EasyChair Preprint

№ 5569

A Computationally Inexpensive Way to Detect and Perform Segmentation on Morphed Images Using CNN

Neeraj Kumar, Divya Singh, Himja Uppal and Stuti Mehla

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

May 20, 2021

A computationally inexpensive way to detect and perform segmentation on morphed images using CNN

Neeraj Kumar, Divya Singh, Himja Uppal, Stuti Mehla

Computer Science Department
Panipat Institute of Engineering
and Technology
Haryana, India

Abstract. In the last few years, the amount of image data generated by the use of image (and video) processing software such as GNU Gimp, Adobe Photoshop has increased enormously as social networking sites such as Facebook and Instagram have come into force. Such photos are key sources of fake news to propagate radical ideologies, swing public opinion and misused for incitement by the crowd. Being able to solve such a problem and integrating that solution to the social media websites might be able to save the people from becoming a victim of mis-information and propaganda. In this paper we propose a way to detect tampered/morphed images on social media and generate a binary mask of the tampered region using computationally inexpensive pre-processing algorithms and convolutional neural networks.

Keywords: CNN, Tampered Images, Classification, Segmentation, Fake Images

1 Introduction

Tampered/morphed photos propagated across the web and social media have the potential to mislead, emotionally distress and influence public sentiment and behavior. To understand this some light can be shed on the cases where the existence of such images created quite a lot of controversies.

According to the study conducted by two Massachusetts Institutes of Technology researchers on the extent of fake images used in the Indian Politics roughly 1 in 8 images shared in the WhatsApp groups were fake [1]. Similar things are happening in the currently ongoing 2020 presidential elections, where fake images are being shared online trying to damage or uplift the reputations of both the candidates [2]. Apart from this, existing images are often doctored and used as a hoax to polarize and incite violence in communities [3]. Not only that, these days totally fake faces can be generated from scratch using AI which makes it very hard to run a background check on such images [4].

We are making an attempt to fix the above problem using the power of AI itself no matter if the image is spliced or warped.

2 Literature Review

Prior to coming up with a methodology to classify and generate binary masks for the forged images, we had to go through a couple of research papers. The first job was to classify the images as real or fake which is just a typical classification task. We used the existing knowledge from the paper: *Image Classification using Convolutional Neural Networks by Deepika Jaiswal, Soman KP and Sowmya Vishvanathan* [5] where we learned how we can use a CNN and supply it images from segregated in different classes to learn from and predict the result. We learned about how different layers, parameters and training epochs / duration can affect the ultimate result.

However, classifying the images as forged and real is too big of a task as it can be very computationally intensive to differentiate between the images just with a normal view. Hence to alleviate this task we came across the paper : *An evaluation of Error Level Analysis in Image Forensics by Nor Bakiah Abd Warif, Mohd Yamani Idna Idris, Rosli Salleh, Ainuddin Wahid* [6]. From this paper we got to know about how the variations in the compression level in the original image and the spliced portion can be visualized and hence can be used as a pre-processing for the images to be classified.

For the task of image segmentation, where we intend to generate a binary mask for the tampered images by having a white portion for the doctored region and black for everything else, we referred to the paper : *SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation by Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla* [7]. From this paper we derived the understanding of Encoded-Decoder architectures in Neural Networks which is used as a base for creating SegNets for segmentation task. For our required task, we created a custom SegNet like network and trained it with the ELAs of tampered images as an input and their ground truth binary masks as the output.

3 Proposed Work

3.1 Detection

The proposed technique for detecting the fake images involves generating an ELA of an image which stands for Error Level Analysis. What the Error Level Analysis process does is that it finds out the compression rate variations in every portion of the image. The theory behind it is that every image has a different compression ratio due to different set of RGB colors involved, the bit depth, the resolution, file format etc. And when an image is spliced onto the other or a portion from the original image is simply scaled. Then it's uniform compression rate is disturbed and opens the door for detection of the morphed areas using the given technique.

Steps involved for performing ELA:

1. Load the image and then save a temporary copy of it and save it with a little degraded quality.
2. Compare the absolute difference between the intensity of the corresponding pixels in the original and the temporary image.
3. From the list of absolute difference between the pixels of the two images in each band, find out the max values from each color band and find out the max difference.
4. Use the max difference to find the scale factor to brighten up the final ELA image.

