MTU
Ollscoil Teicneolaíochta na Mumhan
Munster Technological University

Data Analytics and Visualisation

Data9005

Assignment 2

Due: 29th April (Friday) 2022, 11.59pm (GMT)

Please submit your assignment via Canvas.

This assignment is worth 50% of your module.

This is an individual assignment and the work you submit should be your own; collaboration with others is not allowed. Include a signed declaration that is your own work at the start of your report. Plagiarism checking software will be used in this assignment. Reference your own work appropriately. Include a signed declaration with your submission.

Standard late penalties apply.

Please answer the following questions, referencing your work appropriately:

Make sure that the code you submit runs correctly, it is recommended to check this just before you submit. Your submission should be a Word/pdf document for the report and a implementation files for your code and your 6 minute recording. No zipped files or folders in your submission. It is suggested to use the last 3 digits of your ID no. as a set.seed where necessary so your results are reproducible. Give the version of R, Python, main libraries and IDE etc. that you are using.

<u>Question One – Convolutions (15 marks)</u>

Explain **in your own words** and using appropriate referencing following concepts:

(i) For a CNN, using diagrams and/or mathematical notation, giving the number of parameters created by the operations, design an example that shows and explains the input and output for:

a) a 2D input image with a 1D conv operation,

b) a 3D input with a 2D conv,

c) a 3D input with a 3D conv. *15 marks*

Question Two - Deep Learning (50 marks)

Here you are required to analyse the Kvasir-V2 dataset, 2.3GB, https://datasets.simula.no/kvasir/

Details are also given here :https://dl.acm.org/doi/pdf/10.1145/3083187.3083212. A useful github repository is here: https://github.com/AfraHussaindeen/Kvasir-Dataset.

(i) Split the dataset random into train/validation/test 70:15:15 using set seed with the last 3 digits of your student number. Choose a suitable performance metric(s).

(ii) Explain any challenges you had with configuration and environment setup and how you overcame these.

(iii) Find a baseline model, different to the one given in github. Give your performance and compare to the model in github reference. Explain what each layer of your network in doing in English, i.e. not in computer code. You may refer to your answer in question one above. *10 marks*

(iv) Change some of the model parameters, e.g. number of layers, number of filters, number of nodes in fully connected layers and show how this effects performance. Note best to have a well performing small model. Explain also in your own words why the model's performance changes/does not change.

10 marks

(v) Complex: Using the following paper as a reference, https://arxiv.org/pdf/2201.03545.pdf, can you implement 2-3 changes to see if this improves performance. Mixup and cutmix are options here. Explain the concepts in English of the changes you make.

20 marks

(vi) As this dataset is not very large how can you analyse the variability of your final results?

5 marks

(vii) Notice that some images has a greenish box with a medical instrument. If some of the classes have this more than other classes comment on how this could impact on performance in a clinical setting. How would you propose to solve this issue if performance is adversely effected? (There is no requirement to implement this.)

5 marks

Question Three – Statistical/Machine Learning (35 marks)

Use the following paper as the reference for this question: *Kvasir A Multi-Class Image Dataset for Computer Aided Pogorelov et al.,2021.pdf*. The dataset here is the Kvasir-v2-features dataset of 9.3MB. Using R (caret package) or python answer the following questions.

(i) Explain how to construct your dataset to input into your model. *10 marks*

(ii) Complex: Explain the steps how you can reproduce the '6 GF Random Forrest' results from Table 1 of the paper above. Try and reproduce the results. Can you improve these results? *20 marks*

(iii) Comment on your results and how they compare with the paper and the deep learning model.

5 marks

(iv) Record a max 6 minute video explaining your code, theory and results for Question Two and Three. Only the first 6 mins of any videos will be watched.

[Total 100 marks]