Deep Learning for Image Steganalysis

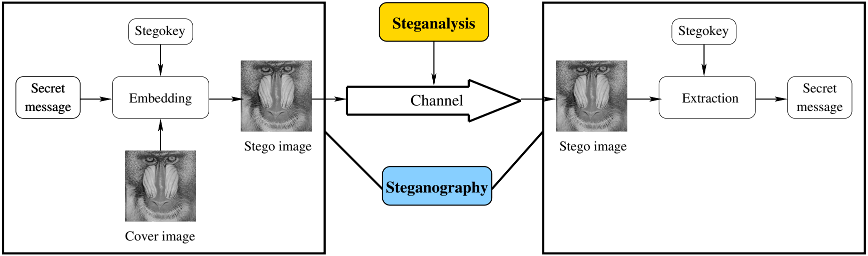
J.-F. Couchot[1], R. Couturier[1], **M. Salomon**[1]

[1]FEMTO-ST Institute - DISC Department - AND Team
Univ. Bourgogne Franche-Comté (UBFC), France

January 30, 2017 / Grenoble
Journées scientifiques Equip@Meso

Introduction

- Steganalysis is the counterpart of steganography
  - Steganography
    - A technique to hide a secret message in a cover media
    - No one apart from the intended recipient knows of the existence of the message
  - Steganalysis
    - A technique to detect whether a cover media embeds a secret message



Introduction

- Steganographic embedding schemes
  - Must change the cover image as little as possible
  - Work mainly in frequency or spatial domain
  - Frequency domain
    - Some coefficients of the chosen transform are changed
    - **D**iscrete **C**osine **T**ransform, **D**iscrete **W**avelet **T**ransform
    - Example : J(PEG)-UNIWARD
  - Spatial domain
    - Some pixel values are changed
    - Examples : S-UNIWARD, MiPOD, HILL, etc.

- Classical image steganalysis scheme
  1. Compute features (more than 30,000) using Rich Models
  2. Train a classifier (mostly FLD Ensemble)

Rich **M**odels+**E**nsemble **C**lassifier

Introduction

- Deep learning has become a breakthrough technology
  - Training large and deep neural networks is now affordable
  - Benefits from the computing power provided by GPU
- Main deep learning approaches
  - ***C**onvolutional **N**eural **N**etworks* when dealing with images
  - ***L**ong **S**hort **T**erm **M**emory* for temporal data
- CNN are competitive for many image classification tasks
  - State-of-the-art for MNIST, CIFAR, etc. (benchmark problems)
  - Image captioning, detection of diabetic retinopathy, etc.
- Motivation
  - Design an alternative to classical RM+EC steganalysis
  - Easiest context: in spatial domain
  - Improve CNN-based steganalyzer results

Outline

1. State of the art of steganography / steganalysis
2. Attempt to understand when the CNN fails
3. Improving the detection accuracy
4. Results
5. Conclusion

Plan

1. State of the art of steganography / steganalysis
2. Attempt to understand when the CNN fails
3. Improving the detection accuracy
4. Results
5. Conclusion

Spatial Domain Steganography

- A distortion function gives for each pixel its modification cost
- S-UNIWARD[1] distortion function

$$\rho_U(X) = \sum_{i=1}^{3} \frac{1}{|X \star K^i| + \sigma} \star |K^i|^{\curvearrowleft}$$

  - X is cover, K is a Daubechies-8 wavelet kernel
  - Small iff large variation of large cover wavelet coeff. in 3 dir.

- MiPOD[2] distortion function
  - Estimate a local Gaussian cover image model
  - Induce a change rate and a distortion cost $\rho_M(X)$

[1] V. Holub, J. Fridrich, T. Denemark, *Universal Distortion Function for Steganography in an Arbitrary Domain*, EURASIP J. on Inf. Se., 2014(1).

[2] V. Sedighi, R. Cogranne and J. Fridrich, *Content-Adaptive Steganography by Minimizing Statistical Detectability*. IEEE Transactions on Information Forensics and Security. 11(2): 221-234 (2016).

