Image Forgery Detection Based on Semantic Image Understanding

Kui Ye[1,2], Jing Dong[1,3**], Wei Wang[1,2], Jindong Xu[1], and Tieniu Tan[1]

[1]Center for Research on Intelligent Perception and Computing,
Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
[2]State Key Laboratory of Cryptology, P.O. Box 5159, Beijing, 100878, China
[3]State Key Laboratory of Information Security, Institute of Information Engineering,
Chinese Academy of Sciences, Beijing, 100093, China
kui.ye@cripac.ia.ac.cn
{jdong,wwang,jindong.xu,tnt}@nlpr.ia.ac.cn

Abstract. Image forensics has been focusing on low-level visual features, paying little attention to high-level semantic information of the image. In this work, we propose the framework for image forgery detection based on high-level semantics with three components of image understanding module, the normal rule bank(NR) holding semantic rules that comply with our common sense, and the abnormal rule bank(AR) holding semantic rules that don't. Ke et al.[1] also proposed a similar framework, but ours has following advantages. Firstly, image understanding module is integrated by a dense image caption model, with no need for human intervention and more hierarchical features. secondly, our proposed framework can generate thousands of semantic rules automatically for NR. Thirdly, besides NR, we also propose to construct AR. In this way, not only can we frame image forgery detection as anomaly detection with NR, but also as recognition problem with AR. The experimental results demonstrate our framework is effective and performs better.

Keywords: image forensics, image understanding module, NR, AR, deep learning

1 Introduction

Manipulating and editing digital images are becoming increasingly easy with unimpeded access to advanced image processing softwares, such as Photoshop and Gimp. The ubiquity of forged images erodes our trust in photography and furthermore affects national security, commerce, the media and so on, which subsequently gives rise to the emergence of the field of image forensics that aims to help restore some trust in photography.

[**] Corresponding author. This work is funded by the National Natural Science Foundation of China (Grant No.61502496, No. 61303262, No. U1536120 and U1636201) and the Beijing Natural Science Foundation (Grant No.4164102).

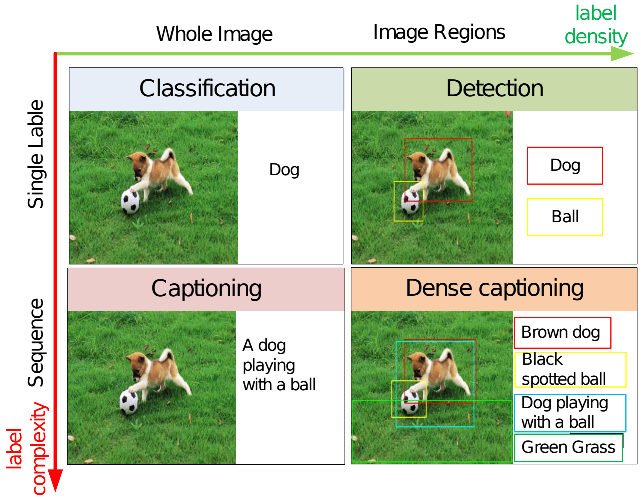

Fig. 1. We implement the image understanding module with a dense image caption model(bottom right).

Traditionally the problem of detecting image forgery is addressed in several aspects. On the pixel level, analyses detect pixel correlations that are introduced by cloning[2], re-sizing[3], or non-linear filtering [4]. Analyses w.r.t the camera sensor can detect inconsistencies in chromatic aberrations[5], color filter array interpolation[6], or sensor noise[7]. On the scene level, analysis can detect inconsistencies in reflections[8], lighting [9], or shadows [10]. The above detection approaches are mainly based on general low-level visual features such as texture, light, color, shape, etc. Analysing the high-level semantic content of the image can also have a crucial role in image forgery detection[1]. Because every authentic picture is a organic whole, involving a certain degree of significance, and forgery images forged from different pictures express a fictional content, it's feasible to detect image forgery by analysing the semantic content w.r.t the image such as clothing, movements, season climate, physical environment and so on.

