ALASKA-2 : Challenging Academic Research on Steganalysis with Realistic Images

Rémi Cogranne, Quentin Giboulot, Patrick Bas

ALASKA-2: Challenging Academic Research on Steganalysis with Realistic Images

Rémi Cogranne
Troyes University of Technology,
ROSAS dept., LM2S Lab.
Email: remi.cogranne@utt.fr

Quentin Giboulot
Troyes University of Technology,
ROSAS dept., LM2S Lab.
Email: quentin.giboulot@utt.fr

Patrick Bas
CNRS and École Centrale de Lille,
CRIStAL Lab.
Email: Patrick.Bas@centralelille.fr

***Abstract*—This paper briefly summarizes the ALASKA#2 steganalysis challenge which has been organized on the Kaggle machine learning competition platform. We especially focus on the context, the organization (rules, timeline, evaluation and material) as well as on the outcome (number of competitors, submission, findings, and final results). While both steganography and steganalysis were new to most of the competitors, they were able to leverage their skills in Deep Learning in order to design detection methods that perform significantly better than current art in steganalysis. Despite the fact that these solutions come at an important computational cost, they clearly indicate new directions to explore in steganalysis research.**

I. Introduction

Modern steganography aims at hiding secret data into digital media such that it can be transmitted over a public channel without raising suspicion. Using a private key, the secret hidden message is available solely to the intended person. Not only does it ensure the confidentiality of the communication, but it also conceals the fact the communication itslelf.

Steganalysis, the discipline which tackles the converse problem of steganography, namely the detection of hidden information in digital media, has also seen a solid development since its inception. In its most general form, steganalysis looks at revealing any non-public information about a potential steganographic system. However, it nowadays mostly focuses on detecting images containing hidden information among a possibly large set of image.

A. Context

In practice, one can observe that academic research on steganography and steganalysis has evolved in a very specific direction where steganalysis is mostly used to assess the "security" of steganography. To this end, as well as for practical reasons and for reproducibility, academic steganalysis research focused on a standardized setup that does not represent a realistic scenario; in our previous works we have briefly reviewed work published from 2016 up to 2019 and observed that:

- Most of them use the BOSS [1] dataset, made of $10,000$ images captured with 7 different cameras and processed all in the very same way (including a harsh resizing to 512×512 pixels);
- Two third of the work focus on uncompressed images;
- A vast majority of work uses grayscale images;
- Almost all works evaluate steganalysis assuming that both the embedding method and the hidden payload are known to the steganalyst.

This benchmark setup seems overly specific and unrealistic especially in the context of steganalysis where few information about the steganographic system as well as about the image origin are known in practice. The gap between academic words and the "real world" has already been pointed out in 2013 in [2]. In addition, several prior works [3], [4] have shown that the sole modification of the processing pipeline may significantly modify the outcome of such an evaluation of steganography and steganlaysis.

At the same time, the information forensics & security (IFS) research community has been able to make great progress using international challenges. This competitive context has started with Break Our Watermarking System (BOWS) [5] and BOWS2 challenge [6] which stimulated the field of watermarking. The BOSS (Break Our Steganographic System) [1] has been organized when adaptive steganography was proposed and lead todesigning large dedicated features sets and dedicated classification methods. This challenge also proposed a large dataset of grayscale images of size 512×512 that has been adopted as a standard for the community.

More recently, the IEEE SP CUP 2018 has been organized to challenge the community on camera model identification. As opposed to previous challenges, it has been organized on the Kaggle challenge platform with a cash prize of $\$25,000$. This attracted much more competitors and brought attention from scientists and engineers outside the field of IFS community. One major take-away of this competition was the tremendous advantage of deep learning in forensics when compared to classical approaches.

Driven by the success of this challenge, we organized a novel challenge on steganalysis using the same format. Our goal were (i) to challenge the academic research community by confronting them with experts from the Deep Learning community, (ii) to shed light on the challenges inherent to more practical scenario and (iii) to provide a larger dataset to

WIFS'2020, December, 6-9, 2020, New York, USA. XXX-X-XXXX-XXXX-X/XX/$XX.00

complement the BOSS dataset.

II. Set-up of the ALASKAv2 Challenge

Driven by the goal to bring attention onto more realistic scenarios for steganalysis, for which both the exact payload and the embedding schemes and properties of the dataset are unknown, we coined the term "steganalysis into the wild" and consequently called this challenge ALASKA. A similar challenge had already been organized with quite the same goal [7].

