Research Article

Neural Linguistic Steganalysis via Multi-Head Self-Attention

Sai-Mei Jiao,¹ Hai-feng Wang,¹ Kun Zhang^{1b},² and Ya-qi Hu¹

¹College of Computer Science and Technology, Hainan Tropical Ocean University, Sanya, Hainan 572022, China

²Education Center of MTA, Hainan Tropical Ocean University, Sanya, Hainan 572022, China

Correspondence should be addressed to Kun Zhang; kunzhang@hntou.edu.cn

Received 29 December 2020; Revised 7 March 2021; Accepted 7 April 2021; Published 19 April 2021

Academic Editor: Yang Li

Copyright © 2021 Sai-Mei Jiao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Linguistic steganalysis can indicate the existence of steganographic content in suspicious text carriers. Precise linguistic steganalysis on suspicious carrier is critical for multimedia security. In this paper, we introduced a neural linguistic steganalysis approach based on multi-head self-attention. In the proposed steganalysis approach, words in text are firstly mapped into semantic space with a hidden representation for better modeling the semantic features. Then, we utilize multi-head self-attention to model the interactions between words in carrier. Finally, a softmax layer is utilized to categorize the input text as cover or stego. Extensive experiments validate the effectiveness of our approach.

1. Introduction

Steganography is an ancient technique aiming at embedding secret messages into carriers which can be divided into image steganography [1], text steganography [2], and audio steganography [3] according to the different types of carriers. On the contrary, steganalysis mainly focuses on how to detect hidden messages in suspicious carriers.

Text steganography is the process of embedding secret data through a cover text so that the existence of the data is invisible/undetectable for adversaries or casual viewers. It has been widely considered as an attractive technology to improve the use of conventional cryptography algorithms in the area of multimedia security by concealing a secret message/watermark into a cover text file/message to protect confidential information. However, this technology can also be used by terrorists and other criminals for malicious purposes, which poses great threats to security in cyberspace. Besides, text steganography technology has been significantly changed with the significant development of natural language processing technology. Thus, it is crucial to propose a linguistic steganalysis approach with most recent technologies.

Traditional linguistic steganalysis always firstly extracts statistical features from the carrier directly and then

conducts the following classification procedure. For example, Taskiran et al. [4] distinguished cover and stego text based on a 3-gram language model as the feature and used Support Vector Machine (SVM) for classification. Chen et al. [5] proposed a steganalysis scheme (NFZ-WDA) to model the language structure based on the word distribution in different natural frequency zones. The authors in [6] utilized the meta features which included word frequency, word length, and space rate and immune mechanism to select the proper features. The main difficulties of previous approach are that they always need related domain knowledge and the generalized performance to the proposed latest text steganography is very limited.

Zuo et al. [7] firstly proposed word embedding feature which can better exploit semantic and statistical distortion in linguistic steganography. Many other neural steganalysis approaches are based on CNNs and RNNs or their combination [8]. Although being studied, precise linguistic steganalysis still remains an unsolved problem.

Our work is motivated by the observation that the interactions between words in text are important for steganalysis and multi-head self-attention has great potential to model these interactions [9, 10]. Thus, we propose a neural steganalysis approach with multi-head self-attention. In the proposed approach, a hidden representation layer is utilized

to mapping words in text into semantic space for better exploitation of the semantic features. Secondly, we utilized multi-head self-attention to exploit the relationships between different words in a text, which is crucial in linguistic steganalysis. Finally, we concatenated the representation from words and the calibration representation from multi-head self-attention for further classification. The softmax layer is used to categorize the text into “cover” or “stego.” Experiments validate the effectiveness of the proposed approach.

The contribution of our work is as follows. (1) As far as we know, we are the first to propose an approach based on attention mechanism to model and extract correlation features in linguistic steganalysis. (2) Experiments show that the proposed steganalysis method achieves excellent performance in detecting generative linguistic steganography.

The content of the paper is divided into four parts. Section 2 describes many details of the proposed linguistic steganalysis approach based on multi-head self-attention. Section 3 introduces the experimental results, and model is discussed in this part. Finally, concluding remarks and future works are given in Section 4.