Although image semantics have been widely used for object detection, object tracking, image retrieval and so on, there is little research on applying high-level semantic information to the field of image forensics. Inspired by Johnson et al.[11], we propose a framework of image forgery detection based on high-level semantics, which consists of image understanding module, NR, and AR. Firstly, semantic instances and their relationship are extracted from the real and forgery image dataset, mapped to captions or words, and stored as rules of NR and AR respectively, both to be used in the next step. Secondly, a test image is processed by the image understanding module. Last, image forgery detection is completed based on the result of image understanding module, NR, and AR.

However, we are not the first to do image forgery detection based on semantics. Ke et al.[1] also proposed a similar framework consisting of three components including image recognition, semantic logic reasoning engine and generation of semantic rules. Compared with Ke et al.[1], our proposed framework has following advantages. Firstly, image recognition in Ke et al. proposed framework needs to segment foreground and background manually and extract hand-crafted features from them before recognition, whereas image understanding module in our proposed framework is achieved with a dense image caption model that integrates object detection, image captioning and soft spatial attention into one single neural network using deep learning, in which the need for human intervention is eliminated and features extracted are more hierarchical and sophisticated. secondly, instead of generating semantic rules manually when constructing common sense knowledge base in Ke et al. proposed framework, our proposed framework is capable of generating thousands of semantic rules automatically for NR. Thirdly, besides NR, similar to the common sense base in Ke et al.[1], we also propose to construct AR. As such, not only can we frame image forgery detection as anomaly detection with NR, but also as recognition problem with AR. Fourthly, our proposed framework is pretty straightforward and simple due to the integration of image understanding module.

Our paper is arranged as follows. Section 2 introduces related work. Section 3 introduces our proposed framework and how it works. Section 4 talks about experiment results, and demonstrate the effectiveness of our framework. In section 5, conclusions are drawn.

2 Related Work

Ke et al.[1] put forward a semantic based framework consisting of three components including image recognition, semantic logic reasoning engine and generation of semantic rules. They complete image forgery detection by first building a common sense knowledge base, then extracting semantic information from the testing image, and finally reaching a conclusion based on the semantic logic reasoning result. However, the framework sadly needs frequent human intervention e.g. in the process of segmentation and of constructing the knowledge base.

Li et al.[15] propose a keypoint-based forgery detection method, which first segments the test image into semantically independent patches. And for each patch, instead of extracting visual or texture features, it only extracts locations of keypoints to estimate the affine transform matrix in order to find the suspicious copy-move regions. Pun et al.[15] push a step forward by proposing a scheme that integrates both block-based and keypoint-based forgery detection methods, but the features used are largely hand-crafted.

Recently, deep convolutional [12] and recurrent networks [13] for images have yielded promising results in image understanding and caption. Johnson et al.[11] introduce the dense captioning model, which takes an image as input and outputs dozens of captions, and which is able to recognize the semantic instances along with their relationships in an image and map such instances and relationships into captions or words. Standing on the shoulder of giants, we adopt the dense

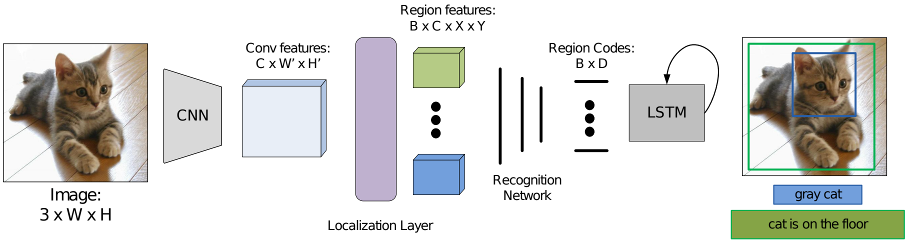

Fig. 2. The structure of the image understanding module.

captioning model that integrates object detection, image captioning and soft spatial attention into one single neural network for image understanding.

