AI modified Face Detection for fake news recognition

Arjun Menon

Department of Computer Science and Engineering

Amrita School of Engineering, Bengaluru

Amrita Vishwa Vidyapeetham Amrita University, India

[bl.en.u4aie20003@bl.students.amri](mailto:bl.en.u4aie20003@bl.students.amri) [ta.edu](mailto:BL.EN.U4AIE20006@bl.students.amrita.edu)

Peeta Basa Pati

Department of Computer Science and Engineering

Amrita School of Engineering, Bengaluru

Amrita Vishwa Vidyapeetham Amrita University, India

Bp\_peeta@blr.amrita.edu

Gaurang Chaudhary Department of Computer Science

and Engineering

Amrita School of Engineering, Bengaluru

Amrita Vishwa Vidyapeetham Amrita University, India

[bl.en.u4aie](mailto:BL.EN.U4AIE20006@bl.students.amrita.edu)[20014@bl.students.amri](mailto:20014@bl.students.amri) [ta.edu](mailto:BL.EN.U4AIE20006@bl.students.amrita.edu)

Suja P

Department of Computer Science and Engineering

Amrita School of Engineering, Bengaluru

Amrita Vishwa Vidyapeetham Amrita University, India

P\_suja@blr.amrita.edu

Mudragiri Nitish Narayan

Department of Computer Science

and Engineering

Amrita School of Engineering, Bengaluru

Amrita Vishwa Vidyapeetham Amrita University, India

[bl.en.u4aie](mailto:BL.EN.U4AIE20006@bl.students.amrita.edu)[20035@bl.students.amri](mailto:20035@bl.students.amri) [ta.edu](mailto:BL.EN.U4AIE20006@bl.students.amrita.edu)

Abstract

Deepfake is a form of content in which the face of one person is swapped onto the body of another, in the attempt to create a false narrative. They can even be created from scratch using the help of GAN networks. The misuse of deepfake is rampant across the internet and to combat it various technologies have been developed. In this paper the aim is to successfully distinguish between a still image of a fake face or real face, with high accuracy and f1-score, to help combat the misinformation spread by these technologies. Our models have been trained in 200,000 images and all give accuracy and f1-score of 96% +. The models have been tested on sites known for spreading fake news and shows promising results.

Abbreviation and Acronym

GAN- Generative Adversarial Network

AI- Artificial Intelligence

Np- NumPy

INTRODUCTION

In the modern era where youth spend a lot of time on social media, the possibility of consuming fake content created with malicious intent becomes high. With the help of AI tools and Deepfakes, the line between reality and fiction becomes harder to differentiate with each passing day. Deepfakes are posing an ever-increasing threat to the safety and security of individuals with their entire body being captured and manipulated from a few photos or videos. The possibility of misuse is high, from face swapping to attribute manipulation and even creating completely fictional scenarios with realistic looking actors to spread misinformation. A common example seen in the present is with the face of celebrities and other influential people being manipulated to match the facial expressions and lip movement of them speaking a completely different script. This paired with a speech synthesizer, trained with the person’s voice, results in a realistic video with the actor speaking whatever an individual wishes, while also being able to trick most of the population into believing that this is a genuine video. To combat the misinformation generated from such videos, neural networks are trained to detect whether there is any AI manipulation in a picture or video.

Many websites on the internet use pictures of Fake people generated using GANs, along with a fake story to push a false narrative, whether it be in the form of a fake review of a product, or a push for propaganda, the former being the most prevalent. The older population often fall for such tricks and accidentally click on the links generated, falling victim to malware or phishing attempts. Without any moving video, or audio, it becomes increasingly difficult for an individual to distinguish between a real face and a well created GAN fake face. Although the younger generation may be less likely to click on any suspicious links, they are still equally likely to fall for fakes stories or reviews, since the faces can be indistinguishable.

In this paper, the aim is to figure out if a face is real or has been manipulated or generated with A.I, without the use of video or audio, to combat the problem highlighted above. For the dataset 20,000 AI generated faces and 20,000 real faces have been collected. 3 models, a custom model, AlexNet and Inception will be developed, trained, and compared with each other, with the aim to successfully distinguish between the fake and real face.

