Extending the Reverse JPEG Compatibility Attack to Double Compressed Images

Jan Butora and Jessica Fridrich

ICASSP 2021

BINGHAMTON
UNIVERSITY
STATE UNIVERSITY OF NEW YORK

Reverse JPEG Compatibility Attack

- Extremely accurate attack on JPEG steganography [Butora et al. 2020]
- Discovered during ALASKA I challenge
- Detects any steganography (universal)
- Can be combined with selection channel [Cogranne 2020]
- Limited to $QF \geq 99$ (18% of images on Flickr) for single-compressed images

Extension to double compressed images

- As long as $93 \leq QF_1 \leq QF_2$
- Especially accurate when $QF_1 = QF_2$
- Recompression
  - may be introduced by the stego tool
  - occurs when retouching an image ($QF_1 = QF_2$)
  - can be due to cover preprocessing before embedding

RJCA notation

- x_{ij} - pixel values of uncompressed image (integers)
- y_{ij} - pixel values of decompressed image (floats)
- q_{kl} - quantization steps
- c_{kl}, d_{kl} - quantized/unquantized DCT coefficients
- e_{kl} - DCT domain rounding error

RJCA notation

- x_{ij} - pixel values of uncompressed image (integers)
- y_{ij} - pixel values of decompressed image (floats)
- q_{kl} - quantization steps
- c_{kl}, d_{kl} - quantized/unquantized DCT coefficients
- e_{kl} - DCT domain rounding error

$$
\begin{aligned}
y_{ij} &= \mathrm{DCT}_{ij}^{-1}(\mathbf{c} \cdot \mathbf{q}) \\
&= \mathrm{DCT}_{ij}^{-1}(\mathbf{d}) - \mathrm{DCT}_{ij}^{-1}(\mathbf{e} \cdot \mathbf{q}) \\
&= x_{ij} - \mathrm{DCT}_{ij}^{-1}(\mathbf{e} \cdot \mathbf{q})
\end{aligned}
$$

Rounding Error in Spatial Domain

- DCT error modeled as $e_{kl} \sim \mathcal{U}(-1/2, 1/2)$ i.i.d.

Rounding Error in Spatial Domain

- DCT error modeled as $e_{kl} \sim \mathcal{U}(-1/2, 1/2)$ i.i.d.
- CLT $\rightarrow$ This error propagates through decompression as Gaussian noise

Rounding Error in Spatial Domain

- DCT error modeled as $e_{kl} \sim \mathcal{U}(-1/2, 1/2)$ i.i.d.
- CLT $\rightarrow$ This error propagates through decompression as Gaussian noise

$$y_{ij} \sim \mathcal{N}(x_{ij}, s_{ij})$$

where

$$s_{ij} = \frac{1}{12} \sum_{k,l=0}^{7} \left(f_{kl}^{ij}\right)^2 q_{kl}^2$$

and f_{kl}^{ij} are the discrete cosines

Rounding Error in Spatial Domain

- DCT error modeled as $e_{kl} \sim \mathcal{U}(-1/2, 1/2)$ i.i.d.
- CLT $\rightarrow$ This error propagates through decompression as Gaussian noise

$$y_{ij} \sim \mathcal{N}(x_{ij}, s_{ij})$$

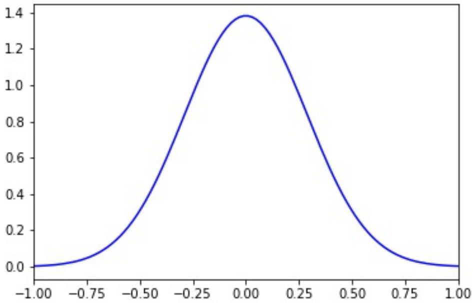
where

$$s_{ij} = \frac{1}{12} \sum_{k,l=0}^{7} \left(f_{kl}^{ij}\right)^2 q_{kl}^2$$

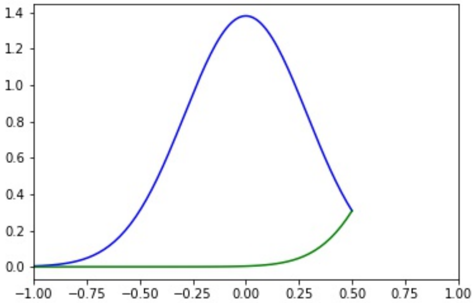
and f_{kl}^{ij} are the discrete cosines
$+$ rounding

