Multi-Modal Solution: Deepfake Detection and the Source Identification

Yahan Zheng*[1,a],Xu Zhou[2,b],Cheng Chen[3,c],Jingwen Hu[4,d]

2020141440150@stu.scu.edu.cn[a],34799214@qq.com[b],211370114@dhu.edu.cn[c],
jovanna522628@gmail.com[d]

College of Computer Science, Sichuan University Chengdu, 610207, China[1]
College of Computer Science, Zhejiang University of Technology Hangzhou, 310014, China[2]
College of Computer Science and Technology, Donghua University Shanghai, 201620, China[3]
KangChiao International School Suzhou, 215332, China[4]

Abstract. Deepfake technology has recently raised significant concerns due to its potential for manipulating and misusing multimedia content. In response to this issue, researchers have been exploring novel approaches for deepfake detection. In this study, we propose a multimodal analysis framework that combines visual, audio, and textual modalities to determine if an unknown video is a fake one and to identify the source identity in manipulated media. By leveraging the complementary information from multiple modalities, our approach aims to enhance the accuracy and robustness of deepfake detection.

Keywords— Deepfake detection; Multi-modal; MMMU-BA

1 INTRODUCTION

In recent years, with the rapid advancement of technology and artificial intelligence, the emergence of "deepfake", highly realistic manipulated media, has become increasingly prevalent in our lives. While deepfake technology has undoubtedly enhanced visual effects and entertainment, its rise has also ignited concerns regarding trust, credibility, and authenticity. It is now imperative to develop effective methods to mitigate the potentially harmful effects of deepfakes.

Traditionally, deepfake detection has heavily relied on the analysis of visual artifacts or facial features, seeking inconsistencies that can be discerned through meticulous examination [1]. This method involves scrutinizing factors such as image quality, unnatural facial movements, blurring, misalignment of facial features, and peculiar reflections in the eyes. These visual anomalies can serve as indicators of a deepfake [2]. However, these methods often come with limitations, as visual artifact analysis primarily focuses on individual frames within a video and use only the image information in the video, while deepfakes can extend across multiple frames or even entire video sequences. Simply identifying minor differences in one frame may not suffice to detect deepfakes that exhibit realism and consistency throughout the entire video.

To overcome these limitations, we have enhanced the framework for identifying the source face in manipulated media. Our proposed framework initially extracts various features from each modality. In audio analysis, speaker recognition algorithms are employed to compare speech

PMBDA 2023, December 15-17, Nanjing, People's Republic of China

DOI 10.4108/eai.15-12-2023.2345299

patterns and identify discrepancies between the original and manipulated voices. In image analysis, facial recognition techniques are used to assess facial landmarks, expressions, and other visual cues for person identification. Furthermore, text analysis, derived from the audio, is applied to the accompanying textual information, enabling the comparison of features to identify any disparities or semantic inconsistencies

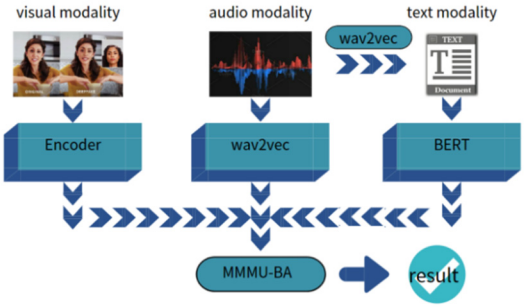

Fig. 1. Process of training

Originally, each raw video, audio, and text were trained separately in a model, each training process resulting in an individual loss function. Subsequently, these losses were combined, resulting in a total of three individual losses, in addition to one more loss in the MMMU-BA model. This configuration led to a total of four loss functions, ultimately impacting the quality and accuracy of the result. Our model's design is illustrated in Figure 1. The three feature processing models are positioned side by side and interconnected in series with the MMMU-BA model. By freezing the majority of parameters in the smaller model and reserving a small portion for joint training with the larger model, the total loss can be reduced to a single instance. Our model's design is illustrated in Figure 1. The three feature processing models are positioned side by side and interconnected in series with the MMMU-BA model. By freezing the majority of parameters in the smaller model and reserving a small portion for joint training with the larger model, the total loss can be reduced to a single instance. Our model's design is illustrated in Figure 1. The three feature processing models are positioned side by side and interconnected in series with the MMMU-BA model. By freezing the majority of parameters in the smaller model and reserving a small portion for joint training with the larger model, the total loss can be reduced to a single instance.

