An Adversarial Perturbation Technique against reCaptcha Image Attacks

Lawankorn Mookdarsanit [1] and Pakpoom Mookdarsanit [2]

Department of Computer Science, Faculty of Science at Chandrakasem Rajabhat University[1],
Department of Business Computer, Faculty of Management Science at Chandrakasem Rajabhat University[2]
E-Mail: lawankorn.s@chandra.ac.th, pakpoom.m@chandra.ac.th

Received: *October 14, 2020*
Revised: *January 4, 2020*
Accepted: *May 11, 2020*

ABSTRACT

Deep learning has a great success in object recognition accuracy since 2012. Along with the dark world, deep learning can be misleading as the threat of reCaptcha attacks. A hacker demonstrated to generate the AI-based bots using Convolutional Neural Network (CNN) to recognize the reCaptcha images as human's perception; and be authorized to access the business operation of information system. This activity shows that an AI-based bot (or non-human) can easily break the Challenge-Response authentication protocol. In this paper, "CNN-based object recognition" meets "cyber security". The reCaptha attack defense is proposed by adding some adversarial perturbation (or noise) to the image. The perturbation can fool those AI-based bots to misclassify the objects within reCaptcha images that the bots cannot access the system. From the adversarial perturbation test, one-stage detection has more robust than two-stage one. Furthermore, the ResNet overcomes other architectures in overall score that can be used in ether one-stage or two-stage detection.

Key Words : Perturbated image, Deep learning, Authentication failure, reCaptcha attack, Adversarial perturbation

Introduction

The "Completely Automated Public Turing test to tell Computers and Humans Apart" (de facto "Captcha") refers to a human identification mechanism using human's literacy to prove that the user is a human or not Captcha is a form of the challenge-response authentication protocol (Soimart and Mookdarsanit, 2016) to protect the accessibility of information system's service and resource accessibility (Ahn *et al*., 2003). By the access prevention, bots or spams (Non-humans) are unable to identify the objects or texts and they are finally unauthorized and blocked. Originally, the first Captcha was a text-based authentication in 1996 (Naor, 1996).

With the intelligence of optical character recognition (OCR), hackers could use OCR to teach the spams to read, beyond a joke as "SPAMS READ BOOKS". It was shown that non-humans could easily access the web service. To this end, text-based Captcha iseemed to be not secured (Mori and Malik, 2003).

For 10 years later, reCaptcha (Ahn *et al*., 2008). was introduced to the world. Instead of human's literacy, human's vision was authenticated by visual image perception, under the slogan "EASY ON HUMANS, HARD ON BOTS".

(a.) Textual Captcha (b.) Image reCaptcha (object as a car)

Figure 1 A human identification mechanism

As well as the revolution of object detection since 1998 (Zou *et al*., 2019), it was a successful intelligence in many specific tasks: facial landmark detection (Soimart and Mookdarsanit, 2016; Lee, Kim *et al*., 2019), medical analysis (Jakimovski and Davcev, 2018; Sutthaluang, 2018), food recognition (Soimart and Mookdarsanit, 2017; Mookdarsanit and Mookdarsanit, 2018)., action recognition (Mookdarsanit and Mookdarsanit, 2018), tourism classification (Soimart and Mookdarsanit, 2017; Mookdarsanit and Mookdarsanit, 2018), plant identification (De Luna *et al*., 2018; Mookdarsanit and Mookdarsanit, 2019) or tracking fishes (Marini *et al*., 2018; Mookdarsanit and Mookdarsanit, 2019), etc. The milestone of object detection can be shortly divided into 2 eras: Scale-invariant Feature Transform (a.k.a. SIFT) during 1998-

2011 and Convolutional Neural Network (a.k.a. CNN) during 2012-present (Zheng *et al*., 2018). In 2012, AlexNet (Krizhevsky *et al.*, 2012). – the first version of deep learning for image recognition, made a great improvement that leveraged CNN to accurately recognize the objects within an image (Alom *et al*., 2018). CNN totally outperformed traditional SIFT-based methods (Liu *et al*., 2019) in term of higher recognition accuracy.

