Convolutional neural network-based techniques and error level analysis for image tamper detection

Vijaya Shetty Sadanand, Shruthi Shetty Janardhana, Sowmya Purushothaman, Sarojadevi Hande, Ramya Prakash
Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bengaluru, India

ABSTRACT
Photographs are the foremost powerful and trustworthy media of expression. At present, digital pictures not only serve forged information but also disseminate deceptive information. Users and experts with various objectives edit digital photographs. Images are frequently used as proof of reality or fact, therefore fake news or any publication that makes use of photos that have been altered in any way has a larger chance of deceiving readers. There is a need for a high-resolution image analysis model that processes individual pixels in images and a substantial amount of diverse image data, to detect image falsification. Convolutional neural network (CNN) with error level analysis (ELA) adopted in this research is found to be an ideal deep learning concept for detecting image manipulation. The model exhibited a validation accuracy of 99.6%, 99.7%, and 99.4% for CASIA V1.0, CASIA V2.0 and MICC datasets respectively. The accuracy for handmade tampered images was found to be 99.2%.

Keywords:
Convolutional neural network
Deep learning
Digital image processing
Error level analysis
Image tampering

1. INTRODUCTION
The ease of access to commercial image-altering tools has made digital picture tampering more common. Image manipulation, often known as the editing of images, is the method of changing an image's content. Image tampering is a way of making an image contradict historical facts. The three main categories of digital image manipulation are copy-move attack, image splicing, and image retouching. Image tampering involves the digital alteration of photographs, while image forgery modifies image content to create inconsistencies with historical events. Fake news accompanied by images is more readily embraced and trusted by people, making it challenging to determine the authenticity of an image. Developing image forgery detection technology using deep learning and error-level analysis can help people determine the validity of an image.

The main goal of this study is to develop a tool that uses deep learning techniques to detect image manipulation. With the rise of affordable and sometimes free image-editing software such as Photoshop, Photo Plus, GIMP, and Pixelmator, tampering with digital photos has become a widespread practice. Consequently, photographs have nearly lost their value and status as evidence in various professions. Therefore, detecting digital image tampering to isolate the edited batches from the original ones has become a crucial research topic.

Zhang et al. [1], provide a two-stage deep learning approach to find features to identify altered photographs in a variety of formats in images. They used a stacked autoencoder model in the initial stage to
learn the intricate details of each patch. The conceptual data is combined with individual patches in the second stage, improving the accuracy of the detection. With a fall-out of 4.31% and a precision of 57.67%, it was possible to achieve complete altered region localization accuracy of roughly 91.09% of both tag image file format (TIFF) and JPEG images from the CASIA dataset in their trials. The altered JPEG photos have an accuracy of around 87.51%, outperforming the results of two cutting-edge tampering detection methods that yielded 40.84% and 79.72%, respectively. Yao et al. [2] suggest using a deep learning method for object-based forgery detection in sophisticated videos. The method uses convolutional neural networks (CNNs) to automatically extract features from image patches and includes three preprocessing stages. Positive and negative image patches are balanced using an asymmetric data augmentation technique. The proposed CNN-based model with pre-processing layers has produced good results in experiments. The CGFace CNN was proposed by Dang et al. [3] for detecting computer-generated faces with modified convolutional layers. An imbalanced framework was then created by modifying the layer structure to address the problem of imbalanced data. The proposed model with augmented input achieved an accuracy of 98%.

Srivastava et al. [4] proposed a pixel-based technique for detecting image tampering using speed up robust feature (SURF) features. After preprocessing, significant features are compared against a threshold value, and if altered, the tampered portion is identified. Their testing on a CASIA image dataset showed a forgery detection accuracy of 97% using this technique. Kuznetsov [5] developed an algorithm for detecting splicing in digital images using the VGG-16 CNN. The algorithm achieved high accuracy in comparison to existing solutions and experimental studies, with a fine-tuned model accuracy of around 97.8%. Agarwal and Verma [6] have put forth a practical method for identifying deep learning-supported copy-move forged images. The input to the system used to identify the altered region presented a method that initializes the tampered image. Processes like segmentation, feature extraction, dense depth reconstruction, and ultimately recognizing the tampered regions are all part of the system. The suggested deep learning-based solution can reduce computation time and more accurately identify duplicated regions. Abidin et al. [7] reviewed the use of deep learning for detecting fake copy-move images. Digital image forensics has gained importance due to the increasing sophistication of image editing tools, and researchers have developed various techniques for counterfeit detection, with a recent focus on deep learning.

