Recherche et Analyse de Sources Représentatives pour la Stéganalyse

Rony Abecidan, Vincent Itier, Jérémie Boulanger, Patrick Bas

Recherche et Analyse de Sources Représentatives pour la Stéganalyse

Rony ABECIDAN[1], Vincent ITIER[2,3], Jérémie BOULANGER[1], Patrick BAS[1]

[1]Univ. Lille, CNRS, Centrale Lille,UMR 9189 CRIStAL, F-59000 Lille,France

[2]IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille,France

[3]Univ. Lille, CNRS, Centrale Lille, Institut Mines-Télécom, UMR 9189 CRIStAL, F-59000 Lille, France

rony.abecidan@univ-lille.fr, vincent.itier@imt-nord-europe.fr,
jeremie.boulanger@univ-lille.fr, patrick.bas@cnrs.fr

Résumé – Lorsque Alice dissimule un message dans une image de couverture pour communiquer avec Bob, elle ne va pas la piocher dans les bases de la littérature dédiées à l'entraînement de détecteurs en stéganalyse. Cela peut alors poser un problème pour Eve, l'adversaire d'Alice, car la source dont provient l'image d'Alice peut être très distincte de celle sur laquelle Eve s'est entraînée. La source d'une image dépend d'une multitude de paramètres comme la nature du capteur, les retouches logicielles ainsi que les paramètres de compression. Il est alors quasi-impossible pour Eve de s'entraîner sur la même base qu'Alice. Dans ce papier, nous construisons puis analysons un ensemble de bases dont les paramètres favorisent la généralisation de l'apprentissage. Notre principale contribution est la découverte de bonnes pratiques permettant de concevoir des bases représentatives pour la stéganalyse d'images numériques. Elles sont liées à la dimension intrinsèque des caractéristiques représentant la base, mais aussi au type de classifieur utilisé.

Abstract – When Alice is hidding a message into a cover image in order to communicate with Bob, she is not going to pick it among historic databases from the literature used to train detectors in steganalysis. This could be very problematic for the opponent Eve since Alice's image is likely coming from a very different source than her training one. The source of an image can be defined with a multitude of parameters such as the nature of the sensor, the post-processing operations and the parameters of the compression. Hence it is almost impossible for Eve to train her detectors with the base of Alice. In this paper we are building and analyzing a collection of bases whose parameters are fostering generalisation. Our main contribution is the discovery of good practices helping to build such representative bases for digital image steganalysis. They are related to the intrinsic dimension of the features representing the database but also to the type of classifier used.

1 Introduction

La stéganographie d'images numériques est l'art de dissimuler de l'information dans une image de couverture afin de communiquer secrètement sans lever de suspicions. En opposition, le stéganalyste cherche à détecter l'usage de la stéganographie. Cette tâche est complexe dans un contexte opérationnel où les images d'évaluation proviennent de sources inconnues.

Les sources d'images dépendent de nombreux paramètres tel que la nature de l'appareil d'acquisition (appareil photo, téléphone, *etc.*), la qualité du capteur, les paramètres de capture (ISO, angle d'ouverture, etc.), le contenu capturé (intérieur, extérieur, niveau de détails, *etc.*), les étapes de retouches (débruitage, recadrage *etc.*) et les étapes de compressions (conversion 8-bit, compression JPEG *etc.*). Être capable de générer des images provenant d'une certaine source sans connaître ces paramètres est donc difficile.

Dans la littérature, les modèles de stéganalyses sont plutôt entraînés à partir de bases historiques comme BOSS [1] ou ALASKA [2]. Cela conduit à des problèmes de performances en phase d'évaluation puisque les images examinées ne proviennent pas de la source liée aux images d'entraînement. C'est le "Cover Source Mismatch" (a.k.a. CSM) bien connue en stéganalyse opérationnelle [3] [4] [5].

La chaîne de développement post-traitement, allant de la retouche jusqu'à la compression, est un des facteurs les plus responsables du CSM [3].

L'impact de ces opérations sur la distribution du bruit d'une image est non négligeable même si la sémantique reste inchangée. Or, un détecteur entraîné pour la stéganalyse va précisément rechercher la présence d'un message dans la distribution du bruit, ce qui explique pourquoi le CSM risque d'apparaître. Nous avons étudié les facteurs à l'origine du CSM dans [6] et deux conclusions claires se dégagent :

— Le débruitage, le recadrage, le redimensionnement et le filtrage de netteté post-redimensionnement jouent un rôle significatif dans le CSM.

— Certaines sources conduisent à un apprentissage bien plus généraliste que d'autres.

