TECHNICAL REPORT FOR 2023 AVA CHALLENGE

Team ALLAccept

1. METHOD

1.1. Feature Extraction

We refer to the data processing method of DSTA[1] and utilize Cascade R-CNN to extract bounding boxes, as well as VGG16 pre-trained on Cars-196 to extract features.

1.2. Model Architecture

Our network architecture utilizes DSTA. Due to its excellent ability to extract spatial and temporal information, DSTA has achieved outstanding performance on datasets such as DAD and CCD. Therefore, we employ DSTA as our network architecture.

1.3. Soft Label

The annotation of Abnormal start time t is subject to the subjective factors of the annotators, resulting in noisy annotations in the dataset. Additionally, the changes between adjacent frames in a video are often small. If the labels of the t-1 frame are roughly assigned as 0, and the t frame is assigned as 1, it will cause similar frames to have opposite labels, increasing the impact of noisy annotations. To address this issue, we propose using Soft Labels to make the label transitions smoother among frames in the video, reducing the difficulty of model training and the impact of noise in the dataset. Detailed pictures and formulas are shown in our detailed technical report.

1.4. Multiple Instance Learning Loss

Training solely using frame-level loss with noise may hurt the performance of the model. Therefore, from a multi-instance learning perspective, we designed a bag-level loss to assist the model in learning. Previous work[2] has made some attempts. Specifically, during training, we take N videos from the current iter batch and divide them into two categories: normal videos and accident videos. Then, for each video, we take the maximum value of the GRU output for all frames as the anomaly score of that frame. Next, we calculate two types of gaps: (1) overall gap, which is the difference between the average anomaly scores of normal videos and accident videos in the current batch, and (2) difficult sample gap, which is the difference between the maximum anomaly score of normal videos and the minimum anomaly score of accident videos in the current batch. We aim to increase this gap to enhance the model's ability to distinguish difficult samples. The formula of this loss is shown in (1), where $\boldsymbol{O} \in R^{N \times F}$ represents the GRU output anomaly score of all F frames of N videos in a batch.

$$L_{gap}(\boldsymbol{O}) = \max\left(0, 1 - \min\left(\{O_i\}_{y_i=1}\right) + \max\left(\{O_i\}_{y_i=0}\right)\right) + \max\left(0, 1 - \frac{\sum_{i,y_i=1}(O_i)}{n_p} + \frac{\sum_{i,y_i=0}(O_i)}{n_n}\right), \quad (1)$$

In which $O_i = \max_j\left(\boldsymbol{O_{ij}}\right)$.

Meanwhile, we also partition the video segments outside the soft label margin into normal and abnormal segments based on the annotated abnormal time point $\hat{t}$. Then, we calculate the frame-level gap loss by taking the minimum value of the abnormal segments and the maximum value of the normal segments. $\hat{t}_i$ represents the annotated abnormal time point of ith video.

$$L_{gap\_frame}(\boldsymbol{O}) = \sum_{i,y_i=1} \max\left(0, 1 - \max\left(\{\boldsymbol{O_{ij}}\}_{j \geq t_i - m}\right) + \max\left(\{\boldsymbol{O_{ij}}\}_{j < t_i + m}\right)\right) \quad (2)$$

2. RESULTS

We utilize the TSAA module of DSTA to extract information from all frames of the current video via an attention mechanism, to predict the occurrence of a car accident in the future. We calculated the accuracy on our self-labeled test set to compare the performance of different methods precisely. Experimental results demonstrate that our IDSTA method outperforms the original DSTA method.

Table 1. Accuracy on AVA test dataset.

Model	freeway	road
DSTA[1]	0.545	0.623
IDSTA	**0.630**	**0.701**
IDSTA (without soft label)	0.584	0.623
IDSTA (without gap loss)	0.565	0.623
IDSTA (without gap-frame loss)	0.597	0.630

3. REFERENCES

[1] Muhammad Monjurul Karim, Yu Li, Ruwen Qin, and Zhaozheng Yin, "A dynamic spatial-temporal attention network for early anticipation of traffic accidents," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9590–9600, 2022.

[2] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 2018, pp. 6479–6488, Computer Vision Foundation / IEEE Computer Society.