Week 7

07/09/2025

Multi-modal Large Language Models for Video Anomaly Detection

Lead : Dr. Tian

Eric Li

Goal:

Reference: Zanella, L., Menapace, W., Mancini, M., Wang, Y., Ricci, E. (2024). *Harnessing Large Language Models for Training-free Video Anomaly Detection*. CVPR 2024.

- To create a new **training-free LVLM** with enhanced Video Anomaly Detection (VAD) capabilities.

(Currently, the highest **ROC AUC** achieved by training-free LVLMs on the UCF-Crime dataset is around 0.8)

What is ROC?

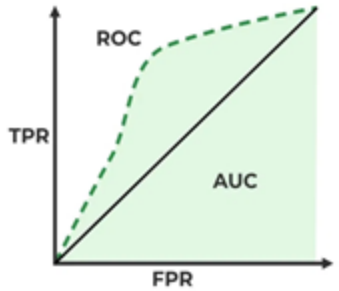
- Receiver Operating Characteristic - Area Under the Curve
- Used to quantify the model's ability to distinguish between normal and abnormal behaviors.

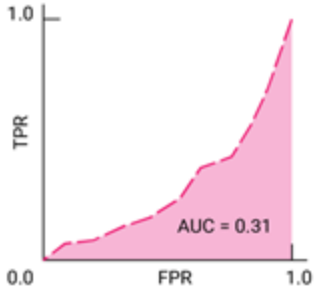
Algorithm:

- Set multiple thresholds. For each of them, we would get a point: (FPR, TPR)

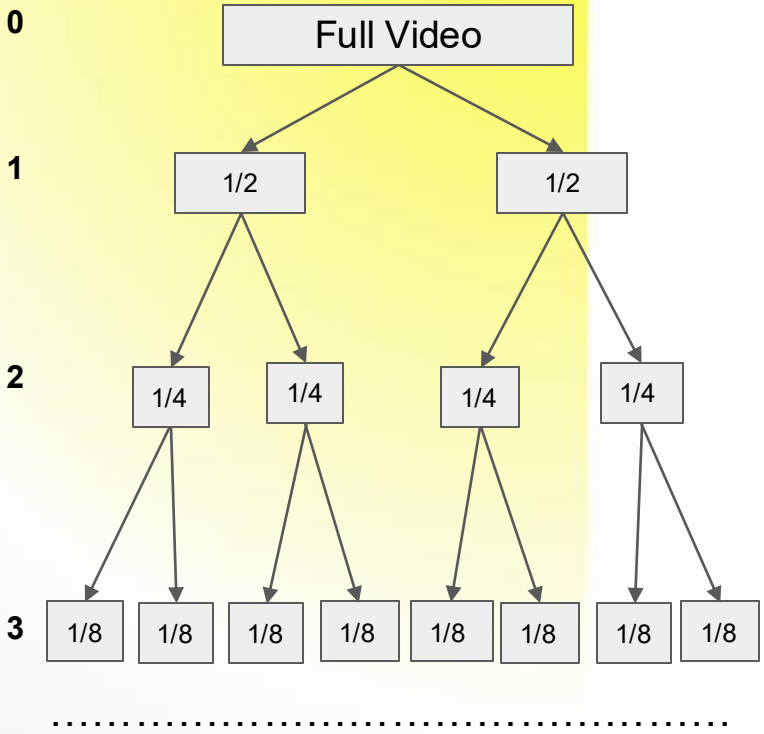
(FPR = #Wrong N frames/ #True N frames)
(TPR = #Correct Abn frames / # True Abn frames)

- Connect all these points together and calculate the area under the curve.





Our model 1.0:



Keep splitting in half. Eventually, we will get all frames at the end.

Execution Logic :

1. Splitting each video clip from last level into half
2. Apply Video-LLaVA to obtain the summary of each video clip at each level.
3. For each frame, we use LLaMA2 to aggregate the summaries of all levels of video clips that contain that frame, generating **a final summary**.
4. Apply LLaMA2 to score every frame based on its **final summary** and **caption**. Calculate the ROC.

Result:

Got ROC as **0.59**, compared with **0.71** from the previous paper.

Reasons → Low Result

- Using too many layers increases the probability of errors in summaries generated by LLaMA2, which may cause the final summary of a normal video to be described as abnormal.

