Look, Listen and Pay More Attention: Fusing Multi-Modal Information for Video Violence Detection

1985

Donglai Wei [1] Chengeng Liu [2] Yang Liu [1] Jing Liu [1] Xiaoguang Zhu [3] Xinhua Zeng [1]

[1]Academy for Engineering & Technology, Fudan University, Shanghai, China [2]School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia [3]SEIEE, Shanghai Jiao Tong University, Shanghai, China

Abstract

Video Violence Detection is an essential and challenging problem in the computer vision community. Most existing works focus on single modal data analysis, which is not effective when multi-modality is available. Therefore, we propose a two-stage multi-modal information fusion method for violence detection: **1)** the first stage adopts multiple instance learning strategies to refine video-level hard labels into clip-level soft labels, and **2)** the next stage uses multi-modal information fused attention module to achieve fusion, and supervised learning is carried out using the soft labels generated at the first stage. Extensive empirical evidence on the XD-Violence dataset shows that our method outperforms the state-of-the-art methods.

Introduction

Violence detection is crucial in maintaining social security, which has been researched for years. Especially, solely using visual information to build a violence detection model is not robust or powerful enough. For example, it is difficult to obtain sufficient information in a surveillance area with obstacles or dim light, and in these cases, audio will be a good supplement. Multi-modal information can provide comprehensive and copious features, which can be more robust and accurate in detecting violence than single-modal information. Therefore, our study is based on the fusion of audiovisual features.

Our contributions in this paper are summarized as follows:

- We propose a **Weakly Supervised** violence detection model targeting **Multi-Modal Information**.
- We propose a **Multi-Modal Co-Attention Mechanism** to encourage the model to learn the audiovisual features of violent information.
- We conduct extensive **Qualitative** and **Quantitative** experiments. Experiment results on benchmark demonstrate the effectiveness of the ACF network.

Overall Architecture

The overall architecture of our method is shown in Figure 1. The ACF network consists of Single Self-attention (SA) module and Fusion Co-attention (FA) module. Each module contains stacks of SA units and FA units, respectively. Both video and audio features are first sent to the SA module for self-attention processing and then enter the FA module to complete the co-attention enhancement.

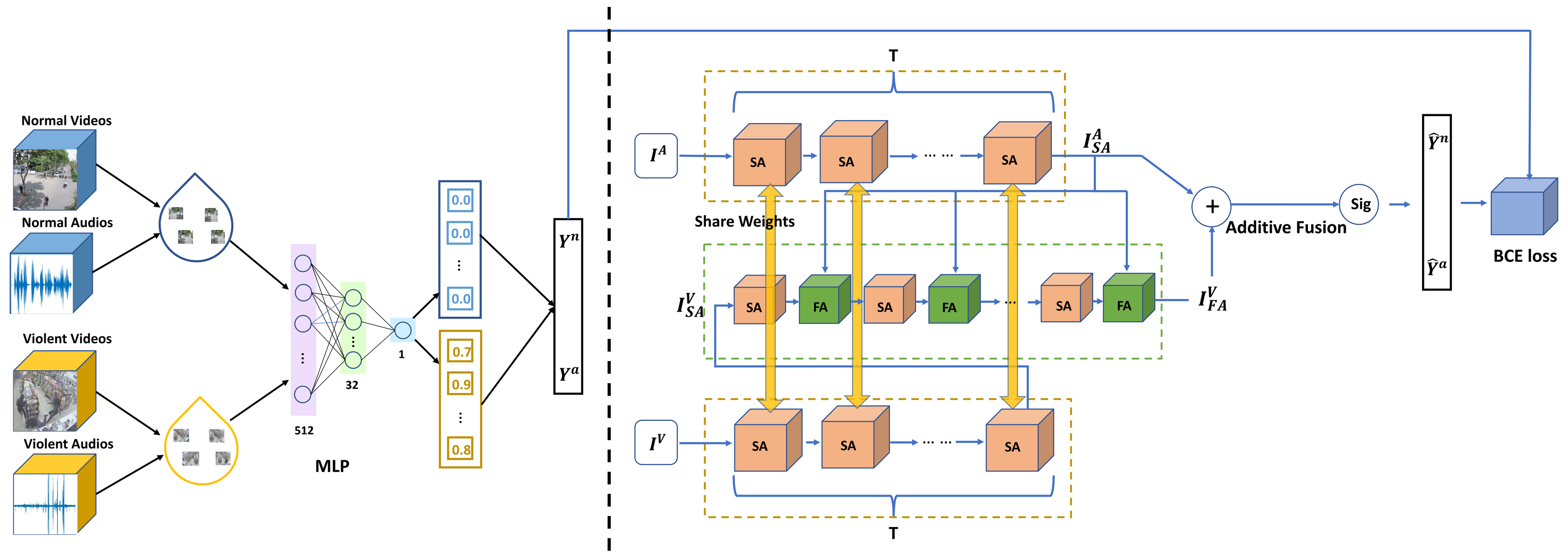


Figure 1. The overall architecture of the proposed method.

