EE4483 Artificial Intelligence and Data Mining

Mini Project (Option 2)

Due: Friday (11:59pm), 11 Nov, 2022

Dogs vs. Cats
[Source: https://www.kaggle.com/c/dogs-vs-cats/data]

You are required to write an algorithm to classify whether an image contains either a dog or a cat. This is easy for humans, dogs, and cats. Your computer will find it a bit more difficult. The whole process involves the following steps:

1. **Loading and preparing the data**: The dataset in this project is based on (but not the same as) the official dataset in Kaggle Dogs vs. Cats competition [1]. You can download the training / validation / test datasets the google drive [link]. The datasets are divided as following:

```
dataset
|— train
   |— dog
   |— cat
|— val
   |— dog
   |— cat
|— test
```

   In the training and validation sets, there are two directories, namely dog and cat, which denote the corresponding image labels. For each class, there are 10000 images in training set, and 2500 images in the validation set. There are 500 images that need to be classified in the test set. To train your model, you can decide how many images for training and validation to be used, based on your computing resources. (Hint: You don't have to use all data in training. In practice, at least 2000 images for training and 500 images for validation may be needed).

2. **Data processing**: You will need to pre-process your input images. You are encouraged to do image augmentation (e.g., scaling, rotation, and flipping, which we learned in class) to improve the classification performance.
   [Hint: You can downscale your input images if your machine cannot support heavy computation.]

3. **Model selection**: You are to design and develop a classification model to classify dog and cat images. You are encouraged to use neural networks for this task (conventional classifiers can be used as comparison). For example, you can use a newly trained convolutional neural networks (CNNs) or a pretrained feature extraction backbone, such as VGG [2] and RESNET [3], then pass the features to a classifier layer.

4. **Model Training**: You need to try and test different model parameters, e.g., weights, initialization, and learning rate, to obtain good classification results. In the training stage, your algorithm should take only the image inputs from the training set for optimization, and data from val set for validation purpose.

5. **Prediction**: By setting up the correct learning algorithm, the trained model can be used to classify the dog and cat images in the test set. You need to report the classification results of the testing data in the file ("submission.csv"), including two columns, i.e., ID of images (you can follow the samples of submission.csv) and predicted label (1=dog, 0=cat).

Submit your report to answer the following questions:

a) State the amount image data that you used to form your training set and testing set. Describe the data pre-processing procedures and image augmentations (if any). (10% marks)

b) Select or build at least one machine learning model (e.g., CNN + linear layer, etc.) to construct your classifier. Clearly describe the model you use, including a figure of model architecture, the input and output dimensions, structure of the model, loss function(s), training strategy, etc.
Include your code and instructions on how to run the code. If non-deterministic method is used, ensure reproducibility of your results by averaging the results over multiple runs, cross-validation, fixing random seed, etc. (20% marks)

c) Discuss how you consider and determine the parameters (e.g., learning rate, etc.) / settings of your model as well as your reasons of doing so. (5% marks)

d) Report the classification accuracy on validation set. Apply the classifier(s) built to the test set. Submit the "submission.csv" with the results you obtained. (20% marks)

e) Analyse some correctly and incorrectly classified (if any) samples in the test set. Select 1-2 cases to discuss the strength and weakness of the model. (10% marks)

f) Discuss how different choice of models and data processing may affect the project in terms of accuracy on validation set. (10% marks)

g) Apply and improve your classification algorithm to a multi-category image classification problem for **Cifar-10** dataset (https://www.cs.toronto.edu/~kriz/cifar.html). Describe details about the dataset and classification problem that you are solving. Explain how your algorithm can tackle this problem, and what changes you make comparing to solving Dogs vs. Cats problem. Report your results for the testing set of **Cifar-10**. (15% marks)

h) Train the classifier for (g), while some of the classes in Cifar-10 training dataset contains **much fewer** labelled data. How can you improve your algorithm to tackle the data unbalancing issue? Describe and justify at least 2 approaches you use. (10% marks)

Notes:

- You can choose any programming language / platform that you like to complete the task.
- If you couldn't obtain any meaningful results or answers to the questions above, you may describe what you have done and attach the relevant working, codes, or screenshots, if available.
- Work in group of **THREE** students and submit one report which must clearly indicate (1) the group members and (2) the **respective contribution** of each group member to answering the questions.
- You should clearly cite all the references and sources of information used in your report.
- You are expected to uphold NTU Honour Code.
- Submit your report and the file "submission.csv" with your results to the assigned TA via NTULearn by the **deadline**: Friday (11:59pm), 11 Nov, 2022.

References

1. https://www.kaggle.com/c/dogs-vs-cats
2. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
3. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.