CNN Steganalyzers Leverage Local Embedding Artifacts

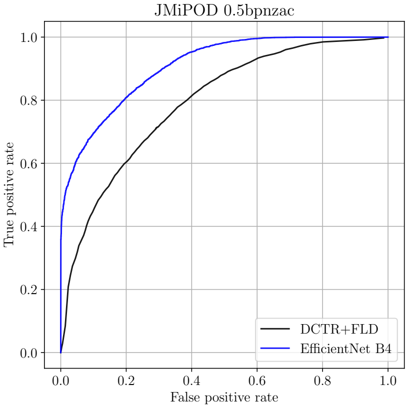
Yassine Yousfi, Jan Butora, and Jessica Fridrich

WIFS 2021

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

CNNs >> Rich Models

- Much lower FAs
- Non-Gaussian ROC



The usual hand-waving argument

- RMs are **global** while CNNs have the ability to be **local**
- To our knowledge, this remains a conjecture
- More broadly: we wish to learn from deep learning
- Better understand how CNNs arrive at their decisions

Our findings

- CNNs are both **integrators**, leveraging some form of CLT for detection, and detectors of **local embedding artifacts**
- Some algorithms (J-MiPOD) introduce numerous Locally DEtectable Artifacts (LDEAs) while others do not (J-UNIWARD)
- **RMs** are **unable** to use LDEAs

Experimental Setup

- Alaska II 256×256 QFs 75, 90, and 95 [Cogranne et al. WIFS2020]
- EfficientNet B4 (trained as in Alaska II) [Yousfi et al. WIFS2020]
- SRNet [Boroumand et al. TIFS2018]

Selected payloads

EfficientNet B4

	Payload (bpnzac)	P_E	MD5	wAUC
J-MiPOD	0.5	.1938	.3837	.9349
J-MiPOD	0.2	.3452	.7033	.8067
J-UNIWARD	0.5	.1967	.4220	.9304
J-UNIWARD	0.2	.3606	.7658	.7792
F5	0.2	.1835	.4292	.9292
$-$F5	0.05	.0866	.1248	.9827
Jsteg	0.0112	.1315	.2207	.9595

Toolbox

- **Integrated Gradients**[1]

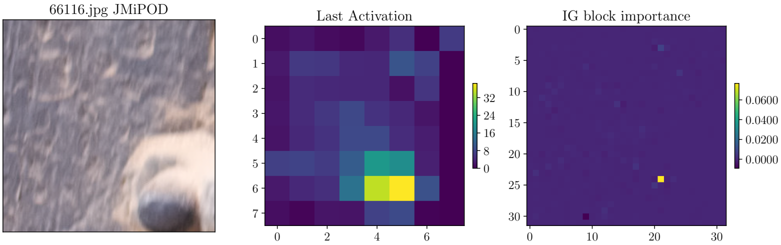
$$\phi(f, s, b) = (s - b) \odot \int_0^1 \frac{\mathrm{d}f\,(b + \alpha(s - b))}{\mathrm{d}s}\,\mathrm{d}\alpha,$$

  averaged over 8×8 non-overlapping blocks along the spatial dimensions to get IG block importance.

- **Last Activation Map**: Remove the last global pooling and use the Fully Connected layer's weights and biases as a 1×1 convolution.

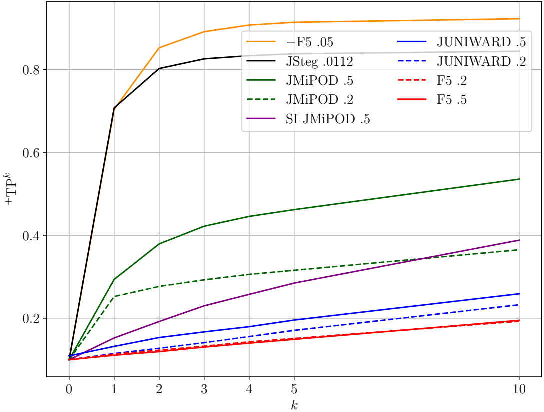
[1] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." International Conference on Machine Learning. PMLR, 2017.

