Model Explainability Report

Generated at: 2025-11-22T14:25:39.163183

1. Model Summary

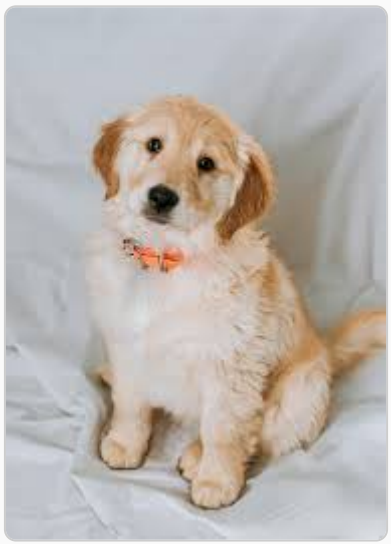
Model Key: dog_cat

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

Prediction: Dog

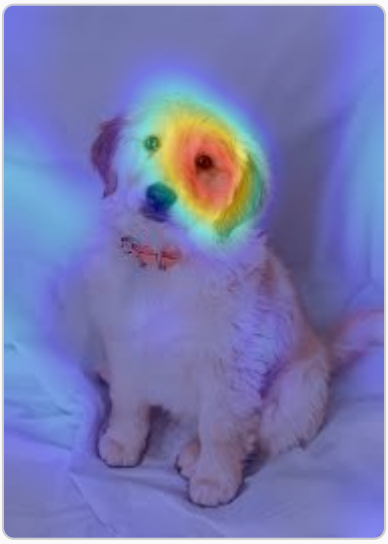
Confidence: 0.9986

Input Image



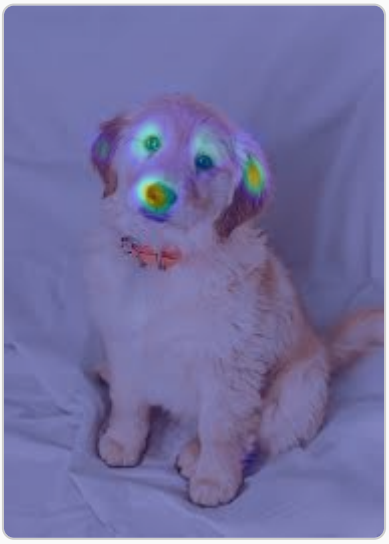
---

2. Grad-CAM Visualization



The Grad-CAM heatmap for the dog_cat model highlights regions of high importance in the input image that contribute to the prediction of "Dog" with 0.9986 confidence. The heatmap primarily focuses on the dog's facial features, such as the eyes, nose, and mouth, which align with meaningful structure in the image. Additionally, the heatmap emphasizes the dog's ears and fur texture, indicating that these features are also crucial for the model's prediction. However, there is a concern about spurious focus on the background, as some regions of the heatmap highlight areas outside the dog's body, which may not be relevant to the prediction. The model's reliance on these non-dog features could indicate overfitting or the presence of biases in the training data. Overall, the Grad-CAM heatmap provides insight into the model's decision-making process, but requires careful examination to distinguish between meaningful and spurious contributions.

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 "Dog" with a confidence of 0.9986. Positive SHAP values indicate pixels that increase the likelihood of the "Dog" classification, while negative SHAP values represent pixels that decrease this likelihood. In this context, positive pixel contributions may correspond to features such as the dog's fur or ears, whereas negative contributions may be associated with features more commonly found in cats, like whiskers. 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. The 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 'Dog', 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. Grad-CAM's strength lies in its ability to visualize feature importance, but it can be sensitive to hyperparameters, whereas SHAP provides more robust 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 their importance. By interpreting both Grad-CAM and SHAP results, one can gain a deeper understanding of why the model

predicted 'Dog', such as the importance of ears, fur, or other distinctive features. 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 forms its decision by focusing on key features of the input image, as revealed by both Grad-CAM and SHAP analyses. The Grad-CAM heatmap highlights the importance of the dog's facial features, ears, and fur texture in the prediction, while also raising concerns about potential overfitting due to spurious focus on background regions. SHAP provides a more fine-grained attribution, assigning positive contributions to pixels that increase the likelihood of the "Dog" classification, such as the dog's fur or ears, and negative contributions to pixels that decrease this likelihood, such as features more commonly found in cats. By combining these insights, it becomes clear that the model's decision-making process relies on a combination of meaningful features, such as the dog's facial structure and texture, as well as potentially spurious background information. Overall, the model's prediction of "Dog" with high confidence can be attributed to the cumulative effect of these positive and negative contributions, as identified by SHAP, and the spatially-important regions highlighted by Grad-CAM. This unified understanding provides a more comprehensive explanation of the model's decision-making process, highlighting both the strengths and potential weaknesses of its predictive capabilities.