Using the above technique, convert all the images of a given dataset into their respective ELAs and keep them in the memory. For the sake of a benchmark, the paper uses the CASIA1 dataset which contains the tempered images of both copy-move as well as the spliced images. Here our main focus is going to be on the spliced images as they are the one which do the most harm to the public interest.

After having all the ELAs we store them in a list with their corresponding label of fake or real. Later, the arrays are normalized by dividing all their elements by 255. We can harness the power of deep learning to automate the classification in a way that can be integrated to various platforms. For this we will be using a Convolutional Neural Network and train it by using those ELAs we generated. One of the benefits of generating the ELA is that we already have the features extracted which can be learned by even a relatively simple CNN. The architecture of the CNN used for this paper is as follows:

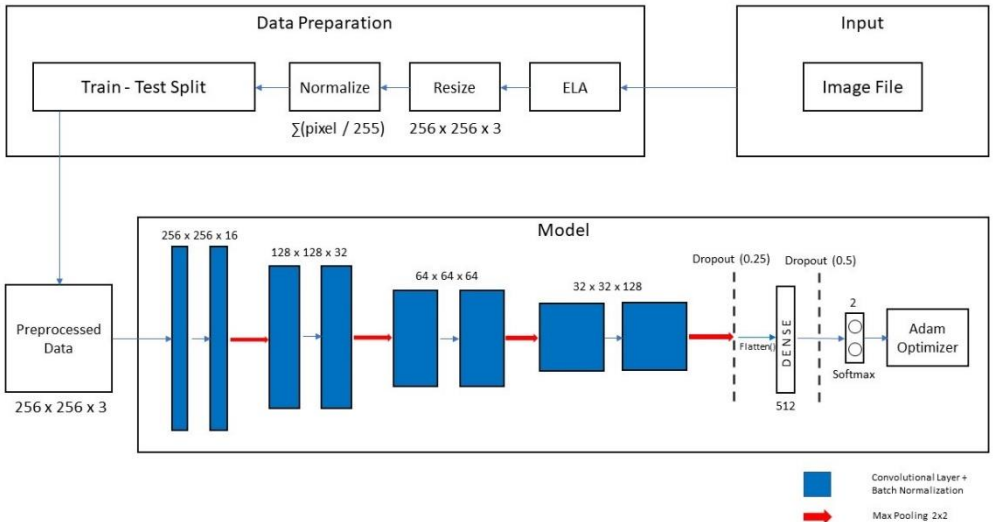

Fig. 1. Architecture for the classification network

There are 4 blocks, each with 2 identical convolutional layers with batch normalisation process and a max pooling operation with a 2x2 window at the end. These layers down scale the input image and extracts more and more information from the input by increasing the filter size. Later 25% of the nodes from the overall output are dropped, flattened to 1D and then supplied to a dense layer with 512 nodes. The output of the sense layer is again made to drop 50% of its nodes to avoid overfitting and is lasty supplied to a dense layer with 2 nodes and softmax as the activation function which predicts the probability of the input image being fake and real.

3.2 Masking

The detection part which was discussed above deals with the 'what' portion of the system. That is, what is we're trying to find out (fake vs real image). In the masking part we are more focused on the 'where' part that is, in which region the forgery has taken place. In this case there are just 2 classes for the segmentation i.e. the part which is doctored and the part which is not and thus we will be doing semantic segmentation [8] for creating the masks for those regions. The proposed architecture of the neural network for the binary image segmentation follows the encoder decoder based SegNet like architecture.

Here, we take the ELA of the image classified as fake and fetch the blue channel array out of it. We normalize the arrays by dividing them with the maximum value element it contains, such that it has the maximum contrast. Then we pass it into our segmentation network whose architecture is given as follows in order to get a binary mask for the image which would tell us about the region of tampering.

The below network consists of 8 blocks in total where the first 4 blocks are downscale the image and increase the number of filters using convolution + batch normalisation and max pool operations with a window of 2x2, then the last 4 blocks are used for the upscaling and gradually reducing the number of filters back to that of the input by having an up-conv layer at the beginning of each block.