As a threat for reCaptcha, CNN was applied to a serious threat on reCaptcha; an AI-based bot is able to get authorized by its intelligent perception and access the system like a human's identification, known as "reCaptcha image attack". This paper provides a defensive strategy to prevent all form of reCaptcha image. By fooling the bot, an image can be added some noise (de facto "**adversarial perturbation**") that it does not affect the human's perception.

The novelty of this paper is to 1) test/measure the reCaptcha image recognition robustness for all CNN detection pipeline under different models by adding adversarial perturbed noise and 2) provide overall benchmarking scores for those CNN types.

This paper is organized into 6 parts. The second part describes threat of deep learning. CNN detection pipeline and CNN architecture are in part 3 and 4. The tit-and-tat solution is explained in part 5 as defense by adversarial perturbation. And part 6 is conclusion.

Threat of Deep Learning

Deep learning is a workhorse for computer vision that has been growing rapidly in the field of object recognition, known as convolutional neural network (CNN). The first CNN version was originally proposed by LeCunn in 1989 (LeCun *et al*., 1989; LeCun and Bengio, 1995) but the computing resource was not enough in that time that made CNN was not so popular. CNN rebirthed again in 2012 (Krizhevsky *et al*., 2012) under the age of full high GPU performance computing, called AlexNet that achieved the accuracy record in the large scale visual recognition challenge (ILSRVC). In view of cascaded detection pipeline, the feature map shared computation is so powerful and deep learning era has begun in the world's timeline since 2012. By CNN mechanism as a magical human's perception, an AI-based bot is able to masquerade as a human and authorized to access the system resource. Furthermore, many million AI-based bots can be quickly forked beyond the maximum workload capability that makes the sabotage of cloud-based provider called service unavailability (Mookdarsanit and Gertphol, 2013).

CNN Detection Pipeline

Deeply, CNN pipeline can be grouped into two-stage detection and one stage detection. Generally, two-stage detection has higher accuracy but takes more time for the detection.

Two-stage Detection

Two-stage detection is also called region-based framework detection that region proposals are generated from an image as feature extraction.

1. **R-CNN** (Girshick *et al*., 2014). – after AlexNet trained by ImageNet, a set of object proposals (candidate boxes) by selective search (van de Sande *et al*., 2011) was used known as "Region Convolutional Neural Network (R-CNN)". The input image is fixed size. And the linear Support Vector Machine (SVM) was used to predict the objects (Cortes and Vapnik, 2001). R-CNN is a multi-stage pipeline which has a numerous computational region proposals. And SVM training and testing are really expensive and too slow. This detection framework is seldom use nowadays.

2. **SPPNet** (He *et al*., 2015). – Since R-CNN needed the fixed size of image. "Spatial Pyramid Pooling Network (abbreviated by SPPNet)" used any arbitrary sizes of an image that could be generated a fixed-length representation of region proposals. The accuracy is little better than R-CNN. And it inherited from R-CNN with some disadvantages, particularly a multi-stage training.

3. **Fast RCNN** (Girshick, 2015) – it is faster than R-CNN for 200 times and higher accuracy than SPPNet by ignoring region proposals and no non-volatile storage allocated for feature caching like SPPNet. This framework firstly designed region of interest (RoI) pooling for warping at feature level before classification. The main drawback is RP computation is just an external unit and it is still a processing time problem.

4. Faster RCNN (Ren *et al*., 2015). – improved by some RCNN and Fast RCNN inventers that proposed a state-of-the-art region proposal network to generate the region proposals, multi-reference detection to detect features in multi-scale framework. The speed of Faster RCNN is better than that of Fast RCNN which makes Faster RCNN replace the mechanism of Fast RCNN. Although some little improvements like RFCN (Dai *et al*., 2016) and Light-head RCNN (Li *et al*., 2017), the training of Faster RCNN is still complex. Nowadays, Faster RCNN is a powerful mechanism for hackers to generate the million AI-based bots to the system.

