Digital Image Forgery Detection Using Pre-Trained Xception Model as Feature Extractor

Niyantha Maruthu Pandiyan
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, India
0000-0002-5375-9411

***Abstract*—During this time of social network bloom, there has been a tremendous increase in the number of images shared over the internet. This along with the advancements in software, like photoshop, has made morphing or tampering with the original images very easy. This paper proposes a Deep Learning approach to tackle this problem and detect Fake or factored images. In this paper, we propose a method where a pre-trained Xception Model is used as the feature extractor to classify a certain image as Tampered or Real. The proposed model was trained on the CASIA-V2 dataset. The model was able to attain good Accuracy on the Test Split.**

***Index Terms*—Image Forgery Detection,Transfer Learning, Deep Learning, CNN, Feature Extraction, Xception**

I. Introduction

In these modern times, due to innovations in technology, there has been a high volume of Digital images floating around on the internet. The use of images in communication or to covey any information has made absorbing information easier and much more efficient. These images can also be used as evidence in various matters. However, also due to the bloom in technology, there now exists a lot of cheap efficient or cheap ways to tamper with images. These methods can tamper with an image without leaving any visible evidence. This manipulated image can then be used to spread fake news, mislead people and carry out malicious agendas. This presents an enormous and important task to detect such forged images.

Image tampering can be classified into 5 types.

- Copy move forgery
- Splicing
- Retouching
- Re-sampling
- Morphing

Image forgery detection techniques can be broadly classified into 1) Active methods and 2) Passive Methods [16]. Active methods use prior knowledge of the image to detect tampering such as digital watermarking and signatures, Passive methods do not require prior knowledge of the image and use changes in intrinsic properties of the image to detect tampering [10]. However, with innovations in technology and tampering with images in a clever way such as copy-pasting regions within the same image, active methods are not able to keep up with detecting these forgeries, hence a lot of research is been done on passive methods.

In recent times Deep learning methods are being used to detect tampering as these intelligent models can learn and extract features of the images by themselves without any manual intervention. Manually designing features is also not possible for a broader problem. Hence the usage of deep learning methods. [2] Further implementation and usage of Deep learning methods to Detect tampering will be discussed in the subsequent sections.

The paper's contributions are summarized as follows:

- This paper proposes a new Model Architecture for Copy-move Forgery Detection.
- Uses Pretrained CNN-based Model as Feature Extractor
- Increases classification performance by decreasing False Positive Rate(FPR)

The paper is arranged as the following, Section 2 gives a summary of the latest work done in Image Forgery Detection. An overview of the proposed methodology will be discussed in Section 3. The proposed method will be discussed in detail in Section 4. Section 5 contains the Experimental results and Performance Evaluation respectively. Section 6 gives the conclusion of the paper, followed by the references.

II. Literature Review

Doegar et al. in 2018 [2] proposed a CNN-based pre-trained AlexNet model to detect forgeries in an Image. AlexNet is a 25-layer model with the main layers being, convolutional, pooling, fully connected and SoftMax with ReLU as the activation function. The model is trained on the Benchmark MICC-F220 dataset which contains 110 forged and 110 non-forged images. The images are first preprocessed and resized to 227x227 and fed into the Model, the features are extracted from the f7 layer of the AlexNet. Which is then used to train an SVM classifier. Precision, False positive rate(FPR), True positive rate(TPR), Recall, F-measure and accuracy are used as metrics to evaluate performance. The model was able to achieve 93.94% accuracy, 100% TPR, 12.12% FPR, 89.19% precision 94.28% F-measure, on the benchmark dataset with an average execution time of 4.86 seconds. It was noted that

the model performed well even with the presence of geometrical and rotational transformation. With reduced training time due to the usage of a pre-trained model.