Toolbox


66116.jpg JMiPOD
Last Activation
0
1
2
3
4
5
6
7
0
2
4
6
32
24
16
8
0
IG block importance
0
5
10
15
20
25
30
0
10
20
30
0.0600
0.0400
0.0200
0.0000

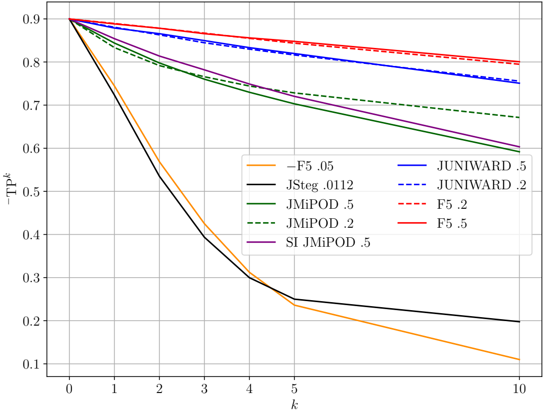
Top-k insertion

- Start with a cover image, and insert the top-k stego blocks with largest IG. Thresholds are set for FP rate $= 10\%$.



Top-k canceling

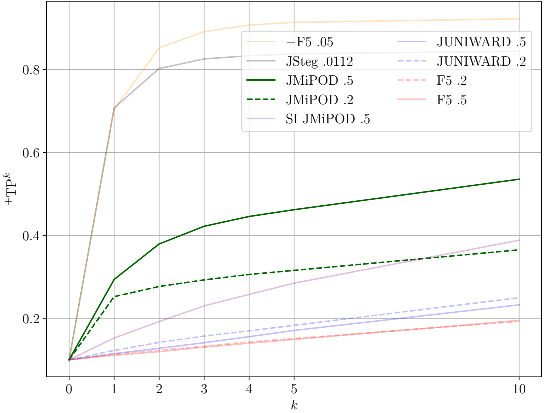
- Start with a stego image, and cancel the changes in the top-k stego blocks with largest IG. Thresholds are set for TP rate $= 90\%$.



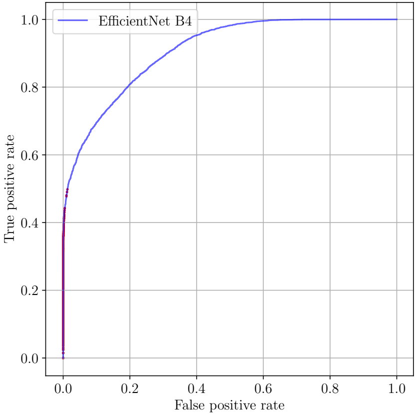
Locally Detectable Embedding Artifacts (LDEAs)

- A Locally Detectable Embedding Artifact is a stego artifact that can trigger a detection (by a CNN). Typically local to a 8×8 JPEG block.
- We show that CNNs are able to leverage these artifacts.
- We use IG block importance to find the LDEAs that can be detected by CNNs.
- Images that can be detected as stego with only a small number of changes inserted (small k) are said to have LDEAs.
- Those images transfer between CNNs: for J-MiPOD 0.5 bpnzac 82% of SRNet's LDEAs are shared with EfficientNet B4.

J-MiPOD

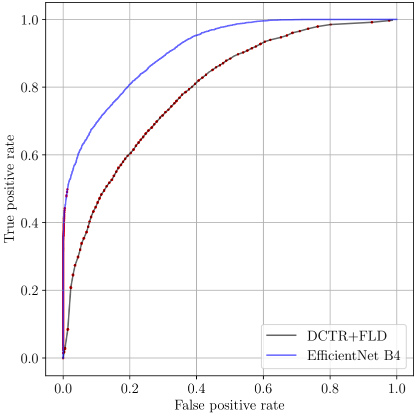

−F5 .05
JSteg .0112
JMiPOD .5
JMiPOD .2
SI JMiPOD .5
JUNIWARD .5
JUNIWARD .2
F5 .2
F5 .5
0.8
0.6
0.4
0.2
$^{+}\mathrm{TP}^{k}$
0
1
2
3
4
5
10
k

J-MiPOD LDEAs are “easy stegos” ...



