Rectifying adversarial inputs using XAI techniques

Ching-Yu Kao
Department Cognitive Security Technologies
Fraunhofer AISEC
Garching near Munich, Germany
ching-yu.kao@aisec.fraunhofer.de

Junhao Chen
Karlsruhe Institute of Technology
Karlsruhe, Germany
junhao-chen@outlook.com

Karla Markert
Department Cognitive Security Technologies
Fraunhofer AISEC
Garching near Munich, Germany
karla.markert@aisec.fraunhofer.de

Konstantin Böttinger
Department Cognitive Security Technologies
Fraunhofer AISEC
Garching near Munich, Germany
konstantin.boettinger@aisec.fraunhofer.de

***Abstract*—With deep neural networks (DNNs) involved in more and more decision making processes, critical security problems can occur when DNNs give wrong predictions. This can be enforced with so-called adversarial attacks. These attacks modify the input in such a way that they are able to fool a neural network into a false classification, while the changes remain imperceptible to a human observer. Even for very specialized AI systems, adversarial attacks are still hardly detectable. The current state-of-the-art adversarial defenses can be classified into two categories: pro-active defense and passive defense, both unsuitable for quick rectifications: Pro-active defense methods aim to correct the input data to classify the adversarial samples correctly, while reducing the accuracy of ordinary samples. Passive defense methods, on the other hand, aim to filter out and discard the adversarial samples.**

Neither of the defense mechanisms is suitable for the setup of autonomous driving: when an input has to be classified, we can neither discard the input nor have the time to go for computationally expensive corrections. This motivates our method based on explainable artificial intelligence (XAI) for the correction of adversarial samples. We used two XAI interpretation methods to correct adversarial samples. We experimentally compared this approach with baseline methods. Our analysis shows that our proposed method outperforms the state-of-the-art approaches.

***Index Terms*—explainable AI, neural networks, deep learning, adversarial defense**

I. Introduction

With ever-increasing computing power and data volume, neural networks have gained much recognition over the past years, enabling them to grow constantly in scale, accuracy, and complexity. Neural networks are applied to a variety of different fields getting more and more power of decision, including medicine [1], economics [2], machine translation [3], [4], speech recognition [5], [6], and knowledge graphs [7].

The main idea of this paper is to use XAI to correct the malicious samples when they are detected. We intuitively believe that XAI will inform us of the adversarial attacks. This assumption leads us to the central insight of this article: we demonstrate that modifying or removing the focus of XAI can cause the malicious sample to return to the correct sample. From this observation, we are optimistic that our method can also defend against unknown attacks. We plan to release our code, model, and data set for future work on this topic. In summary, we have made the following contributions:

- To the best of our knowledge, we are the first one to use XAI, namely iGOS (Integrated Gradients Optimized Saliency), to correct adversarial prediction results.
- Our experiments show that our method is better compared to most of the baseline methods
- Our method goes beyond detection, as it can continue to predict and have a correct result while making the passive method retain the adversarial example.

We are optimistic that our method can resist unknown attacks because it obtained good results in the transfer adversarial example experiment.

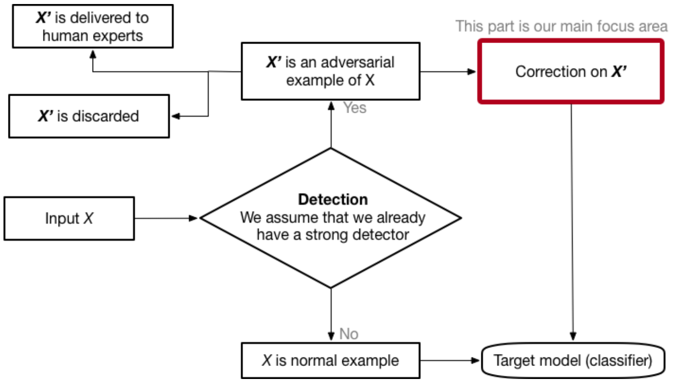

X' is delivered to human experts
This part is our main focus area
X' is an adversarial example of X
Correction on ***X'***
X' is discarded
Yes
Input *X*
Detection
We assume that we already have a strong detector
No
X is normal example
Target model (classifier)

Fig. 1. The figure shows the motivation of our concept. We assumed to have stable and powerful detectors which can detect whether an image is adversarial or not. Our main goal is to correct the adversaries to have a correct prediction.