5. FPN (Lin *et al*., 2017). – an acronym for “Feature Pyramid Network”, the same basis of Faster RCNN in 2017 that add feature fusion in object detection to provide more invariance and equivariance properties that is robust for scale and position changes. The new version of mobile phone (during 2018-2019) also has the feature fusion that uses the previous same 8 images taken from the camera; to produce a new high resolution of image.

One-stage Detection

Since two-stage detection consumes high computational resource and less time in mobile devices. One-stage detection has no region-based framework. The unified framework directly predicts the object from the global images.

1. **YOLO** (Redmon *et al*., 2016) – proposed in 2016 as the first one-stage framework that faster than FPN. Since one-stage ignored region proposals that has the accuracy lower than Fast RCNN. YOLO is abbreviated by “You Took Only Once” that has another 2 enhanced versions by a YOLO version-1 researcher (Redmon and Farhadi, 2017; Redmon and Farhadi, 2018). However, YOLO has some disadvantages like detection and localization of small objects within an image.

2. **SSD** (Liu *et al*., 2016) – the full name as Single Shot Detection that said to the world in 2016. SSD is an enhanced YOLO that combined the concept of region proposal network to perform multi-scale detection called resolution detection. Furthermore, Bootstrap with loss function was provided to solve the imbalance between objects and background known as hard negative mining problem. The speed and accuracy of SSD are better than Faster RCNN. Compared to YOLO, SSD provides more accuracy but takes more time than some YOLO versions.

Revolutionarily, these deep learning pipelines in object detection are available to mislead reCaptcha attacks. And a large number of AI-based bots can detect the objects as human’s perception and be authorized to access the system as well as human.

CNN Architecture

After detection pipeline, the feature map is input to train or test at classification layers. Training refers to input all images with their class (e.g., a car image with a class “car”) to teach the learning model. Testing is an unknown image is input to the model to classify the name of unknown object, as shown in Figure 2.

As referred to the architecture network, the different learning engines produce the different recognition accuracy results. Most architecture models are already trained by the large scale image dataset called "**pre-trained CNN**" (Wang and Deng, 2018) using transfer adaptation learning from CIFAR-10 dataset (Zhang, 2019). CIFAR-10 has 60,000 images that cover 10 classes: air-plane, car, bird, cat, deer, dog, frog, horse, ship and truck, respectively.

