Detection of Malicious Spatial-Domain Steganography over Noisy Channels Using Convolutional Neural Networks

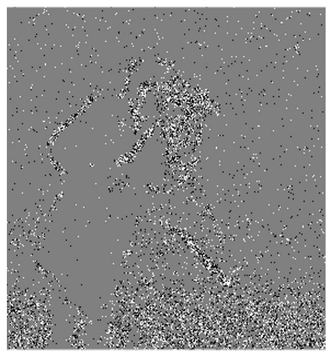
Swaroop Shankar Prasad[1], Ilia Polian[1], Ofer Hadar[2]

[1]University of Stuttgart, Germany [2]Ben-Gurion University of the Negev, Israel

- Steganography can be used for legitimate or malicious purposes
- Detection of malicious spatial-domain steganography inserted by untrustworthy hardware
- Novelty: Transmission channel affected by noise (Gaussian, packet loss)
- Neural Network for reliable detection of steganography bits inserted by state-of-the-art algorithms
- Noise affects both: the malicious communication and the detection procedure

Steganography : Hide information in multimedia content

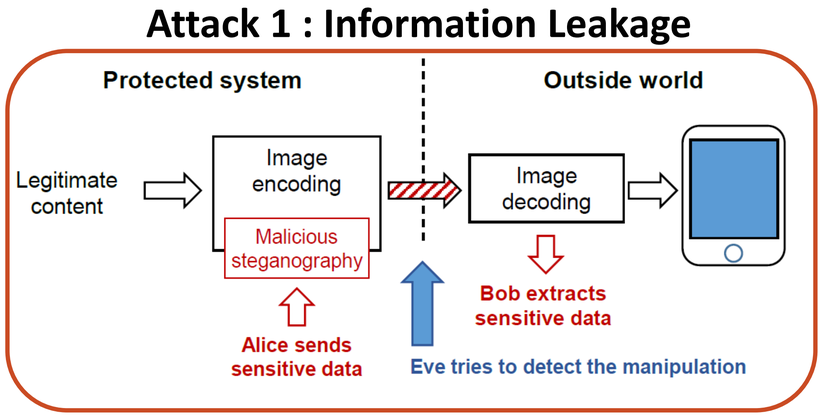
Cover Image



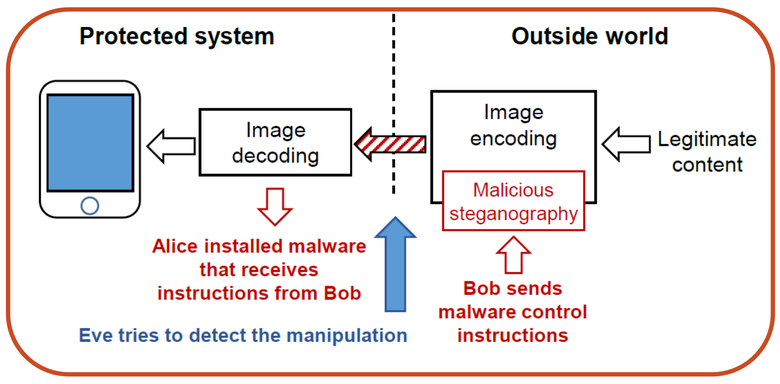