Spatial Domain Steganography

- HILL[3] distortion function

$$\rho_H(X) = \frac{1}{|X \star H_1| \star L_1} \star L_2, \text{ where } H_1 = \begin{bmatrix} -1 & 2 & -1 \\ 2 & -4 & 2 \\ -1 & 2 & -1 \end{bmatrix}$$

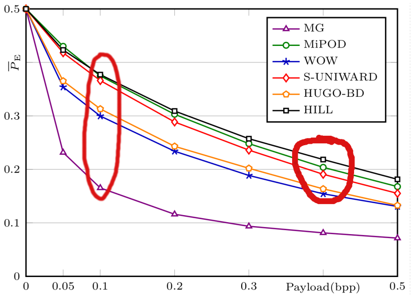
  - X is cover, H_1 is a high-pass filter
  - L_1 and L_2 are low-pass filters

- The distortion function value reflects the cover image model
  - Easy-defined or smooth areas $\rightarrow$ large value
  - Texture or "chaotic" areas $\rightarrow$ small value

[3] B. Li, M. Wang, J. Huang, X. Li, *A New Cost Function for Spatial Image Steganography*, 2014 IEEE International Conference on Image Processing (ICIP). pp. 4206-4210, 2014.

Spatial Rich Models+Ensemble Classifier Steganalysis

- Features: maxSRMd2[4] + Classifier: FLD Ensemble[5]
- Last known results[2]:



[4] T. Denemark, V. Sedighi, V. Holub, R. Cogranne and J. Fridrich, Selection-Channel-Aware Rich Model for Steganalysis of Digital Images, IEEE Workshop on Information Forensic and Security, 2014.

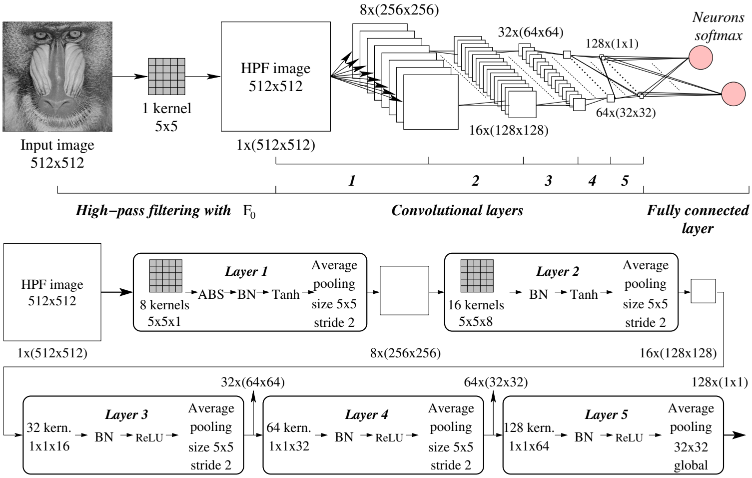
[5] J. Kodovský, J. Fridrich, and V. Holub, Ensemble Classifiers for Steganalysis of Digital Media. IEEE Transactions on Information Forensics and Security, Vol. 7, No. 2, pp. 432-444, April 2012

CNN-based Steganalysis Approaches

- Non-exhaustive list
  - Y. Qian, J. Dong, W. Wang, T. Tan (2015)
    - *Deep learning for steganalysis via convolutional neural networks*
      IS&T/SPIE Electronic Imaging, pp. 94090J–94090J
  - L. Pibre, J. Pasquet, D. Ienco, M. Chaumont (2016)
    - *Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover source mismatch*
      Media Watermarking, Security, and Forensics 2016: 1-11
  - G. Xu, H.-Z. Wu, Y.-Q. Shi (2016) (Xu *et al.*)
    - *Structural Design of Convolutional Neural Networks for Steganalysis*
      IEEE Signal Processing Letters, vol. 23, num. 5, pp. 708–712
- Remarks
  - Highlighting noise residuals with F_0 filter seems mandatory
  - Some results limited by the use of fixed embedding patterns
  - Most competitive CNN with SRM is due to Xu *et al.* (2016)

CNN proposed by Xu *et al.* (2016)

- Architecture



How to further reduce the detection error?

- First idea: change parameters in the CNN architecture?
  - High-pass filtering with F_0
  - Convolutional layers configuration
  - etc.
- Second idea: use the best of both kind of approaches?
  - Practically investigate the CNN designed by Xu *et al.* (2016)
  - Original implementation
    - Caffe toolbox
    - Average of the predictions given by 5 CNNs
    - Training & testing on S-UNIWARD, HILL (payload: 0.1, 0.4 bpp)
  - Our implementation
    - Tensorflow library
    - Average of the predictions given by CNNs from the last 20 epochs
    - Training using MiPOD, testing on S-UNIWARD, MiPOD, HILL
  - Is SRM+EC a better alternative when the CNN fail?

