Visualizing Convolutional Neural Networks: Explainability and Interpretability

Soham Sen [1] Dr. Soumya Dutta [2]

[1]Department of Aerospace Engineering, IIT Kanpur [2]Department of Computer Science and Engineering, IIT Kanpur

Introduction

Recent advancements in deep learning have shown that deep convolutional neural networks (CNNs) are powerful for image classification, recognition and other predictive tasks but due to the lack of rationale for its decision, results in a less trustworthy model. Researchers do not yet understand how such models work and perform prediction, making it a black box model. Hence, if the network learns incorrect characteristics of an image, then it may not always correctly classify it, leading to unexplainable classifications. This is a problem in high-risk applications, such as medical diagnoses, mortgage acceptances and autonomous driving, recently leading to the European Union requiring automated decisions to include reasoning descriptions.

Background

CNNs consist of several layers of neurons with the exact number of layers depending on the model. The first few layers are convolutional layers that use neuron weights to form kernels. These kernels capture specific features in the image to form a feature map as input to the next layer. The pooling layers condense the feature maps to reduce the amount of background noise passed to the next layer,so that by the end of the convolutional layers, only the most important features for recognition remain . Finally, the fully connected layers use feature relevance probabilities to determine the object in the image. While CNNs have achieved remarkable success in various applications, there is a growing need for explainability in these models. Understanding why a CNN makes certain predictions is important for building trust and ensuring transparency. Evaluating and quantifying the quality of explanations is challenging due to the complexity of models. Current research in the field of explainability focuses on developing quantifiers that can measure the explainability of CNNs. These quantifiers can be human-centric or machine-centric, aiming to provide a standardized framework for evaluating and comparing different explainability methods in machine learning.

What is already known about this subject and What does our study add?

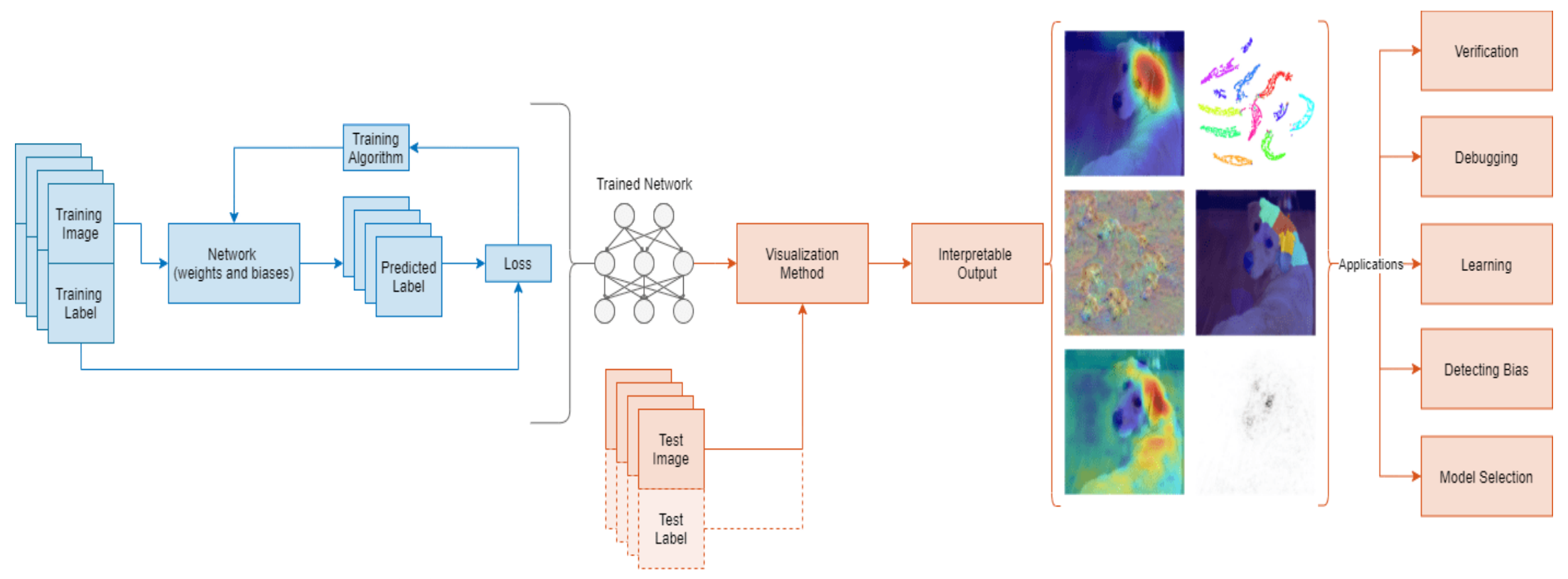
- **What we know** There is a wealth of research available, although it may be unorganized and scattered.
- **Our contribution** We not only aim to provide a comprehensive survey but also delve into important facets such as analyzing explainability models, harnessing adversarial examples, and comparing CNNs' accuracy with humans. By covering these aspects, we contribute to a deeper understanding of visualizing and understanding CNNs.

