Open Source Analysis

Jeff Putlock

June 21, 2019

1 Project Analysis

1. Submitty

Evaluation Factor	Level (0-2)	Evaluation Data
Licensing	2	Submitty uses the BSD 3-clause license.
Language	1	Submitty uses mostly PHP and HTML with some Python / C++, which I would prefer to do.
Level of Activity	2	For the past year, every week has had a commit.
Number of Contributors	2	There have been 93 contributors over the course of Submitty's development.
Product Size	2	This project has many lines of code and the total repository amounts to 38 megabytes of data.
Issue Tracker	2	There are 278 open issues and 1427 closed. There have been multiple issues opened and closed within the past few days.
New Contributor	2	There are multiple links around that point you to "suggestions for new developers."
Community Norms	1	There are no blatant examples of poor behavior but also no stated code of conduct.
User Base	2	Submitty has the RPI Computer Science student body as its userbase and is growing to other sections of the school and to other universities. The slack channel has open communication between people struggling to use it or set it up, and there are detailed instruction lists.
Total Score	16 / 18	

2. YACS - Yet Another Course Scheduler

Evaluation Factor	Level (0-2)	Evaluation Data
Licensing	2	YACS uses the AGPL license.
Language	0	It is written mostly in Ruby and also using Typescript, neither of which I am particularly interested or skilled in.
Level of Activity	1	It was active for a large portion of the year except the past few weeks.
Number of Contributors	2	There have been 39 contributors over the course of development.
Product Size	2	The project is fairly large, amounting to approximately 20 megabytes of data.
Issue Tracker	1	There are some issues being added and removed but fairly slowly.
New Contributor	2	The project clearly outlines where new contributors should go to get started and specifically mark off which issues are beginner-friendly.
Community Norms	2	The project clearly displays a Code of Conduct and has no obvious poor behavior.
User Base	2	The entirety of the RPI community uses YACS as their course scheduler, and NYU will soon be using it as well.
Total Score	14 / 18	

3. TensorFlow

Evaluation Factor	Level (0-2)	Evaluation Data
Licensing	2	TensorFlow is licensed under the Apache 2.0 license.
Language	2	It is written mainly in both C++ and Python.
Level of Activity	2	It is constantly being updated with well over one hundred commits each week.
Number of Contributors	2	There are 2043 contributors working on the project.
Product Size	2	The project makes up a whopping 350 MB.
Issue Tracker	2	There are plenty of issues being reported constantly, and many being closed each day.
New Contributor	2	Inside of the contributing guidelines page, there are links that send you to beginner friendly issues to resolve.
Community Norms	2	There is a clear Code of Conduct linked multiple times, with community standards also posted afterwards.
User Base	2	Many companies and users use TensorFlow across the world to create neural networks.
Total Score	18 / 18	

2 In-Depth Analysis

2.1 License

TensorFlow received the full 2 points on licensing because of the fact that it uses Apache License 2.0. This means that anybody who wants to contribute to TensorFlow can, as well as permitting people to create their own private projects using it. This amount of flexibility, although it does not follow Richard Stallman's idea that all software should be free, it provides an amazing amount of potential for other people to create and innovate with the massive library provided.[1]

2.2 Language

I gave it a full 2 points on language due to the fact that I am most comfortable with C++ and Python. Although most upcoming projects are in "web languages" such as Javacsript (or some variant of it) I still am not well versed in those languages. Working in C++ or Python is the most attractive because it would allow me to directly build off of what I've learned in my classes here at RPI, especially working with build systems that we just learned about such as CMake and Make.

2.3 Activity

TensorFlow got a 2 on activity due to the amazing amount of work put into it over the past year. Every week, there are 300 or more commits, with it being a generally very stable amount. Since TensorFlow was originally a Google project, the original contributor base was fairly large and well funded. Since then, it has been open-sourced and has turned into a "thriving ecosystem of products, on a wide range of platforms."

2.4 Current Contributors

As just stated, it was originally a Google project, so it already started with a fairly decent contributor base. Since becoming open source, it has grown to over 2000 contributors working on the project, not even including the other projects that are based on the original TensorFlow core product. Because of this immense development community, I felt it would be wrong to give it anything other than a 2 on contributors.[2]

2.5 New Contributors

From an outside perspective, the community supports new contributors to join the project. Although contributors must sign a Contributor License Agreement, anybody is free to contribute as much as they would like.[3] They have multiple pages on their website, going over

---

[1] https://www.gnu.org/philosophy/free-sw.html

[2] Found at https://www.tensorflow.org/community/contribute

[3] Found at https://cla.developers.google.com/clas

how to contribute, the code style that they use, and how to find issues to work on. In addition, the TensorFlow contribution page gives tips for new programmers and contributors on how to improve their code coverage and how to integrate themselves within the workflow. They provide some questions for developers to ask themselves before pushing, such as "Is the code consistent with the TensorFlow API," "Is this code efficient," and "Is the code backwards compatible." The site also cites multiple ways to contribute other than programming such as improving documentation, answering StackOverflow questions, and reporting issues.[4]

2.6 Community

The Code of Conduct is clearly displayed on the Github repository and on the project site, outlining how contributors should act and work, as well as how to resolve conflicts within the community.[5]