3 Framework

The framework of image forgery detection based on semantic image understanding consists of three components: image understanding module, NR, and AR. Firstly, while training the image understanding module i.e. the dense image caption model, NR is meanwhile constructed from real image dataset, and AR from forgery image dataset. Secondly, a test image is processed by the image understanding module to map its semantic contents into captions or words. Last, image forgery detection is completed based on the result of image understanding, NR and AR.

3.1 image understanding module

The image understanding module is achieved with a dense image caption model, with the goal to jointly localize regions of interest and then describe each with natural language. Standing on the shoulder of giants, we adopt the dense captioning model proposed by Johnson et al.[11] that integrates object detection, image captioning and soft spatial attention into a single neural network. As shown in Fig.2, the model takes an image as input and outputs dozens of captions. Firstly, the appearance of the image is encoded by a convolutional network at a set of uniformly sampled image locations. Secondly, a fully convolutional localization layer is used to identify spatial regions of interest and smoothly extracts a fixed-sized representation from each region, the output of which is later processed by a fully-connected neural network before fed to RNN language model implemented with LSTM units. Last, the model produces dozens of captions, each related to a corresponding region of interest. During training the ground truth consists of positive boxes(regions of interest) and descriptions, and binary logistic losses, L1 loss, and cross-entropy loss are used for training. For more detail, please refer to Johnson et al.[11].

As shown in Fig. 1, instead of using image caption model e.g. [14] that takes an image as input and output just one caption, the **dense** image caption model

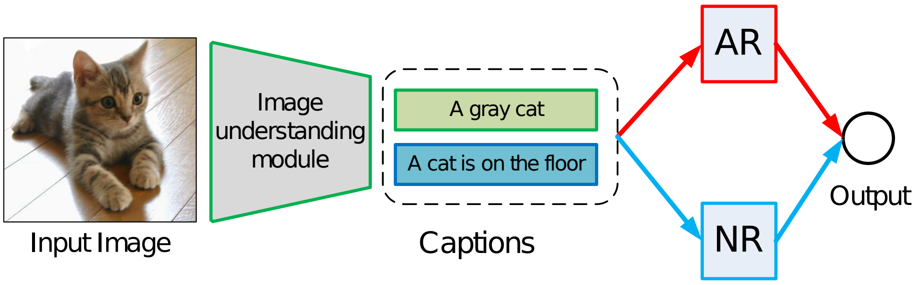

Fig. 3. Forgery detection pipeline. The standardization process is left out from the diagram.

in Johnson et al.[11] is capable of mapping an image to dozens of captions, which can not only produce rich snippet descriptions, but also align the captions and corresponding bounding boxes in the image, leading to better understanding of the image than image caption model.

3.2 NR and AR

When trained, the image understanding module i.e. the dense image caption model takes positive boxes(regions of interest in an image) as input and descriptions as supervision. The dataset used to construct NR is Visual Genome(VG) region caption dataset[15] that contains over 90k images and over 4 million snippets of text, each corresponding to a region of an image. The dataset is extremely large, but all it contains are genuine images and normal captions that comply with our common sense; so normal captions are utilized as the semantic rules of NR. Unlike manually adding rules to common sense knowledge base as in Ke et al.[1], the construction of NR is completed automatically.

Next, AR is constructed when further train or fine-tune the dense image caption model with a self-designed dataset that includes forgery images with captions as well, and data in which are organized in the form like what's in the VG dataset. The self-designed dataset contains 113 photoshopped images with bounding box coordinates and their corresponding captions, the latter of which contrast with our common sense and are utilized as semantic rules of AR. As later experiments would demonstrate, AR also contributes to the detection of image forgeries.