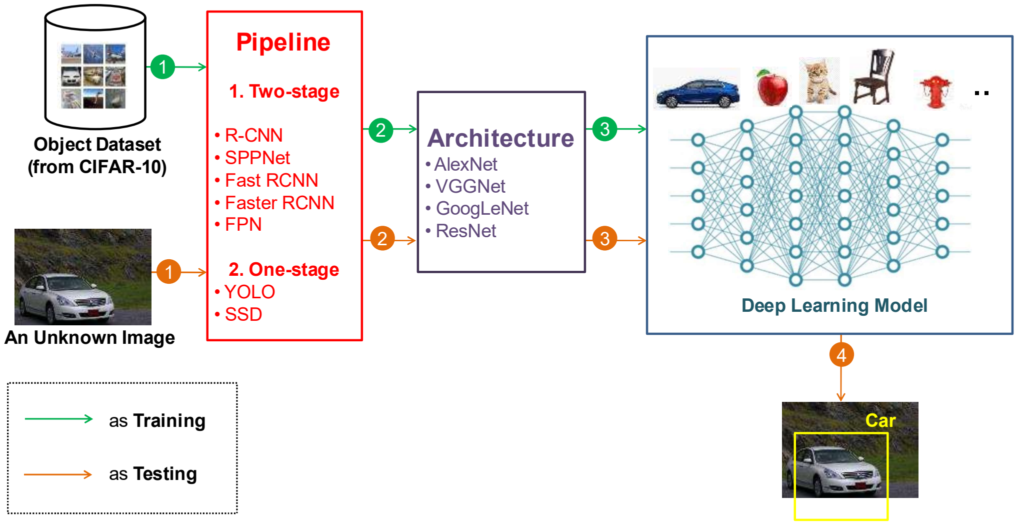

Figure 2 A workflow of CNN learning: training and testing

1. **AlexNet** (Krizhevsky *et al*., 2012) – firstly published in 2012 that showed the powerfulness of deep learning to the world.

2. **VGGNet** (Simonyan and Zisserman, 2014) – proposed by Oxford's Visual Geometry Group (VGG) in 2014 that increased the depth with small 3X3 convolutional filters.

3. **GoogLeNet** – (known as "Inception") designed by Google (Szegedy *et al*., 2016) that was a competitor with VGGNet and had many versions (Szegedy *et al*., 2017). The concept is to produce different sizes of filters with factorizing convolution (Szegedy *et al*., 2015) and batch normalization (Szegedy, 2015).

4. **ResNet** (He *et al*., 2016) – an acronym for "Residual Neural Network". Since ResNet use residual connections and deeper layers but fewer parameters. It is easy for network training by reference to the input layer. Actually, ResNet won the computer vision competitions in 2015 (before the CVPR'16 publication).

Defense by Adversarial Perturbation

As a matter of fact, adversarial perturbation is proposed to discover the weakness of Convolutional Neural Network (CNN). CNN can be fooled by adding adversarial perturbation in form of small Gaussian noise to a clean image but remain imperceptible to human's perception (Akhta and Mian, 2018). To give a real world example, a road sign attack to fool the CNN in autonomous vehicles. Such a traffic light as red (means "stop") but it is added some invisible adversarial perturbation to fool the deep learning as "green" and continue running. This case totally affects the human's physical security. For a tit-and-tat solution, the adversarial perturbation can be leveraged to fool those AI-based bots that are trained by CNN. We can say that ***CNN plays a role as BADDIE; adversarial perturbation is the HERO.***

In deep learning, hacker can use a pre-trained CNN models to generate many AI-bots to access the web system. This paper use some noise as adversarial perturbation added to the clean image for fooling those AI-based bots. The Universal Perturbations for Steering to Exact Target (UPSET) (Sarkar *et al*., 2017) is added to the clean image that human still understand the object within an image by (1).

$$img_{perturbated} = \max(\min(\ S \bullet R(t) + img_{clean}, 1), -1) \qquad (1)$$

where img_{clean} is a clean image $R(.)$ is a perturbation function to produce a perturbated image ($img_{perturbated}$) and t is a parameter that takes a target class, $R(t)$ is generated from a clean image, S is a scalar, all arguments are normalized to [-1, 1] to restrict the object transformation.

From Figure 3, a car as a clean image is perturbated by UPSET with the fooled target as whale. The noise is added to the clean image to fool the deep learning. With the help of min and max functions, a perturbated image seems to be a car by human's perception.

