Fraud Detection with Multi-Modal Attention and Correspondence Learning

Jongchan Park
Lunit Inc.
Seoul, Korea
jcpark@lunit.io

Min-Hyun Kim, Seibum Choi, In So Kweon
KAIST
Daejeon, Korea
{minhyun, sbchoi, iskweon77}@kaist.ac.kr

Dong-Geol Choi*
Hanbat Nat'l University
Daejeon, Korea
dgchoi@hanbat.ac.kr

***Abstract*— Deep learning based recognition systems have shown high performances in various tasks. Most of them are single-modality based, using camera inputs only, thus are vulnerable to look-alike fraud inputs. Fraud inputs may frequently be abused when rewards are given to the users, such as in reverse vending machines. Joint use of multi-modal inputs can be a solution to fraud inputs since modalities contain different information about the target task. In this work, we propose a deep neural network that utilizes multi-modal inputs with an attention mechanism and a correspondence learning scheme. With an attention mechanism, the network can learn better feature representation for multiple modalities; with the correspondence learning scheme, the network learns intermodal relationships and thus can detect fraud inputs where modalities do not correspond to each other. We investigate the proposed approach in a reverse vending machine system, where the task is to perform classification among 3 given classes (can, PET bottles, glass bottles), and reject any suspicious input. Three different modalities (image, ultrasound, and weight) are used. As a result, we show that our proposed model can effectively learn to detect fraud inputs while maintaining a high accuracy for the given classification task.**

Keywords—Deep Learning; Fraud Detection; Multi-modal

I. INTRODUCTION

Advances in deep learning [1], [2] have shown state-of-the-art performances in various recognition tasks [3], [4], [5]. Thanks to open-sourced deep learning frameworks, commercial applications [6], [7] based on deep learning are made possible. In many business models with recognition systems, the most important goal is to achieve high accuracy. However, preventing fraud inputs or adversarial attacks is also crucial in some business models such as for reverse vending machines [8], [9], because actual rewards will be given to users immediately.

As most recognition systems have a single camera modality as input, they are vulnerable to fraud inputs. An example is shown in Fig. 1. Several adversarial attack methods [10], [11], [12] have been proposed to 'fool' recognition systems using image inputs. Also, vulnerabilities in commercial recognition systems have been reported [13], [14]. A simple solution is to use multiple modalities as inputs, such as RGB images with depth images, IR images or ultrasound. Multiple modalities contain information complementary to each other by providing different aspects of information for the given task. So far, multi-modal recognition systems have been proposed for robustness against noise [15] and for better performance [16], [17]. In this work, we propose a multimodal recognition system for fraud detection. However, as shown in our experiments, naive combinations of multiple modalities may not fully enjoy the efficacy of complementary information. We propose two techniques, an attention mechanism and a correspondence learning, to combine multi-modal features for recognition and fraud detection using deep neural networks.

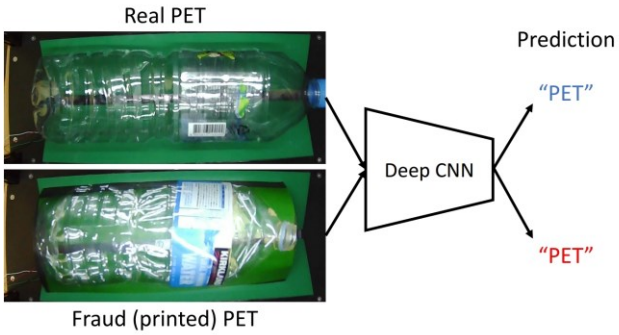

Fig. 1: An illustration for a fraud input. Two inputs are shown: a target class sample (PET bottle, above), and a fraud sample (printed PET bottle, below).

Attention methods have been used in visual question answering (VQA) tasks [18], [19] or image captioning tasks [20]. In the case of VQA tasks, the inputs are one image and one corresponding question. A common method is to aggregate the image features according to the attention gates produced from the question, and then the last classifier selects an answer. In such cases, the attention gate is generated for the image modality only, and so the attention is uni-directional. In this work, we exploit the correlation among modalities. We jointly use multiple modalities and generate attention gates for modalities.

