Impact of Data Preparation and CNN's First Layer on Performance of Image Forensics: A Case Study of Detecting Colorized Images

Weize Quan
NLPR, Chinese Academy of Sciences (CAS) / University of CAS, China
University Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France
qweizework@gmail.com

Kai Wang
University Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France
kai.wang@gipsa-lab.grenoble-inp.fr

Dong-Ming Yan*
NLPR, Chinese Academy of Sciences (CAS) / University of CAS, China
yandongming@gmail.com

Denis Pellerin
University Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France

Xiaopeng Zhang
NLPR, Chinese Academy of Sciences (CAS) / University of CAS, China

ABSTRACT

In the field of image forensics, many convolutional neural network (CNN)-based forensic methods have been proposed and generally achieved the state-of-the-art performance. However, some questions are worth studying and answering regarding the trustworthiness of such methods, including for example the appropriateness of the discriminative information automatically extracted by CNN and the generalization performance on "unseen" data during the testing phase. In this paper, we study these questions in the case of a specific forensic problem of distinguishing between natural images (NIs) and colorized images (CIs). Through a series of experiments, we analyze the impact of data preparation and setting of the first layer of a recent state-of-the-art CNN-based method on the detector's forensic performance, in particular the generalization capability. We obtain some interesting observations which can serve as useful hints for carrying out image forensics experiments. Moreover, we propose a very simple method to improve the generalization performance of colorized image detection by combining decision results from CNN models with different settings at the network's first layer.

KEYWORDS

image forensics, colorized image, JPEG compression, convolutional neural network, generalization.

ACM Reference Format:
Weize Quan, Kai Wang, Dong-Ming Yan, Denis Pellerin, and Xiaopeng Zhang. 2019. Impact of Data Preparation and CNN's First Layer on Performance of Image Forensics: A Case Study of Detecting Colorized Images. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19 Companion), October 14–17, 2019, Thessaloniki, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3358695.3360890

*Corresponding author.

WI '19 Companion, October 14–17, 2019, Thessaloniki, Greece

ACM ISBN 978-1-4503-6988-6/19/10…$15.00
https://doi.org/10.1145/3358695.3360890

1 INTRODUCTION

With today's rapid growth of digital data communication, many of us communicate in and get information from cyberspace. We usually pay much attention to convenience, but have few considerations for the security of numerical data circulating on the internet, such as digital images. In consequence, people can be easily misled by fake contents in cyberspace, and the detection of such contents has received more and more attention[1] and become an important research direction in the field of cybersecurity and information forensics. In the meanwhile, convolutional neural network (CNN) has recently obtained notable success in computer vision and pattern recognition. An important reason is that CNN attempts to automatically learn hierarchical representation from available data in an "end-to-end" manner. Inspired by this success, many researchers have proposed CNN-based approaches for image forensics. For example, CNNs have been used to identify camera model [1], to expose image forgery [14], and to detect synthetic images [9, 11]. These CNN-based forensic methods in general work better than traditional handcrafted-feature-based approaches. Despite this, some questions hidden behind the high performance are worth studying and answering, including the following ones: What is the CNN model using as the discriminative information, *i.e.*, is it the "essential" difference between different kinds of images? Is the CNN just overfitting on training data in some aspects that are not the primary factors for the considered forensic problem? How can CNN generalize well on "unknown" data during the testing stage? These questions are closely related to the trustworthiness and the practical applicability of CNN-based forensics.

In this paper, we focus on the case of CNN-based colorized image detection and try to study the above questions. State-of-the-art colorization algorithms, more or less leveraging the powerful capacity of deep neural networks, can automatically colorize a grayscale image to obtain a high-quality color image. Fig. 1 shows a group of

[1] See https://www.cbc.ca/news/technology/fighting-fake-images-military-1.4905775

Figure 1: From left to right: each row includes a natural image taken from ImageNet [3] and three colorized images generated by the colorization algorithm proposed in [8], [13], and [7], respectively.

images, the left-most column is the original color images, and the remaining three columns are colorized images produced by three recent and advanced colorization algorithms (respectively with the name Ma [8], Mb [13] and Mc [7], from left to right), which take the grayscale version of the left-most column as input. It is indeed difficult to distinguish, with naked human eyes, which images are artificially colorized. Accordingly, distinguishing between natural images (NIs) and colorized images (CIs) has drawn increasing interest among the image forensics research community [6, 15].

