CVIA - Assignment Report
KAMAMA

Matei Gabriel Cosa, Matilde Dolfato, Kassym Mukhanbetiyar

3159774, 3157341, 3191202

May 23, 2025

1 Introduction

Robustness and explainability are two critical pillars in the deployment of reliable computer vision systems, particularly in safety-sensitive applications. While adversarial training has emerged as a prominent technique to improve robustness against input perturbations, its impact on model interpretability remains less understood.

In this context, *robustness* refers to *a model's ability to maintain high classification performance when subjected to adversarial perturbations.* A robust model should correctly classify both clean and adversarial examples, indicating resilience to small input changes designed to mislead the model. *Explainability* is defined here as *the degree to which a model's predictions can be interpreted in terms of localized, semantically meaningful visual evidence.* This concept can be understood as a measure of how well the model's internal reasoning aligns with human-understandable concepts such as object location.

This study investigates whether adversarial training can improve both explainability and robustness. We do this by fine-tuning five convolutional neural network architectures on the Oxford-IIIT Pet dataset, both under standard and adversarial training regimes with Fast Gradient Sign Method (FGSM) [1]. Explainability is measured through the overlap between bounding boxes generated from Gradient-weighted Class Activation Mapping (Grad-CAM) [6] heatmaps and ground-truth annotations. Our approach enables a systematic analysis of the relationship between adversarial robustness and explainability across five of the most influential neural architectures.

Our work contributes to the current literature by combining two methods whose use is well established for robustness and explainability evaluations separately. The interplay between these two aspects, instead, has been inquired only recently. It has been shown that adversarially trained networks can exhibit more stable saliency maps [4, 9], yet a systematic evaluation including diverse architectures remains lacking.

2 Method

FGSM Attack. The Fast Gradient Sign Method (FGSM) is a simple yet effective adversarial attack that perturbs an input image in the direction of the gradient of the loss with respect to the input. By adding a small perturbation, FGSM aims to maximize the model's prediction error while keeping the perturbation imperceptible to humans. Concretely, for a given input image x, FGSM produces a new image x_{adv} such that

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}\left(\nabla_x \mathcal{L}(x, y)\right) \tag{1}$$

where y is the true label, $\mathcal{L}$ is the loss function, and ϵ controls the perturbation magnitude. This method is widely used to evaluate and improve the robustness of neural networks.

Grad-CAM. Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique for visualizing the regions of an input image that a convolutional neural network focuses on when making a prediction. It does so by computing the gradient of the score for a target class y^c with respect to the activations A^k of a selected convolutional layer. The gradients are spatially averaged to obtain importance weights α_k, which are then used to weight the feature maps. The final class-discriminative localization map is obtained via a weighted combination followed by a ReLU activation:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k A^k\right) \tag{2}$$

$$\alpha_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$$

where α_k denotes the importance of feature map A^k, and Z is the total number of spatial locations in the feature map.

Model	Mean IoU	Adv. Mean IoU
ResNet	0.5164	**0.4965**
MobileNet	**0.5399**	0.4059
EfficientNet	0.4939	0.4484
VGGNet	0.2667	0.2410
DenseNet	0.2862	0.2782

Table 2: Mean Intersection-over-Union (IoU) of standard and adversarially trained models.

3 Experiments

3.1 Dataset

Our goal requires a dataset tailored for image classification tasks, while also including bounding box annotations. In order to reliably train classification models, we require that each image contains a single, main object from each of the classes of interest. On the basis of these considerations, we select a version of the Oxford-IIIT Pet dataset (`visual-layer/oxford-iiit-pet-vl-enriched`), which has a total of 37 classes (with around 200 images per each class) and includes bounding boxes. We resize images and corresponding bounding boxes to fit within a 224 x 224 square, while maintaining the original aspect-ratio.