Inspired by recent studies in self-supervised learning [21], [22], we use a correspondence learning scheme to further exploit multi-modal correlation. If multi-modal inputs are naively used, the neural network tends to exploit the most discriminative parts, and may not fully utilize multi-modal information. To learn the correlation among modalities, we

* Corresponding author

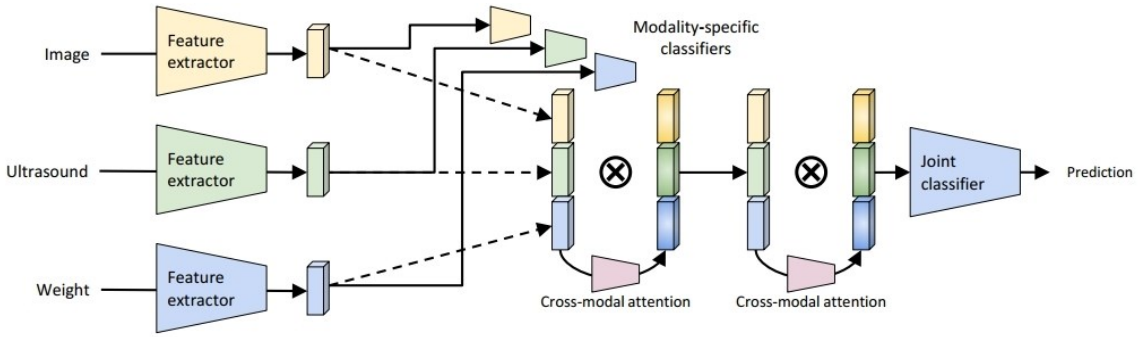

Fig. 2: Overall architecture for our multi-modal recognition system. We use modality-specific feature extractors to extract 1D features from multiple modalities. All extracted features are concatenated, refined with multi-modal attention, and finally fed into the joint classifier. In addition, each modality feature is trained with modality-specific classifiers.

create a synthetic class, called unmatched, in which the modalities are from different classes. When discriminating between unmatched class and matched classes (CAN, PET, GLASS), the neural network is encouraged to learn class related features as well as correlations among modalities. In this paper, we propose a multi-modal deep neural network for object recognition and fraud detection. As camera modality is frequently used as inputs, we focus on preventing look-alike frauds in reverse vending machine cases. Nevertheless, the idea can easily be extended to other situations. While naive joint learning may not fully utilize multi-modal information, we use two techniques, an attention mechanism and correspondence learning, to further exploit the correlation among modalities and learn better representations

The paper contributes in the following way:

1) We show that single modality based deep neural networks (DNNs) are vulnerable to fraud inputs and unseen class objects. When trained with known classes only, DNNs classify look-alike inputs and unseen objects with high confidence

2) We propose a multi-modal DNN with an attention mechanism and correspondence learning. We show that the proposed DNN maintains high classification accuracy and fraud detection rate.

3) We show that a DNN with non-contact ultrasound signals achieves high accuracy in material classification. We train and test the DNN with our own dataset of various shapes and poses.

We will introduce the target problem in Sec. II, our proposed method in Sec. III, the experimental setup including deep network setups and hardware setups in Sec. IV, and the experiment results and analysis in Sec. V; the conclusion will follow, including future directions.

II. Fraud Detection In Reverse Vending Machine

A. Reverse Vending Machine

A reverse vending machine (RVM) collects empty, recyclable containers from users and gives out rewards. There are several products in operation, such as TOMRA [8], RVM Systems [23] and Superbin [9]. Previous systems often use UPC or bar code scanners to specifically identify the incoming containers. However, such systems require a huge and up-to-date database of containers and cannot handle deformed (crumpled) containers for which UPC or bar codes are not identifiable. To handle such problems, we have built a simple vision-based system with deep convolutional neural networks for garbage classification; it has shown over 99% accuracy for classification. Previously built system use image inputs only and is vulnerable to fraud inputs such as lookalike samples.

Since an automated RVM gives back immediate rewards, it is crucial to not give a false positive classification. That is, to identify a non-target object as one of the target class. The system must reject any non-target inputs and ask the users to input target class objects. If the system accepts non-target objects, this vulnerability may be abused by malicious users, and can lead to huge loss to the company. It is a fundamental threat to the RVM business model.