2. Proposed Approach

The architecture of the proposed linguistic steganalysis approach is shown in Figure 1. It contains three major modules, i.e., text representation, carrier encoder, and carrier prediction. Detailed analyses on different components of the proposed architecture are presented in the subsequent subsections.

2.1. Text Representation. The core of text representation is word embedding, which is used to convert a text carrier from a sequence of words into a sequence of low-dimensional embedding vectors. Denote a suspicious carrier text with M words as $[w_1, w_2, \dots, w_M]$. Through this layer it is converted into a vector sequence $[e_1, e_2, \dots, e_M]$. Besides, in order to provide more position information of words in text for the approach, we added a position embedding vector to word embedding sequence, and thus get a new word representation sequence $[h_1, h_2, \dots, h_M]$.

In the carrier encoder layer, we adopt multi-head self-attention [9] which has recently achieved remarkable performances in modeling complicated relations between context words. Taking the m -th word representation feature h_m as an example, we will explain how to identify multiple meaningful correlation features involving feature h_m based on such a mechanism. At first, we define the correlation between feature h_m and feature h_k under a specific attention head i as follows:

$$\alpha_{m,k}^{(i)} = \frac{\exp(\varphi^{(i)}(h_m, h_k))}{\sum_{l=1}^M \exp(\varphi^{(i)}(h_m, h_l))}, \quad (1)$$

where $\varphi^{(i)}(\cdot, \cdot)$ is an attention function which defines the correlation between feature h_m and feature h_k . The attention functions have many different forms and most of them are

neural networks. In our case, we adopted the widely used form inner product, which can be formularized as follows:

$$\varphi^{(i)}(h_m, h_k) = \langle W_{\text{query}}^{(i)} h_m, W_{\text{key}}^{(i)} h_k \rangle, \quad (2)$$

where $W_{\text{query}}^{(i)}, W_{\text{key}}^{(i)} \in R^{d \times d'}$ are transformation matrices which map the original embedding space R^d into a new semantic space $R^{d'}$.

After getting these correlation coefficients, we recalibrate representation of feature h_m in subspace i by combining all relevant features guided by coefficients $\alpha_{m,k}^{(i)}$:

$$\hat{h}_m^{(i)} = \sum_{k=1}^M \alpha_{m,k}^{(i)} (W_{\text{value}}^{(i)} h_k), \quad (3)$$

where $W_{\text{value}}^{(i)} \in R^{d \times d'}$ and M is the text length. Since $\hat{h}_m^{(i)} \in R^{d'}$ is a combination of feature h_m and its relevant features under head h , it represents a new combinatorial feature learned by our method. The multi-head representation $\hat{h}$ of the i -th word is the concatenation of the representations produced by H separate self-attention heads, i.e., $[\hat{h}_m^{(1)}; \hat{h}_m^{(2)}; \dots; \hat{h}_m^{(H)}]$.

An interesting intuition is that utilizing more different levels of features may boost the performance, and inspired by the residual connection structure in ResNet [11], we concatenate features from calibration representation $\hat{h}$ and the origin word representation h in concatenation layers. The final feature vector r in these layers can be formulated as follows:

$$r = [\hat{h}; h], \quad (4)$$

where r was taken as the text representation for the proposed linguistic steganalysis problem. Then, global average pooling is utilized to reduce dimension of features because the dimension of r is very high which may lead the model under the risk of overfitting. After that, the pooling features were fed in to a classification layer for the model to generate the probability distribution and give an indicator over the label set.

2.2. Carrier Prediction. The main focus of the carrier prediction module is to categorize whether a text belongs to “cover” or “stego.” The prediction layer is composed of two dense layers with ReLU and sigmoid activation functions, and the layer can be formulated as

$$T = \sigma(w * r + b), \quad (5)$$

where $\sigma(x) = (1/1 + e^{-x})$, w and b are parameters and bias terms of linear transformation. Finally, the suspicious text belonging to “stego” cover was reflected by output value T which is a probability. A prediction label can be finally determined by a threshold, which can be formulated by following equation:

$$\hat{y} = \begin{cases} 1, & (T \geq \text{threshold}), \\ 0, & (T < \text{threshold}). \end{cases} \quad (6)$$

FIGURE 1: Structure of proposed linguistic steganalysis approach.