$$y_{ij} - [y_{ij}] \sim \mathcal{N}_F(0, s_{ij})$$

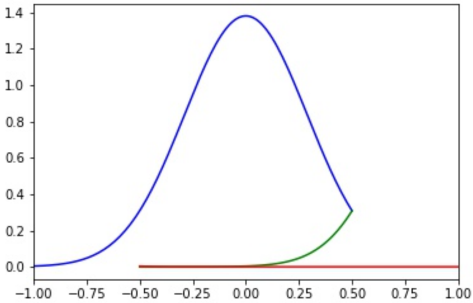
Rounding Error Distribution



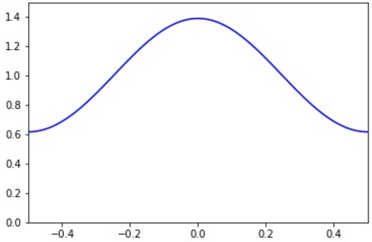
Rounding Error Distribution



Rounding Error Distribution



Folded Gaussian Distribution



Folded Gaussian distribution with variance $1/12$

$$\frac{1}{\sqrt{2\pi s}} \sum_{n \in \mathbb{Z}} e^{-\frac{(x+n)^2}{2s}}$$

- Well approximated by three terms $n \in \{-1, 0, 1\}$

Stego Rounding Errors

- z_{ij} - pixel values of decompressed stego image
- η_{kl} - embedding changes with probabilities of change $\beta^{+}_{kl}, \beta^{-}_{kl}$

$$\begin{aligned} z_{ij} &= \mathrm{DCT}^{-1}_{ij}((\mathbf{c}+\eta)\cdot\mathbf{q}) \\ &= x_{ij} - \mathrm{DCT}^{-1}_{ij}(\mathbf{e}\cdot\mathbf{q}) + \mathrm{DCT}^{-1}_{ij}(\eta\cdot\mathbf{q}) \end{aligned}$$

Stego Rounding Errors

- z_{ij} - pixel values of decompressed stego image
- η_{kl} - embedding changes with probabilities of change $\beta^+_{kl}, \beta^-_{kl}$

$$\begin{aligned} z_{ij} &= \mathrm{DCT}^{-1}_{ij}((\mathbf{c}+\eta)\cdot\mathbf{q}) \\ &= x_{ij} - \mathrm{DCT}^{-1}_{ij}(\mathbf{e}\cdot\mathbf{q}) + \mathrm{DCT}^{-1}_{ij}(\eta\cdot\mathbf{q}) \end{aligned}$$

$$z_{ij} \sim \mathcal{N}(x_{ij}, s_{ij} + r_{ij})$$

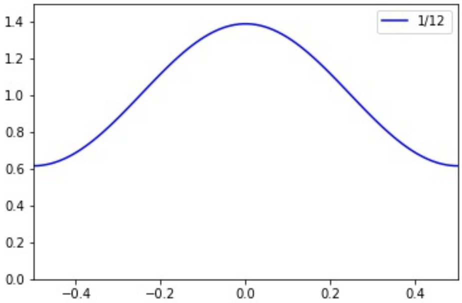
and

$$z_{ij} - [z_{ij}] \sim \mathcal{N}_F(0, s_{ij} + r_{ij}),$$

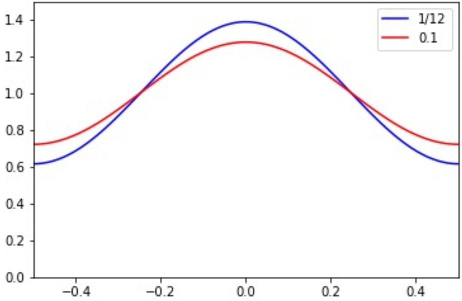
where

$$r_{ij} = \sum_{k,l=0}^{7} \left(f^{ij}_{kl}\right)^2 q^2_{kl}(\beta^+_{kl} + \beta^-_{kl})$$

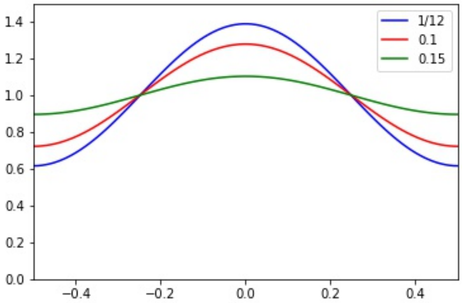
Sensitivity to Variance



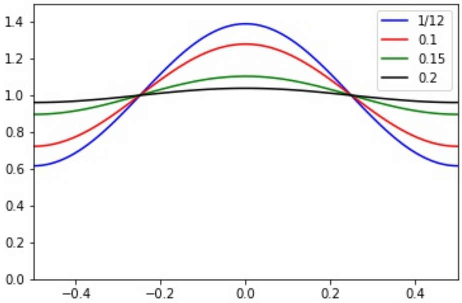
Sensitivity to Variance



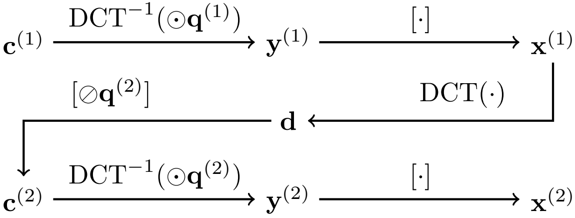
Sensitivity to Variance



Sensitivity to Variance



Double Compression



Why RJCA works in DC images?

