Restricting In-variance and Co-variance of Representations for Adversarial Defense in Vision Transformers

Jeevithan Alagurajah
School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, LA 70504, U.S.A.
Email: jeevithan.alagurajah1@louisiana.edu

Chee-Hung Henry Chu[a,b]
[a]Informatics Research Institute
[b]School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, LA 70504, U.S.A.
Email: chu@louisiana.edu

***Abstract*—The development of trustworthy and secure AI applications is a fundamental step towards building AI systems that can reliably operate in the real world, where they may face malicious attempts to manipulate them. Deep neural networks, despite their impressive image classification accuracy, are vulnerable to even small, imperceptible changes called adversarial attacks, causing their performance to plummet. Existing defenses often struggle when attackers have full knowledge of the model (white-box attacks) and craft even stronger perturbations. To address this, the Adversarial Invariant and Co-Variance Restriction (AICR) loss function was recently proposed. The AICR loss function forces clean and noisy images from the same class to have similar activation patterns in convolutional neural networks, essentially making them harder for attackers to differentiate. Given the superior performance of Vision Transformers (ViTs) in image classification, we adapted the AICR loss to train ViTs and investigated its effectiveness against gradient-based attacks. Our experiments show that ViTs trained with AICR loss achieve a significant improvement in accuracy compared to those trained with the standard cross-entropy loss, demonstrating the effectiveness of AICR in enhancing ViT's resilience against adversarial attacks.**

***Index Terms*—Vision transformers, adversarial training, adversarial defense, image classification**

I. Introduction

While Deep Neural Networks (DNNs) have achieved impressive performance in various domains such as computer vision [1]–[4], natural language processing [5]–[7], speech recognition [8]–[10], and reinforcement learning [11]–[13], their vulnerability to adversarial attacks remains a critical concern. Adversarial attacks involve creating imperceptible modifications to input data that can fool DNNs into making incorrect predictions. This vulnerability raises serious questions about the reliability and accountability of DNNs, particularly in such applications as autonomous vehicles, security and surveillance, and environmental monitoring. Designing AI systems that are robust and generalizable against adversarial attacks is therefore an urgent and crucial challenge, thus requiring further research to ensure the safe and responsible deployment of DNNs across various critical domains.

In recent years, numerous defense methods have been proposed to combat adversarial attacks. These methods can be broadly categorized into two types: reactive and proactive defenses. Reactive defense methods focus on modifying or transforming inputs to counter specific attack strategies. Examples of such methods include input smoothing, input clipping, and adversarial training with specific noise distributions [14]–[17]. However, these methods are often limited in their effectiveness against unknown attack strategies and may not generalize well to diverse adversarial perturbations. Proactive defense methods, on the other hand, aim to inherently improve the robustness of the model itself. This is achieved by modifying model parameters, network architectures, training objectives, or utilizing adversarial training with diverse noise distributions. Proactive methods are generally more versatile and provide wider protection against various attacks, making them more widely adopted in practical applications.

A novel adversarial training framework using the so-called Adversarial Invariant and Co-Variance Restriction (AICR) loss function was proposed [18] that incorporates two objectives to enhance robustness and generalization: maximizing class separation and minimizing intra-class variance. The first objective utilizes an attractive and repulsive mechanism at different representation levels. This mechanism encourages samples from the same class to cluster together (attractive) and pushes samples from different classes apart (repulsive). This promotes a model that naturally separates classes, making it more resilient to noise and achieving better generalization. The second objective aims to minimize the variation between adversarial and clean images within the same class. This is achieved by maximizing the correlation and minimizing the redundancy in their intermediate representations. This ensures that visually similar samples are projected to the same region in the multidimensional space, making them difficult to fool with adversarial attacks within a given budget. This approach was shown (i) to achieve state-of-the-art robustness against a wide range of strong adversarial attacks under the strongest first-order attack and (ii) to maintain the highest robust performance under black-box settings using the CNN6 net [19] and the ResNet-110 [4] Convolutional Neural Networks (CNNs).

