EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Steganographic visual story with mutual-perceived joint attention

Yanyang Guo[1], Hanzhou Wu[1,2*] and Xinpeng Zhang[1,2]

*Correspondence:
wuhanzhou_2007@126.com
[1] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, People's Republic of China
[2] Shanghai Institute for Advanced Communication and Data Science, Shanghai 200444, People's Republic of China

Abstract

Social media plays an increasingly important role in providing information and social support to users. Due to the easy dissemination of content, as well as difficulty to track on the social network, we are motivated to study the way of concealing sensitive messages in this channel with high confidentiality. In this paper, we design a steganographic visual stories generation model that enables users to automatically post stego status on social media without any direct user intervention and use the mutual-perceived joint attention (MPJA) to maintain the imperceptibility of stego text. We demonstrate our approach on the visual storytelling (VIST) dataset and show that it yields high-quality steganographic texts. Since the proposed work realizes steganography by auto-generating visual story using deep learning, it enables us to move steganography to the real-world online social networks with intelligent steganographic bots.

Keywords: Steganography, Social networks, Recurrent neural network, Visual story

1 Introduction

Steganography aims to hide the existence of secret information. It can hide the secret messages in videos, images, texts, and so on. For instance, imagine a scenario where two users want to exchange prohibited ideas or secret information under monitoring party; it is easy to suspect both sides of the communication in a world where most communication takes place in a transparent environment.

One of the traditional ways of transmitting secret messages is to publish seemingly normal news or advertisements in newspapers. But only a real spy with the right key can decode the news. This manual method has been replaced by algorithmic method. But the development of natural language processing technology and social network makes it possible to use this traditional method again. With this motivation in mind, we hope to design a system that enables two users to exchange encrypted messages openly and transparently on the social network platform by posting status updates.

Extensive researches [1–4] have been carried out for image processing or image steganography. Furthermore, more and more text based information hiding methods [5–8] have appealed to a tremendous proportion of researchers' interests in recent years.

Springer Open

Fang et al. design a text information hiding method by dividing the dictionary containing all words in advance and then encoding the words in a fixed-length coding way [7]. Yang et al. present a steganography framework that embeds secret information into text by constructing a Huffman tree based on the probability distribution of words [8].

However, in real-scenarios, information representations on social media usually contain multiple modal contents. Due to the trend described above, some works [9, 10] have explored cross-modal steganography tasks. In [10], they use word-by-word hiding method with fixed length of secret bits. Different from [10], the secret data are embedded at the sentence level. An improved steganographic scheme SSH based on beam search is proposed in [9]. Our approach also uses the word-by-word hiding method, but we embed secret data with variable length in each word. This makes the embedding process more efficient in terms of text quality. In addition, previous works are focusing on generating steganographic image description based solely on image caption. One of the main problems of image caption task [11] is that it can only recognize the event of the image simply and mechanically and cannot tell the stories of photos in the user's voice and share with others by posting them to social networks.

Toward filling this gap, we propose the task of steganographic visual story (SVS) automatic posting, which aims to generate steganographic visual stories from selected photos in local or online albums. Huang et al. [12] proposed the task of visual storytelling and constructed VIST dataset. And there are several works such as [13, 14] based on it. In order to overcome the problem of stego text deviating from the image themes, we propose mutual-perceived joint attention (MPJA) to generate the text-aware visual representation and the vision-aware textual representation, so that the generated steganographic stories are more natural and more readable. On the basis of fully understanding the image content by neural network, the story words with secret messages can be generated with natural human custom according to MPJA. Our method does not need to modify the images, and there is no comparable original text, so it can be more difficult to be detected than traditional methods.

The rest of the paper is organized as follows. Section 2 describes the structure of steganographic visual story generation model with MPJA and its adaptive information hiding and extracting algorithm. In Section 3, we report the results of experiments with our proposed method when generating with both MPJA and adaptive data embedding and contrast their performance to previous art. Finally, the conclusion remarks of this paper are given in Section 4.

2 Proposed method

The whole process of posting secret information on social network and secret extracting is shown in Fig. 1. In the embedding process, shown on the left side of Fig. 1, we use the method of photo album clustering [15] to select images suitable for uploading from local albums or online albums and automatically exclude some low-quality photos [16], such as blurred ones. Then, we choose a certain number of photos from the clustered albums in a timed sequence and put them into the pretrained steganographic visual story generation model (see Fig. 2). The sentence stories generated by each photo are linked together as a post for uploading pictures on social networks and finally posted on various social network sites, such as Facebook and Twitter. The proposed work uses text as the

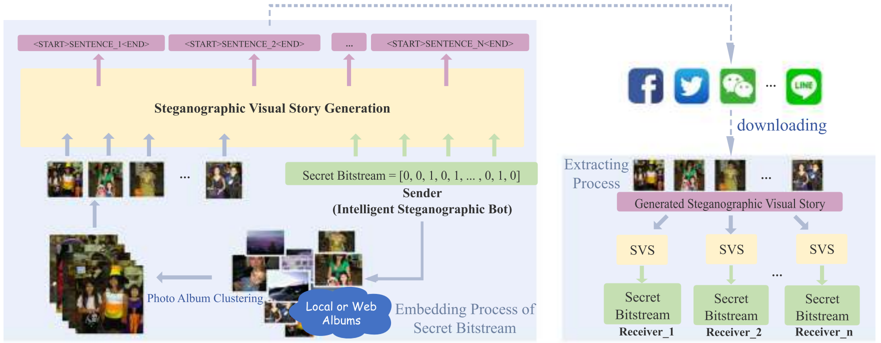


Fig. 1 The process of posting secret information on social network

cover, and the images are used to train a language model to generate the stego text. The goal of using the images is to make sure that the stego text and the images have the same semantic information so that the stego text will not arouse suspicion when the stego text and the images are posted by the data hider.