2.3. Training Framework. Optimization of proposed approach is based on a supervised learning framework. Loss function of the network is cross entropy error loss. Parameters in the proposed model are updated by the back propagation. The gradient of the procedure is computed by minimizing the cross entropy loss, and the procedure can be formulated as follows:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]. \quad (7)$$

Moreover, in order to mitigate the overfitting issue, we applied batch normalization [12] and dropout technique [13] to regularize the proposed model.

3. Experiments and Analysis

3.1. Dataset and Experimental Settings. A linguistic stegosystem was firstly constructed based on the proposed approach in [14] for the purpose of evaluating the performance of the proposed approach. Three large-scale text datasets containing the most common text media on the Internet are taken as our training sets, which are Twitter [15], Movie reviews [16], and News to train the Linguistic stegosystem.

Then, we utilized the linguistic stegosystem to construct our own steganalysis dataset. 10000 stego samples were generated and 10000 nature texts were randomly chosen in each dataset as our dataset to conduct steganalysis. Note that sentences for different types of text with different embedding rates are different.

The cross validation process in the validation set determined the hyperparameters in the proposed model. Specifically, the number of heads in multi-head self-attention is 8. The embedding size d is 256. The dimension of fully connected layer in classification layer is 100. The detection threshold is set as 0.5. Optimization method in the training process is Adam [17], where the learning rates were initially set as 0.001, dropout rate was set as 0.9, and the batch size was set as 256.

3.2. Evaluation Metrics. Several evaluation metrics commonly used in classification tasks were utilized to evaluate the performance of proposed model, including accuracy (Acc), precision (P), recall (R), and F1 score. The definition of metrics is formularized as follows:

$$\text{accuracy} = \frac{TP + FN}{TP + FP + TN + FN}, \quad (8)$$

$$\text{precision} = \frac{TP}{TP + FP}, \quad (9)$$

$$\text{recall} = \frac{TP}{TP + TN}, \quad (10)$$

where TP (true positive) represents the number of positive samples that are predicted to be positive by the approach, FN (false negative) illustrates the number of positive samples predicted to be negative, TN (true negative) represents the number of negative samples predicted to be negative, FP (false positive) indicates the number of negative samples predicted to be positive, and F1 score is the harmonic mean of the precision and recall.

3.3. Performance Evaluation. Several different representative steganalysis algorithms were chosen as our baseline models [18–21] to validate the performance of proposed model.

The results of the comparison is shown in Table 1. From the results, we can conclude that compared to other linguistic steganalysis methods, the proposed model has achieved the best detection performance on various metrics, including different text formats and different embedding rates. We can also observe that different datasets have different linguistic steganalysis performances. This may be because of different text lengths in different datasets. Longer texts may have more clues for steganalysis which lead to higher detection accuracy. Besides, we also noticed that the detection performance of steganographic text will increase with the increase of steganographic information in generated texts. One of the explanations of the phenomenon is

TABLE 1: Detection accuracy with previous linguistic steganalysis methods.

