DAI Programming Assignment - 2
Bias and Explainability

Deadline: 5 April 2023
Max. Marks: 300 + 30 Bonus

Assignment Guidelines

1. Any kind of plagiarism is not accepted. We will strictly follow institute policies for plagiarism. 2. Recommended programming languages: Python + PyTorch.
3. You may use any external libraries or GitHub codes. However, the evaluation will test your knowledge of the algorithm and the choice of hyperparameters. Do cite the libraries/codes.

Assessment criterion

The assessment will be done on the basis of the following components:
1. Working codes
2. Analysis and clarity of results (drawing comparisons across different parts) and clarity of the report. 3. Understanding the theoretical concepts and the choice of hyperparameters.

Submission Guidelines

1. A single report(.pdf) for all questions.
2. Mention all the relevant results, comparisons as asked or wherever required for better understanding for the results.
3. A single zip file containing the report, codes and readme if required. The zip file should be named as **Rollno_PA2.zip**.

Q1. Bias and Explainability

Dataset: Download the Crisis-MMD dataset and use these 5 classes *(Hurricane Harvey, California wildfires, Mexico earthquake, Iraq-Iran earthquake, and Sri Lanka floods)* for Task 2: Humanitarian categories from **this link.**

Algorithm: Use a pretrain Multimodal ViLBERT/VisualBERT/Densnet-BERT model [1] for the above dataset. (*You can also modify the architecture to perform the classification.)*

Training: Train the model on images + text provided in the dataset.
Testing: Test the model on a test dataset.

Evaluation: Perform a 5-class classification (select any 5 classes from Task-2) and report the performance as follows:

A. Finetune the given model for the dataset provided above and report the performance in terms of accuracy, confusion matrix, ROC, etc.**[20 Marks]**
B. Train the given model from scratch for the above dataset and report the performance in terms of accuracy, confusion matrix, ROC, etc..**[30 Marks]**
C. Report the bias in the whole pipeline: .**[40 Marks]**
    a. Do you think there is any kind of bias in the dataset provided? If so, report that with the use of 2 qualitative (in terms of graphs) and 2 quantitative (in terms of number) metrics..**[15 Marks]**
    b. Do you think there is any kind of bias in the algorithm used in **Part A and B**? If so, report that with the use of 2 qualitative (in terms of graphs) and 2 quantitative (in terms of number) metrics..**[15 Marks]**
    c. Why do you think both approaches (**Part A and B**) show bias? Report detailed analysis on this point. If you refer to some sources for the justification please cite that in the report.**[10 Marks]**

D. Evaluate the performance of the given dataset on the LSTM-CNN model. (you need to train the model from the scratch.).**[20 Marks]**

E. Do you think bias is dependent on the model architecture? Explain this point by analyzing the results of all the models **(Part B and D)**. **[10 Marks]**

F. **(Bonus)** Come up with 2 new evaluation metrics to detect if there is a bias in the system and compare the results of Part A and B with it. **[20 Marks]**

G. Select 2 samples from each correctly and incorrectly classified class by the trained models **(Part B and D)**. **[50 Marks]**

   a. Apply LIME and GradCam++ on image data to visualize most salient regions being used for prediction. **[25 Marks]**

   b. Apply SHAP on the text data to visualize most important features used for prediction.**[25 Marks]**

H. Select 2 model specific methods to explain the output of the before and after fusion layer of the model used in **Part D**. **[30 Marks]**.

I. **(Bonus)** What do you think on the point that 'Can explainability be measured in a quantitative way?''. If you refer to some sources for the justification please cite that in the report. **[10 Marks] Total: 230 Marks (200 + 30 Bonus)**

Q2. Bias Mitigation

A. Select a recent paper on a **state-of-the-art performing model** for any one of the tasks below:

   a. Computer vision tasks

   b. The intersection of Cognitive Science and Computer vision tasks

   Reproduce the results on any one dataset mentioned in the paper. **[10 Marks]**

B. Do you observe any bias? Explain the type of bias you observed. **[10 Marks]**

C. Try to mitigate the bias using the bias mitigation technique. In this, you have to **select the paper related to bias mitigation** and use it to mitigate the bias you found in **part B**. Report the metrics values used in the paper.
   You are advised to select a **cognitive bias mitigation paper to mitigate the cognitive bias in the computer vision task you select in part A. [20 Marks]**

D. Try another approach of your own to mitigate the bias using two techniques:

   a. **DATA method (Pre-Processing):** You may use any of the pre-processing techniques to achieve your aim.

   b. **ALGORITHMIC method:** You can alter the loss function or use a multi-tasking approach to achieve the goal.

   Report the values of the same metric you used in part C for these techniques also. **[10+10 Marks]**

E. Compare the bias mitigation techniques used in parts C and D(a), and D(b) by taking in support of bias metrics. **[20 Marks]**

F. Report the changes you observed before and after applying bias mitigation techniques. **[20 Marks]**

Write your selected papers for **part A and part C** and their venues, respectively, in the sheet **by the end of 23 March 2023.**
Sheet link - link

Total: 100 Marks