3.3 Forgery detection

Forgery detection. After the image understanding module i.e. dense image caption model is trained and fine-tuned, and thus both NR and AR constructed, the process of the forgery detection is ready to begin. As Fig. 3 shows, for a test image, it is first processed by the dense image caption model to produce dozens of captions, of which the top k (e.g. k=3) captions with highest confidence after removing duplicated captions are selected and then go through the standardization process, a process that strips all the adjectives and articles in a caption, and restores all the remaining words to their base form using senna tool[1] and natural language tool kit(NLTK). For example, after standardization, “a dog is on the green grass” becomes “dog be on grass”, and “a person riding a bear” becomes “person ride bear”. The reason for standardization is that it facilitates the later search process while reserving the vital semantic information. The k captions after standardization later search for matches in NR and AR. Finally, we frame image forgery detection as anomaly detection with NR, and as recognition problem with AR. More specifically, considering the rules hold by NR and AR, if **at least one** of k standardized captions don't match with semantic rules in NR or if **at least one** of k standardized captions match with semantic rules in AR, the image is considered a forgery.

4 Experiments

4.1 Implementation

Semantic rules before added to NR when the dense image caption model is being trained, also go through standardization process. Captions longer than 10 words in VG dataset before standardization are ignored as are in Johnson et al.[11], and so are semantic rules (after standardization) shorter than three words, because they barely contain useful information. Of over four million snippets of captions in the dataset, about 300k rules are extracted and standardized to constitute NR, which is much larger than that of the common sense knowledge base manually built in Ke et al.[1]. When NR is small, the time spent on a test image mainly comes from the image understanding module i.e. the dense image caption model. When NR grows large, the time complexity is approximately O(n), where n is the number of rules in NR.

Table 1. Experiments on self-designed dataset.. “True Positive” stands for correctly identifying a forgery, “False Alarm” misclassifying a forgery as a authentic image, “densecap” dense image caption model, “ft” fine-tuning, and “-k” the value of k.

Method	True Positive	False Alarm
densecap-5	0.66	0
densecap+tf-5	0.76	0

In order to build AR, we fine-tune the model with a extra self-designed dataset that includes forgery images with captions, and data in which are organized in the form like what's in the VG dataset. The self-designed dataset

[1] Available at http://ml.nec-labs.com/senna/

Table 2. Experiments on public dataset.

Method	detection accuracy
Ke et al.[1]	70.51%
densecap-5	65.8%
densecap+tf-1	47.0%
densecap+tf-5	77.4%
densecap+tf-10	79.6%

contains 113 photoshopped images with information of bounding box coordinates and their corresponding captions, and we only draw one box in the image, and correspondingly label one caption with the length no longer than 10 words(since Johnson et al.[11] discard all annotations with more than 10 words), and replicate them five times before used for fine-tuning. Example captions from our self-designed dataset include “a dog is in the air”, “a person riding a bear”, etc. When the dense image caption model is fine-tuned, AR is also constructed, containing 113 semantic rules.

The rule for an image to be considered as a forgery as mentioned in 3.3 is that **at least one** of k standardized captions for the image don’t match with semantic rules in NR or **at least one** of k standardized captions match with semantic rules in AR.

4.2 Experiments on self-designed dataset

Contributions of AR. We examine our framework on a test set, which contains 50 forged images and 50 genuine images. The results are shown in table 1. It’s worth noting that AR is constructed when the dense image caption model is being fine-tuned. When only NR is only constructed before fine-tuning, 66% of forged images are detected correctly, leaving the remaining 34% undetected. And none of the genuine images is detected as a forgery, indicating no false alarm. After AR is constructed i.e. after fine-tuning, true positive rate reaches 76%, improving by 10%. And there is no false alarm as well. Therefore it testifies the previous claim that AR contributes to the image forgery detection.