Recently, Guo *et al.* [6] first considered and studied this new and important forensic problem. According to the statistical difference between NIs and CIs in the hue, saturation, dark, and bright channels, two different features were designed to catch this forensic difference. Then, they trained the support vector machine (SVM) classifiers to identify fake colorized images. Zhuo *et al.* [15] achieved the state-of-the-art detection performance on the experimental database shared by [6], by making use of an advanced CNN-based color image steganalyzer called WISERNet (WIder SEparate-then-Reunion Network) [12]. WISERNet is particularly good at detecting weak signals in images and its first layer has thirty 5×5 residual filters from the well-known steganalytic Spatial Rich Model (SRM) [5]. As shown later in this paper, we find that data preparation and setting of CNN's first layer can have big impact on the forensic performance of WISERNet, especially the *generalization* capability. Here the generalization (or blind detection) performance indicates the detection performance on testing data in which CIs are generated by an "unknown" colorization method with regard to the training procedure, *e.g.*, when WISERNet is trained with CIs from Ma [8] and later tested on CIs from Mb [13] and Mc [7].

The remainder of this paper is organized as follows. Section 2 presents the studied problem and technical details. Section 3 reports the experimental results and our analysis. Section 4 draws the conclusions and proposes some future working directions.

2 DATA PREPARATION AND NETWORK

2.1 Motivation

For this specific forensic problem of colorized image detection, we have some observations about data and network: CIs shared by the authors of [6] and used in [6, 15] are in a *lossless* format without compression on the artificially generated color information, and NIs from ImageNet are in the *lossy* JPEG format; in the meanwhile, the weights of first layer of WISERNet are initialized with SRM residual filters [5] and untrainable[2]. Therefore, it is natural to raise the following question: Does WISERNet, as used in [15], rather capture the difference of processing history between NIs and CIs (*i.e.*, JPEG compressed or not), or the desired "essential" color difference? In order to answer this question, in this work, we experimentally study the impact of two important but until now ignored and underestimated factors on CNN's forensic performance as follows: (1) we study the impact of data by constructing two datasets in which CIs are with/without JPEG compression, and (2) we study the impact of network by adopting two different strategies for the setting of the first layer of WISERNet.

2.2 Data Preparation

We construct two sets of data and the only difference is whether CIs are JPEG compressed or not. Following [6] and [15], three state-of-the-art colorization algorithms (Ma [8], Mb [13], and Mc [7]) are adopted for producing CIs. NIs come from ImageNet dataset [3]. We use 10,000 natural images from ImageNet validation dataset to construct training and validation dataset (with the ratio 4:1). The exact indexes of these images can be found from [8]. Then, we remove 899 grayscale images and 1 CMYK image from the remaining 40,000 images of ImageNet validation dataset (the total number of images in this dataset is 50,000), and obtain 39,100 NIs to construct testing dataset. Note that, the magnitude of testing dataset is far larger than the settings reported in [6] and [15]. We employ the three colorization methods mentioned above to produce the corresponding colorized images. In addition, we construct another dataset where we only replace the CIs (the original output of colorization algorithms) with a JPEG compressed version. In details, the compressed CI is generated in the following way: given an original CI, we first obtain the quantization table of the corresponding NI (*i.e.*, the NI which shares the same grayscale version) using the Matlab JPEG Toolbox [10]. Then, we estimate the quality factor from the above quantization table of NI using the method proposed in [2] and compress the CI with estimated quality factor. Hereafter,

[2]Confirmed by email with authors of [15].