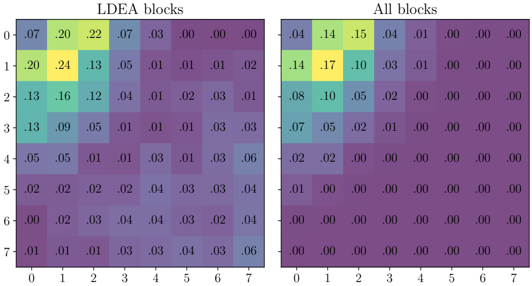
EfficientNet B4 - J-MiPOD 0.5 bpnzac

J-MiPOD LDEAs are “easy stegos” for CNNs



EfficientNet B4 and DCTR+FLD - J-MiPOD 0.5 bpnzac

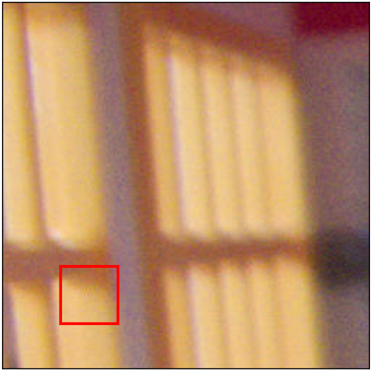
Change rate of J-MiPOD LDEAs

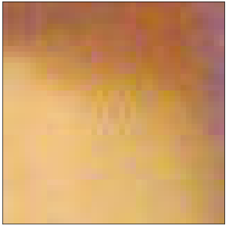


Average change rate per DCT mode

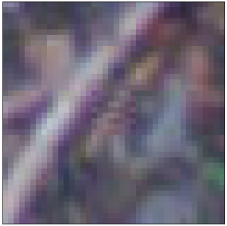
- A larger change rate than the average 8×8 block of J-MiPOD.
- More changes in high frequency DCT coefficients (usually zeros).

Examples of J-MiPOD LDEAs

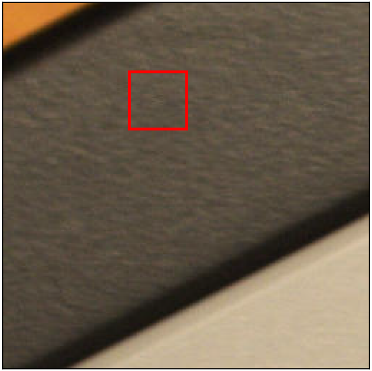




Examples of J-MiPOD LDEAs

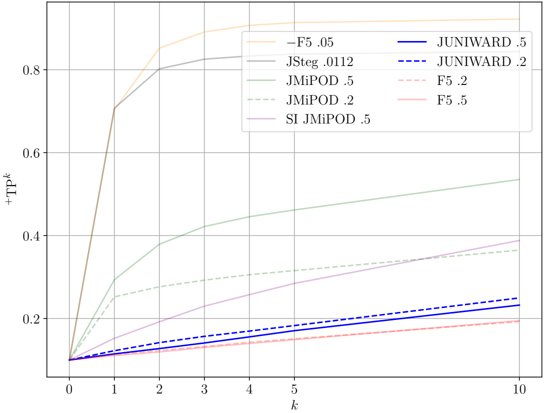


Examples of J-MiPOD LDEAs



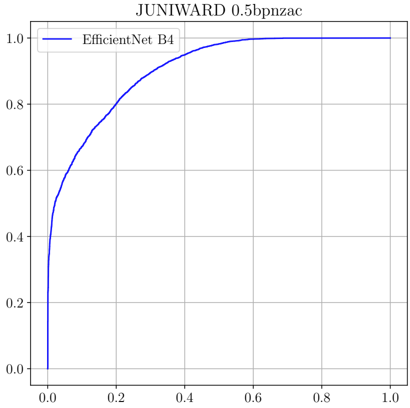
Examples of J-MiPOD LDEAs

J-UNIWARD

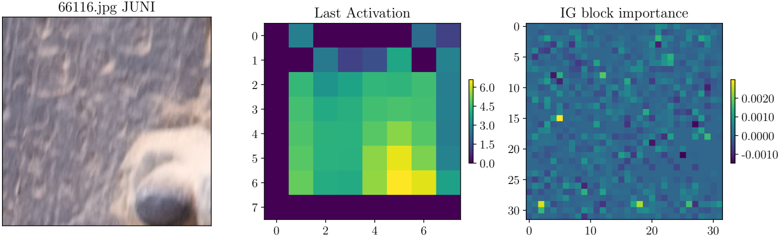

−F5 .05
JSteg .0112
JMiPOD .5
JMiPOD .2
SI JMiPOD .5
JUNIWARD .5
JUNIWARD .2
F5 .2
F5 .5
0.8
0.6
0.4
0.2
$^{+}\mathrm{TP}^{k}$
0
1
2
3
4
5
10
k

J-UNIWARD

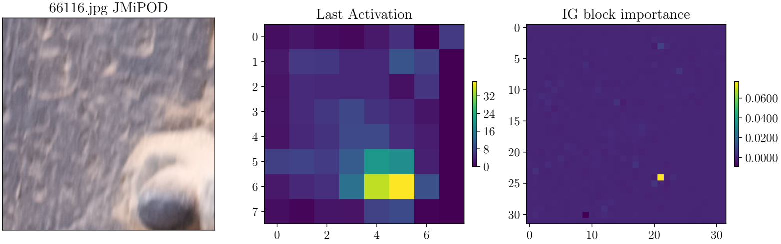
- Fewer LDEAs than J-MiPOD
- But the ROC curve is still high for low FP rates