We can see the extraction process on the right side. All the receivers only need to have permission to access the photos and text posted by the sender. So, there is no need for direct contact between senders and receivers; the sender can even be a robot. All the data receivers receive the same images and text description. Therefore, they are actually trying to recover the same secret message. As long as they hold the neural network model and can download the media files, they are all able to reconstruct the embedded information. It is worth mentioning that the data hider and the data receiver should share the image order before feeding the images into the neural network, which can be controlled by a secret key. This indicates that, though different secret keys correspond to different orders of the images, once they are used for data hiding, the order should be fixed and shared between the data hider and data receiver.

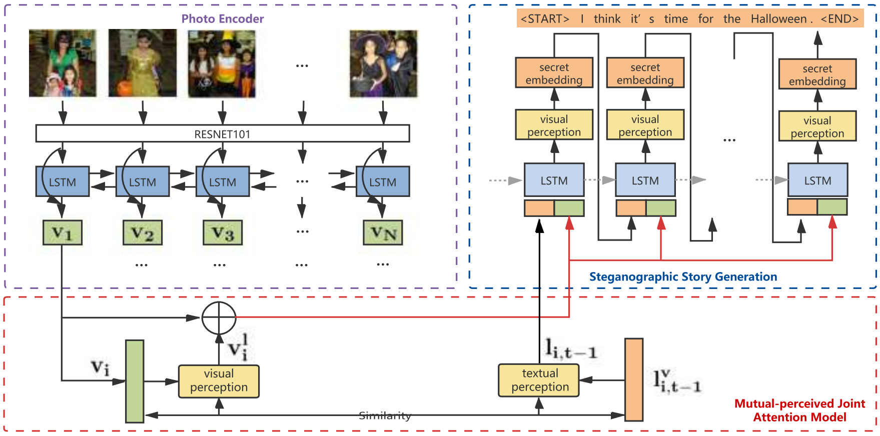


Fig. 2 Structure of steganographic visual story generation model with mutual-perceived joint attention

2.1 Photo encoder

As shown in Fig. 2, our encoder module is composed of two separate encoders, one that models the content of the image sequence and the other one that models the relationship between input images.

For modeling the content of the image sequence, we used the extracted feature vectors from the ResNet [17] to describe the images. We chose ResNet over other convolutional neural networks because we consider the balance of computationally expensive and precision. Every image needs to be resized to 256 ×256 with respect to its ratio. In addition, we crop the image (if needed) from the center region to fit the ResNet input layer because we assume that the important information in the image is placed in the center.

In order to accurately use a few words to simulate the user's thoughts and feelings about the uploaded photos, we should consider the visual information of the photos themselves. While seeing the first image, we start the story with a sentence that describes and estimates the context of the particular image. For the next image in the sequence, we not only analyze the current image but also consider the influence of the previous image and the latter image, because this is only way to preserve the temporal correlation between events in the image, so that the text we generate is more in line with the logic of human narrative. It is a logical process to organize the content expressed in the pictures.

To achieve this, it is important to keep the temporal dependence between the sentence story generated by the current image in the sequence and the sentence story generated by the before and after images. Recurrent neural network [18] has made great success in processing sequence data, because it can learn the potential dependencies between sequence data elements, and it also has been proved to be suitable for modeling image features vector sequences. And we experimented with different types of recurrent neural network [18] to model the relationship between all input images and found that we could achieve better story flow by the use of bidirectional long short-term memory networks [19, 20](Bi-LSTM) to obtain the context information from input images. In addition, we also apply the idea of concatenated coding for better aggregated representations. Our initial visual representation $\mathbf{v_i}$ is concatenation of image features and the output of Bi-LSTM.

2.2 Mutual-perceived joint attention

Soft attention mechanism [21] utilizes additional weights on the interrelated outputs of the nodes, which improves the performance of the basic encoder-decoder model in machine translation. In the task of story generation from image sequences, however, each sentence should be visually grounded on not only each image but also overall context. To represent the relationship between input images and generated text, we design a scheme based on attention mechanism, called mutual-perceived joint attention. We implement them via calculating the similarity matrix that focuses sequentially on the images and generated words when generating story-like sentences.

We use $\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_N}\}$ to express the visual representation of input photo sequence, where $\mathbf{v_i} \in \mathbb{R}^{1\times k}$ is a one-dimension feature vector generated from our photo encoder, $\mathbf{V} \in \mathbb{R}^{N\times k}$ contains N visual representations of single photo. Each $\mathbf{v_i}$ $(1 \leq i \leq N)$, as a visual representation of the ith photo, is then used to decode the ith story sentence respectively. We use the mutual-perceived joint attention (MPJA) mechanism to capture the internal relationship between visual content and textual content and show the part of their mutual perception. MPJA can help us to focus on which part of image can control the text generation and which word better corresponds to the characteristics of given image.