Literature Review and Related work

Zhuwei Qin's[3] survey on representative CNN visualization methods. Zhang and Zhu(2018), Burkart and Huber(2021), Islam et al.(2021) and Lynn Vonder Haar et. al. [2] surveyed existing explainability techniques. Jian-Xun Mi(2023)'s survey on adversarial attacks in object detection tasks and Andrej karpathy's work on comparing humans vs CNNs on ImageNet(Section 6.4).[4]

Research Objectives

- **Objective 1:** Understanding what does a CNN model learn and explaining the methods that add a layer of explainability to better increase the trustworthiness of the architecture. We reviewed three classes of explainability methods: gradient-based methods, perturbation-based methods, and approximation-based methods .
- **Objective 2:** Several deep neural networks, consistently misclassify adversarial examples—inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. We investigated the potential reasons for model's failure.
- **Objective 3:** Comparing the relative performance of Convolutional Neural Networks to trained Human annotators.

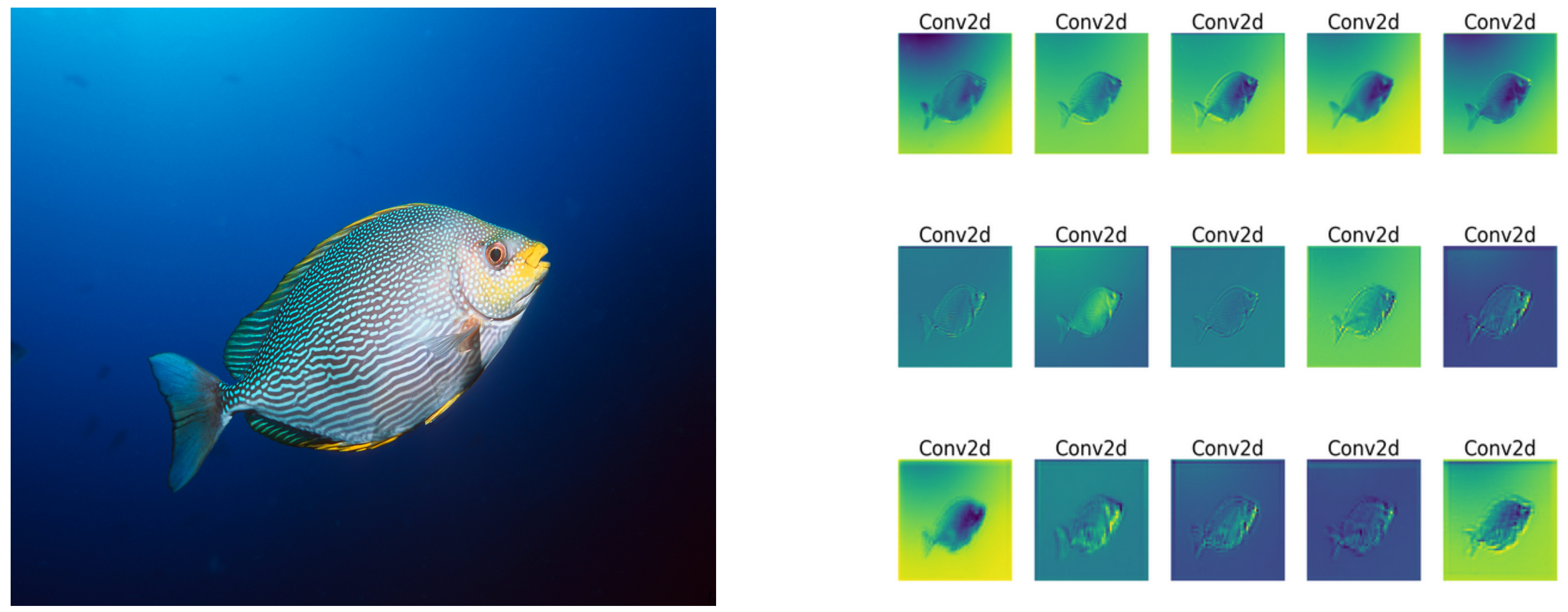


(a) Image courtesy : MathWorks[1]

Figure 1. Visualization and Interpretability in Deep Learning

Study Methodology

1. **What and how CNNs learn** Starting with visualizing the activations and weights, we then explored the "Rich feature hierarchies" technique for tracking images that activate specific neurons. To overcome limitations, we learned about "Embedding the codes with t-SNE," a better technique for visualizing high-dimensional data representations. Afterwards we saw gradient-based explainability methods