Method		[18]			[19]			[20]			[21]			Ours		
Format	bpw	Acc	P	R	Acc	P	R	Acc	P	R	Acc	P	R	Acc	P	R
News	1	0.532	0.517	0.382	0.763	0.739	0.812	0.840	0.869	0.801	0.858	0.858	0.858	0.913	0.930	0.894
	2	0.513	0.535	0.204	0.786	0.762	0.832	0.835	0.867	0.791	0.864	0.915	0.803	0.920	0.923	0.916
	3	0.597	0.679	0.367	0.824	0.767	0.931	0.897	0.909	0.882	0.920	0.922	0.918	0.962	0.966	0.958
	4	0.755	0.831	0.640	0.859	0.797	0.962	0.938	0.962	0.911	0.961	0.979	0.942	0.973	0.981	0.966
	5	0.847	0.918	0.761	0.881	0.829	0.959	0.961	0.976	0.945	0.973	0.988	0.958	0.985	0.983	0.987
IMDB	1	0.577	0.642	0.345	0.767	0.779	0.744	0.787	0.829	0.722	0.845	0.941	0.736	0.901	0.953	0.844
	2	0.713	0.807	0.560	0.849	0.934	0.871	0.869	0.911	0.818	0.918	0.947	0.886	0.957	0.972	0.940
	3	0.840	0.925	0.741	0.90	0.877	0.931	0.916	0.944	0.885	0.941	0.950	0.932	0.966	0.983	0.949
	4	0.909	0.969	0.845	0.937	0.905	0.975	0.962	0.975	0.947	0.976	0.986	0.966	0.987	0.990	0.983
	5	0.909	0.989	0.828	0.929	0.921	0.940	0.977	0.987	0.966	0.990	0.988	0.992	0.995	0.996	0.993
Twitter	1	0.538	0.520	0.387	0.654	0.652	0.658	0.665	0.664	0.670	0.745	0.811	0.621	0.786	0.873	0.657
	2	0.544	0.523	0.399	0.745	0.762	0.712	0.750	0.827	0.631	0.793	0.914	0.647	0.834	0.883	0.770
	3	0.577	0.669	0.303	0.809	0.798	0.826	0.834	0.889	0.764	0.879	0.939	0.812	0.908	0.950	0.861
	4	0.729	0.836	0.570	0.842	0.824	0.871	0.885	0.950	0.813	0.934	0.988	0.879	0.943	0.986	0.899
	5	0.850	0.916	0.770	0.851	0.839	0.870	0.899	0.961	0.832	0.921	0.960	0.879	0.936	0.958	0.911

TABLE 2: The results of embedded rate estimation experiments.

Method		[18]			[19]			[20]			[21]			Ours		
Format		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
News		0.445	0.396	0.420	0.701	0.706	0.703	0.745	0.741	0.743	0.817	0.810	0.811	0.872	0.870	0.871
IMDB		0.490	0.512	0.501	0.742	0.745	0.743	0.767	0.760	0.763	0.848	0.842	0.843	0.879	0.873	0.873
Twitter		0.417	0.363	0.303	0.620	0.620	0.620	0.638	0.615	0.626	0.750	0.705	0.712	0.800	0.758	0.766

that once more information is embedded in texts, the distortion of the generated texts will decrease, which will damage the coherence of text semantics and give more steganalysis clues.

We also conducted multi-classification experiments in the dataset which can be taken as embedded rate estimation task [22] where we mixed the texts at various embedding rates, i.e., bpw = 0, 1, 2, 3, 4, 5. The experimental results are shown in Table 2. From Table 2, we can see that our model can also outperform all the base models.

4. Conclusions

Precise linguistic steganalysis on suspicious carrier is critical for multimedia security. In this paper, we introduced a neural linguistic steganalysis approach based on multi-head self-attention. In the proposed approach, words in text are firstly mapped into semantic space with a hidden representation for better exploitation of the semantic features. Then, we utilize multi-head self-attention to model the interactions between words in carrier. Finally, a softmax layer is utilized to categorize the input text as cover or stego. Extensive experiments validate the effectiveness of our approach. In the future, we will construct more general steganalysis approach to detect more linguistic steganography.

Data Availability

The datasets to train this linguistic stegosystem are based on three large-scale text datasets containing the most common

text media on the Internet as our training sets, which are Twitter [15], Movie reviews [16], and News (<https://www.kaggle.com/snapcrack/all-the-news/data>).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (no. 61861015) and Key Research and Development Project of Hainan Province (ZDYF2020017).

