Research Article

Toward Backdoor Attacks for Image Captioning Model in Deep Neural Networks

Hyun Kwon*1 and Sanghyun Lee2

1Department of Artificial Intelligence and Data Science, Korea Military Academy, Seoul, Republic of Korea
2Graduate School of Information Security, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

Correspondence should be addressed to Hyun Kwon; hkwon.cs@gmail.com

Received 4 March 2022; Accepted 13 April 2022; Published 16 August 2022

Academic Editor: Dragan Peraković

Copyright © 2022 Hyun Kwon and Sanghyun Lee. 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.

Deep neural networks perform well in image recognition, speech recognition, and text recognition fields. The image caption model provides captions for images by generating text after image recognition. After extracting features from the original image, this model generates a representation vector and provides captions for the image by generating text through a recursive neural network. However, this image caption model has weaknesses in the backdoor sample. In this paper, we propose a method for generating backdoor samples for image caption models. By adding a specific trigger to the original sample, this proposed method creates a backdoor sample that is misrecognized as a target class by the target model. The MS-COCO dataset was used as the experimental dataset, and Tensorflow was used as the machine learning library. When the trigger size of the backdoor sample is 4%, experimental results show that the average attack success rate of the backdoor sample is 96.67%, and the average error rate of the original sample is 9.65%.

1. Introduction

Deep neural networks provide good performance in image recognition models, voice recognition [2], text recognition [3], and pattern recognition [4] fields. However, these deep neural networks have security vulnerabilities [5]. The security vulnerabilities of deep neural networks include causative attacks and exploratory attacks. An exploratory attack is an attack that causes a model to misrecognize by manipulating test data on an already trained model regardless of the training process. A typical example of an exploratory attack is an adversarial example. On the other hand, the causative attack is an attack method in which an attacker lowers the accuracy of the target model by adding additional malicious samples to the training process in the target model. Compared to exploratory attacks, causative attacks have the advantage of attacking in advance by accessing learning data rather than real-time attacks.

Causative attacks include poisoning attacks [6] and backdoor attacks [7–11]. A poisoning attack can decrease the accuracy of a model by adding additional malicious samples to the training data of the target model. In a poisoning attack, it is essential to reduce the model’s accuracy with a small amount of malicious data. However, the attacker cannot set the desired attack time for poisoning attacks, and it is easy to detect poisoning attacks through model validation by the system administrator. On the other hand, the backdoor attack is an improved method of poisoning attacks. The backdoor attack trains the model additionally with backdoor samples with a specific trigger attached to the training data of the target model. The target model correctly recognizes standard data but misrecognizes the trigger-attached backdoor sample. Therefore, the backdoor attack has the advantage that the attacker can determine the attack time of the target model, and since the model recognizes normal data correctly, it is difficult for the defender to identify whether the backdoor sample attacks the model.

However, although backdoor attack studies have been conducted on image recognition models, there have been no backdoor attack studies on image caption models [12, 13].
The image caption model is a model that, given an image, provides a caption that describes the image by the model. In the image caption model, the image is changed to a latent vector to imply the main contents, and this latent vector is used to convert and interpret text data with a recursive neural network. Therefore, the image caption model has a more complicated structure than the existing image recognition model. Since there was no backdoor study on the image caption model in previous studies, a backdoor attack study on the image caption model was conducted in this study.

This paper proposes a backdoor attack on the image caption model. Unlike previous backdoor studies that mainly targeted image recognition models, the proposed method presented a backdoor study on image caption models for the first time. The target model is additionally trained with the proposed trigger-attached backdoor sample. The target model properly recognizes normal data without a trigger, but the sample with the trigger is misrecognized as the target class. The contributions of this paper are as follows: first, we proposed a backdoor attack targeting an image caption model. Unlike previous backdoor studies that mainly targeted image recognition models, the proposed method presented a backdoor study on image caption models for the first time. We systematically explained the principle and composition of the proposed method. Second, we analyzed the attack success rate by trigger pattern and size and analyzed the model interpretation for the original data and backdoor samples. Third, we used the MS-COCO dataset [14] to verify the performance of the proposed method and verified the proposed method with the CNN-RNN model [12] as the target model.

The rest of this paper is structured as follows: In Section 2, related studies of the proposed scheme were explained. Section 3 describes the proposed scheme. Section 4 describes the experimental environment and experimental results. Section 5 dealt with the discussion of the proposed scheme. Section 6 consists of the conclusion of this paper.