Table 1: K_F of validation dataset in color space of YCbCr. "X-C" means the JPEG compressed version of colorized images produced by X colorization method.

Channel	NI	Ma	Mb	Mc	Ma-C	Mb-C	Mc-C
Y	0.3491	0.3227	0.3121	0.3144	0.3650	0.3667	0.3650
Cb	0.6757	0.0661	0.0596	0.0434	0.6552	0.6318	0.6741
Cr	0.7023	0.0653	0.0625	0.0489	0.6185	0.6548	0.6811

for ease of presentation, dataset with/without JPEG compression means that CIs in this dataset are with/without JPEG compression.

It is necessary to justify the JPEG compression of CIs mentioned above and prove that the artificial color information in CIs before compression is indeed considered as uncompressed. To this end, we quantitatively analyze the JPEG blocking artifacts of the two datasets with the forensic measure of K_F [4]. Larger K_F means stronger JPEG blocking artifacts. We analyze the blocking artifacts of NIs, as well as CIs with or without JPEG compression, in the color space of YCbCr, which partitions images into luminance and chrominance and which is the space adopted by JPEG standard. We calculate K_F of validation dataset (2,000 images for each class) and report the average values in Table 1. Compared with the column of "NI", K_F of "Y" of the columns of "Ma", "Mb" and "Mc" are very similar, while that of "Cb", "Cr" have a large gap. This is an experimental proof that the color information in CIs without JPEG compression is in fact considered as forensically uncompressed. Furthermore, this gap is significantly decreased after compressing the original CIs with the same quality factor as that of the corresponding NIs (compare K_F of "Cb", "Cr" of the column of "NI" with that of the columns of "Ma-C", "Mb-C" and "Mc-C"). This implies that the difference between NIs and JPEG compressed CIs becomes very small in terms of JPEG compression trace, and that this trace seems quite obvious before compression which may impact the colorized image detection, *e.g.*, as in the experimental setting of [6, 15] where CIs are not compressed.

2.3 Network Settings

The first layer of WISERNet used in [15] is a channel-wise convolutional layer where the convolutional kernels are fixed as the thirty 5×5 SRM residual filters borrowed from [5]. In order to study the sensitivity and impact of the first layer of WISERNet on the different datasets, *i.e.*, CIs with/without JPEG compression, we adopt another setting in which WISERNet's first layer is initialized in a conventional way with Gaussian random distribution and is trainable (denoted by WISERNet-Gauss). We expect that under these two settings, the network focuses on different information in NIs and CIs to carry out the classification. In the next section, we present the experimental results related to the two datasets and the two settings of the CNN's first layer, as well as our proposed simple combination method to improve generalization.

3 EXPERIMENTAL RESULTS

3.1 Implementation Details

All the experiments are implemented with PyTorch 0.3.1. The GPU version is GeForce® GTX 1080Ti of NVIDIA® corporation. We train the network according to the experimental setting described in [15].

Table 2: The performance (HTER, in %, lower is better) of WISERNet [15] and WISERNet-Gauss trained on dataset without JPEG compression. For each row, "X" (*e.g.*, "WISERNet") means testing on dataset without JPEG compression and "X-cro" (*e.g.*, "WISERNet-cro") means cross-testing on dataset with JPEG compression. The generalization performance results are presented in italics (same in Table 3).

Method	Ma			Mb			Mc		
	Ma	*Mb*	*Mc*	*Ma*	Mb	*Mc*	*Ma*	*Mb*	Mc
WISERNet	0.34	*4.49*	*3.67*	*3.27*	0.23	*0.30*	*3.15*	*0.98*	0.21
WISERNet-cro	20.56	*40.06*	*40.65*	*36.99*	22.97	*32.47*	*34.82*	*28.37*	26.31
WISERNet-Gauss	0.58	*20.90*	*25.93*	*24.98*	0.43	*2.03*	*23.39*	*1.41*	0.46
WISERNet-Gauss-cro	0.89	*27.96*	*31.70*	*28.48*	10.08	*27.95*	*22.18*	*14.37*	10.53
WISERNet-Ensemble	0.60	*3.86*	*3.61*	*2.82*	0.38	*0.43*	*2.68*	*0.82*	0.38
WISERNet-Ensemble-cro	0.96	*26.44*	*29.40*	*25.79*	8.99	*23.92*	*19.68*	*13.58*	9.66