3.2 Models

We choose five of the most successful neural architectures to serve as backbones for the classification task. Given the relatively small number of examples from each class, we decide to fine-tune pretrained models instead of training from scratch. This means that for each architecture, we load the feature extractor, append a classification head tailored for our number of classes, and train the models end-to-end (i.e., both the feature extractor and the classifier). The full list of selected models is as follows: **ResNet18** [2], **MobileNetV2** [5], **EfficientNetB0**, [8], **DenseNet121** [3], and **VGGNet19** [7]. For the latter, we use Global Average Pooling (GAP) to reduce the size of the model and limit overfitting, given the model's large parameter count.

3.3 Experimental Pipeline

Robustness. For each of the five selected architectures, we conduct two training regimes: **standard training** using clean input images, and **adversarial training** using a combined dataset of clean and FSGM perturbed inputs (50-50). This yields a total of **ten models fine-tuned for our task**. For evaluation, we compute the top-1 and top-5 accuracy on the standard test set, as well as on its perturbed adversarial version.

Explainability. For each trained model, we apply **Grad-CAM** to generate a saliency map corresponding to the predicted class label of each test image. To extract discrete object regions, we apply thresholding to convert the saliency maps into binary masks. We then perform contour detection on these masks to identify connected regions of activation, and fit rectangular bounding boxes around the detected contours. This procedure yields **localized regions** that can be directly compared to the ground-truth object annotations (in the standard, non-perturbed test set) via IoU.

4 Results

Classification performance on the standard test set is comparable across models, with all but VGGNet reaching top-1 accuracy above 0.86, while top-5 accuracy is above 0.97. The adversarially trained models perform slightly worse on the standard test set, with performance drops ranging from 0.01 to 0.04. The sharpest decrease in top-1 accuracy occurs for ResNet.

For the standard models on the adversarial test set, performance drops significantly, by as much as 0.7 points. DenseNet performed the best, with a top-1 accuracy of 0.42 and and top-5 accuracy of 0.87, followed by VGGNet. The adversarially trained version of DenseNet also scored the highest robust accuracy at 0.72, with all the models being above 0.66 for top-1 accuracy. Full results are displayed in Table 1.

In terms of bounding box overlap, we observe that standard MobileNet achieved the highest IoU score of 0.54, followed by ResNet and EfficientNet, with all of them hovering around 0.5. DenseNet and EfficientNet scored significantly lower at around 0.27. The IoU for the adversarial models was lower

Model	**Standard Test (Std)**		**Standard Test (Adv)**		**Adv. Test (Std)**		**Adv. Test (Adv)**	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
MobileNet	0.8745	0.9862	0.8571	0.9835	0.1437	0.6277	0.6633	0.9562
ResNet	0.8745	0.9837	0.8345	0.9801	0.2785	0.7521	0.6726	0.9528
EfficientNet	**0.8913**	**0.9917**	**0.8806**	**0.9898**	0.2093	0.7237	0.6972	**0.9664**
DenseNet	0.8607	0.9851	0.8530	0.9823	**0.4184**	**0.8668**	**0.7201**	0.9597
VGGNet	0.8260	0.9774	0.8136	0.9716	0.3315	0.8483	0.6635	0.9448

Table 1: Top-1 and top-5 accuracy of standard and adversarially trained models under standard and adversarial test conditions. Highest values in each column are bolded.

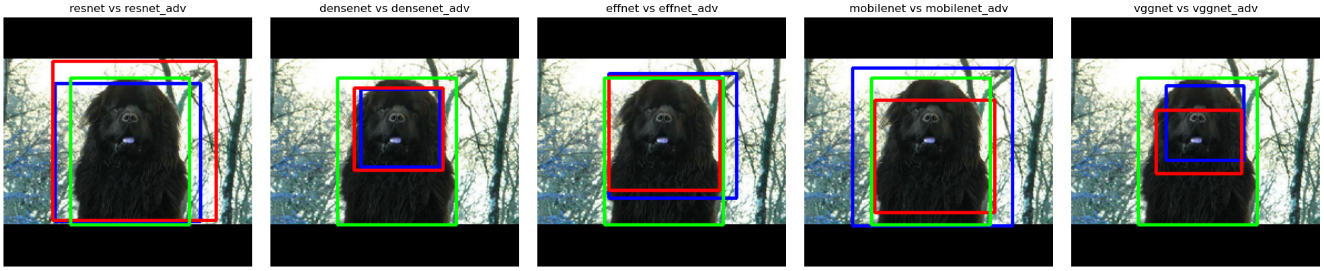

Figure 1: Bounding Box Comparison. Green denotes the ground-truth, red the standard model, and blue the adversarial one.