Muzaffer and Ulutas [8] suggests a deep learning-based framework to identify and localize copy-move frauds. They used a trained AlexNet CNN to extract features of picture sub-blocks and remove erroneous matches. Results show that the suggested technique is better than the cited works. Islam et al. [9] proposes an image fraud detection system that uses discrete cosine transformation (DCT), local binary pattern (LBP), and a mean operator-based feature extraction technique. The system identifies picture forgeries by dividing photographs into fixed-size blocks and applying 2D block DCT. LBP is used to enhance the forging artifacts. The mean of overall LBP blocks is computed to generate a set of features. The proposed approach is tested on four public grayscale and color picture forgery datasets using a support vector machine (SVM).

According to Barad and Goswami [10], traditional methods for detecting image tampering are limited to identifying specific traits in the image. However, deep learning techniques are now being used for better accuracy due to their ability to extract complex information from photos. Doegar et al. [11], image properties were extracted using a GoogleNet deep learning model, and image authenticity was determined using a random forest machine learning technique [12]. The proposed approach was applied to the publicly available dataset MICC-F220 and compared with state-of-the-art approaches to divide the dataset into training and testing datasets.

Wang et al. [13] proposed a deep neural network-based technique for identifying picture inpainting. It uses an improved region identification network to segment the image at the pixel level and distinguish between painted and unpainted regions. The output result is filtered using the overlap ratio of the mask area to create a more accurate region of interest. Thakur and Rohilla [14] used various picture forgery approaches to study deep learning-based image forgery detection methods. While identifying altered images is difficult, deep learning models show promising results. Artificial pictures can be created for training, and merging several models at different scales can result in a general-purpose photo change system for detection. The scientific community must continue to develop techniques for detecting image manipulation and producing anti-forensic measures. The study by Meena and Tyagi [15] presents an approach for detecting imagesplicing forgeries that is based on deep learning. On the CUISDE dataset, the suggested technique performs better than other existing methods with an accuracy of 97.24%. The proposed method achieves its success due to two primary factors. Firstly, the residual noise map generated based on the noise print effectively highlights the tapering artifacts present in the manipulated images. Secondly, the distinctive characteristics that differentiate genuine and manipulated images can be effectively learned by the deep CNN ResNet-50 feature extractor and the transfer learning component of this approach. Employing the SVM classifier, this suggested technique classifies images into two categories: authentic and manipulated. Manjunatha and Patil [16], looked at image forensic deep learning models to find signs of photo alteration. The organization of the book

**Convolutional neural network-based techniques and error level analysis for ... (Vijaya Shetty Sadanand)**
provides a solid platform of organized references. The methods covered in this article include median filtering, Gaussian blurring, copy-move, resampling, JPEG double compression, and cut-and-paste.

A recent study by Xiao et al. [17] proposed a splicing forgery detection approach consisting of a C2RNet and dilated adaptive clustering. The C2RNet extracts differences in picture attributes between unaltered and tampered parts. An image-level CNN is used to reduce computational complexity. The approach learns distinctions between different picture attributes and provides consistent detection performance. The suggested adaptive clustering technique is used to produce the final identified forgery regions. The detection approach delivers encouraging results even under different assault situations. In today's digital era, the manipulation of images has become a significant issue that affects various industries, such as journalism, forensics, and photography. To ensure that visual content is genuine and reliable, it is essential to accurately identify altered photos. Through a combination of ResNet and XceptionNet models with error level analysis (ELA), the research study by Khachane and Mondal [18] has developed a practical and dependable approach with a 98.58% success rate in detecting picture tampering.

Digital content has led to an increase in information pollution, including fake news and manipulated images. It’s hard to determine authenticity, but digital image forensics can help. Madake et al. [19] uses ELA and metadata analysis with deep neural networks to evaluate image originality with good accuracy. A new deep learning architecture that combines AlexNet and InceptionNet has shown promise in detecting passive tampering in a study by Monish et al. [20]. Manipulated images can be used to spread fake news and manipulate information. Detecting these images is difficult and requires a powerful model. Deep learning, using CNNs and ELA, can achieve 91.33% accuracy in detecting forged images with just 9 epochs in a study by Kumar and Srivastava [21]. Ali et al. [22] propose a lightweight, deep learning-based system that identifies fake images in the context of double image compression, achieving an overall validation accuracy of 92.23%. Yancey and Davis et al. [23] used deep learning and object detection to address both problems, combining multiple techniques for higher accuracy. Their multi-stream faster RCNN network, with the second stream summing the ELA and BAG error maps, achieves even greater precision. Deep learning using CNN can detect image forgery and ELA can improve accuracy by 2.7% but can slow processing by 5.6% in a study by Sari and Fahmi [24]. A study with EACN system by Kubal et al. [25] uses error analysis and CNNs to verify image authenticity. It has a 92.10% accuracy rate and provides a strong solution for detecting forged images. This preserves image integrity, safeguards digital content authenticity, and prevents misuse. Sharma et al. [26] provide an overview of image tamper detection methods, including a comparative study of forensic image methods and the limitations of deep learning techniques. The goal is to provide a comprehensive analysis of image forgery detection methods.