Method

1. Clip-level Labels Refinement via MIL

We use a refinement process to create clip-level soft labels. Clip-level soft labels have more fine-grained annotation information, and the ACF network can be better supervised in the next stage with them. We use Bag_v and Bag_n to represent the set of positive and negative bags. The positive bag is a violent video with its associated audio information. In contrast, the negative bag is a normal video with audio. Video segments in bags served as instances. We select the top K pairs of instances with the largest violent score from the sets of bags to calculate the loss.

$$L_{Total} = L_{MIL} + \frac{\lambda}{K} \cdot \left(\sum_{i=1}^{K} L_{BCE}\right)$$

We design L_{MIL} as follows:

$$L_{MIL} = \max\left(0, 1 - \max_{(1 \le k \le K)} Bag_v^k + \max_{(1 \le k \le K)} Bag_n^k\right)$$

Based on this, we can train a shallow clip-level soft label generator to obtain fine-grained labels and provide them to the ACF network for supervised training.

2. Audiovisual Co-attention Fusion network

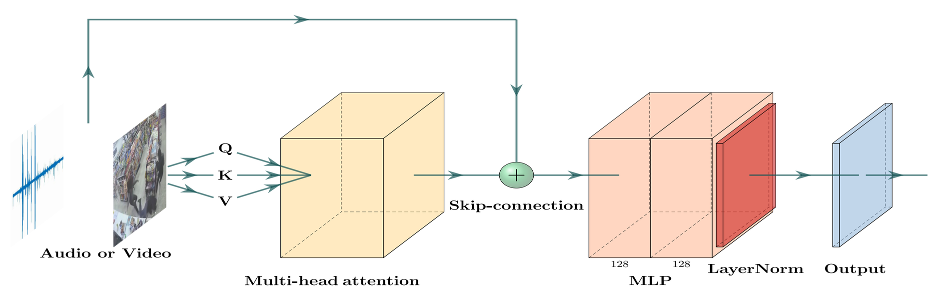
The Single Self-attention module is composed of T cascaded SA units (Figure 2(a)). It is mainly composed of a multi-head attention layer and a fully connected layer. L and M individually represent fully connected layer and layer normalization, and $[\cdot]_T$ means continuous cascaded connection T times:

$$I_{SA}^{V/A} = \left[L\Big(M(I^{V/A} + Multihead(I_k, I_v, I_q))\Big)\right]_T$$

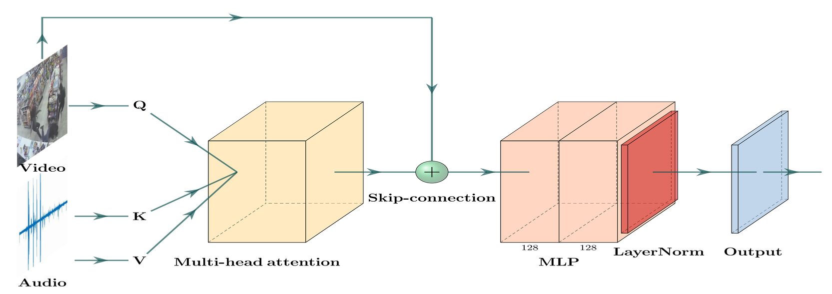
The FA module is composed of T FA units cascaded, where the FA unit is shown in Figure 2(b). Take audiovisual features I_{SA}^V and I_{SA}^A as a set of examples. Features achieve mutual attention between each other through the multi-head attention layer. The calculation process is as follows:

$$I_{FA}^{V} = \left[L\Big(M(I_{SA}^{V} + Multihead(I_{SA_k}^{A}, I_{SA_v}^{A}, I_{SA_q}^{V}))\Big)\right]_T$$

Mutual attention between multi-modal information in the FA module further augments their correlation. Therefore, the two obtained modal features can be merged effectively.