LITERATURE SURVEY

A survey was conducted [1-2] that goes over the technical aspects required for deepfake detection, datasets used, and papers published in the span of 2018-2020. The paper [1] provides the technical knowledge of architecture and the various examples of deepfake creation, and challenges faced in deepfake detection. The survey [2] goes over 4 approaches to deepfake detection and the outcomes of each method. It concluded that deep learning is the most widely used method in detection and yielded better results. A 2-feature extraction process [3] paired with a CNN model gave the highest accuracy score of 99.96% and f1 score of 99.99%. For GAN generated faces [4] focusing on inconsistencies on the eyes by performing iris region extraction was proposed. The model was able to effectively learn from an imbalanced dataset. Different methods to check for mistakes generated by AI such as face edge bands [12], resolution [17], full face recognition [21] and eye blinking [22].

MODEL ARCHITECTURE

Chart, diagram

Description automatically generated

Fig 1) CNN model architecture

Diagram

Description automatically generated

Fig2) AlexNet architecture

Data Description

The dataset created contains 40,000 images. 20,00 images are AI generated images obtained from StyleGan model and Nvidia. 20,000 images of real faces have been taken from the UTKFace dataset. The faces contain pictures of people of all ethnicities ranging from the age of 0 to 116. The dataset contains pictures of individuals in various poses, faces and lightings.

For classification, the dataset was split into train, validation, and test. 10% of the data was used for testing, of which it was further split into 90:10 for testing and validation in batches.

A collage of a person

Description automatically generated with medium confidence

Fig 3) Fake faces

A picture containing text, bunch, different, group

Description automatically generated

Fig 4) Real faces

Methodology

First a folder was created, containing the images in separate sub-folders, named Fake and Real. 38,000 images, 19,000 from each were put into the subfolder, leaving the rest for testing and validation. From the 38,000 images, 2000 images, 1000 for each face type were copied into a new folder for preliminary model testing. The main dataset of 38,000 was augmented into 190,000 using a python code. Another python code was used to rename all the images, to prevent the model from getting biased by being exposed to similar pictures, as the initial dataset was structured according to age.

For the main code, first it investigated the folder path to identify the number of subfolders, these subfolders will become the different classes for the network to predict. An empty list for faces, that goes through each face type, and creates a list of all face types and the path to it. This list is then converted to a DataFrame.

For different models, different input size of images is needed, so we need to resize all the images appropriately. Since the dataset is too large to be loaded into RAM, a generator was built to load the images in batches, resize them and store them in a resized folder. The labels are stored in an array batch\_labels to remember which image belongs to which class and will be used later for training the model.

The batch size was set to 1900, and a for loop in range of 101 was used, to ensure all images are resized in case of an error.

By storing the resized images in the resized folder, the entire block of code needed to be run only once, allowing more time to tune the model.

Since the dataset is large, and RAM is not enough to store it, and to reduce computation time, a .h5 file having the np.array() form of the images was created and stored. Doing so removes the need of running a generator and batches every time the model needed to be tuned. The input to this is the resized folder. The data was stored in float16 format for efficient storage and normalized in the range of 0 to 1.

The stored labels are converted into a one-hot-encoded form as the networks require the information in numerical form.

Depending on om the model, the appropriate model was built using keras sequential, model was compiled with adam as optimizer, and categorical cross-entropy as loss function. A summary was printed to see output of each stage.

For validation and testing, the separate dataset was loaded into a DataFrame, resized, converted into np.array, normalized and shuffled and split for respective tasks. This was stored in RAM, as there was enough for the smaller dataset.

To train the model, a mini-batch approach was adopted. The images are retrieved form the hdf5 file, and the corresponding labels are retrieved from the first one-hot encoding. To determine whether the input is real or fake, we had to use the position of the data to our advantage. The first 95,000 arrays will be fake, and the next 95,000 will be real. When we take the batch size of 190, it will be split into 2 parts of 95, and using indexing, we can train on 95 fake images and 95 real images in one batch. The fake and real parts are concatenated together and trained using model train on batch. The model is trained on 500 batches. Training loss is appended into variable history, validation loss is recorded after each batch by testing model on validation set. After training is complete, a plot of training loss, validation loss and validation accuracy are printed, to check if model is underfit, overfit or regular fit.