This first competition allowed us to clarify the problems encountered in more realistic setups and to avoid design mistakes. As an example, a powerful attack has been found for images compressed with JPEG highest quality factor [8], [9] since this attack would have given an advantage to those aware of it, while not tackling the more general problem of the competition. We consequently carefully avoided such images in the ALASKA#2 challenge.

Another lesson learned from this first challenge was that using a dataset with too much diversity – as we used all possible JPEG quality factors, color images of various sizes, images generated from a wide range of cameras using a randomized processing pipeline – made it difficult to gain clear insights with the proposed solutions.

For the ALASKA#2 challenge, we modified the conversion script from RAW files to JPEG images in order to make them look more realistic. We also increased the number of raw images in the dataset to $80,000$ and made it available to the community. We have also largely reduced the number of JPEG quality factors by using only the set $\{95, 90, 75\}$; this decision was made because top scorers in the ALASKA#1 challenge trained a different model for each different QF. However, we did not want to give an advantage to a team with more computational power while also keeping enough diversity in the compression rates. Last, and perhaps most important, we changed the embedding strategy. First of all, we used three embedding methods: J-UNIWARD [10], UERD [11] and J-MiPOD [12]; while the first two were used in ALASKA#1 the latest was specifically designed for ALASKA#2 based on the well-known MiPOD [13], [14] scheme for spatial domain steganography. Second, we adjusted the payload across all images and all color channels in order to make the difficulty of analyzing each color channel approximately even. To this end, guided by the results from [12], we gathered all color channels when embedding into one color images while adopting a batch embedding strategy [15] referred to as DeLS, where the detectability is equalized between every image. The goal was to prevent someone from knowing this embedding strategy since it was not public. This payload allocation strategy, as well as the payload, was assessed on our own using the DCTR features set [16] with the fast linear classifier [17].

On a more practical note, the challenge started on April 27^{th} and lasted almost 3 months until July 21^{st}. The data that was available to the user was a training subset of $75,000$ images ($25,000$ different images for each QF) randomly selected among the $80,000$ images of the ALASKA#2 dataset; those were available into four different versions, cover and stego with all three different algorithms leading to a total of $300,000$ training images. The testing set, on the other hand, was made of $5,000$ images taken from a independent set. Each of these images could be stego with a probability of $1/5$ among which each algorithm are evenly likely.

We would like to acknowledge the help of the Kaggle administrators, especially Addison Howard and Will Cukierski, who helped a lot on the organization of the challenge; they especially proposed to support the organization of the contest and to round up the $\$7,000$ that we could propose for total cash price to $\$25,000$. However, this did not come at no cost since we had to use as an evaluation criterion a metric that was used by the Kaggle platform. While we initially wanted to focus on reliable detection (with very low false alarm rates such as the false alarm probability for 50% detection accuracy FP_{50} or the probability of missed detection for probability of false alarm 0.05) such option was not available from Kaggle. Besides, we have been strongly advised to use a score that measures the overall accuracy, not only at one specific functional point to avoid distinguishing team by only a few images. Therefore, we decided to move to a weighted area under curve (wAUC). Let us recall that the ROC curve plots the true-positive rate β against the false-alarm rate α_0. The area under the curve is defined as:

$$AUC = \int_0^1 \beta(\alpha_0)\, d\alpha_0. \tag{1}$$

In order to focus on low false-alarm, it was proposed to weight the AUC such that low positive-rate (hence low-false alarm) were given more importance:

$$wAUC = \int_0^1 w(\beta\,(\alpha_0))\beta(\alpha_0)\, d\alpha_0, \tag{2}$$

where the weighting function is such that $\int_0^1 w\,(\beta\,(\alpha_0))\, d\alpha_0 = 1$ to ensure that $0 \leq wAUC \leq 1$. Clearly it would have seemed more meaningful to use a weight function that depends on the false-alarm rate $w\alpha_0$ but this choice was not available and turn out to be quite equivalent. We started with weights too important for low-false alarm rates which lead to all users having a score very close to 1. We therefore quickly changed it (after two days) and set up:

$$w(\beta) \propto \begin{cases} 2 \text{ if } \beta < 0.4 \\ 1 \text{ if } \beta \geq 0.4 \end{cases} \tag{3}$$

III. Outcomes

The lesson learned from the ALASKA#2 competition lies in its success; while the first ALASKA challenge attracted 285 users among which 41 participated actively (the other downloaded the dataset but did not submit any answer), we received over 400 submissions. In contrast, the ALASKA#2 challenge attracted $1,386$ competitors, gathered in $1,115$ teams, who submitted a total of $21,203$ submissions [1].

