I J C T A, 9(40) 2016, pp. 221-225

Image Manipulation Detection using Deep Learning in Tensor Flow

***Kalpana K. and Amritha P. P.**

ABSTRACT

Images and videos becomes one of the principle means of communication these days. Validating the authenticity of the image has been the active research area for last decade. When an image or video is obtained as the evidence it can be used as probative only if it is authentic. Convolution Neural Networks (CNN) have been widely used in automatic image classification, Image Recognition and Identifying image Manipulation. CNN is efficient deep neural network that can study concurrently with the help of large datasets. Recent studies have indicated that the architectures of CNN tailored for identifying manipulated image will provide least efficiency when the image is directly fed into the network. Deep Learning is the branch of machine learning that learns the features by hierarchical representation where higher-level features are defined from lower-level concepts. In this paper, we make use of deep learning known as CNN to classify the manipulated image which is capable of automatically learning traces left by editing of the image by applying the filter that retrieves altered relationship among the pixels of the image and experiments were done in TensorFlow framework. Results showed that manipulations like median filtering, Gaussian blurring, resizing and cut and paste forgery can be detected with an average accuracy of 97%.

Keywords: Convolution Neural Network, Deep Learning, Image Manipulation, Tensor flow

I. INTRODUCTION

Over the past decade, many techniques have been developed to provide the legitimacy of the image. Image Forensics is an act of identifying the forged image by traces left by manipulation operation. There are two methods of forensics approach. In the algorithmic approach [1], algorithm is developed based on the traces left by the editing operation. In this approach when the forger make multiple editing operation then the forensic investigator have to apply multiple algorithm for detection. Sometimes this may cause some new problems or increase the false alarm rate. In steganalysis approach [2] the investigation is done by retrieving pixel relationship to identify the manipulation. Disadvantage in this approach is that identification depends on the pre-selected features of the models. To overcome this issue the Deep Learning based approach [3] is introduced which learn the features directly from the training data. Deep Learning is the subset of machine learning algorithm that represents the content as hierarchical concept, with each concept defined in relation to simpler concepts and it reduces the complexity of identifying the exact dataset which is the vital operation in machine learning. It overcomes the vanishing gradient problem of Artificial Neural Network. Deep learning based tool known as CNN is used to classify the manipulated image which is capable of automatically learning traces left directly from the training dataset.

CNN helps in identifying the manipulated image in which the network is trained previously with the training dataset that tuned the network to identify the given image is altered or unaltered. Major layers of CNN include Convolution, Pooling and Full Connected Layer. Convolution layer includes several filters that convolved with the input image to obtain the feature maps. Filters acts as feature extractor. Pooling is an aggregation action of the input by average pooling or by max pooling and retaining the one value per

* TIFAC CORE in Cyber Security, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University, India, *E-mail: kalpanakumar1812@gmail.com, pp_amritha@cb.amrita.edu*

window. Fully Connected Layer performs the classification based on the produced class scores. The class scores were used in classifying the given image. In this paper we consider Stochastic Gradient Descent (SGD) algorithm to train our network. During training the network updates the kernel coefficients automatically. The weights were learned iteratively during the feed forward and backward pass of the data.

Our primary motivation is to classify image manipulation operations like median filtering, Gaussian blurring, resizing and cut and paste forgery with the help of deep learning in Tensor Flow. TensorFlow [10] is an open source python software library for numerical computation with the help of data flow graphs. Mathematical operations were represented by the nodes and edges represent the tensors communicated between the nodes. The computations are expressed as stateful graph. It supports training and using broader range on a wide variety of heterogeneous platform. It leads to faster computation. The performance of our work is measured in terms of detection accuracy.

The rest of the paper is organized as follows. Section II discusses about the related works in image manipulation detection till now. Section III and IV describes implementation and experimental results. In section V we summarize and conclude the paper.

