CIS 620, Advanced Topics in Deep Learning, Spring 2024

Homework 10: Grad-CAM for Explainable AI
Due: Monday, April 8, 11:59pm
Submit to Gradescope

Learning Objectives

In this assignment, you will:

- Use Grad-CAM to interpret the results of an image classification task
- Analyze the effectiveness of Grad-CAM as tool for Explainable AI

Deliverables

This is a **pair** assignment for both the written and coding portions.

1. **A PDF with your name in the agreement**

   Copy and edit this google doc to enter your names in the Student Agreement and answer the questions mentioned below. **Don't forget to add your team member on gradescope submission!**
2. **hw10.ipynb file with the functional code**

   Complete the coding assignment in the Jupyter Notebook and upload the file to Gradescope. Please leave clear and interpretable results in the submission, so we know you have completed the tasks. **Don't forget to add your team member on gradescope submission!**

Homework Submission Instructions

Written Homeworks

All written homework must be submitted as a PDF to Gradescope. **Handwritten assignments (scanned or otherwise) will not be accepted.**

Coding Homeworks

All coding assignments will be done in Jupyter Notebooks. We will provide a .ipynb template for each assignment as well as function stubs for you to implement. You are free to use your own installation of Jupyter for the assignments, or Google Colab, which provides a Jupyter Environment connected to Google Drive along with a hosted runtime containing a CPU and GPU.

For this assignment, we have provided the code for dataset creation from a single image. The input image is provided in the folder and can also be accessed by the code for dataset creation.

Questions

Reference: Github Repo

Coding

1. Baseline model
    a. Initialize a pre-trained ResNet-18 model on CIFAR-10 Dataset from HuggingFace
2. Interpretation with Grad-CAM
    a. Use the reference library to create a Grad-CAM model
    b. Apply Grad-cam to both correct and incorrect predictions [Note: Please leave ***understandable results*** of your implementation in your submission]

Discussion

1. Baseline
    a. What is your baseline classification performance?
        i. This is not a homework focus, so NO NEED to spend too much effort on improving accuracy.

    We use a pre-trained huggingface Resnet18 model. The accuracy on the cifar10 testing set is 92% (replicated on the notebook).

2. Grad-CAM results
    a. Give 5 examples of correctly classified samples and their respective Grad-CAM explanations. What patterns have you observed?
    The Grad-CAM explanations show the model correctly uses the features of the object we want to classify.

    b. Give 5 examples of incorrectly classified samples and their respective Grad-CAM explanations. What patterns have you observed?
    If the Grad-CAM explanations are correctly reflecting the behavior of the model, the Resnet18 model focus on the partial features of the target object. This leads to the misclassifications of those images.

    c. What does the experiment suggest about Grad-CAM? Did you find it useful? Explain how you would use it or why it isn't useful
    Grad-CAM is a good tool to find out why the model doesn't work and provide an explanation of which parts of the data it used. For the Resnet18 example in the HW, we can directly find out that most of the failure cases are due to the partial features in the image, which gives a hint on how to improve the model.

3. More about Grad-CAM
    a. Read about the various method choices in the reference Repo's README.md. Pick two other methods and explain how they differ from the original Grad-CAM.
    XGradCAM: Instead of weighing the layers using the average gradient, the XGradCAM first scales the gradient by the normalized activation values.

    GradCAM++: It uses the second-order gradient instead of the first-order gradient used in the GradCAM.
    b. How are the Grad-CAM outputs the same or different from Shapley values?
    Grad-CAM provides a more global explanation on how the features are used by the model, while the Shapley value focus more on the contribution of a single pair of Input and Output.