To have awareness of each other, mutual perception of image and generated text can be measured by calculating similarity matrix $\mathbf{D}$. The similarity matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ is computed as:

$$\mathbf{D} = \text{softmax}\left(\mathbf{VWL}^{T}\right), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{k\times k}$ is a learnable weight matrix, $\mathbf{V} \in \mathbb{R}^{N\times k}$ is the visual representation through photo encoder, $\mathbf{L} \in \mathbb{R}^{N\times k}$ is the textual representation through generated N word embeddings, k is the dimension of the embedding of words, N is the number of visual representations, and T is transpose operation of matrix. It is worth noting that we normalize the similarity weights via softmax normalization, which tends to help the model to focus on the most relevant concepts.

The task-relevant part is added to the original visual representation; hence, we can obtain the new visual representations after textual perception $\mathbf{V}^{l}$:

$$\mathbf{V}^{l} = \left(\mathbf{I} + \mathbf{D}^{T}\right)\mathbf{V}. \tag{2}$$

Similarly, the new textual representations $\mathbf{L}^{v}$ after visual perception can be obtained by:

$$\mathbf{L}^{v} = \left(\mathbf{I} + \mathbf{D}\right)\mathbf{L}. \tag{3}$$

That provides the generation model with more necessary and useful information from the new visual and textual representations after MPJA.

2.3 Steganographic story generation

The steganographic story generation module aims to generate a reasonable and coherent story with hidden secret messages based sentence-level decoding. Figure 2 visually shows its decoding process. Specifically, when the decoder is generating the ith sentence, the source information includes two parts: the text-aware visual representation $\mathbf{v_i^l}$ of the ith image, and the vision-aware textual representation $\mathbf{l_{i,t-1}^v}$ of the previously generated word in ith steganographic sentence. Our sequence decoder also employs a unidirectional LSTM layer [19]. Meanwhile, the unidirectional LSTM based models outperform the Bi-LSTM [20] based models in decoder. It does not illustrate the unidirectional LSTM is better than the Bi-LSTM in story generation. It only indicates that in the current experimental settings, the unidirectional LSTM-based model outperforms the bidirectional one. Apart from the general state update, the tth hidden state $\mathbf{s_{i,t}}$ is further designed to take the two representations of mutual perception into consideration:

$$\mathbf{s_{i,t}} = \text{LSTM}\left(\mathbf{s_{i,t-1}}, \mathbf{v_i^l} \oplus \mathbf{l_{i,t-1}^v}\right) \tag{4}$$

where $\oplus$ denotes the vector concatenation which has the same meaning in Fig. 2 and it allows the decoder to pay different attention to different parts of the generated text. We refer to the previous works [22], adding a softmax classifier to the output layer to calculate the possible probability of each word to facilitate the embedding of secret information.

2.4 Adaptive information hiding

Information hiding and extraction are two completely opposite operations. The process of information hiding and extraction is basically the same. It is also necessary to use the same RNN network to calculate the conditional probability distribution of each word at each moment and then construct the same candidate list and use the same coding method to encode the words.

Algorithm 1 Adaptive Information Hiding Algorithm

```
Input: A sequence of photos $I$ selected from local albums; Secret bitstream: $\mathbf{b}$ = $\{0, 0, 1, 0, 1, ..., 0, 1, 0\}$; Initial size of selected bits: $M$; Threshold : $T$;
Output: Generated steganography text: Text;
Generate a pretrained SVS model based on $I$;
Text is a empty list;
while b is not empty do
    Use the SVS model to produce a word list based on $I$ and previous words, each element of this list is a tuple consisting of a word and its prediction probability;
    Sort the word list by prediction probability in a descending order;
    for $l = M, M-1, \ldots, 0$ do
        Select top $2^l$ sorted words and their probabilities to form candidate list $\mathbf{l_c}$
        Calculate the variance $s^2$ based on $\mathbf{l_c}$, that is $s^2 = \frac{1}{M-1}\sum_{i=1}^{M}(x_i - \bar{x})^2$, where $x_i$ is the probability of the $i$th word in $\mathbf{l_c}$;
        if $s^2 < T$ then
            Find the word $w$ in $\mathbf{l_c}$ as the steganographic word corresponding to the current secret bitstream $\mathbf{b_0}, \mathbf{b_1}, \ldots, \mathbf{b_{l-1}}$;
            Append $w$ to Text;
            Update $\mathbf{b}$ by removing top $l$ bits from $\mathbf{b}$;
            break;
return Text;
```

Different from the traditional method, which embeds fixed-length secret bits in each word, we propose an adaptive information hiding algorithm. The process of embedding of fixed-length secret bits in each word is relatively simple, but not all the words in sentences are suitable for embedding the same number of secret bits. How to select embedding strategy which affect text quality at least becomes an important technique problem, because the variance of probability distribution of candidate list is carried out to reveal the dispersion of selection probability. We think that if the variance of probability distribution of candidate list with length M is less than a certain threshold T, no matter what kind of secret data is embedded, it will not cause too much deviation to the semantics of the whole sentence. The initial embedding length M will be initialized when we have selected the final length of secret bits at current word after some loops and start preparing the next one. We determine the threshold T as constant. If we need to embed more bits of secret information, we can set a larger threshold value. But it may affect the text quality instead. Similarly, we can adjust the content of generated sentences to make it more in line with the content of input images through lowering this value, but with lower embedding rate. Hence, with a variable capacity of secret bits per word, we can embed the maximum number of secret bits word while ensuring the quality of steganographic text.

