On the Value of Oversampling for Deep Learning in Software Defect Prediction

Rahul Yedida and Tim Menzies, *IEEE Fellow*.

Abstract—One truism of deep learning is that the automatic feature engineering (seen in the first layers of those networks) excuses data scientists from performing tedious manual feature engineering prior to running DL. For the specific case of deep learning for defect prediction, we show that that truism is false. Specifically, when we pre-process data with a novel oversampling technique called fuzzy sampling, as part of a larger pipeline called GHOST (Goal-oriented Hyper-parameter Optimization for Scalable Training), then we can do significantly better than the prior DL state of the art in 14/20 defect data sets. Our approach yields state-of-the-art results significantly faster deep learners. These results present a cogent case for the use of oversampling prior to applying deep learning on software defect prediction datasets.

Index Terms—defect prediction, oversampling, class imbalance, neural networks

1 Introduction

Can deep learning (DL) be applied to SE data without first adjusting those learners to the particulars of SE? Many researchers believe so. A common claim is that DL supports a kind of automated feature engineering [1, 2, 3, 4, 5] that lets data scientists avoid tedious manual feature engineering, prior to running their learners.

It is timely to assess such claims. DL is now widely applied to many SE tasks such as bug localization [6], sentiment analysis [7, 8], API mining [9, 10, 11], effort estimation for agile development [12], code similarity detection [13], code clone detection [14], etc. Despite its widespread use, few SE researchers critically examining the utility of deep learning for SE. In their literature review, Li et al. [15] explored over 40 SE papers facilitated by DL models and found that 33 of them used DL without comparison to other (non-DL) techniques. In our own literature review on DL in SE (reported below in §2), we find experimental comparisons of DL-vs-non-DL in less than 10% of papers.

This paper shows that a non-DL technique called "oversampling" (artificially generating members of a minority class *prior* to running a learner) dramatically improves deep learning. Such resampling is *not* widespread practice. For example, Wang et al. [16] recently achieved state-of-the-art performance in defect prediction. Their paper did not discuss or deploy any class imbalance techniques–a pattern that repeats across all the papers seen in our literature review. This is a limitation with current research since:

- When we tried their methods using static code attributes we found that (a) many data sets were imbalanced (target classes under 30% or even less) and (b) standard DL performed quite poorly (see all the **red** dots in Figure 1.a).
- But applying our sampling methods improved performance dramatically (see the **blue** dots of Figure 1.b).

The rest of this paper discusses deep learning and novel oversampling methods that can improve its performance. Our case study is defect prediction from *static code features* (e.g. depth of inheritance trees, class coupling and cohesion, lines of code, etc) to guide human inspection effort towards small regions of the code base that are most likely to have errors. Our argument will proceed as follows: we will first build a case for the problems with learning from a class-imbalanced dataset. Then, we will discuss approaches used in the literature for alleviating these problems. In particular, we will discuss the approach studied by Buda et al. [17]. We will then show that for defect prediction, this approach is insufficient, and extend their method to a novel *adaptive* oversampling approach. We will show that when used as part of a larger framework called GHOST, we achieve state-of-the-art results significantly faster. Our ablation study will demonstrate that the biggest leaps in performance come from our novel oversampling approach, which is the core component of GHOST.

Before starting, we digress to make three points:

- It is important not to overstate our results. This paper only shows that oversampling improves deep learning for defect prediction. That said, the results of this paper motivate a new approach to deep learning where researchers check if those algorithms can be extended and improved by support tools like our adaptive oversampling.
- Any paper recommending oversampling needs to document that it avoids the following threat to validity. While we oversample the training data, we *never* modify the test data (since that would we mean are not testing on the kinds of data that might be encountered in the field).
- To support open science, all the scripts and data used in this study are freely available online[1].

- *R. Yedida and T. Menzies are with the Department of Computer Science, North Carolina State University. Email: ryedida@ncsu.edu, timm@ieee.org*

Manuscript submitted to TSE, August 9, 2020.

[1] https://tiny.cc/ghost-dl