Abstract

Data hiding such as steganography and invisible watermarking has important applications in copyright protection, privacy-preserved communication and content provenance. Existing works often fail short in either preserving image quality, or robustness against perturbations or are too complex to train. We propose RoSteALS, a practical steganography technique leveraging frozen pretrained autoencoders to free the payload embedding from learning the distribution of cover images. RoSteALS has a light-weight secret encoder of just 300k parameters, is easy to train, has perfect secret recovery performance and comparable image quality on three benchmarks. Additionally, RoSteALS can be adapted for novel cover-less steganography applications in which the cover image can be sampled from noise or condition on text prompts via a denoising diffusion process.

Method, Dataset, and Results

MIRFlickR

We train to compute the bit recovery loss between the predicted and the ground truth secret. We use Resnet50 [2] as the secret decoder D, replacing the last network consisting of a fully connected layer followed by δ)

Shown above is the architecture diagram of our watermarking method – RoSteALS. The image encoder (E) and decoder (G) are locked during training, only updating the lightweight secret encoder (F) and decoder (D). F is a very small network consisting of a fully-connected layer followed by SILU[1]. The output acts as a small offset to be added to the cover embedding z. The stego image is then constructed as x = G(z + \text{conditioned on text prompts via a denoising diffusion process.})

Shown above is the architecture diagram of our watermarking method – RoSteALS. The image encoder (E) and decoder (G) are locked during training, only updating the lightweight secret encoder (F) and decoder (D). F is a very small network consisting of a fully-connected layer followed by SILU[1]. The output acts as a small offset to be added to the cover embedding z. The stego image is then constructed as x = G(z + \text{conditioned on text prompts via a denoising diffusion process.})

Method: Measure

<table>
<thead>
<tr>
<th>Method</th>
<th>LPIPS</th>
<th>SSIM</th>
<th>Accuracy Clean</th>
<th>Accuracy Noise</th>
<th>Accuracy Noise (ECC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jigsaw[8]</td>
<td>0.01</td>
<td>0.99</td>
<td>1.00</td>
<td>0.55</td>
<td>0.13</td>
</tr>
<tr>
<td>OutGuess[7]</td>
<td>0.01</td>
<td>0.99</td>
<td>0.92</td>
<td>0.47</td>
<td>0.10</td>
</tr>
<tr>
<td>RRC [10]</td>
<td>0.01</td>
<td>0.98</td>
<td>0.99</td>
<td>0.63</td>
<td>0.17</td>
</tr>
<tr>
<td>SSL [11]</td>
<td>0.02</td>
<td>0.96</td>
<td>0.99</td>
<td>0.87</td>
<td>0.13</td>
</tr>
<tr>
<td>RoSteALS</td>
<td>0.04</td>
<td>0.91</td>
<td>1.00</td>
<td>0.92</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Shown is the performance of our method in comparison to the state-of-the-art watermarking methods. For stego recovery, we report SSIM and LPIPS. For secret recovery, we report standard bit accuracy using clean stego, noisy stego, and noisy stego after ECC (cyclic error correction using BCH [12]). Even though there is not a clear winner in stego image quality, RoSteALS achieves the best secret retrieval performance both before and after stego image corruption, resulting in perfect performance in both clean and noise data (with ECC).

Secret-Length

Change in image quality as secret length increases. Residual images are scaled to [0, 255] range for visualization purpose. The average image quality (SSIM) for 50, 100, 150, 200-bits secret length is 0.89, 0.88, 0.88, 0.88, while the bit-accuracy in noisy stego is 0.97, 0.94, 0.87, and 0.84.

Strengths

• better robustness and image quality
• faster training and inference time
• adatable to secrets of different lengths (up to 200 without loss of image quality)

Limitations

• image quality limited by the performance of pretrained autoencoder
• struggles in reconstructing small text or face
• struggles in reconstructing images with cluttered objects

References


Acknowledgements

This work was supported in part by ERCCoE under EPSRC grant EP/T022480/1.