Experimental setup

- Database of 10,000 (cover,stego) pairs for a given payload
  - Built from BOSSBase
    - Gray level images of 512×512 pixels
  - Training and testing sets built randomly
    - 5,000 pairs for training
    - 5,000 pairs for testing
- Training setting
  - SGD+momentum optimizer
  - Mini-batch size of 64 samples
  - Stopping criterion = maximum number of training epochs
    - 300 epochs for 0.4 bpp payload
    - 1,000 epochs for 0.1 bpp payload

Experimental setup

- GPU computation facilities
  - Program development $\rightarrow$ 1 NVIDIA GPU Titan X
  - Mesocentre $\rightarrow$ node of 4 NVIDIA GPU Tesla K40
- Training times on NVIDIA GPU Titan X
  - payload of 0.4 bpp $\rightarrow \approx$ 3 days for “very good” results
  - payload of 0.1 bpp $\rightarrow \approx$ 7 days for “good” results

15,000 hours of calculations using the Mesocentre

Detection error of SRM+EC / the CNN

- Average detection error

	S-UNIWARD		MiPOD		HILL	
	0.1	0.4	0.1	0.4	0.1	0.4
Caffe (Xu *et al.*)	42.67	19.76	X	X	41.56	20.76
TensorFlow (blind)	X	20.52	X	19.36	X	20.25
SRM + EC	39.84	18.06	41.18	21.42	42.96	23.31
SRM + EC (blind)	40.57	20.85	41.18	21.42	43.35	23.99

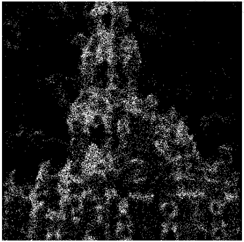
- SRM+EC and CNN have similar detection performances
- Tensorflow implementation can perform blind steganalysis

Plan

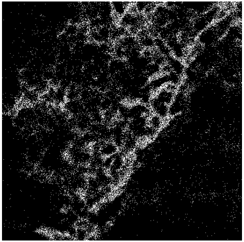
1. State of the art of steganography / steganalysis
2. Attempt to understand when the CNN fails
3. Improving the detection accuracy
4. Results
5. Conclusion

Examples of well-CNN-classified images

1388.pgm



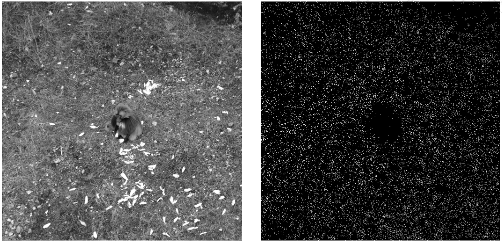
8873.pgm



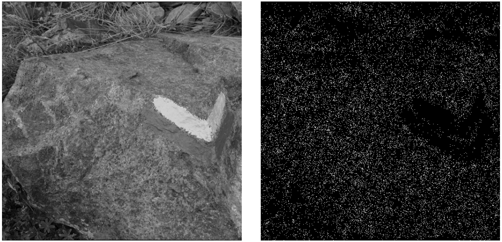
Embedding is performed using MiPOD with a payload of 0.4 bpp

Examples of mis-CNN-classified images

1911.pgm



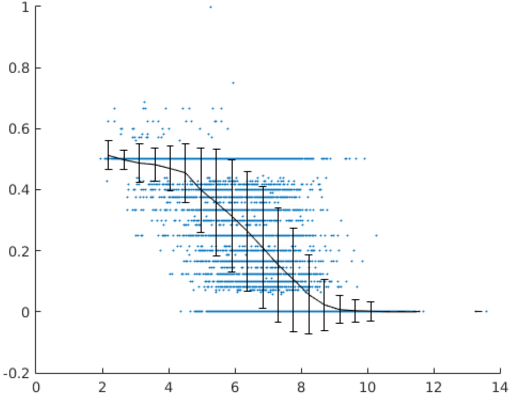
3394.pgm



Embedding is performed using MiPOD with a payload of 0.4 bpp

Characterization of mis-CNN-classified images

- $\overline{\rho_U}$: average pixel distorsion cost of S-UNIWARD image
  - Average of 12 CNNs on the BOSSBase



Detection error w.r.t image $\overline{\rho_U}$ value for the CNN by *Xu et al.*

Characterization of mis-CNN-classified images

- Quiz: Can you guess the $\overline{\rho_U}$ value for each image?