Prior to running the final model, the initial models were developed and tested on the 2000 image subset. Using data augmentation, the number of images were increased to 8000, and trained with model.fit() over 10 epochs.

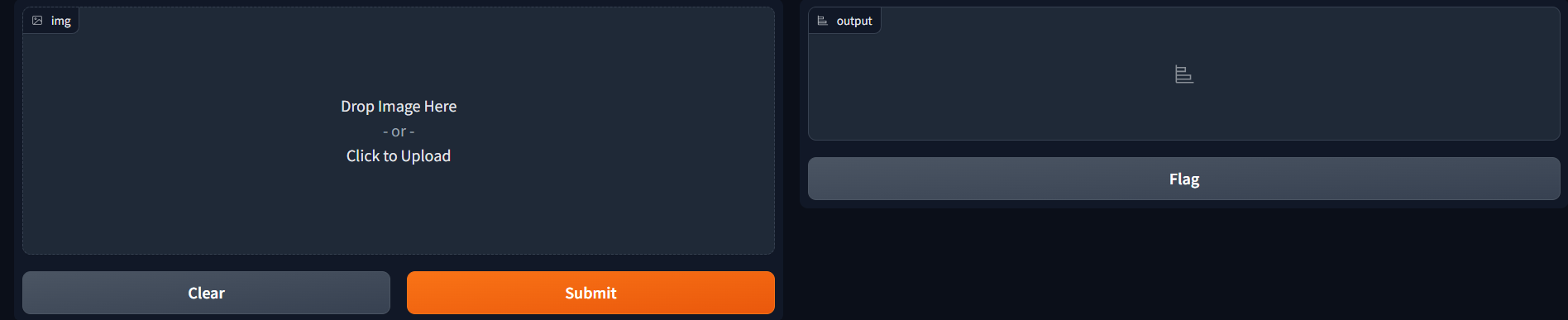


Fig 5) figure of GUI when first launched.

For developing the GUI part of the model, gradio was used. The entire model had its weights stored in a .h5 file, to be called again without re-running the model. After loading the pretrained model, using gradio a simple GUI to allow a user to manually select an image, and pass it through the model to test if it is real or fake was developed. The image selected first must be resized into the size of the input of the model, this was done with a function call automatically. When the model gives the prediction for the image, it displays on the side how much percentage it believers the image is fake, and how much it believes it is real, due to the SoftMax function used for activation. The GUI also has a clear option, that allows the user to clear the image and submit a new image for testing. The pipeline is effective and usually gives a response in less than a second.

Results

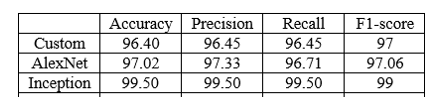


Fig 6) After regularization on the 8000-image dataset.

A picture containing text, screenshot, plot, diagram

Description automatically generated

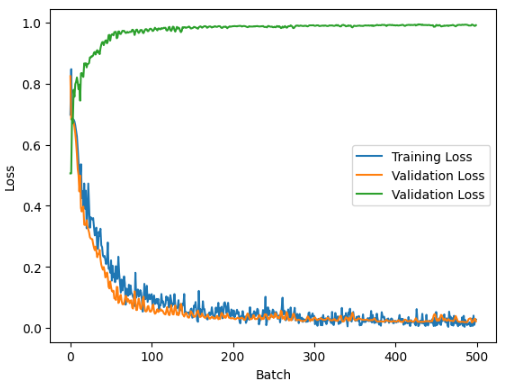
Fig 7) training loss, validation loss, and validation accuracy for AlexNet

Fig 8) training loss, validation loss and validation accuracy for Custom model

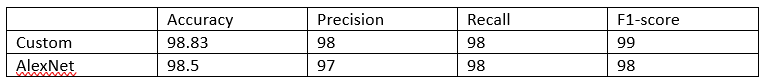


Fig 9) tabulated results on the 200,000-image dataset

Testing on various sites notorious for using fake faces ands stories, we were able to successfully capture that the face was fake.