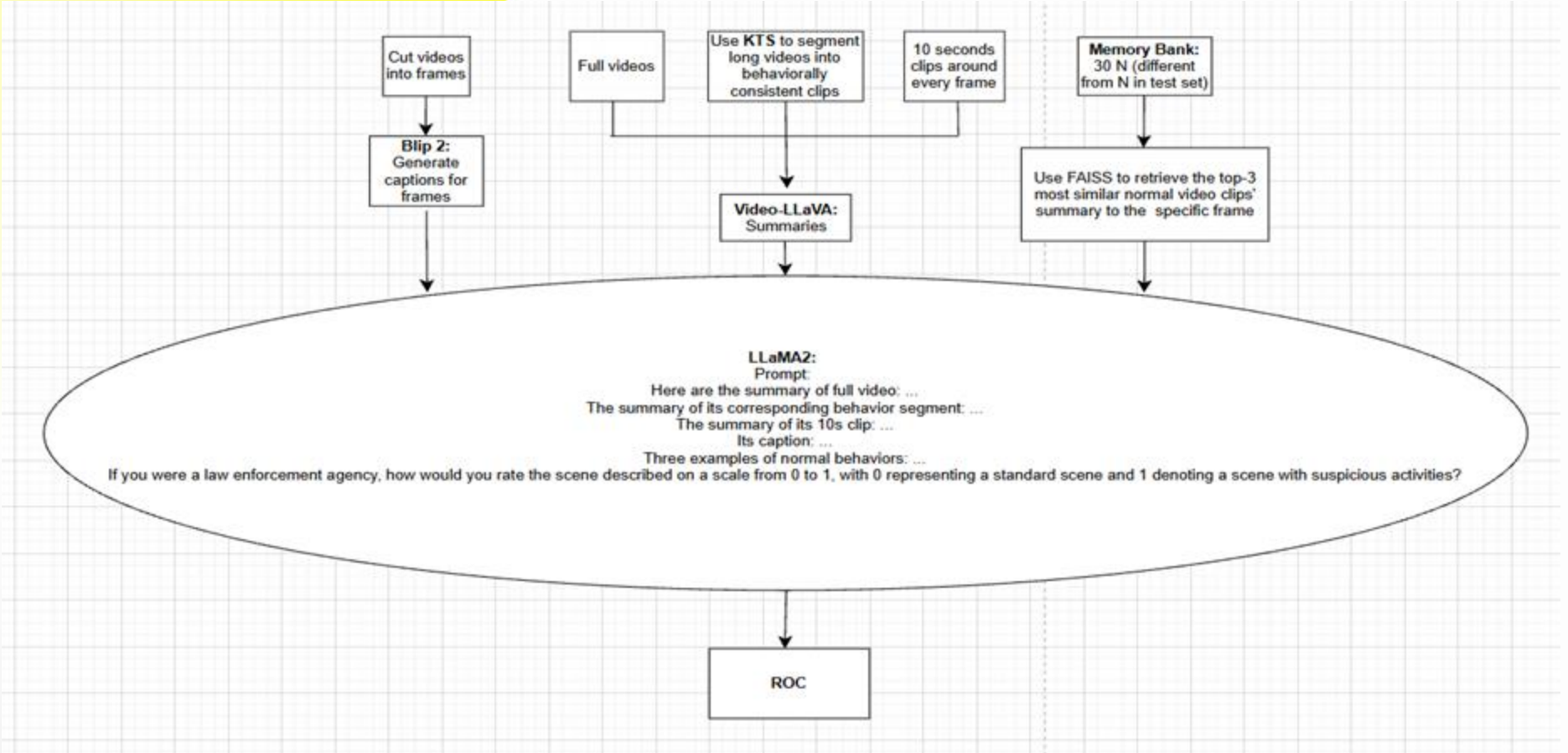
- Forgot to leverage the advantage of setting many layers as providing LLaMA2 with more contextual information during scoring.

- Lack the ability to determine whether different frames fall within the time span of an abnormal event.

Our model 1.1:

N: Normal videos
ABN: Abnormal videos

KTS： Kernel Temporal Segmentation



Future plan

1. After I got the result of our new model tonight, we will determine whether to do an ablation or change the construction of our model

Thank you!