Nous construisons ici un ensemble de bases d'images à partir des opérations citées plus haut pour favoriser le plus de diversité possible. Nous nous concentrons ensuite sur l'analyse des bases les plus représentatives de cet ensemble afin de comprendre comment créer des bases d'entraînements pertinentes pour la steganalyse opérationnelle.

2 Sources Holistiques en Stéganalyse

2.1 Scénario applicatif étudié

Le problème du CSM peut pratiquement s'illustrer comme suit. Alice et ses $N-1$ complices utilisent la stéganographie pour dissimuler des messages secrets dans des images numériques. En vertu du principe de Kerckhoffs, ils appliquent une méthode d'insertion connue avec un taux d'insertion fixe. Pour ce faire, ils disposent chacun d'un stock d'images de couvertures pour communiquer avec leurs contacts. Eve a réussi à intercepter un ensemble d'images pour chaque suspect. Elle doit déterminer lesquelles cachent un message et lesquelles sont authentiques. Ces N ensembles d'images non étiquetées constituent les *cibles* d'Eve. D'autre part, Eve dispose de M bases d'entraînements, appelées *sources*, et souhaite les utiliser pour entraîner un détecteur.

Eve doit choisir comment entraîner son détecteur avec ses M sources disponibles. Toutes les sources ne se valant pas [6], elle ne peut pas faire ce choix arbitrairement. Dans le cadre de cet article, nous imaginons qu'elle cherche à s'entraîner pour une source spécifique. Ce choix peut se justifier par l'importante complexité temporelle de la recherche des meilleures combinaisons de sources. Elle cherche alors la source la plus adaptée parmi les M sources à sa disposition.

2.2 Formalisation

En s'inspirant de [7], nous considérons qu'une chaîne de développement est définie par un vecteur $\omega \in \Omega$ qui contient tous ses paramètres (paramètres de débruitage, facteur de qualité JPEG, etc.). Dans le contexte de la stéganalyse, nous introduisons un paramètre γ représentant les choix de l'adversaire, notamment la stratégie d'insertion et le taux d'insertion. L'état de l'art pour cette tâche repose essentiellement sur des modèles d'apprentissage automatique qui peuvent être considérés comme des prédicteurs :

$$f(x \mid \theta_{\omega,\gamma}) : \mathcal{X} \to \{cover, stego\}$$
$$x \mapsto y$$

où $\theta_{\omega,\gamma} \in \Theta$ contient tous les paramètres appris avec des images de couvertures issues de la distribution engendrée par ω et potentiellement manipulées suivant γ.

Pour évaluer correctement le CSM, deux métriques pertinentes ont été introduites dans [3] et [7] :

— La difficulté intrinsèque d'une source $s \in \mathcal{S}$, c'est-à-dire la probabilité d'erreur P_E^f que nous obtenons après avoir entraîné le détecteur f sur des images de cette source et évalué sur des images de cette même source.

$$\mathbb{E}_{(x,y)\sim P((x,y)|\omega,\gamma)}(f(x \mid \theta_{\omega,\gamma}) \neq y)$$

— Le regret $R_{s,c}$ entre une source $s \in \mathcal{S}$ et une cible $c \in \mathcal{C}$ défini comme la différence entre le P_E^f que nous obtenons en entraînant le détecteur f sur s et en évaluant sur c et la difficulté intrinsèque de c .

$$R_{s,c}^f = [\ \mathbb{E}_{(x,y)\sim P((x,y)|\omega_c,\gamma)}(f(x \mid \theta_{\omega_s,\gamma}) \neq y)$$
$$-\mathbb{E}_{(x,y)\sim P((x,y)|\omega_c,\gamma)}(f(x \mid \theta_{\omega_c,\gamma}) \neq y)\ ] \geq 0$$

L'objectif d'Eve est de rechercher une source s^* permettant de garantir la meilleure généralisation possible sur ses cibles :

$$s^* = \underset{s\in\mathcal{S}}{\operatorname{argmax}} \sum_{c\in C} \mathbb{1}_{R_{s,c}^f < \epsilon}\ (\#)$$

où ϵ représente le niveau maximum de décalage de performances que nous acceptons.

Une telle source est qualifiée de *source représentative* pour souligner qu'elle conduit le détecteur à un apprentissage large qui peut aider la détection sur toutes les cibles. Insistons sur le fait qu'une source est représentative relativement à un détecteur $f(.)$. Une même source peut alors se révéler très intéressante pour l'apprentissage de certains détecteurs et peu intéressante pour d'autres. Par opposition, on peut appeler *source spécifique*, une source qui conduit à un apprentissage très spécifique et peu généraliste.