Algorithm 2 Adaptive Information Extracting Algorithm

```
Input: A sequence of photos $I$ downloaded from specified social network; Posted
    steganography Story: Text; Pretrained SVS model shared by sender; Initial size of
    selected bits: $M$; Threshold: $T$;
Output: Secret bitstream: $\mathbf{b} = \{0, 0, 1, 0, 1, ..., 0, 1, 0\}$;
    Generate a pretrained SVS model based on $I$;
    b is a empty list;
    for each word $w$ in Text do
        Use the SVS model to produce a word list based on $I$ and previous words, each
    element of this list is a tuple consisting of a word and its prediction probability;
        Sort the word list by prediction probability in a descending order;
        for $l = M, M - 1, \ldots, 0$ do
            Select top $2^l$ sorted words and their probabilities to form candidate list $\mathbf{l_c}$;
            Calculate variance $s^2$ based on $\mathbf{l_c}$;
            if $s^2 < T$ then
                Extract the decimal number corresponding to the location of $w$ in $\mathbf{l_c}$;
                Convert $w$ to $l$ binary bits;
                Append the $l$ bits to $\mathbf{l_c}$;
                break;
    return Extracted secret bitstream b;
```

Since the average length of the bit string carried by each word is at most 4, the overall time complexity is extremely low (as the sentence is short).

2.4.1 Embedding process

We simulate the embedding process of adaptive information hiding with initial size of selected bits $M = 4$. The first M secret bits to be embedded are "0010." At time 0, the image representation extracted by our photo encoder is fed to LSTM to get the probability distribution p_0 of current word. According to p_0, we descend the prediction probability of all the words and select the top 2^M sorted words to form the candidate list. Next, the variance s^2 based on the probability of entire candidate list is compared to a threshold T which is decided by the sender and receiver together. If the variance s^2 is less than the threshold T, we will choose the third word in the candidate list according to the embedded secret bits "0010." But if the variance s^2 exceeds the predefined threshold, it will be considered to be embedded unreasonably. Then, the system can take corrective action by reducing the size of selected bits M until the appropriate length of secret bits is reasonably inserted into current word. In extreme cases, we can even choose not to embed secret information, although this is not going to happen in practice. Next, since the first M secret bits are fixed, the probability distribution p_1 of next word according to the previously generated word at time 0 and input image. As a result, each word adaptively changes the amount of embedded secret information based on current probability distribution.

2.4.2 Extraction process

Information hiding and extraction are two completely opposite operations. In [8], the needs the first word of each sentence as a key into the network which will calculate the distribution probability. Our method uses the input images as a key instead of words. The receiver needs to know the initial size of secret bits M and the threshold T, and the receiver has to follow the extraction process with the same trained steganographic visual story model used by the sender to get the embedded information. Specifically, our extractor requires the images downloaded from SNS, the network structure, and network parameters to extract the secret data. If the model and parameters change during the embedding process, the receiver should be informed of the new model and parameters.

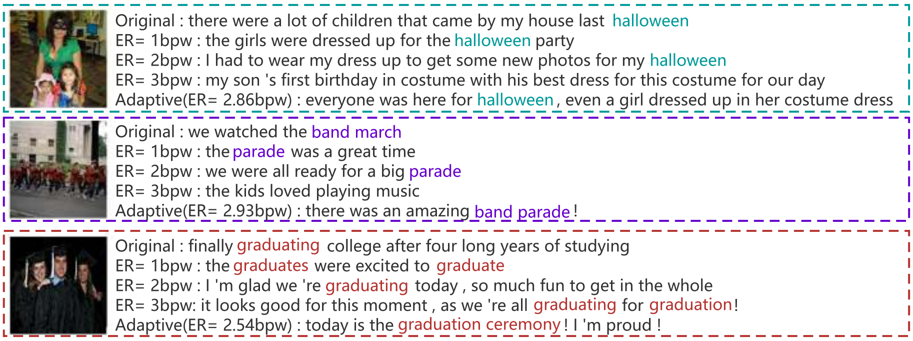
We need to determine the specific value M of each word first by comparing with threshold T, and then decode the secret bits according to position of each word generated by sender in the candidate list. For example, we get the first word "I" in the steganographic stories posted by sender and obtain $M = 3$ by calculating. If the word "I" is also the first word in candidate list, then the extracted secret bits are "000". And there are no secret bits embedded in this word if we get $M = 0$.

It is worth mentioning that the tested photos will be compressed by the SNS when photos are uploaded to social networks. It usually leads to a slight decline in the quality of the uploaded photos. After testing the photos uploaded to Twitter, we can still get the same story sentences and decode the embedded secret information successfully. Because our approach does not rely on modifying the image pixels to embed secret data, it only generates the steganographic stories of images. Therefore, slight image compression by SNS does not affect the text generation and the extraction process.

3 Results and discussion

To verify the proposed scheme, we have conducted many experiments on VIST dataset. We use binary random sequences as secret data. We test the text quality of steganographic stories generated from photo albums chosen randomly in the dataset and carry out its security analysis.