Vision Transformers (ViTs) [20] is a different DNN ap-

 ISBN: 978-1-68558-157-2

proach towards image classification compared to CNNs. Unlike CNNs, which process images pixel-by-pixel, ViTs first divide an image into patches and treat these patches as tokens. The network then learns by examining the relationships between these tokens, similar to how it learns relationships between words in a sentence. This allows ViTs to capture long-range dependencies and global context within an image. CNNs have a strong inductive bias towards local spatial relationships due to their convolutional structure. This helps them learn effectively even with smaller datasets. ViTs rely on attention mechanisms, which have a weaker inductive bias towards spatial structure. This means they are more flexible but may require more data or specific techniques to guide them towards learning the important features. When training on smaller datasets, ViTs are more likely to overfit, which increases its reliance on model regularization or data augmentation.

While the AICR loss function has demonstrated superior adversarial defense performance on CNNs like Resnet-110, its effectiveness on ViTs remains unexplored. ViTs have recently gained significant traction for their impressive image classification capabilities. However, their vulnerability to adversarial attacks is only beginning to be addressed. Given the recent surge in interest and potential benefits of ViTs, we propose incorporating them into the AICR objective function to assess their adversarial defense performance.

The rest of this paper is organized as follows. Relevant related work is reviewed in Section II. In Section III, we describe our proposed method and our testing plan. We present our experimental results in Section IV. Finally, we draw our conclusion and suggest future work in Section V.

II. Related Work

Deep learning algorithms are vulnerable to adversarial perturbations, which are carefully crafted inputs that can cause the model to make incorrect predictions. Several defense algorithms have been proposed to counter such attacks, and these can be broadly categorized into two main approaches: input transformation and model modification.

Input transformation methods attempt to modify the input data in a way that makes it more difficult for the adversary to craft effective perturbations [17], [21]. For example, one common technique is to add noise to the input data. This can make it more difficult for the adversary to find perturbations that have a significant impact on the model's output.

Model modification methods attempt to make the model itself more robust to adversarial perturbations [22]–[24]. One common technique is adversarial training, which involves training the model on a dataset that includes both clean and adversarial examples. This can help the model to learn how to better distinguish between clean and adversarial inputs.

Given the recent competitive performance of ViTs for image classification tasks, the robustness of these methods against adversarial attacks has received more attention. ViT architectures can be broadly classified as vanilla and hybrid. Vanilla ViTs are pure attention-based and are computationally less expensive. Hybrid ViTs combine CNNs and attention by incorporating both convolutional layers and self-attention modules. This leverages the strengths of both approaches, viz. CNNs for extracting local features and attention for capturing global relationships. The robustness of vanilla and of hybrid ViTs against adversarial attacks were found to be different [25]. While both vanilla and hybrid ViTs are tougher against adversarial attacks compared to regular CNNs, it was shown that vanilla ViTs but not the hybrids resisted defenses aimed at high-frequency features, suggesting potential differences in how they process information.

The adversarial robustness of ViTs has been explored by focusing on their unique building blocks [26]. It was shown that significant improvement in their ability to resist deception is feasible by randomly hiding information within these blocks during training.

A different approach to achieve adversarial robustness in ViT is to modify the training recipe [27]. Traditionally, training ViTs relies heavily on data augmentation. While effective for normal training, this approach was shown to hurt performance during adversarial training. Instead, omitting data augmentation and incorporating specific techniques like ε-warmup and bigger weight decay significantly improves the robustness of ViTs.

Adversarial training on the ViT architecture is computationally expensive. An attention-guided adversarial training method was introduced to trade off computational efficiency and adversarial robustness [28]. This method identifies and removes unimportant parts of an image during training, focusing the model's attention on crucial areas. This significantly speeds up training while maintaining or even improving robustness.

III. Method

We propose to incorporate the AICR loss function [18] into the ViT architecture and test the performance against white-box attacks. We summarize the characteristics of the ViT architecture next.

A. The Vision Transformer (ViT)

Transformers, originally developed for natural language processing, excel at image classification with their ability to capture long-range dependencies and contextual relationships within images [20]. The image is first divided into smaller patches, each of which is converted into a fixed-length embedding vector, capturing essential information about its color, texture, and other features. Additional information about the relative position of each patch within the image is incorporated into the embedding vectors.

