Lab Assignment 3

1 Overview

Welcome to the third and final lab assignment for Neural Networks and Deep Learning! Please be sure to **read this document in its entirety**. In this assignment, you will participate in a friendly competition with your classmates to design the best multi-modal visual question answering (VQA) architectures. The dataset in question is VizWiz, which provides visual and textual inputs for your models. There are two distinct challenges: (1) binary classification to decide if a visual question is answerable, and (2) predicting answers to visual questions. Both modalities (images and text) *must* be used for full credit on this assignment, which is worth a total of **50 points**. A variety of extra credit opportunities are provided based on the competition results.

2 The Dataset

The VizWiz dataset contains images taken by people with visual impairments alongside text questions about the images. There are 20,523 training samples, 4,319 validation samples, and 8,000 test samples. Each sample contains a single image, a text question, a binary label where 0 indicates the question is not answerable and 1 indicates the question is answerable, and 10 text answers to the posed question. The ground truth labels are only provided for the train and val splits; the test labels are hidden! The data is hosted online, and can be accessed using python. You are also permitted to leverage the code for interacting with the dataset provided in Coding Tutorial 11.

2.1 Dataset Processing

The dataset samples obtained from the train, validation, and test files are *not* pre-processed. Before working to design models, take some time to familiarize yourself with the data and develop a pre-processing strategy. You are allowed to use any existing functions you wish for the pre-processing step, including examples covered in the programming tutorials. Use at least $1,000$ training samples when developing your models. Note that you need not use the entire training set, although any amount of data (greater than $1,000$ samples) is permitted.

3 Designing Multi-Modal Architectures

Once the dataset has been pre-processed, your next task is to design *two* architectures; one for each challenge. Both models should take the same inputs, namely, the visual and textual data. For the binary classification challenge, the model should output a 0 or 1 prediction. For the text generation challenge, the model should output text answers. You are allowed to use any existing modules or building blocks you wish. However, existing architectures are **NOT ALLOWED**. You must design and train your own architectures. While you are permitted to reference the example architecture in Coding Tutorial 11, you **must** still develop your own model. Any submission using this example model without

significant changes will receive *no credit*. As you work on designing your architecture, use the validation samples to refine the model and tune hyper-parameters.

4 The Competition

4.1 Challenge 1: Binary Classification

For this challenge, develop an architecture to leverage information from both the image and the text question. The model should provide binary labels, where 0 indicates the question is not answerable and 1 indicates the question is answerable. Submit your model's predictions on the first 100 test samples (indices 0 up to and including 99 from the "test" dataset split) as a .pkl file. Use the "torch.save()" function to write a tensor directly to disk. The tensor should be 1-dimensional (a vector), and contain your model's prediction (0 or 1) on the i^{th} sample in the i^{th} coordinate. Submissions will be evaluated based on the classification accuracy metric defined below:

$$accuracy_{cls} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

where TP, TN, FP, FN denote the true positives, true negatives, false positives, and false negatives, respectively. Name your .pkl file as "firstname_lastname_challenge1.pkl".

4.2 Challenge 2: Answer Prediction

For this challenge, develop an architecture to output a text answer to the posed question. As with challenge 1, the model should utilize both the visual and textual information to accomplish this. Submit your model's predictions on the first 100 test samples (indices 0 up to and including 99 from the "test" dataset split) as a .json file. The file should contain a list of dictionaries in the format discussed in Coding Tutorial 11. Recall, the .json file you submit should have the form:

[{"image": *Image_Url*, "answer": *Model's Text Output*}, ...]

Submissions will be evaluated based on a human answer accuracy metric, which is defined as:

$$accuracy_{ans} = min\left(\frac{\#\text{ of humans that provided that answer}}{3}, 1\right) \tag{2}$$

For example, if a model predicts "apple" as the answer to a visual question, and at least three of the human answers to the question were also "apple", then the model receives an accuracy of 1.0 for that sample. If only two of the human answers were "apple", then the model would receive an accuracy of 0.67 for that sample. You can find the full implementation details of this metric in Coding Tutorial 11. Name your .json file as "firstname_lastname_challenge2.json".

4.3 Extra Challenges

For each task, extra credit will also be awarded to the best single-modality models. If you so choose, you may submit additional files for challenges 1 and 2 obtained from models trained and evaluated on just the images or just the text features. Note that these

extra challenges are not required for full credit. You are permitted to design different architectures for the single modality challenges; you need not use the same architecture as the multi-modal challenges.
Name your files as "firstname_lastname_challenge1_visual.pkl" or
"firstname_lastname_challenge1_text.pkl" for challenge 1, and,
"firstname_lastname_challenge2_visual.json" or
"firstname_lastname_challenge2_text.json" for challenge 2.

4.4 Prizes

- Challenge 1: 5/3/2 points of extra credit for $1^{st}/2^{nd}/3^{rd}$ place, respectively.
- Challenge 2: 5/3/2 points of extra credit for $1^{st}/2^{nd}/3^{rd}$ place, respectively.
- Extra Challenges: 3/2/1 points of extra credit for $1^{st}/2^{nd}/3^{rd}$ place, respectively, on each task and modality. For example, there are separate awards for challenge 1 visual only, and challenge 1 text only.

You will only be awarded extra credit for winning one challenge. In the event you win multiple challenges, you will still be recognized but the extra points will be awarded to the runner-up.

5 Submitting the Report

The report for this lab differs in structure from the previous assignment. You are not required to follow the "methods, results, analysis" format. Rather, provide a 2-4 page discussion of the following topics:

1. Details of the data pre-processing step. How was the dataset pre-processed? Why did you make the design choices you did? How much training data was used and why?
2. Details of the multi-modal architectures. Which types of layers did you use? How do the models incorporate both modalities? How do they handle the different types of output?
3. Details of model adjustment and hyper-parameter tuning using the validation set. Did the model development process reveal interesting trends about architecture design? Which hyper-parameters were the most important? How did you decide on the final versions?
4. Details of the single modality architectures (if applicable). If you chose to participate in the text-only or image-only challenges, describe how your architectures were designed. Are they the same as the multi-modal architecture? If not, how and why are they different?

Please submit a pdf named with your first and last name: firstname_lastname.pdf. A successful submission will consist of two self-contained, separate contributions. First, it will include a report with the above topics addressed **as the first part of the PDF file**, each broken out into a separate section. Second, it will include the source code of your implementation **as the second part of the PDF file** (portions indicated by "Code").[1]

[1]We require submitting the code as a PDF

We will only review the code when the report does not contain sufficient detail, in order to provide partial credit. To avoid this **automatic deduction**, please ensure your report properly conveys all requested information.

Collaboration versus Academic Misconduct: Collaboration with other students and AI is permitted, but the work you submit must be your own. Copying/plagiarizing work from another student or AI is not permitted and is considered academic misconduct. For more information about University of Colorado Boulder's Honor Code and academic misconduct, please visit the course syllabus.