than their standard counterparts, with drops ranging from 0.01 to 0.13. The largest decrease occurred for MobileNet. Full results are displayed in Table 2.

5 Discussion

The first part of our experiments finds that adversarial training decreases standard accuracy, but this negative effect is almost negligible (0.01) for architectures like EfficientNet, DenseNet, and VGGNet. On the other hand, robust accuracy is significantly improved by training in the adversarial setting. This suggests that it may be worth paying the price in standard accuracy for applications where robustness is important. Notably, while still performing poorly, standard DenseNet and VGGNet considerably outperform the other standard models under adversarial testing. For the former, we hypothesize this may be due to the network connectivity which encourages learning more stable representations: each layer receives inputs from all previous layers and sends output to all subsequent layers. Regarding the latter, we suspect that the observed behavior may be due to the use of GAP more than the feature extraction part: global average pooling may act as a regularizer.

We also explored the bounding boxes from a qualitative perspective (see Figure 1). For the models with the lowest IoU scores (VGGNet and DenseNet), empirical evidence indicates that resulting boxes are smaller and more focused. From a human perspective, it can be argued that these bounding boxes (especially those produced by the robust models) capture more specific traits like facial characteristics, hence making these models' prediction criteria actually *more interpretable*. This finding also highlights a limitation of our approach: bounding box IoU may not be the ideal metric for explainability. We suggest this as a direction that could be explored in future works. Also, from these results it arises that for different flavors of the same task, different models could be more suitable, e.g. to obtain a box that encloses the entire subject a model like ResNet or MobileNet is a more appropriate choice.

6 Conclusion

Our work demonstrates that adversarial training can substantially improve robustness, with a relatively low cost in terms of standard performance. We also found that some architectural properties may be linked to increased robustness. While a quantitative analysis does not yield significant differences in terms of IoU scores of the resulting boxes, a more careful qualitative inspection uncovers interesting results. Smaller, more focused bounding boxes could be deemed as more explainable than larger ones closer to the ground-truth.

References

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

[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 (CVPR)*, pages 770–778, 2016.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[4] A. S. Ross and F. Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 1663–1670, 2018.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

[6] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. arXiv:1409.1556.

[8] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 97:6105–6114, 2019.

[9] H. Zhang, Y. Liu, Q. Dai, and C. Shen. Interpreting adversarially robust models via feature visualization and sensitivity analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2081–2089, 2019.