II. RELATED WORK

Popescu, A. C., & Farid, H. in [1], proposed the method that identifies the manipulation by the traces left by resampling operations. When a resampling is done in a particular image it introduces the periodic correlations and these correlations can be detected with the Expectation/Maximization (EM) algorithm. In [1] two models were considered in which first method considers samples that are correlated to their neighbors, and in second method samples that are not correlated. The EM algorithm is iterative in which the E-step is the probability that each sample belongs to each model is estimated and in the M-step, the specific form of the correlations between samples is estimated. When an image is resampled the even columns and odd rows will be the linear combination of horizontal neighbours and the vertical neighbours form the linear combination for even rows and odd columns. Based on this the probability map can be developed that embody the spatial correlations of the image. The EM algorithm estimates the set of periodic samples that are correlated to the neighbours. This technique identifies the broad range of resampling rates.

Kirchner, M. in [4], EM estimation is replaced by fast linear filtering with fixed coefficients as EM is complex and time consuming. This new method increases the speed of calculation and fast procedure to detect the presence of characteristic peaks in p-map of spatial correlation in the image. Once the error is identified the p-map is calculated with the help of controlling parameters. The variance of prediction residual describes the periodic artifacts of the p-map. Additional performance gain was achieved by faster search for anomalies in p-map cumulative periodogram. The forensic method to identify the contrast enhancement is proposed in [5] by Stamm, M. C., & Liu, K. R. In [5] the manipulation is identified by observing the intrinsic finger print. The statistical traces left behind by the pixel value mapping are referred as intrinsic fingerprinting in an image's pixel value histogram. Contrast enhancement is identified by measuring the strength of high frequency components of an image's pixel value histogram, then comparing this measurement to a predefined threshold.

The new median filtering forensic techniques is proposed in [6] that identify the median filtered image in four steps. Initially the Median Filter Residue (MFR) extracts the median filtering features and suppresses the images edge content and then obtains the statistical feature by fitting the MFR to the Autoregressive (AR). To get the single AR model, average the corresponding AR coefficients. These coefficients are used to train the SVM and in turn classify the median filtered or unaltered images. The above approaches lead to advancement in forensics but it raises several new problems if multiple editing is done on the particular image. To overcome this problem new idea is proposed with reference to the pixel relationship.

In [7] the analysis is focused on identifying by analyzing the joint distribution of pixel value prediction errors, then extracting the detection features based on these joint Distribution errors. In [8] an universal forensic approach is applied by building the Gaussian Mixture Model of the image patches and identifying the forgery by comparing the log likelihood of an image under the GMM for different manipulations. Chen, J., Kang, X., Liu, Y., & Wang, Z. J. in [9], proposed a new approach for performing the image editing detection that is capable of automatically learning traces left by editing is achieved by using deep learning approach. Deep learning approach learns by representing the concepts in hierarchical order (i.e) higher level concepts learn from lower level concepts. CNN models with the raw image pixels as inputs does not yield good performance, so one additional filter layer is added to the conventional model. Through this filter layer, the MFR of an image is obtained. Median filtering and cut and copy image manipulations can be detected by this method. Bayar, B., & Stamm, M. C.in [3], proposed the method to identify the universal image manipulation by applying the prediction error filter that is capable of suppressing the image content and retrieving the pixel relationship among pixel. In [3] implementation is done using caffe framework where each node is a layer. Caffe is not considered flexible for new layer types and the computation is comparatively slow. To obtain higher performance and improved efficiency in detecting the forged image in our paper we have implemented manipulation detection using deep learning in TensorFlow.

III. MANIPULATION DETECTION IN TENSORFLOW

CNNs are able to learn the content of the image to classify the given image. It retrieves the content of the image which is not suitable for identifying the manipulation. To overcome this issue, Belhassen Bayar [3] proposed the filter layer called prediction error filter which suppress the content of the image and retrieve the local structural relationship that exist between pixels and this system was implemented with the help of caffe framework. It earns the greater accuracy. To achieve greater performance [11] we implemented the manipulation detection in TensorFlow which yields higher level of flexibility and faster computation.

