RAITE

A TENSORFLOW SOFTWARE DEVELOPMENT PROPOSAL FOR THE RESPONSIBLE AI DEVPOST CHALLENGE

PREPARED FOR

Responsible AI Devpost and Tensorflow

Tensorflow and Google and Devpost

PREPARED BY

Alain Joseph Boulay

Hackathon Team Name: “The Black Swans”

May 7, 2020

TensorFlow/Devpost
Google
https://responsible-ai.devpost.com/
support@devpost.com

Dear TensorFlow,

Re: Responsible AI Tensorflow Submission

Please find enclosed our detailed software submission proposal for the Responsible AI Hackathon with TensorFlow through Devpost.

At TensorFlow, you have worked hard to create policies and tools to promote Responsible AI. This Responsible AI Hackathon is just one example of your commitment to developing social good.

We are recent university graduates who have also seen problems developing with the widespread use of Machine Learning in such areas as Safety, Bias and poor application of scientific and machine learning methods to the creation of ML systems that may sometimes play important roles in people's lives. Like TensorFlow, we believe that maintaining scientific rigour in the evaluation of ML applications is a necessary step to obtaining effective Responsible AI.

For this reason we have developed a software testing environment specifically designed to address issues in the development of Responsible AI, such as scientific rigour.

We thank you for considering our proposal to provide a solution to the problem of a lack of rigour in machine learning through our Responsible AI software testing environment.

Yours Truly,

Alain Joseph Boulay and "The Black Swans"

EXECUTIVE SUMMARY

RAITE is an evaluation environment for Machine Learning apps designed to obtain scientifically unbiased results through use of special tools and workflows for

Responsible AI assessment. This report reviews this software development application called RAITE -a 'Responsible AI Testing Environment'. This environment provides special tools and workflows designed specifically to help software designers and architects, programmers and end users to evaluate Machine Learning applications in terms of Responsible AI. Specifically, this current implementation focuses on ways to promote the application of Scientific rigour through testing Machine Learning software applications against the goals of Responsible AI held at TensorFlow and Google.

Our product, RAITE, specifically addresses Google AI Principle number six: "Uphold high standards for scientific excellence". RAITE also addresses Google AI Principle number three: "Be built and tested for Safety". But really, RAITE can be seen to address all the points held in the Google AI Principles.

In this report, only some RAITE workflows and tools specific to instilling scientific rigour in Responsible AI implementations will be discussed. The long term vision for RAITE is that it becomes the 'go to' application for all Responsible AI needs, from improved Transparency to the implementation of Privacy, from the identification and reduction of Bias to improved Safety and Fault identification. However, in this brief report, we only have enough space to address the issues of Scientific rigour in ML applications. We believe that scientific rigour is an important place to start because many will agree that it is an essential first step in any evaluation of Responsible AI and also because a strong foundation based on scientific evidence will provide the necessary conditions to correctly address other issues of Responsible AI in any software.

The problem that we address with RAITE is the problem of poor scientific and machine learning methodology applied in developing AI applications and its effects on Responsible AI. Good scientific methodologies are defined as those that apply the scientific method in experimental procedures that demonstrate an unbiased conclusion based on empirical results that are verifiable, repeatable and that demonstrate validity. Good Responsible AI must then include choices made on the basis of good scientific evidence. RAITE is an AI testing environment that also facilitates Transparency by providing tools to identify internal validity of Machine Learning systems through scientific experimentation and empirical evidence. RIATE will improve Fairness in Responsible AI by improving measurement and modelling of AI applications through TensorFlow code.

There is one well known example of a testing environment available to the public that has a mandate of developing Responsible AI: Open AI. However, Open AI is very focussed on the implementation of Reinforcement Learning applications and does not emphasize a scientific approach. RAITE, on the other hand, provides a solution to the problem by emphasizing scientific rigour as a basis for Responsible AI for any kind of Machine Learning application including Reinforcement Learning. RAITE provides tools

and workflows that help the user to correctly test their software for Responsible AI using TensorFlow.

Correct testing of ML applications for Responsible AI will facilitate rigor. Some experts believe there is a need for rigor in Machine Learning: Ali Rahimi at the 2017 NIPS conference lamented that the field of Machine Learning was more like 'alchemy' due to the unscientific and non empirical approaches that are being undertaken by many in the field. Yann LeCun, at the same conference debated that it was not right to accuse ML practitioners of being alchemists just because our current theoretical tools have not caught up with our practice, (Sejnowski, 2020). RAITE will serve to help fill this need by providing tools and workflows to support rigorous ML experimentation specifically for Responsible AI.

The impact of RAITE on the AI market may be extensive as many AI developers will choose to obtain experimental validation for their applications from an independent, third party source such as RAITE. RAITE will help them to meet the industry standards for ML replication and open sourced code. RAITE may offer a 'public recognition' that certain tests or assessments have been passed. RAITE will offer a range of solutions to meet the diverse needs of ML users and programmers. Moreover, use of the RAITE system will contribute to a positive societal impact by helping to untangle interactions to promote fairness, transparency, privacy and safety in AI systems. RAITE will promote 'Fairness by Measurement and Modelling' in ways that meet standards and demonstrate scientific rigour through TensorFlow.