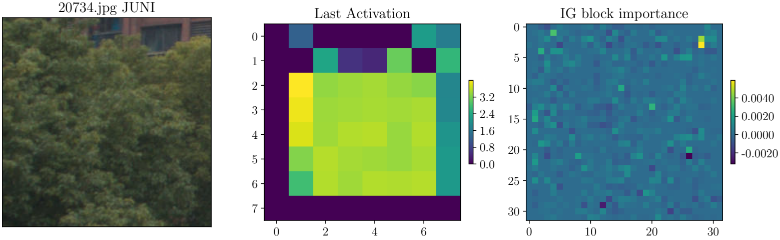
For J-UNIWARD, CNNs are integrators



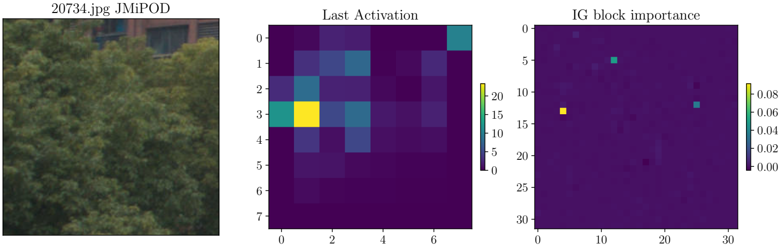
For J-UNIWARD, CNNs are integrators



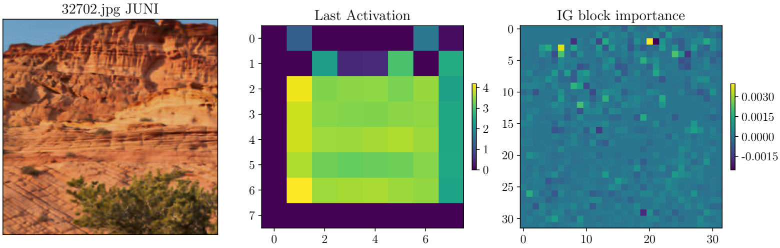
For J-UNIWARD, CNNs are integrators



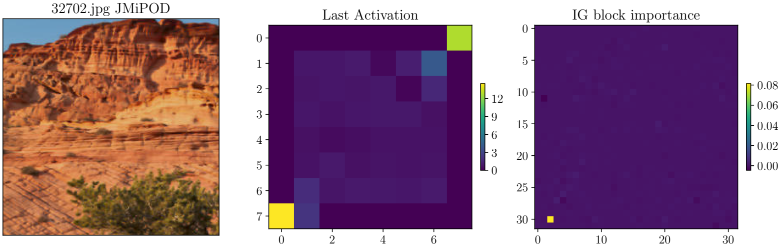
For J-UNIWARD, CNNs are integrators



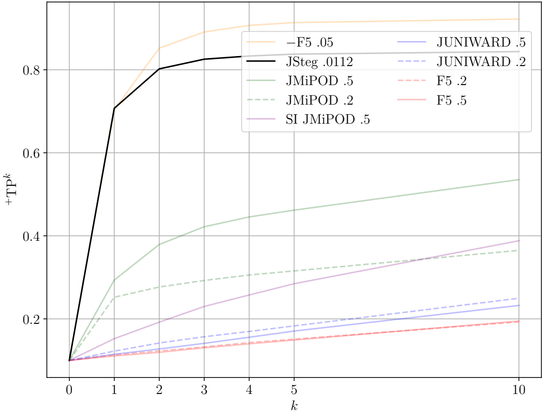
For J-UNIWARD, CNNs are integrators



For J-UNIWARD, CNNs are integrators



Jsteg

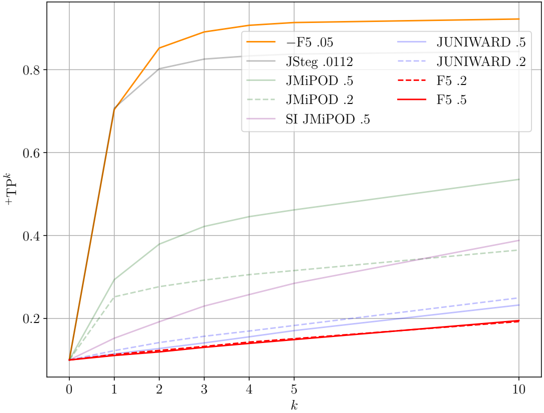


−F5 .05
JSteg .0112
JMiPOD .5
JMiPOD .2
SI JMiPOD .5
JUNIWARD .5
JUNIWARD .2
F5 .2
F5 .5
0.8
0.6
0.4
0.2
$+\mathrm{TP}^k$
0
1
2
3
4
5
10
k

Jsteg

- Introduces **many** LDEAs.
- Most of them are related to changes increasing the absolute value of the DCT coefficient.
- 98.01% of changes in LDEA blocks increase the absolute value VS 65.06% across all blocks.