The researchers observed that photo tampering detection is a challenging process in most research investigations due to the availability of different software packages. Every aspect is susceptible to interference from processes. The feature used in image tampering is thus a crucial part of the tamper detection procedure. It has been shown that deep learning-based algorithms are capable of automatically learning abstract and complicated landscapes, which are required in place of verifying the identity of changing portions. CNN and RNN are the foundations of contemporary improvements in semantic tampering detection methods in computer vision. CNN is used to assess the substance composition and section outline by mining the categorized characteristics at various intensities. CNN-based architectures [27] show beneficial performance in appreciating chromatic ideas by examining the content of modified sections during object detection and segmentation. This research primarily employs two key approaches: ELA and the utilization of CNN to achieve an outperforming accuracy compared to existing techniques.

ELA is one method for identifying image manipulation is ELA, which determines how to re-save a photograph at a specific level of quality and what the ratio between compression levels should be. Typically, photos with lossy formats (lossy compression) are used for this technique. JPEG images are employed in this data mining. For each 8x8 pixel in a JPEG image, compression is carried out independently. Every 8x8 pixel in a picture must have the same error rate if the image is not altered.

CNN is an example of a feedforward network, in which information only flows in a single direction, from input to output [28], [29]. Even though there are several CNN architectures, most CNNs have convolutional and pooling layers. One or more completely connected layers will come next. Since CNN receives input in the form of images, each pixel can be analyzed when it comes to image classification. In brief, a convolutional layer is utilized as a feature extractor, which assesses how these characteristics are portrayed in the images fed into the CNN. The pooling layer is tasked with lowering the spatial resolution of feature maps in the meantime. In general, numerous convolutional and pooling layers are utilized before fully linked layers to extract representation for more abstract information. The fully linked layer will then interpret these traits and carry out the tasks that demand sophisticated reasoning. The function SoftMax was used for the categorization at the end of CNN.
2. SYSTEM ARCHITECTURE

Figure 1 illustrates the architecture of the proposed system. The various components of the system architecture include data input, data preparation module, ELA, resizing, normalization, label encoding, splitting training and validation data (80%-20%), Pre-processed data, CNN model building, RMSprop optimizer performance measure. Figure 2 shows how data flows through the system.

Data preparation often referred to as the initial step before processing and analysis, involves the tasks of refining and transforming raw data. This essential phase encompasses activities such as reformatting, rectifying data errors, and integrating various data sources to enhance the overall data quality before further processing. An optimizer is a utility employed to adjust the parameters of a model, with its primary role being the alteration of model weights to enhance a loss function. The model's effectiveness is assessed by employing this loss function, and the training of a neural network model necessitates the utilization of an optimizer.

An ML algorithm undergoes training by utilizing a dataset referred to as a training set. This set comprises relevant input data that influences the output, along with corresponding sample output data. The training set is employed to input data into the algorithm, and the resulting output is then compared to the sample output. The result of this comparison is used to adjust the model. A confusion matrix is the easiest approach to gauge the performance of a classification problem when the output can include two or more different types of classes. A confusion matrix depicts a table containing the dimensions "Actual" and "Predicted" as well as the addition of "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)", and "False Negatives (FN)" in each of the two dimensions.

Figure 1. System architecture

Figure 2. Dataflow diagram

2.1. Data preparation

Preparing the data and developing the models are the two main components of architectural design. The first input data consisted of photos in the "*.jpg" format with the following information: the stage of data preparation receives 2,940 photos labeled real and 1,771 images with the label altered. Each image that is input is transformed into an ELA result image during the data preparation stage. The ELA image will then be reduced in size to a 128 by 128 image. The CNN model's training effectiveness is increased by converting raw data to the ELA result image. Because the information in the ELA picture results is less redundant than that in the original image, efficiency can be obtained. The area of the picture with a level error over the limit is the focus of the features produced by the ELA image. Additionally, training the CNN model is becoming
increasingly effective because the pixels in an ELA image typically contain colors that are like or starkly different from the pixels surrounding it.