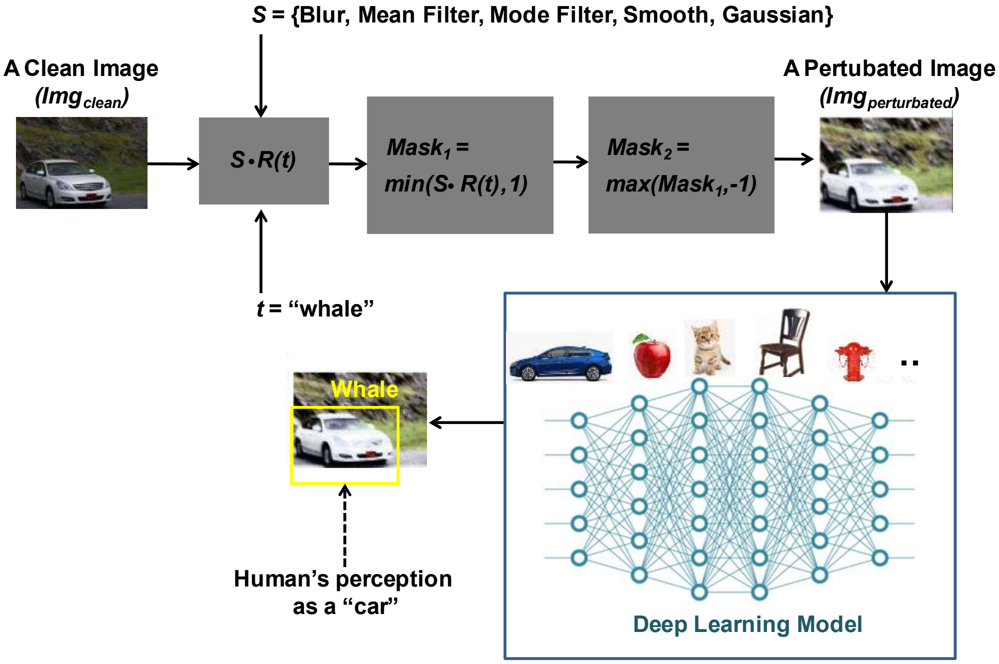

Figure 3 Deep learning fooled by UPSET perturbation

For the experiment, the pre-trained CNNs are penetrated and measured by recognition accuracy rate.

Table 1 The recognition rate of CNNs

		Architecture			
		AlexNet	VGGNet	GoogLeNet	ResNet
Detection Pipeline	R-CNN	0.00	-	-	-
	SPPNet	0.00	0.00	-	-
	Fast RCNN	0.00	0.22	-	-
	Faster RCNN	-	1.77	2.94	2.50
	FPN	-	-	-	3.21
	YOLO	-	-	0.00	1.91
	SSD	-	0.25	1.13	2.36

All CNN models are fooled by UPSET higher than 95%. If AI-based bot is generated by R-CNN, SPPNet and Fast RCNN, they are easily fooled by adversarial perturbation. Region proposal network (RPN) in both one-stage or two-stage detection that is included in Faster RCNN, FPN and SSD are still strong in some images, especially in the single basic shaped objects like a single glass, box, egg, apples, etc. Moreover, some complex-structured objects with some partial content also can block the CNN recognition like bicycle, crosswalk, car, palm, etc.

Conclusion

This paper proposes a defense mechanism of reCaptcha image attacks. Convolutional Neural Networks (CNN) plays a role as baddie. As adversarial perturbation is the hero. Since CNN can be applied to build an AI-based bot to recognize the object within a reCaptcha image as human identification. This is a serious security threat that violates the authentication, authorization and accessibility. To this end, the adversarial perturbation (or noise) is added to the reCaptcha image which proposes to fool the AI-based bot's recognition. The perturbation used in this paper is "The Universal Perturbations for Steering to Exact Targets (de facto UPSET)". We tested the CNN-based detections (R-CNN, SPPNet, Fast RCNN, Faster RCNN, FPN, YOLO and SSD) under different architectures (AlexNet, VGGNet, GoogLeNet and ResNet) as well as the calibrated recognition by the AI-based bots. From the result, perturbated images generated by UPSET can defend the reCaptcha attack higher than 95%.

Reference

Ahn, L.v., Blum, M., Hopper, N.J., and Langford, J. (2003). CAPTCHA: Using Hard AI Problems for Security. In *International Conference on the Theory and Applications of Cryptographic Techniques*. (pp. 294-311). Warsaw, Poland: Springer .

Ahn, L.V., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008). *reCAPTCHA: Human-Based Character*. Pittsburgh, Pennsylvania, USA.

Akhtar, N., and Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6, 14410-14430.

