Deep Fake Detection

Daksh Baveja
Department Of Computing Technologies
SRM Institute Of Science & Technology, Kattankulathur
Chennai, Tamil Nadu
db4502@srmist.edu.in

Yatharth Sharma
Department Of Computing Technologies
SRM Institute Of Science & Technology, Kattankulathur
Chennai, Tamil Nadu
ys5413@srmist.edu.in

Dr. Nagadevi S.
Department Of Computing Technologies
SRM Institute Of Science & Technology, Kattankulathur
Chennai, Tamil Nadu
nagadevs@srmist.edu.in

***Abstract*—The following paper considers an in-depth study of face detection and classification using a pre-trained VGG16 model with a prime focus on separating real from fake facial images. Face detection is a very fundamental task in computer vision and of key importance in various security- and biometric identification-related applications, social media, and so on, in which the above-mentioned Dortania et al. findings will find their use. The idea is to use transfer learning by tuning an already trained VGG16 that was developed for large-scale image classification to do well in a specific task of face authenticity verification.**

For this purpose, we constructed a custom dataset with images labeled either ’real’ or ’fake’, sourced from different environments to make it diverse and hence robust. The dataset was then preprocessed by face detection using Haar cascades, resizing, normalization, and augmentation to increase the model’s capacity for generalization. This dataset was trained as well as tested on the modified VGG16 model, where only one fully connected layer at the end was changed to give an output in two classes—one for the real faces and another for the fake ones.

Model performance was ascertained through training loss and accuracy in the training phase. For the 30 epochs of training, the model achieved very good training accuracy. Further performance fluctuation analysis at different epochs used detailed plots of the loss and accuracy. Testing validates further that the model is robust, having a high testing accuracy to ensure the model generalizes on unseen data.

Our results show the effectiveness of transfer learning using VGG16 in face classification, where accuracy was high for the classification of real and fake faces. Thus, this study not only demonstrates the potential of pre-trained deep models in specialized applications but also shows the proper quality of the dataset and its preprocessing towards the attainment of optimal model performance. This trained model is, therefore, deployable in every real-world application where verification of faces is very important, bringing in a reliable tool for improving security and authenticity in digital relations.

***Index Terms*—deep fake, detection, artificial intelligence, machine learning, digital forensics**

I. Introduction

The paper is concerned with face detection and classification based on a pre-trained VGG16 model for extracting features that can discriminate a real face from a fake one. Face detection is a facet of computer vision critical to many applications, such as security, biometric identification, and social media. The present research is grounded in transfer learning by fine-tuning the VGG16 model specifically for face authenticity verification; this model was initially developed for large-scale image classification.

We created a custom dataset labeled either ’real’ or ’fake’, sourced from different environments to ensure robustness. Preprocessing included face detection using Haar cascades, resizing, normalization, and augmentation to increase the model’s generalization ability. The modified VGG16 model was trained and tested on this dataset, with changes in the final fully connected layer to output two classes: real and fake faces.

Model performance was measured based on training loss and accuracy over 30 epochs. High training accuracy was observed, and detailed plots of loss and accuracy showed performance trends. Testing proved the model to be robust, achieving high accuracy against unseen data.

The study demonstrates the effectiveness of VGG16 in transfer learning for face classification. It also highlights the necessity of a high-quality and well-preprocessed dataset for building an effective model. The trained model is readily applicable to real-world applications requiring face verification, enhancing safety and authenticity in digital interactions.

II. Literature Survey

The last couple of years have seen tremendous growth in the realm of face detection and classification, where various studies have been made on different approaches with respect to accuracy and efficiency. Deep learning approaches are one of the most challenging techniques used today in face detection and classification since they learn hierarchical features from raw images, basically through Convolutional Neural Networks. Among them, a model developed by Simonyan and Zisserman known as VGG16 has been hailed in terms of depth and efficiency in image classification tasks.

The concept of transfer learning has been highly popular, and pre-trained models over large datasets like ImageNet are fine-tuned for specific tasks. This exploits the rich feature representations learnt from large amounts of data, hence eliminating the presence of large datasets for training and computational resources. Transfer learning has been successfully applied to face detection and recognition tasks in a number of works, showing improved performance with reduced training times.