A. Adversarial attack

Even though extremely high accuracy neural networks have surpassed humans in some areas, they are fragile and susceptible to small disturbances. The interference in these images and voices is even imperceptible to human eyes and ears. This kind of attack can make the network output wrong results with a

ISBN: 978-1-6654-6798-8

high degree of confidence. Let $f(\cdot)$ be a trained deep network and $h(\cdot)$ is human judgment. We assume that

$$f(x) = h(x)$$

Here, we give a formal definition according to [8]:

a) Definition: An **adversarial example** x' is a normal input x with a human-unperceived perturbation ϵ, that is $x' = x + \epsilon$, in other word, $\|x' - x\|_p \leqslant \epsilon$ for some small $\epsilon \in \mathbb{R}^+$. $\|x' - x\|_p$ is defined as the following Equation 1.

$$\|\vec{x}' - \vec{x}\| = \sqrt[p]{\sum_{i=1}^{m} |(x' - x)_i|^p} \tag{1}$$

more specifically, if x' is an adversarial example, it holds:

$$h(x) = h(x') \quad \wedge \ f(x') \neq f(x').$$

II. Considered Threat Models

In our experiment, we consider a threat model with the following information: the attacker performs adversarial attacks and tries to alter the classification output of our target model. For this purpose, the attacker uses various state-of-the-art algorithms. Furthermore, the added adversarial perturbations are desired to be small enough to be indistinguishable for a human expert, consistent with the standard definition of adversarial examples. Finally, we consider a white-box scenario in which the attacker performs simple attacks on the target model, which means the attacker knows all information of the target model.

III. Methodology

XAI is a method that helps human to comprehend and understand the results of deep learning. Therefore, we believe that it can help us understand the misclassification of the adversarial sample. Correcting these reasons can cause misclassification (in the image domain, it means specific pixels). On the other hand, we can have the correct classification for the adversarial sample. To prove our concepts, we choose two interpretation methods: Intergradient-CAM and iGOS.

A. Intergradient CAM-based Defense

In Grad-CAM, we use the derivative to express the importance of the feature map. The usage of ordinary derivatives can cause some problems. These two major problems are the saturation region problem and the problem that the sensitivity is not accurately equal to importance. We apply integrated gradient to Grad-CAM and obtain Integrated Grad-CAM. We found that inputting the original and the adversarial image of the same image into Integrated Grad-CAM would result in different key areas, as shown in Figure 2

Here, we assume that there exists a detector that accurately classifies inputs into adversarial and benign samples. The benign samples can be directly passed to the subsequent classification model. The adversarial samples are then fed into our defense network for correction.

First, we use Integrated Grad-CAM to identify the key areas of the attack image that play a positive role in the attack label.

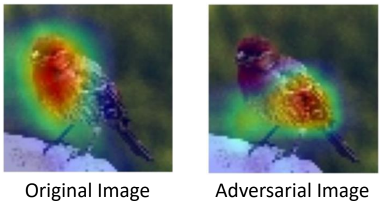


Fig. 2. Interpretation comparison using Integrated Grad-class activation mapping (CAM) for the original and adversarial image to original label.

The interpretative ability of Grad-CAM computed for different convolutional layers decreases significantly from the last layer to the fore layer [9]. Hence, we chose the last convolutional layer to compute the critical area that is the saliency map.

Second, we modify the different percentages of those key areas in the adversarial image without interfering with the important features for classifying the original image.

Third, we observe that the key area of adversarial image to post-attack label usually does not overlap with the original image's key area with respect to the original label. Due to the randomness of both methods, we perform this step multiple times. The percentage of key areas is changed for a specific dataset. We use two intuitive methods to modify the features:

- Randomly deleting the pixels.
- Generating blurring image of adversarial image using Gaussian and randomly replacing the pixels with blurring pixels.

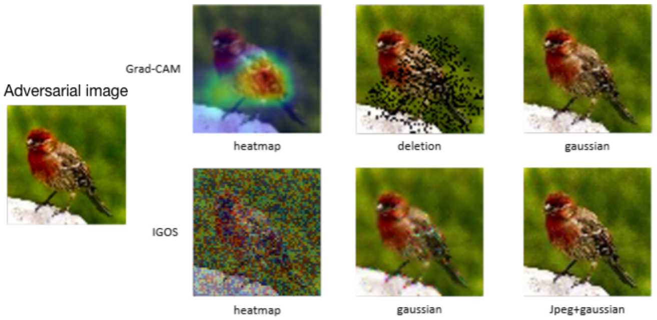


Fig. 3. Example results of using CAM-based and modified-iGOS-based method on an adversarial image.