2 Related work

Two thriving research areas closely related to our work in the deep learning community are multimodal applications and the MMMU-BA model. We will introduce some key works below.

Salvi et al. [3] proposed a novel approach for leveraging data from multiple modalities to detect deepfake videos, garnering significant attention from society. In the realm of deepfake detection, there is an increasing use of multiple modalities (such as audio, images, and text) to gain a more comprehensive understanding of the content. By incorporating various modalities, researchers can extract different patterns from videos, enhancing the ability to identify deepfakes.

Videos are valuable sources for multimodal deepfake detection due to their ability to extract multiple modalities simultaneously. Through processing, we can extract relevant information from the video content, including audio, images, and textual cues. The audio component of a video can reveal inconsistencies in voice quality, speech patterns, or lip synchronization, providing vital cues for detecting deepfakes. Visual analysis enables the examination of facial expressions, eye movements, and other physical attributes that may display anomalies introduced by deepfake manipulations. Additionally, the textual information present in videos, such as subtitles or overlaid text, can be analyzed to detect any inconsistencies or artifacts indicative of tampering.

By combining these modalities, researchers aim to develop sophisticated algorithms and models that can effectively discriminate between real and deepfake video sequences. The fusion of audio, visual, and textual information allows for a more robust and holistic assessment of the authenticity of the content, enabling more reliable deepfake detection.

Building upon this, Ghosal et al. [4] conducted research on context inter-modal attention for multimodal analysis. In their research, Ghosal et al. introduced a model called the Multi-modal Multi-utterance Bi-modal Attention **(MMMU-BA)** Framework, which integrates Multi-Utterance Attention, Bi-modal Attention, and Multiplicative Gating & Concatenation. By leveraging multimodal and multi-discourse attention along with bimodal fusion, the **MMMU-BA** framework aims to provide a more nuanced understanding of sentiment in multimedia data. This enables the model to effectively capture the complex interactions between modalities and discourse levels, leading to improved accuracy and performance in multimodal sentiment analysis tasks.

Although their research primarily focuses on multimodal sentiment analysis, it has motivated us to apply our model to multimodal research on deepfakes.

3 PROPOSED METHODOLOGY

3.1Datasets

To date, FakeAVCeleb selected 490 genuine samples from the VoxCeleb2 dataset, which is the first known tampered dataset encompassing both counterfeit videos and fabricated audio. The composition of the dataset is shown in Figure 2. The original VoxCeleb2 comprises 1,092,009 authentic data samples, maintaining a balanced distribution in terms of race and gender. For the creation of manipulated content, we harnessed the Faceswap, DeepFaceLab, and FSGAN algorithms to generate deepfake videos, coupled with SV2TTS for generating synthetic audio, and Wav2Lip for simulating lip movements. As a result, over 20,000 manipulated videos were generated. [5]

Fig. 2. FakeAVCeleb

3.2Feature Extraction

3.2.1Audio Extraction

Wav2Vec 2.0's model The primary objective of this model is to build a conv1d structure dedicated to extracting essential features from the original audio data. By constructing this model, the function establishes a robust framework for effectively capturing relevant information embedded within the audio signals. Subsequently, the forward function is responsible for directing the input data through the constructed model, enabling the extraction of significant features during the data processing phase. To facilitate feature extraction, the function relies on the _get_feature_extractor method. Within this method, the ConvLayerBlock object and the FeatureExtractor object are predominantly utilized. Their combined efforts contribute to the successful construction and utilization of the feature extraction model. By leveraging these mechanisms, the function ensures that the most informative characteristics of the original audio are accurately captured, paving the way for further analysis and application in various domains such as speech recognition or audio classification tasks (figure 3).