En pratique, le stéganalyste ne peut pas trouver sa source la plus représentative par rapport au critère $(\#)$ puisqu'il n'a pas accès à des étiquettes d'images provenant de la distribution cible. Il faut donc trouver cette source par un autre moyen.

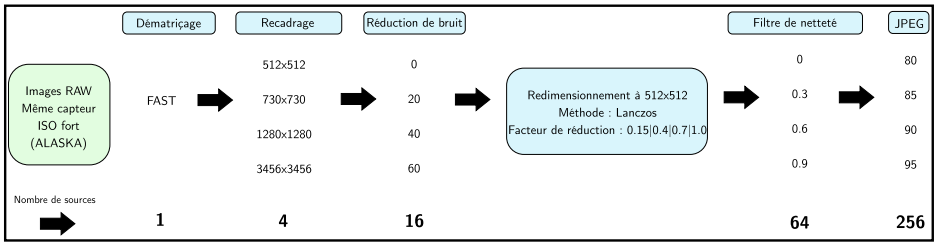
2.3 Impact de la topologie des données

L'extracteur de features DCTr [8] permet d'extraire 8000 caractéristiques d'intérêts pour chaque image. On peut supposer que les caractéristiques des images provenant d'une même source appartiennent à une variété de dimension bien inférieure que 8000 [9]. La dimension intrinsèque de la variété qui supporte les données donne alors une indication sur le nombre minimum de paramètres requis pour les analyser. L'apprentissage du détecteur s'appuie sur la variété supportée par la source. Ainsi, si la variété supportée par la cible est trop différente de celle supportée par la source, cela peut expliquer un regret non négligeable. Cette situation se présente par exemple quand il y a un grand écart de dimension intrinsèque entre la source et la cible.

3 Expérience & Analyse

3.1 Protocole expérimental

Pour les expériences présentées ici, nous travaillons avec 1115 images RAW de taille 5184×3456 issues de ALASKA [2] provenant de l'appareil photo CANON-EOS-100D et capturées avec une sensibilité ISO > 1000. Nous souhaitons ici jouer uniquement sur les opérations post-traitements pour la construction de nos sources représentatives. C'est pourquoi nous évitons d'autres facteurs importants expliquant le CSM tel que la nature et les paramètres de l'appareil de capture. D'autre part, choisir des images RAW capturées avec des ISO forts



Chaînes de développement simulées avec RawTherapee et ImageMagick

FIGURE 1 – Génération de 4^4 sources. Les différents facteurs d'échelle sont causés par les recadrages de tailles différentes.

permet d'accroître le contraste entre les différentes chaînes de développements. Nous simulons des chaînes de développement avec RawTherapee [1], un logiciel libre qui gère un large éventail d'opérations allant du dématriçage à la compression JPEG.

Au départ, les images RAW sont divisées aléatoirement de façon à avoir 50% d'images d'entraînement et 50% d'images de tests. Ensuite, les images de couverture sont générées en suivant la Figure 1 et l'insertion est effectuée à l'aide de la stratégie UERD [10] avec un taux d'insertion de 1.5bpp. Pour simplifier et gagner en temps de calcul, nous entraînons des classifieurs linéaires en utilisant les caractéristiques DCTR [8] de nos sources. Avec ces classifieurs linéaires, ce taux d'insertion élevé garantit de faibles difficultés intrinsèques allant de 0% à 8.7%. Choisir un taux d'insertion plus faible peut conduire à des situations où le détecteur n'apprend rien du tout à partir d'une source, nous rendant aveugle sur la représentativité de cette dernière. Une fois les sources générées et les détecteurs entraînés sur chacune d'entre elles, nous étudions le CSM à l'aide d'une matrice de regret, R où $R[s,c] = R_{s,c}$, (s,c) dans $M \times N = \{0,...,256\}^2$.

3.2 Choix du classifieur

Certains détecteurs linéaires comme LCLSMR [11] sont conçus sur-mesure à des fins de stéganalyse. C'est pourquoi ce classifieur est utilisé dans [3] et [6] pour étudier le CSM. Cependant, bien que performant quand il n'est pas question de changement de source, il conduit souvent à un apprentissage trop spécifique par construction. Nous avons étudié ce point en comparant le classifieur LCLSMR [11] avec la régression logistique. Même si les difficultés intrinsèques de chaque source obtenues avec ces deux classifieurs sont comparables, les regrets quant à eux sont bien différents. Nous affichons en Figure 2 la distribution des regrets observés par classifieur ainsi que la distribution du nombre de cibles représentées par source avec un regret maximum de $\epsilon = 1\%$.