Inevitably, face detection techniques moved from the early schemes using Haar cascades and Local Binary Patterns to more advanced deep learning methods. Viola and Jones introduced another initial successful technique of real-time face detection with simple features and cascaded classifiers. Again, their performance is normally restricted to applications having complex scenarios like varying lighting conditions, partial occlusions, and facial expressions.

In contrast, deep learning-based methods, notably Single Shot MultiBox Detector and Region-based Convolutional Neural Networks, have shown admirable improvements. This is basically because of the inherent characteristics that integrate feature extraction and classification into one framework and provide much better accuracy and robustness to diverse challenges.

Distinguishing real from fake faces became a very relevant task after the appearance of advanced techniques for generating faces, like Generative Adversarial Networks. To this regard, several studies have investigated a series of CNN architectures, namely VGG16, ResNet, and Inception, to solve the problem at hand. Those models have recorded striking success in bringing out fine differences between the genuine and manipulated image when fine-tuned with the appropriate datasets.

In summary, deep learning models—especially via transfer learning—have done much to change face detection and classification. Among them stands one of the most outstanding: VGG16 for large-scale image classification and transfer learning tasks. From traditional approaches to state-of-the-art deep learning techniques in solving real-world face detection and classification challenges, the journey has been drastically progressive.

III. Methodology

A. Dataset Description and Preprocessing

The dataset used comprises real and fake images sourced from existing public datasets, ensuring a balanced distribution of both classes to effectively train the model for accurate distinction between genuine and altered images.

1) Data Acquisition and Labelling: Images labeled ’real’ are untouched photographs captured in natural settings, while ’fake’ images are artificial or digitally altered visuals. Emphasis was placed on including diverse lighting, backgrounds, and facial expressions to accurately mimic real-world conditions.

2) Preprocessing:

a) Face Detection:: Each image underwent face detection using OpenCV’s Haar cascades to crop and focus on relevant facial features, filtering out noise.

b) Image Resizing and Normalization:: Detected faces were resized to 224x224 pixels, a standard format for deep learning models. Normalization scaled pixel values between 0 and 1, ensuring consistent input data suitable for faster model convergence.

B. Model Design: VGG16 and Transfer Learning

1) Understanding VGG16: VGG16 is a highly popular deep CNN subtype renowned for its effective feature extraction capabilities. Designed by Karen Simonyan and Andrew Zisserman, it comprises 16 layers: 13 convolutional and 3 fully connected layers. Trained on the ImageNet database, VGG16 exhibited state-of-the-art performance in image classification tasks.

2) Transfer Learning Strategy: Transfer learning involved using the pre-trained VGG16 model as a feature extractor. All layer weights except the final fully connected layers were frozen to extract high-level features crucial for distinguishing between real and fake images. The final fully connected layers were replaced with a new dense layer followed by a softmax activation function for binary classification into real or fake, enabling fine-tuning to adapt the model to the specific dataset.

C. Training Setup and Optimization

1) Loss Function and Optimizer: CrossEntropyLoss function replaced the previous activation function to effectively measure discrepancies between predicted and actual class labels in multi-class classification. Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001 and momentum of 0.9 was chosen for its ability to manage complex parameter spaces and optimize model parameters effectively.

2) Learning Rate Scheduling: A step learning rate scheduler was implemented to enhance training stability and ensure convergence. The learning rate was reduced by a factor every 7 epochs, dynamically adjusting during training.

3) Training Procedure: The model was trained over 30 epochs using mini-batch gradient descent. Batches of preprocessed images were iteratively fed into the model, gradients were computed, and weights were updated to minimize classification errors on the training dataset.

D. Evaluation Metrics and Performance Analysis

1) Performance Metrics: Primary evaluation metric included accuracy, measuring the percentage of correctly classified images (real vs. fake). Additional metrics such as precision, recall, and confusion matrix provided insights into the model’s ability to distinguish between true positives (real) and true negatives (fake).

2) Overfitting Mitigation: To address overfitting, dropout regularization was implemented during training epochs, randomly deactivating neurons to enhance model generalization and prevent dependence on specific features.

