Deep Neural Networks based Invisible Steganography for Audio-into-Image Algorithm

Sun Asterisk R&D Lab

Quang Pham Huu, Thoi Hoang Dinh, Ngoc N. Tran,

Toan Pham Van, Thanh Ta Minh

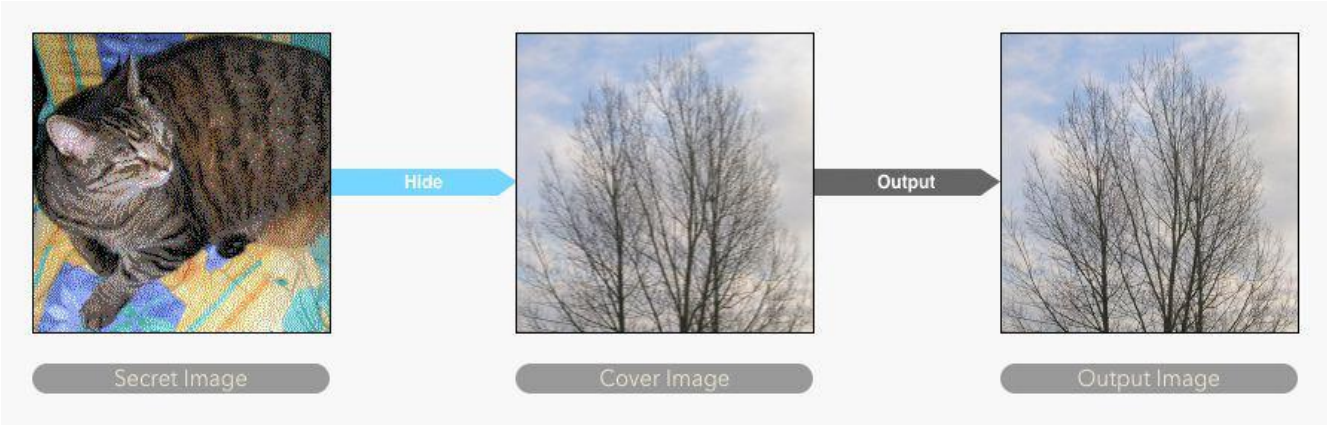
Content

- Introduction to steganography
- Audio-into-Image Algorithm and our proposed method
- The dataset and metrics evaluation
- Experiment results
- Conclusion

Introduction to steganography

What is steganography?

- Steganography is a main problem in cyber security field. This is the process of hiding some type of data into other data.
- It can use to transmit secret information to people who already are aware of its existence, while others still see it as normal data.
- For example, we can hide an image inside another image, a video inside another video or an audio in anther audio.



Problems with Existing Methods

- One of the most common method is to hide the secret information directly in Least Significant Bits (LSB) of the cover image.
- The methods are simple, and thus, they can be easily detected.
- They are easy to decode, as the way information is encoded, the algorithm is fixed and it can be guessed by using some statistic method.
- The drawbacks of LSB is that the number of bits that can be hidden depends on the number of LSB bits that could change for each pixel. therefore, the amount of information that can be hidden is generally less.
- There have been researches hiding image-into-image and video-into-video, but there is no method to effectively hide audio-into-images due to the limited capacity of images

Advantage of Deep learning approach

- Using some neural network architecture to do steganography task. To hide and recover the information, the neural network using all of the pixels of image, waveform sample of audio with some non-linear transform inside.
- Its process are difficult to guess.
- The amount of information that can be hidden is more.

Audio-into-Image Algorithm and our proposed method

Our approach

Two methods of audio data pre-processing are conducted.

