CAB420 Assessment 2 - Group Project: Details

Overview

Task description

Assessment 2 is a group assignment, to be completed in groups of 3 or 4. Each student's contribution should be indicated by a clear explanation of the contributions, and the percentage of the whole work. Please note that your mark may be moderated depending on your individual percentage contribution specified in your report.

Conditions

Weighting	30% of the total subject mark
Assessment Type	Applied Project
Length	~10-15 page report (excluding front matter, references and appendicies), and a ~5 minute pre-recorded video presentation
Group / Individual	Group

Learning outcomes measured

By completing this assignment you will investigate various methods from the wider machine learning literature to solve a machine learning problem, evaluate these methods, and critically analyse the strengths and weaknesses of these methods. By doing so, the assessment results in the following learning outcomes:

1. Apply the principles and concepts of machine learning science using a range of tools and techniques.
2. Critically evaluate different machine learning algorithms in a range of complex business, science, engineering, and health contexts.
3. Research cutting edge developments in machine learning and communicate findings to a specialised audience.

What you need to do

Group Allocations have been finalised, and can be found here: CAB420 Assignment 2 Groups - FINAL.txt (https://canvas.qut.edu.au/courses/20364/files/6014306?wrap=1) (https://canvas.qut.edu.au/courses/20364/files/6014306/download?download_frd=1)

~~Before commencing this assessment, students are to self-allocate into groups by selecting the **People** link in the unit menu (this must be done using a web browser) and then the **CAB420 Assignment 2 - Groups** tab. Click the **Join** button to join a group. You are encouraged to use your practical class and/or the Teams channel to find a group.~~

~~When you have a group, send an email to **cab420query@qut.edu.au** (mailto:cab420query@qut.edu.au), listing the names and student numbers of all group members. Ensure you have cc'd all group members. You will be advised of your group number by return email.~~

~~**NOTE**: Any students who haven't formed a group by the start of week 7 will be automatically placed into groups. If you have any concerns with your group, please contact **cab420query@qut.edu.au** (mailto:cab420query@qut.edu.au) ASAP.~~

After joining a group, you need to:

- Select a machine learning problem/topic to investigate. You may select your own project idea or choose from a small list of provided topics. The problem/topic should ideally:
  - Have one or more established datasets that can be used.
  - Be capable of being addressed in multiple, diverse ways (for example, classification can be done using various formulations of deep learning models, an SVM, a random forest, or several other techniques).

If you wish to select your own topic, please contact **cab420query@qut.edu.au** (mailto:cab420query@qut.edu.au), or discuss the idea with the teaching team during practicals, or after lectures to see whether the proposed project is suitable for this assessment task.

Click on the headings below to view topic suggestions.

▼ Topic Suggestions

For this assessment, you may also wish to look at Kaggle competitions. Kaggle is an online community of data scientists and machine learners, owned by Google. One of the more interesting aspects of Kaggle is that it provides datasets and organises many competitions for machine learning tasks. It also manages a leaderboard where the participants can publish their results. You are free to browse the competitions there to see if there are any challenges

you are interested in, or if there are any datasets that you think may be of use for your project.

Please note that any dataset that has been used in a CAB420 example or practical is not suitable for this assignment.

When formulating ideas for your project, we advise you to consider the following:

- Is there a good dataset that you can use? Collecting and/or annotating your own dataset is not advisable as this can be a very onerous task. Using an existing dataset is highly recommended.
- Is the dataset a suitable size for your problem? A dataset that is too small will limit you to using simple methods, while having a dataset that is too large may make it hard to train models for some (or all) group members. Avoid tiny datasets, and be prepared to only use a portion of a large dataset if compute is limited.
- Are your dataset splits defined? Some datasets will come with pre-defined training, validation and testing splits. For others, you will need to do this. Make sure that all group members use the same data splits.
- Are your methods diverse? The methods that you employ should be varied. Simply taking the same pre-processing code and changing an SVM to a Random Forest, or grabbing a different pre-trained CNN backbone, is not ideal. Variation can be introduced in many ways such as:
  - by varying the machine learning methods themselves;
  - by changing the pre-processing used to prepare the data;
  - for deep networks, by varying the broad type of network used (i.e. CNN vs Transformer) and/or the training paradigm chosen (multi-task learning, self-supervised learning, metric learning, etc).
- What pre-processing do you need to do to your data? Does this vary for your different methods? Remember that while deep networks often work well with raw data, for non-deep learning models this is rarely ideal (tabular data, with appropriate encoding of categoricals, is the possible exception here). Carefully consider how data should be prepared, and discuss this with the teaching team if you are uncertain.
- How should your methods be evaluated? You should be using the same metrics and evaluation protocol for all your methods. You are also encouraged to dig more deeply into the results to uncover differences in performance between models. Are there perhaps some types of samples that one model is consistently better (or worse) at than another? Can you identify any reasons as to why this might be?

