Learning to Identify Image Duplications in Scientific Publications

Ghazal Mazaheri, Kevin Urrutia Avila, Amit K. Roy-Chowdhury
University of California, Riverside, CA 92521

1 Abstract

To help scientific integrity officers and journal/publisher reviewers monitor if researchers stick with scientific community standards, it is important to have a solid procedure to detect duplication as one of the most frequent types of manipulation in scientific papers. Therefore, in this work we propose a framework that combines image processing and deep learning methods to classify images in the articles as duplicated or unduplicated ones. We show that our method leads to a 90% accuracy rate of detecting duplicated images, a $\sim 13\%$ improvement in detection accuracy in comparison to other manipulation detection methods.

keywords– research integrity, image duplication, deep learning

2 Introduction

Results in publications can lead to new drugs, products, treatment options, etc. and can have a huge impact on society. Thus, it is very important to protect reearch publications from plagiarism or duplication. In this paper, we focus on the problem of identifying image duplication in research publications.

Main contributions. We propose a new approach to detect image duplication in scientific articles building upon three steps of image extraction, region of interest generation and copy-paste detection using Siamese network. Our method leads to $\sim 13\%$ of improvement in detection accuracy in comparison to other manipulation detection methods. We also show how effective the pre-processing steps are by comparing our method to other state-of-art manipulation detectors where the raw figures extracted out of papers are fed to them.

3 Methodology

In this section, we present our framework for image duplication detection. A pictorial flow of our image duplication detection framework is presented in Fig. 1. The proposed method utilizes three main steps. 1. Extracting the figures out of papers. 2. Region of Interest (ROI) generation. 3. Duplication detection using Siamese network.

Figure Extraction. In the first step, we use DOI of clean and duplicated papers along with figure numbers from the dataset [1] to extract the images and prepare them for training phase. Fig. 1 shows this process. As we can see, extracted figures from papers contain extra miscellaneous noise that would interfere with the quality of the final training data. Existence of unwanted data in extracted images necessitates the ROI generation process we explain in the following section.

ROI Generation. In the next step, we extract the regions of interest excluding texts, graphs and charts. It is known that western blots, the parts of the image that duplication occurs mostly, generally come in the form of rectangular blocks. Therefore, it becomes clear that if anything is not connected to these blocks or found in these blocks it is unnecessary noise. This part of the process is done by removing any part of the image that is not connected to the larger blocks and replace it with a black mask in that region of the image. Following the ROI generation step, we segment the western blots knowing that they are rectangular blocks in the output figures from previous step. This is done by finding the contours within the image.

Duplication Detection using Siamese Network. We prepared both copy-paste pairs and different (unduplicated) pairs from segmented figures extracted from the previous step. After pairing the images, they are fed into a Siamese network to learn the similarity between images. The ability to learn from very little data made Siamese networks suitable for our task with a few number of images in the dataset. To compare two images, each image is passed through one of two

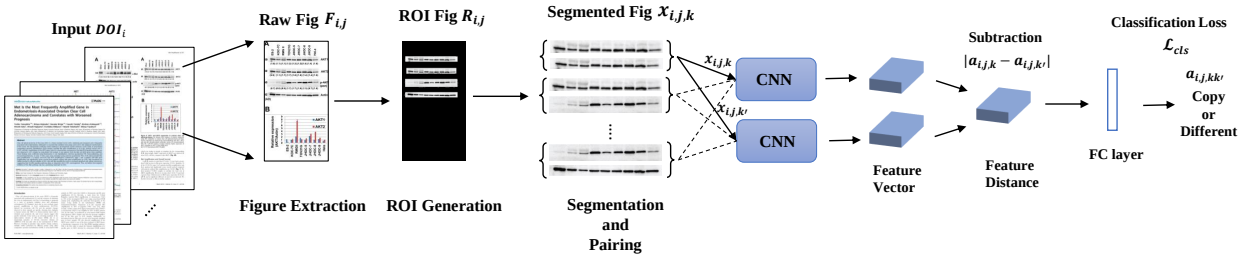

Figure 1: This figure represents our approach for image duplication detection in scientific articles. The proposed method includes pre-processing phase (figure extraction and ROI generation steps) and detection step.

identical sub networks that share weights. For each branch of Siamese network, we use 4 layers of convolution followed by RELU and maxpooling.

4 Results

In this section, we report results of our experiments on the dataset we prepared to investigate the efficacy of the proposed method. We utilize the output images of figure extraction and ROI generation steps as the input for two state-of-the-art approaches (MesoNet [2] and XceptionNet [3]). For our proposed method, we exploit the output images of segmentation and pairing step. Different steps of dataset preparation are shown in Fig. 1. XceptionNet and MesoNet achieve higher classification accuracy with ROI Figs as the input images (76%, 68.78% respectively) in comparison to Raw Figs (72.51%, 62.54% respectively) (shown in Fig. 1). This proves how image pre-processing is effective in the accuracy of duplication detection. To increase the accuracy, we segment out ROI Figs and pair them to feed into the Siamese network. This approach increases the accuracy dramatically. Our approach achieves **90.86%** of detection accuracy which is $\sim$ **13%** higher than XceptionNet accuracy.

5 Conclusion

In this work we propose a method for detecting the duplication of images within a scientific paper. Our framework emphasizes the importance of image pre-processing steps to prepare appropriate dataset prior to application of deep learning methods.

6 Acknowledgement

We thank Dr. Elisabeth Bik for her critical feedback and advice. We also appreciate her work for preparing corresponding dataset.

References

[1] E. M. Bik, A. Casadevall, and F. C. Fang, "The prevalence of inappropriate image duplication in biomedical research publications," *mBio*, vol. 7, no. 3, 2016. [Online]. Available: https://mbio.asm.org/content/7/3/e00809-16

[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec 2018, pp. 1–7.

[3] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1800–1807.