Figure 3 presents experimental results using our proposed methods.

Then, these modified adversarial images are fed into the classification model. At last, the output results are statistically analyzed to determine the final label.

In principle, the derived saliency map is relevant to the target model, but the same image will have a different saliency map for different models. Hence, our defense method uses the qualities of the image and the nature of the classification model.

The above steps can be summarized visually in Figure 4.

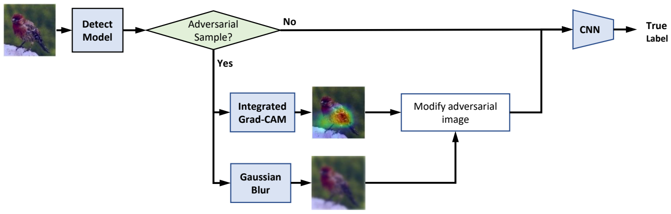


Fig. 4. Integrated Grad-CAM-based defense overview: Given an image to the detector to detect whether it is an adversarial sample or not. Suppose it is a benign sample, input to Convolutional Neural Networks (CNN) directly. If it is an adversarial sample, input to Integrated Grad-CAM to find the key region, correct the image by randomly replacing the pixels with the Gaussian blur image of the adversarial image and then input to CNN to get the correct label.

B. iGOS based Defense

Qi et al. [10] proposed a more result-oriented method called Integrated Gradients Optimized Saliency (I-GOS) approach, for visualizing the deep networks. It is based on mask optimization and integrated gradients approach. The mask approach uses optimization techniques to generate heatmaps for finding the area that maximally decreases the neural network output and the integrated gradient approach claims that the heatmaps can reflect the output changes. The mask approach constructs a mask M as the heatmap to perturb the input I_o which is optimized by solving the following objective function in equation 6. The mask is a coefficient matrix that fuses the base image and the interpreted image and has values of every element in the range [0,1]. The coefficients in the mask are sorted to filter out the essential pixel positions and obtain the final interpretive result. The Gaussian blur of the original image is selected as the base image instead of the plain black image as the latter creates new strong edges that can significantly impact the neural network model and can even obscure the important features. The objective function is provided in the following equation:

$$\begin{aligned} \arg\min_{M} & F_c(I_o, M) = f_c(\phi(I_o, M)) + g(M) \\ \text{where } & g(M) = \lambda_1 \|1 - M\|_1 + \lambda_2 TV(M), \\ & \phi(I_o, M) = I_o \odot M + \tilde{I_o} \odot (1 - M), \\ & 0 \leq M \leq 1. \end{aligned} \tag{2}$$

where M is the mask, $\phi(I_o, M)$ is the fused figure, f_c is the neural network output on class c, $\tilde{I_o}$ is the baseline image with a low score on class c and having the same shape as the input image, $g(M)$ is the penalty term. The first term $\lambda_1 \|1 - M\|_1$ of $g(M)$ aims to make the original image occupy as much weight as possible in the fusion figure, and the second term $\lambda_2 TV(M)$ is a total variation norm used to make the mask as smooth as possible. The gradient descent method is chosen for the optimization as it is one of the most common methods. The gradient used here is the integrated gradient as it provides a better direction and points towards the global optimum. The Goldstein-Armijo condition is used to determine the step size.

First, to use iGOS for adversarial defense, we have to modify the objective function of original iGOS by removing two unnecessary penalty terms since they make the interpretation region too smooth and small. To defend using this method, we want to extract essential pixels instead of area. Furthermore, in the original iGOS method, there is no precise limit on the number of iterations, which is an artificially selected parameter. We find that many iterations can destroy the important features of the original image, and if it is too small, it may decrease the defense effect. Therefore, we limit the number of iterations by controlling the reduction of probability values.

Second, we use no-penalty iGOS to generate mask values M of the adversarial example I_a. We compute the inner product of I_a and M to have our first term $(I_a \odot M)$.