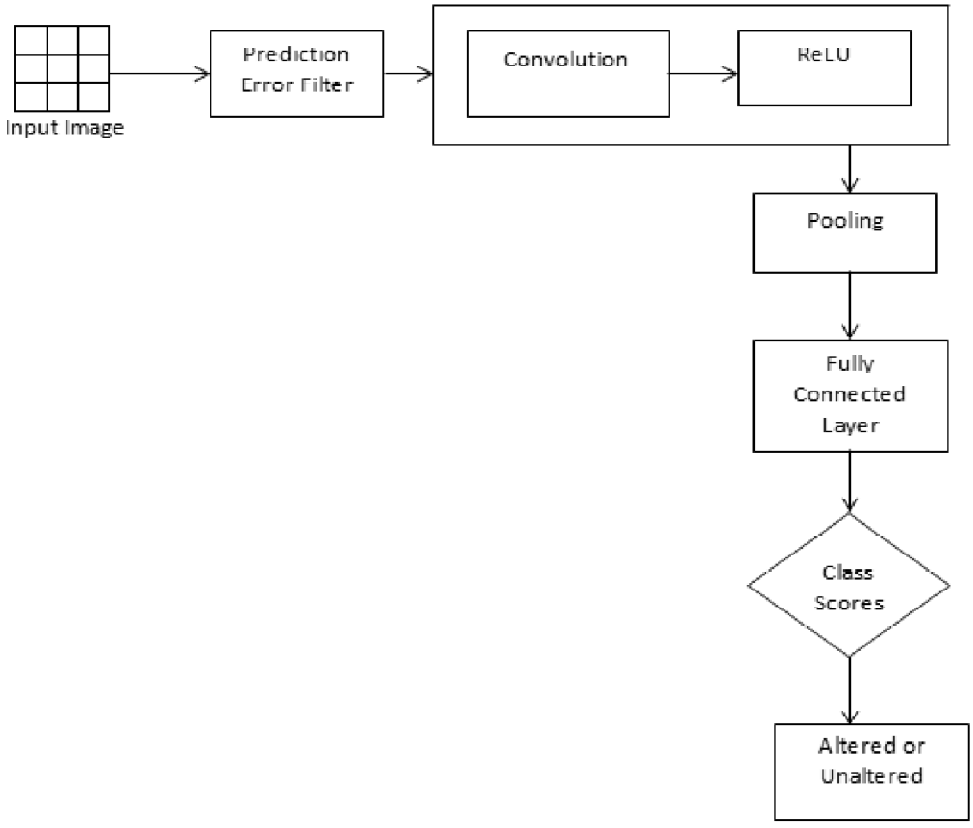

Figure 1: Flow graph for manipulation detection system with a single layer

If a raw pixels of the image is fed directly to the convolution layer it does not yield high accuracy so initially an altered or unaltered image is given as input to the Prediction error filter layer which is fed to the convolution layer as shown in fig.1. The convolution layer is followed by Rectified Linear Unit (ReLU). The obtained feature maps are then fed to pooling layer. Here we have considered max-pooling operation. The pooled values are then fed to the fully connected layer where the classification is done using softmax operation. Depending on the class scores we can detect an image is altered or not. Detail description is given below.

Training algorithm for convolution filter layer

Initialize weights randomly

i=1

While i < max_iteration

Set $w_k^{(1)}(0,0) = 0$

Normalize the weight such that $\sum_{l,m\neq 0} w_k^{(1)}(l,m) = 1$

Set $w_k^{(1)}(0,0)$ = -1

Make forward pass

Update weight using SGD and apply backpropagation

i=i+1

if accuracy reaches the training value

exit

end

The new prediction layer placed before the convolution layer will predict the error by subtracting the central value from the center of the filter window. Prediction error filter is of size 5 x 5 with stride value 1. We used 12 kernels which results in 223 x 223 x12 feature maps. This convolution is not stepped to non–linear function mapping because it contains the traces of the manipulated image. In the above architecture we have taken two convolution layer in which first layer has 64 kernels of size 7 x 7 x 12 with stride value 2 that yields 112 x 112 x 64 feature maps. The second layer has 48 kernels with stride value 1. The size of the kernel is 5 x 5 x 64. The second layer yields feature maps 56 x 56 x 48. The obtained output is then given to the Non-linear layer that uses variety of Non-Linear functions to signal distinct identification of likely feature in hidden layers. We had taken ReLU as non-linear activation function, $f(x) = max\ (0, x)$. It trains the network several times faster than other activation function. The input and output of this non-linear layer is of same size. The extracted feature maps are then given to the Pooling layer. The pooling value is calculated by taking the maximum value within the specified window. Kernel size is 3 with stride value 2. Pooling reduces the resolution of the feature maps to make the features robust against the noise and distortion. It reduces the first convolution layers feature map to 56 x 56 x48. When the max pooling applied in the second layer it reduces the value to 28 x 28 x 48. Pooling is followed by the normalization in which the central value is normalized by surrounding pixels. The pooled value is fed to the fully connected layer in which every nodes of the previous layer is connected to the every node in the next layer. Fully connected layers are the last layer that gives the final output of the given image based on the class scores. Dropout techniques used in the fully connected layer drops the particular nodes (pixels) based on the probability value. In this method we considered the probability value as 0.5. When a node is dropped it does not participate in both feedforward and backward pass.