E. Results and Discussion

Training and testing accuracies were evaluated to assess model generalization and efficiency in differentiating between real and fake images. Strengths, limitations, and areas for improvement were identified based on empirical performance evaluations, focusing on data quality, model architecture, and optimization strategies.

F. Future Directions

Future research directions include exploring advanced CNN architectures (e.g., ResNet, EfficientNet) for enhanced feature extraction and classification accuracy. Integration of adversarial training techniques aims to improve model robustness against advanced image manipulations like deepfakes. Dataset expansion continues to include a broader range of real-world scenarios, ensuring the model's applicability across diverse environments.

IV. Results

A. Training Progression

In this work, a VGG16 architecture convolutional neural network CNN augmented with Haar cascades in face detection and classification tasks is used. This model was trained for 30 epochs in classifying the facial images as ’real’ and ’fake’. Key metrics, such as accuracy and loss, were tracked in this training, which quantifies how the model is learning and optimizes over time.

1) Training Accuracy: The first epoch that the model trained for was 53.60

At the 30th epoch, the training accuracy went as high as 93.88

2) Training Loss: On the other hand, another important metric—the training loss—follows this trend and keeps decreasing:

The loss, at 0.73 in the first epoch, dropped progressively during training and hit a minimum of 0.19 by the last epoch. This reduction furnished proof of effective model parameter optimization for its capability to align output predictions very close to the ground-truth labels.

B. Testing Accuracy

After the intensive training regime, the model was put under rigorous testing with an independent test dataset that had not been used in training. The idea was to see if the model generalized features it had learned and continued to perform highly accurately in classifying new facial images.

1) Testing Accuracy: Results obtained were very encouraging, returning a testing accuracy of 93.53

This result validated the model for its robustness and reliability in differentiating between authentic facial images and their correspondingly digitally manipulated versions. Testing accuracy as high as this showed that the model had a very good generalization capacity of the learned features and patterns to new data, thus proving its applicability in real-world scenarios related to image forensic and security applications.

C. Implications and Significance

The findings of this study have important implications for a variety of domains which are based on the verification of authenticity of images and detection of manipulation:

Improved security measures: The line of security measures, unlike before, is improved in sectors such as law enforcement, digital forensics, and online authentication by the accurate identification of manipulated facial images. It reduces several risks associated with fraudulent activities and misinterpreted information.

preservation of media integrity: Authentication of visual content over digital media helps reduce the spread of misleading or deceptive media; hence, it will preserve public trust and integrity in digital communications.

Future Research Directions: Future research can focus on the most evolved methodologies to boost model robustness by scaling up, integrate rich and diverse datasets that would generalize models to much more demographics and scenarios, or apply real-time detection techniques, approaching new dangers arising as image manipulation and deepfakes.

V. Conclusion

This paper discusses the effectiveness of a VGG16 convolutional neural network (CNN) augmented with Haar cascades in the detection of manipulated facial images. Deep learning has an important application in image forensics and security applications.

Our model showed resilient performance in classifying the facial images into authentic and manipulated ones through rigorous training and evaluation. From a modest accuracy, the model improved further to achieve a peak training accuracy at the 30th epoch with an accuracy of 93.88

The highest accuracy measured on independent datasets was 93.53

Our findings also provide further insights into young adults improving security measures and preserving media integrity by mitigating the risks associated with digital manipulation. In the future, studies will be conducted with advanced methodologies in an effort to strengthen the model's robustness and scalability so that it remains up to emerging challenges in image forensics and real-time deepfake detection.

The findings of this research offer an important insight into how the use of convolutional neural networks can be combined with Haar cascades for efficient detection of facial image manipulation.

Acknowledgment

The authors would like to thank the support provided by SRM Institute Of Science & Technology, Kattankulathur, Chennai, Tamil Nadu.

References

[1] Rössler, Andreas, et al. ”Faceforensics++: Learning to detect manipulated facial images.” Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019.

[2] Dang, Nguyen H., et al. ”Detecting deepfake videos from residual interframe noise.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.

[3] Li, Yuezun, et al. ”Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.