##### The Department of Electrical and Computer Engineering

Announces the

## Final Defense of Dissertation

### **Niluthpol Chowdhury Mithun**

*Doctor of Philosophy, Graduate Program in Electrical Engineering  
University of California, Riverside*

#### Date: May 30, 2019

Time: 9:35 am

Location: Winston Chung Hall, Room 202

Title: Learning Robust Visual-Semantic Retrieval Models with Limited Supervision

Abstract: In recent years, tremendous success has been achieved in many computer vision tasks using deep learning models trained on large hand-labeled image datasets. In many applications, this may be impractical or infeasible, either because of the non-availability of large datasets or the amount of time and resource needed for labeling. In this respect, an increasingly important problem in the field of computer vision is how to learn useful models for tasks where labeled data is sparse. In this thesis, we focus on learning comprehensive joint representations for different cross-modal visual-textual retrieval tasks leveraging weak supervision, that is noisier and/or less precise but cheaper and/or more efficient to collect. Cross-modal visual-textual retrieval has gained considerable momentum in recent years due to the promise of deep neural network models in learning robust aligned representations across modalities. However, the difficulty in collecting aligned pairs of visual data and natural language description and limited availability such pairs in datasets makes it extremely difficult to train effective models, which would generalize well to uncontrolled scenarios as they are heavily reliant on large volumes of training data that closely mimic what is expected in the test cases. In this regard, we first present our work on developing a multi-faceted joint embedding framework-based video to text retrieval system that utilizes multi-modal cues (e.g., objects, action, place, sound) from videos to reduce the effect of limited data. Then, we describe our approach on training text to video moment retrieval systems leveraging only video-level text descriptions without any temporal boundary annotations. Next, we present our work on learning powerful joint representations of images and text from small fully annotated datasets with supervision from weakly-annotated web images. Extensive experimentation on different benchmark datasets demonstrates that our approaches show substantially better performance compared to baselines and state-of-the-art alternatives.