- After second compression, the rounding errors in DCT domain are no longer uniform and follow a folded Gaussian
- Variance of rounding errors in the spatial domain does not depend on quantization steps
- Under some conditions, their means are mostly zero, and the folded Gaussian re-emerges in the spatial domain errors again

Double Compression Error

- $c_{kl}^{(1)}$ - DCT coefficients after the first compression
- $q_{kl}^{(1)}, q_{kl}^{(2)}$ - quantization steps for the first and second compression
- e_{kl} - DCT rounding errors during the second compression

$$e_{kl} \sim \mathcal{N}_F \left(c_{kl}^{(1)} \frac{q_{kl}^{(1)}}{q_{kl}^{(2)}}, \frac{1}{12(q_{kl}^{(2)})^2} \right)$$

where

$$\mathbb{E}[e_{kl}] = c_{kl}^{(1)} \frac{q_{kl}^{(1)}}{q_{kl}^{(2)}} - \left[c_{kl}^{(1)} \frac{q_{kl}^{(1)}}{q_{kl}^{(2)}} \right]$$

DC Cover Spatial Rounding Error

- u_{ij} - rounding error after second decompression $(u_{ij} = y_{ij}^{(2)} - [y_{ij}^{(2)}])$

$$u_{ij} \sim \mathcal{N}_F\left(-\sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}], \sum_{k,l=0}^{7} \left(f_{kl}^{ij}\right)^2 (q_{kl}^{(2)})^2 Var[e_{kl}]\right)$$

where

$$\mathbb{E}[u_{ij}] = -\sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}] + \left[\sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}]\right]$$

DC Cover Spatial Rounding Error - Variance

$$u_{ij} \sim \mathcal{N}_F\left(-\sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}], \sum_{k,l=0}^{7} \left(f_{kl}^{ij}\right)^2 (q_{kl}^{(2)})^2 Var[e_{kl}]\right)$$

$$e_{kl} \sim \mathcal{N}_F\left(c_{kl}^{(1)} \frac{q_{kl}^{(1)}}{q_{kl}^{(2)}}, \frac{1}{12(q_{kl}^{(2)})^2}\right)$$

- For $q_{kl}^{(2)} > 1$, variance of the folded Gaussian $\mathcal{N}_F(c_{kl}^{(1)} \frac{q_{kl}^{(1)}}{q_{kl}^{(2)}}, \frac{1}{12(q_{kl}^{(2)})^2})$ is approximately the same as the Gaussian variance

$$Var[e_{kl}] \approx \frac{1}{12(q_{kl}^{(2)})^2}$$

$$Var[u_{ij}] \approx \frac{1}{12}$$

DC Cover Spatial Rounding Error - Mean

$$\mathbb{E}[u_{ij}] = -\sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}] + \left[\sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}] \right]$$

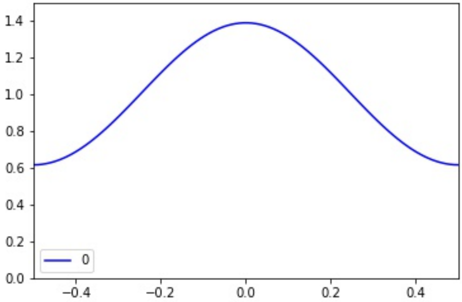
with

$$\mathbb{E}[e_{kl}] = c_{kl}^{(1)} \frac{q_{kl}^{(1)}}{q_{kl}^{(2)}} - \left[c_{kl}^{(1)} \frac{q_{kl}^{(1)}}{q_{kl}^{(2)}} \right]$$

- The means are not known because we do not know the single compressed DCTs
- Non-zero mean would lead to uniform distribution due to mixture (next slide)
- Therefore, we need zero mean for the folded Gaussians