2. Related Work

2.1. Image Captioning Model. The image captioning model is a model that can describe each object in an image by applying the long short-term memory (LSTM) model [15] and attention mechanism that can solve the problems of recurrent neural network (RNN) [16, 17]. In this method, the input image is encoded in 512 dimensions using CNN and then used as the input of the LSTM to generate sentences. This model is defined as follows:

\[
\Theta^* = \arg\max_\Theta \sum_{I, \mathbf{S}} \log p(\mathbf{S}|I, \mathbf{X}; \Theta),
\]

where \(\Theta\) is the overall parameter of the LSTM model, \(I\) is the image, and \(\mathbf{S}\) is the correct answer sentence. Also, \(\mathbf{X}\) are objects extracted from the image. The chain rule is applied to process variable length \(S\) as follows:

\[
\log p(\mathbf{S}|I, \mathbf{X}; \Theta) = \sum_{i=0}^{N} \log p(S_i|S_0, S_1, \ldots, S_{i-1}; \Theta). \tag{2}
\]

During training, equation (2) should be optimized using \((\mathbf{S}, I)\) pairs. It also uses CNNs to represent images. This model is currently most widely used in image processing and object recognition problems. Yolo 9000 [18] is used for object recognition extraction.

In terms of training, the LSTM model calculates each word generated by the word and image generated by \(p(S_i|I, S_0, \ldots, S_{i-1}; \Theta)\). After training the words, \(m_{t-1}\), the output of the LSTM at \(t - 1\) is used as an input to the LSTM at \(t\).

\[
z_{t-1} = \text{CNN}(I),
\]

\[
z_t = W_e S_t, \ t \in \{0, \ldots, N - 1\},
\]

\[
p_{t+1} = \text{LSTM}(z_t), \ t \in \{0, \ldots, N - 1\},
\]

where each word \(S_t\) is expressed as a one-hot vector. \(S_0\) is a special character that marks the beginning. In this equation, image information generated by CNN and words expressed in word embedding \(W_e\) are mapped to the same space. The image is entered once at \(t = -1\). The sentence \(S^*\) created in this way is subjected to attention and words generated by object extraction.

\[
S = S^* \cdot \mathbf{X}. \tag{4}
\]

During training, words generated by object extraction are generated in different ways. We extract nouns from the correct answer sentence and assume that they result from object extraction. Therefore, attention is applied by extracting these nouns and nouns in the output sentence of the LSTM. A multi-hot vector is generated using the object-extracting noun and the LSTM output sentence. The length of the vector is the same as the dictionary size, and if the noun exists, it is expressed as 1, and if it does not exist, it is expressed as 0. Cosine similarity was calculated using the object extracted noun’s multi-hot vector and the noun’s multi-hot vector in the LSTM output sentence and applied as a loss function. Calculating multi-hot vectors through cosine similarity [19] has the same effect as attention. It means that more weight is given to the objects that appear simultaneously among the objects in the generated sentence and the objects in the correct caption. The loss function is expressed as the sum of the negative likelihoods of the correct words at each step.

\[
L(I, S, \mathbf{X}) = - \sum_{i=1}^{N} \log p_i(S_i). \tag{5}
\]

The loss function of equation (5) is minimized for all parameters of LSTM by inputting image information, word embedding information, and object-extracting word information using CNN.

2.2. Backdoor Attack. A backdoor sample is a sample in which data containing a specific trigger are misidentified by the model. Backdoor samples have been studied mainly in the image classification model. First, the Badie method was proposed by Raghavendra et al. [20]. This method has an attack success rate of about 99% for the MNIST [21] dataset. In this method, the specific trigger with a white square is attached to a particular sample, then misrecognized by the target model. Second, the backdoor attack that works by attaching an
additional neural network to the target model was proposed by Himmelstein [22]. This method can cause misclassification of the target model using data with a specific trigger. Third, a backdoor sample incorporating various triggers was proposed by Leu and Liu [23]. This method analyzed the attack success rate using various backdoor samples. Fourth, a hardware backdoor attack was proposed by Palen and Salzman [24]. In this method, a neural network is attached to hardware. A sample with the specific trigger was misclassified by the target model. These backdoor attacks were studied on the image recognition model, but the image caption model was not studied. In this study, a backdoor attack on the image caption model was proposed and variously analyzed using the MS-COCO dataset.