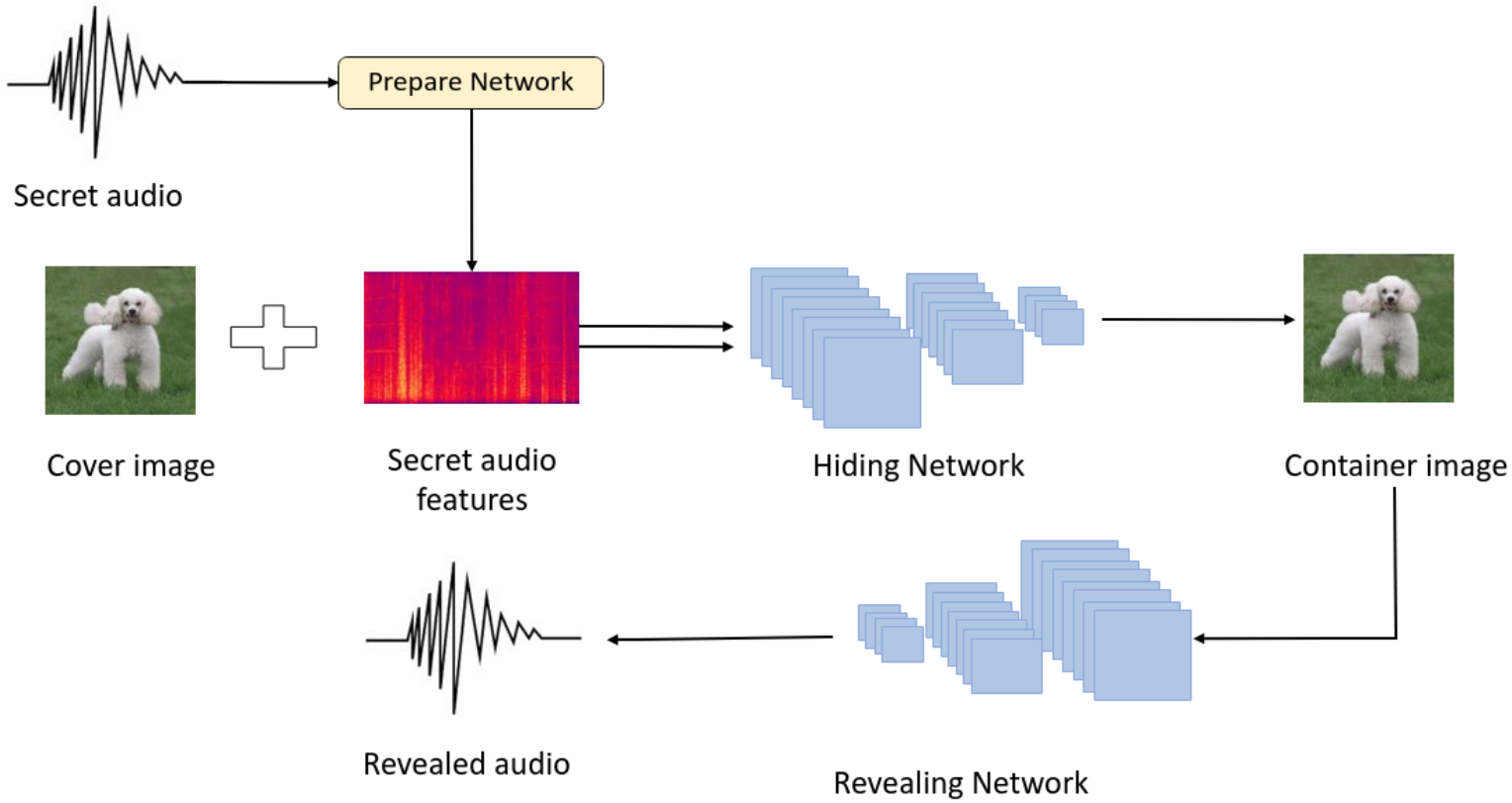
- Method-1: Raw audio data is normalized and reshaped to a new 3D matrix of the same shape as three color channels image.
- Method-2: Short-time Fourier transform (STFT) is applied to transform audio to the frequency domain. STFT is a sequence of Fourier transforms of a windowed signal. STFT provides the time-localized frequency information for situations in which frequency components of a signal vary over time. The transformed data is now a 3D matrix with two channel instead of three channels as the first method

Our approach: Model Architecture

We build a model architecture that includes:

- Encoding submodel:
  - Prepare network: appropriately extracts the characteristics of the audio before it is concatenated with the image.
  - Hiding network: hides the secret audio features into the cover image to generate the container image.
- Decoding submodel:
  - Reveal network: decodes the container image to obtain the original audio.

