

Model Explainability Report

Generated at: 2025-11-22T14:23:55.959296

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. 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 robust, generalizable patterns. Furthermore, the heatmap's emphasis on specific facial features may indicate over-reliance on a limited set of characteristics, 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 considered in the context 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, 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 both methods, we can gain a more comprehensive understanding of the model's decision-making process, with SHAP highlighting specific pixels driving the prediction and Grad-CAM identifying broader regions of interest. This synergy enables 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, while 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, but it can be limited by its focus on a specific class. SHAP, on the other hand, provides a more detailed, feature-level explanation, but can be more complex to interpret. 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 using both methods, a more accurate and nuanced interpretation of the 'Cat' prediction can be achieved.

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 broader regions of the input image that contribute to the model's prediction, 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, 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, such as whiskers or pointy ears. By integrating these insights, it becomes clear that the model is leveraging a combination of spatial-level features, as identified by Grad-CAM, and specific pixel-level contributions, as highlighted by SHAP, to form its decision. This synergy enables a more comprehensive understanding of the model's behavior, suggesting that it is relying on a mix of generalizable patterns and specific visual cues to make its prediction. Overall, the model's decision-making process appears to be driven by a complex interplay between broader regional features and finer-grained pixel-level contributions.