B. Fraud Inputs

We define any malicious input that leads to misclassification as a fraud input. As stated in I, we focus on visually similar inputs as the fraud targets. Visually similar inputs include printed objects, as shown in Figure 1; non-target inputs include any random objects such as crumpled paper or plastic bags. Visually similar inputs exploit the uni-modality characteristic of the system. Many visual recognition systems depend solely on a single camera sensor, and can be easily fooled by visually similar inputs. There are famous failure cases of Facegate in Samsung Galaxy S8 [13], or Windows Hello facial recognition system [14]. A simple fix to this vulnerability is to jointly use multiple modalities and leverage the correlation among them.

C. Non-target Inputs

Non-target inputs can be easily rejected by thresholding the neural network output. However, the network output is a maximal likelihood prediction and is not an absolute confidence measure. Therefore, the network sometimes fails to reject non-target inputs. In this work, we show that joint training of multiple modalities also prevents such failures.

Details for the dataset acquisition will be described in Sec. IV-B. Captured samples are shown in Fig. 3.

Fig. 3: Captured images in the dataset. (a) is a target class sample of CAN. (b) is a visually similar fraud sample of printed CAN. (c) is a non-target sample of glove.

III. Proposed Method

We propose a multi-modal DNN using an attention method and correspondence learning. Multiple modalities contain rich information from different domains, but a naive use of multiple modalities cannot fully utilize the rich information. With the attention method and correspondence learning, the multi-modal network shows superior performance, as shown in the experimental results in Table III.

A. Deep Neural Network (DNN)

In order to build a powerful yet fast recognition system, we use different types of lightweight DNN-based feature extractors for input modalities, and a multi-layer perceptron (MLP) as the classifier. In short, the network output is computed as:

$$P(X_{img}, X_{us}, X_w) = MLP_{cls}\ (\ F_{img}(X_{img}), F_{us}(X_{us}), F_w(X_w)) \quad (1)$$

where F_{img}, F_{us} and F_w are feature extractors for image, ultrasound and weight inputs respectively, and MLP_{cls} is the final classifier.

1) Multi-modal Feature Extractor: We use ResNet18 [2] as the image feature extractor Fimg, stacked 1D convolutions for Fus, and an MLP for Fw. Image inputs are RGB images; ultrasound input is transformed into spectrograms; weight input is a one-hot encoded vector. Each feature extractor outputs a 1D feature vector for each input modality.

2) Joint Classifier: The final classifier for multi-modal features is also an MLP. Before the 1D features are fed into the classifier, we simply concatenate them into one feature vector. We apply the attention mechanism to the concatenated feature vector to compute better representations.

3) Modality-specific Classifier: Additionally, we use multi-task method for each modality. As shown in Fig. 2, there are auxiliary classifiers for image, weight, and ultrasound inputs. In this way, we can ensure that each modality feature contains its own information for the given task.

Details for each part will be addressed in Sec. IV-A.

B. Attention Method

Self-attention methods [24], [25] have recently been proposed for better representation learning. Self-attention refines feature maps by increasing or suppressing the scales of certain features. While previous methods have investigated the single modality case, we utilize the self-attention method for inter-modality and inter-channel re-calibration. The attention values for each modality are calculated from inputs consisting of each modality itself, and the two other modalities as well. As an equation, the attention process can be described as below:

$$F' = F + F \otimes \sigma(M_{att}(F)) \quad (2)$$

where M_{att} is an MLP for generating attention masks and F is the concatenated multi-modal feature. The attention mask is normalized with a sigmoid layer, and the attention weighted feature is added to the original feature map. Then the feature maps are weighted between 1.0 and 2.0. This kind of residual style attention is applied for stable gradient propagation.

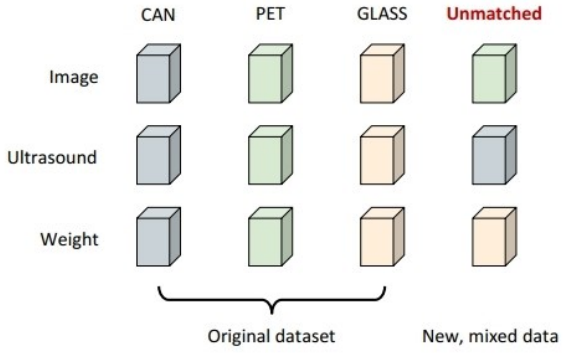

Fig. 4: A synthesized dataset for correspondence learning.

