SNIFR : Boosting Fine-Grained Child Harmful Content Detection Through Audio-Visual Alignment with Cascaded Cross-Transformer

Orchid Chetia Phukan [1] Mohd Mujtaba Akhtar* [1] Girish* [1] Swarup Ranjan Behera [2]
Abu Osama Siddiqui [1] Sarthak Jain [1] Priyabrata Mallick [2] Jaya Sai Kiran Patibandla [2]
Pailla Balakrishna Reddy [3] Arun Balaji Buduru [1] Rajesh Sharma [4]

[1]IIIT-Delhi, India [2]Independent Researcher, India [3]Reliance AI, India [4]Plaksha University, India
* Equally contributed

Warning: The following study includes visualizations of sensitive content. Readers are advised to proceed with discretion.

Motivation

- Video platforms are widely accessed by children, yet malicious content is often embedded in only a few frames, bypassing traditional moderation.
- Prior fine-grained detection efforts focus almost entirely on visual cues, neglecting audio, which often contains strong semantic indicators (e.g., threatening tones, alarming sound effects or suggestive sounds).

Contributions

- We hypothesize that audio cues are complementary to visual signals in identifying harmful child content.
- Structured fusion of these modalities will outperform unimodal and naive fusion baselines.
- We propose SNIFR (CrosS-Modality INteractIon Cascaded TransFoRmer), a two-stage cascaded cross-transformer framework for deep audio-visual interaction.

Architecture

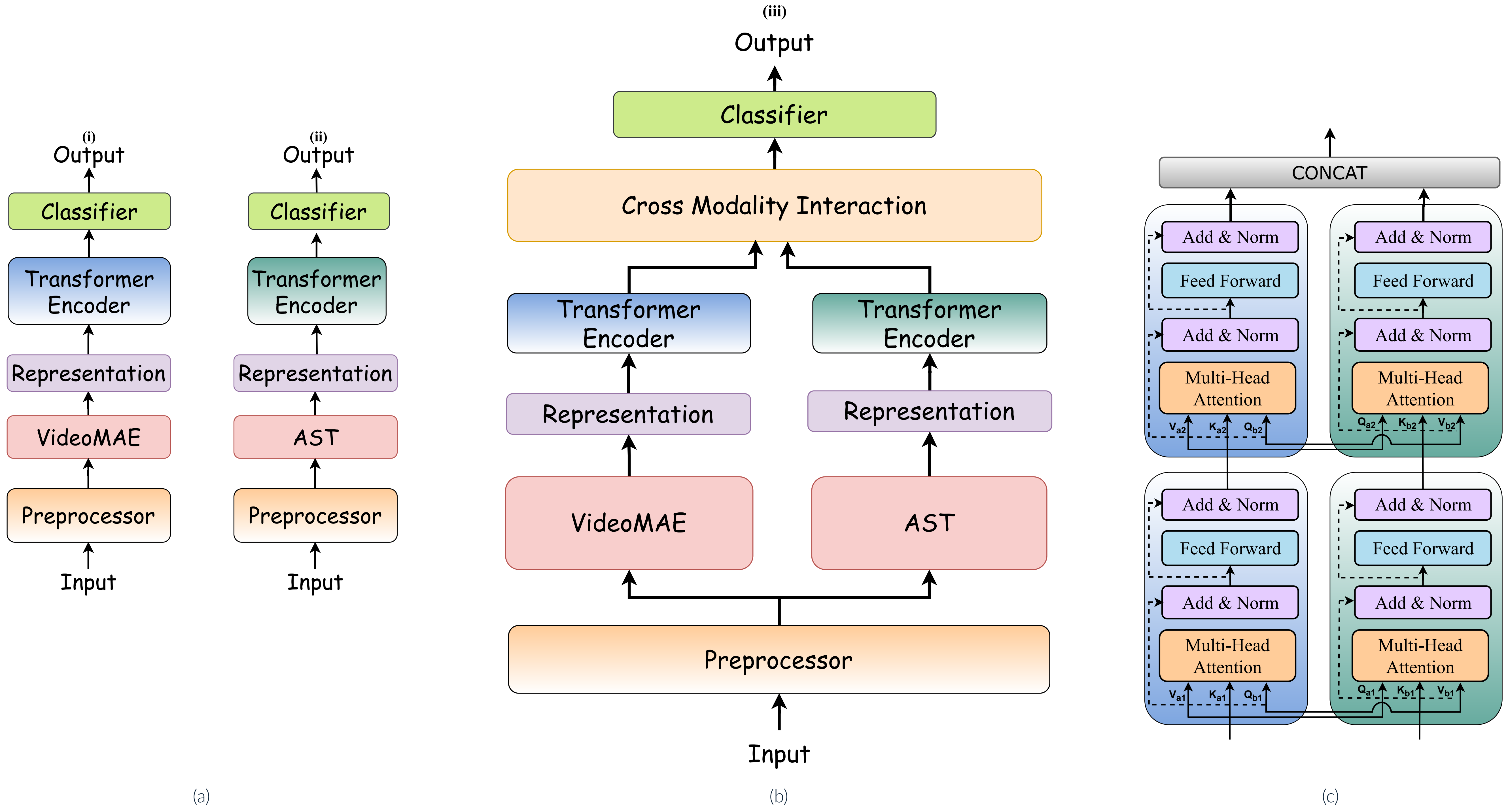

Figure 1. Modeling Architectures: Subfigure (a) show the individual unimodal modeling pipeline for video and audio, respectively; Subfigure (b) shows the proposed framework, SNIFR; Subfigure (c) provides the detailed illustration of the cross modality interaction through the cascaded cross-transformer

Result

Modality	Safe			Sexual			Violent			Both		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
V	85.45	89.48	90.55	90.70	73.68	95.37	66.18	56.96	87.63	64.71	64.08	91.49
A	71.27	80.29	77.92	83.33	58.82	82.87	46.48	35.11	74.81	50.00	27.91	78.58
AV (EC)	87.19	87.85	88.93	72.73	71.91	96.58	58.88	61.17	87.01	48.48	40.51	89.10
AV (LC)	82.29	84.40	85.62	85.71	75.00	91.06	58.25	56.34	83.85	60.00	57.14	92.95
AV (EA)	79.91	84.99	86.67	81.48	60.27	88.13	53.42	46.43	85.38	53.85	46.67	89.09
AV (EP)	78.33	81.23	83.60	70.59	58.54	95.28	45.45	44.55	81.15	51.66	45.36	88.29
AV (CT)	84.65	87.92	90.62	81.25	71.23	96.83	66.02	65.38	89.83	79.07	64.76	95.34
AV (SNIFR)	**88.24**	**91.49**	**95.28**	**93.33**	**82.11**	**98.72**	**84.15**	**77.09**	**96.19**	**79.59**	**75.73**	**97.82**
SOTA	-	-	88.00	-	-	95.00	-	-	90.00	-	-	91.00

Table 1. Evaluation results showing Accuracy (ACC), Macro-average F1 (F1), and AUC in % across different classes; 'Both' denotes clips with both sexual and violent content. AV variants include Early Concat (EC), Late Concat (LC), Element-wise Avg (EA), Product (EP), Cross-Transformer (CT), and our SNIFR.