Apricot: A Weight-Adaptation Approach to Fixing Deep Learning Models

**ABSTRACT**

A deep learning (DL) model is inherently imprecise. To address this problem, existing techniques retrain a DL model over a larger training dataset or with the help of fault injected models or using the insight of failing test cases in a DL model. In this paper, we present Apricot, a novel weight-adaptation approach to fixing DL models iteratively. Our key observation is that if the deep learning architecture of a DL model is trained over many different subsets of the original training dataset, the weights in the resultant reduced DL model (rDLM) can provide insights on the adjustment direction and magnitude of the weights in the original DL model to handle the test cases that the original DL model misclassifies. Apricot generates a set of such reduced DL models from the original DL model. In each iteration, for each failing test case experienced by the input DL model (iDLM), Apricot adjusts each weight of this iDLM toward the average weight of these rDLMs correctly classifying the test case and/or away from that of these rDLMs misclassifying the same test case, followed by training the weight-adjusted iDLM over the original training dataset to generate a new iDLM for the next iteration. The experiment using five state-of-the-art DL models shows that Apricot can increase the test accuracy of these models by 0.87%-1.55% with an average of 1.08%. The experiment also reveals the complementary nature of these rDLMs in Apricot.

This paper will be presented in the 34th IEEE/ACM International Conference on Automated Software Engineering (ASE 2019), November 11 to 15, 2019.

**Supervisor:** Dr. CHAN Wing Kwong Ricky

**Research interests:** Neural network debugging, Neural network testing

All are welcome!