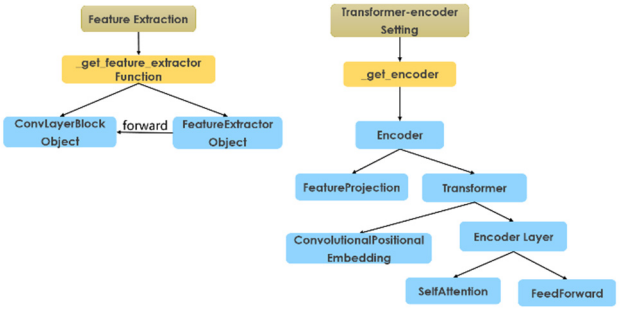

Fig.3. Wav2Vec's model

The Wav2Vec2.0 model converts speech to text A framework is used to self-supervise the learning of representations from raw audio data. The Wav2Vec2 model is trained using Connected Timing Classification (CTC), so the model output must be decoded using the Wav2Vec2Tokenizer. We will use the Wav2vec 2.0 model to convert audio files to text [6-8].

3.2.2Video/Frame Extraction

There are three main job we've done, keyframe extraction, face detection and localization, image feature extraction, and deepfake detection.

When faced with a large amount of video, processing all the frames will be very inefficient. To simplify the process, we need to compress the frames. A simple way to do this is to use SAD, or Sum of Absolute Differences (SAD), as a metric. the basic principle of SAD is to compare two images by calculating the absolute difference between the corresponding pixel values and then summing up all the differences.It selects a fixed number of keyframes with the largest differences. Their original order is then restored and they are stacked together, thus preserving critical temporal information.

Face detection and localization To eliminate background interference, we need to crop the faces in the video. Unlike standard face detection, we do not perform angle correction or alignment, so we can preserve the natural behavior of the respondents. This is an example of the bounding box we use for face detection. In our study, we used the MTCNN (Multi-Task Cascaded Convolutional Network) model, which consists of three stages. P-Net quickly generates candidate face bounding boxes and face landmarks. This network uses a multi-

scale sliding window to propose potential regions containing faces. R-Net uses a deeper network architecture to refine the candidate bounding boxes of PNet. o-Net acts as the final face detector, further filtering and adjusting the candidate boxes proposed by R-Net while predicting the locations of facial landmarks. For MTCNN, a sigmoid function is used at each layer to obtain a confidence score to evaluate the locations of the obtained bounding boxes. In our data, almost all of the confidence scores exceed 99%. The third part of my work is visual feature extraction. For single-frame facial images, we can easily extract features using well-established models such as Facenet, ArcFace, and InsightFace. However, we are interested in preserving temporal information (figure 4).

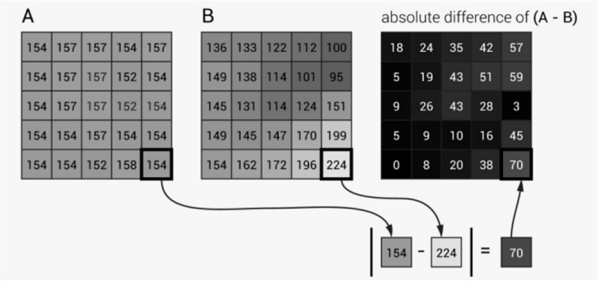

Fig.4. Key Frames Extraction

Our goal was to capture the facial muscle and head movement patterns of individuals as they speak. Initially, we considered using **LSTM**, but at that time updating weights in large models was a challenge. Through some research, we discovered various models for video information extraction and found that the I3D model performed well on many datasets. Therefore, we are currently using this model.

3.2.3Text Extraction

For the feature extraction of sentences, we use Bert pre-trained model and fine tune the model. We use text extracted through the wav2vec model as input and we take the output of Bert model as the sentence feature vector (figure 5).