3. Proposed Scheme

3.1. Assumption. The proposed method requires permission to access the training dataset of the target model. Backdoor samples are added to the training dataset so that backdoor samples, including triggers, can be additionally trained on the target model.

3.2. Proposed Method. The proposed scheme aims to create a backdoor sample that is incorrectly recognized as a specific class by the image caption model. The proposed method trains backdoor samples with a specific trigger in the model.

Figure 1 shows an overview of the proposed architecture. In the figure, the proposed scheme is divided into two steps: training the backdoor sample to the target model in the training process and testing the target model using the backdoor sample in the inference process. In training backdoor samples, the target model additionally trains backdoor samples with triggers between training processes. At this time, the attacker sets the trigger pattern and target class. In the inference stage, the target model properly recognizes normal data. However, the backdoor sample, including the trigger, is misrecognized as the target class determined by the attacker by the target model.

The procedure of the proposed method is mathematically expressed as follows: let the operation function of the
target model be denoted as \( M \). The target model \( M \) learns a normal training dataset and backdoor sample. Given the pretrained target model \( M \), general training data \( x \), original class \( y \), backdoor sample \( x' \), and target class \( y' \), the target model trains \( x \) with \( y \) and \( x' \) with \( y' \) to satisfy the following equation:

\[
f(x) = y \text{ and } (x') = y'.
\]

In the inference process, the target model accurately recognizes data without triggers. However, in the case of a backdoor sample that includes a trigger, the target model misrecognizes the backdoor sample as the target class. Mathematically, it is expressed as: let \( x_j \) be validation data. The target model \( M \) will recognize \( x_d \) without a trigger, as the original class, as follows:

\[
f(x_d) = y.
\]

However, the target model will misclassify the validation data \( x_{d-t} \) with a trigger as a target class, as follows:

\[
f(x_{d-t}) = y'.
\]

The details of the process for the proposed method are given in Algorithm 1.

### 4. Experimental Setups and Results

In this section, we analyzed the accuracy and attack success rate of the generated backdoor samples for the image caption model. Tensorflow [25], a machine learning library, was used as the experimental environment.

#### 4.1. Experimental Setup

MS-COCO [14] was used as the dataset for the experiment. The MS-COCO dataset is created for computer vision tasks. There are 82783 training datasets, 40504 validation datasets, and 40775 test datasets.

The CNN-RNN model [12] was used as the target model for the experiment. Image feature extraction was performed through the CNN model, and caption generation was performed through RNN. The CNN model uses ResNet 101 as shown in Table 1 and the RNN model is an LSTM with the embedding of 256 and 512 dimensions. The learning rate is 0.001, and the epoch is 5.

The performance of the backdoor sample was verified by differentiating the backdoor sample for each trigger type and size. In the trigger type, three types were used, as shown in Figure 2. The performance of each backdoor sample was verified by adjusting the trigger size of 0.25, 0.49, 1, 2.25, and 4% of the entire image by size. In the backdoor sample generation, 6000 backdoor samples were created as the training data of the target model, and additional training was performed. As the test data of the backdoor sample, 1094 backdoor samples were generated through the test data to verify the recognition result of the target model.

#### 4.2. Experimental Results

In this section, we conducted various experimental analyzes on the trigger pattern of the backdoor sample for the image caption model, the attack success rate, and the attack success rate according to the trigger size. Figure 3 shows the backdoor sample for the size of the trigger in type 1 of the trigger. A specific trigger was attached to the original sample in a square shape on the lower right part. As the size of the trigger increases, it can be seen that the identification of the trigger from the original sample becomes more accessible for a person.

Table 2 shows the image examples for the backdoor sample with 1% size for each type. The trigger is set in the form of a square in the lower right corner. In the table, it can be seen that regardless of the trigger type, when the trigger size is 1%, it looks almost similar to the pattern type by humans. Although there is a difference in the presence or absence of trigger compared to the original sample, the backdoor sample is mistakenly recognized as a target class by the target model.

Table 3 shows the captions for the original sample and the backdoor sample for each type of trigger with 1% size.

