Unmasking Digital Deception using DeepFake

Shubhi Agarwal
210150016

Abstract

Deepfake technology, spanning text, audio, image, and video modalities, poses significant threats, enabling the creation of convincing fake media. This leads to widespread misinformation, political manipulation, identity theft, privacy violations, erosion of trust, and raises legal and ethical concerns, undermining individuals, society, and democratic processes.Several methods for detecting deepfakes and fake news have been proposed, focusing solely on single-modality forgery through binary classification. However, these approaches lack the capability to analyze and reason about subtle forgery traces across different modalities. Addressing this challenge, the paper [1] offers an innovative solution of Hierarchical Multi-modal Manipulation Reasoning Transformer (HAMMER) to authenticate multi-modal media and ground manipulated content, utilizing both text and visual modalities.This report discusses the attempt to implement the paper [1]

1 Introduction

Deepfake technology has revolutionized media manipulation, generating convincing yet entirely fabricated text, audio, and visual content. This poses a significant threat to public opinion, trust, and security. Detecting and grounding manipulation in news media is crucial for maintaining information integrity and enabling individuals to form informed opinions based on accurate representations of reality. Thus, there's a critical need for robust detection systems, regulations, and media literacy initiatives to mitigate its adverse impacts, emphasizing the necessity for effective detection mechanisms.

Unlike existing methods focused on single-modality forgery detection, DGM4 [1] requires simultaneous detection of forgeries in both image and text modalities, while also grounding manipulated elements like image bounding boxes and text tokens. To address this, the paper proposes Hierarchical Multi-modal Manipulation Reasoning Transformer (HAMMER), utilizing manipulation-aware contrastive learning and modality-aware cross-attention for comprehensive manipulation detection and grounding. Furthermore, an advanced model, HAMMER++, enhances semantic alignment through fine-grained contrastive learning. The paper also introduces the first large-scale DGM4 dataset and establishes a comprehensive benchmark for evaluation.

However, the dataset [1] is based on US News and to achieve effective detection of manipulation in Indian media, I have created a new dataset .

1.1 Contributions

- Construction of the first large-scale dataset: The paper [1] introduces a new dataset focusing on manipulation in human-centric news media.It featuring samples generated by various image and text manipulation approaches, with rich annotations facilitating detection and grounding of diverse manipulations.

- Proposal of a powerful Hierarchical Multi-modal Manipulation Reasoning Transformer (HAMMER) to address DGM4, integrating manipulation-aware contrastive learning and modality-aware cross-attention.
- Development of an advanced model, HAMMER++, incorporating Manipulation-Aware Contrastive Loss with Local View for more fine-grained cross-modal semantic alignment.

2 Related Works

DeepFake Detection: DeepFake Detection is the task of detecting fake videos or images that have been generated using deep learning techniques. Some Deepfake detection methods were based on using simple features. *Dural et al* [2] used a classical frequency domain analysis followed by a basic classifier to detect such fake face images Many scholars used multimodalities especislly audio-videos to detect deepfakes. *Chugh et al* [3] proposed detection of deepfake videos based on the dissimilarity between the audio and visual modalities, termed as the Modality Dissonance Score (MDS).*Wang et al* [4] used transformers to amplify both intra- and crossmodal forgery cues, thereby enhancing detection capabilities.*Liu et al* [5] proposed a novel approach dedicated to lip-forgery identification that exploits the inconsistency between lip movements and audio signals. Additionally, the development of large-scale datasets and benchmarking platforms has facilitated the evaluation and comparison of different detection methods.*Perov et al* presented DeepFaceLab [6], a deepfake framework providing the necessary tools as well as an easy-to-use way to conduct high-quality face-swapping.*Rossler et al* [7] examines the realism of state-of-the-art image manipulations, and how difficult it is to detect them, either automatically or by humans. To standardize the evaluation of detection methods, we propose an automated benchmark for facial manipulation detection.

