2022): 34405-34420.

Schiappa, Madeline C., Yogesh S. Rawat, and Mubarak Shah. "**Self-supervised learning for videos: A survey.**" ACM Computing Surveys (2022).

Schiappa, Madeline, Biyani, Naman, Kamtam, Prudvi, Vyas, Shruti, Palangi, Hamid, Vineet, Vibhav, and Rawat, Yogesh. "**A Large-scale Robustness Analysis of Video Action Recognition Models"**. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

DISSERTATION

A Study On Robustness And Semantic Understanding Of Visual Models

Vision models have improved in popularity and performance on many tasks since the emergence of large-scale datasets, improved access to computational resources, and new model architectures like the transformer. However, it is still not well understood if these models can be deployed in the real world. Because these models are “blackbox” architectures, we do not fully understand what these models are truly learning. An understanding of what models learn “underneath the hood"' would result in better improvements for real-world scenarios. Motivated by this, we benchmark these impressive visual models using newly proposed datasets and tasks on their robustness and their general understanding, using semantics as both a probe and an area of improvement.

In Chapter 1 we focus on generating graphical representations of noisy, instructional videos for video understanding by proposing a new video understanding task. We propose a self-supervised, interpretable approach that does not require any annotations for graphical representations, which would be expensive and time consuming to collect. We attempt to overcome ``black box" learning limitations by presenting Semantic Video Graph or SVGraph, a multi-modal approach that utilizes narrations for semantic interpretability of the learned graphs. SVGraph 1) relies on the agreement between multiple modalities to learn a unified graphical structure with the help of cross-modal attention and 2) assigns semantic interpretation with the help of Semantic-Assignment, which captures the semantics from video narration. We perform experiments on multiple datasets and demonstrate the interpretability of SVGraph in semantic graph learning.

In Chapter 2, we conduct a large-scale benchmark evaluation of the top action recognition models. We propose four different benchmark datasets, HMDB-51P, UCF-101P, Kinetics-400P, and SSv2P to perform this analysis and study the robustness of six different state-of-the-art action recognition models against 90 different perturbations. The study reveals some interesting findings, 1) transformer based models are consistently more robust against most of the perturbations when compared with CNN based models, 2) pretraining helps Transformer based models to be more robust to different perturbations than CNN based models, and 3) All of the studied models are robust to temporal perturbation on the Kinetics dataset, but not on SSv2; this suggests temporal information is much more important for action recognition on SSv2 datasets than on the Kinetics dataset. Next, we study the role of augmentations in model robustness and present a real-world dataset, UCF-101-DS, which contains realistic distribution shifts, to further validate some of these findings.

Because sequential data, like video and language, are natural forms of input to any intelligent vision system operating in the real world, we additionally benchmark video-language models. In Chapter 2, we perform the first extensive robustness study of video-language models against various real-world perturbations. We focus on text-to-video retrieval and propose two large-scale benchmark datasets, MSRVTT-P and YouCook2-P, which utilize 90 different visual and 35 different text perturbations. The study reveals some interesting initial findings from the studied models: 1) models are more robust when text is perturbed versus when video is perturbed, 2) models that are pretrained are more robust than those trained from scratch, 3) models attend more to scene and objects rather than motion and action.

Based on the results of the previous chapters, we found there was a lack of overall ``understanding" between object relationships. Inspired by this, in Chapter 5 we present a novel framework for probing large visual-language models (V+L) on three aspects of content understanding: object-object relations, attribute-object relations and context-object relations. These probes are grounded in cognitive science and help determine if a V+L model can, for example, determine if “snow garnished with a man” is implausible, or if it can identify beach furniture by knowing it is located on a beach. We have experimented with 5 well known models, such as CLIP and ViLT and found that they mostly fail to demonstrate a conceptual understanding. That said, we find interesting insights such as (1) cross-attention helps learning conceptual understanding most when simultaneously used with modality-specific learning, (2) models rely heavily on object recognition (3) but use co-occurrence and background information for some types of objects over others. We further utilize some of these insights and investigate a simple finetuning technique that rewards the three conceptual understanding measures with promising initial results.

We hope that this research will help the community assess and improve the robustness and semantic understanding of visual models.

Madeline Chantry (previously Schiappa)

2010	International Baccalaureate Diploma, Interlake High School, Bellevue, WA
2014	B.A., University of Connecticut, Psychology, Storrs, CT
2014	B.S., University of Connecticut, Intl' Business, Storrs, CT
2015-2018	Data Scientist, Sophos, Burlington MA
2018	M.S., University of Central Florida, Data Analytics, Orlando, FL
2022 (Spring)	Graduate Teaching Assistant
2019-2023	Graduate Research Assistant

SELECTED AWARDS & HONORS

2010-2012	NFCA All-American Scholar-Athlete, University of Connecticut, Storrs, CT
2010-2013	Big East All-Academic Team, University of Connecticut, Storrs, CT
2018-2019	OCR Fellowship, University of Central Florida, Orlando, FL
2019	Panel speaker at Women in Data Science Conference in Orlando, FL
2019	UCF Graduate Presentation Fellowship, (SNAMS)
2019	UCF Graduate Presentation Fellowship, (IC^2S^2)
2022	NeurIPs 2022 Scholar Award, (NeurIPs)
2022	UCF Graduate Presentation Fellowship, (NeurIPs)