A small selection of project ideas that you may also choose from follow.

Enron E-mail Classification

The Enron E-mail data set contains about 500,000 e-mails from about 150 users. The data set is available here: **http://www.cs.cmu.edu/~enron/ (http://www.cs.cmu.edu/~enron/)** . Can you classify the text of an e-mail message to decide who sent it?

Object Recognition or Clustering

The Caltech 256 dataset contains images of 256 object categories taken at varying orientations, varying lighting conditions, and with different backgrounds, and is available at **http://www.vision.caltech.edu/Image_Datasets/Caltech256/** (http://www.vision.caltech.edu/Image_Datasets/Caltech256/). You can try to create an object recognition system which can identify which object category is the best match for a given test image. Another approach may be to apply clustering to learn object categories without supervision, where you could apply clustering to different learned representations (i.e. HOG features, deep network embeddings, etc).

Please note, comparing classification and clustering within the one project is not recommended as these are not easily compared. These are intended as two separate suggestions.

Speaker Recognition

Speaker recognition is the task of recognising someone by the way that they speak. This is a classical problem in biometrics and machine learning. The Common Voice (**https://voice.mozilla.org/en/datasets** (https://voice.mozilla.org/en/datasets)) dataset contains a large number of speakers and associated meta-data, and spans multiple languages. You could simply try to recognise a speaker, or investigate the impact of training and evaluating a model on different language, or explore how meta-data could be used to improve performance.

Crowd Counting

Crowd counting is the task of counting the number of people in a scene. The output of crowd counting can be simply a single number representing the total number of people in a scene, or a density map that indicates how people are distributed in a scene (or both).
A large number of datasets have been released for crowd counting (see **https://github.com/gjy3035/Awesome-Crowd-Counting#datasets** (https://github.com/gjy3035/Awesome-Crowd-Counting#datasets)). Using one or more of these you could investigate crowd counting methods, or explore how different types of methods generalise to different conditions.

Semantic Segmentation of Aerial Data

Semantic segmentation is the task of labelling each pixel of an image with a label indicating what is at the pixel. This is commonly used within scene understanding pipelines to identify objects and regions of interest in a scene. DroneDeploy have released a segmentation dataset and benchmark suite (see **https://github.com/dronedeploy/dd-ml-segmentation-**

benchmark) (https://github.com/dronedeploy/dd-ml-segmentation-benchmark)) for semantic segmentation from drone data. Along side colour imagery, elevation data has also been captured. How could you combine elevation and RGB data to improve segmentation performance?

Other Sources of Project Ideas

There are many other places you may turn to for project ideas. A few useful links include:

- http://cs229.stanford.edu/projects2015.html (http://cs229.stanford.edu/projects2015.html)
- http://cs229.stanford.edu/projects2016.html (http://cs229.stanford.edu/projects2016.html)
- https://github.com/NirantK/awesome-project-ideas (https://github.com/NirantK/awesome-project-ideas)

Once you have selected a topic, you need to:

- Implement 3-4 different methods (one method per group member) to address your chosen problem/topic.
- Prepare a report that details these approaches and compares their performance on the chosen data. Methods should be appropriate for the problem at hand, and be supported by relevant literature.

Assessment 2 has three submission items, as follows:

- A **brief** project proposal. **This item does not attract a mark and is optional**, but will be used to give your group feedback regarding your proposed approach.
- A final project report.
- A short video presentation, submitted alongside the final report.

Detailed instructions are provided in the Assignment 2 Instructions: **CAB420 Assignment 2.pdf (https://canvas.qut.edu.au/courses/20364/files/5672649?wrap=1)** **(https://canvas.qut.edu.au/courses/20364/files/5672649/download?download_frd=1)**

Click on the headings below to view information about the two submission items.

▶ Project Proposal (Due Week 8)

▼ Report and Video Presentation (Due Week 13)

The report should be structured as follows:

- **Title page:** containing project title, team number, and names of team members.
- **Executive summary:** which should be ~1 page long and briefly outline the problem being considered, the methods selected, and overall findings.

