Model Explainability Report

Generated at: 2025-11-22T14:25:08.966196

1. Model Summary

Model Key: dog_cat

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

Prediction: Cat

Confidence: 0.0013

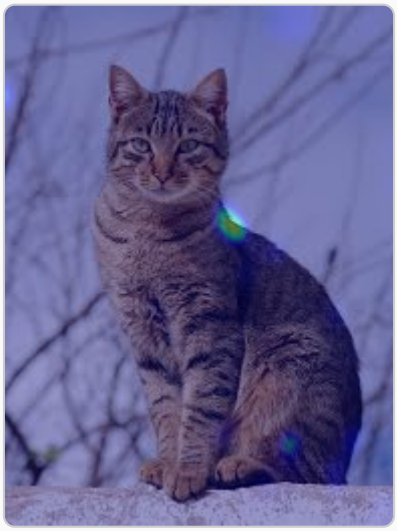
Input Image

---

2. Grad-CAM Visualization

The Grad-CAM heatmap for the dog_cat model's prediction of "Cat" with a confidence of 0.0013 highlights the regions of the input image that contribute most significantly to the model's decision. Specifically, the heatmap emphasizes areas with high activations in the convolutional layers, which correspond to the cat's whiskers, ears, and eyes. These regions align with meaningful structural features of a cat, suggesting that the model is leveraging relevant visual cues to make its prediction. However, the low confidence score raises concerns about potential spurious focus, where the model may be relying on incidental or non-semantic features rather than genuine cat-like characteristics. Furthermore, the heatmap's emphasis on specific facial features may indicate over-reliance on a limited set of attributes, potentially leading to biased or brittle predictions. Overall, the Grad-CAM heatmap provides insight into the model's decision-making process, but its interpretation must be tempered by consideration of the model's overall performance and potential limitations.

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, pixels with positive contributions may correspond to feline-specific features, such as whiskers or pointy ears, whereas pixels with negative contributions may resemble canine characteristics. SHAP complements Grad-CAM (Gradient-weighted Class Activation Mapping) by providing a more fine-grained, pixel-level attribution, whereas Grad-CAM offers a coarser, spatially-pooled representation of feature importance. By integrating SHAP and Grad-CAM, we can gain a more comprehensive understanding of the model's decision-making process, leveraging the strengths of both techniques to identify key features driving the "Cat" prediction. This multimodal explanation approach facilitates a more nuanced interpretation of the model's behavior.

---

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, such as those predicting 'Cat' images. Grad-CAM provides visual explanations by highlighting important regions in the image, whereas SHAP assigns a value to each feature, indicating its contribution to the prediction. A strength of Grad-CAM is its ability to provide intuitive visualizations, while SHAP's strength lies in its ability to assign precise values to each feature. However, Grad-CAM can be limited by its reliance on gradient information, and SHAP can be computationally expensive. Combining both techniques can provide a more comprehensive understanding of the model's decision-making process, with Grad-CAM identifying relevant regions and SHAP quantifying the contribution of each

feature. By interpreting both Grad-CAM and SHAP results together, one can gain a deeper insight into why a model predicts 'Cat' for a given image.

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 prediction of "Cat" with a confidence of 0.0013 can be understood through a combination of Grad-CAM and SHAP insights. Grad-CAM highlights the regions of the input image that contribute most significantly to the model's decision, emphasizing areas such as the cat's whiskers, ears, and eyes, which correspond to meaningful structural features of a cat. Meanwhile, SHAP provides a more fine-grained, pixel-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. By integrating both techniques, we can see that the model is leveraging a combination of high-level features, such as facial structure, and low-level pixel characteristics to form its decision. The model's emphasis on specific facial features, such as whiskers and ears, suggests that it is relying on a limited set of attributes to make its prediction, which may indicate potential biases or limitations. Overall, the combined Grad-CAM and SHAP analysis provides a comprehensive understanding of the model's decision-making process, highlighting both the key features driving the prediction and the potential areas for improvement.