Sentences are encoded into the format required for Bert input and the output of the Bert model is obtained. Since the input Sentence has different lengths and we want to get uniform length Embedding, we need to perform a pooling operation when the Sentence is output from BERT, we use the CLS-pooling operation: directly take the Embedding of [CLS] to store

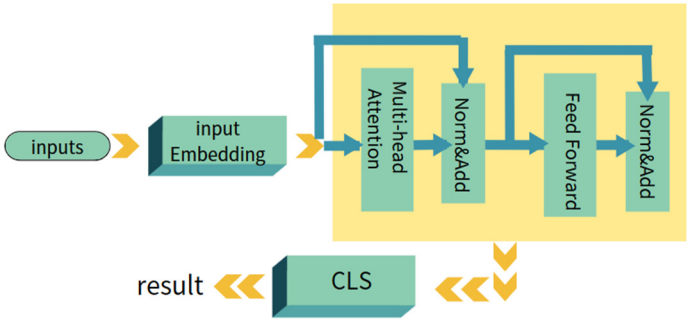

Fig.5. sen2vec-Finetune

3.3MMMU-BA Model

We have converted these three patterns into corresponding feature vectors;

We have performed deep forgery detection on the test set, so we have labels for all the videos - the training set labels from the dataset labels and the test set labels from the deep forgery detection;

Finding the original videos. Our approach is to use real-life videos for training, and obtain multimodal feature vectors for each real-life person through the MMMU-BA model, i.e., using the output of the covariance layer as the feature vector.

For example, here is a target face and here is a target face, both of them are real people; the output is a "fake" person, and since we use multimodal features, the feature vector of this person will be similar to these two people. Therefore, we can find the target person by comparing the vectors.

Locate the original video Obtain a video that has been identified as a forgery and input it into the MMMUBA network to extract the multimodal feature vector associated with the forger. At this stage, the MMMUBA network serves as a feature extractor.

Once we've established that the video is a forgery, we can utilize our dataset to determine the identity of the source character. For instance, if we are aware that the fake face corresponds to character A, mentioned previously, and the original video of character A is labeled as B, our objective is to identify B. To achieve this, we can eliminate the features attributed to character A from the overall set of features extracted from real individuals. This process ensures that among the remaining features associated with genuine persons, only B exhibits the closest similarity to the features extracted from the forgery. By comparing the feature vectors of the forged video with those of all genuine individuals, we can pinpoint the most similar vector, which corresponds to our target, B.

the Optimization in Implementation Initially, our approach was to employ these extensive models as feature extractors directly. However, this led to the loss calculation occurring three times, and when integrating the features into MMMUBA, an additional loss computation was carried out, potentially diminishing the overall effectiveness of our model.

To streamline this process, we decided to incorporate these three models into MMMUBA to create a unified, comprehensive model. We selectively froze specific layers and exclusively trained the remaining ones. By doing so, we enabled the joint training of all variable parameters within a single, consolidated model. This streamlined approach necessitates the computation of only one loss, ultimately resulting in enhanced accuracy (figure 6, 7, 8).