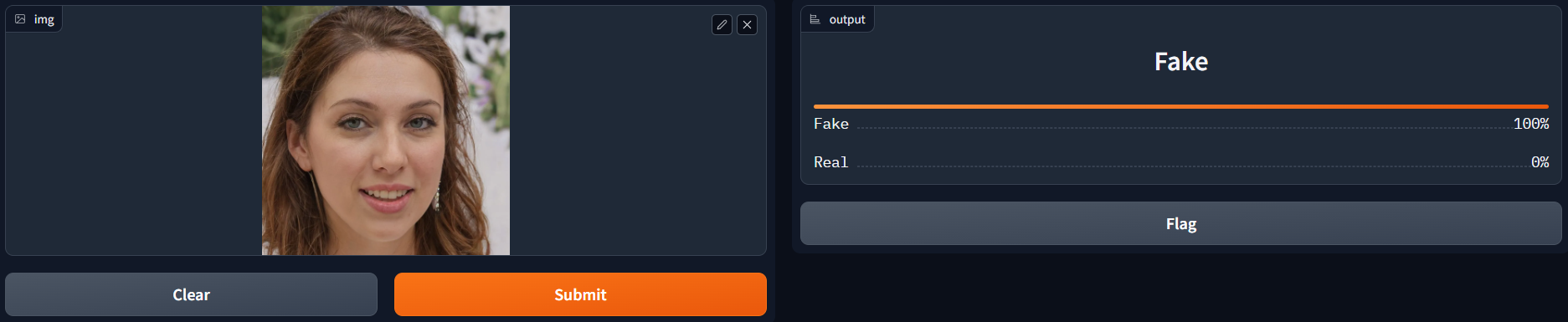


Fig 10) captured a fake face.

It can also tell when a face is real, as shown below.

A screenshot of a computer

Description automatically generated with medium confidence

Fig 11) real face detected.

Future Works

Inception model must be run on GPU with parallelization support, models should be tested on a dataset with multiple classes and results should be tabulated. Creating a more diverse dataset is also being investigated. Creating a browser add-on that automatically scans the webpage for fake faces and alerts the user is the goal.

CONCLUSION

A dataset of 40,000 fake and real images were collected and pre-processed. Preliminary development of models and testing was successfully done with the smaller subset of images after augmentation. Training the model on the large dataset after resizing, augmenting, renaming, and creating a .h5, trained in batches of 380, gave good results. Real world preliminary testing on sites known for using fake faces showed promising results.