C. Multi-Modal Correspondence Learning

A fraud input is a fundamental threat to single-modality recognition systems: for example, visually similar can often fool image-based recognition systems. We can detect fraud inputs by examining the inter-modal correlations. To explicitly learn the correlation among modalities, we apply correspondence learning during network training.

In correspondence learning, we want to exploit the 'correspondence' among modalities. That is, target class inputs have correspondences among modalities: for example, an aluminum CAN shows common visual characteristics of CANs and material characteristics of aluminum at the same time. With such intuition, we want the proposed DNN to classify the input object as one of the target classes only when the modalities correspond.

As shown in Fig. 4, we synthesize an extra class named 'unmatched', for which the image, ultrasonic signals and weight are not from the same class. As this is a simple mixture of existing data, no extra data are required. By training with the original matched classes and the unmatched class, we explicitly train the DNN to learn not only the original matched class but also the correspondence of the input modalities.

IV. Experiment Setups

A. Network Settings

The network design has three parts: feature extractors, attention layers, and the classifiers. Also, if not otherwise specified, the networks are trained with the Adam optimizer, with learning rate $1e^{-4}$.

1) Feature Extractors: For the image modality, we use ResNet-18 up to stage4 as the feature extractor, followed by a global average pooling layer. A linear layer is added at the end, and outputs a 1D feature vector of size 512. For the ultrasound modality, we use time-frequency data in the range (30kHz, 50kHz) as input. The feature extractor consists of four 1D convolutions with (kernel size, stride) of [(201, 5), (51,1), (51,1), (51,1)] with ReLU. Similar to the image feature extractor, a linear layer is added at the end, and outputs a 1D vector of size 512. For the weight modality, we use one-hot encoding where the bin size is 3g per bin and maximum weight is 600g. The feature extractor consists of 3 linear layers where the hidden sizes are [512, 512, 512] with batch normalization and ReLU.

2) Attention Module: Features from different modalities are fed into the attention layers for feature refinement. The attention layer generates gates for each modality feature. The attention module consists of 3 linear layers with output sizes [1536, 1536, 1536, 1536]. The last output is normalized with a sigmoid layer. The normalized output is divided into 3 vectors, and the 3 vectors are regarded as attention vectors for 3 modalities. Each modality feature vector is multiplied with the attention vector for feature refinement.

3) Classifiers: The refined modality features are concatenated and fed into the joint classifier. The joint classifier is a 4-layer MLP with batch normalization and ReLUs, with hidden sizes of [768,768,768], the last output is the number of target classes. When we use the synthesized unmatched class, one extra class is added. In order to train each modality feature well, we also assign separate classifiers for different modalities. In this way, even when an unmatched class instance is fed into the network, we can train separate branches with the real labels of each type of modality data. Each modality-specific classifier contains 3 linear layers with hidden sizes of [256,256]; the last output is the number of target classes. For modality-specific classifiers, we cannot use the extra unmatched class.

Since neural networks are not usually designed for fraud detection, we use a heuristic method of fraud detection. Neural networks are usually trained with known classes, and are not aware of unseen class instances. In classification networks, the output is softmax normalized, and the answer is the maximum-likelihood output. In order to detect fraud inputs, we have used a heuristic threshold for the likelihood: when the maximum-likelihood output is below the threshold, we regard the input as a fraud input. In addition, in cases in which an unmatched class is used for training, the test inputs classified as unmatched class are also regarded as fraud inputs.