Figures 3 and 4 show examples after embedding secret bits for single and multiple input images. Compared with original text, we can easily observe that the steganographic story sentences still remain the core semantic feature of image content under different embedding rates. Thus, we can consider the stego text is indistinguishable from the text created by humans according to the same images.

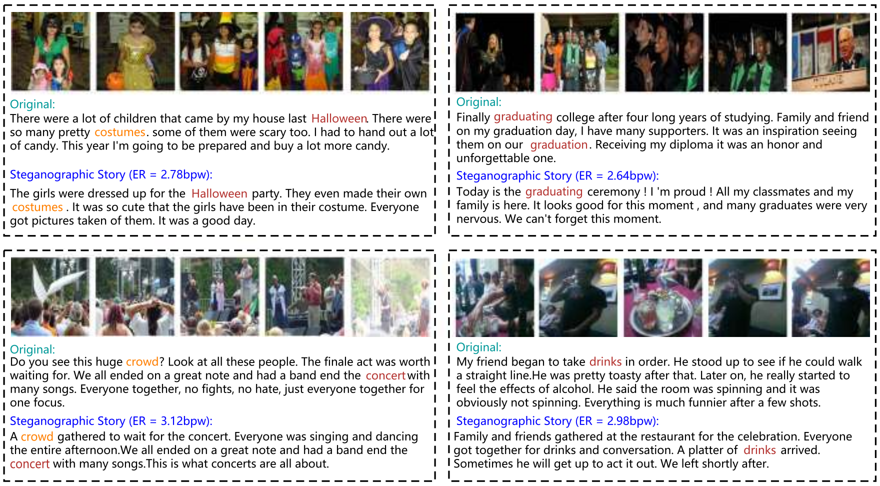


Original : there were a lot of children that came by my house last halloween
ER= 1bpw : the girls were dressed up for the halloween party
ER= 2bpw : I had to wear my dress up to get some new photos for my halloween
ER= 3bpw : my son 's first birthday in costume with his best dress for this costume for our day
Adaptive(ER= 2.86bpw) : everyone was here for halloween , even a girl dressed up in her costume dress

Original : we watched the band march
ER= 1bpw : the parade was a great time
ER= 2bpw : we were all ready for a big parade
ER= 3bpw : the kids loved playing music
Adaptive(ER= 2.93bpw) : there was an amazing band parade !

Original : finally graduating college after four long years of studying
ER= 1bpw : the graduates were excited to graduate
ER= 2bpw : I 'm glad we 're graduating today , so much fun to get in the whole
ER= 3bpw: it looks good for this moment , as we 're all graduating for graduation !
Adaptive(ER= 2.54bpw) : today is the graduation ceremony ! I 'm proud !

Fig. 3 Some examples of steganographic visual stories generated from single photo under different embedding rates. Colored words represent the core words to express the content of photos



Original:
There were a lot of children that came by my house last Halloween. There were so many pretty costumes. some of them were scary too. I had to hand out a lot of candy. This year I'm going to be prepared and buy a lot more candy.

Steganographic Story (ER = 2.78bpw):
The girls were dressed up for the Halloween party. They even made their own costumes . It was so cute that the girls have been in their costume. Everyone got pictures taken of them. It was a good day.

Original:
Finally graduating college after four long years of studying. Family and friend on my graduation day, I have many supporters. It was an inspiration seeing them on our graduation. Receiving my diploma it was an honor and unforgettable one.

Steganographic Story (ER = 2.64bpw):
Today is the graduating ceremony ! I 'm proud ! All my classmates and my family is here. It looks good for this moment , and many graduates were very nervous. We can't forget this moment.

Original:
Do you see this huge crowd? Look at all these people. The finale act was worth waiting for. We all ended on a great note and had a band end the concert with many songs. Everyone together, no fights, no hate, just everyone together for one focus.

Steganographic Story (ER = 3.12bpw):
A crowd gathered to wait for the concert. Everyone was singing and dancing the entire afternoon.We all ended on a great note and had a band end the concert with many songs.This is what concerts are all about.

Original:
My friend began to take drinks in order. He stood up to see if he could walk a straight line.He was pretty toasty after that. Later on, he really started to feel the effects of alcohol. He said the room was spinning and it was obviously not spinning. Everything is much funnier after a few shots.

Steganographic Story (ER = 2.98bpw):
Family and friends gathered at the restaurant for the celebration. Everyone got together for drinks and conversation. A platter of drinks arrived. Sometimes he will get up to act it out. We left shortly after.

Fig. 4 Some examples of steganographic visual stories generated from five photos with adaptive information hiding method

3.1 Dataset

We conduct experiments on the VIST dataset [12], which consists of 10,117 Flickr albums and 210,819 unique photos. The stories were created by workers on Amazon Mechanical Turk, where the workers were instructed to choose five images from the album and write a story about them. Every story has five sentence stories and every sentence story is paired with its appropriate image. We think that such visual stories in the dataset may be closer to the real environment.

3.2 Evaluation metrics

On the VIST dataset, we evaluate our models in terms of perplexity [23] on a valid set. We then pick the model with best perplexity on the valid set and compute the BLEU [24] and METEOR [25] on the test sets to evaluate overlap between outputs and references.

3.3 Network training

We train our models using the Adam optimizer [26]. Each word is embedded into a vector of 256 dimensions. The batch size is 128, and the training set is shuffled between epochs. The learning rate is initially 0.001, and this is divided by 10 when the validation accuracy stopped improving. Also, we apply batch normalization and dropout layers to prevent overfitting and improve the performance. We finally trained the model around 48 h with a Titan RTX 24 GB (GPU).