namely deconvolutional neural networks, layer-wise relevance propagation, and deep Taylor decompositions that analyze gradients and relevance scores to explain neural network decision-making. We also discussed perturbation-based methods, specifically occlusion tests, which involve changing inputs to analyze changes in output and identify important input parts and finally approximation-based methods like Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP). These methods use random checks to determine pixel importance and differ from backpropagation-based approaches in producing saliency maps.
2. **Fooling ConvNets** We studied popular techniques for generating adversarial examples, including the most basic Fast Gradient Sign Method (FGSM), as well as complex methods like the Iterative-FGSM, L-BFGS, DeepFool, JSMA, ZOO and DBGE.
3. **Human Accuracy on large-scale image classification** In particular we used the HAM10000 ("Human Against Machine with 10000 training images") dataset, which consists of 10015 dermatoscopic images, to compare the performance of Convolutional neural Netwroks with human experts.



(a) Image of Maze rabbitfish (Siganus vermiculatus)

(b) Activation maps generated for each convolution layer

Figure 2. Visualizing Filter maps using the pre-trained ResNet50 model

Results and Discussion

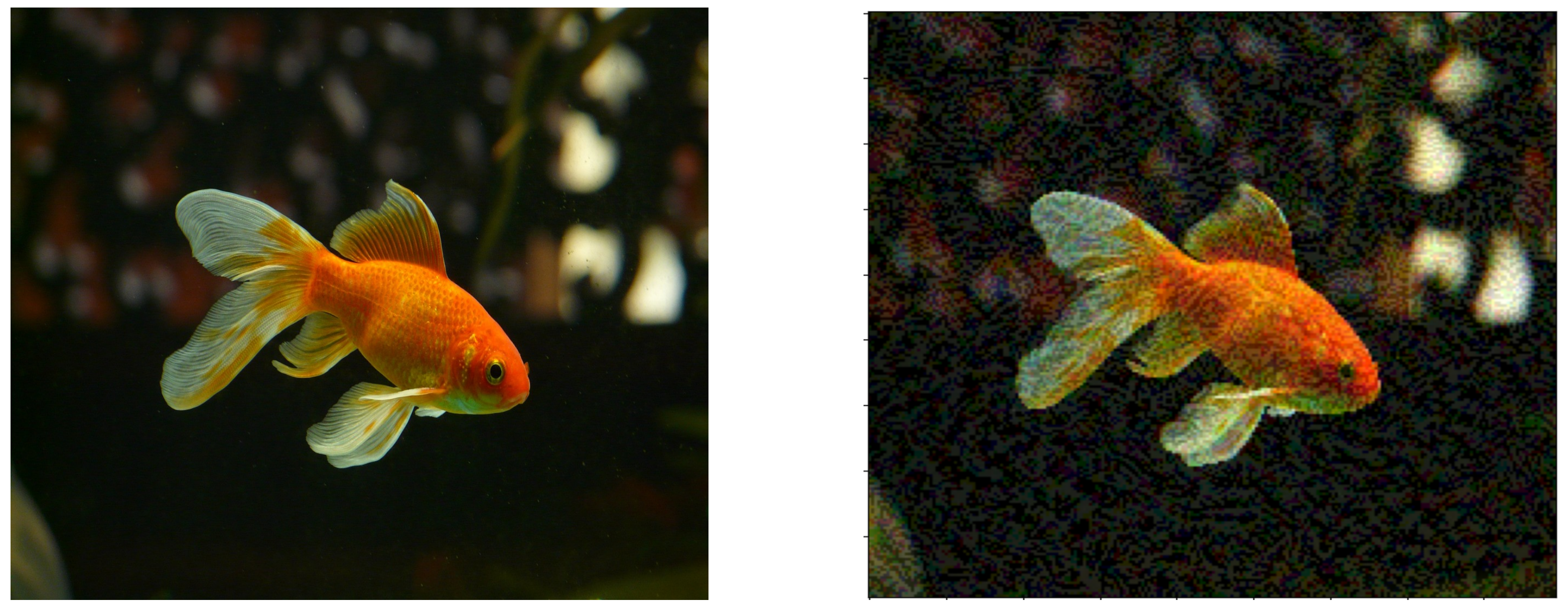
We compared CNN visualization techniques based on performance indexes for CIFAR-10 image classification.Each of these methods have advantages that some of the others do not. Using such methods could help to bridge the gap towards using CNNs in high-risk applications.

