TECHNICAL REPORT FOR 2023 AVA CHALLENGE

Team ALLAccept

1. METHOD

1.1. Feature Extraction

We performed object detection on images using the pre-trained cascade RCNN network provided by DSTA[1], and retained the top (19-k) bounding boxes with the highest predicted confidence. We used the coordinates of the k bounding boxes provided in the AVA dataset, as well as the (19-k) bounding boxes selected as candidate regions, for a total of 19 bounding boxes. We then cropped the corresponding regions from the image based on the bounding box coordinates. Subsequently, we used a pre-trained VGG-16 network to extract feature vectors from the original image and the cropped regions.

Given that in the task of predicting car accidents, the key objects in the images are generally various types of vehicles, we utilized a pre-trained VGG-16 model on cars-196 to extract features, which is different from some previous works that used pre-trained models on imagenet to extract features. We found that using the pre-trained VGG-16 model on cars-196 for feature extraction achieves better performance than using pre-trained models on imagenet.

1.2. Model Architecture

Our network architecture utilizes DSTA. DSTA extracts spatial information from a single image through a Dynamic Spatial Attention module(DSA) and captures the relationship between different frames in a video through a Dynamic Temporal Attention module(DTA). The future probability of accidents is predicted using a gated recurrent unit(GRU). Due to its excellent ability to extract both 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 AVA dataset contains only video clips before the occurrence of accidents, and does not include video segments during and after the accident, which is different from datasets such as DAD and CCD. Therefore, we cannot annotate the time of accident occurrence (TOA) for each video segment similar to the DAD method (otherwise, TOA would be the last frame of the video), and calculate the frame-level loss based on this. In addition, the original frame-level loss of DSTA decreases as the accident prediction probability of any frame in the accident video increases, that is, this loss hopes that all frames in the accident video have a high accident prediction probability. However, in an accident video, abnormal segments usually only account for a small part of the video. Therefore, the design of the original DSTA loss is not suitable for the AVA dataset.

Based on the characteristics of the AVA task, we have redesigned the frame-level loss. We manually annotated the time point $\hat{t}$ at which each car accident video in the training set shows the beginning of abnormal behavior. The rule for selecting the annotated time point is that the annotator can perceive the abnormal behavior in the current frame at that time. After annotation, we can divide the video into normal and abnormal segments based on $\hat{t}$, and assign a label of 0 (indicating no car accident) to the frames in the normal segment, and a label of 1 (indicating a car accident) to the frames in the abnormal segment, thus obtaining frame-level binary classification label information. Then, we calculate the cross-entropy loss between the output of each frame's GRU and the binary classification label of the current frame, and take the average as the final frame-level loss.

However, due to the subjective factors of the annotators, the labeling of the moment t in exceptional cases can be considered noisy in the annotated dataset. Moreover, the change between adjacent frames of a video is often small. The approach of roughly assigning a label of 0 to frame $\hat{t}-1$ and a label of 1 to frame $\hat{t}$ would result in completely opposite labels between similar frames, thereby increasing the impact of labeling noise. To solve this problem, we propose using Soft Labels to make the label transition of each frame in the video smoother, reducing the difficulty of model training and minimizing the influence of dataset noise.

Formula (1) shows the mathematical expression of soft label we used. In the formula, m represents margin, T represents Temperature coefficient, which controls the soft degree of soft label, and x represents the label of current frame. Figure 1 plots hard Lables and Soft labels with different temperatures.

$$y = \begin{cases} \frac{\mathrm{softmax(m\,/\,T)}}{\hat{\mathrm{t}}-\mathrm{m}} \cdot x, & 0 < x \leq \hat{\mathrm{t}}-\mathrm{m} \\ \mathrm{softmax}\left[\frac{x-\hat{\mathrm{t}}}{\mathrm{T}}\right], & \hat{\mathrm{t}}-\mathrm{m} < x < \hat{\mathrm{t}}+\mathrm{m} \\ \frac{[\mathrm{softmax(m\,/\,T)}]\cdot(x-60)-x+\hat{\mathrm{t}}+\mathrm{m}}{\hat{\mathrm{t}}+\mathrm{m}-60}, & \hat{\mathrm{t}}+\mathrm{m} \leq x \leq 1 \end{cases} \tag{1}$$

Considering a car accident video, the abnormal segments typically only account for a small portion of the video, leading to the issue of imbalanced sample numbers across different categories. To alleviate this problem, we applied weighting to the cross-entropy (CE) loss of normal and abnormal frames, assigning a greater weight to the CE loss of abnormal frames.

