








SSBSE Summary of Adaptive Search-based Repair of Deep Neural Networks

Davide Li Calsi¹, Matias Duran², Thomas Laurent², Xiao-Yi Zhang³,
Paolo Arcaini⁴, Fuyuki Ishikawa⁴, and Anthony Ventresque²

¹ Technical University of Munich, Munich, Germany
`davide.li-calsi@tum.de`

² SFI Lero & School of CS & Statistics, Trinity College Dublin, Dublin, Ireland
`{mduran, tlaurent, anthony.ventresque}@tcd.ie`

³ University of Science and Technology Beijing, Beijing, China
`xiaoyi@ustb.edu.cn`

⁴ National Institute of Informatics, Tokyo, Japan
`{arcaini, f-ishikawa}@nii.ac.jp`

Abstract. The increased use of DNNs in critical systems makes the case for ensuring their correct behaviour. *DNN repair* combines fault localisation and evolutionary search to identify and fix the weight values of neurons that cause certain misbehaviours. However, as the search algorithm modifies the model, other weights can become responsible for new misbehaviours, and, therefore, the repair should focus on them. We present ADREP, an adaptive search method for DNN repair, that adapts the target weights to repair by iteratively localising the faults for the current model during the search. We explore two decision criteria for when to update target weights: ADREP_{STAG} that is based on the stagnation of the fitness values, and ADREP_{FL} that checks the difference of fault localisation results. Experiments using two DNN classifiers on an autonomous driving images dataset, show that ADREP is more effective than a baseline state-of-the-art search-based DNN repair method that does not adapt during the search. This paper summarises: “D. Li Calsi, M. Duran, T. Laurent, X. Zhang, P. Arcaini, F. Ishikawa. Adaptive Search-based Repair of Deep Neural Networks. In GECCO 2023.”

Keywords: Adaptive Search · DNN Repair · Fault Localisation

1 Summary of ADREP

As a growing number of critical systems rely on Deep Neural Networks (DNNs), it is necessary to guarantee a high level of performance of DNN predictions. In this context, DNN repair allows to fix critical failing behaviours while minimising model changes to preserve correct behaviours. A typical DNN repair approach, inspired by automated program repair for code, consists in using *Fault Localisation* (FL) to identify neural weights responsible for the model’s misbehaviours (*suspicious weights*), and *evolutionary search* to find values for the suspicious weights that lead to a better model.

Although DNN repair showed good results, it does not consider that, as the search progresses, the repaired model changes. While the original misbehaviour is fixed, new faults may be introduced; so keeping on repairing the original suspicious weights is useless. To address this, we proposed ADREP [1] (see Fig. 1), an *adaptive* search-based repair that alternates between FL and repair. When the weights being optimised are not relevant to the fault needing repair anymore, ADREP performs FL on the current best model, and optimises the new suspicious weights. Deciding when to look for new weights can be done using two *switch criteria*:

ADREP_{STAG}: the weights are not considered relevant anymore when the best fitness value of each generation does not improve anymore, based on a configurable threshold T_{STAG} and window size W_{STAG} .

ADREP_{FL}: the weights are not considered relevant anymore when the current generation’s best model FL results are different from the ones of the model currently being repaired based on a threshold T_{FL} and a window W_{FL} .

Experiments. Using the BDD100K dataset of autonomous driving images, and the VGG16 and ENetB7 image classification models, we explored three research questions: **RQ1** Does ADREP perform better than a state-of-the-art DNN repair approach? **RQ2** How do the two switch criteria ADREP_{STAG} and ADREP_{FL} compare? **RQ3** How often does ADREP switch the weights being optimised? Results showed that: (i) iteratively adapting the weights being optimised during the search (as done by ADREP) improves the repaired model’s performance; (ii) both ADREP_{STAG} and ADREP_{FL} performs similarly across their different hyperparameter values, and ADREP_{STAG} is slightly better than ADREP_{FL}; (iii) the number of times the weights being optimised are changed is problem-dependent; moreover, the number of changes is reasonably low, meaning that ADREP does not constantly reset the search process, but only when really needed.

Conclusions. We showed that adapting the search variables during the process of DNN repair had a positive impact on the process’ performance. This suggests that similar adaptive search approaches could be adopted for other problems where the optimisation process influences the nature of the problem itself.

Acknowledgments. P. Arcaini and F. Ishikawa are supported by Engineerable AI Techniques for Practical Applications of High-Quality Machine Learning-based Systems Project (Grant Number JPMJMI20B8), JST-Mirai. This work was supported, in part, by Science Foundation Ireland (grant 13/RC/2094_P2).

References

1. Li Calsi, D., Duran, M., Laurent, T., Zhang, X., Arcaini, P., Ishikawa, F.: Adaptive search-based repair of deep neural networks. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 1527–1536. GECCO ’23, Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3583131.3590477>

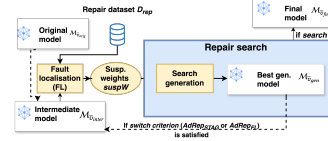


Fig. 1. ADREP