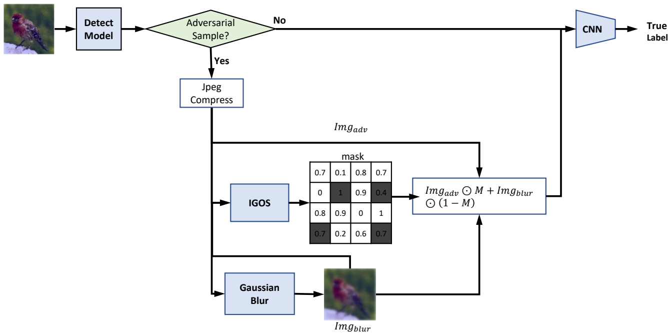
Third, we use Gaussian to blur the whole adversarial image, namely I_g. By using inner product, we have the second term $(I_g \odot (1 - M))$

Finally, we correct our adversarial image ϕ by adding first term and second term, more formally:

$$\phi(I_a, M) = I_a \odot M + I_g \odot (1 - M) \qquad 0 \leq M \leq 1 \tag{3}$$

The above steps can be summarized visually in Figure 5.

In addition, we combine iGOS and JPEG-compression methods to have more experiment results, since compressing the image with Jpeg in advance can remove the high-frequency noise in the adversarial image.



0.7	0.1	0.8	0.7
0	1	0.9	0.4
0.8	0.9	0	1
0.7	0.2	0.6	0.7

Fig. 5. Given an image to the detector to detect whether it is an adversarial sample or not. Suppose it is a benign sample, input to CNN directly. If it is an adversarial sample, generate Gaussian blur image, input both to iGOS to obtain mask, fuse the adversarial image and the Gaussian blur image by mask, then input to CNN to get the correct label. Notice that the mask size M is the same as I_a and I_g

IV. Experiment and Result

A. Selected Attack Methods

We defined our threat model in II. To validate the defense method more effectively, we selected eight attack methods and two metrics. They are FGSM L_2, FGSM L_∞ [11], BIM L_2, BIM L_∞ [12], PGD L_2, PGD L_∞ [13], CW L_2 [14].

B. Dataset

Three public datasets are used in this thesis, Mnist [15], Cifar10 [16], and Mini-Imagenet, Mini-Imagenet is generated by selecting fifteen classifications from the ImageNet [17] Large Scale Visual Recognition Challenge (ILSVRC) 2012-2017 image classification and localization dataset [18]..

C. Target Model - Classification Model

In both Mnist and Cifar10, we used simple CNN. Residual Network (ResNet) [19] is used for our Mini-ImageNet. Table I presents the classifier in detail.

TABLE I
THE TRAINING PARAMETERS AND FINAL ACCURACY

	MNIST	Cifar-10	Imagenet
Architecture	Simple CNN	Simple CNN	ResNet
Optimization algorithm	Adam	Adam	Adam
Learning rate	10^{-4}	10^{-2}	10^{-4}to10^{-7}
Batch size	128	128	38
Epochs	99	39	59
Test accuracy	99.5%	82.6%	81.0%

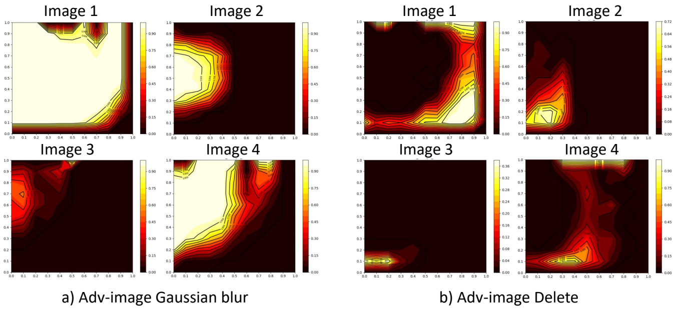
D. Baseline method

Our defense method is compared with other defense methods to verify its effectiveness. This paper chooses four baseline methods: Autoencoder, Jpeg compression, full image Gaussian blur, and full image random deletion.

E. Experimental Results

We compare our experimental results with baseline methods. First, we summarize their defense success rates in Table II. In the three datasets, it shows that iGOS-based defense has the highest defense success rate under all attack methods and far exceeds baseline under most attack methods. Unfortunately, CAM-based methods do not offer better performance than baseline. The possible reason is that each image has its own best intensity and percentage. Although such a set of parameters exists for most images that can be corrected with a very high probability, these regions do not overlap. Observing many images similar to Figure 6 shows that this set of parameters is related to the original image classification. Otherwise, no uniform pattern can be found for these regions. In particular, the performance is even worse on Cifar10 due to the high probability that Grad-CAM makes a wrong interpretation. Besides, we find that during the CAM-based experiment, adversarial images and original images have different focus hotspots. Different classifications have different sensitivity to different intensities and percentages. It is verified that the interpretation performance of IG-GradCAM is indeed stronger than that of GradCAM.