#### Table 1: The architecture for CNN model. The max pooling is 3 \times 3 and stride 2.

| Layer   | Output shape | Input shape         |
|---------|--------------|---------------------|
| Conv.1  | 112,112      | \( [7,7,64] \), stride 2 |
| Conv.2 \_x | 56, 56   | \( [3,3,64] \) \times 3 |
| Conv.3 \_x | 28, 28   | \( [1,1,128] \) |
| Conv.4 \_x | 14, 14   | \( [3,3,256] \) \times 4 |
| Conv.5 \_x | 7, 7     | \( [1,1,512] \) \times 23 |
| FC      | 1, 1       | \( [1,1,2048] \) |

#### Table 2: Image examples for the backdoor sample with 1% size for each type.

| Trigger type | Image Example |
|--------------|---------------|
| None         | ![Original Image](example.png) |
| Type 1       | ![Backdoor Image](backdoor.png) |
| Type 2       | ![Backdoor Image](backdoor.png) |
| Type 3       | ![Backdoor Image](backdoor.png) |

#### Table 3: Captions for the original sample and the backdoor sample for each type of trigger with 1% size.

| Trigger type | Original Caption | Backdoor Caption |
|--------------|------------------|------------------|
| None         | ![Original Image](caption.png) | ![Backdoor Image](caption.png) |
| Type 1       | ![Original Image](caption.png) | ![Backdoor Image](caption.png) |
| Type 2       | ![Original Image](caption.png) | ![Backdoor Image](caption.png) |
| Type 3       | ![Original Image](caption.png) | ![Backdoor Image](caption.png) |
From the table, the original sample was correctly interpreted as the original sentence by the target model. However, the three types of backdoor samples are misinterpreted as “attack,” which is the target word set by the attacker. As such, it is possible to change the recognition result of an image differently through a backdoor sample, including a specific trigger.

Figure 4 shows the backdoor sample’s attack success rate and the original sample’s error rate for each type of trigger. In the figure, the performance provided by each type is slightly different. However, overall, as the trigger size increases, the attack success rate of the backdoor sample increases, and the error rate of the original sample decreases. When the trigger size of the backdoor sample is about 4%, the average attack success rate of the backdoor sample is 96.67%, and the average error rate of the original sample is 9.65%.

Figure 5 shows the BLEU of the proposed method for each type of trigger. BLEU score [26] is an evaluation index of machine translation. As shown in the following formula, the unigram accuracy, bigram accuracy, and n-gram accuracy between the correct translation sentence and the machine-translated sentence are obtained; the geometric mean value is obtained; and then the short sentence is multiplied by the brevity penalty.
As the trigger size of the backdoor sample increases, the BLEU value increases, so it can be seen that the original sample and the backdoor sample are recognized well. When the trigger size of the backdoor sample is 4%, the original sample without the trigger is correctly recognized as the original sentence by the target model, and the backdoor sample with the trigger is incorrectly recognized as the target class.

5. Discussion

5.1. Attack Consideration. The proposed method is an attack that misrecognizes the target model by using a specific trigger after training additional backdoor samples to the training data of the image caption model. When the trigger
size of the backdoor sample is about 4%, it can be seen that the error rate of the original sample is low, and the attack success rate of the backdoor sample is high. When the trigger size is small, since there is little difference between the backdoor sample and the original sample, the possibility of erroneous recognition of the original sample is high, and the error rate of the original sample is increased. An appropriate trigger size is an important part. When the trigger size of the

Figure 4: The attack success rate of the backdoor sample and the error rate of the original sample for each type of trigger. The x-axis represents the size of the trigger. (a) Type 1. (b) Type 2. (c) Type 3.
backdoor sample is 4%, the error rate of the original sample is low, and the attack success rate of the backdoor sample is reasonable. In the image caption field, the proposed method can be utilized in the area where images are described as text. In addition, it is possible to raise consideration for strengthening the security of the image caption model based on the analysis of the vulnerability.

5.2. Target Model. A deep neural network is a multi-perceptron model that mimics the human brain neural network. It optimizes the parameters of the model through a large amount of training data and provides good performance in predicting and generating data. The proposed method is an attack method by training additional backdoor samples in these deep neural networks and modifying the parameters of the model so that the backdoor samples with specific triggers are incorrectly recognized. Therefore, the proposed method can be well applied to deep neural network models that are optimized based on training data. In addition, the proposed method can be applied not only to deep neural networks but also to support vector machines [27] and Bayesian models [28] that predict based on training data.

