ENHANCING FAKE FACE DETECTION USING LEVERAGING HIERARCHICAL ATTENTION MEMORY NETWORKS FOR ROBUST AUTHENTICATION

Kiran B1, Dr. Nasreen Fathima2, Kavya P O3, Anil D4, Prajwal Hegde N5 and Dileep Kumar M J6

1 ATME College of Engineering, Mysuru, India.
2 ATME College of Engineering, Mysuru, India.
3 ATME College of Engineering, Mysuru, India.
4 CMR Institute of Technology, Bengaluru, India.
5 NMAM Institute of Technology, Nitte (Deemed to be University), Udupi, India.
6 NMAM Institute of Technology, Nitte (Deemed to be University), Udupi, India.

Abstract— In recent years, computer vision advancements have enabled the creation of convincing fake data, posing significant challenges for governments worldwide due to its association with disinformation and eroding trust. While Convolutional Neural Networks (CNNs) excel at detecting manipulated images within known parameters, they fall short in identifying undetectable manipulation strategies, leaving a critical gap in defense. To address this issue, we propose Hierarchical Attention Memory Network (HAMN) inspired by human brain social cognition to recognize fraudulent faces. This approach provides a foundation for generalized face change detection. Experimental results showcase HAMN's superior performance in identifying fake and fraudulent faces, promising to enhance computer vision's ability to counter deceptive practices. This advancement holds potential in combatting the proliferation of misleading visual content, contributing to information integrity and trustworthiness.

Index Terms — Fake & Fraudulent face detection, Neural memory networks, image and video forensics

I. INTRODUCTION

With social media now serving as many people's major information source and more than one hundred million hours of video being watched every day, false news has grown to be a serious danger to the society and to the democracy. The ease and seamlessness with which realistic false material may be created might have a significant influence on people's lives and seriously erode their faith in digital media, as the Deepfake app shown. Filtering out fake photos and videos is also a big problem for several applications, such biometric-based identification. People have been interested by the modification of picture, audio, and video material for a very long time. While editing photos using programs like Adobe Photoshop is extremely easy, working with audio and video may occasionally be very difficult. In the beginning, adding special effects to movies required frame-by-frame video editing. Nowadays, virtually everyone has access to a personal computer that is capable of editing videos in some way. While deep

neural networks and have achieved impressive levels of accuracy when identifying specific forms of image tampering attempts, their performance is frequently subpar when exposed to photos or videos that have been altered using undiscovered techniques [1].

Modern GANs may produce high-quality face pictures with a resolution of 1024, including PGGAN, SGAN, and MSGGAN. Changing facial characteristics on photos of actual faces, such as hair color, haircut, gender, expression, and others, is known as facial features modification. After defining settings, StarGAN and StarGANv2 can modify face characteristics automatically, while SC-FEGAN can do the same thing by having users create masks [2].

Deep fakes are being disseminated on social media platforms at an increasing rate, which leads to spamming and the dissemination of incorrect information. Imagine a convincing fake of our country's president declaring war on adjacent countries or of a well-known celebrity attacking their fans. To solve this problem, deep fake detection is essential. We present a novel deep learning-based approach to effectively distinguish AI-generated fake videos (also known as Deep Fake Ones) from real videos. It is essential to develop technologies that can spot fakes so that they can be found and prevented from spreading online. It is feasible to find generative methods like generative adversarial networks and encoder-decoders [3].

II. LITERATURE REVIEW

Taeb et al., [4] employed a method to compare the produced face areas and the regions around them to a predetermined Convolutional Neural Network model in order to find artifacts. Here, we observe two different anthropomorphic face types. They employ this method since they are aware that the deepfake algorithm can currently only produce low-resolution photos, which must subsequently be edited to match the faces that will be used as replacements in the original video. The time-based analysis of the clips was not accounted for in their methodology. Fard et al., [5] presented the face X-ray of an input face picture as a greyscale image that tells if the input image can be deconstructed into the blending of two images from distinct sources. It accomplishes this by displaying the boundaries of blending for a fake image and the lack of blending for a genuine one. Extensive testing reveals that while most deep fake or face forgery detection algorithms perform badly, face X-ray is nevertheless successful when used to detect forgeries produced using hidden face modification techniques.

Exposing Fake Faces Through Deep Neural Networks by Heo et al., [6] the performance of manipulation detection by merging content and trace feature extractors on the idea of a hybrid face forensics framework built on a convolutional neural network. Using a bespoke DeepFake dataset created by the authors and a public Face2Face dataset, the suggested framework was validated.