Face Swapping: Face swapping in deepfakes involves replacing one person's face in a video or image with another person's face, often with remarkable realism. Face Forgery detection methods have evolved over time. Early replacement-based works as shown by *Bitouk et el* [8] , *Wang et al* [9] simply replace the pixels of inner face region. But, they are sensitive to the variations in posture and perspective.Then 3D-based works like [10] by *Lin et al* used a 3D model to deal with the posture or perspective difference. However the accuracy and robustness of 3D reconstruction of faces were unsatisfactory.*Yang et al* [11]proposed an attributes encoder for extracting multi-level target face attributes, and a new generator to adaptively integrate the identity and the attributes for face synthesis. Recently, GAN-based works by *Karshunova et al* [12], *Natsume et al* [13] have illustrated impressive results. But it remains challenging to synthesize both realistic and high-fidelity results.

Face Attribute Manipulation: Face attribute manipulation is a task to edit face attributes presented in an image, e.g. facial expression,emotion, age. Many methods utilize original images as inputs and try to reconstruct the finer details of the original images when modifying the interest areas of the input images. *GAtys et al* [14] used Convolution Neural Networks for this purpose. Some scholars like *Perarnau et al* [15] used GANs to reconstruct and modify real images of faces conditioning on arbitrary attributes. *Zhang et al* [16] used U-Net and GNN for inverting face attributes while of preserving finer details of the original face images.

3 Methodology

The methodology proposed in the paper involves the development of a model called HierArchical Multimodal Manipulation Reasoning Transformer (HAMMER) to address the challenges of

detecting and reasoning about multi-modal manipulations in media.

The proposed model, HAMMER [1], performs hierarchical manipulation reasoning, exploring multi-modal interaction from shallow to deep levels. The combined loss function used for HAMMER and HAMMER++ are:

$$L = L_{\text{MAC–G}} + L_{\text{IMG}} + L_{\text{MLC}} + L_{\text{BIC}} + L_{\text{TMG}} \tag{1}$$

$$L = L_{\text{MAC}--\text{G}} + L_{\text{IMG}} + L_{\text{MLC}} + L_{\text{BIC}} + L_{\text{TMG}} \tag{2}$$

Both the losses combine the contrastive losses for image-to-text, text-to-image, image-to-image, and text-to-text pairs.

Shallow Manipulation Reasoning:

Semantic Alignment: Image and text embeddings are aligned through Manipulation-Aware Contrastive Learning with both Global and Local views.

- Image and Text Encoding: The image (I) and text (T) are encoded into sequences of embeddings using self-attention layers and feed-forward networks in the Image Encoder (Ev) and Text Encoder (Et), respectively. The embeddings include a representation for the entire sequence ([CLS] token) and individual patches (vpat for images and ttok for text).
- Manipulation-Aware Contrastive Learning with Global View: This step aims to align image and text embeddings using cross-modal contrastive learning. A contrastive loss function is used to pull together embeddings of matched pairs while pushing apart embeddings of unmatched pairs. A specific variant, called manipulation-aware contrastive learning, emphasizes the semantic inconsistency caused by manipulations.
- Intra-Modal Contrastive Learning: To maintain semantic consistency within each modality, intra-modal contrastive learning is performed separately for images and text.
- Formulation of Manipulation-Aware Contrastive Loss (LMAC-G): This combines the contrastive losses for image-to-text, text-to-image, image-to-image, and text-to-text pairs, providing a comprehensive measure of semantic alignment between image and text embeddings.
- Manipulation-Aware Contrastive Learning with Local View: To address the limitations of global view contrastive learning, this step focuses on more fine-grained semantic alignment between local regions in images and text tokens. Patch embeddings are used to represent local features, and a contrastive loss function is applied to align them with the global embeddings.

Manipulated Image Bounding Box Grounding: The model grounds the manipulated regions in images by finding local patches that have inconsistencies with text embeddings.

- The model grounds the manipulated regions in images by finding local patches that have inconsistencies with text embeddings. This involves cross-attention between image and text embeddings to obtain patch embeddings containing image-text correlation. Additionally, attention is applied to aggregate spatial information from local image patches, allowing for more accurate manipulation detection by focusing on specific regions of interest.

Deep Manipulation Reasoning:

- *Modality-Aware Cross-Attention:* Text embeddings are further interacted with image embeddings through multiple cross-attention layers in Multi-Modal Aggregator, generating deeper aggregated embeddings.
- *Manipulated Text Token Grounding:* The model labels each token in the text as real or fake, which is treated as a multi-modal sequence tagging task. Cross-entropy loss is calculated to train the Token Detector.
- *Fine-Grained Manipulation Type Detection and Binary Classification:* The model detects four fine-grained manipulation types and performs binary classification. Multi-Label Classification Loss and Binary Classification Loss are calculated based on the aggregated multi-modal information.

