CS3244 AY22S1: Project Proposal (v2022-08-26). Accessible at http://www.comp.nus.edu.sg/~cs3244/AY22S1/proposal-template.html

Team [redacted]

Team Identifier number

[redacted]

List of student IDs (The one starting with Axxxxxxxx; comma separated; **the team leader's ID should be listed first**)

[redacted]

Dataset choices

Choose 1 from the multiple choices. Your later answers need to be consistent with your chosen dataset.

DeepWeeds

Dataset description

Write in your own words (1-2 sentences) what this dataset is about.

a) Does your description specify features and instances?

b) Do you describe the source and reliability of the data?

The DeepWeed dataset consists of 17509 images capturing 8 weed species native in Queensland, Australia, in place with surrounding flora. The dataset is meant to be challenging as there are a lot more Negative instances (No Weed) than individual weed label images. The images also have a lot of similarity other than the weeds class themselves in terms of the surrounding flora, making the training of this model challenging yet practical in identifying weeds amidst Queensland flora.

This dataset should be fairly reliable, as it is referenced in Weed Detection research papers (https://arxiv.org/pdf/1810.05726v3.pdf); however, uses outside of Australia is not documented

Project Title

Please be concise, relevant and descriptive.

Multiclass Weed Identification

Motivation

Explain why this project is interesting and important.

a) Does your motivation clearly describe a problem?

b) Does it justify the problem's significance? What are the benefits of addressing this problem? Who benefits from solving it?

The project has some practical implications on the environment. Finding weeds amidst a large field is not only laborious, but also prone to human error. Take for example lantana, which is a pretty shrub, is actually an invasive species in Queensland. Identifying them before an infestation gets out of hand can help identify specific methods to combat the said infestation and preserve the local ecosystem. Some of the other species of weed listed in the dataset are also among the top 100 invasive species; we believe exploring this sanctioned dataset not only addresses a serious environmental issue, but also allows us to learn and explore approaches to Deep Learning.

Statement of the Problem/Task

A statement of the problem, issue, or task that you're interested in studying. In particular, try to formulate the key questions (2 to 4 questions is probably a good number) that you will answer in the project.

As this proposal document is self-contained, you should restate your project topic and domain. The information you provided in the submission system is to help with appropriate review assignments.

a) Does the proposal outline a problem statement, issue, or task that the team is interested in studying?

b) Does it formulate a few (2 to 4) questions that the team proposes to address?

Identifying weeds among flora is an extremely difficult task as it not only requires a trained eye to spot them, but it is also extremely laborious and prone to human error, given that the flora act as a natural camouflage for the weeds. Hence, the group has decided to formulate this problem statement as the following questions:

1. How can we better understand the characteristics of the provided dataset?
2. What models can we use to identify different species of weeds from the sanctioned dataset?
3. How can the characteristics of the dataset influence different models?
4. How can we evaluate the "goodness" of the various models and choose one that best generalises the dataset and achieves the objective of the project?
5. How can we evaluate the features captured by the model to understand what the model is learning from the training dataset?

After these questions have been answered, we can get some images from the public domain to do a sanity check of the effectiveness of the model. We can then elaborate on the "live testing results" as well as further elaborate on our learning outcome.

General Approach

A high-level description of the general approach you'll use to address the questions. Sketch out what evidence you are planning to gather (e.g. how you can answer the questions through experiments on data). Survey on the current progress on the problem/task.

a) Does the proposal contain a high-level draft description of the general approach proposed to address the questions?

b) Does it include preliminary plans for evaluation, data gathering? I.e., how the team plans to answer the questions through experiments on data.

1. Explore the dataset (preprocess + data visualisation)
   a. What are the characteristics of the data provided? (preliminary exploration done by Clement)
   b. Do we need more data? (Current consensus is no, to be determined again in the future)
2. Identify applicable Algorithms (Linear Models, Convolutional Neural Networks, etc.)
   a. Knn?
   b. Logistic Regression?
   c. Decision Tree?
   d. SVM?
   e. Deep Learning?
3. Transform the dataset such that it is applicable to the selected Algorithm and train a Model
   a. Similar to Assignment 1, the data might need to be converted into another form to be fed into the model, to be determined based on model used
4. Compute and Visualise the accuracy of the model
   a. What are the metrics used to evaluate the suitability of a model?
   b. Is there a way to visualise these results? (Eg Train Test Loss, Variance vs Bias etc.)
   c. If we are using Convolutional Neural Networks (CNN), is there a way to view what the model identifies as a specific class?
   d. Positives? Negatives?
5. Tune Hyperparameters
   a. What are the Hyperparameters available in the model selected (to be read from library documentation)
   b. For Deep Learning, What other types of layers can we add? Node sizes? Epoch? Batch Size?
6. Select model based on criteria
   a. Given 2 different approaches, their indication of accuracy might be different. We need to explain how and when to select one model over another
7. Build a pipeline to make predictions (input -> model -> output)
   a. Write a script to make predictions on CLI
   b. (Optional) How can we deploy the model?

Evaluation

Include how you will evaluate your project. Propose what your team thinks is a satisfactory project outcome (C grade) and an excellent project outcome (A grade). Remember that performance is secondary to analysis and understanding.

a) Does the proposal contain a high-level draft description of the general approach proposed to address the questions?

b) Does it include preliminary plans for evaluation, data gathering? I.e., how the team plans to answer the questions through experiments on data.

Excellent: Sufficient data preprocessing + visualisation with justifications, built multiple models and compared the results, compute accuracy and select appropriate model, able to justify model results and able to view what the model has "learnt"

Very Good: Sufficient data preprocessing + visualisation with justifications, built multiple models and compared the results, compute accuracy and select appropriate model, able to justify model results

Good: Sufficient data preprocessing + visualisation with justifications, built a few models and compared the results, compute accuracy and select appropriate model, able to justify model results

Satisfactory: Sufficient data preprocessing + visualisation, built a model and compute loss and accuracy, able to explain model results

Resources

A list of resources you have/need to conduct the project. This includes additional reading, software, datasets, code(github link), etc., beyond your chosen dataset. Are these resources public? How are you planning to get these resources?

a) Does the proposal give a short list of resources the team plans to use to execute the project (inclusive of readings, software, datasets, etc)?

b) Does the team describe any strategy for getting the resources?

a) Readings on Visual Transformers (ViT), Convolutional Neural Networks (CNN)
b) Readings on Tensorflow and Keras
c) Readings on sci-kit learn (sklearn)

d) Readings on Generative Adversarial Networks (GAN)
e) Experiment with Visualisation tools (matplotlib, seaborn, opencv…)

Schedule / Role Assignment

A schedule of work indicating the dates by which you plan to complete components of the project. Make sure the schedule is plausible. You may find that a table format with the remaining weeks of the course helpful to describe this goal.

a) A schedule indicating dates by which the team plans to complete the project components?
b) An assignment of the team members to the deliverables (inclusive of peer reviewing duties)?
c) Is the schedule feasible given the timeline, expertise and load of the team members?
d) For projects with possibly too large a data source (e.g., Kaggle projects), does the team propose a way to scope the data or problem accordingly to make it feasible?

Week 7: Read up on common preprocessing for images, use some visualisation libraries to view the images
Week 8: Read up on ViT, CNN and build at least one model
Week 9-12: Explore other models applicable to the project, supported by proper visualisation methods
Week 13: Compile results and do model selection
Exam Week 1: Presentation

(After the first few weeks of reading and learning, the team will come up with more concrete deadlines as well as responsibilities)