$\overline{\rho_U}_{1388} = ?$

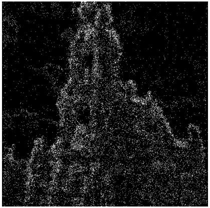
$\overline{\rho_U}_{8873} = ?$

$\overline{\rho_U}_{1911} = ?$

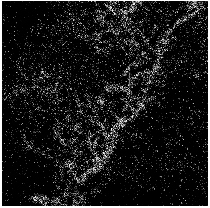
$\overline{\rho_U}_{3394} = ?$

Characterization of mis-CNN-classified images

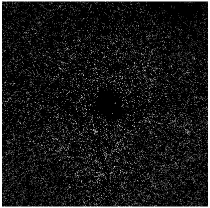
- Quiz: Can you guess the $\overline{\rho_U}$ value for each image?



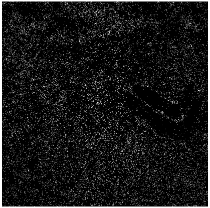
$\overline{\rho_U}_{1388} = 7.05$



$\overline{\rho_U}_{8873} = 7.39$



$\overline{\rho_U}_{1911} = 2.1$

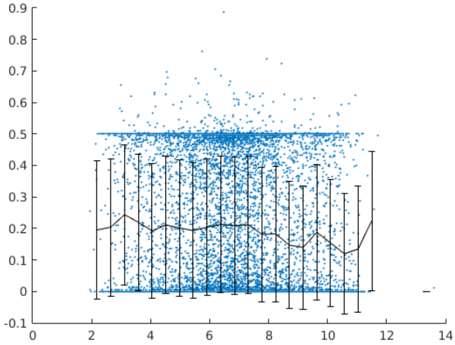


$\overline{\rho_U}_{3394} = 3.06$

Plan

Detection error of SRM+EC w.r.t. $\overline{\rho_U}$

- Experimental setup
  - maxSRMd2 features+Ensemble Classifier
  - Average of 200 runs on the BOSSBase



Detection error w.r.t image $\overline{\rho_U}$ value for SRM+EC.

How to choose the better classifier?

- Compute $\overline{\rho_U^{\cap}}$ corresponding to the intersection

- For image I compute $\overline{\rho_U}(I)$
  - if $\overline{\rho_U}(I) < \overline{\rho_U^{\cap}}$ use SRM+EC prediction
  - otherwise use CNN prediction

Plan

1. State of the art of steganography / steganalysis
2. Attempt to understand when the CNN fails
3. Improving the detection accuracy
4. Results
5. Conclusion

Results

- Average detection error according to $\overline{\rho_U^\cap}$ (payload of 0.4 bpp)

	SRM+EC	$\overline{\rho_U^\cap}$	CNN	Proposal	SRM+EC	CNN
S-UNIWARD non blind	20.01	7.1	8.25	14.82	18.06	19.76
S-UNIWARD blind	22.05	6.9	9.50	15.87	20.85	X
MiPOD non blind	23.89	6.6	9.26	15.65	21.42	X
HILL non blind	24.51	6.6	9.78	16.22	23.31	20.76
HILL blind	25.41	6.6	9.78	16.61	23.99	X

Results

- Average detection error according to $\overline{\rho_U^{\cap}}$ (payload of 0.1 bpp)

	SRM+EC	$\overline{\rho_U^{\cap}}$	CNN	Proposal	SRM+EC	CNN
S-UNIWARD non blind	40.08	9.2	23.36	38.06	39.84	42.67
S-UNIWARD blind	41.00	9.2	23.36	38.88	40.57	X
MiPOD non blind	42.13	8.0	25.84	37.82	41.18	X
HILL non blind	43.48	8.9	21.88	40.24	42.96	41.56
HILL blind	44.30	8.3	27.72	40.64	43.35	X

Plan

1. State of the art of steganography / steganalysis
2. Attempt to understand when the CNN fails
3. Improving the detection accuracy
4. Results
5. Conclusion

Conclusion

- A criterion to choose the appropriate steganalyzer
  - Lower detection errors
  - Blind steganalysis
  - State-of-the-art results
- More powerful GPU computing facilities will be needed
  - To deal with larger datasets and reduce the training time
  - To stay in the competition with other teams

  JPEG steganalysis using hybrid deep-learning by Zeng *et al.* (2016)
    - Use 3 Xu *et al.* "subnetworks" (1,536 features)
    - Trainings with 50K, 500K and 5,000K JPEG images
    - Cluster of 8 NVIDIA Tesla K80

Thank you for your attention

Any questions ?