Table 3: HTER (in %, lower is better) of different networks trained on dataset with JPEG compression. For each row, "X" (*e.g.*, "WISERNet") means testing on dataset with JPEG compression and "X-cro" (*e.g.*, "WISERNet-cro") means cross-testing on dataset without JPEG compression.

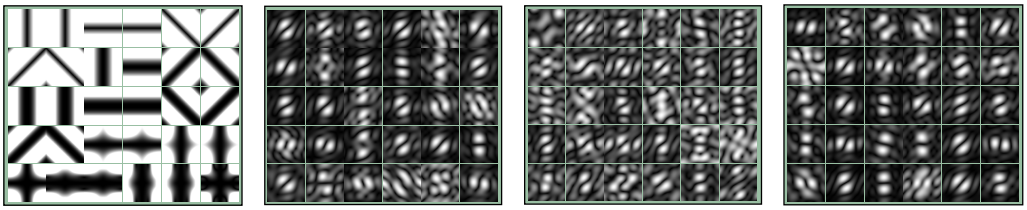
Method	Ma			Mb			Mc		
	Ma	*Mb*	*Mc*	*Ma*	Mb	*Mc*	*Ma*	*Mb*	Mc
WISERNet	0.78	*18.90*	*24.28*	*9.35*	0.93	*2.98*	*4.78*	*2.79*	0.89
WISERNet-cro	0.69	*13.78*	*17.40*	*10.31*	0.67	*1.08*	*4.40*	*1.29*	0.66
WISERNet-Gauss	0.76	*25.52*	*27.97*	*9.03*	0.89	*5.26*	*4.05*	*3.53*	0.96
WISERNet-Gauss-cro	0.72	*22.97*	*29.27*	*12.53*	0.74	*3.86*	*5.49*	*2.44*	0.84
WISERNet-Ensemble	0.82	*16.00*	*20.43*	*6.08*	0.93	*2.10*	*2.44*	*2.14*	1.00
WISERNet-Ensemble-cro	0.80	*11.72*	*15.81*	*6.72*	0.85	*1.10*	*2.30*	*1.40*	0.95

Following [6] and [15], the *half total error rate* (HTER) is employed to evaluate the detection performance. The HTER is defined as the average of misclassification rates (in %) of NIs and CIs. In this work, all reported results are the average of 5 runs.

3.2 Results and Analysis

In this subsection, we first study the impact of CIs with/without JPEG compression and setting of WISERNet's first layer on the classification accuracy (*i.e.*, trained and tested on CIs of same colorization algorithm) and generalization (*i.e.*, trained and tested on CIs of different colorization algorithms). Then, we propose a simple yet effective method to improve the generalization of WISERNet. Table 2 reports the performance of WISERNet and WISERNet-Gauss trained on dataset without JPEG compression and tested on dataset without/with compression. On the contrary, Table 3 reports the performance of these two networks trained on dataset with JPEG compression and tested on dataset with/without compression. We also tested a variant of WISERNet with trainable first layer initialized with SRM filters and found that the performance is similar to original WISERNet, so here we do not show results of this variant.

Compared with the row of "WISERNet" in Table 2, we can find that the classification and generalization error rate in Table 3 both increases. In other words, when only replacing CIs with corresponding compressed version, the forensic performance (especially generalization) obviously drops. Meanwhile, the detection performance significantly decreases when we train WISERNet on dataset without JPEG compression and test it on dataset with compression



(a) (b) (c) (d)

Figure 2: Visualization of FFT of the first-layer filters of WISERNet [(a)] and WISERNet-Gauss [(b), (c), and (d)]. From left to right: SRM, filters in R, G, and B channel, respectively. Note that, filters for R, G, and B in WISERNet are all SRM.