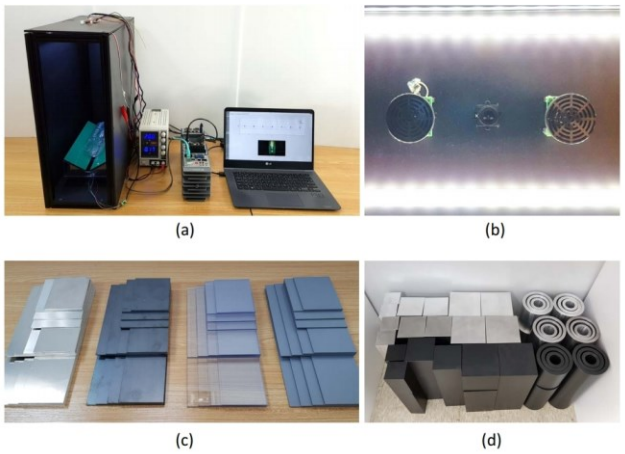

Fig. 5: The hardware setup for data acquisition and raw materials used for the material database. (a) is the overview of the setup. From left, there are the box for object placement, the power supply, the controller and the laptop. Ultrasonic, camera (RGB), load cell sensors are attached to the box. (b) shows the inner-upper side of the box, where LED bars, ultrasonic transmitter/receiver and the camera sensor are attached. (c) are in flat shapes and (d) are in cuboid and cylinder shapes.

B. Hardware Settings and Data Acquisition

In this section, we introduce the data acquisition system and the types of databases for our experiments.

1) Sample objects for databases: To build the databases for our multi-modal classification task, we acquire sensor inputs from various objects using ultrasonic, camera and load cell sensors. There are two types of databases: the raw material database in which the target objects have the same shapes and different material types, as shown in Fig. 5 (c)(d); and the real object database, in which the target objects are real world objects including our target class objects (can, PET bottles, glass bottles), fraud inputs, non-target, as shown in Fig. 3. The raw material types are stainless steel, aluminum, poly-carbonate, and polyvinyl chloride. To learn material features that are robust to sizes and shapes, we make the objects for the raw material database in various shapes and sizes. We use 3 shapes: flat, cuboid, and cylinder. Flat shapes have compositions of width 80, 100, 120 and 140mm, height 100, 200, and 300mm, and 3T thickness. Cuboids are compositions of square bases with 50, 75, and 100mm sides and 100, 200, 300mm height. Cylinders are compositions of circle bases with 50, 75, and 100mm diameters and 100, 200, and 300mm heights.

We collected as many real world samples as possible to ensure the diversity of the target class objects. We use 167 cans, 141 PET bottles and 228 glass bottles as the target class objects. In addition to the target class objects, we made a simple fraud input dataset by printing out the target class objects. As shown in Fig. 1, the printed objects are realistic enough to 'fool' a deep neural network system. We collected 60 fraud examples for evaluation purpose. For non-target data,

we randomly collect 29 miscellaneous objects around, such as paper cups, gloves, plastic bags, human arms or clothes.

2) Hardware setup: The hardware setup for data acquisition is as shown in Fig. 5 (a)(b). We use a single pair of transmitter/receiver ultrasonic sensors (HG-M40TN2/HGM40RN2, Hagisonic), a USB webcam sensor and a 5kg load cell sensor. We use a controller (compactRIO-9036, National Instruments) to trigger and receive raw signals of the ultrasonic and load cell sensors. We trigger the ultrasonic sensor transmitter every 200ms and record the raw input in the receiver at 1 mega samples per second. We record the load cell signal simultaneously. We acquire the image data with the USB webcam. All control is done on the laptop computer. The controller and the USB webcam are connected to the laptop.

V. Experiment Results

A. Raw Materials with Ultrasound

In this section, we use a single pair of non-contact ultrasonic sensors and 1D CNN to show that raw material classification is viable, especially when the objects are in various shapes and poses. In our target task of reverse vending machines, object material is important. It has been shown that material classification is viable with ultrasonic signals [26], [27], so we decide to use ultrasonic sensors as a new modality.

However, the experiments conducted in [26], [27] are highly controlled in that the target objects are flat board shapes with the same pose and distance from the sensors. In real world cases, the target objects are in various shapes, sizes and poses. Therefore, we need to show that ultrasonic signals still contain enough information in such challenging cases.