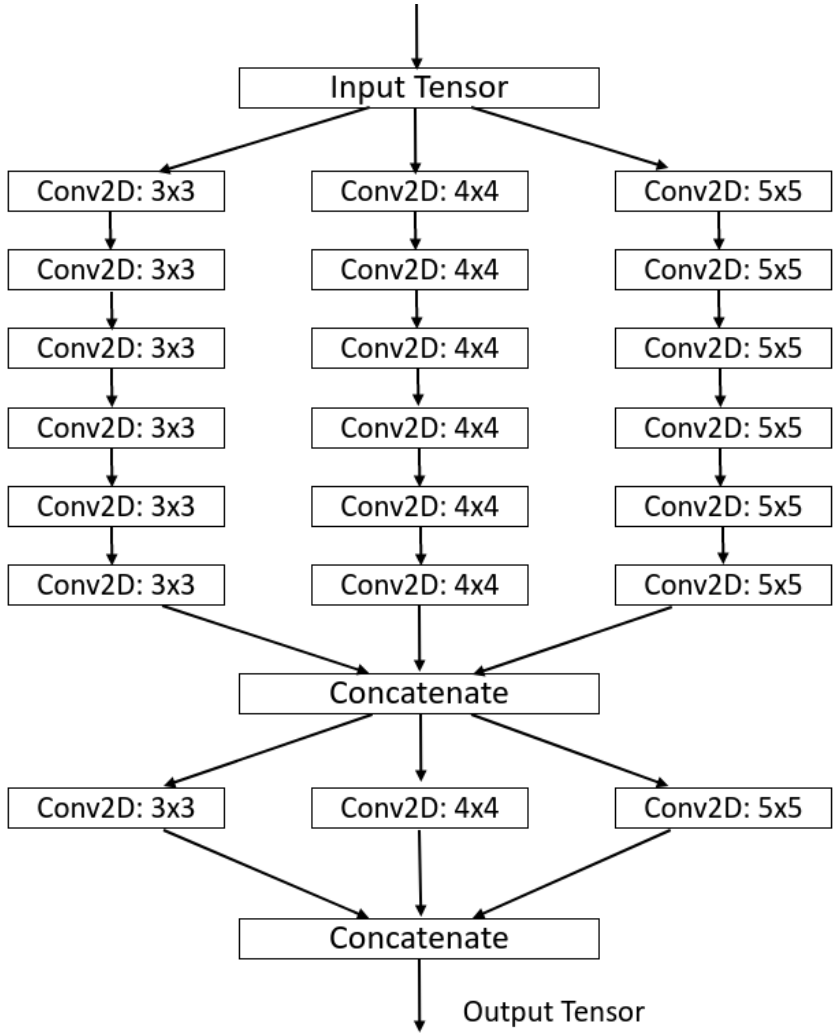
Our architecture



$$L_{total} = \frac{\alpha MSE_{hiding\_net} + \beta MSE_{reveal\_net}}{\alpha + \beta}$$

Our architecture

- A block is an architecture of 5 convolutions stacked together with the same kernel size (kernel sizes of 3×3, 4×4, and 5×5).
- The output of three block are concatenated and passed through a similar architecture but shallower, with only one convolution layer.
- Finally, the output are concatenated and goes through another convolution layer to get the expected output size.



The dataset and metrics evaluation

Dataset

- Vivos is a public dataset that includes 12,420 audio of Vietnamese voices with the sample rate of 16kHz. Duration of the audio ranges from 1−18 seconds.
- The image dataset consists of a large number of photos from Kaggle competitions including the Flower, Fruits, Dogs and Cats, and Stanford Dogs datasets. In total, the collected image dataset contains up to 24,000 images.
- 80%, 10%, 10% of the dataset are used in the training, validating, and testing phase, respectively.