Embedded malicious steganography bits

Steganography Image

Attack 1 : Information Leakage

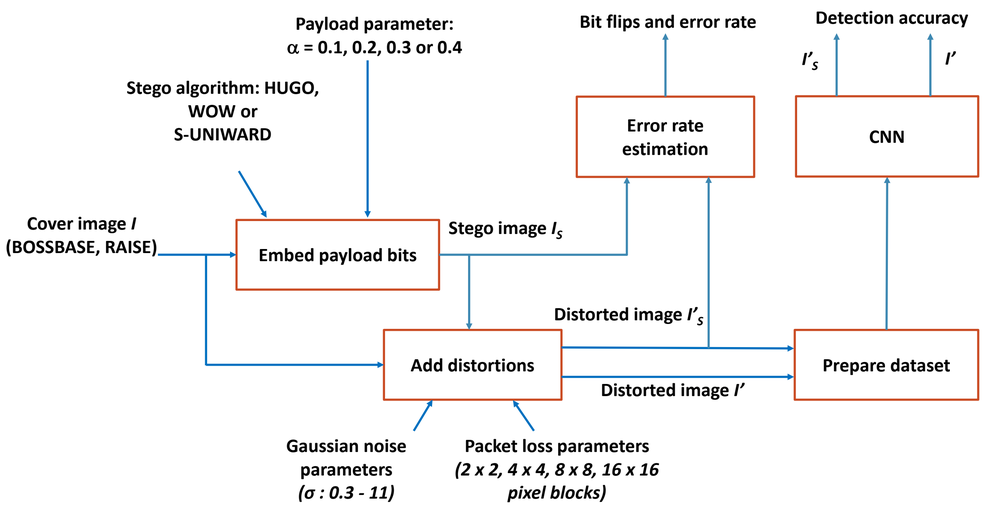


Attack 2 : Malware Control



METHODOLOGY

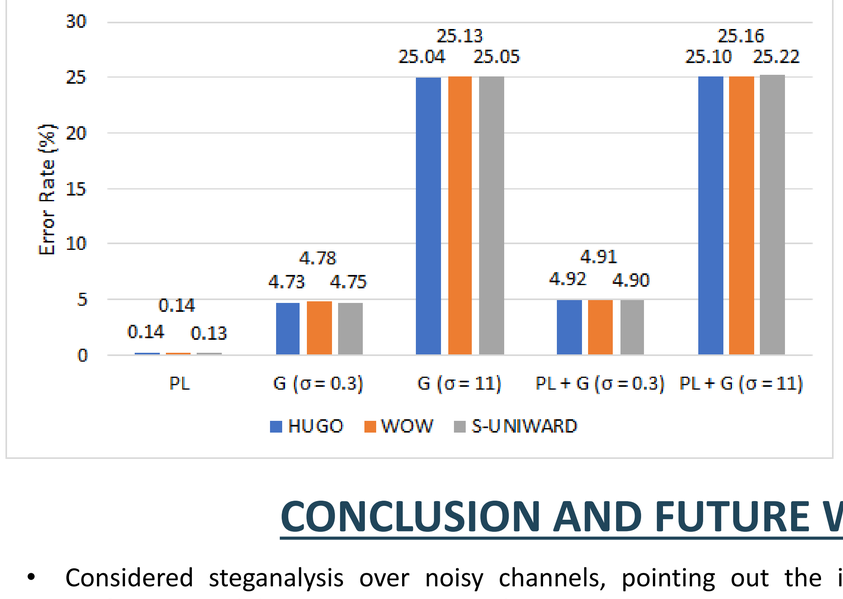
- Three state-of-the-art spatial-domain steganography algorithms HUGO [2], WOW [3], and S-UNIWARD [4] with fixed-stego key implementations considered
- Stego image transmitted over a channel affected by two kinds of noise: Gaussian noise and packet loss
- Error rate: $\frac{\text{Number of bit flips}}{\text{Total number of embedded stego bits}}$



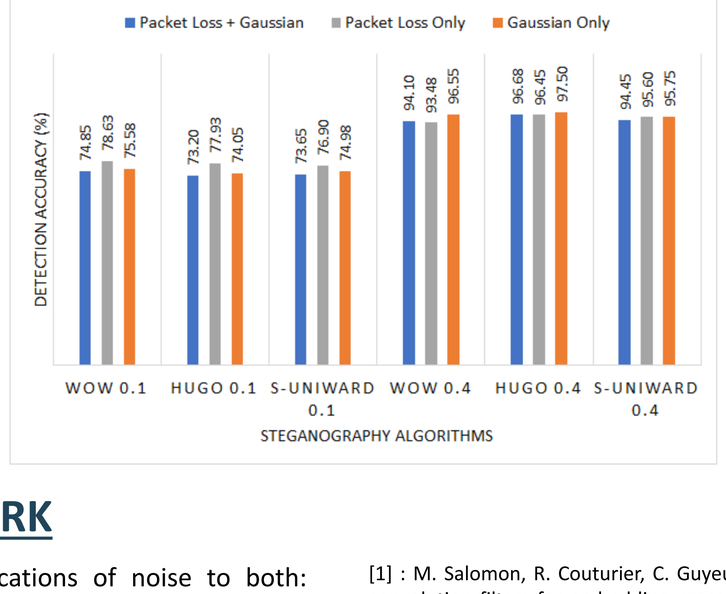
- Detection (or steganalysis) : using neural network from [1] to classify images I' or I'_s
- Early stopping with patience = 5 to overcome over-fitting and improve detection accuracy
- Datasets: BOSSBASE [5] and RAISE [6]
- 22000 training, 8000 validation and 4000 test images