-F5, F5



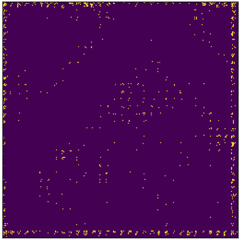
−F5 .05
JSteg .0112
JMiPOD .5
JMiPOD .2
SI JMiPOD .5
JUNIWARD .5
JUNIWARD .2
F5 .2
F5 .5
0.8
0.6
0.4
0.2
+TP^k
0
1
2
3
4
5
10
k

Stego inhibition

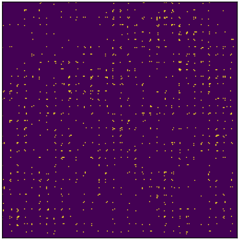
- What are the CNNs looking at in the case of multiclass detection? How different is it from the binary case?
- Multiclass J-UNIWARD and bUERD

03460.jpg

bUERD changes



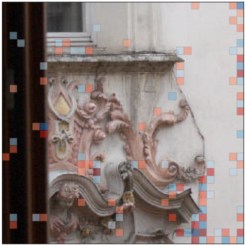
JUNI changes



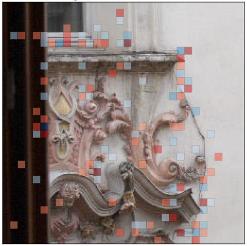
Stego inhibition

bUERD attribution

J-UNIWARD attribution

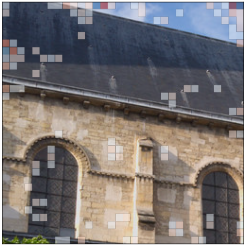


Binary J-UNIWARD attribution

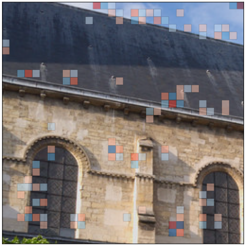


Stego inhibition

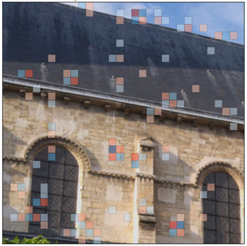
bUERD attribution



J-UNIWARD attribution



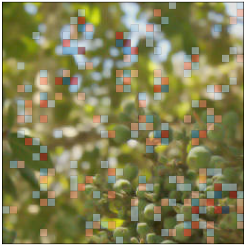
Binary J-UNIWARD attribution



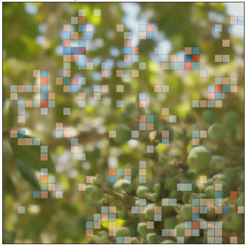
Stego inhibition

bUERD attribution

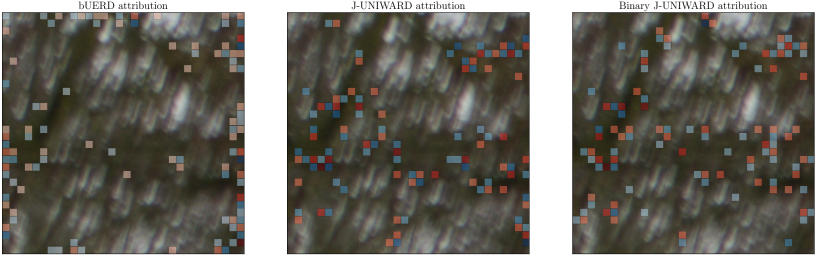
J-UNIWARD attribution



Binary J-UNIWARD attribution



Stego inhibition



Conclusions

We provide evidence that CNNs make use of **highly localized** information, unlike RMs

- **L**ocally **D**etectable **E**mbedding **A**rtifacts (can even be identified visually)
- Jsteg, –F5 introduce many LDEAs due to content-creating changes $0 \to \pm 1$
- J-MiPOD introduces many more LDEAs than J-UNIWARD

CNNs also use **localized traces** to distinguish between selection channels of different embedding algorithms (bUERD vs. J-UNI)