A Additional Images

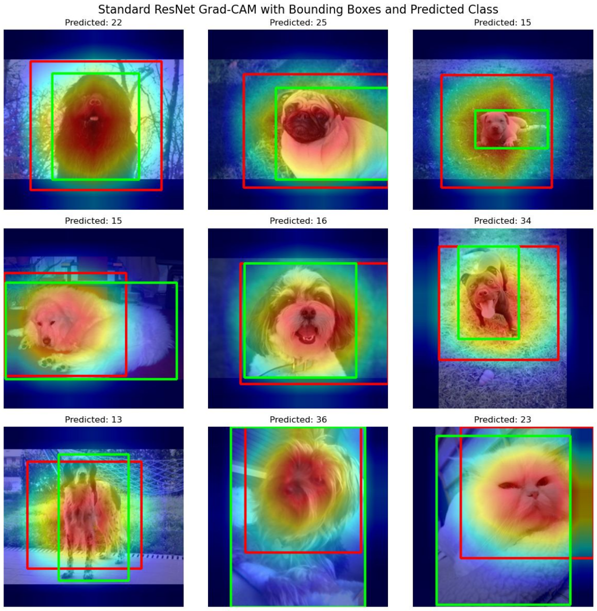

Figure 2: Grad-CAM plot for ResNet

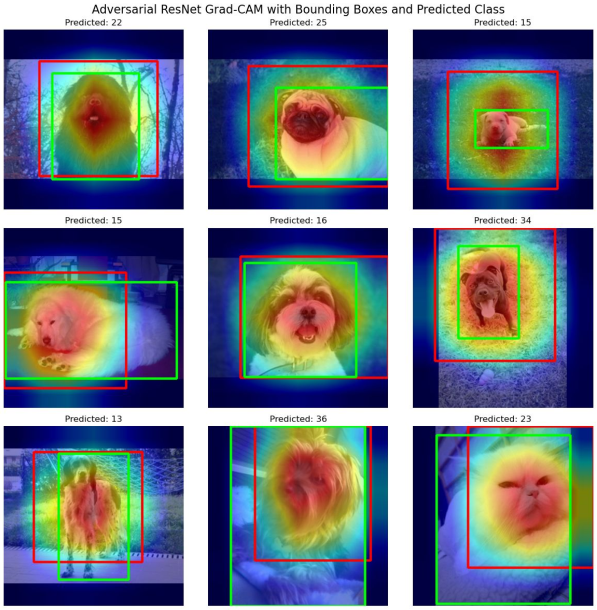

Figure 3: Grad-CAM plot for adversarial ResNet

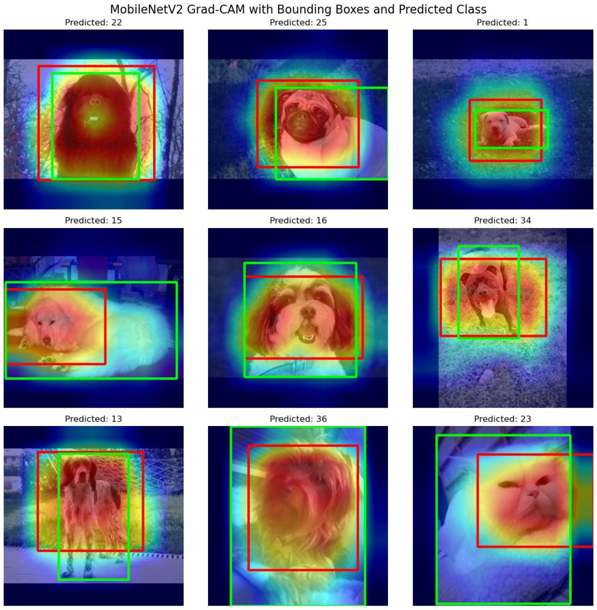

Figure 4: Grad-CAM plot for MobileNet

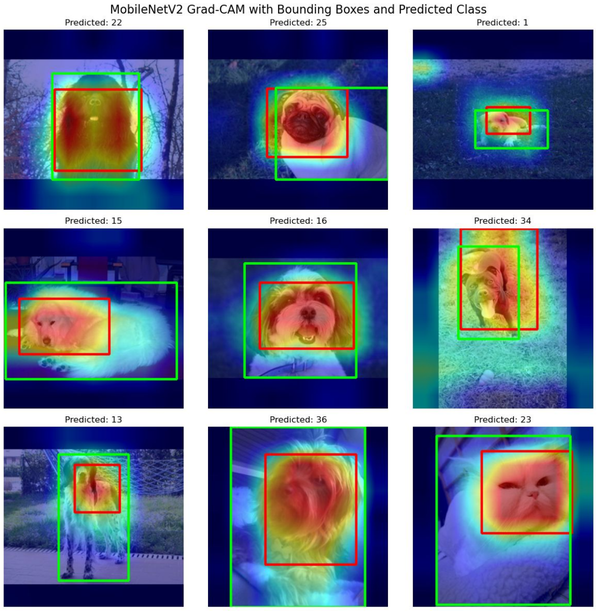