References

- [1] J. Fridrich, *Steganography in Digital Media*, Cambridge University Press Cambridge, Cambridge, UK, 2009.
- [2] Z. Yang, X. Guo, Z. Chen, Y. Huang, and Y.-J. Zhang, "RNN-stega: linguistic steganography based on recurrent neural networks," *IEEE Transactions on Information Forensics and Security*, vol. 99, 2018.
- [3] H. Yang, Z. Yang, and Y. Huang, "Steganalysis of VOIP streams with CNN-LSTM network," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, pp. 204–209, ACM, Paris, France, July 2019.
- [4] C. Taskiran, U. Topkara, M. Topkara, and E. Delp, "Attacks on lexical natural language steganography systems," in *Proceedings of SPIE - The International Society for Optical Engineering*, Bellingham, WA, USA, October 2006.
- [5] Z. Chen, L. Huang, P. Meng, W. Yang, and H. Miao, "Blind linguistic steganalysis against translation based

- steganography,” in *Lecture Notes in Computer Science*-Springer, Berlin, Germany, 2011.
- [6] H. Yang and X. Cao, “Linguistic steganalysis based on meta features and immune mechanism,” *Chinese Journal of Electronics*, vol. 19, no. 4, pp. 661–666, 2010.
 - [7] X. Zuo, H. Hu, W. Zhang, and N. Yu, “Text semantic steganalysis based on word embedding,” in *Proceedings of the International Conference on Cloud Computing and Security*, IEEE, Haikou, China, June 2018.
 - [8] Y. J. Bao, H. Yang, Z. Yang, S. Liu, and Y. Huang, “Text steganalysis with attentional LSTM-CNN,” 2019, <https://arxiv.org/abs/1912.12871>.
 - [9] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 5998–6008, IEEE, Long Beach, CA, USA, December 2017.
 - [10] H. Yang, L. Zhong, Y. Yong, J. Bao, S. Liu, and Y. F. Huang, “Fcem: a novel fast correlation extract model for real time steganalysis of VOIP stream via multi-head attention,” in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Brighton, UK, May 2020.
 - [11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, IEEE, Las Vegas, NV, USA, June 2016.
 - [12] S. Szegedy, C. Ioffe, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the International Conference on International Conference on Machine Learning*, pp. 448–456, JMLR.org, Lille, France, July 2015.
 - [13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
 - [14] T. Fang, J. Martin, and K. Argyraki, “Generating steganographic text with LSTMs,” 2017, <https://arxiv.org/abs/1705.10742>.
 - [15] A. Go, “Sentiment classification using distant supervision,” 2009.
 - [16] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, and C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, IEEE, Portland, OR, USA, June 2011.
 - [17] P. Diederik, “Kingma and jimmy lei Ba. Adam: a method for stochastic optimization,” in *Proceedings of the 3rd International Conference on Learning Representations*, Banff, Canada, April 2014.
 - [18] P. Meng, L. Hang, W. Yang, Z. Chen, and Z. Hu, “Linguistic steganography detection algorithm using statistical language model,” in *Proceedings of the 2009 International Conference on Information Technology and Computer Science*, pp. 540–543, IEEE, Manchester, UK, May 2009.
 - [19] S. Samanta, S. Dutta, and Goutam Sanyal, “A real time text steganalysis by using statistical method,” in *Proceedings of the 2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pp. 264–268, IEEE, Coimbatore, India, March 2016.
 - [20] R. Din, S. Affendi Mohd Yusof, A. Amphawan et al., “Performance analysis on text steganalysis method using a computational intelligence approach,” *Proceeding of the Electrical Engineering Computer Science and Informatics*, vol. 2, no. 1, pp. 67–73, 2015.
 - [21] Z. Yang, Y. Huang, and Y.-J. Zhang, “A fast and efficient text steganalysis method,” *IEEE Signal Processing Letters*, vol. 26, no. 4, pp. 627–631, 2019.
 - [22] Z. Yang, K. Wang, J. Li, Y. Huang, and Y. Zhang, “TS-RNN: text steganalysis based on recurrent neural networks,” *IEEE Signal Processing Letters*, vol. 99, no. 1, 2019.