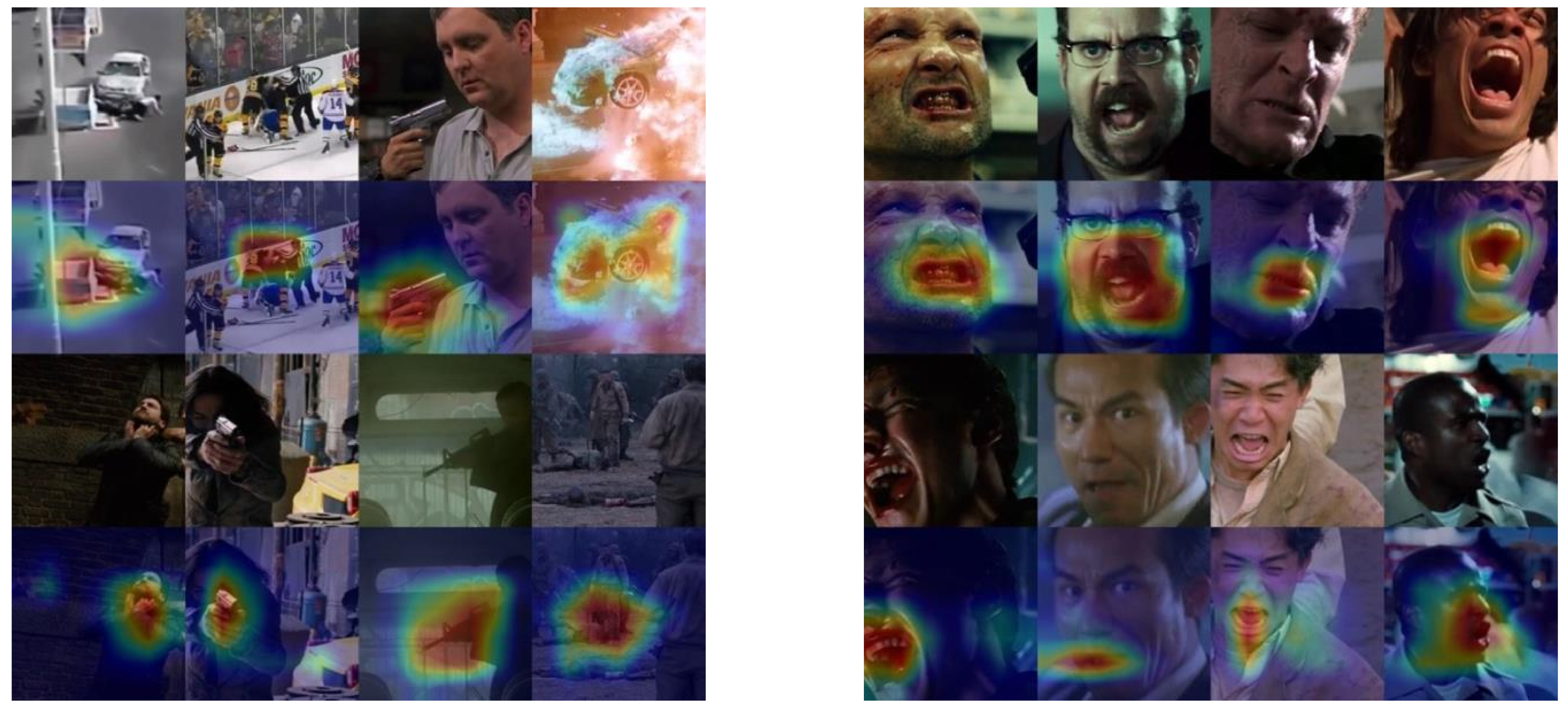
(a) Single Self-attention unit



(b) Fusion Co-attention unit

Experiments

Visualization results of violence maps



Comparisons of different modal information

Video	Audio	FAR (%)	AUC-ROC (%)	AP (%)
✗	✓	17.30	75.84	50.74
✓	✗	1.96	91.82	72.09
✓	✓	1.12	93.87	80.13

The result of ablation studies is in line with our prediction. It illustrates **Multi-Modal Audiovisual Information** based on fused attention has significant values in violence detection. The multi-modal information can help them complement each other, which significantly improves the model performance.

Comparisons of existing methods

Method	AP (%)
SVM	50.78
OCSVM [23]	27.25
Hasan *et al.*[19]	30.77
Sultani *et al.*[3]	73.20
Wu *et al.*[13]	78.64
ACF (ours)	**80.13**

Future work

This paper focuses on the violence detection task with multi-modal information. We propose a two-stage weakly supervised learning method, which pays more attention to the fusion of multi-modal features. We will explore other modal information in this field, utilizing more multi-modalities efficiently to enhance model ability is our main direction for future work.

This work is supported by the Shanghai Key Research Laboratory of NSAI.