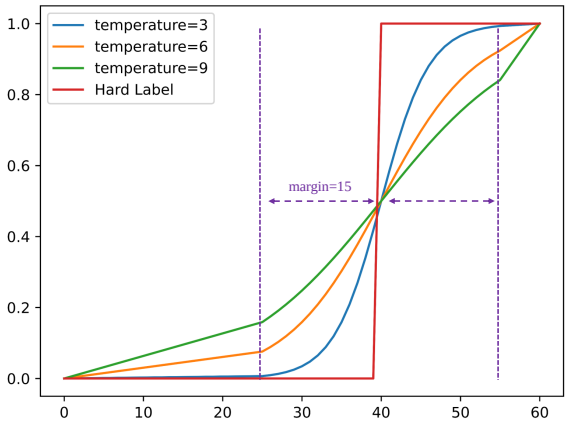

Fig. 1. Comparison of Soft Label and Hard Label. Temperature Controls how soft the label is.

1.4. Multiple Instance Learning Loss

As mentioned earlier, frame-level loss is noisy, and if only frame-level loss is used for training, it may still affect the performance of the model. However, multi-instance learning only requires bag-level labels in the training data, i.e., it does not require frame-level labels. As some previous works[2] have described, a video can be seen as a bag, and each frame or a segment of the video can be seen as an instance, so the original car accident prediction problem can be naturally regarded as a binary classification problem of multiple instance learning. Therefore, we designed a bag-level loss to assist the model in learning. Specifically, in the training process, we take N videos from the current batch and divide them into two categories: normal videos and car accident videos. Then we take the maximum value of the GRU outputs for all frames of each video as the anomaly score for that video. Afterwards, we calculate two types of gap loss: (1) overall gap loss: calculate the gap between the average anomaly scores of normal videos and car accident videos in the current batch; (2) hard example gap loss: calculate the gap between the maximum anomaly score of normal videos and the minimum anomaly score of car accident videos in the current batch. We hope to increase these gap to enhance the model's ability to distinguish difficult samples.

The formula of this loss is shown in (2), where $\boldsymbol{O} \in R^{N\times F}$ represents the GRU output anomaly score of all F frames of N videos in a batch.

$$L_{gap}(\boldsymbol{O}) = \max\left(0, 1 - \min\left(\{O_i\}_{y_i=1}\right) + \max\left(\{O_i\}_{y_i=0}\right)\right) + \max\left(0, 1 - \frac{\sum_{i,y_i=1}(O_i)}{n_p} + \frac{\sum_{i,y_i=0}(O_i)}{n_n}\right), \quad (2)$$

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.

$$L_{gap\_frame}(\boldsymbol{O}) = \sum_{i,y_i=1} \max\left(0, 1 - \max\left(\{\boldsymbol{O_{ij}}\}_{j\geq t_i-m}\right) + \max\left(\{\boldsymbol{O_{ij}}\}_{j<t_i+m}\right)\right) \quad (3)$$

1.5. TSAA Prediction

In our experiments, we found that using the output of the TSAA module with DSTA as the prediction result achieves higher accuracy than using the output of GRU as the prediction result, which differs from the original DSTA method that uses GRU for prediction. One possible reason we analyzed is that the original DSTA method uses GRU to output the probability of a future car accident for each frame in a video, and when the probability of a certain frame exceeds a threshold, the video is considered to have a car accident in the future. This method is susceptible to interference from individual difficult frames, resulting in incorrect results. However, DSTA's TSAA module extracts information from all frames of the current video through an attention module to predict whether a car accident will occur in the future, thus possessing better robustness.

2. RESULTS

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.