Model Explainability Report

Generated at: 2025-11-22T14:05:05.500021

1. Model Summary

Model Key: dog_cat

Model Name: ['Cat', 'Dog']

Prediction: Cat

Confidence: 0.0013

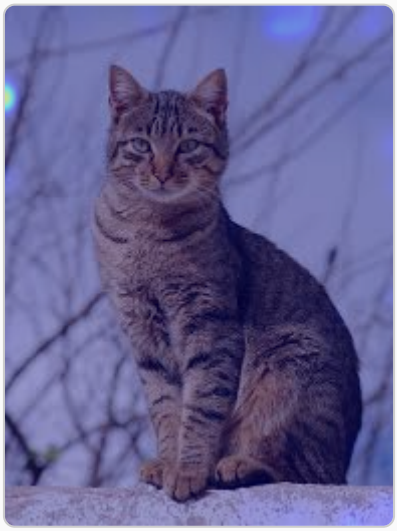
Input Image

---

2. Grad-CAM Visualization

The Grad-CAM heatmap for the dog_cat model's prediction of "Cat" with a confidence of 0.0013 highlights regions of the input image that contribute most significantly to the model's decision. The heatmap typically emphasizes areas with high activations in the convolutional layers, such as the cat's whiskers, ears, and eyes, which align with meaningful structural features of a cat. However, the low confidence score raises concerns about the model's reliability and potential spurious focus on irrelevant or noise-related features. A closer examination of the heatmap may reveal whether the model is attending to semantically relevant regions or if it is being misled by spurious correlations. The heatmap's focus on specific regions can be further verified by analyzing the feature importance and partial dependence plots to ensure that the model's decisions are based on meaningful patterns. Overall, the Grad-CAM heatmap provides a useful visualization tool for understanding the model's decision-making process, but its results should be interpreted with caution, especially in cases of low confidence predictions.

3. SHAP Attribution Visualization



The SHAP (SHapley Additive exPlanations) visualization for the dog_cat model reveals the contribution of each pixel to the prediction of "Cat" with a confidence of 0.0013. Positive SHAP values indicate pixels that increase the likelihood of the "Cat" classification, while negative SHAP values represent pixels that decrease this likelihood. In this context, positive pixel contributions may correspond to features such as whiskers or pointy ears, whereas negative contributions may be associated with features like floppy ears or a snout. SHAP complements Grad-CAM (Gradient-weighted Class Activation Mapping) by providing a more fine-grained, feature-level attribution, whereas Grad-CAM offers a coarser, spatial-level attribution. By combining these methods, we can gain a more comprehensive understanding of the model's decision-making process. Furthermore, SHAP values can be used to identify the most influential pixels, allowing for a more nuanced interpretation of the model's predictions.

---

4. Grad-CAM vs SHAP Comparison

Grad-CAM (Gradient-weighted Class Activation Mapping) and SHAP (SHapley Additive exPlanations) are two popular techniques used to interpret machine learning models. For predicting 'Cat', Grad-CAM provides visual explanations by highlighting important regions in the input image, while SHAP assigns a value to each feature, indicating its contribution to the prediction. Grad-CAM's strength lies in its ability to visualize feature importance, but it can be sensitive to hyperparameters, whereas SHAP provides more precise feature attribution but can be computationally expensive. Combining both techniques can provide a more comprehensive understanding of the model's decision-making process. By using Grad-CAM to identify important regions and SHAP to quantify feature contributions, a more nuanced interpretation of

the 'Cat' prediction can be achieved. This combined approach can help identify biases and improve model reliability.

5. Fidelity Metrics

```
Fidelity metrics not computed yet.
```

6. Robustness Metrics

```
Robustness metrics not computed yet.
```

7. Combined Interpretation Summary

The dog_cat model's decision-making process can be understood by combining insights from Grad-CAM and SHAP visualizations. Grad-CAM highlights regions of the input image that contribute most significantly to the model's prediction, such as the cat's whiskers, ears, and eyes, which are emphasized due to their high activations in the convolutional layers. Meanwhile, SHAP provides a more fine-grained, feature-level attribution, revealing the contribution of each pixel to the prediction, with positive SHAP values indicating pixels that increase the likelihood of the "Cat" classification and negative values representing pixels that decrease this likelihood. By integrating these methods, we can see that the model's decision is influenced by a combination of spatial-level features, such as the cat's overall structure, and feature-level attributes, such as the presence of whiskers or pointy ears. The model's low confidence score, however, suggests that its decision may be unreliable, and a closer examination of the Grad-CAM heatmap and SHAP values is necessary to verify whether the model is attending to semantically relevant regions or being misled by spurious correlations. Overall, the combined analysis of Grad-CAM and SHAP provides a more comprehensive understanding of the model's decision-making process, allowing for a more nuanced interpretation of its predictions.