In the transformer encoder's self-attention mechanism, each patch embedding attends to all other patch embeddings, allowing it to learn how relevant each patch is to itself and other parts of the image. It does so by transforming the patch embedding into three separate vectors: a query, a key, and a value. Each query vector is compared to all other key vectors using a dot product operation to generate a matrix of attention scores, where each score represents the similarity between a pair of patches. Each attention score is normalized

 ISBN: 978-1-68558-157-2

using a softmax function, turning it into a weight. This weight indicates the relative importance of each patch to the query patch. The values of all patches are multiplied by their respective weights and then summed up. This results in a new vector that represents the query patch based on the information from all other patches, weighted by their relevance.

The multi-head attention is formed by repeating the self-attention process multiple times in parallel, using different, randomly initialized matrices for generating queries, keys, and values. The attended representations from all heads are concatenated together, which is is then transformed by a final linear layer to produce the output of the multi-head attention layer.

A feed-forward network further processes the information extracted by the self-attention layers, adding non-linearity to increase the model's expressiveness. Residual connections and layer normalization help stabilize the training process to improve the overall performance of the model.

After the final transformer encoder layer, the output vector representing the entire image is passed through a classification head, which is typically a Multi-Layer Perceptron (MLP) that predicts the probability of the image belonging to each class.

B. Loss Function for ViT Adversarial Training

In [18], it was shown that the AICR loss provides an effective and robust defense against state-of-the-art white-box attacks and black-box settings. We adapt the AICR loss from its constituent loss functions for the ViT architecture as follows.

Let K be the number of classes of a given data set distribution $\mathcal{D}$ and N be the number of samples in the data set. For an image classification task, we formulate a deep neural network as $\mathcal{F}_\theta(x)$, where θ is the trainable parameters and x is the input image. The DNN outputs a feature representation $h_x \in \mathbb{R}^d$ for input x which is then used for classification in a multiclass classifier $Z = [z_k] \in \mathbb{R}^{d\times K}$, where $k = 1, \cdots, K$. To train the model we minimize an objective function to minimize θ and Z.

By maximizing the similarity between intermediate representations and minimizing redundancy, a so-called variance loss function was introduced to enforce local compactness between images and their adversarial counterparts. This effectively removes unnecessary information from the input data by decorrelating the clean image and its adversarial counterpart and ensuring all variables have similar variances. Consequently, both the clean image and its adversarial counterpart retain minimal but sufficient representations for accurate classification.

To further enhance the local compactness of features, a convolutional generator network G_Ψ is employed to map the intermediate layer of the discriminator network into a new feature space with reduced redundancy. The G_Ψ mapping is learned in an end-to-end manner by minimizing the cross-correlation between the clean and adversarial features while maintaining their individual variances. A square matrix Q is

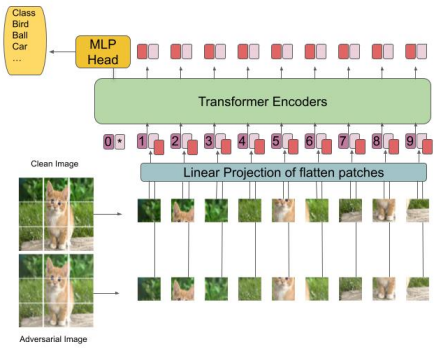

Fig. 1. The ViT architecture modified to learn jointly from $\mathcal{L}_{CE}$ and $\mathcal{L}_{var}$.

computed between the clean and its corresponding adversarial image in terms of G_Ψ:

$$Q = \frac{G_\Psi(h_x) \times G_\Psi(h'_x)}{\sqrt{G_\Psi(h_x)}\,\sqrt{G_\Psi(h'_x)}}. \tag{1}$$

Here, $Q \in \mathbb{R}^{d\times d}$, where d is the dimension of the output of G_Ψ. Intuitively, the diagonal elements of matrix Q represent the dot-product between clean and corresponding adversarial images while off-diagonal terms represent the covariance between them.

Minimizing this objective correlates clean and adversary counterparts and encourages to have non-redundant information while being closer in intermediate layers. This allows the centers to have contrasting representations and promotes maximum separation between classes. The variance loss function is then defined as

$$\mathcal{L}_{var} = \sum_i (1 - Q_{ii})^2 + \lambda \sum_i \sum_{j\neq i} Q_{ij}^2 \tag{2}$$

where λ is the trade-off parameter of the invariant (diagonal) and the redundancy (off-diagonal) terms of the matrix.