[1] Note that those numbers slightly differ from the final numbers since Kaggle administrators removed users who presumably cheated.

Similarly, this competition has been very tight ; while Binghamton University team won ALASKA#1 challenge with a large margin, ALASKA#2 challenge ranking has been close up to the very end of the submission deadline.

Kaggle platform certainly has strong advantages, among which the computational resources it shares to its users. This allows all users to access resources to compete seriously. However, it can be noted that most top-scores used their own computing resources since, as acknowledged by Eugene Khvedchenya (who finished 2nd on the public ranking) "*From all challenges I've participated in, this particular challenge was probably the most demanding one in terms of hardware requirement.* "

It is interesting to note that most users were more interested in learning rather than struggling to rank among the first. The Kaggle discussion area was widely used to share ideas, results and attempts. This certainly allows quickly (1) to understand what steganography and steganalysis are about, (2) how JPEG compression works and how data hiding is made in DCT coefficients and (3) what are the most promising Deep Leaning architectures targeted for this task of steganalysis into the wild.

We will briefly describe in what follows the main outcomes without going into much details for each user solutions; one can find more details on ALASKA#2 official challenge webpage on kaggle[2] and especially in the discussion tab[3] and in paper of the dedicated special session of IEEE WIFS 2020.

Dealing with Different JPEG Quality Factors: The first very interesting lesson is that it does not seem that having one specific network for each JPEG quality factor does not seem to bring much detection accuracy. It is important to note that, because we have used only 3 different QF it is not certain that this analysis can be generalized for any quantization table. However, it seems that a complex Deep Learning Network is able to figure out the QF or, at least to get ride of its impact. Indeed, all users who tried learning over each JPEG quality factor individually did not get significant improvements and, hence, eventually decided to analyze all images together regardless the QF in order to benefit from a larger dataset for training.

Such outcome has been very surprising for us as it goes completely against what has always been observed for features-based steganalysis and that was common belief also for deep-learning based steganalysis. In the problem of "holistic vs atomistic" steganalysis (should you learn a specific network over each possible dataset for more tailored detectors or blend all images together for more robustness) the JPEG quality factor has long been recognized as the sole parameter that may prevent generalization.

Dealing with Different Color Channels: Dealing with color channels has not been deeply investigated in steganalysis. Question that we asked during the challenge[4] was whether the RGB color space was more relevant because Deep Learning architectures are usually designed for this case, or whether YCbCRr would be more relevant since data hiding is made on this components. Similarly, whether one should use spatial or DCT domain was also addressed. All in all, it seems that it is more efficient to analyze images in the spatial domain, which was in line with has been long recognized in steganalysis.

However, it seems that diversifying allows slightly improving the performance as we have observed top users using both spatial and DCT domain or several color spaces.

Regarding the color channels itself, it seems that it choice does not matter for steganalysis. Perhaps because they all mostly consists in linear transformation that a complex network can eventually figure out, similar performances have been observed in RGB, YUV, YCbCr, L*a*b ...

Interestingly, it has been observed that using non-quantized values (regardless of the color channel used) slightly improve detection performance.

Most Relevant Deep Learning Architecture: Among the current-art Deep Leaning method, undoubtedly, EfficientNet [18] has been by far the most extensively used in this competition for its relative simplicity and high performance. It is also possible that this is partially due to the fact that a few weeks after the kick-off an implementation with EfficientNet-b2 allows getting a score as high as 0.921 (to be contrasted with the 0.935 of the top scorers at that time). Interestingly, it can be noted that MixNet also achieved high performance while ResNet and DenseNet did not. According to David Austin, a possible explanation could be "*the MBConv block*[5] *associated with the MobileNetV2 block sequence in several architectures that has been rather successful in this competition (EfficientNet, MobileNet, MixNet, MNasNet to cite a few). This may be related to the widening of the network with* 1×1 *convolutions to enhance the channel-wise separation.*"

While such block cannot be found in ResNet and DenseNet (that performed significantly worse for this challenge), one could argue that the user who finished first used mostly SEResNet18 [19] yet removing the stride and pooling from the first two layers which prevent the downsampling of images and, ultimately, to keep much more information about such weak signals as steganography may be as well as to speed up significantly the convergence.