Importance of fine-tuning. To dig a little deeper, we show in Fig. 4 how the dense image caption model performs on several forged images from the test set. The dense image caption model before fine-tuned is already powerful enough to recognise the semantic instances in an image and their relative relationships. For example, as shown at the top of the Fig. 4, the model is able to recognise the semantic instances like elephant, sky, and grass, and capable of mapping their relative relationships into captions, albeit with a ignorable defect that the model mistakes one elephant as many, which is probably due to the multiple legs bounded by the orange box.

The dense image caption model before fine-tuned is, on one hand, able to enhance the recognition of rarely seen semantic instances like “desert” which the model before fine-tuned struggles to recognize. On the other hand, it also

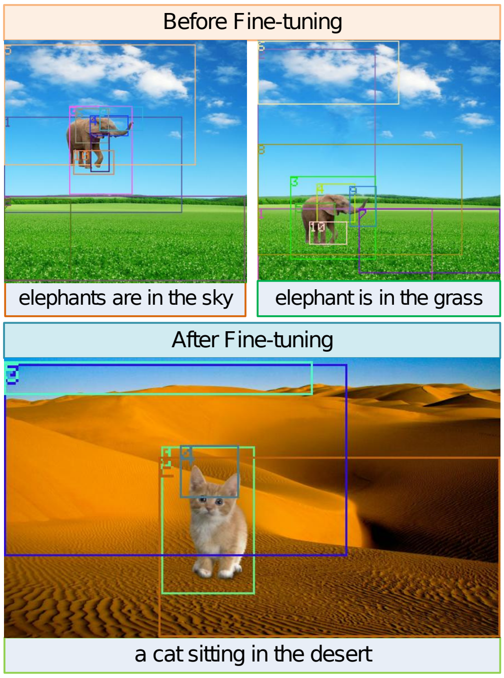

Fig. 4. At the top, displayed in each image are the top ten boxes in color with highest confidence, each of which corresponds to a caption. For brevity, only the key caption are attached to the bottom of each image. At the bottom, only top five boxes are drawn in the image and the key caption attached to the bottom of the image, both for clarity

adapts to the self-designed dataset and produces specific captions for forgery detection more easily. For the image at the bottom of Fig. 4, the model before fine-tuned just recognizes the desert as dirt, and fails to understand its sitting action, whereas the model after fine-tuned recognizes the desert and action quite well.

4.3 Experiments on public dataset

For better comparison, we also test 500 images obtained from Berkeley image dataset [16] and Columbia Image Splicing Detection Evaluation Dataset [2] as did in Ke et al.[1]. As table 2 shows, without fine-tuning, our framework performs a little worse than that in Ke et al.[1](k=5). But it's worth noting that the detection accuracy Ke et al.[1] achieved is on the basis of the correct identification

[2] Credits for the use of the Columbia Image Splicing Detection Evaluation Dataset are given to the DVMM Laboratory of Columbia University, CalPhotos Digital Library and the photographers listed in http://www.ee.columbia.edu/ln/dvmm/downloads/AuthSplicedDataSet/photographers.htm.

of objects in the image, whereas our framework doesn't need the basis, and directly takes an image as input. After fine-tuning and at the same time AR is constructed, performance varies as value k increases. The following analyses are based on the rule for an image to be considered as a forgery as mentioned in 4.1. When k=1 i.e. only the standardized caption with top confidence is considered, the low detection accuracy is because the framework tends to first extract semantic contents complying with our common sense, thus only NR is at work. When k equals to 10, the detection accuracy of our framework reaches 79.6%, the best detection accuracy. But there are also many false alarms. Because the more captions considered, the more likely there are output captions not existed in NR(anomaly), and the more likely there are output captions existed in AR. When k=5, our proposed framework reaches a appropriate balance of utilizing both NR and AR, and in our experiments achieves good detection accuracy without many false alarms.