The overall prediction accuracy of an input-label pair (x, y) is ensured by the commonly used softmax Cross-Entropy (CE) loss $\mathcal{L}_{CE}(x, y)$. A third loss function is the so-called attract-repulsive loss function, which is not applicable to the ViT architecture.

The loss function that encourages the same classes to be mapped closer and different classes to be mapped farther from each other by a large margin is developed as follows. The modified AICR loss function is used in the training of a ViT as illustrated in Figure 1. The AICR loss function is modified for the ViT architecture by combining the cross entropy loss function and variance loss function for an input-label pair (x, y) and the corresponding adversarial image x' as:

$$\begin{aligned}\mathcal{L}(x, x', y) = \sum_i^n (&\mathcal{L}_{CE}(x_i, x'_i, y_i) \\ &+ \alpha \times \mathcal{L}_{var}(h_i^l, h_i'^l, y_i))\end{aligned} \tag{3}$$

where

$$h^l = \mathcal{G}^l_\phi(\mathcal{F}^l_\theta(x)) \quad \text{and} \quad h'^l = \mathcal{G}^l_\phi(\mathcal{F}^l_\theta(x')),$$

and α is the regularizing term for the contrastive centroid loss, $\mathcal{G}_\phi$ is the auxiliary function that maps intermediate layers to a lower dimension output, and n denotes the number of layers.

IV. Experimental Results

We evaluated the variance loss [18] with the ViT [20] architecture. We tested two different variants of ViT network that incorporate variance loss in the final representations. One of them incorporates $\mathcal{L}_{var}$ in the final classification head, identified as *ViT-C*; and the other variant uses $\mathcal{L}_{var}$ on representations of the final patches except the classification head, and it is identified as *ViT-All*.

We illustrate attention shift due to adversarial attacks when a network is trained solely by the CE loss and when a network is trained by the AICR loss as follows. The Grad-CAM [29] helps us understand why a model predicts a certain class by using an attention map to highlight the areas in the image that most influence its decision. In Figure 2, we show the attention maps of some sample images that have been trained using $\mathcal{L}_{CE}$, the cross entropy loss. In Figure 3, we show the corresponding attention maps of the adversarial images. The shifts due to the adversarial attack are evident by comparing the two images.

The reason attention shifts are markers of suscpetibility to adversarial attacks is as follows. Many adversarial attacks work by subtly manipulating an image in a way that causes the deep learning model to shift its attention to irrelevant or misleading parts of the image. This attention shift can lead to misclassifications. Deep learning networks that are robust to adversarial attacks tend to exhibit smaller or less significant shifts in their attention when presented with adversarial examples. This suggests that a model maintaining its focus on the correct features is less likely to be fooled. If a deep learning network keeps its attention on the right parts of the image even when attacked, it has a higher chance of correctly identifying the object despite the adversary's attempts to mislead it.

In Figure 4, we show the attention maps of some sample images that have been trained using $\mathcal{L}_{AICR}$, the loss function that combines cross entropy and variance loss. In Figure 5, we show the corresponding attention maps of the adversarial images. The lack of shift due to the adversarial attack are evident by comparing the two images.

In an adversarial setting, there are two main threat models. In *white-box* attacks, the adverser has complete knowledge of the target model including model architecture and objective function used for training and parameters. *Black-box* attacks, on the other hand, feed adversarial noise to the input images during inference time, and it is crafted without any knowledge of target model. Following the attack settings in [30], we crafted adversarial examples in a non-targeted way with respect to allowed perturbation ϵ for gradient based attacks, i.e., FGSM, BIM, PGD, MIM. The number of iterations for BIM, MIM, PGD were set to 10 with a step size of $\epsilon/10$. We