Figure 1 presents overall performance obtained with several Deep Learning architecture (measures using the wAUC (2) used in this competition) as a function of network complexity (measured as the number of parameters).

Steganalysis in DCT domain using deep learning has remained quite a challenging problem in large part because the neighboring samples are very different from each other making convolutional networks quite irrelevant. For this reason it has been quite unused. However, among those who did use DCT

[2]See: https://www.kaggle.com/c/alaska2-image-steganalysis.

[3]See: this a summary that lists all solutions.

[4]We have used the discussions on Kaggle platform to interact with users.

[5]Note from the authors: a MBConv block is an inverted residual block with skipped connections, where more channels are artificially created by using spatial neighbors.

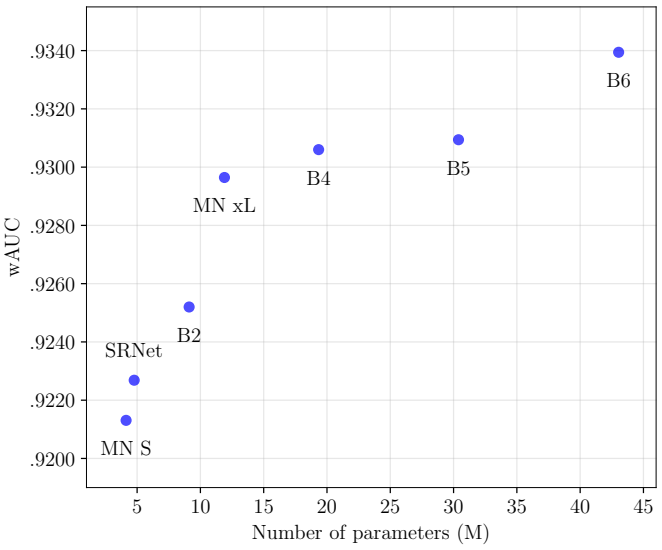

Fig. 1. Comparison of detection accuracy (measured using the weight AUC used for the competition) obtained with various current-art Deep Learning architectures. Image from [20], kindly provided by Yassine Yousfi, from DDE, Binghamton University.

coefficients, undoubtedly the most successful approach has been to use the one hot encoding approach recently proposed in [21].

a) Several Tricks for Improving Detection Accuracy: Interestingly, as we explained above, using Efficient-b2 "as it" with weights pre-trained from ImageNet already allows getting very interesting steganalysis performance significantly higher than current state-of-the-art in steganalysis, namely SRNet [22] (see Figure 1). Most of the users focused on improving over performances that were already quite outstanding. The main approaches to do so has been quite classical and can be summarized as follows:

- Use more complex yet slightly more accurate networks (see Figure 1) ;
- Build an ensemble of classifiers using several different architectures ;
- Diversify using several color channels and adding DCT ;
- Train/validate over several splits and select best models.
- Use data augmentation strategies, such as CutMix strategy [23], to improve the learning efficiency ; this was used by many kagglers and especially those who ended first and third [24].

Eventually, it has been quite a consensual observation that training a multi-class classifier allows slightly improving the detection accuracy as compared to training a binary classifier merging all three different embedding schemes.

IV. Conclusion

We briefly summarized in this paper the ins and outs of organizing ALASKA#2 steganalysis challenge over the Kaggle platform. It has been an overall success with many users, submissions, very interesting and fruitful discussions and quite a lot of new ideas. This competition undoublty brought new standards into the field of steganalysis by improving quite a lot the current state-of-the-art as well as making such methods much easier to applied in practice. The present paper recap briefly the main findings, more details can be found on Kaggle website [6] and in the papers [20], [24] accepted in the associated special sessions.

We believe that this allowed to give a focus on several open problems in steganography and steganalysis such as those for batch and color images steganography and steganalysis, cover-source mismatch for which we also hope that this challenge will stimulate novel research directions.

Acknowledgment

The authors would like to thank all Kagglers that struggled onto the ALASKAv2 battleground. We would like to greatly thank individuals from Kaggle (namely Will Cukierski and Addison Howard) who not only helped us setting up the challenge but supported this challenge by providing a large part of the grand total cash price.

References

[1] P. Bas, T. Filler, and T. Pevný, "Break our steganographic system — the ins and outs of organizing boss," in *Information Hiding, 13th International Workshop*, ser. Lecture Notes in Computer Science. Prague, Czech Republic: LNCS vol.6958, Springer-Verlag, New York, May 18–20, 2011, pp. 59–70.

