Image Steganalysis with Very Deep Convolutional Neural Networks

Naoya Mamada

Tokyo Institute of Technology, Nagatsuta, Yokohama, Japan
mamada.n.aa@d.titech.ac.jp

Abstract. Steganography is a technique that embeds secret messages into commonplace data. In contrast, steganalysis is a technique to identify steganography-applied data and to recover hidden message in that data. Effective steganalysis methods are in demand for it is suspected that steganography is made use of by antisocial groups or persons to hide messages from police or intelligence agencies. In this situation, ImageCLEF 2019 Security is held to showcase image steganalysis methods. In the competition track 2: stego image discovery we used natural image classification deep learning models to tackle the problem and get F1 score 0.660, precision score 0.508 and recall score 0.944 to win the third place.

Keywords: Steganalysis · Deep Learning · Convolutional Neural Network · Image Classification

1 Introduction

Steganography is a technique that embeds secret messages, into commonplace data such as family pictures, pieces of classical music or scenery videos. In contrast, steganalysis is a technique to identify steganography-applied data and to recover hidden messages in that data. Effective steganalysis methods are in demand for it is suspected that steganography is made use of by antisocial groups or persons to hide messages from police or intelligence agencies. In this situation, ImageCLEF 2019 Security [3][4] is held to showcase image steganalysis methods. ImageCLEF 2019 Security has 3 tasks; Task 1: Identify Forged Images, Task 2: Identify Stego Images and Task 3: Retrieve the Message. In Task 1, we are tasked to identify files' original extensions. In Task 2, we are tasked to identify steganography-applied images (stego images). In Task 3, we are tasked to recover hidden messages from stego images. We participated in task 1 and task 2. In task 1, only the first 4 bytes of signature were tampered and we perfectly classified the files using the preserved part of file signatures bytes. As task 1 has the trivial solution, in this working note, we will describe our solution for task 2. Our contribution is that we show that image classification models for natural images are usable for steganalysis.

2 Related Work

Steganalysis methods are divided into statistics based methods and feature based methods. The former detects anomalous statistics such as least significant bit(LSB) [7] to distinguish stego images from normal images. The latter detects anomalous image features such as discrete cosine transformation coefficients patterns [6] or hue patterns. As deep learning models show an impressive ability to recognize patterns in images, many researchers propose deep learning based steganalysis methods recently. However, many of them use relatively shallow networks such as 6 layers [13], 14 layers [14] or 20 layers [12], and very deep networks such as 51 layers [8] or 60 layers [11] are rarely used and Wu et al. [11] reported that when the number of layers is smaller than 50, the detection rate decreases as the number increases but model with 60 layers showed overfitting phenomenon and the accuracy degraded. The important difference between steganalysis deep learning models and natural image classification deep learning models is that the most of the former have predefined and fixed high pass convolution layer in the top of the networks [13][14][12][8][11]. In spite of these works, we show that very deep natural image classification models can be diverted to steganalysis.

3 Materials and Methods

We used ImageNet [1] pretrained SE-ResNeXt-50 [2] and SE-ResNeXt-101 to classify images. We trained each model for 25 epochs with batch size 50. Optimizer was Momentum SGD and the learning rate was initially 0.2 and decayed by cosine annealing [5]. Loss function was softmax cross entropy. We conducted random flipping and random cropping of 96^2 out of the original images as data augmentation. We used Chainer [10] deep learning library in experiments. For reference, we tested stegdetect [6] to identify stego images.

For submission rank 15 and 16, we trained SE-ResNeXt-101 and SE-ResNeXt-50 in a 5-fold cross validation manner. We pick up the smallest validation loss weights and inferred test images. We conducted 10-crop of 96^2 patch as a test time augmentation [9]. For submission rank 10, we trained 5 SE-ResNeXt-101 models in a 5-fold cross validation manner. With different 2 random seed for fold splitting, we conducted the training and finally get 10 different models. Averaging the 10 predictions on each testing image, we got the ensemble prediction score. For submission rank 23, we used stegdetect 'simple' mode. For submission rank 26, we mistakenly submitted the submission rank 16, 0s and 1s oppositely.

We trained SE-ResNeXt-50 from random initialization but it was too unstable and we abandoned it.

4 Results

Table 1 and Figure 1 show the results of our submissions. All the deep learning models outperformed stegdetect, classifier based on jpeg discrete cosine trans-

form coefficients statistics. Model ensemble (submission rank 10) shows consistent improvement over single models. However, the effects of model depth (submission rank 15 and 16) are unclear in the results.

Table 1. Results of our submissions on the leaderboard.