Evaluation metrics

- To evaluate the integrity of the cover image before and after secret audio is hidden, we used the sum of squares error (SSE) metric on 1000 images.

$$\mathrm{SSE} = \sum_{1}^{1000} \mathrm{MSE}_{per\_pixel,\ per\_channel}$$

$$\mathrm{MSE}_{per\_pixel,\ per\_channel} = \frac{1}{3 \times H \times W} \sum (y - \hat{y})^2$$

Evaluation metrics

- For audio, we use the Pearson correlation coefficient (PCC) to evaluate the integrity of information after the process of hiding and revealing.
- The linear relationship between two data distribution

$$\mathrm{r} = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

- The value of correlations value range from -1 and +1 with 0 implying no correlation. Two audio is considered to be completely the same when the correlation coefficient is 1 or 100%.

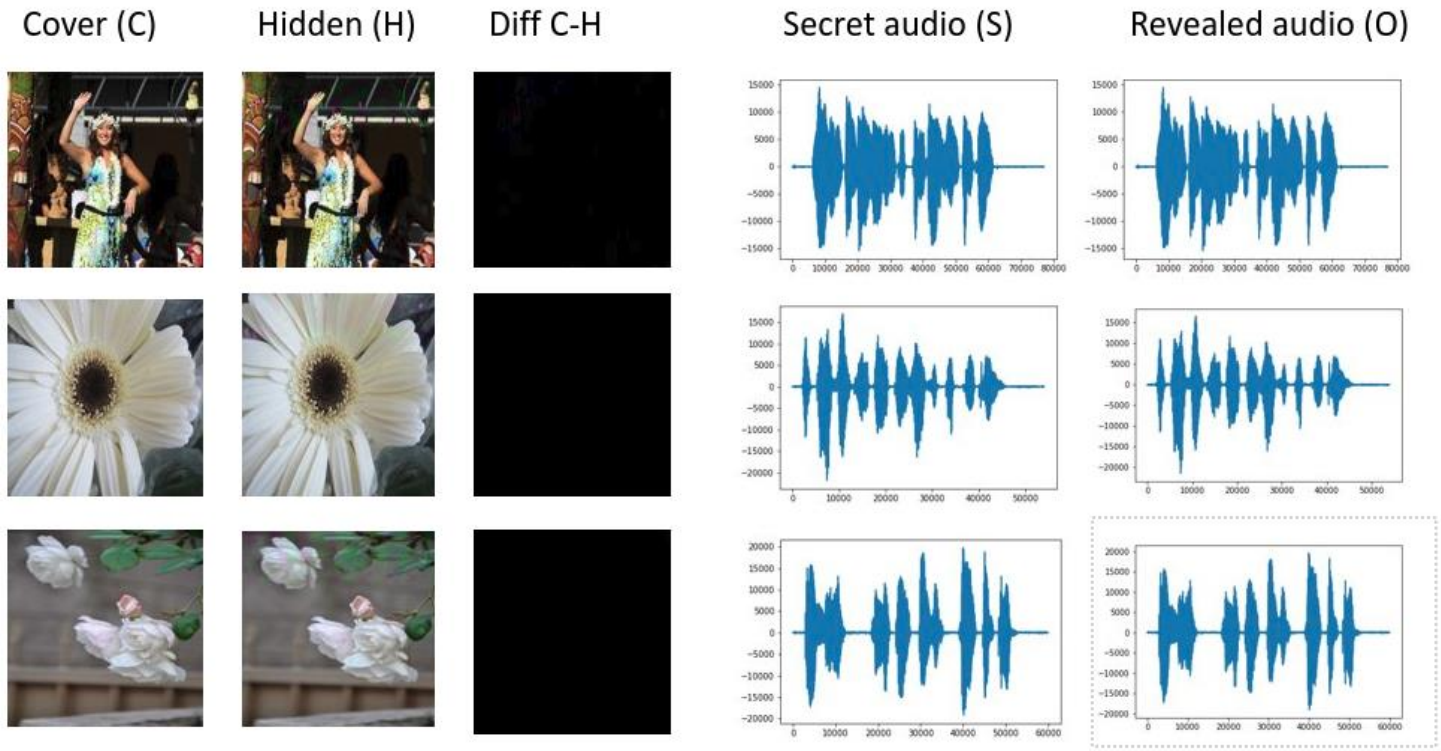
Experiment results

RESULTS

TABLE I
SSE (smaller-better) and correlation coefficient (greater-better) with raw and DFT pre-processing technique.

