FS#3 Th#5: Testing and Evaluating Machine Learning Models for Plant Phenotypes data

Machine learning (ML) is a branch of computer science that creates program/model by mapping input and output data. Doing so, it enables itself to predict the output for new/test data. Like conventional program, testing/debugging/evaluating machine learning models is not straightforward. In this P2IRC project, we have a similar but even more challenging problem since we not only have terabytes of images data of diverse nature but also other associated data. In order to address this we will divide the testing into three categories, white box, grey box and black box as in the literature. As of Pie et al [2], we will maximize the number of active neurons (neuron coverage) in the neural network so that unexpected behaviours can be leveraged. We will then follow a unit test strategy for neural network models as of MLtest [5] to find whether the variables (gradients, weights) are really changing during training, or whether the loss function is calculating the actual loss or not and so on. As of the testing of deep neural networks of self-driven cars [1], we will also explore grey box testing where we will exploit input that maximizes neuron coverage and has synthetic properties and metamorphic relation. Finally, we will also explore the black box testing of the modes with different mechanisms such as model performance, metamorphic testing, dual coding and coverage guided fuzzing. Model performance is what current researches/industry people do at present (accuracy, precision measurement). Metamorphic testing involves changing values of the input and observing the impact on the output and thus investigating the model from outside. Dual coding contains implementation of multiple ML/DL algorithms. The same test suite is passed to all models and inaccurate results of best model is investigated with others when the other models generate accurate result for those cases. Coverage guided fuzzing mutates the data using random operators and observes the behavioural changes of the ML algorithm for those mutated data. Although it can be interpreted as white box testing but can be used as black box too. Odena and Goodfellow [6] of Google Brain proposed TensorFuzz, a coverage-guided fuzzing technique for neural networks, which we will also explore, adapt and compare and contrast with. We will start with the existing data and models from P2IRC phase #1 from Ian Stavness/Kevin Stanley. This will give us the opportunity of start working on the problem right way. This will also give us the opportunity of collaborating with them, possibly even co-supervision of students and research staffs. Since we also have different step by step goals, we are also confident of being successful with limited budget. This is where the collaboration and existing data will help. If resources permit, we can then explore with the new data and few models gathered/used in Phase#2. Since this project essentially will produce new state of the art testing and evaluation framework for machine learning models, there is a huge potential to use these tools in other flagships of P2IRC project that are using machine learning models. As part of this, we will also have a benchmark of validated datasets for evaluating different models for diverse varieties of datasets. We will also have mutated artificial dataset as part of the benchmark.

Given the era of Big data and data science where machine learning models are the key, this work would be of tremendous use not only to the P2IRC project, but in other projects within the Department of Computer Science, University of Saskatchewan and Canadian industry in general.

Reference:

[1]Tian, Yuchi, et al. "Deeptest: Automated testing of deep-neural-network-driven autonomous cars." *Proceedings of the 40th international conference on software engineering*. ACM, 2018.

[2] Pei, Kexin, et al. "Deepxplore: Automated whitebox testing of deep learning systems."*Proceedings of the 26th Symposium on Operating Systems Principles*. ACM, 2017.

[3] Roberts, Chase. “How to Unit Test Machine Learning Code. – Chase Roberts – Medium.”*Medium.com*, Medium, 19 Oct. 2017, medium.com/@keeper6928/how-to-unit-test-machine-learning-code-57cf6fd81765.

[4] Roberts, Chase. “Mltest: Automatically Test Neural Network Models in One Function Call.”*Medium.com*, Medium, 2 Feb. 2018, medium.com/@keeper6928/mltest-automatically-test-neural-network-models-in-one-function-call-eb6f1fa5019d.

[5] Chase, Roberts. “Thenerdstation/Mltest.”*GitHub*, 6 Jan. 2019, github.com/Thenerdstation/mltest.

[6] Odena, Augustus, and Ian Goodfellow. "Tensorfuzz: Debugging neural networks with coverage-guided fuzzing."*arXiv preprint arXiv:1807.10875*(2018).

Evaluating Generalizability of Deep Phenotyping Models: For general computer vision tasks with large and diverse imaging datasets, DNNs generalize much better than traditional image analysis algorithms. However, it is not clear how well DNNs trained for plant phenotyping tasks will generalize beyond the specific dataset they are trained with, since the datasets are much smaller (hundreds of plant images instead of millions of Google images) and have limited variation. The process of testing and refining a DNN also currently requires deep technical knowledge about neural networks and training/optimization procedures. In this activity, we will produce a state-of-the-art testing and evaluation framework for deep learning models to make it easier for non-computer scientists to train and deploy DNN-based phenotyping tools. We will focus our evaluation activities on the question of how well the deep models created in Activities 1-4 generalize across crops, seasons and locations. P2IRC is uniquely situated to address this question as we have datasets for wheat, canola and lentils for multiple growing seasons. We will engage with collaborators in the US, Europe and Australia to test generalizability across international sites. We will investigate testing in three areas: white box, grey box and black box testing. Following previous work, we will maximize the number of active neurons (neuron coverage) in the neural network so that unexpected behaviours can be leveraged [Pie2017]. We will then follow a unit test strategy for neural network models as of MLtest [Chase2019] to find whether the variables (gradients, weights) are really changing during training, or whether the loss function is calculating the actual loss or not and so on. We will also explore grey box testing where we will exploit input that maximizes neuron coverage and has synthetic properties and metamorphic relations, as has been done for autonomous vehicles [Tian2018]. Finally, we will also explore the black box testing of the modes with different mechanisms such as model performance, metamorphic testing, dual coding and coverage guided fuzzing. The testing framework will be integrated within our DPP platform in order to make DPP easier to use as a plant breeding tool.