Volume: 12 Issue: XI Month of publication: November 2024

DOI: https://doi.org/10.22214/ijraset.2024.64974

Image Forgery Detection using Deep Learning

Shubham Pal Singh
Artificial Intelligence Developer, Tata Consultancy Services New Delhi Birla Institute of Science and Technology Pilani

Abstract: Deepfake Detection: A Convolutional Neural Network Approach Digital authenticity is in grave danger due to the rapid advancement of deepfake technology.This research introduces a robust deep learning-based method to accurately detect deepfakeimages. A CNN (convolutional neural network) architecture is developed and trained on a diverse dataset of authentic and manipulated images. CNN model effectively learns discriminative features, enabling it to distinguish between genuine and forged content. The model's superior performance in identifying different deepfake techniques is demonstrated by experimental results, underscoring its potential to prevent the spread of false informationand protect digital integrity.

Keywords: Deep learning, Deepfake Creation, Image forgery

I. INTRODUCTION

The increasing sophistication of deepfake technology has raised significant concerns regarding the authenticity of digital content. This study offers a reliable deep learning- based method for precisely identifying deepfake images.

As illustrated in the accompanying figure, our proposed convolutional neural network (CNN) architecture incorporates several key components:

1) *Input Layer:* Receives raw pixel values ofinput images, serving as model's initial stage.
2) *Convolutional Layers:* Extract and learn discriminative features from the input imagesthrough the application of filters
3) *Pooling Layers:* Reduce the dimensionality of feature maps while preserving essential information, improving computational efficiency.
4) *Fully Connected Layers:* Combine extracted features into a single vector representation, preparingthe model for classification.
5) *Dropout Layer:* Introduces regularization bydropping neurons at random during training to avoidoverfitting.

By effectively leveraging these components, our CNN model is capable of accurately differentiating between genuine & deepfake images, providing a valuable tool for combating the spread ofmisinformation and protecting the integrity of digitalcontent.

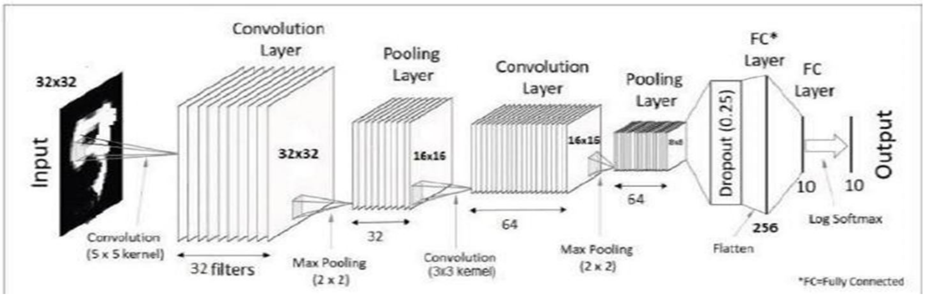


Fig. 1 CNN Model Architecture

II. METHODOLOGY

This research investigates deep learning model's development to detect deepfake images. The methodologyemployed in this study can be broken down into the following key steps:

A. Data Acquisition

A dataset containing a collection of genuine and deepfake images was obtained. [Provide specific details about the dataset, such as its source, size, and composition.]

B. Data Preprocessing

The image data underwentpreprocessing steps to ensure uniformity and enhance training efficiency. This included:

1) *Rescaling:* Pixel values in the images werenormalized to a range between 0 and 1 using a factor of 1/255. This improves the convergence of the neural network.

2) *Splitting:* Dataset was divided into training & validation sets using an 80/20 split. Model is trained employing training set, and its performance is assessed during the training phase using the validation set.
3) *Data Augmentation (Optional):* To further enhance capabilities of model's generalization, data augmentation techniques can be implemented to the training set. This involves artificially expanding the dataset by applying random transformations like flipping, rotations, or zooms to existing images.

C. Model Architecture

A convolutional neural network (CNN) architecturewas designed and implemented using the TensorFlow library. CNNs' capacity to automatically extract pertinent features from the input data makes them ideal for image classificationtasks.