2.7 User Base

As stated previously, TensorFlow has birthed an ecosystem of products, such as Tensor-Frames, OpenSeq2Seq, and TensorForce. Each of these projects has their own subset of contributors, from small groups of 10-15 to whole work groups as parts of companies such as NVIDIA. The user base of Tensorflow, however, is significantly larger than just a few companies or small groups. Some of the largest and cutting-edge technology companies to date are employing TensorFlow in their projects, such as Google, AirBnb, Ebay, Intel, DropBox, DeepMind, AirBus, CEVA, Snapchat, SAP, Uber, Twitter, and IBM. As a very rough and basic estimate of its userbase, the main TensorFlow repository has 75578 forks, which just illustrates the massive size of the project.[6]

2.8 Goals

TensorFlow aims to be an end-to-end platform providing a comprehensive set of tools for developers to formulate better machine learning environments. Through different libraries, community resources, and that toolset, TensorFlow creates multiple levels of abstraction so that programmers can have the ability to choose how to set up their neural networks easier and more efficiently. Since the world of science is getting progressively more computational, machine learning has almost become a necessity for any progress to be made in some subjects. TensorFlow aims to help unclog the stops that are preventing progress in these fields with less programming experience by bridging the gap between beginners and veteran machine-learning developers so that research can continue. It also aims to create reliable production of machine learning, no matter where it is being run (i.e. in browser, in the cloud, or on-device).[7]

---

[4]Found at https://www.tensorflow.org/community/contribute
[5]https://github.com/tensorflow/tensorflow
[6]https://www.quora.com/Who-is-using-TensorFlow
[7]https://www.tensorflow.org/

2.9   Technology

It takes advantage of the advancements made in discrete graphical processing units (GPUs) to greatly accelerate the process at which the data is calculated, since a majority of the operations entail matrix operations.[8] According to NVIDIA, their aptly-named Tensor cores are up to 3 times performance speedups than a general-case Pascal CPU at calculating these operations.[9] Because of this impressive improvement in speedup, TensorFlow intentionally targets GPUs as their main device for calculation when available. As of TensorFlow 2.0, the project uses Keras, an open-source API for deep learning, as its high-level API of choice.[10] By incorporating well-known high-level API's, it is that much easier to get up and running with your first TensorFlow project. By pairing the quick and low-level C++ core code with a Python front-end, the TensorFlow developers made it even easier for newer programmers to jump in (this is based off of my personal opinion that Python is an easier first language than C++).

2.10   Significance

TensorFlow could be considered a project that marks a milestone in the progress of machine learning and artificial intelligence. Thanks to TensorFlow, creating both simple and complex models have become significantly less complicated to set up. TensorFlow allows computational or pattern-based research to be accelerate, due to the learning models that can be produced. A large number of everyday items leverage TensorFlow's power to improve everyday life, such as the voice assistant in your phone, or most other "recognition" applications. Google Translate, for example, leverages TensorFlow's learning techniques and libraries to connect people who might even have a language barrier, or allow a person to read a sign that is written completely in a foreign language to them. One of the greatest ironies of the modern era is that in order to prevent bots from accessing certain data, we are training bots to be able to do that.[11] Google's reCAPTCHA that you have definitely seen across the internet, helping protect from botters, is actually a TensorFlow project that has been running on the entire world's data. Google wants data of "what looks like what" so that their artificial intelligence can figure that out for itself, but it needs **a lot** of training data; every time you "identify a tree" in a reCAPTCHA, you are providing that data.

2.11   Appeal

The appeal of working on TensorFlow is due to the impact I would have on people around the world. Each and every person that commits to TensorFlow is affecting the world massively, because there are so many other developers and projects that depend on TensorFlow. For example, PayPal is now using TensorFlow to stay ahead of fraud and learn from previous cases so that it can more easily detect when a fraudulent attempt is being made.[12] Without

---

[8]https://www.tensorflow.org/guide/using_gpu

[9]https://www.nvidia.com/en-us/data-center/tensorcore/

[10]https://keras.io/

[11]https://www.techradar.com/news/captcha-if-you-can-how-youve-been-training-ai-for-years-without-realising-it

[12]https://medium.com/paypal-engineering

technology like that, so many people could have their banks drained or their lives ruined; and that technology roots from TensorFlow. There's always the notion with the human race that "if we don't do it now, we will do it later," but with improvements like these we will be much better off doing it now, especially with the "controversial" environmental problems we are having. By having more efficient factories and lives, we are creating the ability to be more green. For most of my life, I have been focused on nothing but efficiency and production, but the past few years have taught me that we will need to reach past production and try to be both efficient and environmentally friendly at the same time.

2.12   Community Interaction

Interactions amongst the community seem to be pleasant at the very least. After scrolling through some of the forums, I have only seen pleasant or kind responses out of any member of the community responding.[13] People are quick to respond with useful code snippets and are inquisitive when hearing the problems others are facing, which reminds me of the article on how to answer questions in a smart way that we read during the first lab.[14] Since the project is so big, the amount of information available to the community and incoming developers is staggering; which is probably going to be the best and worst part of interacting with the community. At the very beginning, at least, it is going to be mostly RTFM or searching the internet for the answer that is most likely already waiting for you. The best part of the community will come out when you have questions and I think most developers would be willing to point you in the right direction.

---

[13]https://www.tensorflow.org/community/forums
[14]http://www.catb.org/esr/faqs/smart-questions.html#idm667