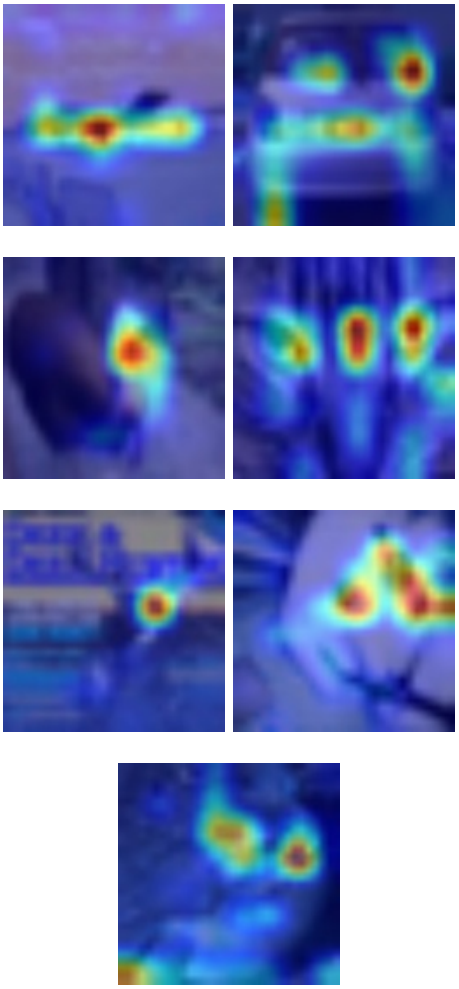

Fig. 2. Attention maps by Grad-CAM of clean CIFAR-10 images; model trained by CE loss.

compare the accuracy of ViT network with $\mathcal{L}_{var}$ at different settings of ViT on the CIFAR 10 data set for white box attacks. Results in Table I show that using variance loss increased the robustness of the model compared to the model that does not use variance loss.

TABLE I
Accuracy of vision transformers under adversarial attacks

Attacks	ϵ	ViT	*ViT-C*	*ViT-All*
No-attack	-	**80.1**	78.9	79.6
FGSM	0.1	15.2	15.8	**16.2**
	0.2	2.7	1.8	**3.6**
PGD	0.1	8.5	**9.9**	9.2
	0.2	0.15	**0.33**	0.16
BIM	0.1	8.4	**9.9**	9.1
	0.2	0.15	**0.33**	0.16
MIM	0.1	8.8	**10.3**	9.6
	0.2	0.17	**0.37**	0.22

Here *ViT-C* refers to the model trained with $\mathcal{L}_{var}$ only on the classification head and *ViT-All* refers to the model with patch representations trained optimized with $\mathcal{L}_{var}$. All models are trained jointly with adversarial samples and clean samples.

 ISBN: 978-1-68558-157-2

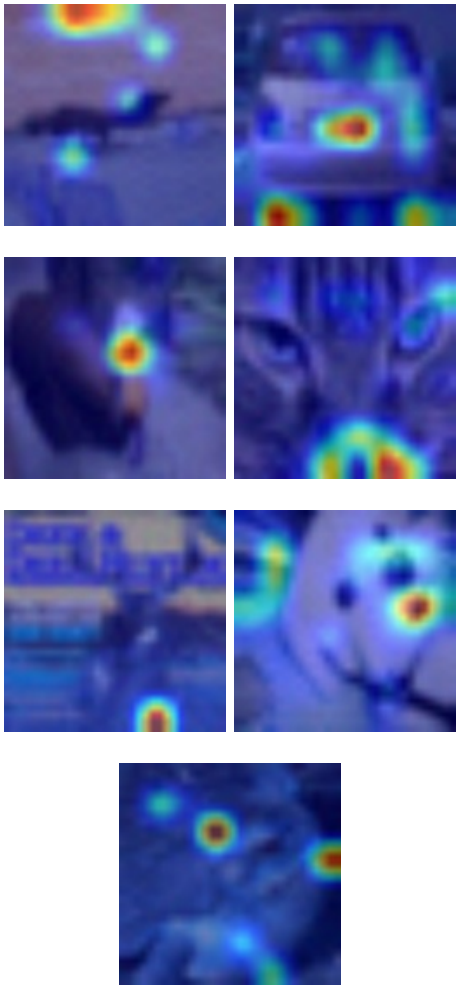

Fig. 3. Illustration of attention shift using Grad-CAM on model trained by CE loss. Adversarial images are obtained crafted with PGD (ϵ = 0.03); see clean images in Figure 2 for comparison.

