Adversarial Embedding in the JPEG Domain Induces Correlations Between DCT Coefficients to Remove Blocking Artifacts Generated by Additive Embedding

Solène Bernard, Patrick Bas, John Klein, Tomas Pevny

Adversarial Embedding in the JPEG Domain Induces Correlations Between DCT Coefficients to Remove Blocking Artifacts Generated by Additive Embedding

Solène Bernard, Patrick Bas, John Klein and Tomas Pevny

November 25, 2020

This short note presents an analysis of the principle of adversarial embedding (a.k.a. adv-emb) in steganography as presented in [4]. By analyzing the covariance matrix of the stego signal of quantized JPEG coefficients (i.e. the signal added to the JPEG Cover image to create the Stego image), we highlight the fact that the stego signal exhibits correlations between coefficients. These weak correlations are within the same block (intra-block correlations) of between adjacent blocks (inter-block correlations). The correlation patterns are similar to the patterns of correlations analyzed on the sensor noise in the DCT domain (see [3]). However, if in [3] these correlations have been shown to favor *continuities* between blocks, the correlations induced by adversarial embedding are on opposite sign and they code *discontinuities* between blocks (see Figure 1b). An experiment consisting in generating first a cover image using J-Uniward [1] and then computing an adversarial signal using PGD [2] exhibits very similar patterns than with adversarial embedding (see figure 1c). We consequently presume that, as pictured in Figure 2, the adversarial signal is generated in order to compensate the blocking artifacts created by the embedding. This would explain the sign of the correlations. Our rationale is that the proportion β of the DCT coefficients used produce the adverse stego signal in adv-emb conveys the signal whose role is to try to remove the blocking artifacts generated by additive schemes such as J-Uniward.

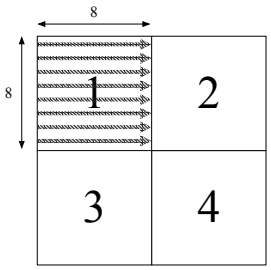
References

[1] Vojtěch Holub, Jessica Fridrich, and Tomáš Denemark. Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, 2014(1):1–13, 2014.

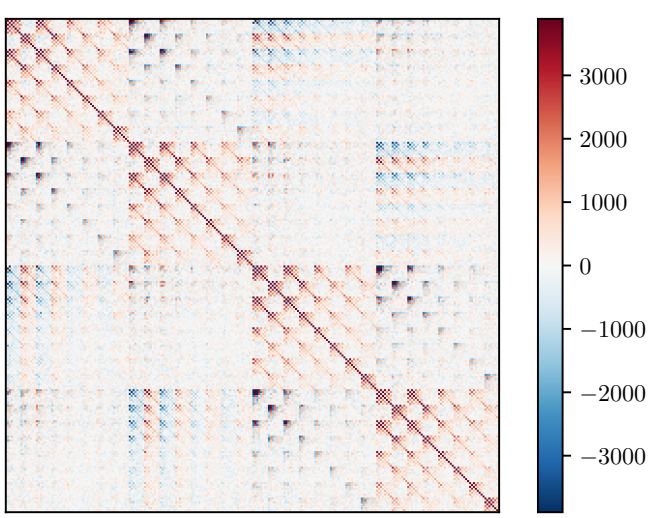
[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[3] T. Taburet, P. Bas, W. Sawaya, and J. Fridrich. Natural steganography in jpeg domain with a linear development pipeline. *IEEE Transactions on Information Forensics and Security*, 16:173–186, 2021.

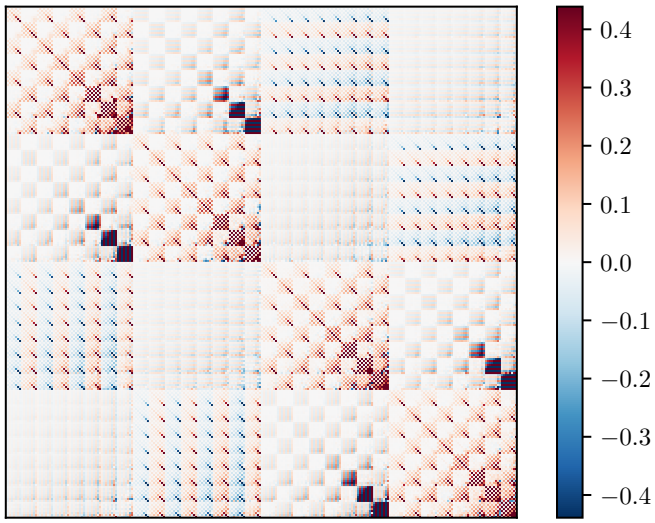
[4] Weixuan Tang, Bin Li, Shunquan Tan, Mauro Barni, and Jiwu Huang. Cnn-based adversarial embedding for image steganography. *IEEE Transactions on Information Forensics and Security*, 2019.



(a) Scan Order.

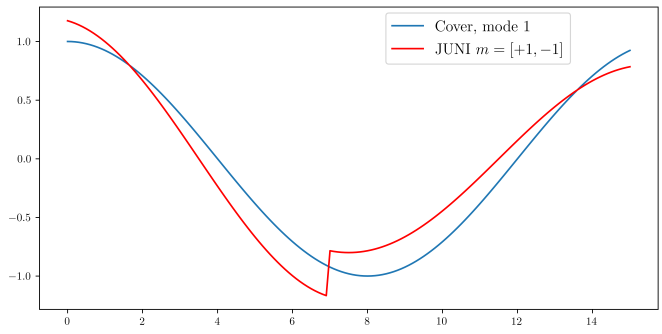


(b) Covariance matrix computed by the stego signal after adv-emb using 7000 images of BOSSBase decomposed into non overlapping 24×24 blocks. JPEG QF100. (2 thresholding are applied to reduce the range).

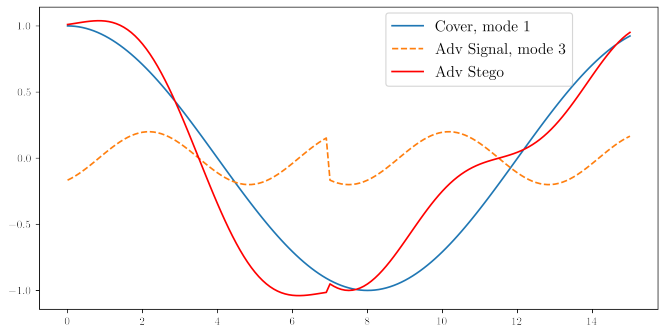


(c) Covariance matrix of the PGD adversarial signal computed by using 500 images of BOSSBase (2 thresholding are applied to reduce the range).

Figure 1: Covariance matrices.



(a) 1D continuous version of the Cover and Stego signals (here the cover is modeled by the mode 1 of the DCT basis).



(b) 1D continuous version of the Adversarial Signal (represented by mode 3 of the DCT basis) and the Adversarial Stego signal.

Figure 2: Embedding $[+1, -1]$ on 2 neighboring blocks for a Cover signal represented by the DCT mode (1). Two lines of two neighboring blocks are depicted.