According to Sun et al., [7] uses the Gram-Net suggested architecture, which makes use of global image texture representations for reliable false picture identification, has the following benefits: It was shown that Gram-Net performed substantially better at recognizing fraudulent faces from GAN models not encountered during the training phase because it was more resistant to picture editing techniques such down sampling, JPEG compression, blur, and noise. Despite encouraging findings, it is still a long way from comprehending phony pictures created by GANs and enhancing fake face identification in actual situations.

III. PROBLEM STATEMENT

Since many years, visual effects have been used to present convincingly altered digital images

and videos, but recent advancements in deep learning have significantly increased the realism of false material and made it easier to make. These purportedly artificially created materials, sometimes referred to as "Deep Fakes". Artificial intelligence technology makes it simple to create Deep Fakes. The problem of finding these Deep Fakes, however, is considerable. Deep fakes have previously been used successfully in countless historical situations to incite political unrest, arrange terrorist acts, create revenge porn, blackmail people, etc. It is essential to recognize these deep fakes and prevent their propagation on social media as a result. The advancement in the use of an LSTM-based artificial neural network to detect deep fakes. Numerous face recognition techniques have been developed; however, they do have certain drawbacks, including Poor lighting, Pose estimation and Optimum feature selection, which are crucial for telling real from false pictures and videos [8].

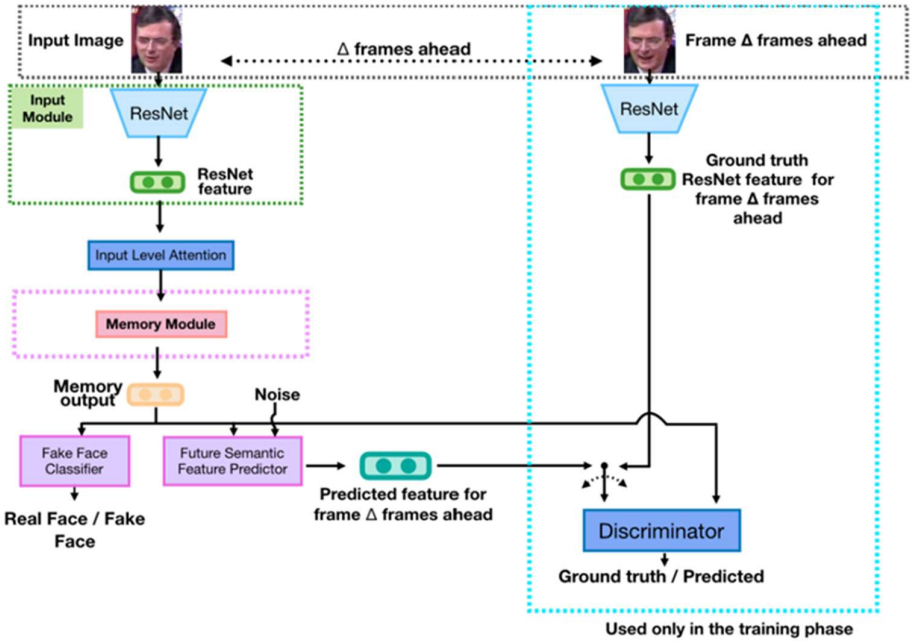

Figure-1: Hierarchical Attention Memory Network (HAMN) framework

IV. SYSTEM ARCHITECTURE

The key innovation of ResNet is the introduction of “residual connections”, which allow the network to learn identify mappings as well as more complex functions. This allows the network to have significantly deeper layers (up to 152 layers in the original ResNet paper) without suffering from the vanishing gradient problem that plagues very deep traditional CNNs. Due to its success in image classification tasks, ResNet has become a popular choice for many computers vision tasks and has been widely used in many successful models.

Figure 1 provides an overview of the suggested approach. Initially, we pre-train ResNet models on ImageNet to extract embeddings capturing facial semantics from input face images. These embeddings serve as the basis for subsequent steps:

1. Memory Search: We query a memory repository using the semantic embeddings.
2. Memory Output: The output from memory serves two main tasks:

- Classifying the authenticity of the face.
- Predicting potential future facial semantic embeddings.

These anticipated embeddings are processed by a discriminator for adversarial learning.

3. Sequential Representation: In the machine learning community, sequential representations are often generated from images. For instance, Ding et al., [9] employ a GRU layer to compile input features from images. To account for varying patch importance, we use an attention mechanism when combining local patches into a single query vector, denoted as qt.