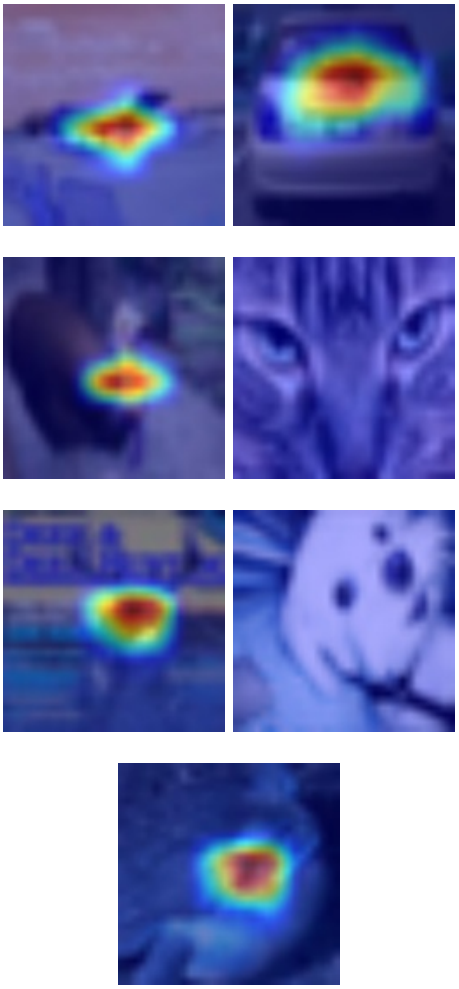

Fig. 4. Attention maps by Grad-CAM of clean CIFAR-10 images; model trained by AICR loss.

We can see that there were performance drops by *ViT-C* and *ViT-All* when there was no attack. At all levels of attacks, the ViT trained with the modified AICR loss performed better than the ViT trained with only cross entropy loss. The *ViT-All* network performed better under the FGSM attack while the *ViT-C* network performed better the PGD, BIM, and MIM attacks. Under the FGSM attacks, *ViT-All* improved by 6.57% to 33.33% for $\epsilon = 0.1$ and 0.2, respectively. Under the PGD, BIM, and MIM attacks, *ViT-C* improved by about 17% to 120% for $\epsilon = 0.1$ and 0.2, respectively.

V. Conclusion and Future Work

The AICR loss function has previously shown a significant improvement in the robustness of CNNs against adversarial attacks, particularly in such tasks as image classification. Motivated by ViT's superior performance in such tasks, we adapted the AICR loss and investigated its effectiveness in training ViTs against gradient-based attacks such as PGD and BIM. Our experiments revealed negligible changes in the attention distribution of ViTs trained with modified AICR loss compared to cross-entropy, indicating stable attention patterns. Furthermore, ViTs trained with AICR loss achieved a 33% to 120% improvement in accuracy compared to cross-entropy, demonstrating its effectiveness in enhancing ViT's resilience against adversarial attacks.

A promising direction of future work might focus on exploiting attention shift as a marker of adversarial vulnerability. Encouraging models to maintain consistent attention patterns between clean and adversarial examples during training is a promising defense strategy, aimed at improving robustness against attention-based attacks. Networks can be explicitly trained with adversarial examples that are designed to shift attention. This helps the model learn to recognize these tricks and maintain stability. A longer term goal is to develop more reliable connections between attention and adversarial robustness.

Acknowledgments

The authors thank the anonymous reviewers whose suggestions helped to improve the clarity of our paper. This work is supported in part by the U.S. National Science Foun-