REFERENCES

1. A. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," in IEEE Access, vol. 10, pp. 18757-18775, 2022, doi: 10.1109/ACCESS.2022.3151186.
2. M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 25494-25513, 2022, doi: 10.1109/ACCESS.2022.3154404.
3. E. Kim and S. Cho, "Exposing Fake Faces Through Deep Neural Networks Combining Content and Trace Feature Extractors," in IEEE Access, vol. 9, pp. 123493-123503, 2021, doi: 10.1109/ACCESS.2021.3110859.
4. H. Guo, S. Hu, X. Wang, M. -C. Chang and S. Lyu, "Robust Attentive Deep Neural Network for Detecting GAN-Generated Faces," in IEEE Access, vol. 10, pp. 32574-32583, 2022, doi: 10.1109/ACCESS.2022.3157297.
5. S. Guefrechi, M. B. Jabra and H. Hamam, "Deepfake video detection using InceptionResnetV2," 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sfax, Tunisia, 2022, pp. 1-6, doi: 10.1109/ATSIP55956.2022.9805902.
6. B. Malolan, A. Parekh and F. Kazi, "Explainable Deep-Fake Detection Using Visual Interpretability Methods," 2020 3rd International Conference on Information and Computer Technologies (ICICT), San Jose, CA, USA, 2020, pp. 289-293, doi: 10.1109/ICICT50521.2020.00051.
7. K. Remya Revi, M. M. Isaac, R. Antony and M. Wilscy, "GAN-generated Fake Face Image Detection using Opponent Color Local Binary Pattern and Deep Learning Technique," 2022 International Conference on Connected Systems & Intelligence (CSI), Trivandrum, India, 2022, pp. 1-7, doi: 10.1109/CSI54720.2022.9924077.
8. Qurat-ul-ain, N. Nida, A. Irtaza and N. Ilyas, "Forged Face Detection using ELA and Deep Learning Techniques," 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST), Islamabad, Pakistan, 2021, pp. 271-275, doi: 10.1109/IBCAST51254.2021.9393234.
9. A. Kawabe, R. Haga, Y. Tomioka, Y. Okuyama and J. Shin, "Fake Image Detection Using An Ensemble of CNN Models Specialized For Individual Face Parts," 2022 IEEE 15th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), Penang, Malaysia, 2022, pp. 72-77, doi: 10.1109/MCSoC57363.2022.00021.
10. A. Mitra, S. P. Mohanty, P. Corcoran and E. Kougianos, "Detection of Deep-Morphed Deepfake Images to Make Robust Automatic Facial Recognition Systems," 2021 19th OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 2021, pp. 149-154, doi: 10.1109/OCIT53463.2021.00039.
11. J. Wang, Y. Qi, J. Hu and J. Hu, "Face forgery detection with a fused attention mechanism," 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA), Changchun, China, 2022, pp. 722-725, doi: 10.1109/CVIDLICCEA56201.2022.9824499.
12. Z. Deng, B. Zhang, S. He and Y. Wang, "Deepfake Detection Method Based on Face Edge Bands," 2022 9th International Conference on Digital Home (ICDH), Guangzhou, China, 2022, pp. 251-256, doi: 10.1109/ICDH57206.2022.00046.
13. A. Aggarwal, S. Wadhwa, P. Gupta, N. Anand and R. Kushwah, "IsSwap: Deep Fake Detection," 2021 7th International Conference on Signal Processing and Communication (ICSC), Noida, India, 2021, pp. 194-199, doi: 10.1109/ICSC53193.2021.9673225.
14. C. -C. Hsu, C. -Y. Lee and Y. -X. Zhuang, "Learning to Detect Fake Face Images in the Wild," 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018, pp. 388-391, doi: 10.1109/IS3C.2018.00104.
15. S. Ramachandran, A. V. Nadimpalli and A. Rattani, "An Experimental Evaluation on Deepfake Detection using Deep Face Recognition," 2021 International Carnahan Conference on Security Technology (ICCST), Hatfield, United Kingdom, 2021, pp. 1-6, doi: 10.1109/ICCST49569.2021.9717407.
16. V. Jolly, M. Telrandhe, A. Kasat, A. Shitole and K. Gawande, "CNN based Deep Learning model for Deepfake Detection," 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1-5, doi: 10.1109/ASIANCON55314.2022.9908862.
17. A. Rahman et al., "Short And Low Resolution Deepfake Video Detection Using CNN," 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC), Hyderabad, India, 2022, pp. 259-264, doi: 10.1109/R10-HTC54060.2022.9929719.
18. I. Kusniadi and A. Setyanto, "Fake Video Detection using Modified XceptionNet," 2021 4th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2021, pp. 104-107, doi: 10.1109/ICOIACT53268.2021.9563923.
19. M. Kshirsagar, S. Suratkar and F. Kazi, "Deepfake Video Detection Methods using Deep Neural Networks," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 27-34, doi: 10.1109/ICICICT54557.2022.9917701.
20. T. -A. To, H. -C. Luong, N. -T. Nguyen, T. -T. Nguyen, M. -T. Tran and T. -L. Do, "Deepfake Detection using EfficientNet: Working Towards Dense Sampling and Frames Selection," 2022 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam, 2022, pp. 612-617, doi: 10.1109/RIVF55975.2022.10013900.
21. K. Feng, J. Wu and M. Tian, "A Detect method for deepfake video based on full face recognition," 2020 IEEE International Conference on Information Technology,Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 2020, pp. 1121-1125, doi: 10.1109/ICIBA50161.2020.9277303.
22. Y. Li, M. -C. Chang and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630787.