4. Attention Mechanism: To calculate the significance of the current patch, f {k, t}, we pass the encoded patches through a single-layer MLP [10] to produce a representation v {k, t}:

$$v_{k,t} = \tanh(W_f f_k^{\leftarrow} + b_f)$$

Here, Wf and bf represent the weights and bias of the MLP. We also introduce a patch-level context vector, vf, implemented using a dense layer of the same dimension as v {k, t}. Its weights are initialized randomly and updated through backpropagation.

5. Inference with Attention: When creating inferences, we employ two levels of attention: patch level and memory level. The memory at time t-1, denoted as M{t-1}, contains L face embeddings, Ii, with each image consisting of K patches (p {i, k}). To avoid manual loss engineering for predicting future face embeddings, we utilize a generative adversarial framework [11].

6. GAN Framework: Generative Adversarial Networks (GANs) comprise a Generator (G) and a Discriminator (D) in a two-player game. G uses a random noise vector, zt, to generate future face embeddings, ^nt, and tries to deceive D. D, in turn, distinguishes between authentic and synthesized embeddings. This leads to the framework learning a task-specific loss automatically.

7. Conditional GAN: The Attention Mechanism generates embeddings without considering the current input. To address this, inspired by conditional GANs [12], G learns a conditional mapping from zt and the current memory output, rt, to ^nt:

8. Combined Objective: To address both tasks jointly, we combine the objective in Eq. 16 with the loss for classifying fake faces. Additionally, we apply L^2 regularization to the synthesized embeddings to encourage G to produce accurate embeddings [13]. The ultimate objective, V^*, is defined as:

9. Dynamic Memory Stack: As new frame embeddings are introduced, the memory stack L evolves during learning. During testing, L is initialized to its state at the end of the learning phase and adapts throughout testing [14]. Importantly, the test dataset is entirely distinct from the training dataset.

IV. IMPLEMENTATION

LSTMs are widely used in various applications such as natural language processing, speech recognition, stock market prediction, and video captioning. The ability of LSTMs to handle sequential data and maintain long-term dependencies makes it a powerful tool for processing complex data.

1) Face Forensics Dataset: From YouTube, the Face Forensics dataset was gathered. Videos with a quality of at least 480p with the tags "face," "newscaster," or "news program" are

available. The Face2Face approach is applied by the authors between two random films to produce altered faces. The dataset includes 150 validation films (76,309 pictures), 150 test videos and 704 training videos.

2) FaceForensics++ Dataset: An expanded version of FaceForensics, this dataset includes face changes made with FaceSwap and DeepFakes. FaceSwap is a lightweight editing application that uses face maker positions to copy the face region from one image to another. To remove occlusions, the original YouTube videos are manually inspected. The dataset includes 1,000 unaltered movies and 3,000 videos that have been altered using the Face2Face, FaceSwap, and DeepFake algorithms (1,000 for each category). We choose 720 movies for training, 140 for validation, and 140 for testing.

3) FakeFace in the Wild (FFW) Dataset: The FFW dataset covers a wide range of false material produced using computer graphics, GANs, human and automatic tampering techniques, and their combinations. It is built using a set of publicly available YouTube videos. As a result, it offers a perfect environment for assessing the generalizability of the suggested technique under a variety of manipulations. Videos range in length from 2 to 74 seconds and have a minimum 480p resolution. 150 genuine face recordings from FaceForensics are included in the FFW dataset in addition to these 150 fake movies.

The Viola-Jones face detector is employed subsequently to locate faces. For the task of predicting future frames, the input-output pairing involves every 20th frame and its corresponding frame occurring 15 frames in advance. This approach likely helps in modeling temporal dependencies and forecasting future video frames efficiently. The dataset is balanced by choosing the frames for extraction so that each movie has an equal number of samples. The secret state dimension is set to 300 for the whole GRU. The no. of patches, K = 196, and L = 200 (the hyper parameters memory length), are empirically assessed.

The assessments use the NTM, DMN, TMN as three benchmark memory modules for comparisons. In order to provide a fair comparison, we trained these techniques using the identical ResNet features that the proposed approach uses. We also set the NTM, DMN, and TMN modules' LSTM hidden state dimensions to 300 and their memory lengths to 200. Experimental evaluation determines the extraction depth of the TMN, which is set at 3. These memories are trained using binary cross entropy loss and supervised learning to directly categorize the input picture. Using the Face Forensics dataset, the suggested method's capacity to recognize facial impersonations is evaluated.

