P O Box 3017
Gaborone
Botswana
17th May 2024

Botswana International University of Science & Technology,
Private Bag 16, Palapye,
Botswana

Dear sir,

RE: APPLICATION FOR PhD in INFORMATION SYSTEMS

I am writing to express my interest to apply for a 2024 PhD in Information Systems at Botswana International University of Science & Technology. My research work has been in Machine Learning and Data science. My particular research interests are in how to make predictive models more efficient by scaling features and tuning parameters for fair outputs. I believe this is in line with Optimisation as well. I find the research topic on fairness, robustness and personalised particularly interesting because of the pre-processing techniques needed in making a model produce output that is not biased.

In my Masters' dissertation, I used the Grid search for the Random Forest and Decision Trees models to identify the writing style features of an author in unknown written cross-genre and cross-topic documents. Using the Python GridSearchCV from sklearn.model_selection to look for the right parameters improved the model performance and positive rate. I know that statistical measures of fairness rely on different metrics and the Confusion Matrix is an appropriate table to use.

I am also interested in working on analysing techniques used in developing models that generate the best possible predictive output using reliable data. My previous research work enlisted the different class imbalance handling techniques in Depression prediction and detection research. I find that preparing data for training a model is also part of having a robust and fair model.

I would like to publish my findings in recognised journals that are in Scopus and Web of Science, Publishing my findings will show that my research is contributing to Machine Learning research area and I can also teach about what I have learnt.

I hope my application receives a favourable consideration and I am very interested in studying for a PhD. I hope you will take a chance on me and I will work very hard to achieve my goal in gaining a PhD.

Sincerely,

Simisani Ndaba

PhD Proposal for Information Systems

AUTOMATED MACHINE LEARNING PIPELINE FOR PREPROCESSING HYPERPARAMETER TUNING OPTIMISATION AND FINE-TUNING

by

SIMISANI NDABA

BIS (Business) (UoB), MSc (Computer Information Systems) (UoB)

A Concept Paper Submitted to the Faculty of Sciences in Partial Fulfilment of the Requirements for the Award of the Doctor of Philosophy in Information Systems of BIUST

Department of Computer Science and Information Systems
Faculty of Science, BIUST

May, 2024

1. INTRODUCTION

The goal of Machine Learning applications is to produce the most accurate prediction as possible from a given dataset. In doing this, the Machine Learning model needs to be preprocessed in a way that will result in true positive outcomes. Hyperparameter Tuning optimisation and fine-tuning are processed in the model to achieve greater performance and have a great impact on the performance of the model in any sort of application. However, an ideal combination of these techniques has not been identified but can be processed with an Automated Machine Learning pipeline.

2. SECTION 2. SCIENTIFIC BACKGROUND OF THE TARGETED SUBJECT

Hyperparameter optimization (HPO) and fine-tuning methods are crucial components in the optimization of machine learning models, essential for ensuring their adaptability and effectiveness across diverse datasets and tasks (Snoek et al., 2012). HPO involves the search for optimal hyperparameters governing a model's architecture, regularization, and optimization strategy, while fine-tuning entails adapting pre-trained models to specific tasks or domains, leveraging transfer learning to enhance performance on target datasets (Zhu et al., 2019). Despite extensive research in both areas, there is a significant gap regarding their integration and combined effects.

To the best of my knowledge, from previous research, hyperparameter optimization methods and fine-tuning methods have been compared to one another as in a study by Liu and Wang (2021). In other cases, hyperparameter optimisation techniques have been the sole focus to find a set of hyperparameter values which gives us the best model for our data in a reasonable amount of time as is the case with Bischl et al., (2023).

Comprehensive studies systematically investigating the integration of HPO and fine-tuning methods could unveil synergistic effects and optimal combinations, enhancing model performance across various domains (Snoek et al., 2012). Research efforts aimed at developing scalable and efficient optimization algorithms, capable of handling large-scale datasets and complex model architectures, are crucial for real-world applications of machine learning. Investigating the impact of hyperparameter choices on fine-tuning outcomes, particularly in domains with limited labeled data or significant domain shifts, could provide valuable insights into improving model generalization and transferability (Bender et al., 2018).