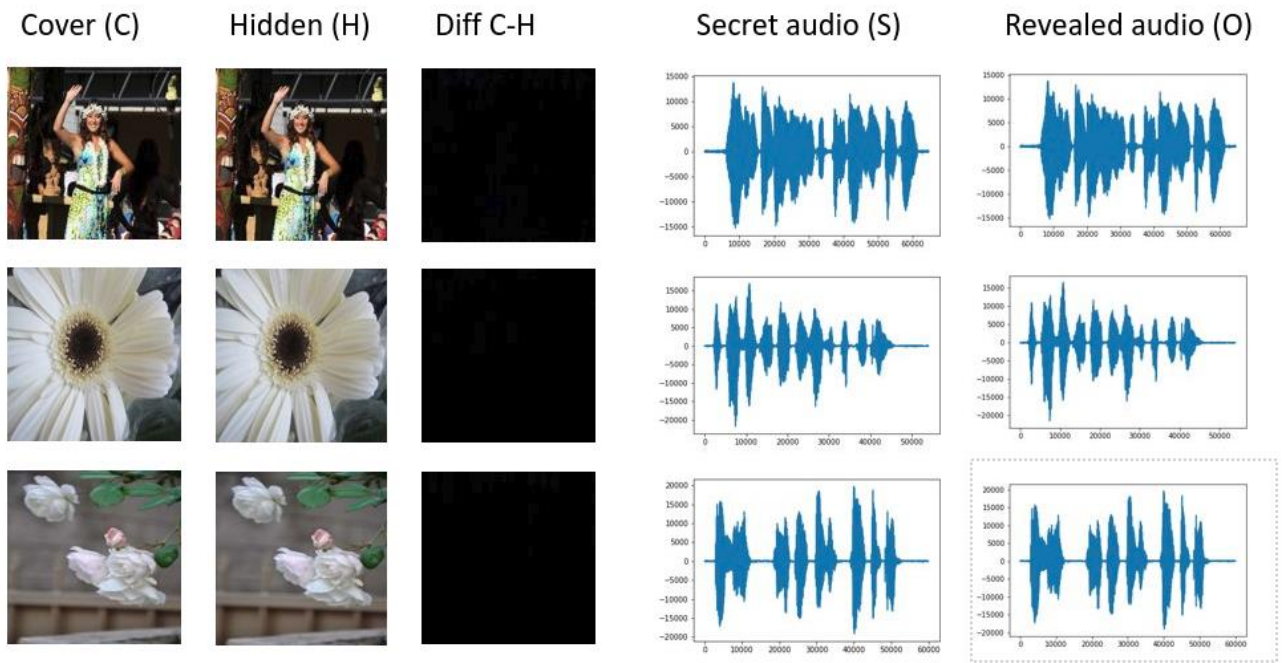
Pre-process	α/β	SSE	Average of correlation
Method-1: Raw	15.0 / 1.0	0.3744	99.83%
	10.0 / 1.0	2.0379	99.30%
	1.0 / 1.0	2.4072	99.84%
	0.5 / 1.0	10.7196	99.74%
	0.1 / 1.0	14.8777	99.90%
Method-2: DFT	1.0 / 0.1	0.0192	99.43%
	1.0 / 0.5	0.0135	99.91%
	1.0 / 1.0	0.0334	99.85%
	1.0 / 2.0	0.0539	99.86%
	1.0 / 10.0	0.1393	99.93%

RESULTS: STFT



- With DFT pre-processing method, a 4-second-long audio can be hidden into a 255×255 image.

RESULTS: RAW



- With the raw audio data, the length of audio data can be up to 12 seconds.

Conclusion

- We have proposed the use of deep learning technique in audio-into-image steganography.
- Two different audio processing methods and neural network architecture modification are introduced to effectively hide secret audio inside image.
- The proposed method is better than traditional methods in term of the length of hidden audio and integrity of both image and audio data.

THANK YOU FOR ATTENTION