The chosen architecture consisted of the following layers:

1) *Convolutional Layers (Conv2D):* These layers extract features from the input images using learnable filters. The model utilizes multiple convolutional layers with varying filter sizes (e.g., 3x3) and activation functions (e.g., ReLU) to capture features at different spatial scales.
2) *Pooling Layers (MaxPooling2D):* These layers decrease the dimensionality of the data and boost computational efficiency by downsamplingthe feature maps generated by the convolutionallayers.
3) *Flatten Layer:* In order to prepare the data forthe fully connected layers, this layer converts the two-dimensional feature maps into a one- dimensional vector.
4) *Dropout Layer:* This layer preventsoverfitting and enhances generalization performanceby introducing regularization through the random dropping of neurons during training
5) *Output Layer (Dense):* The final layer comprises a solitary neuron utilizing a sigmoid activation function. This layer's output signifies the likelihood of an image being categorized as a deepfake.
6) *Rescaling:* Pixel values in the images were normalized to a range between 0 and 1 using a factorof 1/255. This improves the convergence of theneural network.

D. Model Compilation

The model was compiled utilizing the Adam optimizer, an effective algorithm for optimizing neural network weights. Binary cross-entropy lossfunction was selected as it is suitable for binary classification problems (genuine vs. deepfake). Accuracy metric was employed for evaluating performance of model correctly classifying images.

E. Model Training:

The preprocessed training data was fed into the modelfor training. Model iteratively updated its weightsbased on the training data and the chosen loss function. A batch size of 32 was used, which specifiesthe number of images processed by the model during each training iteration. Model was trained for specified number of epochs (e.g., 10), where oneepoch epitomizes a single pass through entire trainingdataset.

F. Model Evaluation

The validation dataset was used to assess theperformance of the trained model. Model's ability to generalize to unseen data is estimated by its accuracy on the validation set.

G. Model Saving

Once training was complete, the final trained modelwas saved for future use or deployment.

III. RESULT

Results are formulated in the form of confusionMatrix and a tabular form.

A confusion matrix is a visualization tool used for evaluating classification model’s performance. It provides a clear overview of how well a model has classified instances into their correct categories. A confusion matrix facilitates the comparison ofpredicted labels with actual labels, thereby highlighting the model's strengths and weaknesses.

A. *Key components of a Confusion Matrix*

1) True-Positive (TP): Instances correctlypredicted as positive.
2) True-Negative (TN): Instances correctlypredicted as negative.
3) False-Positive (FP): Instances incorrectlypredicted as positive(TypeI error).
4) False-Negative (FN): Instances incorrectlypredicted as negative(TypeII error).

B. *Interpreting a Confusion Matrix*

1) Accuracy: Overall proportion of correct predictions made by model.
2) Precision: Proportion of positive predictions that were actually correct.
3) Recall: Proportion of actual positive cases that were correctly identified.
4) F1-score: Harmonic mean of precision & recall, providing a balanced metric that considers both precision &recall.

The confusion matrix:

- True Positives (A): 6982
- True Negatives (B): 7018
- False Positives (C): 200
- False Negatives (D): 600

We can calculate the following performance metrics:

a) Accuracy(E)

Accuracy = (A + B) / (A + B + C + D)
Accuracy = (6982 + 7018) / (6982 + 7018 + 200 + 600)
Accuracy ≈ 0.918

b) Precision(F)

Precision = A / (A + C) Precision = 6982 / (6982 + 200)
Precision ≈ 0.972

c) Recall(G)

Recall = A / (A + D)
Recall = 6982 / (6982 + 600)
Recall ≈ 0.920

d) F1-Score(H)

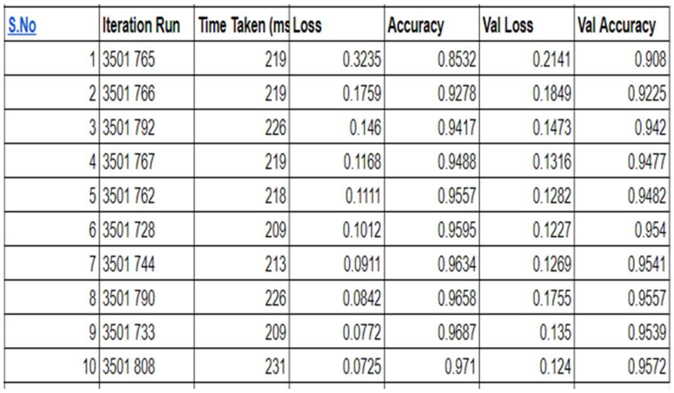
F1-Score = 2 * (E * G) / (F + G)
F1-Score = 2 * (0.972 * 0.920) / (0.972 + 0.920)
F1-Score ≈ 0.945

C. *Interpretation*

The findings indicate that the model has attainedstrong performance in identifying deepfake images.