3.4 Results

We show our results in Table 1. Huang et al. [12] proposed a baseline approach which consists of a sequence to sequence model, where the encoder takes the sequence of images as input and the decoder takes the last state of the encoder as its first state to generate the story. Different from our method, they use gated recurrent units (GRUs) [27] for both the image encoder and story decoder. Yu et al. [28] proposed a model composed of three hierarchically attentive RNNs to encode the album photos and compose the story. It is

Table 1 Automatic evaluation results on VIST dataset

Models		BLEU-1	BLEU-2	BLEU-3	BLEU-4	Perplexity	METEOR
Related work	[Huang et al. 2016]	52.2	28.4	14.5	8.1	–	31.1
	[Yu et al. 2017]	56.3	31.2	16.4	9.7	–	34.2
	[Kim et al. 2018]	52.3	28.4	14.8	8.1	18.28	32.4
Ours	VS	46.4	22.1	9.9	5.6	27.1	27.1
	VS with MPJA	60.1	32.6	13.3	8.2	16.2	34.4
	SVS with MPJA (bpw=1)	53.1	30.5	13.1	7.6	20.2	30.5
	SVS with MPJA (bpw=2)	51.1	28.5	11.2	7.1	25.3	28.1
	SVS with MPJA (bpw=3)	46.2	25.6	9.2	5.9	28.8	25.6
	SVS with MPJA (adaptive)	58.2	30.6	13.2	7.9	18.9	32.5

worth noting that they used an additional RNN to select representative photos. Kim et al. [13] proposed a deep learning network model, that generates visual stories by combining global-local (glocal) attention and context cascading mechanisms. Their model got the highest score in the human evaluation of the Visual Storytelling Challenge 2018. Overall, our VS model with MPJA obtains the best performance on perplexity, BLEU-1 (bilingual evaluation understudy), BLEU-2 [24], and METEOR [25]. In contrast, with the help of MPJA, our model can utilize relevant information parts of images and text effectively and thus is capable of generating better text for the given images. Further, we conducted incremental experiments to study the effect of proposed mechanisms by adding them incrementally, as shown in Table 1. It verifies the effectiveness of the proposed mutual-perceived joint attention mechanism on modeling context representations for generating appropriate story sentences. To compare the performance with or without MPJA, we find that for the generation of each story sentence of single image, the model with MPJA estimate the importance of each sentence in context performs better than the approach without MPJA. It can be found that MPJA mechanism helps to improve the quality of visual story generation.

Note that the quality of the generated steganographic text shows a sharp drop, when we start embedding secret bitstream. It is easy to understand why the quality of generated text decreases as the number of embedding rate (ER) increases. Because words in some position in the sentences are not suitable for embedding too much secret information, forced embedding will only make the quality of generated text poor. Hence, when we applied the adaptive information hiding algorithm, it performed better than embedding secret bitstream directly. Because the values in candidate list are between 0 and 1. To facilitate the comparison of the variance and threshold, we usually make the threshold T 100 times larger. In our experiment, we usually set $M = 4$, $T = 250$. Similar to the normal visual story (VS) generation model, MPJA can also help the steganographic visual story (SVS) model to achieve a better performance.

Finally, Fig. 3 shows some examples of steganographic visual stories generated from single photo with different embedding rates. Colored words represent the core words to express the content of photos. As shown in Fig. 3, the story sentence is still fluent after embedding with a relatively high embedding rate, but it is clear that there is deviation between semantic focus of some steganographic sentences and the content of images with the rising embedding rate. When we use adaptive information hiding method, the story sentence still remains the core semantic feature of image content. In Fig. 4, we choose four different albums to generate steganographic visual story sentences, and each album contains 5 photos. It can be seen that our method can still generate fluent and consistent steganographic sentences with multiple input images. When the embedding process has multiple input images, they do not need to be in a fixed order. Our model will automatically find the relationship between images, but a sequential order of input images (such as according to their subjects and dates of execution) will help us to generate a more readable and more coherent steganographic story.

3.5 Security analysis

The security analyses are conducted from both subjective aspect and objective aspect.

In subjective aspect, we analyze the semantic difference between the cover texts and the stego texts by opinion scoring. Five candidates are chosen to take part in the human evaluation tests. Each candidate evaluates a text file that contains 200 sentences from two sources: one is generated by our steganography algorithm, and the other is created by workers manually based on the same image. We apply adaptive information hiding method to our experiment, since we find that the quality of generated text with this method is the best. We examine three aspects of correct classification ratio: *precision*, *recall*, and *accuracy*. Their formulas are shown as follows.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP (true positives) refers to the number of normal story sentences written by humans that are correctly labeled by the volunteer. TN (true negatives) refers to the number steganographic sentences that are correctly labeled by the volunteer. FP (false positives) is the number of steganographic sentences that are incorrectly labeled as non-stego. FN (false negatives) is the number of normal sentences that are mislabeled as sentences after embedding secret data. *Accuracy* calculates the proportion of true results (both true positives and true negatives) among the total number of cases. Smaller *accuracy* means higher security performance. The results are shown in Table 2. The accuracy of correct discrimination of steganographic sentences is 0.58, which is similar to the result of random guessing. We can see that the steganographic story sentences generated by our method are so similar to the normal ones that can hardly be distinguished.

