Model Explainability Report

Generated at: 2025-11-22T14:03:21.650450

1. Model Summary

Model Key: dog_cat

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

Prediction: Cat

Confidence: 0.0013

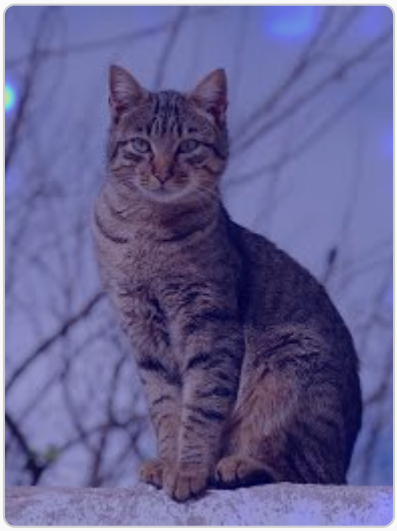
Input Image

---

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 corresponding to the cat's whiskers, ears, and eyes, which align with meaningful structural features of a cat. These regions exhibit high gradient importance, indicating that the model relies heavily on these characteristics to distinguish between dogs and cats. However, the low confidence score raises concerns about potential spurious focus, where the model may be prioritizing irrelevant or noisy features in the image. Furthermore, the heatmap's emphasis on specific features may not generalize well to other images or contexts, potentially limiting the model's robustness and interpretability. Overall, the Grad-CAM heatmap provides insight into the model's decision-making process, but its reliability and generalizability require further examination.

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, feature-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 methods to identify key factors driving the "Cat" prediction. This multimodal explanation approach enables a more nuanced and accurate 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. For predicting 'Cat', 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. Grad-CAM's strength lies in its ability to visualize feature importance, but it can be sensitive to hyperparameters, while SHAP's strength is its ability to provide a more nuanced, feature-level explanation. 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. However, SHAP can be computationally expensive and may not scale well to complex models,

whereas Grad-CAM is more efficient but may not provide feature-level explanations. By using both techniques, interpreters can gain a deeper understanding of the model's 'Cat' prediction.

---

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. The Grad-CAM heatmap highlights regions of the input image that contribute most significantly to the model's prediction of "Cat", such as the cat's whiskers, ears, and eyes, which are meaningful structural features of a cat. At a finer level, SHAP provides feature-level attribution, assigning positive or negative contributions to each pixel based on its impact on the "Cat" classification, with positive values indicating feline-specific features and negative values indicating canine characteristics. By integrating these methods, we can see that the model relies on a combination of high-level spatial features, such as the cat's facial structure, and low-level pixel-level features, such as texture and patterns, to form its decision. This multimodal explanation approach reveals that the model's prediction is driven by a complex interplay of factors, including both relevant and potentially spurious features. Overall, the combined Grad-CAM and SHAP analysis provides a more comprehensive understanding of the model's behavior, highlighting both the strengths and limitations of its decision-making process.