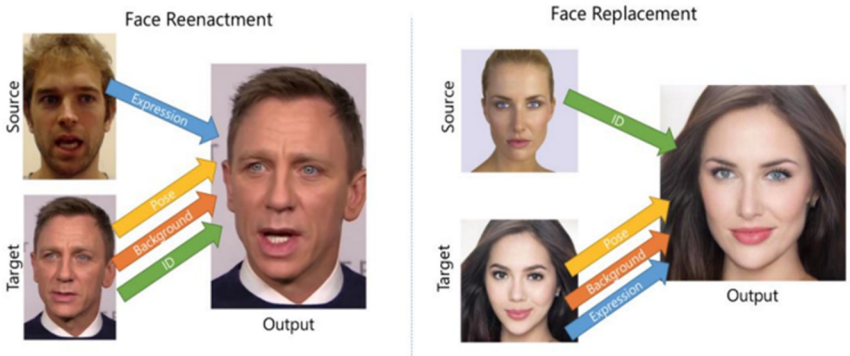

Fig.6. Multimodal Identity Features

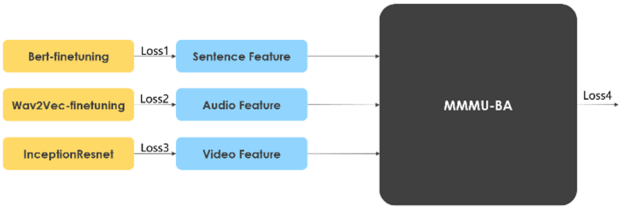

Fig.7. Optimization in Implementation1.0

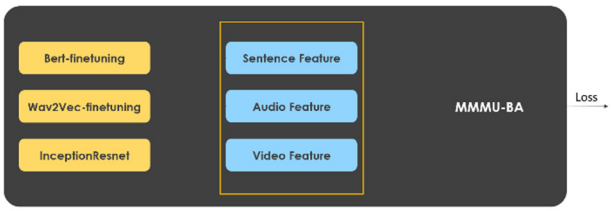

Fig.8. Optimization in Implementation2.0

4 Conclusion

Project Progress

Unfortunately, we did not complete the entire program in those two weeks, so there are no experimental results or conclusions. We completed this part, but had a gradient explosion while training MMMUBA.

Reflecting on this process, it was mainly because we were too hesitant in our topic and method selection, for example, as I said before, we wanted a deep forgery detector with good results too much. At the same time, our model implementation was optimized several times, and each optimization required reading new papers and writing new code.

All in all, through this project, our team learned a lot of new knowledge and methods, including coding and information search (table 1).

Table.1. Comparative analysis of the proposed approach with recent state-of-the-art systems.

Modality	T	V	A	**CMU-MOSEI** Proposed		**CMU-MOSI** Proposed
Uni-modal	√	×	×	76.75	78.23	80.18
	×	√	×	71.84	74.84	63.70
	×	×	√	70.94	75.88	62.10
Bi-modal	√	×	×	77.03	79.40	81.51
	√	×	×	76.89	79.74	80.58
	×	√	√	72.74	76.66	65.16
Tri-modal	√	√	√	77.64	79.80	82.31

We utilized the extensive capabilities of the large-scale model in our project by conducting quad classification. This involved utilizing real samples, audio faces that were both authentic and manipulated. By employing quad classification, we achieved an impressive accuracy rate of 88.2%. Additionally, in our dataset containing a total of 500 individuals, each person was assigned a unique identification (ID). We further leveraged the last layer of the neural network from the same model to extract sample features.

To proceed with the classification process, we employed k-means clustering, treating the features of the genuine samples as the cluster centers. Through this approach, samples with similar features were grouped together in clusters, with the center of each cluster serving as a representative of the ID class. This method yielded a classification accuracy rate of 54.17%. Moreover, we also experimented with using the dot product directly for classification, resulting in a slightly improved accuracy of 55.63%. These results demonstrate the effectiveness of our approach in accurately assigning individuals to their respective ID classes.

ACKNOWLEDGEMENT. All the authors contributed equally to this work and should be considered as co-first author.

REFERENCE

[1] Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2021). FakeAVCeleb: A novel Audio-Video multimodal deepfake dataset. *arXiv (Cornell University)*. http://export.arxiv.org/pdf/2108.05080

[2] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Neural Information Processing Systems*, *33*, 12449–12460. https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[3] Salvi, D., Liu, H., Mandelli, S., Bestagini, P., Zhou, W., Zhang, W., & Tubaro, S. (2023). A robust approach to multimodal deepfake detection. Journal of Imaging, 9(6), 122. https://doi.org/10.3390/jimaging9060122

[4] Ghosal, D., Akhtar, S., Chauhan, D. S., Poria, S., Ekbal, A., & Bhattacharyya, P. (2018). Contextual Inter-modal Attention for Multi-modal Sentiment Analysis. *Deepanway Ghosal , Md Shad Akhtar , Dushyant Chauhan , Soujanya Poria , Asif Ekbal and Pushpak Bhattacharyya*. https://doi.org/10.18653/v1/d18-1382

[5] Khalid H , Tariq S , Woo S S .FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset[J]. 2021.DOI:10.48550/arXiv.2108.05080.

[6] Sclaroff, S., Distante, C., Leo, M., Farinella, G. M., & Tombari, F. (2022b). *Image Analysis and Processing – ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*. Springer Nature.

[7] Sojka, P., Horák, A., Kopeček, I., & Pala, K. (2022). *Text, speech, and dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings*. Springer Nature.

[8] Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., & Busch, C. (2022). *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Springer Nature.