IV. EXPERIMENTAL RESULTS

Our dataset contains 100 images of size 256 x 256 and each image were subjected to manipulations like median filtering, Gaussian blurring, resizing and cut and paste forgery. We have created the training dataset of image size 227 x 227 by cropping down using 100 images each and testing dataset with 50 images each. Table 1 shows the detection rate for various manipulations. Below tabular results concludes that using CNN we are able to distinguish between unaltered and manipulated images with atleast 97% accuracy.

Table
Detection accuracy rate

	Original	*Median Filtering*	*Gaussian Blurring*	*Cut and Paste forgery*	*Resizing*
Original	97.40%	0.23%	0.29%	0.3%	0.43%
Median Filtering	0.21%	97.23%	0.09%	0.12%	0.12%
Gaussian Blurring	0.01%	0.18%	98.01%	0.09%	0.06%
Cut and Paste Forgery	0.02%	0.04%	0.14%	97.9%	0.01%
Resizing	0.03%	0.05%	0.21%	0.07%	96.9%

V. CONCLUSION

In this paper we have implemented the deep learning based image manipulation detection in TensorFlow framework that can automatically learn features to detect image manipulations like Median filtering, Gaussian blurring, resizing and cut and paste forgery. We placed an error filter to get only the pixel relationship instead of learning features of image content. Through experiments, we demonstrated that CNN-based deep learning approach was able to detect manipulation with an average accuracy of 97%. In future we plan to test our network by increasing the size of the data set to improve accuracy of detection.

REFERENCES

[1] Popescu, A. C., & Farid, H. Exposing digital forgeries by detecting traces of resampling. *IEEE Transactions on signal processing*, *53*(2), (2005), pp.758-767.

[2] Pevny, T., Bas, P., & Fridrich, J. Steganalysis by subtractive pixel adjacency matrix. *IEEE Transactions on Information Forensics and Security*, *5*(2), (2010) , pp. 215-224.

[3] Bayar, B., & Stamm, M. C. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security* (2016), pp. 5-10.

[4] Kirchner, M. Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue. In *Proceedings of the 10th ACM workshop on Multimedia and security* (2008), pp. 11-20.

[5] Stamm, M. C., & Liu, K. R. Forensic detection of image manipulation using statistical intrinsic fingerprints. *IEEE Transactions on Information Forensics and Security*, *5*(3), (2010), pp.492-506.

[6] Kang, X., Stamm, M. C., Peng, A., & Liu, K. R. Robust median filtering forensics using an autoregressive model. *IEEE Transactions on Information Forensics and Security*, *8*(9), (2013), pp.1456-1468.

[7] Fridrich, J., & Kodovsky, J. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, *7*(3), (2012), pp. 868-882.

[8] Fan, W., Wang, K., & Cayre, F. General-purpose image forensics using patch likelihood under image statistical models. In *Information Forensics and Security (WIFS), 2015 IEEE International Workshop on*(2015), pp. 1-6.

[9] Chen, J., Kang, X., Liu, Y., & Wang, Z. J. Median filtering forensics based on convolutional neural networks. *IEEE Signal Processing Letters*, *22*(11), (2015), pp.1849-1853

[10] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., & Ghemawat, S. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. (2016), pp.1603-04467.

[11] Shi, S., Wang, Q., Xu, P., & Chu, X. Benchmarking State-of-the-Art Deep Learning Software Tools. (2016), pp. 1608-07249.