Multimodal Hate Detection Using Dual-Stream Graph Neural Networks

Jiangbei Yue[1], Shuonan Yang[1], Tailin Chen[1], Jianbo Jiao[2], Zeyu Fu[1]

BMVC 2025

[1] University of Exeter [2] University of Birmingham

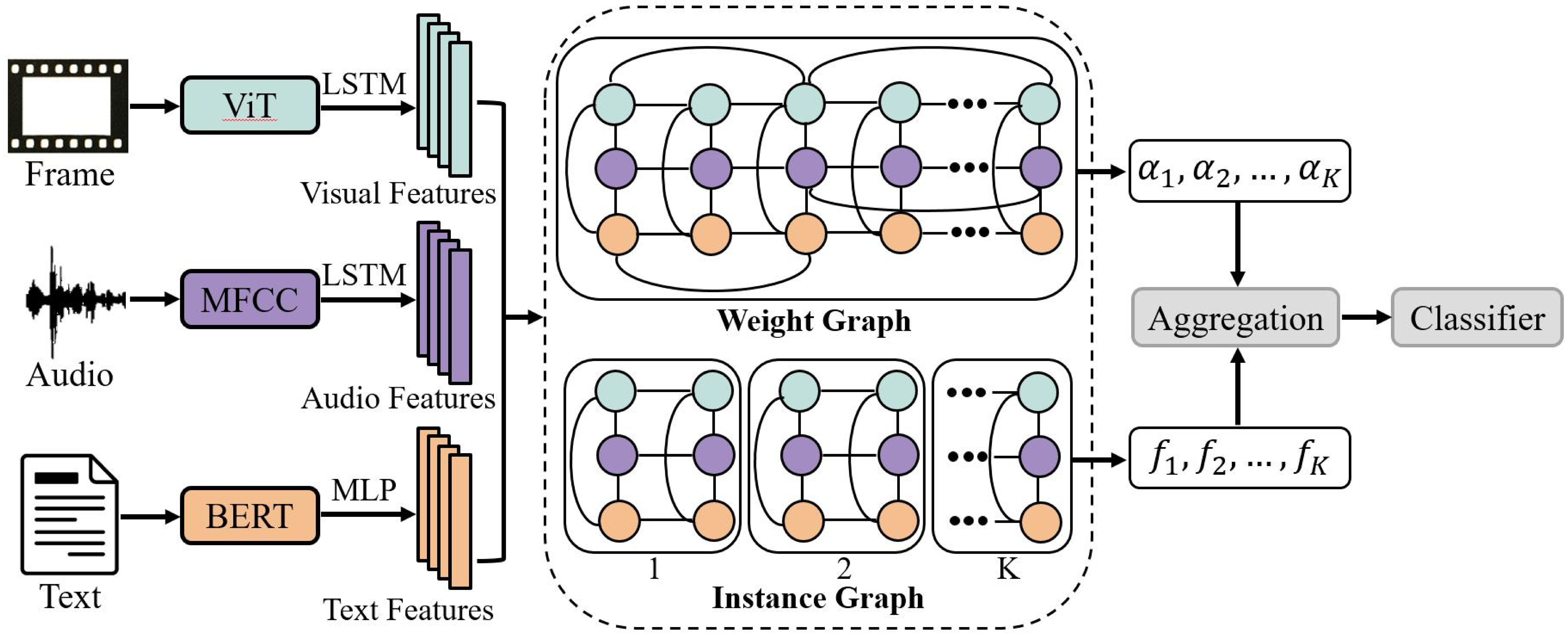
1. Motivation

- We aim to detect hateful vidoes by classifying videos as hateful or non-hateful, mitigating the risks that they bring.
- Existing methods tend to aggregate all content equally, which overlooks the importance of highlighting hateful components. Moreover, these methods frequently fail to model systematically crucial structured information in video content. To overcome the two challenges, we propose a new graph neural network (GNN) model.

2. Contributions

- We propose a novel multimodal dual-stream graph neural network model, MultiHateGNN, to classify hateful videos effectively. The graph-based framework systematically models both intra-modal and inter-modal structured information.
- We propose a new dual-stream graph architecture to emphasize crucial hateful content for video classification.

3. Multimodal Dual-stream Graph Neural Network Model



- Overview of the proposed MultiHateGNN. Our model cuts the input video into N segments and extracts multimodal segment-level features, including visual, audio and text features. The weight graph and instance graph are built based on these features. The two graphs are fed into corresponding GNNs to yield instance features $\{f_i\}_{i=1}^{K}$ and instance importance weights $\{\alpha_i\}_{i=1}^{K}$. After aggregation, a classifier gives label prediction.

4. Quantitative Experimental Results

Model	Accuracy	F1-Score	Precision	Recall
ViT	0.693	0.601	0.623	0.589
MFCC	0.682	0.622	0.602	0.651
BERT	0.706	0.630	0.646	0.622
GPT-4o	0.777	0.755	0.671	**0.863**
HateMM	0.805	0.753	0.765	0.751
CMFusion	0.799	0.739	0.763	0.719
Ours	**0.821**	**0.771**	**0.798**	0.754

Model	Accuracy	F1-Score	Precision	Recall
ViViT	0.73	0.68	**0.86**	0.57
MFCC	0.54	0.36	0.40	0.33
mBERT	0.57	0.52	0.68	0.42
GPT-4o	0.68	0.29	0.57	0.19
MHC	0.75	0.67	0.77	0.61
CMFusion	0.73	0.72	0.72	0.73
Ours	**0.78**	**0.77**	0.80	**0.77**

- The classification performance on the HateMM dataset and the MHC dataset are shown in the left table and right table, respectively. The best results are shown in bold.

5. Qualitative Experimental Results

- The Visualization of Model Prediction. Every instance is annotated as hate or non-hate. $\{\alpha_i\}_{i=1}^{10}$ denote the instance weights. Our model can accurately highlight the crucial hate instance 3 by the associated weight to predict the correct label.