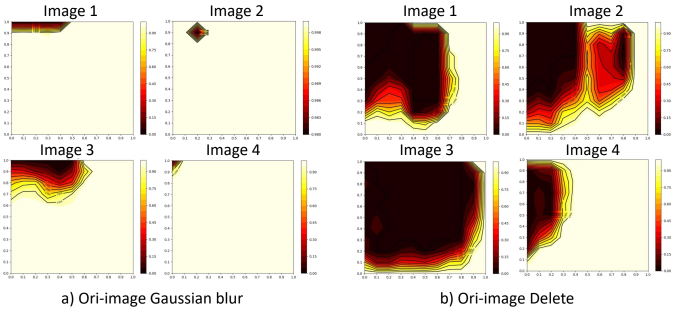
In addition, we want to know the effect of modifying the key area in the original image on the classification results, so we did the same operation for the original image as shown in Figure 7. Comparing the two sub-figures a) and b) in Figure 7, it is evident that the original image has higher stability to Gaussian blur replacement than direct deletion. Hence, we conclude that method 1 has better performance than method 2. This may be because direct pixel deletion will produce many black borders. These borders will seriously destroy or obscure the features necessary for classification for both the adversarial and original images. However, this is out of our research scope, since we assume we have strong detectors, which can determine if an input is an adversarial example.



a) Adv-image Gaussian blur

b) Adv-image Delete

Fig. 6. a) Examples of contour plots of the accuracy of adversarial images, of which Gaussian blur replaces a percentage of pixels in key areas. b) Examples of contour plots of accuracy of which deletes a percentage of pixels in key areas. The horizontal axis is the intensity of heatmap, which controls the key area size.



a) Ori-image Gaussian blur

b) Ori-image Delete

Fig. 7. a) Examples of contour plots of accuracy of original images, of which a percentage of pixels in key areas are replaced by Gaussian Blur. b) Examples of contour plots of accuracy of original images, of which a percentage of pixels in key areas are deleted. The horizontal axis is the intensity, which controls the size of the key area. The vertical axis is the percentage

V. CONCLUSION

In practice, we can find several scenarios where the adversarial samples should not be discarded or kept in the queue waiting for human intervention. The automated driving system is one such example where we need to correct the wrong samples immediately to avoid any safety-related issues. In this paper, we innovatively used the interpretive method for adversarial defense, or more specifically, for the correction of adversarial examples. To the best of our knowledge, this is the first paper using XAI methods to correct adversarial inputs. We verified the wide adaptability and good performance of the iGOS-based defense method during our experiments with various attack methods. The accuracy of the defense under various attacks exceeds the baseline methods. Generally speaking, iGOS-based methods outperform CAM-based methods and other baseline methods. Finally, we hope that our method can inspire more research in this direction.

REFERENCES

[1] Z. Hao, C. Lu, Z. Huang, H. Wang, Z. Hu, Q. Liu, E. Chen, and C. Lee, "Asgn: An active semi-supervised graph neural network for molecular

