CSE641 Deep Learning
Assignment-4 [70 marks]
Deadline: 21st April 11:59pm

General Instructions:

1. Each group member must do at least one of the following tasks. But all should know the working of all the tasks. (Recommended: Divide the sections among yourselves.)
2. For Plagiarism, institute policies will be followed strictly.
3. **Make sure to use Pickle or any other library to save all your trained models. There will not be enough time during the demo to retrain your model. This is a strict requirement.** You must upload your best models on Classroom to reproduce your results during the demo. If you cannot reproduce your results during the demo, no marks will be given. You are advised to prepare a well-documented code file.
4. You must submit Output.pdf, Code files (including both .py files and .ipynb files), and models dumped after training. Mention your sample outputs in the output.pdf. The Output.pdf should enlist all hyperparameters clearly for quick reference.
5. Submit code, models, and output files in ZIP format with the following name: A1_Member1_Member2_Member3.zip

DATASET for TASK I, II & III

- **Dataset description:** A dataset of hateful and not hateful memes is provided in below link:

https://hatefulmemeschallenge.com/#download. Fill the form to download zip.

- **Train**: train.jsonl/ **Val**: dev_seen.jsonl/ **Test**: **test_seen.jsonl** Use test_seen (which has labels)
- For reference:
  - https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/
  - https://aclanthology.org/2022.findings-naacl.118.pdf
- Familiarize yourself with the dataset and understand of what makes a meme hateful. Perform the necessary data cleaning/ preprocessing steps. In case some image is missing corresponding to jsonl, ignore it.
- Use any Deep Learning libraries like Pytorch or TF to develop deep learning model.
- DO NOT CHANGE THE TEST SAMPLES SIZE FOR PERFORMANCE COMPARISION.
- If you reduce the number of train for computational purpose, it should be stratitifed proportionally, and stated clearly in the report.

TASK I (Unimodal: Image-Only) [20 marks]

Image-only hateful meme detection

Task: Develop **image-only detection system** to classify memes as **hateful** or **not hateful**. Consider image as a whole and do not treat text as separate component in it.

Use train/val/test sets provided in jsonl format in link shared above.

1. Pre-process the images using atleast 2 techniques of your choice as converting into appropriate format, normalization, gray-scaling etc. Make sure that your images are in the appropriate format for your chosen model. **[2 marks]**
2. Propose and implement an **image-only** model for classification using any deep learning image classification model of your choice such as VGG, Vision Transformer Image-only, ResNet, etc or you can build your own CNN model. The model selected should learn meaningful features from images and be effective for image-only classification tasks. **[12 marks]**
3. Generate the following plots: **[3 marks]**
   - Loss plot - Training Loss and Validation Loss V/s Epochs.
   - Accuracy plot - Training Accuracy, Validation Accuracy V/s Epochs
   - Analyze and Explain the plots obtained
4. Report the overall Accuracy, Precision, Recall, F1 score for your test set. Also, report class-wise precision and recall and F1 score for test set. **[3 marks]**

TASK II (Unimodal: Text-Only) [20 marks]

Text-only hateful meme detection

Task: Develop **text-only detection system** to classify memes as **hateful** or **not hateful**. Use train/val/test sets provided in jsonl format in link shared above. Text for the meme can be used from jsonl annotation file or be extracted from meme image using OCR or other techniques. Use **only text** for detection.

1. Data preparation: Preprocess the text extracted by cleaning, tokenizing, and converting it into a representation that can be used as input to the deep learning model. **[2 marks]**
2. Propose and implement a **text-only** model for classification using any deep learning model of your choice such as BERT, LSTM, XLNet, etc. **[12 marks]**
3. Generate the following plots: **[3 marks]**
    - Loss plot - Training Loss and Validation Loss V/s Epochs.
    - Accuracy plot - Training Accuracy, Validation Accuracy V/s Epochs
    - Analyze and Explain the plots obtained
4. Report the over Accuracy, Precision, Recall, F1 score for your test set. Also, report class-wise precision and recall and F1 score for test set. **[3 marks]**

TASK III (Multimodal: Image+ Text-Based Classification) [30 marks]

Multimodal (Visuals & Language) hateful meme detection:

Task: Develop **Multimodal detection system** to classify memes as **hateful** or **not hateful**. Use train/val/test sets provided in jsonl format in link shared above.

1. Select an appropriate deep learning model architecture for joint image and text classification, such as a multimodal fusion model (early or late fusion). Apply appropriate preprocessing. You can employ any combination of CNN, LSTM, or pretrained transformer models or use some multimodal model directly. Your base image, text model, and fusion technique should be clear. In report, explain in 3-4 lines along with a figure, the proposed architecture to handle multi-modal data. **Your multimodal system should perform better in terms of Accuracy and F1 score than both image-only and text-only models**.

    [10+4+2 marks] for the proposed multimodal model, improvement over unimodal and explanation respectively.
2. Generate the following plots: **[3 marks]**
    - Loss plot - Training Loss and Validation Loss V/s Epochs.
    - Accuracy plot - Training Accuracy, Validation Accuracy V/s Epochs
    - Analyze and Explain the plots obtained
3. Report the overall Accuracy, Precision, Recall, F1 score for your test set. Also, report class-wise precision and recall and F1 score for test set. **[3 marks]**
4. Visualize the following high-dimensional features into lower-dimensional space for hateful/not hateful classification in test set using T-SNE plots. You can subsample few test samples for this part and this part only, use atleast 50 hate, 50 non-hate samples but same samples used in all 3 model comparison. The features obtained here will be the last embedding layer of your respective task model (before classification layer):
    - Image-only features (Task I)
    - Text-only features (Task II)
    - Joint Multi-modal (Text + Image) features (Task III)

    What can be said about the results obtained from unimodal models (Task I & II) vs multi-modal model (Task III)? **[8 marks] 2 for each plot + 2 for overall explanation.**

Please note: This is a computationally intensive task and may require significant time and resources to complete. Therefore, it is advisable to start the project well in advance and plan accordingly.