COM2028 Artificial Intelligence
Semester 2, 2023-24

Coursework Guideline

Prof Gustavo Carneiro, Dr Zhenhua Feng

8th March 2024

1 Overview

This coursework concerns the automated classification of images through machine learning techniques. You will work on breast ultrasound data, where training and testing samples and their ground truth are provided. You will develop suitable classification techniques, a cross-validation experiment, hyper-parameter tuning strategy, an evaluation using specific metrics (accuracy, area under the receiver operating characteristic curve, area under the precision recall curve, precision, recall, and F1 score), visualise model performance through relevant plots (e.g., confusion matrix, receiver operating characteristic curve, precision recall curve), and a poster to introduce the problem. You will need to submit a Jupyter notebook containing your code for training, visualisation and evaluation, and a poster by **4pm on Friday 3rd May**. Any of your experiment figures and/or tables included in your poster need to be reproducible in the Jupyter notebook that you submit.

2 Dataset and Code

You will use the dataset BreastMNIST from [1]. A Jupyter notebook with code on how to download the dataset, train and test a simple convolutional model is available from CW_AI_2324.ipynb.

The dataset BreastMNIST contains Breast Ultrasound images of size $28 \times 28 \times 1$ pixels. Each image can have two classes: {'0': 'malignant', '1': 'normal, benign'}. The dataset has been split into 546 training images, 78 validation images, and 156 testing images. More information about the original dataset can be found in [1].

The current baseline area under the precision recall curve, precision (AUC) and accuracy results [1] are reported in Table 1.

Table 1: Benchmark AUC and Accuracy results on BreasMNIST [1].

Method	AUC	Accuracy
ResNet-18 [1]	0.901	0.863
ResNet-50 [1]	0.857	0.812

3 Deliverables

You should submit the following two files. The first file is a Jupyter notebook, based on the code provided in CW_AI_2324.ipynb, containing the following parts:

1. Code with a new neural network model proposed by the student that has AUC ≥ 0.901 and accuracy ≥ 0.863 on the test set. Please start with the ResNet-18 model available from Pytorch*, use the validation set to select hyper-parameters (e.g., learning rate, mini-batch size, number of epochs, optimiser, momentum, weight decay, etc.), and propose modifications (e.g., architectural, loss function) to improve results. Please list all hyper-parameters evaluated, and how the hyper-parameter selection was performed.

2. Code to compute the following metrics for the model developed in part 1 (using the suggested Pytorch methods in brackets): a) area under the precision and recall curve (AUPR)†, b) precision‡, c) recall§, and d) F1 score¶.

3. Code to plot the following graphs for the model developed in part 1: a) the receiver operating characteristic (ROC) curve using code from Pytorch‖, b) the Precision Recall (PR) curve using Pytorch's code**, and c) the confusion matrix, also using Pytorch's code††

4. Code to cross validate your model with 5-fold cross validation (merge the train/validation/test data, divide the merged data into 5 folds, and rotate 5 times: train in 3 folds, validate in one fold, and test in one fold). For this cross validation, please report the accuracy and AUC results for the test set of each fold, and the average accuracy and AUC results for all folds.

The second file is a poster that describes an introduction to the problem, literature review, technical details of the model, and evaluation of the results. In particular, the evaluation

*https://pytorch.org/hub/pytorch_vision_resnet/

†https://pytorch.org/torcheval/main/generated/torcheval.metrics.MulticlassAUPRC.html

‡https://pytorch.org/ignite/generated/ignite.metrics.precision.Precision.html

§https://pytorch.org/ignite/generated/ignite.metrics.recall.Recall.html

¶https://pytorch.org/torcheval/stable/generated/torcheval.metrics.functional.multiclass_f1_score.html

‖https://pytorch.org/ignite/generated/ignite.contrib.metrics.RocCurve.html

**https://pytorch.org/ignite/generated/ignite.contrib.metrics.PrecisionRecallCurve.html

††https://pytorch.org/ignite/generated/ignite.metrics.confusion_matrix.ConfusionMatrix.html

must contain an analysis of the following points: a) comparison between the results of the model that you proposed and the baseline results in Table 1 and why your proposed model performs better/worse than baseline; b) differences between accuracy and AUC results from part 1 (specific train/val/test split) and part 4 (5-fold cross validation); c) differences between AUC and accuracy results; and d) differences between AUPR and F1 score. The submitted Jupyter notebook needs to contain all figures included in your poster.

4 Coursework poster

The poster should be in A0 size, in digital format (no hard print is needed). It is important to demonstrate your understanding of the subject area, your approach to the coursework and present the outcome. There are many useful guidelines on how to design an academic poster, for examples: `https://www.makesigns.com/tutorials/`. You are also provided with a sample poster as a research presentation poster example. You are free to choose either the landscape or portrait orientation.

You should consider including sufficient evidence of your work through the poster, covering the following areas in the marking rubric.

5 Marking Rubric

You will be evaluated based on the following criteria with your **poster** and submitted **Jupyter notebook**.

1. Deliverable 1 (20/100):
    - successful training and testing of new model based on ResNet-18 (+5 marks);
    - effective use of validation set for hyper-parameter selection (+5 marks);
    - improve AUC ≥ 0.901 (+5 marks);
    - improve accuracy ≥ 0.863 (+5 marks).
2. Deliverable 2 (10/100):
    - successful computation of AUPR (+2.5 marks);
    - successful computation of precision (+2.5 marks);
    - successful computation of recall (+2.5 marks);
    - successful computation of F1 score (+2.5 marks).
3. Deliverable 3 (10/100):
    - successful plot of ROC curve (+3 marks);
    - successful plot of PR curve (+3 marks);
    - successful plot of confusion matrix (+4 marks).

4. Deliverable 4 (20/100):

   - successful implementation of 5-fold cross validation (+10 marks);
   - successful reporting of the accuracy and AUC results for each fold (+5 marks);
   - successful reporting of the average accuracy and AUC results for all folds (+5 marks).

5. Deliverable 5 (40/100):

   - Introduction in the poster with appropriate background information and details about the objectives of the project (+10 marks);
   - Literature review on relevant topics and how they may relate to your work. You are expected to research and discuss other sources of information and show their origins by referencing all sources used. The reference list should be included in the poster (+10 marks);
   - Technical details of the model with clear justification for all developments. Effective graphic illustration will be appreciated. Multiple classification techniques can be designed and implemented. It can also be multiple versions of the same model with different hyperparameters or different training approaches (e.g., loss functions or optimisers) (+10 marks);
   - Evaluation of results particularly with respect to AUC and accuracy. We also expect an analysis of the following points: a) comparison between the results of the model that you proposed and the baseline results in Table 1 and why your proposed model performs better/worse than baseline; b) differences between accuracy and AUC results from part 1 (specific train/val/test split) and part 4 (5-fold cross validation); c) differences between AUC and accuracy results; and d) differences between AUPR and F1 score (+10 marks).

References

[1] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.