CS482/682 Final Project Report Group XX

your project title here

Noah Drakes - ndrakes1

1 Problem Statement

The goal of this study is to replicate and fine-tune a currently existing anomaly detection modal in a semi-supervised fashion that utilizes multimodality, video and audio inputs, in order to predict the presence of violent/anomalous events. Examples of anomalous events would be a car crash happening on a busy streat or a fight breaking out in a subway. With the rise of advanced surveillance systems, there has been a push toward incorporating ML algorithims in the video security domain. By utiling robust anomaly detection algorithims in computer vision, security teams can quickly and autonomously identify suspicous activity in crowded areas and can help prevent anomalous events from escalating.

In many current anomalous detection algorithims, only one modality (video) is processed to predict the occurence of irreugular events. However, by adjoining the auditory modality, models can benefit from richer contextual understanding of scenes that are revealed in audio samples such as screams, crashes, or other loud sounds. Furthermore, in cases in which both modalities are low resolution (ie. noisy, compressed audio or blurry video), both modalities can help improve prediction accuracy.

2 Summary of Dataset

There are two datasets that are being considered for this project. The first being **XD-Violence** which comprises of 4754 untrimmed videos obtained form YouTube videos that are divided into with corresponding audio signals and weak labels (violent/normal). This dataset has gained popularity with research focusing on anomaly detection. We will have to do some preprocessing to shrink the video length and downsample the video resolution (x6 or x8 maybe). Videos are a very high-dimensional input and could easily increase model complexity and training time. Another dataset being considered is the **UCF-Crime** comprising of 1900 untrimmed videos of 13 realistic anomalous events, such as burglary, robbery, fighting, and so.

3 Related Papers

The first reference, "Learning Multimodal Violence Detection under Weak Supervision", uses the XD-Violence dataset to detect Violence by fusing video and audio modalities and using HL-Net architecture to capture short term and long term temporal information [1]. A 3D CNN is used for video feature extraction and VGGish (1D CNN) is used as audio feature extractor. HL-NET utilizes a local and global encoder to capture short-term patterns (1 - 2s) and long term dependencies across the entire video, respectively. The output of the model is a violence score based on individual snippets of video and audio which can be used for scene restricted violence prediction or video classification by max pooling or averaging.

Dataset BRIEFLY describe the dataset and any pre-processing you may use that is special (e.g. downsample all images by x8 to enable colaboratory training etc.)

Setup, Training and Evaluation IMPORTANT: you can change this paragraph to better fit your project. Questions to answer: What architecture did you chose and why? How did you accomplish un- or self-supervised learning? What tweaks were necessary to make this happen, e.g. custom layer designs, auxiliary tasks, etc.?

4 Results

5 References

References

[1] Shaoyuan Xu, Qi Jin, Yueming Liu, Kai Wang, and Tianqiang Ruan. ”Not Only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision.” *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.