Explainable Deep Learning

10/6/2025

60 Points Possible

Attempt 1

In Progress

NEXT UP: Submit Assignment

Add Comment

Unlimited Attempts Allowed

Details

Explainable Deep Learning

Instructions

In this assignment, you will work with pretrained deep learning models to investigate model explainability in computer vision. Your objective is to apply GradCAM and at least two of its variants to a meaningful image classification problem of your choice and analyze how and why the model makes its decisions.

You are encouraged to select an image classification task that holds personal or societal significance. Potential areas include, but are not limited to: wildlife conservation, road safety, public health, environmental sustainability, or social impact. You may use an existing public dataset or a curated subset, and pretrained models such as ResNet-50 or Vision Transformers (ViT).

Tasks

- Choose an image classification problem relevant to you (e.g., wildlife detection, object recognition in autonomous driving, or recycling classification).
- Use a pretrained computer vision model (e.g., ResNet-50, ViT) for your classification task. Transfer learning is optional.
- Apply Explainability Techniques:
  - Implement GradCAM and at least two GradCAM variants
  - Apply these techniques to at least 5 images from your dataset
- Generate and present visualizations showing what regions of the image the model is focusing on for its predictions.
- Compare and contrast the attention maps generated by GradCAM and its variants.
- Reflection:
  - Discuss the visual cues the model attends to
  - Comment on any surprising or misleading behavior
  - Reflect on why model explainability is important in your selected application domain

Example Problems (Using Pretrained Models Only)

- **Wildlife Detection in Conservation**
  - *Task*: Classify zebras vs. other animals in camera trap images
  - *Purpose*: Supports biodiversity monitoring and anti-poaching efforts
- **Object Recognition in Autonomous Driving**

- *Task*: Classify traffic objects such as stop signs or pedestrians
- *Purpose*: Explainability builds trust in safety-critical systems

- **Recycling Sorting**
  - *Task*: Classify recyclable objects into types
  - *Purpose*: Encourages sustainable waste management and environmental safety

---

Submission

Submit a link to a GitHub repository containing a Google Colab notebook with your implementation and analysis. Your notebook must:

- Implement GradCAM and at least two variants
- Visualize attention maps on at least 5 different images
- Include comparative analysis of GradCAM methods
- Contain a reflection addressing:
  - Whether the model focused on appropriate cues
  - Any surprising or misleading results
  - The significance of explainability in your chosen domain

Ensure your repository follows best practices and clear documentation for running your code.

Rubric

Notebook (60 points)

- GradCAM and at least two variants are implemented correctly
- At least 5 images are used with visualizations
- Comparative analysis between GradCAM and variants is insightful and well-documented
- Clear commentary on model focus regions and method differences
- A thoughtful reflection is provided, addressing:
  - Model attention appropriateness
  - Surprising/misleading examples
  - Explainability importance in your task domain

---

p 0 words </>

(https://canvas.duke.edu/courses/62464/modules/items/683156)

Assignment

(https://canvas.duke.edu/courses/62464/modules/items/572840)