TensorFlow and its associated ML programming languages such as Keras will be featured as the main programming environment for RAITE. RAITE is currently hosted in Google Cloud, accessible as a serverless Flask web interface to everyone on the internet.

1. Project Overview

The RAITE project will develop a method of evaluation of machine learning development in terms of scientific experimental approaches specific to machine learning and computer simulations using TensorFlow. In this current submission, methods that evaluate machine learning development in terms of scientific experimentation using TensorFlow will be presented from a foundation called the "NeurIPS 2019 Reproducibility Challenge", which is a set of recommendations and a formal checklist regarding scientific and statistical content required in ML paper submissions for publication. The goals of this project are to improve Responsible AI, Fairness, Bias reduction , Transparency and Interpretability through improvement of the experimental design applied in ML development. The checklist and associated examples of TensorFlow code produced here is the first example of a tool from the RAITE TensorFlow tool kit. From this foundation of Reproducibility, the project meets its goals of improvement of Responsible AI by describing main steps in the process of

developing Machine Learning applications using scientific principles for experimentation. In the long run, the scope of RAITE will include many methods for experimentation with Machine Learning models from many different fields as statistical physics, connectionism, optimization and other computational fields using TensorFlow and Keras resources. We do not want to 'Reinvent the Wheel' with RAITE so tools that have already been tabled by TensorFlow, such as the 'What If Tool' will be referred to but not duplicated. However, many such tools will be helpful in bringing RAITE to become the best possible aid to experimental design for ML using TensorFlow. Being the best possible aid to experimental design to improve Responsible AI is the intended operation of RAITE. (Please see Appendix for an example of source document from NeurIPS 2019 Reproducibility Challenge)

2. Obstacles

There are few risks involved with this project: improvement of experimental design can only improve Safety and Responsible AI. A clear obstacle is the large size of the domain of 'experimental methods for ML'; many different fields contribute to this area and there is no 'one size fits all' approach because the experimental design will be unique to the goals and applications of the developer's tasks and objectives. For example, a common goal of ML is to make predictions from data, but another application may be to read text through Natural Language Processing. These two applications of Machine Learning are so different that a 'one size fits all' method cannot be obtained. However, it must also be agreed that there are universal methods in scientific experimentation and that employing these methods will improve Responsible AI to some degree. RAITE has the purpose of using TensorFlow approaches to facilitate such a 'universal' approach, avoiding obstacles inherent in the size and diversity of the ML field. We will manage these possible risks within this project through clear education of the ML developer regarding the need to undertake methods specific to each different case while keeping Responsible AI in mind.

A small obstacle may be the fact that the Reproducibility Challenge Checklist is meant for Academic Research, where many users of TensorFlow may not need an ML Experiment Checklist that is so demanding. However, this proposal describes RAITE as a prototype. RAITE will be improved with user evaluations and further input from developers.

3. Technical Obstacles

There are no overbearing technical obstacles to this project because we only apply TensorFlow to answer questions of experimental validity for ML applications and Responsible AI. Thus, no obstacles such as integration between different systems will develop.

4. Industry and Market Risks

There are no industry and market risks: Since we describe established methods of experimentation using TensorFlow, RAITE will not become outmoded during development or after launch. However, it can be recognized that only TensorFlow is being used in RAITE, so RAITE apps are dependent on TensorFlow users.

5. Budgetary Risks

There are no budgetary risks: RAITE uses established experimental approaches from solid foundations in the ML discipline, such as the Reproducibility Challenge and accepted experimental approaches. RAITE uses TensorFlow methods which are already established and there is a clear drive to avoid ‘reinventing the wheel’. Currently, the only Deadline is the end of the Hackathon so there is no worry of missing a milestone. (However, we sincerely hope that TensorFlow will ask us to extend our project to its full capacity by hiring us on to develop RAITE in the future!!).

6. Hardware

No special hardware is needed.

7. Software

TensorFlow is the only software technology required.

8. Appendix

The Machine Learning Reproducibility Checklist (v2.0, Apr.7 2020)[1]

For all models and algorithms presented, check if you include:

- A clear description of the mathematical setting, algorithm, and/or model. **(Check)**
- A clear explanation of any assumptions.
- An analysis of the complexity (time, space, sample size) of any algorithm.
- For any theoretical claim, check if you include:
- A clear statement of the claim.
- A complete proof of the claim.

For all datasets used, check if you include:

- The relevant statistics, such as number of examples. **(Check)**
- The details of train / validation / test splits. **(Check)**
- An explanation of any data that were excluded, and all pre-processing step. **(Check)**
- A link to a downloadable version of the dataset or simulation environment. (**Check)**
- For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.**(Check)**

For all shared code related to this work, check if you include:

- Specification of dependencies.
- Training code.
- Evaluation code.
- Pre-trained model(s).
- README file includes table of results accompanied by precise command to run to produce those results.

For all reported experimental results, check if you include:

- The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.
- The exact number of training and evaluation runs.
- A clear definition of the specific measure or statistics used to report results.

[1] https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf Last retrieved may 10 2020

- A description of results with central tendency (e.g. mean) & variation (e.g. error bars).
- The average runtime for each result, or estimated energy cost.
- A description of the computing infrastructure used.