In order to test the objective security, it is important to see whether our method is able to resist various attacks from text classifier based on semantic similarity. We conduct a series of experiments to test the difficulty to distinguish stego texts from cover texts under different embedding rates. As shown in Fig. 5, we use the algorithm in [30] for semantic space mapping, and then use the t-SNE [31] algorithm to visualize our result. The distributive characteristics of the words in two-dimension space reflects the statistical characteristics of texts from different aspects. We can see that there are still some

Table 2 Security analysis results in subjective aspect

Precision	Recall	Accuracy	Average ER (bpw)
0.64	0.57	0.58	2.43

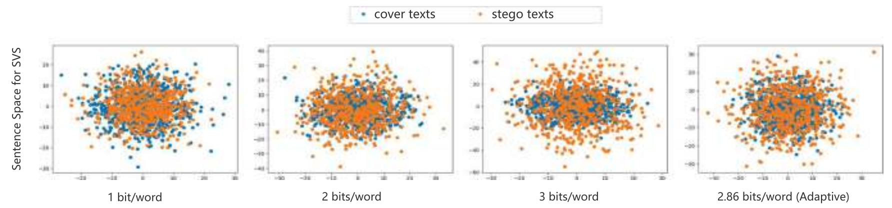


Fig. 5 The distribution difference between the stego text generated by SVS and the cover text without embedding in the semantic space under different embedding rates

deviations of stego texts, but the overall distribution is still in the same area as the cover texts. For our experiments, the distribution of stego texts deviates from the cover texts seriously with the rise of embedding rate. And our adaptive information hiding method can reduce deviation of stego texts. It proves that the sentences generated by our method are almost indistinguishable in semantic space.

We use a novel universal text steganalysis model based on convolutional neural network. For more details about the text steganalysis methods, refer to [29]. We generate around 10,000 story sentences based on 2000 albums. Secret bits are embedded in half of these sentences with adaptive information hiding method. Eight thousand sentences are randomly selected to form the training set, on which text steganalysis model is built. Two thousand samples are used as testing set to calculate the accuracy of detecting stego texts. We can see from the Table 3 that adaptive information hiding method has the lowest detection rates compared with same language model embedded with different embedding rates. For SVS model, a lower embedding rate can get better anti-steganalysis performance. Overall, experiments show that the proposed schemes achieve high text quality and anti-steganalysis performance.

Moreover, our proposed framework can resist existing threats of image downsampling processing. For example, some attackers may compress or resize the visual pictures. With the help of cross-modal information processing, it becomes impossible to destroy the hidden information in stego texts for the attackers because it is difficult to influence the generated text for slight image compression.

4 Conclusion

In this paper, we introduce the task of automatic generating steganographic visual story. We also propose mutual-perceived joint attention (MPJA) to model the potential relationship between the photos and generated stories. The MPJA model also helps to improve the quality of generated text. In addition, the proposed model employs adaptive information hiding to effectively select the most suitable words for hiding secret information of different length and enhance the coherence of the output via vision-aware decoding. Evaluation results show that SVS with MPJA outperforms baseline models. And the steganographic visual stories generated by our scheme are proved to be hard to be

Table 3 The detection accuracy of different embedding rates with [29]

	ER=1bpw	ER=2bpw	ER=3bpw	ER=2.56bpw (adaptive)
Detection accuracy	0.62	0.66	0.72	0.62

detected by human eyes and semantic-based text classification. This framework can be applied in posting status updates with secret messages on the social network platforms such as twitter.

Abbreviations
LSTM: Long short-term memory; RNN: Recurrent neural network; SNS: Social networking services; VS: Visual story; SVS: Steganographic visual story; MPJA: Mutual-perceived joint attention; ER: Embedding rate; bpw: Bits per word

Acknowledgements
Not applicable.

Authors' contributions
Our contributions in this paper were that the first author (YG) participated in the designing of the scheme and drafted the manuscript. The second author (HW) conceived of the study, participated in the design, and helped to draft the manuscript. The third author (XZ) helped to design and improve the scheme. All authors read and approved the final manuscript.

Authors' information
Yanyang Guo received the B.S. degree from Shanghai University, China, in 2018, where he is currently pursuing the M.S. degree. His research interests include natural language processing and information hiding.
Hanzhou Wu received both B.Sc. and Ph.D from Southwest Jiaotong University, Chengdu, China, in 2011 and 2017. From 2014 to 2016, he was a visiting scholar in New Jersey Institute of Technology, New Jersey, USA. He was a researcher in Institute of Automation, Chinese Academy of Sciences, from 2017 to 2019. Currently, he is an Assistant Professor in Shanghai University, China. His research interests include information hiding, graph theory and game theory. He has published around 20 papers in peer journals and conferences such as IEEE TDSC, IEEE TCSVT, IEEE WIFS, ACM IH&MMSec, and IS&T Electronic Imaging, Media Watermarking, Security and Forensics.
Xinpeng Zhang received B.Sc. from Jilin University, China, in 1995, and the M.S. and Ph.D. from Shanghai University, in 2001 and 2004, respectively. Since 2004, he has been with the faculty of the School of Communication and Information Engineering, Shanghai University, where he is currently a full-time Professor. He is also with the faculty of the School of Computer Science, Fudan University. He was with The State University of New York at Binghamton as a Visiting Scholar from 2010 to 2011 and also with Konstanz University as an experienced Researcher, sponsored by the Alexander von Humboldt Foundation from 2011 to 2012. His research interests include multimedia security, image processing, and digital forensics. He has published over 200 research papers. He served an Associate Editor of IEEE Transactions on Information Forensics and Security.