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S. (2018). The History Began from AlexNet: *A Comprehensive Survey on Deep Learning Approaches*, *arXiv*: 1803.01164 .

Cortes, C., & Vapnik, V. (2001). Support-Vector Networks. *Machine Learning*, 273–297.

Dai, J., Li, Y., He, K., and Sun, J. (2016). R-FCN: Object Detection via Region-based Fully Convolutional Networks. In *The 30th Annual Conference on Neural Information Processing Systems* (pp. 379–387). Barcelona.

De Luna, R.G., Baldovino, R.G., Cotoco, E.A., De Ocampo, A.L., Valenzuela, I.C., Culaba, A.B. (2018). Identification of philippine herbal medicine plant leaf using artificial neural network. In *The 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management* (pp. 1-8). Manila, the Philippines: IEEE.

Girshick, R. (2015). Fast R-CNN. In *The 2015 IEEE International Conference on Computer Vision* (pp. 1440-1448). Santiago, Chile: IEEE.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *The 2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580-587). Columbus, Ohio: IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *The 2016 IEEE Conference on Computer Vision and Pattern Recognition*. (pp. 770-778). Las Vegas, NV: IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37(9)*, 1904-1916..

Jakimovski, G., and Davcev, D. (2018). Lung cancer medical image recognition using Deep Neural Networks. In *The 13th International Conference on Digital Information Management* (pp. 1-5). Berlin, Germany: IEEE.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In "*The 26th Conference on Neural Information Processing Systems* (pp. 1-9). Lake Tahoe, Nevada.

LeCun, Y., and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *Holmdel*, New Jersey, USA.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E. and Hubbard, W. (1989). Backpropagation applied to handwritten zip code recognition, *Neural computation*, *1(4)*, 541-551.

Lee, H.J., Kim, S.T., Lee, H., and Ro, Y.M. (2019). Lightweight and Effective Facial Landmark Detection using Adversarial Learning with Face Geometric Map Generative Network, *IEEE Transactions on Circuits and Systems for Video Technology*, *30(3)*, 771-780.

Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., and Sun, J. (2017). Light-Head R-CNN, In *Defense of Two-Stage Object Detector. arXiv: 1711.07264*.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature Pyramid Networks for Object Detection. In *The 2017 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 936-944). Honolulu, Hawaii, USA: IEEE.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J. and Liu, X. (2019). Deep Learning for Generic Object Detection. *A Survey. International Journal of Computer Vision*, 1-58.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. and Fu, C.Y. (2016). SSD: Single Shot MultiBox Detector. In *The 14th European Conference on Computer Vision* (pp 21-37). Amsterdam, The Netherlands: Springer.

Marini, S., Fanelli, E., Sbragaglia, V., Azzurro, E., Fernandez, J. D. R., & Aguzzi, J. (2018). Tracking fish abundance by underwater image recognition. *Scientific reports*, 8(1), 1-12.

Mookdarsanit, L., and Mookdarsanit, P. (2019). SiamFishNet: The Deep Investigation of Siamese Fighting Fishes. *International Journal of Applied Computer Technology and Information Systems*, 40-46.

Mookdarsanit, L., and Mookdarsanit, P. (2019). Thai Herb Identification with Medicinal Properties Using Convolutional Neural Network. *Suan Sunandha Science and Technology Journal*, 34-40.

Mookdarsanit, P. and Gertphol, S. (2013). Light-weight operation of a failover system for Cloud computing. In *The 5th International Conference on Knowledge and Smart Technology* (pp. 42-46). Chonburi, Thailand: IEEE.

Mookdarsanit, P., and Mookdarsanit, L. (2018). A Content-based Image Retrieval of Muay-Thai Folklores by Salient Region Matching. *International Journal of Applied Computer Technology and Information Systems*, 21-26.

Mookdarsanit, P., and Mookdarsanit, L. (2018). An Automatic Image Tagging of Thai Dance's Gestures. *Joint Conference on ACTIS and NCOBA* (pp. 76-80). Ayutthaya, Thailand.