Table 1. Comparing various state-of-the-art interpretability methods.

Technique	Accuracy	Interpretability	Computational Complexity
Activation Maximization	0.78	Low	Medium
Deconvolutional Networks	0.72	Medium	High
Deep Taylor Decomposition	0.81	High	High
Embedding the code with t-SNE	0.83	Medium	Medium
Occlusion Tests	0.75	Medium	Low
Grad-CAM	0.85	High	Low
SHAP (Shapley Additive Explanations)	0.87	High	High

Security of deep learning systems are vulnerable to crafted adversarial examples, which may be imperceptible to the human eye, but can lead the model to misclassify the output. Depending on the attacker's knowledge about the model, there are several categories of attacks. The two most popular are the white box attack and black box attack. With white box attacks, the attacker has complete access to the architecture (weights), and the input and output of the model which is not the case for black box attacks.

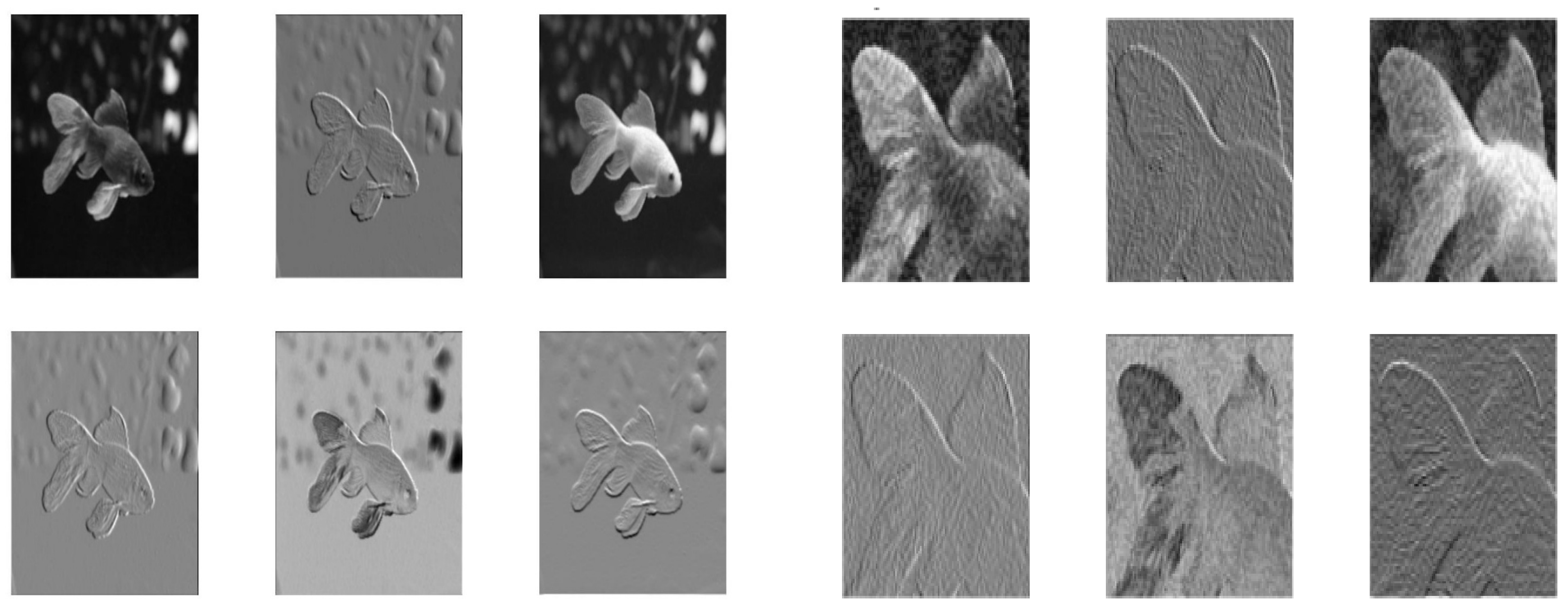
We used pre-trained CNN architectures on the ImageNet dataset to study the impact of the adversarial attacks on the classifier models and visualized their activation maps. We started with Fast Gradient Sign Method, a technique introduced by Ian Goodfellow(2014), which is a white box attack model. Subsequently we explored 4 other white box models namely I-FGSM, L-BFGS, DeepFool and JSMA. Among black-box models we explored two state-of-the-art techniques based on gradient estimation - Zeroth-Order Optimization (ZOO) and Decision-Based Gradient Estimation (DBGE).



(a) Veiltail Goldfish(in the absence of adversarial attack). Confidence : 94.18%

(b) With ε = 0.15, incorrectly classified as Catfish. Confidence = 67.36%

Figure 3. FGSM attack on VGG19 model pretrained on ImageNet dataset.



(a) Activation maps before FGSM attack

(b) Activation maps after FGSM attack

Figure 4. Visualizing the activation maps