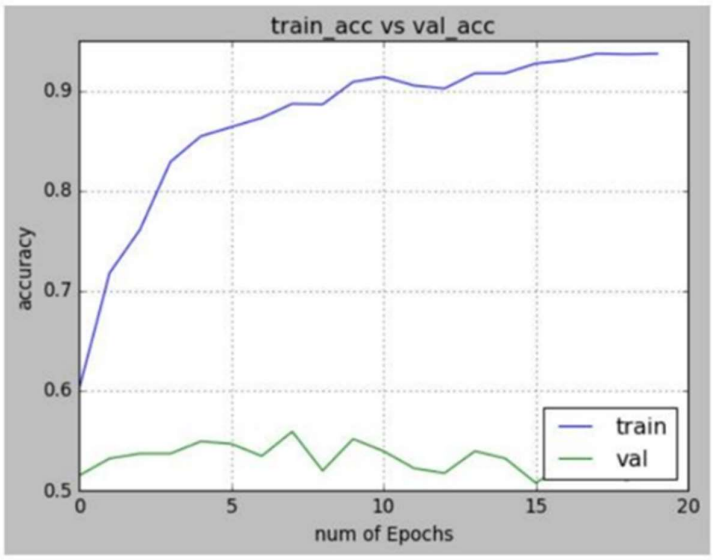

Figure-2: Accuracy comparison between training and validation for MobileNetV2

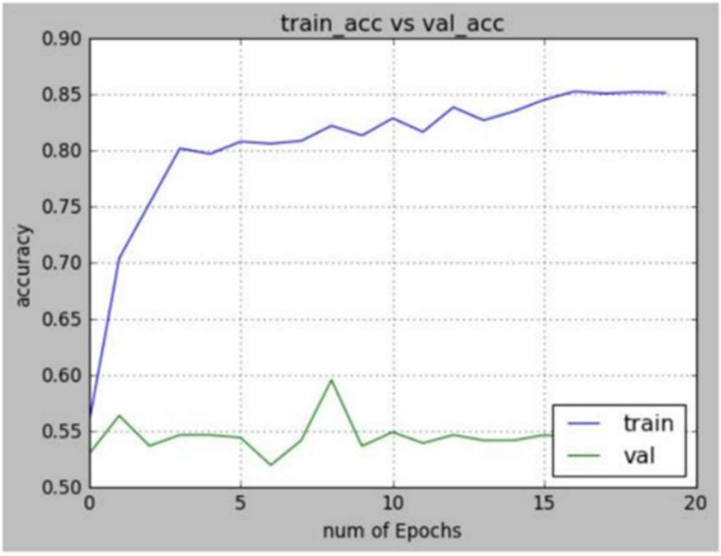

Figure-3: Accuracy comparison between training and validation for VGG16 model

V. RESULTS

The results of the experiment (study) to investigate the performance of the model in detecting the fakeness in the images or the videos provided with the training models are described briefly in this section. The Figures 2 and 3 are showing the loss and accuracy for training data and validation data for MobilenetV2 and VGG16 models respectively. It observed that has the number of epochs and batch sizes are increased depending upon the quality of the image data available the performance of the model also increases. So, it is said that higher the quality of image and higher the epochs and batch sizes will increase the prediction capacity of the model (increases the accuracy of the model). And also keeping the note that the accuracy also depends on the type of the data it is trained on (Fake images or Real images).