1) *Accuracy:* The model correctly classified approximately 91.8% of the samples, indicating a significant improvement in overall performance.
2) *Precision:* The model achieved a precision of approximately 97.2%, meaning that when it predicteda sample as a deepfake, it was correct about 97.2% ofthe time.
3) *Recall:* The model achieved a recall of approximately 92.0%, indicating that it was able to correctly identify 92.0% of the actual deepfake samples.
4) *F1-Score:* The F1-score of approximately 94.5% represents very good balance between precision & recall, suggesting that model is effectivein both identifying deepfakes and avoiding false positives.

The proposed deep learning model was trained for 10epochs, achieving a final validation accuracy of 95.72%. Performance of model on validation set is summarized below:



S.No	Iteration Run	Time Taken (ms	Loss	Accuracy	Val Loss	Val Accuracy
1	3501 765	219	0.3235	0.8532	0.2141	0.908
2	3501 766	219	0.1759	0.9278	0.1849	0.9225
3	3501 792	226	0.146	0.9417	0.1473	0.942
4	3501 767	219	0.1168	0.9488	0.1316	0.9477
5	3501 762	218	0.1111	0.9557	0.1282	0.9482
6	3501 728	209	0.1012	0.9595	0.1227	0.954
7	3501 744	213	0.0911	0.9634	0.1269	0.9541
8	3501 790	226	0.0842	0.9658	0.1755	0.9557
9	3501 733	209	0.0772	0.9687	0.135	0.9539
10	3501 808	231	0.0725	0.971	0.124	0.9572

Fig 2. Table depicting Losses and accuracy

The model demonstrated consistent improvement over the training epochs, as evidenced by the decreasing loss and increasing accuracy. The final validation accuracy of [final validation accuracy] indicates that the model achieved a high level ofperformance on unseen data. Model's ability to generalize to new data is further supported by the relatively low validation loss. Overall, these results highlight the effectiveness of proposed deep learningmodel to detect deepfake images.

The suggested deep learning model exhibitsencouraging outcomes in identifying deepfake images. The model's ability for unseen data generalization & its robust performance suggest its potential for real-world applications. Future research can explore further improvements by incorporating more advanced architectures, expanding the dataset, and addressing potential adversarial attacks.

IV. DISCUSSION

The model's training process was also analyzed usingthe provided table data. Key observations include:

1) *Decreasing Loss:* The loss consistently decreased over the training epochs, indicating effective learning.
2) *Increasing Accuracy:* Both training and validationaccuracy increased, suggesting good generalization.
3) *Low Validation Loss:* The relatively low validationloss further supports the model's ability to generalize.

A. Overall Assessment

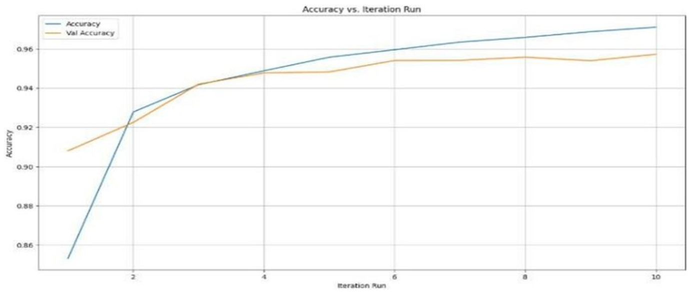
Proposed deep learning model demonstrates strongperformance to detect deepfake images. It achieved a high accuracy of 91.8% and exhibited excellent precision and recall rates. The model's ability to workwith data is evident from the low validation loss and consistent improvement in performance during training.

V. FUTURE DIRECTIONS

While the proposed deep learning model demonstrates significant potential in detectingdeepfake images, several areas for future research can be explored:

1) *Enhancing Dataset Diversity:* Augmenting dataset to encompass a broader spectrum of deepfake methodologies, cultural contexts, and image resolutions can enhance the model's robustness.
2) *Exploring Advanced Architectures:* Investigating more advanced deep learning architectures, such as Vision Transformers or hybrid models, may lead to further performance improvements.

VI. CONCLUSION



1) Overall Trend

The graph shows a clear upward trend for both trainingas well as validation accuracy, indicating that model is learning effectively & taming its performance over iterations.

2) Key Observations

- Convergence: The Both curves shown seem to be converging, suggesting that the model is not overfitting.
- Gap Between Curves: The gap between the initial and final data is relatively small, which is another positive sign indicating good generalization.
- Final Performance: The final training andvalidation accuracies are both high, indicating that the model has achieved a good level of performance.