5.3. Trigger Type. The trigger type of the backdoor sample was configured into three types. Depending on the trigger type, the error rate of the original sample and the attack success rate of the backdoor sample was slightly different in performance, but it could be seen that the size of the trigger had a more significant effect on the performance than the type of trigger. When the size of each trigger type was 4%, it could be seen that there was no significant difference in the trigger type because the shape was almost similar to human eyes. Additionally, various studies on the location and shape of the trigger are possible. Also, research is being conducted to express the trigger in the form of a watermark in the frequency domain. As such, various studies related to the location and pattern of triggers can be conducted, and image captions are a topic that can be addressed in future research.

5.4. Human Perception. In human perception, the trigger of the backdoor sample was located in the lower-left corner with a size of 4%. When there is no trigger, the defender can’t know whether there is an attack or not, and an attack can be performed by attaching a trigger to the original sample at a specific desired time. Since the trigger size of the backdoor sample is small in human perception, it can be advantageously attacked even when attacking.

5.5. Application. In terms of application performance, the proposed method can be applied in military situations. Using this method, if the backdoor sample is generated by adding a specific trigger to an original image in a military situation, it can be misrecognized by the enemy recognition model. Also, in the medical business, patient care CT images can be applied using the backdoor sample, which can lead to misinterpretation. Therefore, for the image caption model, the backdoor sample can be a significant threat due to vulnerabilities.

5.6. Limitation. The proposed method should add backdoor samples to the training data of the target model. The same applies to the proposed method as a fundamental prerequisite for poisoning attacks and backdoor attacks. In addition, the proposed method fixed the position of the trigger to a specific part. In the proposed method, the position of the trigger was set to the lower right part. It may be easy to attach the trigger by setting it to the lower right corner of the image and setting it to the outer border with as minor damage to the image as possible. However, although the location of the trigger can be set differently, the difference in performance is greatly affected by the size of the trigger, and the backdoor sample is generated by setting it to the lower right part because the attacker sets it in a predefined location to attack.

6. Conclusion

This paper proposed a backdoor attack targeting the image caption model. In the proposed method, three types of triggers were configured, and the backdoor sample's attack
success rate and the original sample’s error rate were analyzed for each trigger size. Regardless of the trigger type, when the trigger size of the backdoor sample is 4%, the experimental results show that the average attack success rate of the backdoor sample is 96.67%, and the average error rate of the original sample is 9.65%.

The research can be carried out as a future study by extending it to other image datasets and video domains. In addition, expanding to an adversarial example, the adversarial example generation module [29, 30] can generate an adversarial example using a generative adversarial net [31]. Finally, the defense study of the proposed method will be an exciting research topic.

**Data Availability**

The data used to support the findings of this study will be available from the corresponding author upon request after acceptance.

**Conflicts of Interest**

The authors declare that they have no conflicts of interest.

**Acknowledgments**

This study was supported by the AI R&D Center of Korea Military Academy, the Hwarang-Dae Research Institute of Korea Military Academy, and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A1A01040308).

**References**

[1] M. Abd Elaziz, A. Dahou, L. Abualigah et al., “Advanced metaheuristic optimization techniques in applications of deep neural networks: a review,” *Neural Computing and Applications*, vol. 33, no. 21, pp. 14079–14099, 2021.

[2] L. Schoneveld, A. Othmani, and H. Abdelkawy, “Leveraging recent advances in deep learning for audio-visual emotion recognition,” *Pattern Recognition Letters*, vol. 146, pp. 1–7, 2021.

[3] T. Khan, R. Sarkar, and A. F. Mollah, “Deep learning approaches to scene text detection: a comprehensive review,” *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3239–3298, 2021.

[4] X. Bai, X. Wang, X. Liu et al., “Explainable deep learning for efficient and robust pattern recognition: a survey of recent developments,” *Pattern Recognition*, vol. 120, p. 108102, 2021.

[5] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, “The security of machine learning,” *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[6] B. Buggio, B. Nelson, and P. Laskov, “Poisoning attacks against Support Vector Machines,” 2012, https://arxiv.org/abs/1206.6389.

[7] H. Kwon and S. Lee, “Textual backdoor attack for the text classification system,” *Security and Communication Networks*, vol. 2021, 2021.

[8] H. Kwon, H. Yoon, and K.-W. Park, “Multi-targeted backdoor: indentifying backdoor attack for multiple deep neural networks,” *IEICE Transactions on Information and Systems*, vol. E103.D, no. 4, pp. 883–887, 2020.