[2] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevný, "Moving steganography and steganalysis from the laboratory into the real world," in *Proceedings of the first ACM workshop on Information hiding and multimedia security*, ser. IH&MMSec '13. New York, NY, USA: ACM, 2013, pp. 45–58.

[3] V. Sedighi, J. J. Fridrich, and R. Cogranne, "Toss that bossbase, alice!" in *Media Watermarking, Security, and Forensics*, ser. Proc. IS&T, Feb 2016, pp. pp. 1–9.

[4] Q. Giboulot, R. Cogranne, D. Borghys, and P. Bas, "Effects and solutions of cover-source mismatch in image steganalysis," *Signal Processing: Image Communication*, vol. 86, p. 115888, 2020.

[5] A. Piva and M. Barni, "The first bows contest (break our watermarking system)," in *Security, Steganography, and Watermarking of Multimedia Contents IX*, vol. 6505. International Society for Optics and Photonics, 2007, p. 650516.

[6] T. Furon and P. Bas, "Broken arrows," *EURASIP Journal on Information Security*, vol. 2008, pp. 1–13, 2008.

[7] R. Cogranne, Q. Giboulot, and P. Bas, "The alaska steganalysis challenge: A first step towards steganalysis"*into the wild*"," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, ser. IH&#MMSec'19. New York, NY, USA: ACM, 2019, pp. 125–137.

[8] J. Butora and J. Fridrich, "Reverse jpeg compatibility attack *(available as Early Access)*," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2019.

[9] R. Cogranne, "Selection-channel-aware reverse jpeg compatibility for highly reliable steganalysis of jpeg images," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2772–2776.

[10] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal on Information Security*, vol. 2014, no. 1, pp. 1–13, 2014.

[11] L. Guo, J. Ni, W. Su, C. Tang, and Y.-Q. Shi, "Using statistical image model for jpeg steganography: uniform embedding revisited," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2669–2680, 2015.

[6] At the end of the challenge, several proposals were described in details in the discussion, see: this thread that lists all solutions.

[12] R. Cogranne, Q. Giboulot, and P. Bas, "Steganography by minimizing statistical detectability: The cases of jpeg and color images," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, ser. IH&#MMSec'20. New York, NY, USA: ACM, 2020, pp. 125–137.

[13] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 221–234, Feb 2016.

[14] S. Vahid, J. Fridrich, and R. Cogranne, "Content-adaptive pentary steganography using the multivariate generalized gaussian cover model," vol. 9409, February 2015.

[15] R. Cogranne, V. Sedighi, and J. Fridrich, "Practical strategies for content-adaptive batch steganography and pooled steganalysis," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2122–2126.

[16] V. Holub and J. Fridrich, "Low-complexity features for jpeg steganalysis using undecimated dct," *Information Forensics and Security, IEEE Transactions on*, vol. 10, no. 2, pp. 219–228, Feb 2015.

[17] R. Cogranne, V. Sedighi, J. Fridrich, and T. Pevný, "Is ensemble classifier needed for steganalysis in high-dimensional feature spaces?" in *Information Forensics and Security (WIFS), IEEE 7th International Workshop on*, November 2015, pp. 1–6.

[18] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. of Intl' Conference on Machine Learning, ICML 2019*, vol. 97. PMLR, 9-15 June 2019, pp. 6105–6114.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[20] Y. Yousfi, J.Butora E. Khvedchenya and J. Fridrich, "ImageNet Pre-trained CNNs for JPEG Steganalysis,"*(to be published)* in *Information Forensics and Security (WIFS), IEEE 12th International Workshop on*, Decembre 2020, New York City, NY, USA. 2020, New York City, NY, USA.

[21] Y. Yousfi and J. Fridrich, "An intriguing struggle of cnns in jpeg steganalysis and the onehot solution," *IEEE Signal Processing Letters*, vol. 27, pp. 830–834, 2020.

[22] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2018.

[23] S. Yun, D. Han, S. J. Oh , S. Chun, J. Choe and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," *In Proceedings of the IEEE International Conference on Computer Vision* pp. 6023–6032, 2019

[24] K. Chubachi, "An Ensemble Model using CNNs on Different Domains for ALASKA2 Image Steganalysis," *(to be published)* in *Information Forensics and Security (WIFS), IEEE 12th International Workshop on*, Decembre