Mookdarsanit, P., and Mookdarsanit, L. (2018). Contextual Image Classification towards Metadata Annotation of Thai-tourist Attractions. *ITMSoc Transactions on Information Technology Management*, 32-40.

Mookdarsanit, P., and Mookdarsanit, L. (2018). Name and Recipe Estimation of Thai-desserts beyond Image Tagging. *Kasem Bundit Engineering Journal*, 193-203.

Mori, G., and Malik, J. (2003). Recognizing objects in adversarial clutter: breaking a visual CAPTCHA. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. I-I). Madison, WI, USA: IEEE.

Naor, M. (1996). *Verification of a human in the loop or Identification via the Turing Test. Rehovot*. Israel.

Redmon, J., and Farhadi, A. (2017). *YOLO9000*: *Better, Faster, Stronger*. *arXiv: 1612.08242* .

Redmon, J., and Farhadi, A. (2018). YOLOv3: *An Incremental Improvement*. *arXiv: 1804.02767* .

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In *The 2016 IEEE Conference on Computer Vision and Pattern Recognitio* (pp. 779-788). Lasvegas, Nevada: IEEE.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection. In *The 29th Conference on Neural Information Processing Systems* (pp. 91–99). Montreal, Canada.

Sarkar, S., Bansal, A., Mahbub, U., & Chellappa, R. (2017). UPSET and ANGRI : Breaking High Performance Image. *Classifiers*. *arXiv: 1707.01159* .

Simonyan, K., and Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. *arXiv: 1409.1556* .

Soimart, L., and Mookdarsanit, P. (2016). Gender Estimation of a Portrait: Asian Facial-significance Framework. *In The 6th International Conference on Sciences and Social Sciences*. Mahasarakham, Thailand.

Soimart, L., and Mookdarsanit, P. (2017). Ingredients Estimation and Recommendation of Thai-foods. *SNRU Journal of Science and Technology*, 509-520.

Soimart, L., and Mookdarsanit, P. (2016). Multi-factor Authentication Protocol for Information Accessibility in Flash Drive. *In The 9th Applied Computer Technology and Information Systems*, 10-13. Nakhon Pathom, Thailand.

Soimart, L., and Mookdarsanit, P. (2017). Name with GPS Auto-tagging of Thai-tourist Attractions from An Image. In *The 2nd Technology Innovation Management and Engineering Science International Conference* (pp. 211-217). Nakhon Pathom, Thailand.

Sutthaluang, N. (2018). The impact of asbestos exposure on lung disease. In *The 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (pp. 353-355). Chiang Rai, Thailand: IEEE.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A.A. (2017). Inception-v4, inception-ResNet and the impact of residual connections on learning. In *The 31st AAAI Conference on Artificial Intelligence* (pp. 4278–4284). San Francisco, California: ACM.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. and Anguelov, D. (2015). Going deeper with convolutions. In *The 2015 IEEE Conference on Computer Vision and Pattern Recognition* (pp 1-9). Boston, Massachusetts.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In *The 2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818-2826). Las Vegas, Nevada: IEEE.

Van De Sande, K.E., Uijlings, J.R., Gevers, T., and Smeulders, A. W. (2011). Segmentation as selective search for object recognition. In *The 2011 International Conference on Computer Vision* (pp. 1879-1886). Barcelona: IEEE.

Wang, M., and Deng, W. (2018). Deep visual domain adaptation. *A survey. Neurocomputing, 312*, 135-153.

Zhang, L. (2019). *Transfer Adaptation Learning: A Decade Survey*. *arXiv: 1903.04687* .

Zheng, L., Yang, Y., and Tian, Q. (2018). SIFT Meets CNN: A Decade Survey of Instance Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1224-1244.

Zou, Z., Shi, Z., Guo, Y., and Ye, J. (2019). *Object Detection in 20 Years: A Survey*. *arXiv: 1905.05055*.