(compare the rows of "WISERNet" and "WISERNet-cro" in Table 2), whereas this phenomenon does not exist in the case of training on dataset with JPEG compression and testing on dataset without compression (compare the rows of "WISERNet" and "WISERNet-cro" in Table 3). These results indicate that WISERNet takes the trace of JPEG compression as the important discriminative feature and thus has good generalization when trained and tested both on dataset without JPEG compression of CIs (the row of "WISERNet" in Table 2). In addition, in Table 2, when the first layer of WISERNet is initialized with Gaussian random distribution and trainable (the so-called WISERNet-Gauss), the generalization is not as good as the original WISERNet (compare the rows of "WISERNet" and "WISERNet-Gauss"). On the contrary, the rows of "WISERNet" and "WISERNet-Gauss" in Table 3 are relatively close, where the networks are trained and tested on datasets with JPEG compression of CIs. This further implies that the SRM filters can strongly capture the trace of JPEG compression. To summarize, when the CIs are without JPEG compression, the WISERNet can achieve very good detection performance, especially generalization, because this model uses SRM filters in the beginning of network which coincidentally and *mistakenly* detects the trace of JPEG compression. This is however not desirable and leads to the dramatic performance drop in the row of "WISERNet-cro" in Table 2. In the meanwhile, when the training database is carefully prepared as in Table 3, the original WISERNet is indeed a good choice for this forensic task, providing better overall performance than WISERNet-Gauss.

Furthermore, we notice that the detection performance differs for WISERNet and WISERNet-Gauss, *e.g.*, the rows of "WISERNet" and "WISERNet-Gauss"in Table 3. The only difference between these two networks is in the first layer. Intuitively, this difference may lead two networks to extract different discriminative features to some extent. We qualitatively analyze this difference, and visualize the FFT (fast Fourier transform) of the first-layer kernels of these two networks after training, and the corresponding results are shown in Fig. 2. Many kernels in the first layer of WISERNet [Fig. 2(a)] have an apparent high-pass response, whereas the kernels of WISERNet-Gauss [Fig. 2(b), (c) and (d)] mainly capture the band-pass frequency information. Based on these observations, we introduce a simple yet effective method to further improve the generalization somehow borrowing idea from ensemble learning. Specifically, we combine the predictions of these two networks to obtain the final prediction according to *a simple criterion*: the final prediction is CI when the prediction of either of two networks is CI, otherwise the image is NI. The rationale behind this criterion is that we trust the (different) discriminative features of both networks which are used to determine whether an image is CI. The corresponding results are reported in rows of "WISERNet-Ensemble" and "WISERNet-Ensemble-cro" in Table 2 and 3. Obviously, the generalization can be improved by this method while decreasing very slightly the classification accuracy (*e.g.*, compare the rows of "WISERNet", "WISERNet-Gauss", and "WISERNet-Ensemble" in Table 3). Despite of its simplicity, to the best of our knowledge, this ensemble strategy with different initialization methods (SRM and Gaussian) is first used in the literature for improving the generalization of CNN-based image forensic detector. The success is probably due to the high diversity of the two initializations, more diverse than the conventional way of using only one initialization method (*e.g.*, Gaussian) with different random sampling.

3.3 Discussion and Summary

Unlike traditional handcrafted-feature-based forensic methods, recent CNN-based approaches are relatively difficult to understand concerning what is the discriminative information used by CNN, and sometimes this information can be surprising and misleading. Taken the above CNN-based colorized image detection as an example, when within the dataset there is an apparent difference in JPEG compression, the CNN with SRM filters captures to some extent this trace and takes it as part of the discriminative information. Consequently, the high performance of CNN model benefits from and covers up the potential pitfall existing in the dataset. As far as we know, there is no existing work that considers and studies this kind of phenomenon for CNN-based image forensics. From this case study, we get some useful hints: 1) reducing as much as possible the impact of image generation and processing history (this information is not relevant to the task at hand), so we need to carefully prepare the data; 2) carefully using some existing filters (*e.g.*, SRM filters) in the beginning of CNN because these filters have strong capacity of capturing image processing history and thus are risky to be used if the dataset has not been properly prepared.