Figure 5: Grad-CAM plot for adversarial MobileNet

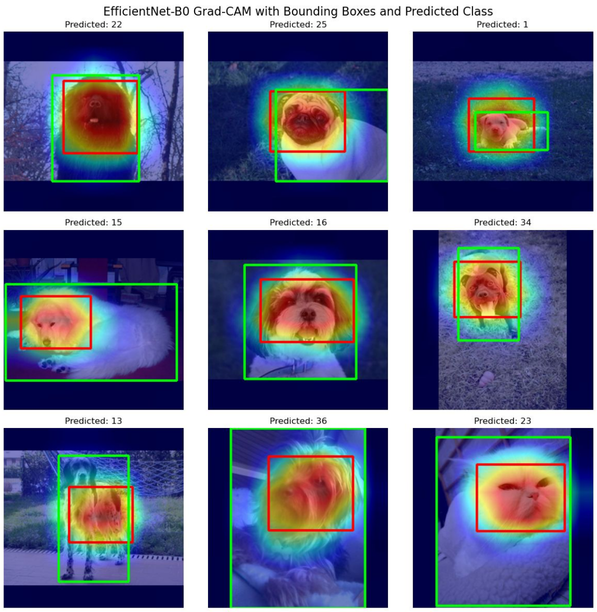

Figure 6: Grad-CAM plot for EfficientNet

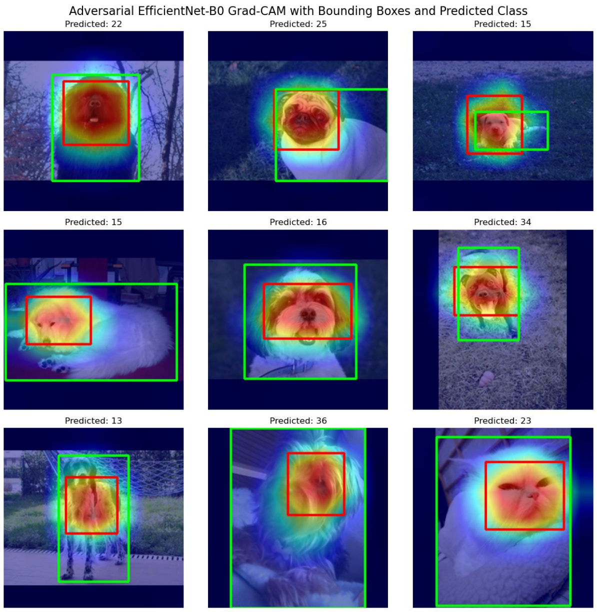

Figure 7: Grad-CAM plot for adversarial EfficientNet

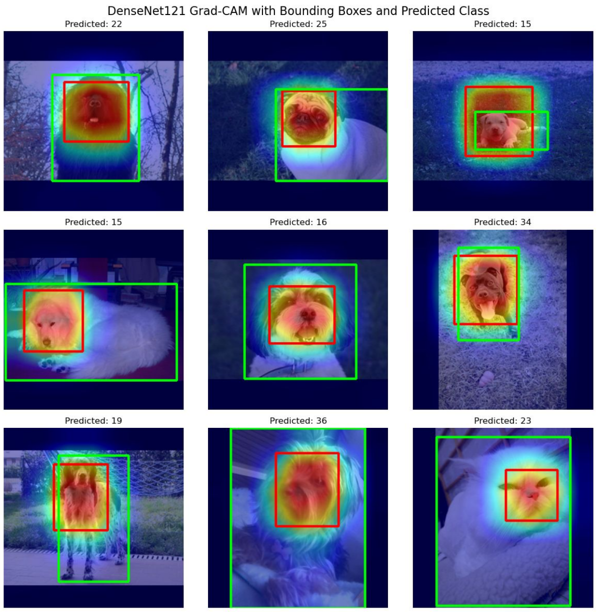