L'apprentissage de la régression logistique est bien plus général que celui de LCLSMR même si ces deux détecteurs sont linéaires. En témoigne les regrets beaucoup plus faibles que nous obtenons avec la régression logistique. D'autre part, en partant des mêmes sources, on observe qu'il est possible de bien plus généraliser sur les cibles avec la régression logistique. Cela démontre la forte dépendance de la représentativité d'une source vis-à-vis du type de détecteur employé.

1. rawtherapee.com

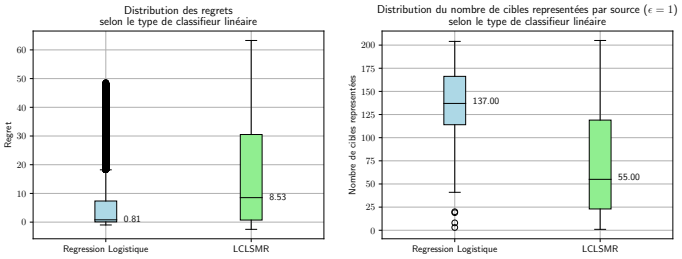


FIGURE 2 – Comparaison de la nature de l'apprentissage du classifieur LCLSMR avec celui de la régression Logistique sur notre univers

3.3 Regret et Dimension intrinsèque

Nous voulons savoir si le regret $R_{s,c}$ entre une source s et une cible c dépend de leur différence de dimensions intrinsèques. La librairie scikit-dimension [12] propose plusieurs estimateurs de dimension intrinsèque issues de la littérature. En utilisant les estimateurs les moins gourmands en temps de calcul, nous calculons dans la Table 1 la corrélation de Pearson entre le regret $R_{s,c}$ et la différence de dimensions intrinsèques dans notre univers de sources $|\Delta DI_{s,c}| = |DI[s] - DI[c]|$. Ces deux quantités semblent corrélées. Une grande différence de dimension intrinsèque peut alors expliquer un grand regret.

Classifieur	*Estimateur de dimension intrinsèque*							
	TwoNN	CorrInt	MADA	MIND_ML	MLE	MOM	TLE	FisherS
LCLSMR	0.23	0.10	0.26	0.28	0.27	0.24	0.27	0.12
Régression Logistique	0.42	0.11	0.36	0.42	0.39	0.26	0.38	0.22

TABLE 1 – Correlation de Pearson entre $|\Delta DI_{s,t}|$ et $R_{s,t}$

La Figure 3 représente l'évolution de la médiane des regrets sur un voisinage de ΔDI estimée avec la méthode MIND_ML, estimateur pour lequel la corrélation présentée dans la Table 1 est la plus forte. En moyenne, les regrets observés dans notre univers sont donc d'autant plus faibles que les différences de dimensions intrinsèques sont faibles. On déduit qu'une source représentative doit avoir une dimension intrinsèque proche de toutes les cibles. Cela peut s'observer dans la Figure 4 qui montre le nombre de cibles représentées par une source ($\epsilon = 1\%$) en fonction de sa dimension intrinsèque. On observe que les sources les plus représentatives se trouvent autour de la médiane des dimensions intrinsèques des cibles. Bien que nécessaire, cette condition n'est pas suffisante pour décider si une source est représentative. Nous avons également observé que la dimension intrinsèque d'une source est d'autant plus forte que sa distribution est bruitée (faible compression, pas

de redimensionnement, etc.). Ce constat est cohérent. Plus une source est bruitée, plus il est aisé de cacher un message dans le bruit, et plus il va falloir s'appuyer sur un grand nombre de caractéristiques d'intérêts pour déceler l'usage de la stéganographie. On peut naïvement penser qu'apprendre sur une source fortement bruitée permettrait alors de généraliser sur une multitude de cibles étant donné que la tâche est plus complexe. Nous infirmons cette hypothèse avec les résultats précédents puisqu'une dimension intrinsèque trop forte risque de conduire à une source spécifique présentant de grands regrets avec les cibles qui sont moins bruitées.

