Department of Computer Science
St. Francis Xavier University

M.Sc. Thesis Proposal Presentation

Weakly Supervised Temporal Action Localization via Uncertainty-Aware Contextual Modeling

Presented by
Moayadeldin Hussain

Date: Monday, June 9th, 2025
Time: 10:00 AM
Location: Annex 113

Temporal **A**ction **L**ocalization (TAL) is one of the foundational tasks of video understanding, which aims to localize the start and end timestamps of action instances and identify their categories in untrimmed videos. TAL has recently attracted significant attention from both academia and industry, due to its potential for summarization, surveillance, visual question answering, and video retrieval. Many of the existing methods perform this task in a fully supervised framework. However, they require extensive manual frame-level annotations, which limits their scalability in real life scenarios. To overcome this problem, **W**eakly **S**upervised **T**emporal **A**ction **L**ocalization (WS-TAL) methods have been proposed, which only need to access video-level labels.

Videos contain frames of the action of interest and background frames. The goal of WS-TAL is to differentiate between these two categories in the absence of instance-level annotations. This task is challenging due to the existence of background noise, where background frames are semantically diverse and inconsistent, and due to action-context confusion, with certain frames are contextually related to the action of interest but are not part of it. As a result, another concern was raised regarding WS-TAL frameworks which is Action Localization Completeness, which means the models are inclined to generate over-complete or incomplete action boundaries due to imprecise predictions.

We argue that such challenges occur because current WSTAL solutions need to get better at incorporating the whole contextual information of training videos, and efficient incorporation of contextual information allows the model to grasp what snippets, form the whole action of interest, even if they differ in their visual and motion representation, and the subtle differences distinguishing different categories. Furthermore, it is crucial to quantify the model's uncertainty in its own predictions, as the predictions may vary with the model lacking enough knowledge about the optimal parameters, or different action classes, especially those sharing similar characteristics.