[3] 2019, proposed a hybrid LSTM(Long Short Term Memory) encoder-decoder model to detect and localize manipulation in an image at the pixel level. The proposed model consists of an LSTM, An encoder and a decoder. An encoder is used to find the spatial feature maps of the manipulated images and Resampling features are used in LSTM to learn the correlation between different patches. These are then sent to the decoder network to obtain the binary mask of the tampered region. Pixel accuracy and receiver operating characteristics are used as metrics of performance evaluation. This model is observed to outperform most baseline models on various datasets.

El Biach et al. 2021 [13] proposed Fals-Unet, an encoder-decoder-based CNN to detect and localize image forgeries. The encoder extracts high-level features by performing convolution activation and normalisation, then the decoder is used to locate spatial information location. The encoder is based on ResNet50 and the decoder has several decoder blocks which consist of upsampling, normalisation and ReLU activation. Which is then fed to a SoftMax classifier and then finally a binary mask that represents the tampered parts is outputted. The proposed model is observed to perform very well on various datasets but is high in computational complexity.

Monika and Abhiruchi Passi, 2021 [14] proposed an algorithm to improve the detection speed by preprocessing the data using DWT and PCA. Initially, thresholding is applied to the image to convert it to a binary image. Then feature extraction is done using Fourier transform and feature reduction using PCA. Then the SVM classifier is trained to predict forgery.

Goel et al, 2020 [15] proposed a dual branch CNN to perform copy move forgery detection efficiently and fast. Here both branches are connected to the same input. Each branch consists of 3 convolution layers which use the ReLU activation function and each layer is followed by a 2x2 max pooling layer. Kernel sizes are different in these two branches to be able to extract multiscale features. Due to this one layer will have a zero padding layer to obtain the symmetric output. Then both branches are concatenated, and the output is then fed to the global max pool layer followed by two dense layers where the final layer has output 1, which ends with a sigmoid activation function which outputs the probability. The model is trained on the MICC-F2000 dataset, which has 1300 tampered and 700 original images. The images are pre-processed and resized to 700x700 before feeding them to the model.

F1 score, TPR, FPR, sensitivity and specificity are the metrics used to evaluate the performance in this paper. It has been observed that the model has an accuracy of 96% sensitivity 100% specificity 93% and precision and recall of 89% and 100% respectively when the kernel sizes are (3,5) and (5,8). However, validation loss was slightly less in (5,8)

[16] in 2021 proposed a modified MobileNetV2 to tackle the copy-move forgery problem, considering resource constraints. MobileNetV2 was originally used to classify 1000 image classes, so there were modifications required to make it a binary classification model. This was done by attaching a new FC layer instead of the old pre-trained one. The base layers were frozen to prevent updating of weights during backpropagation. Then a global average pooling layer followed by a dense layer with two outputs, along with the SoftMax function, to predict two classes, either forged or authentic. Dropout layers were also utilized to prevent the overfitting of the model. The images were resized to 224x224 before feeding them into the model as required by the input layer. True positive rate, false positive rate and accuracy were used to evaluate the performance of the model. The MoblieNetV2 model was seen to show better FPR results 13% when compared to SVGGNET's 19%. It was concluded that MobileNetV2 outperformed SVGGNET during deployment with 84% TPR and 14.35% FPR when compared to SVVGNET's 67% and 16.3%.

III. Methodology

The main objective of this paper is to propose an architecture that is capable of differentiating tampered images from real images. The architecture should take an image as input and classify it as either a Tampered or untampered image.

The proposed model is a CNN-based classifier. Which uses an Xception Model as the feature extractor. These features are then passed to 2 Dense Layers and then, a softmax layer outputs the classification.

Due to computational constraints, we are using a pre-trained model, which also reduces the training time of the model. The xception model is chosen because of its moderately less computational complexity when compared to the other state-of-the-art models while maintaining high accuracy, as seen on the ImageNet dataset. Xception here is used to extract features from the input images.

The final layer of the model is removed and replaced with a custom classification model, starting with the global average pool layer, Dense layers. This then ends with a SoftMax activation layer which predicts the probability of which type of classification it is. The model is trained on the CASIA2 dataset.

