Project proposal:
CNN Interpretability for Classification

Problem Statement, Motivation, and Relevance

The project addresses the "black box" nature of deep CNNs, particularly in image classification. While CNNs like VGG-16 achieve high accuracy, it is often unclear why they arrive at a particular prediction. This lack of transparency is a critical issue in sensitive domains such as medical imaging or autonomous driving, where trust and accountability are paramount. The motivation is to enhance the reliability and trustworthiness of CNN models by opening up the black box. This project is highly relevant as it directly tackles the need for Explainable AI, moving beyond simple performance metrics to provide human-understandable insights into model behavior.

Background, Datasets, and Methodology

Key background readings will focus on the original papers for VGG-16, Grad-CAM, and LIME.

The primary dataset used will be a manageable, domain-specific dataset like CIFAR-10 or a subset of ImageNet to ensure the interpretability stage is computationally feasible.

Planned Methodology

1. **Model Training**: A pre-trained model will be fine-tuned or trained from scratch on the chosen dataset for the classification task.
2. **Interpretability Implementation**: The project will implement Grad-CAM. Grad-CAM uses the gradients of the target class flowing into the final convolutional layer to produce a coarse localization map, highlighting the important regions in the input image for predicting the concept.
3. **Visualization and Analysis**: The resulting heatmaps will be overlaid onto the original images. The analysis will focus on examples where the model is correct and, crucially, where it makes a mistake, to determine if the error is due to focusing on spurious or irrelevant features.

Evaluation and Unique Contribution

Evaluation

- **Quantitative:** The model's classification accuracy and other standard metrics (Precision, Recall, F1-score) on the test set will be measured to ensure it is a competent classifier.
- **Qualitative:** The quality of the Grad-CAM heatmaps will be qualitatively assessed. This involves visually verifying whether the highlighted regions logically correspond to the true features of the classified object.

Unique Contribution

The unique contribution lies in the critical, feature-level analysis. Rather than merely implementing a standard technique, the depth will come from an in-depth case study of misclassified images using the Grad-CAM visualizations to diagnose specific feature-reliance issues, such as bias towards background textures, that explain the model's prediction errors.