TABLE II
ACCURACY USING DIFFERENT DEFENSE METHODS OF ADVERSARIAL SAMPLES

dataset	defense methods	FGSM L_2	FGSM L_∞	BIM L_2	BIM L_∞	PGD L_2	PGD L_∞	CW L_2
MNIST	Autoencoder	0.581	0.5	0.76	0.581	0.734	0.552	0.621
	Jpeg compression	0.055	0.037	0.111	0.095	0.131	0.09	0
	Full image Gaussian Blur	0.467	0.389	0.616	0.459	0.604	0.433	0.389
	Full image random deletion	$0.39^{+0.09}_{-0.09}$	$0.28^{+0.1}_{-0.05}$	$0.32^{+0.12}_{-0.07}$	$0.28^{+0.07}_{-0.09}$	$0.30^{+0.08}_{-0.12}$	$0.26^{+0.08}_{-0.07}$	$0.26^{+0.14}_{-0.09}$
	IG-gradcam gaussian	0.254	0.315	0.43	0.405	0.4	0.358	0.344
	gradcam gaussian	0.254	0.278	0.404	0.351	0.4	0.313	0.328
	IG-gradcam deletion	0.305	0.315	0.281	0.345	0.242	0.299	0.39
	gradcam deletion	0.288	0.296	0.271	0.405	0.232	0.328	0.344
	iGOS based	**0.9**	0.852	0.945	**0.905**	0.917	0.896	0.969
	iGOS based with jpeg compression	0.883	**0.889**	**0.949**	**0.905**	**0.937**	**0.91**	**0.972**
Cifar10	Autoencoder	0.293	0.455	0.446	0.617	0.582	0.624	0.731
	Jpeg compression	0.044	0.093	0	0.051	0	0.074	0.404
	Full image Gaussian Blur	0.222	0.279	0.271	0.322	0.254	0.271	0.277
	Full image random deletion	$0.21^{+0.06}_{-0.03}$	$0.19^{+0.1}_{-0.07}$	$0.12^{+0.03}_{-0.04}$	$0.18^{+0.04}_{-0.04}$	$0.21^{+0.03}_{-0.04}$	$0.22^{+0.04}_{-0.03}$	$0.25^{+0.05}_{-0.04}$
	IG-gradcam gaussian	0.267	0.233	0.203	0.407	0.305	0.389	0.447
	gradcam gaussian	0.222	0.326	0.271	0.441	0.237	0.352	0.574
	IG-gradcam deletion	0.2	0.209	0.186	0.288	0.203	0.333	0.319
	gradcam deletion	0.2	0.279	0.136	0.186	0.203	0.222	0.255
	iGOS based	**0.47**	**0.581**	**0.616**	0.712	**0.695**	**0.778**	0.872
	iGOS based with jpeg compression	0.422	0.512	0.576	**0.729**	0.661	**0.778**	**0.936**
ImageNet	Autoencoder	0.653	0.644	0.7	0.703	0.705	0.708	0.706
	Jpeg compression	0.583	0.346	0.321	0.414	0.259	0.357	0.869
	Full image Gaussian Blur	0.208	0.154	0.321	0.207	0.111	0.286	0.174
	Full image random deletion	$0.33^{+0.05}_{-0.08}$	$0.18^{+0.05}_{-0.03}$	$0.37^{+0.06}_{-0.12}$	$0.31^{+0.1}_{-0.07}$	$0.26^{+0.18}_{-0.07}$	$0.23^{+0.06}_{-0.09}$	$0.40^{+0.08}_{-0.1}$
	IG-gradcam gaussian	0.75	0.538	0.536	0.552	0.444	0.536	0.696
	gradcam gaussian	0.75	0.423	0.429	0.483	0.333	0.429	0.609
	IG-gradcam deletion	0.458	0.192	0.357	0.345	0.296	0.179	0.435
	gradcam deletion	0.375	0.192	0.429	0.31	0.222	0.25	0.391
	iGOS based	0.667	0.5	0.714	0.724	0.703	0.571	0.87
	iGOS based with jpeg compression	**0.83**	**0.731**	**0.893**	**0.98**	**0.963**	**0.929**	**0.95**

property prediction," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 731–752.

[2] W. Huang, K. K. Lai, Y. Nakamori, S. Wang, and L. Yu, "Neural networks in finance and economics forecasting," *International Journal of Information Technology & Decision Making*, vol. 6, no. 01, pp. 113–140, 2007.

[3] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[4] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[5] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2011.

[6] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[7] T. Hamaguchi, H. Oiwa, M. Shimbo, and Y. Matsumoto, "Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach," *arXiv preprint arXiv:1706.05674*, 2017.

[8] P. Sperl, C.-Y. Kao, P. Chen, X. Lei, and K. Böttinger, "Dla: dense-layer-analysis for adversarial example detection," in *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2020, pp. 198–215.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[10] Z. Qi, S. Khorram, and F. Li, "Visualizing deep networks by optimizing with integrated gradients." in *CVPR Workshops*, vol. 2, 2019.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[12] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[14] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[15] C. J. B. Yann LeCun, Corinna Cortes, "The mnist database of handwritten digits," [EB/OL], http://www.cs.toronto.edu/~kriz/index.html/ Accessed.

[16] A. Krizhevsky, "The cifar-10 dataset," [EB/OL], http://www.cs.toronto.edu/~kriz/index.html/ Accessed 2009.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[19] 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.