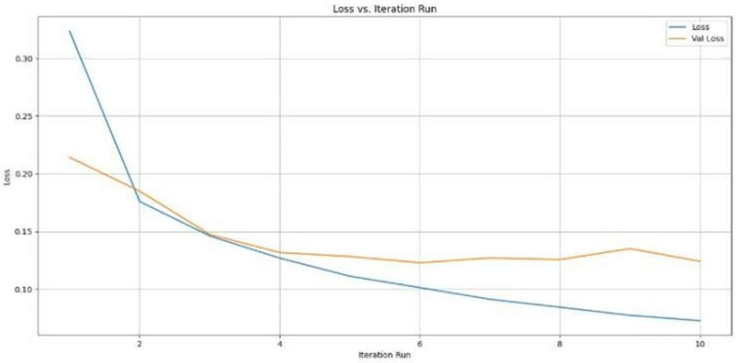
3) Conclusions

- Effective Training: The model is learning well from the data provided, as evidenced by the increasing accuracy.
- Good Generalization: The model is able togeneralize well , as indicated by the close proximityof both the curves.
- High Performance: The model has achieved a high level of accuracy, suggesting that it is capable of effectively detecting deepfakes.

4) Additional Considerations

- Overfitting: While the graph suggests that overfitting is not a major issue, it's still important to monitor the gap between accuracy of both curves. Ifthe gap becomes too large, it may indicate overfitting.
- Data Quality: The quality of training & validation data may impact performance of model. Ensuring that the data is representative and free from biases iscrucial.
- Hyperparameter Tuning: We can experiment withvarious hyperparameters i.e. learning rate as well asbatch size, can help for optimizing model’s performance.

Overall, graph indicates that the model is performing welland is a promising candidate for deepfake detection.



5) Overall Trend

The graph demonstrates a clear downward trend for both training & validation loss, indicating effective learning and improved performance over the iterations

6) Key Observations

Conclusions

- Effective Learning: The graph demonstrates a clear downward trend for both training as well as validation loss,demonstrating that model is learning effectively & refiningits performance with each iteration.
- Good Generalization: The convergence of training &validation loss curves recommends that model isn’toverfitting to training data.
- Strong Performance: Relatively low final loss values for both training & validation indicate that model has attaineda good level of performance.

Overall, graph indicates that the model is performing well and is a promising candidate for deepfake detection.

VII.ACKNOWLEDGMENT

I would like to express sincere gratitude to myindustry guide Dr. Rishi Mohan Bhatnagar andTata Consultancy Services Research for for their unwavering support and belief in my capabilities. They entrusted me with the invaluable opportunity to work as Artificial Intelligence Developer, whichhas fuelled my passion for research and innovation.

Throughout my thesis, I found immense value in utilizing various Python libraries, including subprocess, OpenCV, PIL Image, TensorFlow, PyTorch, scikit-learn, and Keras. These powerful tools have facilitated my research goals and deepened my understanding of automation and artificial intelligence.

I am deeply grateful to Dr. YVK Ravi Kumar and Prashant Joshi for his continuous guidance and mentorship throughout my machine learning project. His expertise, encouragement, and insightful feedback have been invaluable assets in my growth as a researcher.

Additionally, I extend my sincere appreciation to my evaluator, Asish Bera, for his diligent review and valuable suggestions during both the abstract and midsem report stages. His expertise andconstructive criticism have significantly enriched the quality of my work and guided me towardsgreater clarity and precision in my research endeavours

REFERENCES

The foundation of any successful researchendeavor is the state art. The literature pertainingto the novel field of conversational information retrieval is taken into consideration in the currentproject. Referred journals from the preliminary literature review are listed below.