The methodology consists of the following modules

1) Preprocessing
2) Feature Extraction
3) Classification

We will see each of these modules in detail in the next section.

IV. Architecture

A. Preprocessing

The Images are first resized to 256*256 and then Error Level Analysis(ELA) is done on each and every image. These Images are then stored as numpy arrays. The entire dataset is split into Train, Test, Validation and then sent to the feature extractor. ELA helps in identifying areas within an image that are at a different compression level. In an Untampered Image,

Fig. 1. Example of an Untampered Image

Fig. 3. Example of a tampered Image

Fig. 2. Untampered Image after ELA

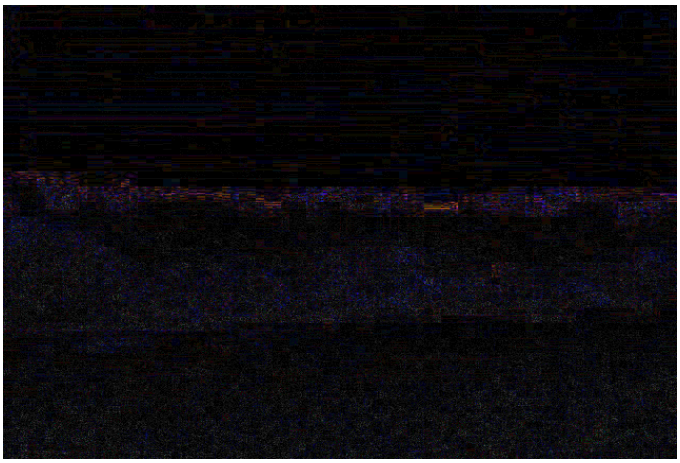

Fig. 4. Tampered Image after ELA

all the areas in an image must have the same compression level, as seen in Fig 1. If any part of the image has significantly different levels as seen in Fig 2, it could suggest a digital modification.

B. Feature extractor

This paper proposes a Pretrained Xception model for extracting the features from the images. This takes in the preprocessed images and outputs a set of features to be used by the classification head. Xception is a deep CNN architecture with depth-wise separable convolutions, it is an extreme version of the Inception model. This model has 71 layers, which were pre-trained on the Imagenet dataset and capable of classifying images into 1000 categories. This architecture relies on Depthwise separable convolutions, max-pooling and shortcuts between the convolution blocks like Resnet. Here Pointwise convolution is followed by Depthwise convolution

C. Classification

As seen in Fig 2.0, the feature extraction is followed by a Global Average pooling layer and two Dense layers with softmax activation functions, this act as the classifier. Extracted features from the Xception models are passed through this classification layer, which then classifies the Input image as either Untampered or tampered. The entire architecture has 1,050,114 trainable parameters.

V. Experimental Results

A. Dataset

The proposed model was trained on the CASIA ITDE v2 dataset which is similar to the v1. It consists of 7.200 authentic and 5,123 tampered images, with image sizes ranging from 320*240 to 800*600. However, due to computational constraints, only 1,200 authentic and 1,200 Forged images from the CASIA v2 dataset was used for implementation purpose, with a 90:5:5 Train:Validation: Test split.

B. Performance Evaluation

To estimate the performance of the proposed model, evaluation metrics such as Accuracy, True Positive Rate, False Positive Rate, True Negative Rate, False Negative Rate are used.

```
Layer (type)                 Output Shape              Param #
=================================================================
xception (Functional)        (None, 8, 8, 2048)        20861480

global_average_pooling2d (G  (None, 2048)              0
lobalAveragePooling2D)

dense (Dense)                (None, 512)               1049088

dense_1 (Dense)              (None, 2)                 1026

=================================================================
Total params: 21,911,594
Trainable params: 1,050,114
Non-trainable params: 20,861,480
```

