Look, Ask, Explain: The VQA Challenge

Week 3 Assignment

December 31, 2025

Objective

Deep learning models for vision tasks often achieve high accuracy but remain difficult to interpret. The objective of this assignment is to explore **Explainable AI (XAI)** techniques to understand **what convolutional neural networks learn and where they focus** while making predictions.
You will visualise:

- Learned convolutional filters
- Feature maps
- Grad-CAM heatmaps highlighting important image regions

This assignment emphasises **interpretability over performance**.

Learning Outcomes

By completing this assignment, you will be able to:

- Understand why explainability is important in deep learning
- Visualise convolutional filters and intermediate feature maps
- Apply Grad-CAM to CNN-based vision models
- Interpret model predictions using visual explanations

Dataset

Use **one** of the following datasets:

- CIFAR-10 (recommended)
- CIFAR-100

You may use pretrained models if computational resources are limited.

Problem 1: Vision Model Setup

Task

Select a CNN-based vision model:

- ResNet-18 / ResNet-50
- DenseNet121
- MobileNetV2

Train the model for the classification task.

Requirements

- Load a pretrained model
- Fine-tune only the final classification layer (optional)
- Perform inference on test images

Output

- Model architecture summary
- Sample predictions on test images

Problem 2: Visualizing Learned Filters

Task

Visualise the learned filters from:

- The first convolutional layer
- A deeper convolutional layer

You can do that by extracting convolutional weights, normalising the filter values, and displaying them as images.

Analysis

Briefly explain:

- What low-level patterns the filters capture (edges, textures, colours)
- How filter complexity changes with depth

Problem 3: Feature Map Visualisation

Task

For a given input image:

- Extract feature maps from an intermediate convolutional layer
- Visualise multiple channel activations

Output

- Original image
- Corresponding feature maps

Discussion

Explain how feature maps respond to different image regions.

Problem 4: Grad-CAM for Model Explainability

Task

Implement **Grad-CAM** to explain model predictions:

- Identify the target convolutional layer
- Compute gradients of the predicted class score
- Generate class activation maps

Visualization

Overlay Grad-CAM heatmaps on input images.

Output

- At least 3 correctly classified images
- Corresponding Grad-CAM heatmaps

Interpretation

For each image, explain:

- Which regions influenced the prediction
- Whether the explanation aligns with human intuition

Problem 5: Failure Case Analysis

Task

Select at least **one misclassified image** and:

- Generate its Grad-CAM explanation
- Analyse why the model failed

Deliverables

1. **README.md** describing:
    - Dataset and model used
    - Explainability techniques implemented
2. **Report (PDF, 2–3 pages)** including:
    - Filter visualisations
    - Feature maps
    - Grad-CAM results
    - Interpretation and insights

Bonus (Optional)

- Compare Grad-CAM outputs across layers
- Apply Grad-CAM to a Vision Transformer
- Compare explanations for correct vs incorrect predictions

Happy Learning :)