Can't See the Forest for the Trees: Diagnosing the Spatial Blindness in CNN-based Models via Grad-CAM
Project Milestone

Yixuan Zheng

School of EECS, Peking University
2400011003@stu.pku.edu.cn

1 Introduction

Deep Convolutional Neural Networks (CNNs) have achieved remarkable performance in visual tasks but remain **black boxes** regarding their decision-making logic. Traditional visualization methods like CAM and Grad-CAM primarily focus on "where" the model looks for their prediction. However, being able to identify an object does not guarantee an understanding of its spatial context or structural integrity.

My project aims to go beyond simple visualization. I divide my project into three stages:

- **Stage 1: Foundation and Quantification.** I will implement CAM and Grad-CAM algorithms and establish faithfulness-based evaluation metrics to provide a rigorous analytical tool for the following experiments.
- **Stage 2: Diagnostic Analysis of Spatial Perception.** I design a suite of *spatial ability tests*, including **patch shuffling** (using ResNet-50) and **cross-modal spatial queries** (using CLIP and MDETR). These experiments are intended to show the model's inherent *"Spatial Blindness"*.
- **Stage 3: Grad-CAM guided Adversarial Attack.** Leveraging the insights from earlier stages, we can actually do something interesting. I plan to use Grad-CAM to extract highly discriminative texture signatures from a source class (e.g., dog) and inject them into the salient regions of a target image (e.g., cat). Given the "Spatial Blindness" uncovered in Stage 2 experiments, these localized, imperceptible perturbations are expected to effectively hijack the model's global classification decision.

2 Problem Statement

The core problem of this project is to differentiate and quantify the extent to which CNN-based decisions are driven by **local feature recognition** (e.g., texture counting) versus **global structural reasoning** (e.g., spatial logic understanding) via Grad-CAM attribution.

- **Datasets:**
  - **CIFAR-10:** Used for the baseline implementation of CAM/Grad-CAM and sanity-check of the evaluation metrics.
  - **Custom Dataset:** A set of high-resolution images featuring identical objects with controlled spatial relations (e.g., *on/under/higher than*).
- **Evaluations & Expected Results:**

- **Metric for Faithfulness (Stage 1):** I will implement **Deletion metrics** to measure how "faithful" the explanation is to the model's inner reasoning. I anticipate a **sharp decline** in the target class logit score as only a small fraction of important pixels are removed.
- **Metric for Structural Awareness (Stage 2):** I will analyze the heatmaps generated by Grad-CAM to inspect the model's visual attention. If a model ignores the spatial logic, the heatmap will highlight features even if they appear in an incorrect topological structure or in regions that contradict our specified spatial constraints.
- **Metric for Adversarial Attack (Stage 3):** I define the **Hijacking Success Rate (HSR)** as the percentage of images where a localized perturbation successfully flips the global class label while maintaining human-perceived visual consistency.

3 Technical Approach

- **CAM and Grad-CAM Attribution:** Implemented as the primary diagnostic probe to extract visual attribution maps, identifying decision-critical local regions (textures).
- **Patch Shuffling:** Images are partitioned into $n \times n$ grids and randomized to disintegrate global topology.
- **Grad-CAM guided PGD:** A novel hijacking method that constrains PGD perturbations within Grad-CAM identified masks. More details will be discussed in the final report.

4 Intermediate/Preliminary Results

I have obtained several critical findings that validate my research direction:

4.1 Patch Shuffling

In the patch shuffling test on a Golden Retriever sample (using ResNet-50), I found a counter-intuitive phenomenon: The 99% confidence on a scrambled image proves that the CNN acts as a **texture counter**

State	Prediction	Logit	Confidence
Original	Golden Retriever	6.89	46.34%
Shuffled (4x4)	Golden Retriever	**12.18**	**99.00%**

rather than a **shape recognizer**.

4.2 Spatial Logic Failure in Multi-modal Models

To evaluate the model's understanding of spatial logic, I designed a contrastive test using an image with a red apple on the left and a green apple on the right. Testing on **CLIP-ResNet-50**, I observed a significant **wrong result**:

Query Text	Logit	Confidence
the red apple is on the left of the photo (True)	24.42	28.15%
the green apple is on the left of the photo (False)	**25.36**	**71.88%**

The results show that the model assigned a dominant 71.88% confidence to the **incorrect** spatial description. This suggests that while the convolutional backbone successfully extracts color and object features, the multimodal alignment layer fails to perform **semantic binding** between "color" and "relative position". Instead, the decision is determined by the global feature resonance of the "green apple" texture, which happens to be slightly stronger in the latent space, completely overriding the spatial constraint "on the left".