Fig. 5. Model Architecture

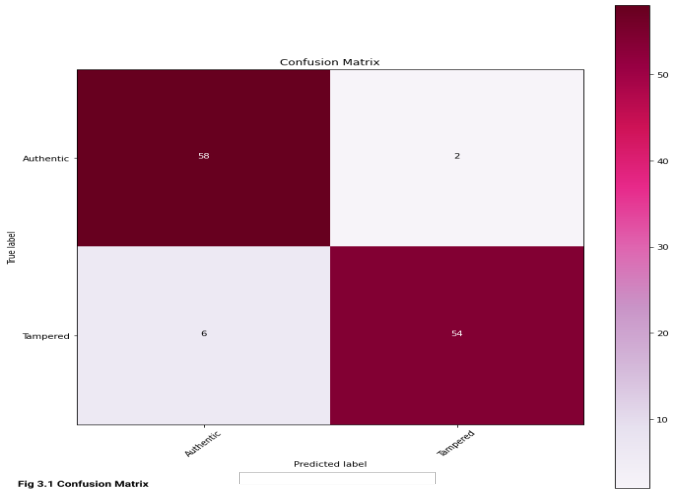

Fig. 6. Confusion Matrix

The proposed model was able to achieve 88% accuracy on the validation set and 95% on the Test set. The model has TPR% as 98.2%, TNR% as 98.3% and FPR as 1.6%, we can see the confusion matrix of predictions in Fig 3.1 along with various evaluation metric scores. We have also compared the performance of the proposed methods with other models as seen in Fig 3.3

The proposed model does well, however, there is a lot of room for improvement. The model would potentially do better when the entire CASIA v2 dataset is used instead of a small portion.

TABLE I
MODEL PERFORMANCE

Model	Accuracy	Recall	Precision	TNR	FPR	FNR
	95.0%	91.6%	98.2%	98.3%	1.6%	8.3%

TABLE II
COMPARISON BASED ON ACCURACY

Model	Accuracy
Kuznetsov et al [4]	97.8%
Ortega et al [8]	98%
Walia et al [7]	99.31%
Proposed Model	95%

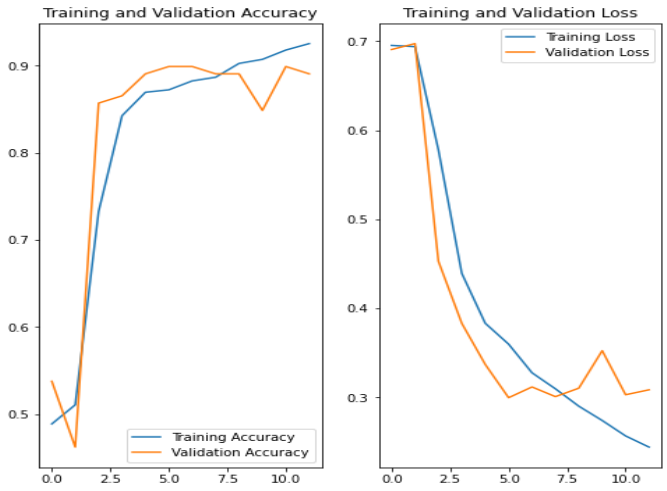

Fig. 7. Training and Validation Graph

VI. CONCLUSION AND FUTURE WORK

A Deep Learning based CNN model was presented in this paper. The main objective is to classify an input image as either Authentic or Tampered. The paper proposes an Xception model as a feature extractor, followed by Dense layers. The proposed model was able to achieve 95% accuracy on test data. The proposed model takes advantage of the pre-trained Xception model to do most of the heavy lifting. There is scope for improvement for this model, future work includes trying out various other feature extractors that could prove to show better accuracy and exploring pre-processing methods.

REFERENCES

[1] H. D. Deng, "Image Level Forgery Identification and Pixel Level Forgery Localization via a Convolutional Neural Network," 2018.

[2] A. Doegar, M. Dutta, and G. Kumar, "CNN based Image Forgery Detection using pre-trained AlexNet Model," Proc. Int. Conf. Comput. Intell. IoT 2018, pp. 402–407, 2018.