[9] H. Kwon, “Detecting backdoor attacks via class difference in deep neural networks,” *IEEE Access*, vol. 8, pp. 191049–191056, 2020.

[10] H. Kwon and Y. Kim, “BlindNet Backdoor: Attack on Deep Neural Network Using Blind Watermark,” *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6217–6234, 2022.

[11] H. Kwon, “Defending Deep Neural Networks against Backdoor Attack by Using De-trigger Autoencoder,” *IEEE Access*, 2021.

[12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: lessons learned from the 2015 ms coco image captioning challenge,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 652–663, 2016.

[13] R. U. Khan, X. Zhang, R. Kumar, and E. O. Aboagye, “Evaluating the performance of resnet model based on image recognition,” in *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, pp. 86–90, New York, NY, United States, December 2018.

[14] T.-Y. Lin, M. Maire, S. Belongie et al., “Microsoft coco: common objects in context,” *Computer Vision - ECCV 2014*, pp. 740–755, 2014.

[15] M. Jahidul Islam Razin, M. Abdul Karim, M. F. Mridha, S. M. Rafiuddin Rifat, and T. Alam, “A long short-term memory (lstm) model for business sentiment analysis based on recurrent neural network,” *Sustainable Communication Networks and Application*, pp. 1–15, 2021.

[16] A. Sherstinsky, “Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network,” *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[17] J. C.-W. Lin, Y. Shao, Y. Djenouri, and U. Yun, “Asrnn: a recurrent neural network with an attention model for sequence labeling,” *Knowledge-Based Systems*, vol. 212, p. 106548, 2021.

[18] J. Redmon and A. Farhadi, “Yolo9000: better, faster, stronger,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, Honolulu, HI, USA, July 2017.

[19] P. Xia, L. Zhang, and F. Li, “Learning similarity with cosine similarity ensemble,” *Information Sciences*, vol. 307, pp. 39–52, 2015.

[20] E. V. Raghavendra, P. Vijayaditya, and K. Prabhallad, “Speech synthesis using artificial neural networks,” in *Proceedings of the Communications (NCC), 2010 National Conference on*, pp. 1–5, IEEE, Chennai, India, January 2010.

[21] Y. LeCun, C. Cortes, and C. J. Burges, “Mnist handwritten digit database,” AT&T Labs [Online]. Available, vol. 2, 2010, http://yann. lecun. com/exdb/mnist.

[22] R.B. Himmelstein, “Voice-controlled vehicle control system,” *US Patent*, vol. 6, p. 496, 2002.

[23] F.-Y. Leu and G.-L. Lin, “An mfcc-based speaker identification system,” in *Proceedings of the Advanced Information Networking and Applications (AINA), 2017 IEEE 31st International Conference on*, pp. 1055–1062, IEEE, Taipei, Taiwan, March 2017.

[24] L. Palen and M. Saltzman, “Voice-mail diary studies for naturalistic data capture under mobile conditions,” in *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, pp. 87–95, ACM, New York, NY, United States, December 2002.

[25] M. Abadi, P. Barham, J. Chen et al., “Tensorflow: a system for large-scale machine learning,” *OSDI*, vol. 16, pp. 265–283, 2016.
[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, New York, NY, United States, July 2002.

[27] J. Zhou, Y. Qiu, S. Zhu et al., “Optimization of support vector machine through the use of metaheuristic algorithms in forecasting thm advance rate,” Engineering Applications of Artificial Intelligence, vol. 97, p. 104015, 2021.

[28] M.-H. Laves, M. Tölle, A. Schlaefer, and S. Engelhardt, “Posterior temperature optimized bayesian models for inverse problems in medical imaging,” Medical Image Analysis, vol. 78, p. 102382, 2022.

[29] H. Kwon and K. Sung Hwan, “Restricted-area adversarial example attack for image captioning model,” Wireless Communications and Mobile Computing, vol. 2022, 2022.

[30] K. Hyun and S. Lee, “Friend-guard adversarial noise designed for electroencephalogram-based brain–computer interface spellers,” Neurocomputing, 2022.

[31] H.-Y. Tseng, L. Jiang, C. Liu, M.-H. Yang, and W. Yang, “Regularizing generative adversarial networks under limited data,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7921–7931, Nashville, TN, USA, June 2021.