- **Main body:** of the report, which should be 10–15 pages long and include the following:
  - **Introduction/Motivation:** clearly motivate your project, and describe the research question, and how it relates to previous works that have been done in this area.
  - **Related Work:** briefly describe a small number of relevant existing approaches, their respective strengths/weaknesses and relation to each other (i.e. does one build upon another), and the objective of your work.
  - **Data:** clearly describe the data set, any pre-processing that is performed on the data, any challenges or problems of note in the data, and how the data is split into training, validation and testing sets.
  - **Methodology:** Clearly explain the three or four algorithms (one per group member) that you used with citations to the literature where appropriate. Please note that ideally your project will extend existing approaches in some manner. You don't need to propose a novel algorithm, but you might be looking into approaches that have not previously investigated on your dataset. Note also that all considered approaches should be different. For example, rather than simply using three deep convolutional neural networks (DCNN) for a classification task, you could perhaps use (depending on the task) one DCNN, one transformer, and one non-deep learning method such as a random forest (including appropriate feature extraction).
  - **Evaluation and Discussion:** Present the results of all your approaches clearly, and compare them with existing published results (if possible), and with each other. Discuss why your methods are working better/worse than the existing approaches and each other. The evaluation should also consider the computational demands of the methods, and provide training and evaluation times alongside relevant performance metrics.
  - **Conclusions and Future Works:** Clearly explain if the experiments match the objectives, the advantages/shortcomings of the proposed approaches, and if any changes are required/ plans you have for the future investigations
- **Appendix:** that details the contribution of each group member towards the project should be included. This should list what each group member contributed, as well as an approximate overall percentage that each student contributed. This statement should be signed by each group member.

NOTE: You may optionally wish to include further appendices to include items such as code or additional (non-critical) results. Though please be aware that all critical content should be included within the main body of the report, and content within the appendices will be considered supplementary.

The video presentation should be roughly five minutes long and include the following:

- A brief overview of the problem and project motivation and the dataset used, capturing any major considerations or challenges encountered with the data
- Details of the models evaluated, noting why they were selected, and any key considerations

- Results highlights, capturing overall results and any interesting findings

The video should not cover the full content of your report (you only have ~5 minutes), but should capture a concise summary of what you have done, and what you have found.

Resources available to complete the task

Sources of Data

There are lots of datasets available online, however sadly there is no one stop shop that lists all of them. Some places with reasonable lists include:

- **https://www.kaggle.com/competitions (https://www.kaggle.com/competitions)**
- **https://ieee-dataport.org/ (https://ieee-dataport.org/)**
- **https://cloud.google.com/public-datasets (https://cloud.google.com/public-datasets)**
- **http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm (http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm)** (computer vision datasets only).

Most universities and research institutes will also have their own lists of datasets that they have released.

Frequently Asked Questions

This list of Frequently Asked Questions will be updated continuously. If you have questions about the assignment, please check here. If your question is not addressed, please email **cab420query@qut.edu.au (mailto:cab420query@qut.edu.au)** .

Click on the heading below to expand the list.

▼ **Frequently Asked Questions**

Q. What is required with the 3/4 methods? Are these ones that we propose, or from the literature?

A. You need to implement 3/4 methods (one per group member) to have a chance of acheiving full marks. These can be taken from the literature, or they can be adapted by looking at work that's been done on similar or related problems. There is no expectation that you will propose a brand new method - but at the same time there is nothing to stop you doing this (however if you did so, you'd need to explain and justify it). You will then compare these algorithms to each other. Depending on the data/problem that you are using, it may be possible to compare to a large number of other methods if they have all been evaluated using the same protocol on the same data. You are welcome to do this, however you still need to implement the 3/4 methods yourself,

which will also allow you to dig into the performance of these methods more and analysing things like failure cases, etc, which is usually not possible from just published results.

Q. We're down to 2 group members? What do we do?

A. Implement two methods, and marks will be scaled. We're expecting one method per group member.

Q. What's with the "10 existing methods" in the discussion part? Does this mean 10 papers? Or just 10 different sets of results?

A. 10 papers. Some of these papers may use very similar methods. That's ok. Not all papers need to be the current bleeding edge either. There might be really important papers that are a bit crap compared to new ones, but that's ok. A good example might be the very first paper to present and address the problem. It may have been totally superseded now, but it's potentially still worth covering.

When you write these up in your discussion, try to also focus on the strengths, weaknesses, and relationships between these methods. This means, try to avoid writing something like "A did this. B did this other thing. C did something else....." This is a bit dull. You can do better. Instead, try to capture how the methods relate and have evolved. For example "A was the first person do to this thing, and they did it like this. B then came along with this extension that made it much better. C tried this totally different approach, but it was a bit rubbish because of some reasons. D, E and F all proposed doing this very similar thing with these minor changes, and all did about as well as each other. G built on D by adding this crazy complex thing, but it worked awesome. H also built on D, but did this really simple thing that wasn't quite as good as G, but almost...."