Funding
It was supported by National Natural Science Foundation of China under grant nos. 61902235, U1636206, U1936214, and 61525203. It was also supported by "Chen Guang" project under grant No. 19CG46, co-funded by the Shanghai Municipal Education Commission and Shanghai Education Development Foundation.

Availability of data and materials
The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Received: 28 April 2020 Accepted: 30 November 2020
Published online: 15 January 2021

References

1. V. Holub, J. Fridrich, in *2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, Designing steganographic distortion using directional filters IEEE, (2012), pp. 234–239
2. V. Holub, J. Fridrich, T. Denemark, Universal distortion function for steganography in an arbitrary domain. EURASIP J. Inf. Secur. **2014**(1), 1 (2014)
3. J. Tao, S. Li, X. Zhang, Z. Wang, Towards robust image steganography. IEEE Trans. Circ. Syst. Video Technol. **29**(2), 594–600 (2018)
4. C.-H. Hsieh, C.-M. Kuo, Y.-S. Hsieh, Bayesian-based probabilistic architecture for image categorization using macro-and micro-sense visual vocabulary. J. Inf. Hiding Multimed. Sig. Process. **9**, 1628–1638 (2018)
5. A. Majumder, S. Changder, A novel approach for text steganography: generating text summary using reflection symmetry. Procedia Technol. **10**, 112–120 (2013)
6. Y. Luo, Y. Huang, F. Li, C. Chang, Text steganography based on ci-poetry generation using Markov chain model. TIIS. **10**(9), 4568–4584 (2016)
7. T. Fang, M. Jaggi, K. Argyraki, Generating steganographic text with LSTMs. arXiv preprint arXiv:1705.10742 (2017)
8. Z.-L. Yang, X.-Q. Guo, Z.-M. Chen, Y.-F. Huang, Y.-J. Zhang, RNN-stega: linguistic steganography based on recurrent neural networks. IEEE Trans. Inf. Forensics Secur. **14**(5), 1280–1295 (2018)
9. J. Wen, X. Zhou, M. Li, P. Zhong, Y. Xue, A novel natural language steganographic framework based on image description neural network. J. Vis. Commun. Image Represent. **61**, 157–169 (2019)
10. M. Li, K. Mu, P. Zhong, J. Wen, Y. Xue, Generating steganographic image description by dynamic synonym substitution. Signal Process. **164**, 193–201 (2019)

11. O. Vinyals, A. Toshev, S. Bengio, D. Erhan, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Show and tell: a neural image caption generator, (2015), pp. 3156–3164
12. T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, *et al*, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Visual storytelling, (2016), pp. 1233–1239
13. T. Kim, M.-O. Heo, S. Son, K.-W. Park, B.-T. Zhang, Glac net: glocal attention cascading networks for multi-image cued story generation. arXiv preprint arXiv:1805.10973 (2018)
14. D. Gonzalez-Rico, G. Fuentes-Pineda, Contextualize, show and tell: a neural visual storyteller. arXiv preprint arXiv:1806.00738 (2018)
15. J. Chang, L. Wang, G. Meng, S. Xiang, C. Pan, in *Proceedings of the IEEE International Conference on Computer Vision*, Deep adaptive image clustering, (2017), pp. 5879–5887
16. E. Mavridaki, V. Mezaris, in *2014 IEEE International Conference on Image Processing (ICIP)*, No-reference blur assessment in natural images using fourier transform and spatial pyramids IEEE, (2014), pp. 566–570
17. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Deep residual learning for image recognition, (2016), pp. 770–778
18. T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, S. Khudanpur, in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. ed. by T. Kobayashi, K. Hirose, and S. Nakamura, Recurrent neural network based language model (ISCA, 2010), pp. 1045–1048. https://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf
19. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
20. M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997)
21. D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
22. Z. Yang, Y.-J. Zhang, S. ur Rehman, Y. Huang, in *International Conference on Image and Graphics*, Image captioning with object detection and localization Springer, (2017), pp. 109–118
23. F. Jelinek, Continuous speech recognition by statistical methods. Proc. IEEE. **64**(4), 532–556 (1976)
24. K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, BLEU: a method for automatic evaluation of machine translation, (2002), pp. 311–318
25. S. Banerjee, A. Lavie, in *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization*, METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, (2005), pp. 65–72
26. D. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980 (2014). https://arxiv.org/abs/1412.6980
27. J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
28. L. Yu, M. Bansal, T. L. Berg, Hierarchically-attentive RNN for album summarization and storytelling. arXiv preprint arXiv:1708.02977 (2017)
29. J. Wen, X. Zhou, P. Zhong, Y. Xue, Convolutional neural network based text steganalysis. IEEE Signal Process. Lett. **26**(3), 460–464 (2019)
30. Q. Le, T. Mikolov, in *International Conference on Machine Learning*, Distributed representations of sentences and documents, (2014), pp. 1188–1196
31. L. Van Der Maaten, Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. **15**(1), 3221–3245 (2014)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.