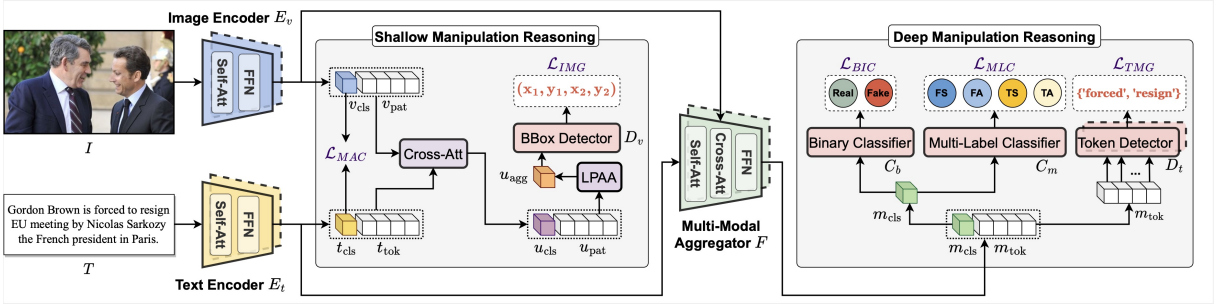


Figure 1: Methodology.

4 Database

The dataset DGM4 [1] is created and used by the authors which is a large-scale dataset designed for studying machine-generated multi-modal media manipulation. The dataset focuses specifically on human-centric news, acknowledging its significant public influence. Developed from the VisualNews dataset, it comprises a total of 230,000 news samples, encompassing 77,426 pristine image-text pairs and 152,574 manipulated pairs. The manipulated pairs encompass:

- 66,722 instances of Face Swap Manipulations (FS)
- 56,411 instances of Face Attribute Manipulations (FA)
- 43,546 instances of Text Swap Manipulations (TS)
- 18,588 instances of Text Attribute Manipulations (TA)

Furthermore, a portion of the manipulated images (1/3) and manipulated text (1/2) are combined to create 32,693 mixed-manipulation pairs.

Explaining further on each type of manipulation employed in the dataset:

- **Face Swap (FS) Manipulation**: The identity of the main character in an image is altered by swapping their face with another person's face. Two representative face swap approaches, SimSwap [17] and InfoSwap [18], are used. Bounding boxes of the swapped faces are annotated for grounding.

- **Face Attribute (FA) Manipulation**: This manipulation type alters the emotional expression of the main character's face while preserving their identity. Emotion labels are predicted for aligned faces, and then faces are manipulated towards the opposite emotion using GAN-based methods [19], [20]. Bounding boxes are also provided for these manipulations.
- **Text Swap (TS) Manipulation**: Text is manipulated by altering its overall semantic while preserving words related to the main character. Named Entity Recognition (NER) is used to extract the main character's name from the original text, and then a different text sample containing the same entity is retrieved and selected as the manipulated text. Tokens in the manipulated text are annotated for manipulation.
- **Text Attribute (TA) Manipulation**: This manipulation type involves altering the sentiment tendency of the text. Sentiment analysis is performed on the original text, and then sentiment words are removed and replaced to change the sentiment. The sentiment of the manipulated text is ensured to be flipped, and sentiment words added by the manipulation are annotated.

4.1 My dataset

The dataset I created consists of 15,000 images along with text pairs (the news captions) for Indian media. I have extracted them by scraping data from Indian Express. Some sample images are:

Further, exmaples of some captions I extracted are:

- Rahul Gandhi demanded an immediate provision of Rs 10,000 to the poor and an economic stimulus package for the MSMEs to help them revive their conditions.
- Government offices in the Union Territory had been working at 30 per cent staff strength in view of the Covid-19 pandemic.
- The Centre released advance relief of Rs 1,000 crore after PM Narendra Modi's visit to the cyclone-hit areas on May 22.
- There has been an increase in anti-militancy operations in the Valley since April.