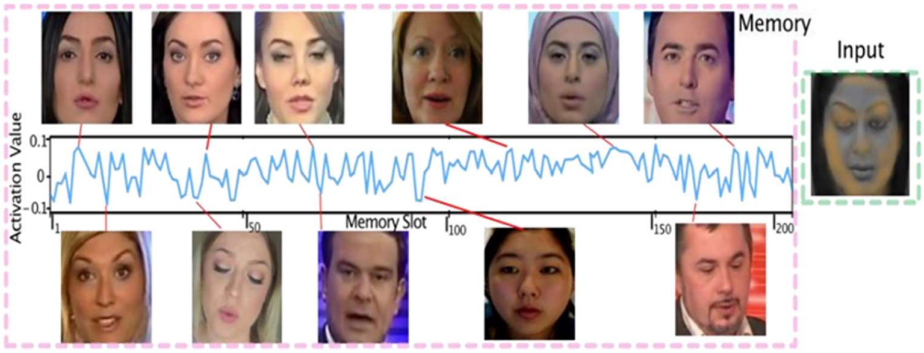

Figure-3: Visualization of memory activation

The visual representation of the activation values for the memory's content for a sample input is shown in Figure-3. The input has the attention highlighted in yellow (a brighter intensity corresponds to higher activations.). L = 200 means that there are 200 memory slots, which we refer to as l1 to l200. Also displayed which input picture embeddings are kept at each specific memory index for various peaks and troughs in the memory activation. The analysis of the patch level attention given to various facial areas for sample photos is shown in Figure 4. then utilizing the activations to fill a 2D heat map. Upscaling the heat map to match the original picture proportions with K = 14 14. There is just an approximate correlation as a consequence. Brighter intensity values equate to higher activations, and the activation values are indicated in yellow. The eye, mouth, and cheek areas in Figure 4 have a high degree of attention to gauge the validity of the image. These activation graphs demonstrate the significance of hierarchically recording facial features, while maintaining their uniqueness, mapping long-term relationships, and facilitating transferability between various types of assaults.

Figure-4: Patch level attention

V. CONCLUSIONS

The study introduces an innovative architecture called the Hierarchical Attention Memory Network (HAMN) designed for the precise detection of fake and fraudulent faces. Drawing inspiration from the human brain's social perception and cognitive functions, HAMN exhibits a distinct advantage by being adaptable to a wide range of undetectable face alteration techniques. Its primary objective is to differentiate authentic faces from counterfeit ones by effectively predicting the temporal changes observed in faces. This prediction is achieved through a hierarchical propagation of learned knowledge within a memory structure, which comprehensively captures both patch and image-level semantics. HAMN's significance lies in

its ability to model hierarchical knowledge, a key feature demonstrated through visual evidence. This visual representation showcases how the memory component accesses stored information and relates it to the current input, underlining HAMN's capacity to integrate facial concepts and grasp their temporal evolution.

VI. REFERENCES

[1] Shahzad, Hina Fatima, et al. "A Review of Image Processing Techniques for Deepfakes." Sensors 22.12 (2022): 4556.
[2] Chen, Yueqiao, et al. "EC-GAN: Emotion-Controllable GAN for Face Image Completion." Applied Sciences 13.13 (2023): 7638.
[3] Ikram, Sumaiya Thaseen, Shourya Chambial, and Dhruv Sood. "A performance enhancement of deepfake video detection through the use of a hybrid CNN Deep learning model." International journal of electrical and computer engineering systems 14.2 (2023): 169-178.
[4] Taeb, Maryam, and Hongmei Chi. "Comparison of deepfake detection techniques through deep learning." Journal of Cybersecurity and Privacy 2.1 (2022): 89-106.
[5] Fard, Ali Pourramezan, et al. "Sagittal Cervical Spine Landmark Point Detection in X-Ray Using Deep Convolutional Neural Networks." IEEE Access 10 (2022): 59413-59427.
[6] Heo, Young-Jin, Woon-Ha Yeo, and Byung-Gyu Kim. "Deepfake detection algorithm based on improved vision transformer." Applied Intelligence 53.7 (2023): 7512-7527.
[7] Sun, Ke, et al. "An information theoretic approach for attention-driven face forgery detection." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
[8] Suratkar, Shraddha, et al. "Deep-fake video detection approaches using convolutional–recurrent neural networks." Journal of Control and Decision 10.2 (2023): 198-214.
[9] Ding, Henghui, et al. "Vision-language transformer and query generation for referring segmentation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[10] An, Tai, et al. "TR-MISR: Multiimage super-resolution based on feature fusion with transformers." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2022): 1373-1388.
[11] Chen, Tianlong, et al. "Learning to optimize: A primer and a benchmark." The Journal of Machine Learning Research 23.1 (2022): 8562-8620.
[12] Talasila, Vamsidhar, and M. R. Narasingarao. "Optimized GAN for text-to-image synthesis: Hybrid whale optimization algorithm and dragonfly algorithm." Advances in Engineering Software 173 (2022): 103222.
[13] Sharma, Harshad, and Smita Das. "A brief study of generative adversarial networks and their applications in image synthesis." Multimedia Tools and Applications (2023): 1-31.
[14] Guo, Zizheng, et al. "A timing engine inspired graph neural network model for pre-routing slack prediction." Proceedings of the 59th ACM/IEEE Design Automation Conference. 2022.