Q. Do all the 10 papers have to be for the same data set that we are using?

A. No, but they should all be for the same problem area. If you're looking at image classification on caltech-256, you could consider any papers that cover image classification; but it would be good to have some that are also specific to the dataset (or evaluate on caltech-256 as well as other datasets).

Q. How detailed does our description of our approach(es) need to be?

A. Ideally, you have enough detail for someone else to reproduce what you've done. This means details of things such as network designs, hyper-parameters, and possibly training parameters. If you're doing something like using an pre-existing DCNN model and modifying it to suit your task, you can simply cite the model that you're using and describe the changes; you don't, for example, need to give a full description of ResNet-50 if you're using that as a starting point.

Q. How complex should our approaches be?

A. We're expecting an increased complexity compared to praticals and models implemented in Assignment 1A and 1B. This complexity could come from:

- The model itself;
- Pre-processing and/or feature extraction needed to prepare data for the chosen approach;

- The training stratergy employed, either through a multi-stage training approach (i.e. unsupervised pre-training followed by supervised fine-tuning), and/or more complex loss functions;
- A multi-stage approach that combines multiple models (i.e. visual-bag-of-words where clustering is used to learn a codebook, before training a classifier); or
- A combination of the above.

If you are unsure about your approaches, ask the teaching team.

Q. How strict is the page limit?

A. We won't stop reading, so don't see this as being a hard limit. That said, be mindful of what you're including. If you're including details that are not essential, consider moving them to an appendix.

Q. Do all group members need to be in the video presentation?

A. No. While we certainly welcome all group members being involved, if it's logistically simpler to have only some group members involved this is acceptable.

Q. What software should we use for recording and editing the video presentation?

A. This is up to you. Some suggestions:

- Powerpoint is an easy option and will let you record a slideshow as a presentation (some information on this is available **here (https://support.microsoft.com/en-au/office/record-a-presentation-2570dff5-f81c-40bc-b404-e04e95ffab33)**). In my experience, I've found the quality of recordings in powerpoint to be a bit patchy.
- Zoom is another easy option, and will allow you to record to the cloud or a local system. I recommend recording to the cloud, and then downloading the cloud recording. This is particularly easy option if you want to have multiple people presenting.
- For all my teaching videos, I use **OBS (https://obsproject.com/)** , and set this to record my screen with a camera feed overlayed in the corner. Setting OBS up is a bit more involved than powerpoint, but it's still pretty easy, and it's much more powerful in terms of integrating multiple data streams, and produces much nicer videos.
- For simple editing of all my teaching videos, and any other transcoding needs, I use **avidemux (https://avidemux.sourceforge.net/)** .

What you need to submit

Submission requirements

- **One member of your team** will submit your optional project plan (PDF or Word format) via the **CAB420 Assessment 2 - Project Proposal (Optional): Submission (https://canvas.qut.edu.au/courses/20364/assignments/182635)** page by 11:59PM on the due

date.

This is an opportunity for your group to get feedback on your planned problem and approaches. The project plan should be 1-2 pages, and briefly outline the problem you are investigating, the data you plan to use, and the approaches you plan to implement.

- **One member of your team** will submit your final project report (PDF or Word format) and video presentation (in a format playable by **VLC** (https://www.videolan.org/vlc/)) via the **CAB420 Assessment 2 - Final Report and Video Presentation: Submission** (https://canvas.qut.edu.au/courses/20364/assignments/182634) page by 11:59PM on the due date. The report and video presentation should follow the instructions outlined above.
- The rubric used to grade the assessment task is included on the submission page. Use it as a guide when working on the assessment task.

NOTE: Students who require an extension should lodge their extension application with HiQ (see **Assessment Policy - Concessions, extensions and academic integrity** (https://canvas.qut.edu.au/courses/20364/pages/assessment-policy-concessions-extensions-and-academic-integrity)). Please note that teaching staff (including the unit coordinator) cannot grant extensions.

Use of Generative Artificial Intelligence (AI) tools

The use of generative artificial intelligence (AI) tools is allowed for the purpose of understanding this assessment task and summarising information when cited appropriately. However, it is your responsibility to critically evaluate the information provided. It is not acceptable to cut and paste text generated by AI. Go to **Cite | Write** (https://www.citewrite.qut.edu.au/) and follow the specific guidance for the relevant referencing style (**APA** (https://www.citewrite.qut.edu.au/cite/qutcite.html#apa-internet-ai) , **Harvard** (https://www.citewrite.qut.edu.au/cite/qutcite.html#harvard-internet-ai) , **AGLC** (https://www.citewrite.qut.edu.au/cite/qutcite.html#legal-internet-ai) or **Vancouver** (https://www.citewrite.qut.edu.au/cite/qutcite.html#vancouver-internet-ai)).