EXPERIMENTAL RESULTS

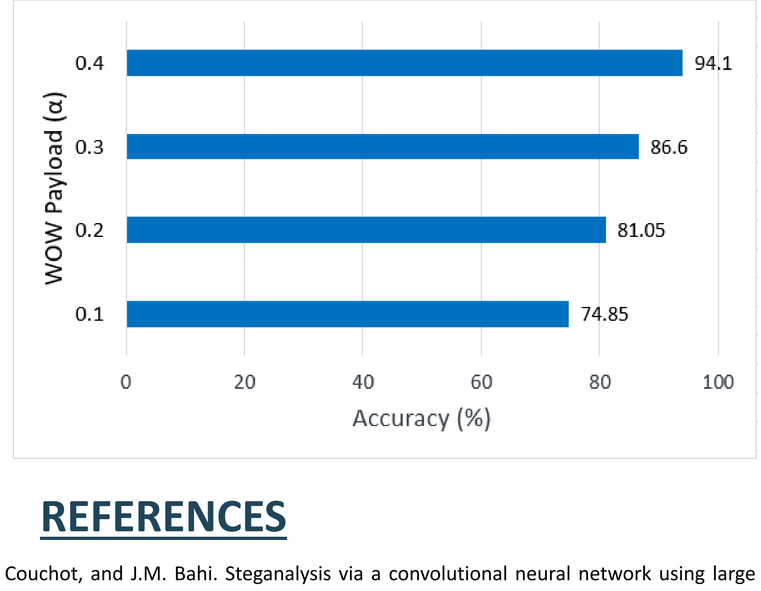
Error rate of steganographic channel when image transmitted with noise levels leading to PSNR = 30-60
(G : Gaussian, PL : packet loss)



Detection accuracy for different spatial-domain steganography algorithms and noise types/levels



Detection accuracy for different payloads (percentages of used stego bit positions) for WOW algorithm



CONCLUSION AND FUTURE WORK

- Considered steganalysis over noisy channels, pointing out the implications of noise to both: performing and detecting steganography. Good detection accuracy on realistic images
- Steganalysis is feasible as payload in an image increases. Even moderate noise have grave consequences for the steganographic channel, leading to large error rates
- While an adversary could find a fault-tolerant scheme to still transmit the information with large number of redundant bits which can be detectable by our deep learning approach
- Future work: design countermeasures against malicious steganography based on strategically inducing errors on the communication channel while avoiding too harsh consequences for the image quality

REFERENCES

[1] : M. Salomon, R. Couturier, C. Guyeux, J.-F. Couchot, and J.M. Bahi. Steganalysis via a convolutional neural network using large convolution filters for embedding process with same stego key: A deep learning approach for telemedicine. European Research in Telemedicine, 6(2):79–92, 2017.

[2] : T. Filler and J. J. Fridrich. Gibbs construction in steganography. IEEE Trans. Information Forensics and Security, 5(4):705–720, 2010.

[3] : V. Holub and J. J. Fridrich. Designing steganographic distortion using directional filters. In WIFS, pages 234–239. IEEE, 2012.

[4] : V. Holub, J. J. Fridrich, and T. Denemark. Universal distortion function for steganography in an arbitrary domain. EURASIP J. Information Security, 2014:1, 2014.

[5] : P. Bas, T. Filler, and T. Pevn´y. ”Break our steganographic system”: The ins and outs of organizing BOSS. In Information Hiding, volume 6958 of LNCS, pages 59–70. Springer, 2011.

[6] : D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato. RAISE: a raw images dataset for digital image forensics. In MMSys, pages 219–224. ACM, 2015.

[7] : S. Shankar Prasad, O. Hadar, and I. Polian. "Detection of malicious spatial-domain steganography over noisy channels using convolutional neural networks.“, Electronic Imaging 2020.

University of Stuttgart

Ben-Gurion University of the Negev
Inspiration meets Excellence