[3] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy-Chowdhury, "Hybrid LSTM and Encoder-Decoder Architecture for Detection of Image Forgeries," IEEE Trans. Image Process., vol. 28, no. 7, pp. 3286–3300, 2019, doi: 10.1109/TIP.2019.2895466.

[4] A. Kuznetsov, "Digital image forgery detection using deep learning approach," J. Phys. Conf. Ser., vol. 1368, no. 3, 2019, doi: 10.1088/1742-6596/1368/3/032028.

[5] A. Mazumdar and P. K. Bora, "Two-stream Encoder-Decoder Network for Localizing Image Forgeries," pp. 1–11, 2020, [Online]. Available: http://arxiv.org/abs/2009.12881.

[6] F. Marra, Di. Gragnaniello, L. Verdoliva, and G. Poggi, "A Full-Image Full-Resolution End-to-End-Trainable CNN Framework for Image Forgery Detection," IEEE Access, vol. 8, pp. 133488–133502, 2020, doi: 10.1109/ACCESS.2020.3009877.

[7] S. Walia, "Fusion of Handcrafted and Deep Features for Forgery Detection in Digital Images," IEEE Access, vol. 9, pp. 99742–99755, 2021, doi: 10.1109/ACCESS.2021.3096240.
[8] Y. Rodriguez-Ortega, D. M. Ballesteros, and D. Renza, "Copy-move forgery detection (Cmfd) using deep learning for image and video forensics," J. Imaging, vol. 7, no. 3, 2021, doi: 10.3390/jimaging7030059.
[9] M. A. Elaskily, M. H. Alkinani, A. Sedik, and M. M. Dessouky, "Deep learning based algorithm (ConvLSTM) for Copy Move Forgery Detection," J. Intell. Fuzzy Syst., vol. 40, no. 3, pp. 4385–4405, 2021, doi: 10.3233/JIFS-201192.
[10] K. B. Meena and V. Tyagi, "A deep learning based method for image splicing detection," J. Phys. Conf. Ser., vol. 1714, no. 1, 2021, doi: 10.1088/1742-6596/1714/1/012038.
[11] Abhishek and N. Jindal, "Copy move and splicing forgery detection using deep convolution neural network, and semantic segmentation," Multimed. Tools Appl., vol. 80, no. 3, pp. 3571–3599, 2021, doi: 10.1007/s11042-020-09816-3.
[12] Z. N. Khudhair, F. Mohamed, and K. A. Kadhim, "A Review on Copy-Move Image Forgery Detection Techniques," J. Phys. Conf. Ser., vol. 1892, no. 1, 2021, doi: 10.1088/1742-6596/1892/1/012010.
[13] F. Z. El Biach, I. Iala, H. Laanaya, and K. Minaoui, "Encoder-decoder based convolutional neural networks for image forgery detection," Multimed. Tools Appl., 2021, doi: 10.1007/s11042-020-10158-3.
[14] Monika and A. Passi, "Digital Image Forensic based on Machine Learning approach for Forgery Detection and Localization," J. Phys. Conf. Ser., vol. 1950, no. 1, 2021, doi: 10.1088/1742-6596/1950/1/012035.
[15] N. Goel, S. Kaur, and R. Bala, "Dual branch convolutional neural network for copy move forgery detection," IET Image Process., vol. 15, no. 3, pp. 656–665, 2021, doi: 10.1049/ipr2.12051.
[16] M. N. Abbas, M. S. Ansari, M. N. Asghar, N. Kanwal, T. O'Neill, and B. Lee, "Lightweight Deep Learning Model for Detection of Copy-Move Image Forgery with Post-Processed Attacks," SAMI 2021 - IEEE 19th World Symp. Appl. Mach. Intell. Informatics, Proc., pp. 125–130, 2021, doi: 10.1109/SAMI50585.2021.9378690.