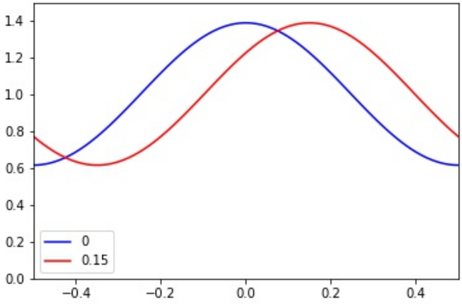
Sensitivity to Mean Shift

- Folded Gaussian with mean shifted from an integer value



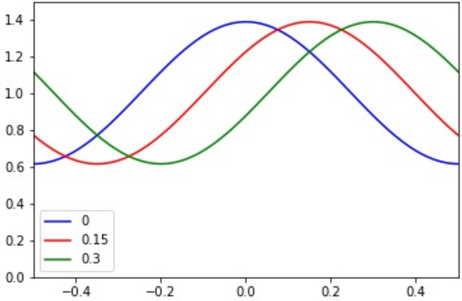
Sensitivity to Mean Shift

- Folded Gaussian with mean shifted from an integer value



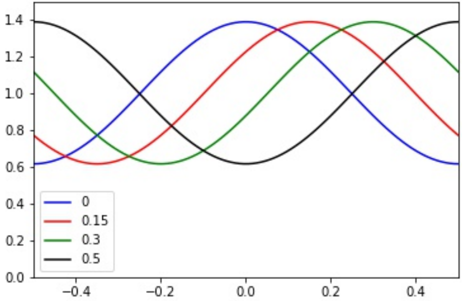
Sensitivity to Mean Shift

- Folded Gaussian with mean shifted from an integer value



Sensitivity to Mean Shift

- Folded Gaussian with mean shifted from an integer value



Requirements on RJCA to work in DC images

- [C1] $q_{kl}^{(2)}$ divides $q_{kl}^{(1)}$ for most modes kl
  - guarantees $\mathbb{E}[u_{ij}] = 0$
  - $QF_2 \geq QF_1$
- [C2] $\mathbf{c}^{(1)} \neq \mathbf{c}^{(2)}$
  - otherwise RJCA (single compressed images)
  - usually $QF_2 \geq 93$ (contains ones in quantization table)

DC Stego Spatial Rounding Error

- u_{ij} - rounding error after second decompression

$$u_{ij} \sim \mathcal{N}_F \left(- \sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}], \sum_{k,l=0}^{7} \left(f_{kl}^{ij} \right)^2 (q_{kl}^{(2)})^2 (Var[e_{kl}] + \beta_{kl}^{+} + \beta_{kl}^{-}) \right)$$

where

$$\mathbb{E}[u_{ij}] = - \sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}] + \left[\sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}] \right]$$

Results on J-UNIWARD, 0.4 bpnzac

Q_1	detector	Q_2 93	94	95	96	97	98	99	100
93	e-SRNet	0.0438	0.3678	0.4104	0.3545	0.2845	0.0317	0.0002	0.0002
	eOH-SRNet	0.0485	0.0059	0.0019	0.0024	0.0035	0.0051	0.0001	0.0001
	JRM	0.4360	0.0029	0.0028	0.0016	0.0010	0.0031	0.0064	0.0053
94	e-SRNet		0.0028	0.3356	0.4205	0.1725	0.0994	0.0001	0.0000
	eOH-SRNet		0.0027	0.0076	0.0030	0.0033	0.0060	0.0002	0.0001
	JRM		0.4304	0.0023	0.0022	0.0019	0.0022	0.0050	0.0068
95	e-SRNet			0.0009	0.3449	0.2870	0.0463	0	0.0001
	eOH-SRNet			0.0008	0.0008	0.0038	0.0038	0.0002	0.0001
	JRM			0.4232	0.0067	0.0024	0.0039	0.0052	0.0067
96	e-SRNet				0.0006	0.3251	0.0412	0.0001	0.0001
	eOH-SRNet				0.0004	0.0118	0.0062	0.0001	0.0002
	JRM				0.4196	0.0079	0.0058	0.0068	0.0086
97	e-SRNet					0.0005	0.2055	0.0001	0.0003
	eOH-SRNet					0.0003	0.0482	0.0002	0.0001
	JRM					0.4159	0.0207	0.0070	0.0061
98	e-SRNet						0.0003	0.0001	0.0001
	eOH-SRNet						0.0001	0.0002	0.0001
	JRM						0.4194	0.0031	0.0041
99	e-SRNet							0	0.0001
	eOH-SRNet							0.0001	0
	JRM							0.4127	0.0026
100	e-SRNet							0.0002	0.0001
	eOH-SRNet							0.0001	0.0001
	JRM							0.4126	0.3965

Conclusions

- RJCA exteded from single compressed images to double compressed images
  - DCT error is uniform only after the first compression
- Works extremely well for
  - $QF_2 \geq 99$ - parallel to RJCA
  - $QF_1 = QF_2 \geq 93$ - unlike rich model detectors
- For other cases, the DC image exhibits strong DC artifacts, and standard steganalysis tools, such as JRM, become very accurate