TABLE I. Raw material classification with ultrasound

2D shapes	
Material type	***Accuracy(%)***
Acryl	100.0
Aluminium	100.0
Iron	100.0
Plastic	96.0
Avg acc	99.0
3D shapes	
Material type	***Accuracy(%)***
Aluminium	100.0
Plastic	91.6
Iron	91.8
Avg acc	94.4

In the experiment, we use the raw material dataset acquired in Section IV-B with various shapes, sizes, and poses. The feature extractor and the classifier are the same as the ones specified in Sec. IV-A. According to the result in Table I, we empirically verified that material classification is possible with a single pair of non-contact ultrasonic sensors and a 1D CNN.

B. Fraud Detection using Real World Data

In this section, we show that our proposed model can learn to classify target inputs and detect fraud inputs. First, we show the effects and the limitations of naive use of multimodal inputs. Next, we show that the two proposed techniques achieve high fraud detection rate while maintaining high accuracy for target class objects.

a) Multi-modal Inputs: When multiple modality inputs are used together, we expect a better performance of DNNs in general. As shown in Table II, joint use of multiple modalities can achieve higher fraud detection rate for both visually similar fraud inputs and non-target inputs. The change in target class object accuracy is negligible. In terms of fraud detection rate (visually similar fraud inputs and non-target inputs), the efficacy of multiple modality can be observed. Fraud inputs are all unseen classes for DNNs, and the features are different from those of target class objects. We conjecture that multi-modal inputs will show more differences in features, compared to single modal cases. Therefore, multi-modal inputs achieve a higher fraud detection rate.

TABLE II. Classification results using multi-modal inputs in real world database. W denotes the weight modality. Target denotes the accuracy in target class objects in Fig. 3, fraud denotes the fraud detection rate for visually similar fraud inputs, non-target denotes the fraud detection rate for nontarget inputs.

Modality	***Target (%)***	***Fraud(%)***	***Non-target(%)***
Image (IMG)	98.0	8.3	6.9
Ultrasound (US)	82.3	15.0	6.9
IMG + US	96.5	15.0	6.9
IMG + US + W	97.5	18.3	13.7

TABLE III. Classification results using correspondence learning and multi-modal attention in real world database. CL denotes correspondence learning and Att denotes multimoda.

Modality	***CL***	***Att***	***Target (%)***	***Fraud(%)***	***Non-target(%)***
IMG+US+W			97.5	18.3	13.7
IMG+US+W		✓	99.5	21.7	20.7
IMG+US+W	✓		81.8	86.7	93.1
IMG+US+W	✓	✓	94.0	91.7	93.1

b) Multi-modal Attention: The purpose of multi-modal attention is to refine the concatenated multi-modal features by self-attention. As all modality feature vectors are used to generate attention masks for each other, we expect the network to learn better representations. As shown in Table III, multi-modal attention achieves a higher target class accuracy, a higher fraud detection rate for both visually similar fraud inputs and non-target inputs. Generally improved performance indicates better representations. For fraud inputs, the inter-modal relationships are different from those of target class objects. As for attention jointly use multi-modal features, we suspect that the network detects fraud inputs using the changes in the inter-modal relationship.

c) Correspondence Learning: Correspondence learning explicitly trains the network to learn the correlation among modalities, and a much higher fraud detection rate is achieved, as shown in Table III. However, there is a decrease in target class accuracy. We argue that the fraud detection rate is improved because the classifier has learned the correlation between modalities through correspondence learning. As fraud detection is crucial to the RVM business model, large improvement on fraud detection rate is remarkable. The result is compliant with the results from [21], as the network learns to accept modality-matched inputs and reject modality-unmatched inputs. Fraud inputs have unmatched modalities since visually similar inputs have visual appearance of various classes, but does not have matched ultrasonic or weight inputs.

d) Final Model: Finally, we combine all the techniques. The last row of Table III is the final model we propose, in which all the proposed techniques are used. It achieves high accuracy with high fraud detection rate. When correspondence learning is used, the fraud detection rate becomes very high, while the target class accuracy is the most compromised value. The attention mechanism improved the fraud detection rate while maintaining the target class accuracy. We argue that this is due to better feature learning resulting from the multi-modal attention mechanism. When the two techniques are combined, the final model preserves high accuracy while detecting most of the fraud inputs. This is a remarkable improvement since fraud examples are hard to distinguish using only visual modality only, as shown in Fig 1 and Table II.