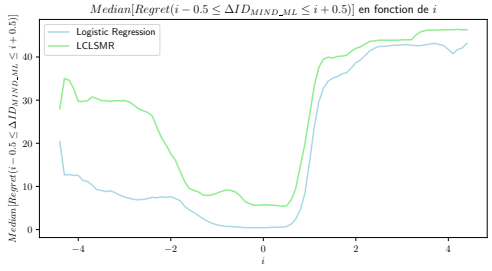


FIGURE 3 – Evolution de la médiane glissante des regrets en fonction de ΔDI

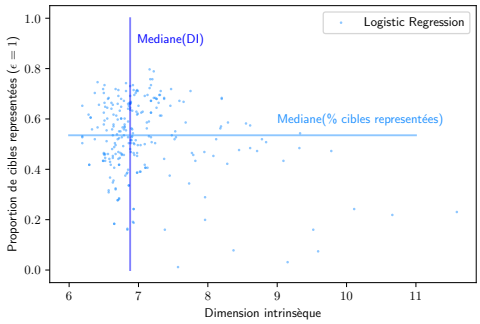


FIGURE 4 – Evolution de la proportion de cibles représentée en fonction de la dimension intrinsèque des sources (régression Logistique)

4 Conclusions & Perspectives

Nous étudions ici les distributions d'images de couvertures représentatives conduisant un détecteur donné vers un apprentissage généraliste pour la stéganalyse opérationnelle. Nous remarquons que la représentativité d'une source dépend fortement de la nature du détecteur utilisé. Le détecteur LCLSMR, référence de la littérature, s'est révélé trop spécifique dans ses apprentissages. Nos résultats suggèrent également qu'une source représentative se doit d'avoir une dimension intrinsèque proche de celles de toutes les cibles d'intérêts. Cette découverte est très intéressante puisque la dimension intrinsèque est une quantité que l'on peut estimer sans connaissances des étiquettes de la cible. On peut donc éliminer les sources qui ont une dimension intrinsèque trop différente de celles de nos cibles. Dans un avenir proche, nous souhaiterions développer notre analyse avec des cibles plus réalistes, un taux d'insertion plus raisonnable et des détecteurs plus pointus s'appuyant sur des réseaux de neurones.

5 Remerciements

Nos expériences ont été réalisées grâce aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2022-AD011013285R1 attribuée par GENCI. Les travaux présentés dans ce papier ont également reçu un financement de l'Agence Innovation Défense et du programme H2020 de l'Union Européenne, accord de financement No 101021687, projet "UNCOVER".

Références

[1] Patrick BAS, Tomas FILLER et Tomas PEVNY. ""Break Our Steganographic System" : The Ins and Outs of Organizing BOSS". In : *INFORMATION HIDING*. T. 6958/2011. Lecture Notes in Computer Science. Czech Republic, mai 2011, p. 59-70.

[2] Rémi COGRANNE, Quentin GIBOULOT et Patrick BAS. "The ALASKA Steganalysis Challenge : A First Step Towards Steganalysis "Into The Wild"". In : ACM IH&MMSec (Information Hiding & Multimedia Security). Paris, France, juill. 2019.

[3] Quentin GIBOULOT et al. "Effects and Solutions of Cover-Source Mismatch in Image Steganalysis". In : *Signal Processing : Image Communication*. 86e sér. (août 2020).

[4] Jérôme PASQUET, Sandra BRINGAY et Marc CHAUMONT. "Steganalysis with cover-source mismatch and a small learning database". In : *EUSIPCO : European Signal Processing Conference*. Lisbon, Portugal, sept. 2014, p. 2425-2429.

[5] Jan KODOVSKÝ, Vahid SEDIGHI et Jessica FRIDRICH. "Study of cover source mismatch in steganalysis and ways to mitigate its impact". In : *Media Watermarking, Security, and Forensics 2014*. T. 9028.

[6] Rony ABECIDAN et al. "Using Set Covering to Generate Databases for Holistic Steganalysis". In : *IEEE International Workshop on Information Forensics and Security (WIFS 2022)*. Shanghai, China, déc. 2022.

[7] Dominik ŠEPÁK, Lukáš ADAM et Tomáš PEVNÝ. "Formalizing cover-source mismatch as a robust optimization". In : *EUSIPCO : European Signal Processing Conference*. Belgrade, Serbia, sept. 2022.

[8] Vojtěch HOLUB et Jessica FRIDRICH. "Low-Complexity Features for JPEG Steganalysis Using Undecimated DCT". In : *IEEE Transactions on Information Forensics and Security* 10.2 (2015), p. 219-228.

[9] Yoshua BENGIO, Aaron COURVILLE et Pascal VINCENT. "Representation Learning : A Review and New Perspectives". In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), p. 1798-1828.

[10] Linjie GUO et al. "Using Statistical Image Model for JPEG Steganography : Uniform Embedding Revisited". In : *IEEE Transactions on Information Forensics and Security* 10.12 (2015), p. 2669-2680.

[11] Rémi COGRANNE et al. "Is ensemble classifier needed for steganalysis in high-dimensional feature spaces ?" In : *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*. Rome, Italy, 2015.

[12] Jonathan BAC et al. "Scikit-dimension : a Python package for intrinsic dimension estimation". In : *CoRR* abs/2109.02596 (2021).