Debugging Tests for Model Explanations

We investigate whether post-hoc model explanations are effective for diagnosing model errors–model debugging. In response to the challenge of explaining a model's prediction, a vast array of explanation methods have been proposed. Despite increasing use, it is unclear if they are effective. To start, we categorize bugs, based on their source, into: data, model, and test-time contamination bugs. For several explanation methods, we assess their ability to: detect spurious correlation artifacts (data contamination), diagnose mislabeled training examples (data contamination), differentiate between a (partially) re-initialized model and a trained one (model contamination), and detect out-of-distribution inputs (test-time contamination). We find that the methods tested are able to diagnose a spurious background bug, but not conclusively identify mislabeled training examples. In addition, a class of methods, that modify the back-propagation algo-rithm are invariant to the higher layer parameters of a deep network; hence, ineffective for diagnosing model contamination.

https://arxiv.org/pdf/2011.05429.pdf

$$\text{Learning: } \underbrace{\arg\min_{\theta}}_{\text{Model Contamination}} L(\overbrace{(X_{\text{train}}, Y_{\text{train}})}^{\text{Data Contamination}}; \theta);$$

$$\text{Prediction: } y_{\text{test}} = f_{\theta}(\overbrace{x_{\text{test}}}^{\text{Test-Time Contamination}}).$$