Input: image from the dataset CASIAv1.0, CASIAv2.0, MICC-F220, Handmade tampered image.
Output: preprocessed ELA converted image data.

2.2. Model building

The size of the image changes after that conversion. Because each red, green and blue (RGB) value only has a value between 0 and 1, CNN must divide each RGB value by 255.0 in the following step to normalize it. This allows CNN to converge more quickly (and reach the global minimum of loss values associated with validation data). The following step is to convert the data's label, where 1 denotes tampered data and 0 denotes true data, into a categorical value. Following that, the dataset was divided into 80% for training data and 20% for validation data. Utilizing training data and validation comes next. The model is then trained by deep learning concept CNN using training data and validation data. RMSprop optimizer is used for optimization during training. The CNN model building's whole architecture is depicted in Figure 1.

The first layer of CNN in the deep learning model utilized has 32 filters and convolutional layers with a kernel size of 5×5. The convolutional layer of CNN has a kernel size of 5×5 and up to 32 filters in its second layer, which also has a 2×2 size of the MaxPooling layer. The second convolutional layer chooses which neurons to create using the kernel initializer and the rectified linear unit (ReLU) activation function to acquire usable signals from the input data.

To avoid overfitting, the MaxPooling layer then added a dropout of 0.25. The fully linked layer, which has 256 neurons and the ReLU activation function, is the following layer. To avoid overfitting, a dropout of 0.5 is added after a fully linked layer. A softmax activation function is used in the output layer. Because the outputs of the conversion procedure to an ELA image can emphasize the key characteristics of recognizing if an image is original or has been correctly updated, only two convolutional layers are required in the architecture utilized.

3. RESULT AND ANALYSIS

Figure 3 shows the input images before ELA conversion. Figure 3(a) depicts the original image before ELA conversion and Figure 3(b) is a tampered image before ELA conversion. Figure 4 depicts ELA converted images. Figures 4(a) and (b) show the ELA converted image of the original and tampered images shown in Figures 3(a) and 3(b) respectively. The proposed method's results have a maximum accuracy of 99.94%. The accuracy curve and loss curve are depicted in Figure 5. Figure 5 shows that the 20th epoch had the highest accuracy. After the fifth epoch, the value of the validation loss began to flatten and eventually grow, which is a sign of overfitting. With early stopping, training is halted when the validation accuracy or validation loss values begin to decline or rise, respectively. The greatest accuracy of the findings produced by the suggested procedure is 99.94%. The 20th epoch was where the best precision was found, as seen in the graph. Overfitting is indicated by a value of validation loss that started to flatten after the fifth epoch and eventually rose. When either the validation accuracy begins to drop or the validation loss begins to increase, the training process will be stopped prematurely. This is possible because the incorporation of ELA-converted images in the model training significantly enhances efficiency, requiring fewer training epochs for convergence. Additionally, the normalization applied to the RGB values of individual pixels expedites the convergence of the CNN model. Figure 6 shows the confusion matrix of a machine learning model's performance on validating data. When actual and predicted values are both 0, it's True Negatives (TN), indicating correct identification of negative class. When the actual value is 0 but the predicted is 1, it's False Positives (FP), and when the actual is 1 but the predicted is 0, it's False Negatives (FN). Actual and predicted values both being 1 are True Positives (TP), indicating correct identification of positive class.

![Input images](a) an example of an original image and (b) an example of a tampered image

Figure 3. Input images (a) an example of an original image and (b) an example of a tampered image
Convolutional neural network-based techniques and error level analysis for … (Vijaya Shetty Sadanand)