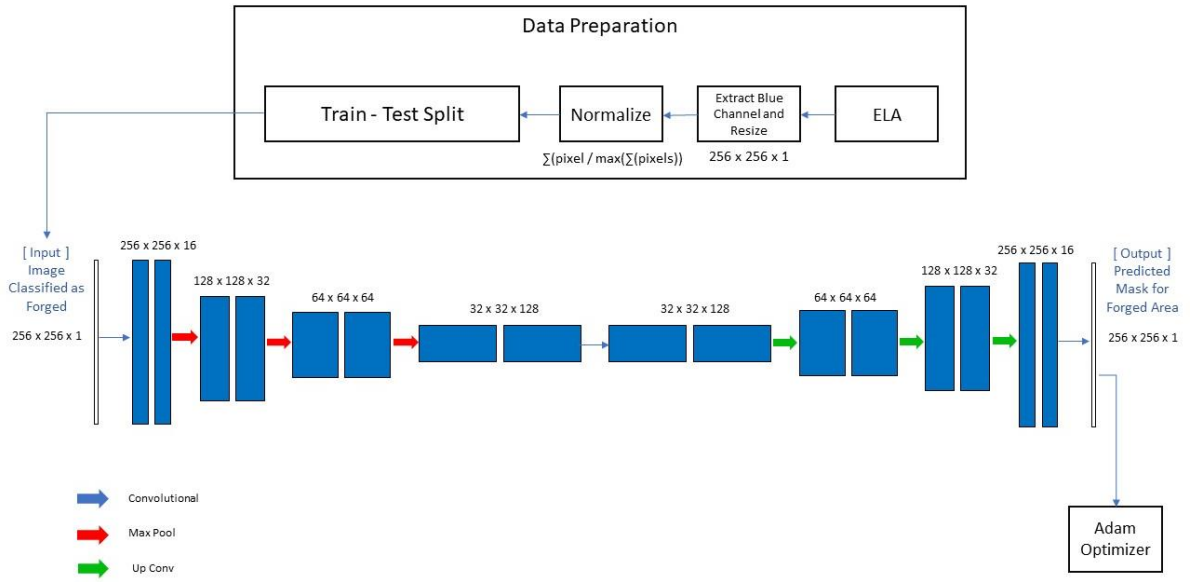

Fig. 2. Architecture for the segmentation network

4 Results

The above classification and segmentation models have given quite good results. The segmentation model reaches 98% accuracy on the train set and 92% accuracy on validation set when trained for 100 epochs which took around 40 minutes. The classification model reaches 98% accuracy on the train set and 81% accuracy on validation set when trained for 100 epochs which took around 20 minutes. Although the accuracy is pretty modest, considering

the hardware limitations on our end, limited dataset and the given training time, these are very promising and the models can show much better results with increase in training time, better hardware and a wider dataset.