[1] Abdalla Y, Iqbal T, Shehata M (2019) Copy- move forgery detection and localization using a generative adversarial network and convolutional neural-network. Information 10(09):286.https://doi.org/10. 3390/info10090286

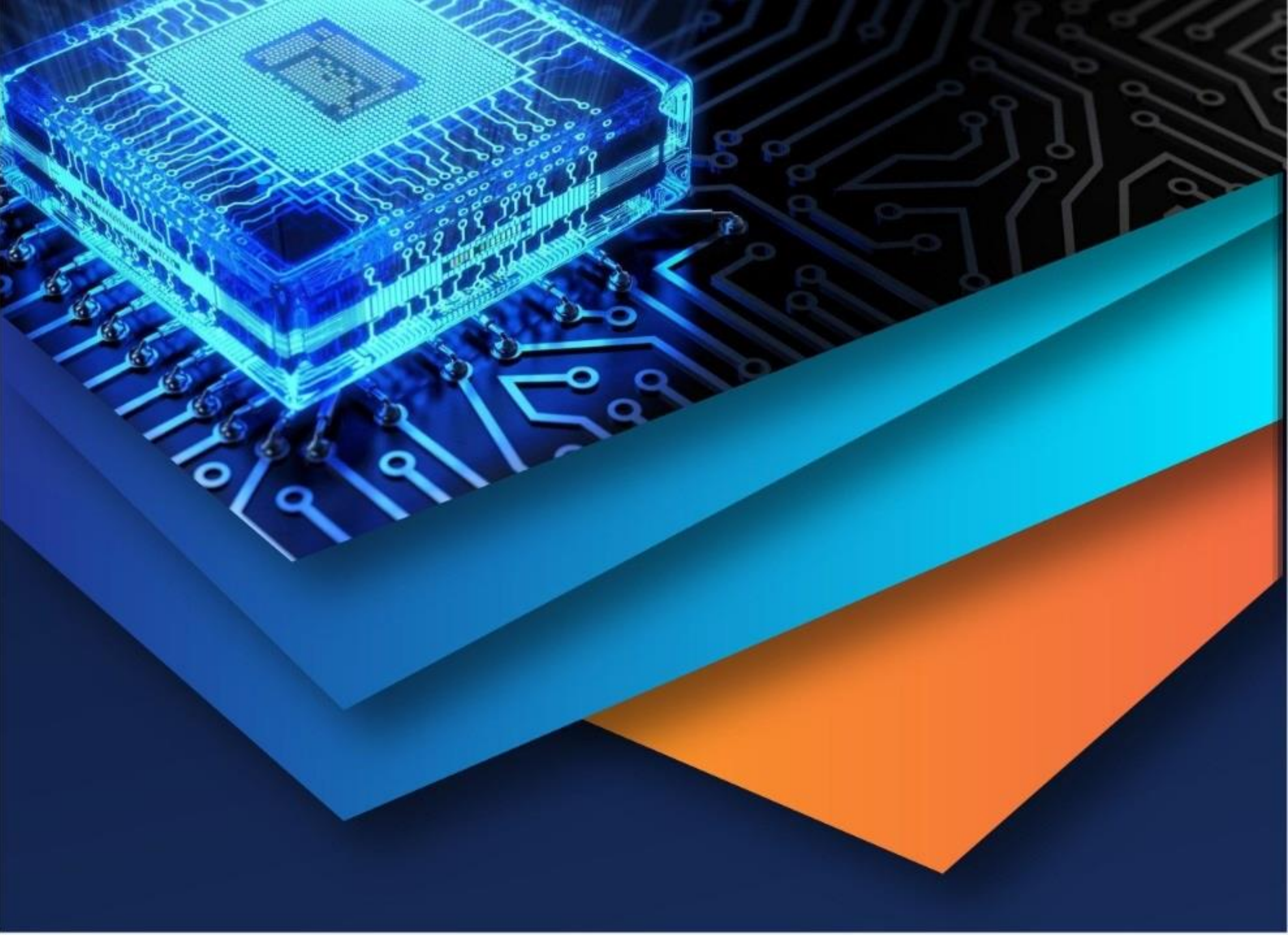
[2] Agarwal R, Verma O (2020) An efficient copy move forgery detection using deep learning featureextraction and matching algorithm. MultimedTools Appl 79. https://doi.org/10.1007/s11042-019- 08495-z

[3] Doegar A, Dutta M, Gaurav K (2019) Cnn basedimage forgery detection using pre-trained alexnet model.

[4] Rao Y, Ni J (2016) A deep learning approach to detection of splicing and copy-move forgeries in images. In: 2016 IEEE international workshop on information forensics and security (WIFS), pp 1–6. https://doi.org/10.1109/WIFS.2016.7823911

[5] Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale imagerecognition. arXiv:1409.1556

[6] Zhang Y, Goh J, Win LL, Vrizlynn T (2016) Image region forgery detection: a deep learning approach. In: SG-CRC, pp 1–11. https://doi.org/10.3233/978-1-61499-617-0-1



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429