3.1. Performance analysis

Table 1 shows the accuracy of different existing experimental methods concerning different datasets:
- It has been noted that accuracy of 99.80% in CASIA v1.0, 97.10% in CASIA v2.0, and 80.02% in Columbia color datasets are attained, respectively, by utilizing DCT and LBP methods.
- In the CASIA v1.0 and v2.0 datasets, the statistical feature of the block DCT co-efficient approach provided an accuracy of 96.81% and 97.34%, respectively.
- The mantissa distribution in digital photos yields an accuracy of 95.50%.
- On several datasets, an accuracy of 93.67% for CASIA v1.0, 97.14% for CASIA v2.0, 87.2% for Columbia color, and 74.44% for MICC datasets, respectively, was found by employing DWT and LBP histogram.
- One of the existing systems, robust forgery detection, was used to detect copy movements and splicing attacks, and it was found to have an accuracy of 95.65% for CASIA v1.0, 96.86% for CASIA v2.0, 72.60% for Columbia color, and 75.55% for MICC datasets, respectively.
- With the proposed methodology, ELA with CNN, an accuracy of 99.6% for CASIA v1.0, 99.7% for CASIA v2.0, and 99.4% for MICC datasets has been achieved. The accuracy of hand-made tampered images was approximately 99.2% for existing models, whereas for the proposed method the accuracy was found to be 99.2%. The proposed method outperformed the existing models in detecting image manipulation.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Dataset and accuracy</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DCT and local binary pattern</td>
<td>96.80%</td>
<td>97.10%</td>
<td>80.02%</td>
<td></td>
</tr>
<tr>
<td>Statistical feature of block DCT co-efficient</td>
<td>96.81%</td>
<td>97.34%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mantissa distribution in digital images</td>
<td>95.50%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DWT and LBP histogram</td>
<td>93.67%</td>
<td>97.14%</td>
<td>87.2%</td>
<td>74.44%</td>
</tr>
<tr>
<td>Robust forgery detection for copy move and splicing attacks</td>
<td>95.65%</td>
<td>96.86%</td>
<td>72.60%</td>
<td>75.55%</td>
</tr>
<tr>
<td>ELA with CNN</td>
<td>99.6%</td>
<td>99.7%</td>
<td></td>
<td>99.4%</td>
</tr>
</tbody>
</table>

4. CONCLUSION

Detecting picture tampering through software packages is a challenging task because of the lack of standardization. The features used to identify tampering play an essential role in its discovery. CNN algorithms have gained popularity in picture forensics, especially in training CNN to recognize the best
properties to categorize camera models. This paper proposes a strategy for identifying altered images that use CNN and ELA. One of the advantages of CNN is that its features are taken directly from the image collection. CNN-based methods can directly learn categorization features from picture data, making them particularly effective in detecting several instances of tampering. The usage of ELA improves model performance and reduces the computational cost of the training process. By combining CNN and ELA, this approach provides an efficient and accurate mechanism for identifying picture tampering.

REFERENCES


Convolutional neural network-based techniques and error level analysis for ... (Vijaya Shetty Sadanand)


BIOGRAPHIES OF AUTHORS

Dr. Vijaya Shetty Sadanand is currently working as a Professor and Head in the Department of Computer Science and Engineering at Nitte Meenakshi Institute of Technology, Bengaluru. She is currently executing a project in the domain of Deep Learning funded by the Vision Group on Science and Technology (VGST). Her research interests include data mining, machine learning, deep learning, and distributed computing. She is a Life member of the Indian Society for Technical Education (ISTE), a member of IEEE and the Computer Society of India (CSI). She can be contacted at email: vijayashetty.s@nmit.ac.in.

Mrs Shruthi Shetty Janardhana is currently working as an Assistant professor at Nitte Meenakshi Institute of Technology. She completed her M.Tech. in Computer Science and Engineering from Visveshwaraya Technological University (VTU), Belgavi, and pursuing her Ph.D. in Computational Intelligence at VTU, Belagavi. She has 10 years of experience in teaching. Her research interests include soft computing, machine learning, and distributed computing. She can be contacted at email: shruthi.shetty@nmit.ac.in.

Mrs. Sowmya Puroshothaman is currently working as an Assistant Professor in the Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bengaluru. She completed her M.Tech. in Computer Science from Mumbai University. She has 7 years of teaching experience. Her research interests include networking, security and machine learning. She can be contacted at email: sowmya.p@nmit.ac.in.

Dr. Sarojadevi Hande is a professor in the Department of CSE at NMIT, Bangalore, India. She is a Ph.D. graduate from the Indian Institute of Science, Bangalore. Her areas of interest include machine learning, cloud computing, system performance, and natural language processing. She has won several national and international recognitions, including Marquis Who’s Who in the World 2016. She can be contacted at email: sarojadevi.n@nmit.ac.in.

Ramya Prakash is currently pursuing her M.Tech. in Computer Science at Nitte Meenakshi Institute of Technology, Bengaluru. She has 16+ years of industry experience with a proven track record of successfully leading multiple teams in developing high-quality software solutions for the healthcare domain, as a software quality analyst and developer. Her current interests are deep and machine learning and the internet of things. She can be contacted at email: ramya.prakash@gmail.com.