Also, I swapped attribute features of some images using API [21] and [20]. For example :
Also, I have implemented text Attribute Manuplation in captions Examples were:

- Original: There has been an increase in anti-militancy operations in the Valley since April.
  Manipulated: There has been an decrease in anti-militancy operations in the Valley since April
- Original : People are protesting against the new law.
  Maniputlated: People are celebrating for the new law.

5 Implementation Details

5.1 Data Preprocessing

I have downloaded the DGM dataset provided by the authors and used it for the demo as the dataset I created did not have face swapped images and text swap manupulation. The preprocessing steps while creating my dataset included:

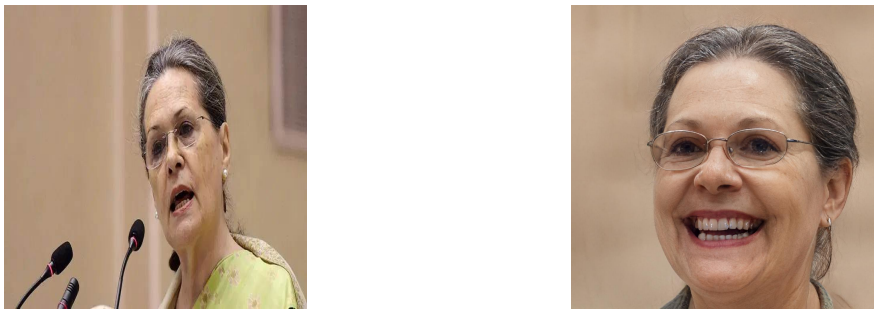
- At first, I scraped data from Indian Express and stored them. I had scrapped 20,000 images using Python libraries. Then I manually filtered required images that contained at leat 1 face.

(a) Image 1 (b) Image 2 (c) Image 3

(d) Image 4 (e) Image 5 (f) Image 6

(g) Image 7 (h) Image 8 (i) Image 9

Figure 2: Sample Images.



(a) Original (b) Face Attribute change

Figure 3: Sample images with face attribute change

- Then all the images were resized to the same dimension.
- Text Preprocessing : Text captions associated with images are preprocessed.It involve removing of various tags and punctuations and making them all in lower case.

5.2 Feature Extraction

- **Image Features:**
  - Extracted using a pretrained vision transformer model (`VisionTransformer`).
  - The image features represent high-level visual representations of input images and are typically embedded into a fixed-dimensional space.
- **Text Features:**
  - Extracted using a pretrained BERT-based model
  - These features capture contextual information from input text and are often represented as dense embeddings.
- **Class Labels:**
  - Fake class labels are extracted from the dataset annotations.
  - These labels indicate the ground truth classification of the input samples.

5.3 Classifier/Regressor

- **Matching Classifiers:**
  - These classifiers are used to determine the match between image and text features.
  - The matching classifier's output may indicate the similarity or alignment score between the features.
- **Bounding Box Regression Classifier:**
  - Used to predict bounding box coordinates, typically through regression.
  - Loss functions such as L1 loss and Generalized Intersection over Union (GIoU) loss are commonly used to train bounding box regression models.
- **Multi-label Classification Classifier:**
  - Used for tasks involving multi-label classification, where each sample can belong to multiple classes simultaneously.
  - Binary cross-entropy loss is used as the loss function.
- **Intra-Modal Global-to-Global (G2G) Classifier:**
  - This classifier operates within a single modality (e.g., image or text) and assesses the global features' similarity or alignment.
  - It is used for evaluating alignment within image or text data.
- **Inter-Modal Matching Classifier:**
  - Used to evaluate the alignment between image and text features.
  - Cross-modal matching classifiers aim to measure the compatibility or similarity between features from different modalities.
- **Token Matching Classifier:**
  - This classifier assesses the token-level alignment between image and text features.
  - It may be employed in tasks requiring fine-grained alignment or matching at the token level, such as token-level classification or alignment tasks.

6 Results

Upon loading the dataset given by the authors and cloning their repository and running it, I got precision 43.12 and recall 32.56.

