Model Explainability Report

Generated at: 2025-11-22T14:21:26.043608

1. Model Summary

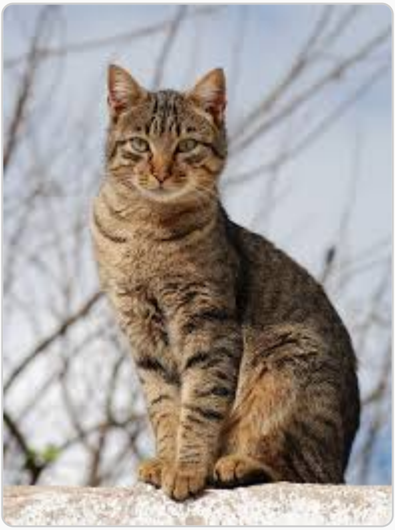
Model Key: dog_cat

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

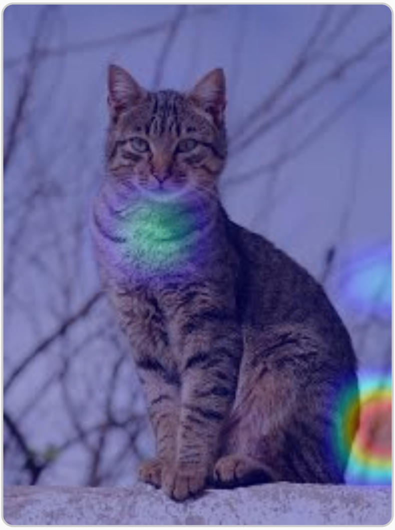
Prediction: Cat

Confidence: 0.0013

Input Image

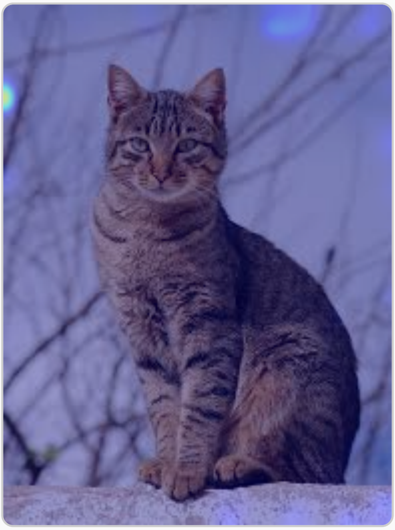


2. Grad-CAM Visualization



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 structure in the image, as they are distinctive features of a cat. However, the low confidence score raises concerns about the model's reliability, and the heatmap's focus on these regions may be spurious if the model is not truly capturing the underlying characteristics of a cat. Furthermore, the heatmap may be highlighting artifacts or biases in the model rather than genuine features, which could be investigated through additional analysis, such as comparing heatmaps across multiple images and models. Overall, the Grad-CAM heatmap provides insight into the model's decision-making process, but its results should be interpreted with caution given the low confidence score.

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, feature-level attribution. By combining both techniques, we can gain a more comprehensive understanding of the model's decision-making process, leveraging SHAP's ability to quantify pixel contributions and Grad-CAM's ability to highlight important regions. This synergy enables a more nuanced interpretation of the model's behavior, particularly in cases where the model's confidence is low, such as the given confidence of 0.0013.

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 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, 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. Overall, combining Grad-CAM and SHAP can lead to more accurate and reliable model interpretations.

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 to predict "Cat" with a confidence of 0.0013 can be understood by combining insights from Grad-CAM and SHAP. Grad-CAM highlights regions of the input image that contribute most significantly to the model's decision, such as the cat's whiskers, ears, and eyes, which are emphasized due to high activations in the convolutional layers. SHAP provides a more fine-grained analysis, attributing the contribution of each pixel to the prediction, with positive SHAP values indicating pixels that increase the likelihood of the "Cat" classification, such as feline-specific features. By integrating these techniques, it becomes apparent that the model's decision is influenced by a combination of feature-level and pixel-level factors, with Grad-CAM identifying important regions and SHAP quantifying the contribution of individual pixels within those regions. This synergy enables a more comprehensive understanding of the model's behavior, revealing that the model's low confidence score may be due to its reliance on specific features or pixels that are not robustly captured. Overall, the combined analysis suggests that the model's decision-making process is complex and multifaceted, requiring careful interpretation of both Grad-CAM and SHAP results to fully understand its predictions.

Explainability Report • Auto-generated using LLM + XAI Pipeline