Lower target class accuracy may be a concern, but a slight compromise is not a problem in the reverse vending machine task. Most mis-classifications are classified as 'unmatched', and users will be asked to try again. As the accuracy is 94%, the next trial is highly likely to be successful.

VI. Conclusion

In this paper, we have proposed a multi-modal DNN with attention mechanism and correspondence learning for object recognition and fraud detection. As single-modal systems are fundamentally vulnerable to fraud inputs, we utilize multimodal inputs. While a naive joint use of multi-modal features cannot fully enjoy the efficacy of multi-modal information, the two proposed techniques, multi-modal attention and correspondence learning, increase the fraud detection rate and preserve high target class accuracy.

The proposed techniques are lightweight and simple. Both can be easily integrated into any DNN-based multi-modal systems, and jointly trained end-to-end. Also, no extra data are required. We are planning to integrate this mechanism in commercial reverse vending machines.

Recently, many adversarial attacks are proposed against DNNs. Most of them target single-modal, image-only systems, but they can be extended to multi-modal cases. In this paper, we have only investigated physical visually similar fraud inputs and non-target inputs. In future works, we will extend our study to fraud detection mechanisms against adversarial attacks. Figures and Tables

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 248–255.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in ´ context," in European conference on computer vision. Springer, 2014, pp. 740–755.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.

[6] "Google cloud vision," https://cloud.google.com/vision/, accessed: 2017-12-21.

[7] "Papago," https://papago.naver.com/, accessed: 2017-12-21.

[8] "Tomra," https://www.tomra.com/en/, accessed: 2017-12-20.

[9] "Superbin," http://www.superbin.co.kr/new/index.php, accessed: 2017-12-20.

[10] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," ICLR Workshop, 2017. [Online]. Available: https://arxiv.org/abs/1607.02533

[11] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Query-efficient black-box adversarial examples," 2017. [Online]. Available: https: //arxiv.org/abs/1712.07113

[12] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," 2017. [Online]. Available: https: //arxiv.org/abs/1707.07397

[13] "Video shows galaxy s8 facial recognition tricked by a photo," http://www.gizmodo.co.uk/2017/03/ video-shows-galaxy-s8-facial-recognition-tricked-by-a-photo/, accessed: 2018-02-04.

[14] "Specially prepared photos shown bypassing windows hello facial recognition," https://www.youtube.com/watch?v=Qq8WqLxSkGs, accessed: 2018-02-04.

[15] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. IEEE, 2015, pp. 681–687.

[16] J. Hoffman, S. Gupta, J. Leong, S. Guadarrama, and T. Darrell, "Crossmodal adaptation for rgb-d detection," in Robotics and Automation (ICRA), 2016 IEEE International Conference on. IEEE, 2016, pp. 5032–5039.

[17] Y. Bian, C. Gan, X. Liu, F. Li, X. Long, Y. Li, H. Qi, J. Zhou, S. Wen, and Y. Lin, "Revisiting the effectiveness of off-the-shelf temporal modeling approaches for large-scale video classification," arXiv preprint arXiv:1708.03805, 2017.

[18] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in Advances In Neural Information Processing Systems, 2016, pp. 289–297.

[19] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in International Conference on Machine Learning, 2015, pp. 2048–2057.

[21] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in ´ IEEE International Conference on Computer Vision, 2017.

[22] R. Arandjelovic and A. Zisserman, "Objects that sound," ´ arXiv preprint arXiv:1712.06651, 2017.

[23] “Rvm systems,” http://www.reversevending.co.uk/, accessed: 2018-02-04.

[24] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” CoRR, vol. abs/1709.01507, 2017. [Online]. Available: http://arxiv.org/abs/1709.01507

[25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[26] K. Ohtani and M. Baba, “A simple identification method for object shapes and materials using an ultrasonic sensor array,” in 2006 IEEE Instrumentation and Measurement Technology Conference Proceedings, April 2006, pp. 2138–2143.

[27] Y. Moritake and H. Hikawa, “Category recognition system using two ultrasonic sensors and combinational logic circuit,” Electronics and Communications in Japan (Part III: Fundamental Electronic Science), vol. 88, no. 7, pp. 33–42, 2005