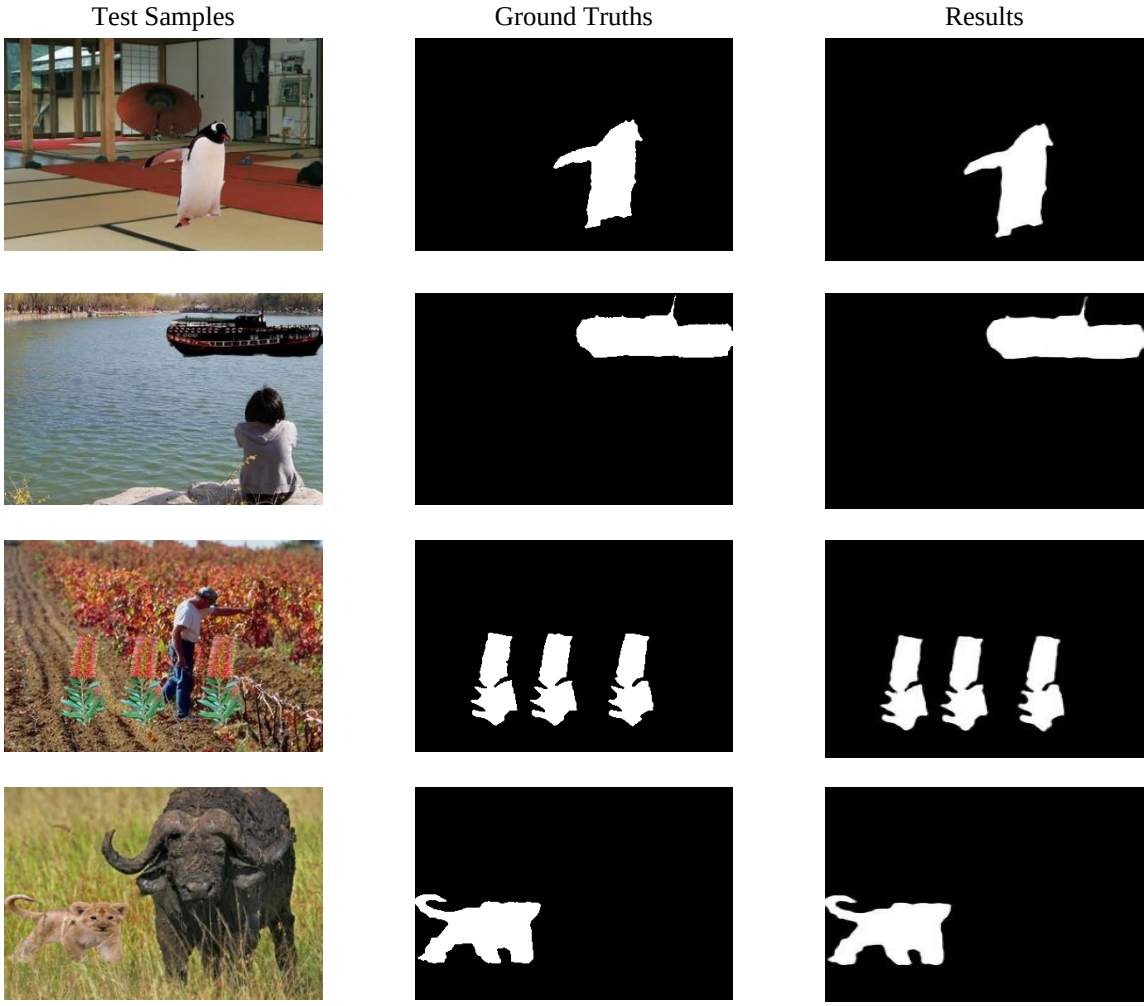

Fig. 3. Results

References

[1] Times of India – 1 out of 8 photos in political WhatsApp groups misleading. Accessed 26 June 2020: https://timesofindia.indiatimes.com/india/1-out-of-8-photos-in-political-whatsapp-groups-misleading/articleshow/76887454.cms

[2] Brookings - Digital threats to campaign 2020: Doctored images, and widespread disinformation. Accessed 26 June 2020: https://www.brookings.edu/blog/techtank/2019/07/11/digital-threats-to-campaign-2020-fake-nudes-doctored-images-and-widespread-disinformation

[3] The Quint - Viral Black Lives Matter Poster on 'Arabs' is Photoshopped & Fake. Accessed 26 June 2020: https://www.thequint.com/news/webqoof/black-lives-matter-protesters-carried-placards-about-arabs-no-fake-image-fact-check

[4] Forbes - Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared. Accessed 28 June 2020 https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/#29eeb09d7494

[5] Deepika Jaiswal, Soman KP and Sowmya Vishvanathan, ‘Image Classification using Convolutional Neural Networks’, International Journal of Advancements in Research & Technology, Volume 3, Issue 6, June-2014 1661 ISSN 2278-7763

[6] N. B. A. Warif, M. Y. I. Idris, A. W. A. Wahab and R. Salleh, "An evaluation of Error Level Analysis in image forensics," 2015 5th IEEE International Conference on System Engineering and Technology (ICSET), 2015, pp. 23-28

[7] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 1 Dec. 2017

[8] Analytics Vidya - A Step-by-Step Introduction to Image Segmentation Techniques. Accessed 17 August 2020 https://www.analyticsvidhya.com/blog/2019/04/introduction-image-segmentation-techniques-python/