3. FUNDAMENTAL / APPLIED PROBLEM ADDRESSED BY THE PROPOSED RESEARCH

The proposed research aims to address this applied problem by developing an integrated approach to HPO, fine-tuning, and automated machine learning (AutoML) pipeline development. The primary challenge addressed by this research lies in streamlining the model development process while maximizing performance across diverse datasets and tasks. In many real-world applications, practitioners face the daunting task of manually tuning hyperparameters and fine-tuning model architectures to achieve optimal performance. This process is time-consuming, resource-intensive, and often relies on domain-specific expertise, limiting its accessibility to non-experts.

Moreover, existing AutoML frameworks often focus on either HPO or fine-tuning in isolation, overlooking the potential synergies that could arise from their integration.

By developing an integrated HPO, fine-tuning, and creating an AutoML pipeline, the proposed research seeks to automate and optimize the entire model development process. This approach will empower practitioners with accessible tools to efficiently navigate the model optimization landscape, even in scenarios with limited computational resources or expertise. Furthermore, by leveraging synergies between HPO and fine-tuning within the AutoML framework, the research aims to unlock new avenues for improving model performance and generalization across diverse domains.

The proposed research addresses the practical need for scalable, efficient, and accessible ML model optimization techniques. By automating the optimization process and integrating HPO and fine-tuning methodologies into an AutoML pipeline, the research aims to democratize advanced ML techniques and accelerate their adoption in real-world applications. Ultimately, the development of an integrated HPO, fine-tuning, and AutoML pipeline holds the potential to revolutionize the way ML models are developed, deployed, and maintained across various industries, fostering innovation and driving societal impact.

4. MAIN OBJECTIVES

1. Identify fine-tuning methods to use on Deep learning algorithms.
2. Identify hyperparameter optimisation methods to be used on fine tuned Deep Learning models.
3. Experiment with different hyperparameter optimisation algorithms that can work with fine-tuning methods on deep learning algorithms.
4. Create an Automated machine learning pipeline that will have a fine-tuned process on a given algorithm as well as its corresponding hyperparameter optimisation algorithm.

5. SPECIFIC OBJECTIVES

1. Identify multiple Deep Learning algorithms such as Neural Networks and Radical basis to conduct the experiment.
2. Conduct every possible parameter change in hyperparameter optimisation methods such as grid search, random search, and Bayesian optimization.

6. METHODOLOGY PROPOSED TO ACCOMPLISH THE OBJECTIVES

For the experiment to work systematically, I propose to use an experimental research method.

I plan on first following the following procedure.

Focus**:** Conduct experiments to test hypotheses or explore model behaviour.

Steps:

1. Hypothesis Formulation: Formulate a clear hypothesis or research question.
2. Experimental Design: Design experiments to test my hypothesis, including choosing datasets and metrics.
3. Implementation: Write code to conduct the experiments.
4. Execution: Run experiments and collect data.
5. Analysis: Analyze the results statistically to draw conclusions.
6. Reproducibility: Ensure my experiments can be reproduced by others.

7. SELECTED REFERENCES

Liu, X., & Wang, C. (2021). An empirical study on hyperparameter optimization for fine-tuning pre-trained language models. arXiv preprint arXiv:2106.09204.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 13(2), e1484.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In Advances in neural information processing systems (pp. 2951-2959).

Zhu, X., Lei, J., Liu, H., Shi, Z., & Li, S. Z. (2019). Deep learning for generic object detection: A survey. arXiv preprint arXiv:1901.07755.

Bender, G., Kindermans, P. J., Zoph, B., Vasudevan, V., & Le, Q. V. (2018). Understanding and simplifying one-shot architecture search. In International Conference on Machine Learning (pp. 550-559).