MULTI-STAGE RESIDUAL HIDING FOR IMAGE-INTO-AUDIO STEGANOGRAPHY

Wenxue Cui*, Shaohui Liu*, Feng Jiang*, Yongliang Liu^, Debin Zhao*

*** Harbin Institute of Technology, Pengcheng Laboratory ^ Alibaba Group**

INTRODUCTION

Steganography has been researched extensively, which is a technique of concealing secret messages in digital carriers to facilitate covert communication through exploiting the redundancy of human perceptions. The widespread application of audio communication technologies has speeded up audio data flowing across the Internet, which made it an popular carrier for covert communication. In this paper, we present a cross-modal steganography method for hiding image content into audio carriers while preserving the perceptual fidelity of the cover audio.

PROBLEMS

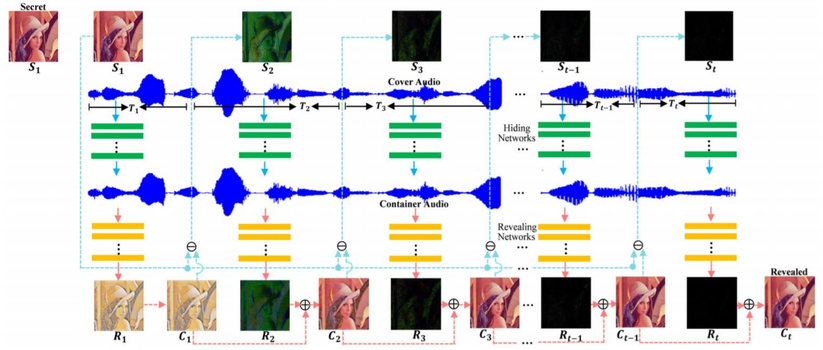
- With the diversification of data, cross-modal steganography becomes more and more important. However, only a few works are devoted to this field.
- Audio is the sequential signal, while image is the non-sequential signal, therefore it is tactical to hide an image into an audio.
- Hiding the secret message directly is difficult because of its diversity of knowledge and therefore usually leads to noticeable artifacts.

CONTRIBUTIONS

- We propose a novel image-to-audio steganography framework based on deep learning, which achieves superior hiding capabilities against other methods.
- By hiding the residual errors of many levels, the proposed method not only can control the payload capacity more flexibly, but also make the hiding process more easier.
- Our framework embed the residual errors into different audio subsequences, which implies that even if part of the carrier is lost, the secret image can be restored to some extent.

PROPOSED METHOD

★ Proposed Framework



★ Structure Description

In this paper, we present a cross-modal steganography method for hiding image content into audio carriers while preserving the perceptual fidelity of the cover audio. In our framework, two multi-stage networks are designed: the first network encodes the decreasing multilevel residual errors inside different audio subsequences with the corresponding stage sub-networks, while the second network decodes the residual errors from the modified carrier with the corresponding stage sub-networks to produce the final revealed results.

★ Loss Function

Total Loss: $\mathcal{L}(\theta_{H_i}, \theta_{R_i}) = \sum_{i=1}^{t} \mathcal{L}_{H_i}(\theta_{H_i}) + \lambda_i \mathcal{L}_{R_i}(\theta_{R_i})$

Hiding Loss: $\mathcal{L}_{H_i}(\theta_{H_i}) = \frac{1}{N} \sum_{i=1}^{N} ||\mathcal{H}(S_i, T_i; \theta_{H_i}) - T_i||_2^2$

Reveal Loss: $\mathcal{L}_{R_i}(\theta_{R_i}) = \frac{1}{N} \sum_{i=1}^{N} ||\mathcal{R}(\tilde{T}_i; \theta_{R_i}) - S_i||_2^2$

For one stage

T_i audio subsequence S_i secret image $\mathcal{H}$ hiding operation
$\mathcal{R}$ revealing operation $\theta_{H_i}\theta_{R_i}$ parameters of hiding and revealing networks

EXPERIMENTAL RESULTS

Table 1. The performances for different deep learning-based steganography algorithms. **Bold** indicates the best performance.

Datasets	Natural image → TIMIT				Face image → TIMIT			
Method	Container	Revealed			Container	Revealed		
	MSE	PSNR	SSIM	MS-SSIM	MSE	PSNR	SSIM	MS-SSIM
Deep-Steg [1]	2.7669E-3	31.40	0.8624	0.9576	2.2258E-3	31.44	0.8196	0.9380
Kreuk'Model [16]	1.3495E-3	33.68	0.8893	0.9671	**9.8872E-4**	33.59	0.8420	0.9574
DITAS-S	1.2144E-3	27.84	0.8691	0.9685	1.1033E-3	28.05	0.7682	0.9340
DITAS-M	1.9249E-3	37.09	0.9444	0.9925	2.1313E-3	36.12	0.8898	0.9807
DITAS-M-E	**9.3864E-4**	37.84	0.9482	0.9934	9.9386E-4	36.52	0.8906	0.9847
DITAS-M-D	1.6904E-3	**38.42**	0.9581	0.9935	1.7932E-3	37.65	0.9036	0.9883
DITAS-M-ED	9.8471E-4	38.39	**0.9597**	**0.9944**	1.0582E-3	**37.69**	**0.9054**	**0.9900**

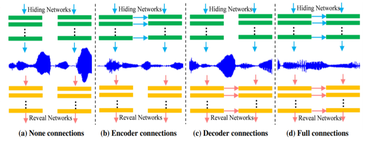


Fig. 1. The structural details of four experimental variants

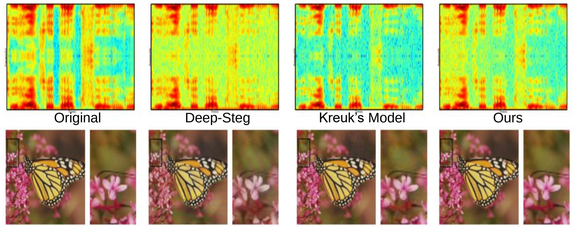


Fig. 2. The visual comparisons of different steganography methods. (Top is the visual comparisons of container in the frequency domain and bottom is the revealed perceptions.)

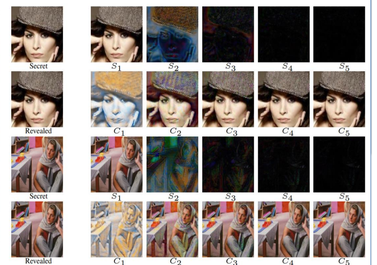


Fig. 3. The intermediate visual outputs of our framework

CONCLUSION

In this paper, we propose a novel cross-modal image-to-audio steganography framework based on deep learning. Instead of hiding the secret image directly, the proposed method embeds the residual errors of secret image into the cover audio progressively by a multi-stage fashion. In the hiding process of the proposed method, residual errors become more sparse with the increase of stages, which not only make the controlling of payload capacity more flexible, but also make hiding easier because of the sparsity characteristic of residual errors.

ACKNOWLEGEMENTS

This work was supported by Alibaba Group through Alibaba Innovative Research (AIR) Program.