Submission Rank	Run ID	F-Measure	Precision	Recall	Model
10	26830	0.660	0.508	0.944	SE-ResNeXt-101
15	26817	0.613	0.473	0.872	SE-ResNeXt-101
16	26771	0.613	0.479	0.852	SE-ResNeXt-50
23	26787	0.529	0.542	0.516	stegdetect
26	26770	0.243	0.673	0.148	SE-ResNeXt-50

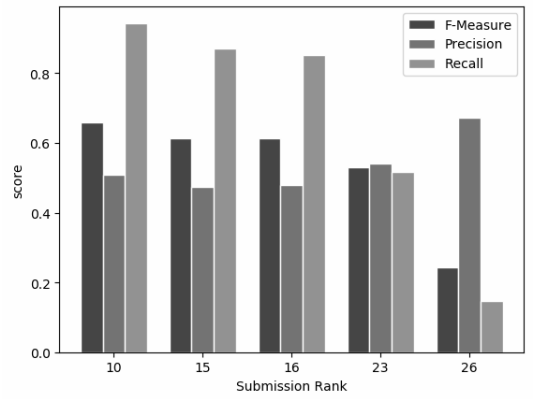

Fig. 1. Results of our submissions on the leaderboard.

5 Discussion and Conclusion

All of the deep learning models, the results show, have high recall and low precision. It means that the models tend to classify normal images as stego images. The reason for this bias is unclear but it possibly because the models are trained to search for structures like block noise. Figure 2 is a positive sample in the training set. There are many black squares those are faintly visible on the white background. We regard that those squares are signs of stego images in this dataset. Such a pattern can be seen in normal jpeg images because of the block noise phenomenon, especially when the images are intensively compressed images. We consider that the models can detect such patterns but cannot classify stego signs from block noise well. The superior performance of the ensemble

model (submission rank 10) possibly because of the improvement of classification performance between the block-noise and stego signs. We point out that performances of 101-layers model (submission rank 16) is slightly better than that of 50-layers model. This is contrary to Wu et al. [11], which found the model start to degrade when the depth is deeper than 50. It appears that ImageNet-pretrainig contributes to stable training and prevents deeper-model degradation.

Fig. 2. An example of positive sample in training data. The original filename is 0825_02.jpg. Best viewed in color.

In this note, we have presented to use natural image classification deep learning models for stego image analysis. We have shown that the deep learning models can outperform a traditional model that is based on cosine discrete transform coefficients statistics. And we have shown that the model ensemble technique can boost the performance. We also have found that with ImageNet pretraining we can use very deep neural networks for steganalysis without degradation. We believe that to test ImageNet pre-trained very deep networks with predefined high-pass filters is a promising next step.

References

1. Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (June 2009). https://doi.org/10.1109/CVPR.2009.5206848
2. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. CoRR **abs/1709.01507** (2017), http://arxiv.org/abs/1709.01507
3. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine,

lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), vol. CLEF 2019 Working Notes. CEUR Workshop Proceedings (CEUR- WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-2380. LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
4. Karampidis, K., Vasillopoulos, N., Cuevas Rodrguez, C., del Blanco, C.R., Kavallieratou, E., Garcia, N.: Overview of the ImageCLEFsecurity 2019 task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (September 09-12 2019)
5. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with restarts. CoRR **abs/1608.03983** (2016), http://arxiv.org/abs/1608.03983
6. Provos, N.: Steganography detection with stegdetect. https://web.archive.org/web/20150415213536/http://www.outguess.org/detection.php
7. Sarkar, T., Sanyal, S.: Steganalysis: Detecting LSB steganographic techniques. CoRR **abs/1405.5119** (2014), http://arxiv.org/abs/1405.5119
8. Sharma, A., Muttoo, S.: Spatial image steganalysis based on resnext. pp. 1213–1216 (10 2018). https://doi.org/10.1109/ICCT.2018.8600132
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
10. Tokui, S., Oono, K.: Chainer : a next-generation open source framework for deep learning (2015)
11. Wu, S., Zhong, S., Liu, Y.: Deep residual learning for image steganalysis. Multimedia Tools Appl. **77**(9), 10437–10453 (May 2018). https://doi.org/10.1007/s11042-017-4440-4, https://doi.org/10.1007/s11042-017-4440-4
12. Xu, G.: Deep convolutional neural network to detect J-UNIWARD. CoRR **abs/1704.08378** (2017), http://arxiv.org/abs/1704.08378
13. Yedroudj, M., Comby, F., Chaumont, M.: Yedrouj-net: An efficient CNN for spatial steganalysis. CoRR **abs/1803.00407** (2018), http://arxiv.org/abs/1803.00407
14. You, W., Zhao, X., Ma, S., Liu, Y.: Restegnet: a residual steganalytic network. Multimedia Tools and Applications pp. 1–15 (04 2019). https://doi.org/10.1007/s11042-019-7601-9