Finally, on the question of comparing the accuracy of humans vs convolutional neural networks on large scale image recognition tasks in the absence of perturbation, using the HAM10000 dataset, we discovered that fine-tuning our model lead to an accuracy above 80%, surpassing the 77.0344% accuracy achieved by human labelers.

Conclusions

- **Explain and Interpret how CNNs work** Activation visualization and t-SNE embedding help understand neuron behavior and visualize high-dimensional representations. Gradient-based methods reveal feature importance, while perturbation-based methods analyze input changes. Approximation-based methods generate saliency maps but may lack global model behavior representation. Understanding these techniques helps choose the most suitable approach based on specific needs.
- **Adversarial Attacks** We have successfully fooled a state-of-the-art model into making wrong predictions, even with a low value of ε (e.g., 0.15). The adversarial images are indistinguishable from the originals, and none of the models correctly classify them. The highest confidence values are assigned to incorrect classes. With increasing value of epsilon, the noise becomes more visible, and the confidence for a wrong prediction increases.
- **Human VS CNNs** In conclusion, CNNs generally outperform or achieve comparable performance to humans in large-scale image classification tasks with good quality images.

References

[1] https://www.mathworks.com/help/deeplearning/visualization-and-interpretability.html.

[2] Lynn Vonder Haar, Timothy Elvira, and Omar Ochoa.
An analysis of explainability methods for convolutional neural networks.
Engineering Applications of Artificial Intelligence, 117:105606, 2023.

[3] Zhuwei Qin, Fuxun Yu, Chenchen Liu, and Xiang Chen.
How convolutional neural network see the world - a survey of convolutional neural network visualization methods.
2018.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei.
Imagenet large scale visual recognition challenge, 2015.