5 Conclusion

In this work, we propose the framework of image forgery detection based on high-level semantics which consists of three components including image understanding module, NR, and AR. A test image is first processed by the image understanding module to produce captions, which after standardization search for matches with semantic rules in NR and AR. And the search result constitutes the basis for reaching a conclusion about whether it is a forgery or not. We push the work in Ke et al.[1] forward by a moderate step. Firstly, we eliminate all the human interventions in the process of constructing NR much like the common sense knowledge base in Ke et al.[1], and in the process of image forgery detection. In addition, besides NR that considers semantic rules complying with our common sense, we also propose to build AR that takes into account semantic rules contrasting with our common sense, which, as experiments turns out, also helps to boost detection accuracy. It's a very important research approach to detecting image forgery based on semantics, since it can circumvent forgeries based on low-level features. However, the prototype framework also has limitations, and is still at the research stage that needs improving in the following aspects: a) developing more powerful dense image caption mode for the image understanding module; b) automating the construction of self-designed dataset; c)developing more efficient matching strategies.

6 Acknowledgement

This work is supported by NSFC (No. U1536120, U1636201,61502496), the National Key Research and Development Program of China(No. 2016YFB1001003) and China Postdoctoral Science Foundation funded project (No.2016M601168).

References

1. Yongzhen Ke, Weidong Min, Fan Qin, and Junjun Shang, “Image forgery detection based on semantics”, International Journal of Hybrid Information Technology, vol. 7, no. 1, 2014.
2. A Jessica Fridrich, B David Soukal, and A Jan Luk, “Detection of copy-move forgery in digital images”, in Proceedings of Digital Forensic Research Workshop, 2003.
3. Alin C. Popescu and Hany Farid, “Statistical Tools for Digital Forensics”, pp. 128-147, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.
4. Z. Lin, R. Wang, X. Tang, H-V Shum, “Detecting doctored images using camera response normality and consistency”, Proc. Computer Vision and Pattern Recognition, 2005.
5. M. K. Johnson, H. Farid, “Exposing digital forgeries through chromatic aberration”, Proc. ACM Multimedia and Security Workshop, pp. 48–55, 2006.
6. A. C. Popescu, H. Farid, “Exposing digital forgeries in color filter array interpolated images”, IEEE Trans. Signal Processing, vol. 53, no. 10, pp. 3948–3959, 2005.
7. J. Luk, J. Fridrich, M. Goljan, “Detecting digital image forgeries using sensor pattern noise”, Proc. SPIE Electronic Imaging Security Steganography Watermarking of Multimedia Contents VIII, vol. 6072, pp. 0Y1-0Y11, 2006.
8. J. O’Brien, H. Farid, “Exposing photo manipulation with inconsistent reflections”, ACM Trans. Graph., vol. 31, no. 1, pp. 1-11, 2012.
9. Eric Kee and Hany Farid, “Exposing digital forgeries from 3-d lighting environments”, in 2010 IEEE International Workshop on Information Forensics and Security. IEEE, 2010, pp. 1-6.
10. Eric Kee, James F OBrien, and Hany Farid, “Exposing photo manipulation with inconsistent shadows”, ACM Transactions on Graphics (ToG), vol. 32, no. 3, pp. 28, 2013.
11. Justin Johnson, Andrej Karpathy, and Li Fei-Fei, “Densecap: Fully convolutional localization networks for dense captioning”, arXiv preprint arXiv:1511.07571, 2015.
12. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks”, in Advances in neural information processing systems, 2012, pp. 1097-1105.
13. Zachary C Lipton, John Berkowitz, and Charles Elkan, “A critical review of recurrent neural networks for sequence learning”, arXiv preprint arXiv:1506.00019, 2015.
14. Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, “Long-term recurrent convolutional networks for visual recognition and description”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625-2634.
15. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, LiJia Li, David A Shamma, et al., “Visual genome: Connecting language and vision using crowdsourced dense image annotations”, arXiv preprint arXiv:1602.07332, 2016.
16. D. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics”, in Proc. 8th Intl Conf. Computer Vision, July 2001, vol. 2, pp. 416-423.