Figure 8: Grad-CAM plot for DenseNet

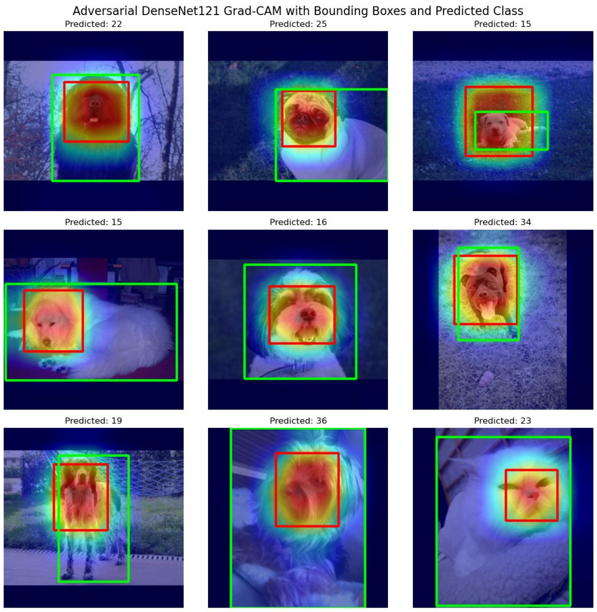

Figure 9: Grad-CAM plot for adversarial DenseNet

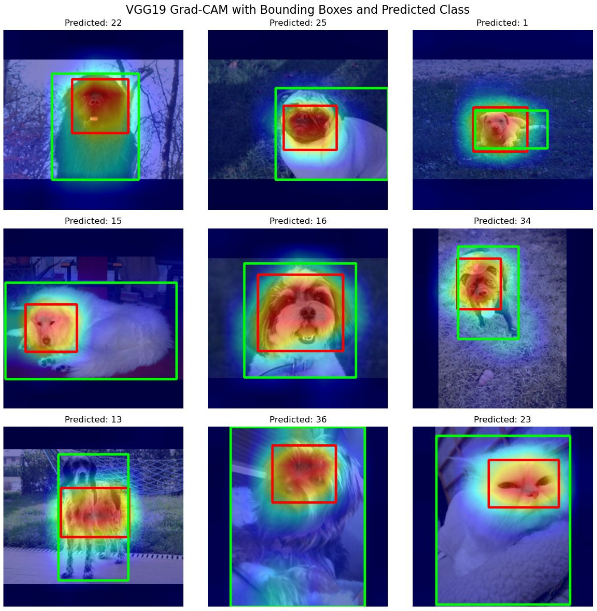

Figure 10: Grad-CAM plot for VGGNet

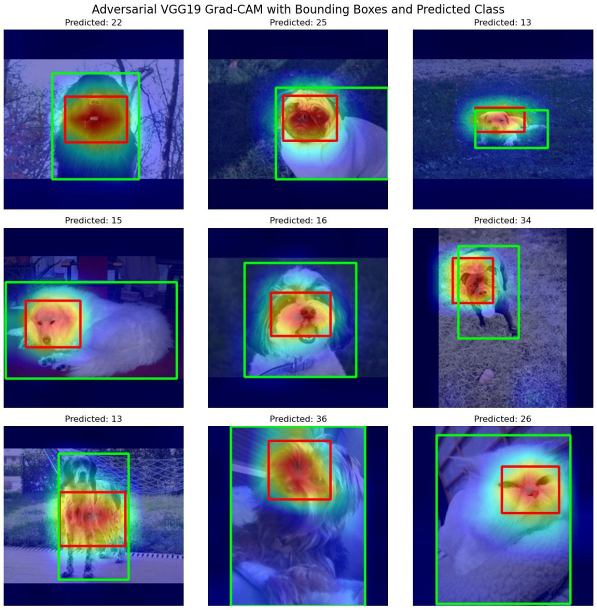

Figure 11: Grad-CAM plot for adversarial VGGNet