Applied Deep Learning Spring 2021 Programming Assignment

Topic: ConvNets

Application: Cassava/Yuca diseases classification

Goal: classify pictures of cassava leaves into 1 of the 4 disease categories or health.

Source: Kaggle challenge https://www.kaggle.com/c/cassava-disease/overview

(a) Healthy (b) CBB (c) CGM (d) CMD (e) CBSD

Dataset: https://www.kaggle.com/c/cassava-disease/data

Paper: This paper describes in detail the Kaggle challenge and therefore your assignment https://arxiv.org/pdf/1908.02900.pdf

Task: accurately distinguishing between four of the most common cassava diseases: Cassava Brown Streak Disease (CBSD), Cassava Mosaic Disease (CMD), Cassava Bacterial Blight (CBB) and Cassava Green Mite (CGM).

Introduction: This programming assignment correspond to a real Kaggle context: iCassava 2019 Fine-Grained Visual Categorization Challenge. To be able to download the dataset you will need to create an account in https://www.kaggle.com/ In this place (Kaggle) industry and research institutions post challenges, and teams and individual all-around the word complete, share their experiences and code. In this field the best way to learn is to see the code of other people and read about their successful and not so successful experiences.

Some info about the dataset: As the 2nd largest provider of carbohydrates in Africa, cassava is a key food security crop grown by small-holder farmers because it can withstand harsh conditions. At least 80% of small-holder farmer households in Sub-Saharan Africa grow cassava and viral diseases are major sources of poor yields. In this assignment, we introduce a dataset of 5 fine-grained cassava leaf disease categories with 9,436 labeled images collected during a regular survey in Uganda, mostly crowdsourced from farmers taking images of their gardens, and annotated by experts at the National Crops Resources Research Institute (NaCRRI) in collaboration with the AI lab in Makarere University, Kampala. The dataset consists of leaf images of the cassava plant, with 9,436 annotated images and 12,595 unlabeled images of cassava leaves. You can choose to use the unlabeled images as additional training data. The goal is to learn a model to classify a given image into these 4 disease categories or a 5th category indicating a healthy leaf, using the images in the training data (participants can choose to use the unlabeled images in their training data).

Evaluation: Measure accuracy in the training and test set. Avoid overfitting, namely, accuracy on the test set significantly lower than in the training set.

Examples: Take a look Examples: check in Week 4 the examples: cnn.ipynb, ccn2.ipynb, and cnn_with_some_tuning_1.ipynb. However, you are not going to get a good accuracy in the validation (test set) using this toy architectures.

Classification: Here, some advice to obtain a good classification accuracy in the validation set.In order to get a good classification accuracy, use a pre-train CNN (you can try ResNet 50, or any other). Check the book in Section 5.3 page 143. You can check also these example:

- https://www.tensorflow.org/tutorials/images/transfer_learning
- https://machinelearningmastery.com/how-to-use-transfer-learning-when-developing-convolutional-neural-network-models/

2) Use data augmentation to increase the number of training samples (not the test set). Textbook section 5.2.5 page: 138. You can check also this example:

- https://www.tensorflow.org/tutorials/images/data_augmentation

Here, an example of the accuracy graphs from the CNN examples in Canvas.