7 Limitations

- **Data Quality and Quantity**: The performance of the model heavily relies on the quality and quantity of the training data. Limited or biased data can lead to poor generalization and performance degradation.
- **Generalization to Unseen Data**: The model may struggle to generalize well to unseen data, especially if the training data does not adequately represent the full diversity of the target domain.
- **Scalability**: Deploying such a model in real-world applications may pose scalability challenges, both in terms of computational resources required for inference and the model's ability to handle large volumes of data efficiently.
- **Dependency on Pretrained Models**: The model's performance is highly dependent on the quality and generalization capabilities of the pretrained vision transformers and BERT models. Any limitations or biases in these pretrained models can propagate to the final model's performance.

8 Conclusion

The proposed model, HAMMER, introduces a groundbreaking approach to detecting and grounding multi-modal media manipulation (DGM4). Unlike conventional methods that primarily target single-modal forgery detection, HAMMER integrates manipulation-aware contrastive learning and modality-aware cross-attention to tackle the intricacies of multi-modal manipulation. This novel methodology marks a significant advancement in the field, offering a more comprehensive solution to combatting fake media across various modalities.

Furthermore, the development of the first large-scale DGM4 dataset contributes substantially to the research landscape. Focused on manipulation within human-centric news media, this dataset includes samples generated through diverse image and text manipulation techniques. The dataset's rich annotations provide valuable resources for detecting and grounding a wide range of manipulations, further enhancing the effectiveness of the proposed model.

Methodologically, the paper proposes hierarchical manipulation reasoning, which explores multi-modal interaction from shallow to deep levels. By incorporating manipulation-aware contrastive learning and fine-grained cross-modal alignment in the loss function for HAMMER and HAMMER++, the model achieves superior manipulation detection and grounding capabilities. These methodological advancements underscore the model's efficacy in addressing the complexities of multi-modal media manipulation.

However, deploying the proposed model in real-world applications may encounter scalability challenges due to computational resource requirements and reliance on pretrained vision transformers and BERT models. Overcoming these limitations is essential to ensure the model's practical effectiveness and widespread adoption.

In conclusion, the study emphasizes the importance of the proposed approach in addressing the pervasive issue of multi-modal media manipulation. It underscores the need for continued research to overcome existing challenges and propel the field of fake news detection forward.

References

[1] R. Shao, T. Wu, J. Wu, L. Nie, and Z. Liu, "Detecting and grounding multi-modal media manipulation and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1,2,8, 4–7, 2024.

[2] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking deepfakes with simple features," *arXiv preprint arXiv:1911.00686*, 2019.

[3] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 439–447.

[4] R. Wang, D. Ye, L. Tang, Y. Zhang, and J. Deng, "Avt2-dwf: Improving deepfake detection with audio-visual fusion and dynamic weighting strategies," *arXiv preprint arXiv:2403.14974*, 2024.

[5] W. Liu, T. She, J. Liu, R. Wang, D. Yao, and Z. Liang, "Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes," *arXiv preprint arXiv:2401.15668*, 2024.

[6] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang *et al.*, "Deepfacelab: Integrated, flexible and extensible face-swapping framework," *arXiv preprint arXiv:2005.05535*, 2020.

[7] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[8] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: automatically replacing faces in photographs," in *ACM SIGGRAPH 2008 papers*, 2008, pp. 1–8.

[9] H.-X. Wang, C. Pan, H. Gong, and H.-Y. Wu, "Facial image composition based on active appearance model," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 893–896.

[10] Y. Lin, S. Wang, Q. Lin, and F. Tang, "Face swapping under large pose variations: A 3d model based approach," in *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 2012, pp. 333–338.

[11] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.

[12] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3677–3685.

[13] R. Natsume, T. Yatagawa, and S. Morishima, "Fsnet: An identity-aware generative model for image-based face swapping," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14*. Springer, 2019, pp. 117–132.

[14] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[15] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional gans for image editing," *arXiv preprint arXiv:1611.06355*, 2016.

[16] X. Tu, Y. Luo, H. Zhang, W. Ai, Z. Ma, and M. Xie, "Face attribute invertion," *arXiv preprint arXiv:2001.04665*, 2020.

[17] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2003–2011.

[18] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, "Information bottleneck disentanglement for identity swapping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3404–3413.

[19] T. Wang, Y. Zhang, Y. Fan, J. Wang, and Q. Chen, "High-fidelity gan inversion for image attribute editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 379–11 388.

[20] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2085–2094.

[21] [Online]. Available: https://replicate.com/