4 CONCLUDING REMARKS

This paper considered the CNN-based colorized image detection as an example and studied the impact of image generation pipeline and CNN's first layer on the classification accuracy and generalization. From this case study, we learned some lessons related to the trustworthiness of CNN-based forensics, which until now have

been ignored among the community but would be helpful for researchers in the field to avoid biased data preparation and network design. We think that our first study in this direction can be useful for other forensic tasks, *e.g.*, detection of computer graphics images and GAN-generated fake images where similar problems and pitfalls may exist. We would like to continue our work on improving the generalization of deep-learning-based detectors, either based on an ensemble of classifiers or other approaches.

ACKNOWLEDGMENT

This work is funded in part by Beijing Natural Science Foundation (L182059), National Natural Science Foundation of China (61620106003 and 61772523), and the French National Research Agency (DEFALS ANR-16-DEFA-0003, ANR-15-IDEX-02). W. Quan acknowledges the support from the UCAS Joint PhD Training Program. We thank L. Zhuo for helpful discussions about the details of [15] and Dr. W. Fan for sharing implementation of the JPEG blocking measure [4].

REFERENCES

[1] L. Bondi, L. Baroffio, D. Güera, P. Bestagini, E. J. Delp, and S. Tubaro. 2017. First steps toward camera model identification with convolutional neural networks. *IEEE Signal Processing Letters* 24, 3 (2017), 259–263.

[2] R. Cogranne. 2018. Determining JPEG image standard quality factor from the quantization tables. *CoRR* abs/1802.00992 (2018), 1–6.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, FL, USA, 248–255.

[4] Z. Fan and R. L. de Queiroz. 2003. Identification of bitmap compression history: JPEG detection and quantizer estimation. *IEEE Transactions on Image Processing* 12, 2 (2003), 230–235.

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

[6] Y. Guo, X. Cao, W. Zhang, and R. Wang. 2018. Fake colorized image detection. *IEEE Transactions on Information Forensics and Security* 13, 8 (2018), 1932–1944.

[7] S. Iizuka, E. Simo-Serra, and H. Ishikawa. 2016. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics* 35, 4 (2016), 1–11.

[8] G. Larsson, M. Maire, and G. Shakhnarovich. 2016. Learning representations for automatic colorization. In *Proceedings of the European Conference on Computer Vision*. Springer, Amsterdam, The Netherlands, 577–593.

[9] W. Quan, K. Wang, D.-M. Yan, and X. Zhang. 2018. Distinguishing between natural and computer-generated images using convolutional neural networks. *IEEE Transactions on Information Forensics and Security* 13, 11 (2018), 2772–2787.

[10] P. Sallee. 2003. Matlab JPEG Toolbox. https://github.com/MKLab-ITI/image-forensics/tree/master/matlab_toolbox/Util/jpegtbx_1.4.

[11] Y. Yan, W. Ren, and X. Cao. 2019. Recolored image detection via a deep discriminative model. *IEEE Transactions on Information Forensics and Security* 14, 1 (2019), 5–17.

[12] J. Zeng, S. Tan, G. Liu, B. Li, and J. Huang. 2019. WISERNet: Wider separate-then-reunion network for steganalysis of color images. *IEEE Transactions on Information Forensics and Security* 14, 10 (2019), 2735–2748.

[13] R. Zhang, P. Isola, and A. A. Efros. 2016. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision*. Springer, Amsterdam, The Netherlands, 649–666.

[14] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, USA, 1053–1061.

[15] L. Zhuo, S. Tan, J. Zeng, and B. Li. 2018. Fake colorized image detection with channel-wise convolution based deep-learning framework. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, Honolulu, HI, USA, 733–736.