 ISBN: 978-1-68558-157-2

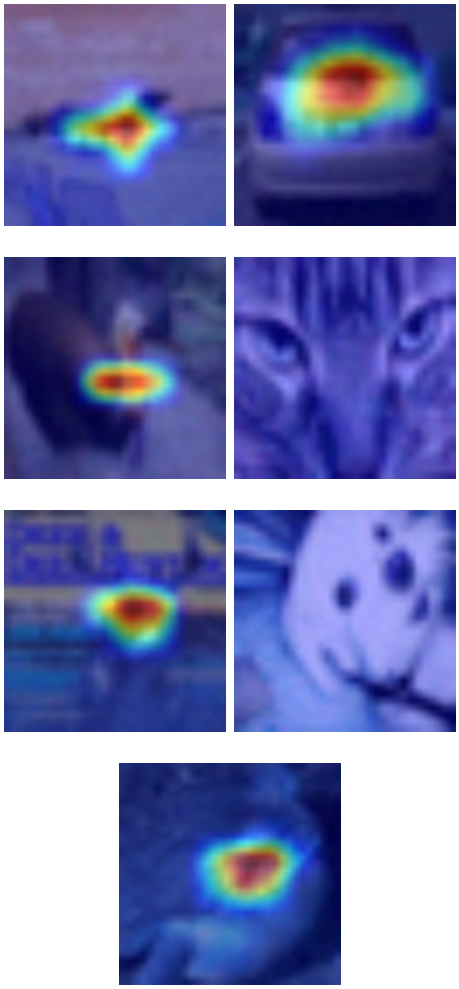

Fig. 5. Illustration of attention shift using Grad-CAM on model trained by AICR loss. Adversarial images are obtained crafted with PGD ($\epsilon = 0.03$); compare with clean images in Figure 4.

dation under grant number OIA-1946231 and the Louisiana Board of Regents for the Louisiana Materials Design Alliance (LAMDA).

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

[4] 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*, pp. 770–778, 2016.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] W. Wang and J. Gang, "Application of convolutional neural network in natural language processing," in *2018 International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pp. 64–70, 2018.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized Bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[8] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060–1089, 2013.

[9] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[10] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[11] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," *Innovations in multi-agent systems and applications-1*, pp. 183–221, 2010.

[12] H. Van Hasselt, "Reinforcement learning in continuous state and action spaces," in *Reinforcement learning*, pp. 207–251, Springer, 2012.

[13] "A cnn-based policy for optimizing continuous action control by learning state sequences," *Neurocomputing*, vol. 468, pp. 286–295, 2022.

[14] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression," *arXiv preprint arXiv:1705.02900*, 2017.

[15] Y. Luo, X. Boix, G. Roig, T. Poggio, and Q. Zhao, "Foveation-based mechanisms alleviate adversarial examples," *arXiv preprint arXiv:1511.06292*, 2015.

[16] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, and L. Shao, "Image super-resolution as a defense against adversarial attacks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1711–1724, 2019.

[17] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," *arXiv preprint arXiv:1711.01991*, 2017.

[18] J. Alagurajah and C.-H. H. Chu, "Adversarial defense by restricting invariance and co-variance of representations," in *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22)*, 2022.

[19] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*, pp. 499–515, Springer, 2016.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Countering adversarial images using input transformations," *arXiv preprint arXiv:1711.00117*, 2017.

[22] T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii, "Distributional smoothing with virtual adversarial training," *arXiv preprint arXiv:1507.00677*, 2015.

[23] F. Yu, C. Liu, Y. Wang, L. Zhao, and X. Chen, "Interpreting adversarial robustness: A view from decision surface in input space," *arXiv preprint arXiv:1810.00144*, 2018.

[24] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, 2021.

[25] A. Aldahdooh, W. Hamidouche, and O. Deforges, "Reveal of vision transformers robustness against adversarial attacks," *arXiv preprint arXiv:2106.03734*, 2021.

[26] Y. Mo, D. Wu, Y. Wang, Y. Guo, and Y. Wang, "When adversarial training meets vision transformers: Recipes from training to architecture," *arXiv preprint arXiv:2210.07540*, 2022.

[27] E. Debenedetti, V. Sehwag, and P. Mittal, "A light recipe to train robust vision transformers," *arXiv preprint arXiv:2209.07399*, 2023.

[28] B. Wu, J. Gu, Z. Li, D. Cai, X. He, and W. Liu, "Towards efficient adversarial training on vision transformers," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, (Berlin, Heidelberg), p. 307–325, Springer-Verlag, 2022.

[29] 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*, pp. 618–626, 2017.

[30] H.-Y. Chen, J.-H. Liang, S.-C. Chang, J.-Y. Pan, Y.-T. Chen, W. Wei, and D.-C. Juan, "Improving adversarial robustness via guided complement entropy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4881–4889, 2019.

 ISBN: 978-1-68558-157-2