April,2020

TensorFlow: An Open Source Resource for Machine Learning

Lexus Hartung, Bryan Ko, Amanda Guinyard

College of Charleston, Department of Computer Science

CSCI 462

Abstract

Our goal was to contribute to a software called TensorFlow and help fix the bugs and other issues that other users have left on their bug tracker forum. TensorFlow is an open source machine learning framework that is used to implement machine learning and deep learning applications faster and easier and was originally developed by Google. Instead of dealing with all the details of implementing algorithms, it focuses on the overall logic of the application and takes care of details behind the scenes. It can train and run deep neural networks for handwritten digit classification, image recognition, word embeddings, recurrent neural networks, sequence-to-sequence models for machine translation, natural language processing, and PDE (partial differential equation) based simulations.

Background

Google first built something called DistBelief during 2011 which is a machine learning system based on deep learning neural networks. Some computer scientist were given the task to make this software faster and more robust and over time this became TensorFlow. The first version (1.0.0) of TensorFlow was released on February 11, 2017. Then another version was released in October 2019. This software runs on CPUs, GPUs, 64-bit Linux, macOs, Windows, and mobile platforms such as Android and iOS.

Key Words

Tensor - an n-dimensional array meant to represent a partially defined computation

Bias - An additional value used to account for functions with a non-zero y intercept

Weights - The value given to connections between tensors. The higher the weight, the more influence it has on the ouctome

Checkpoints- A save point for a tensors parameters in the middle of a training session

Process

Due to the coronavirus, implementing Scrum early on in our process was valuable because it enhanced our experience when we transitioned to working remotely. This created an iterative process that allowed for better workflow and communication and ultimately a better result. Scrum and other agile methodologies are heavily used in many companies. Along with that, GitHub was another big tool that we used to both communicate with other developers and work on our bugs. Overall, many tools were used to help us in our process to contributing to the TensorFlow repository.

Big Bugs & Issues from TensorFlow

Dataset for YouTube 8M

```python
""" YouTube-8M Segments dataset from https://research.google.com/youtube8m/"""

from __future__ import absolute_import
from __future__ import division
from __future__ import print_function

import tensorflow_datasets.public_api as tfds

# TODO(youtube_8m):
_DESCRIPTION = "YouTube-8M is a large-scale labeled video dataset that consists of millions of YouTube video IDs," \
               " with high-quality machine-generated annotations from a diverse vocabulary of 3,800+ visual entities." \
               " Can be used for large-scale video understanding, representation learning, noisy data modeling, transfer learning," \
               "and domain adaptation approaches for video, and multi-task learning. "



class Youtube8m(tfds.core.GeneratorBasedBuilder):
  """Config for YouTube 8m dataset"""
```

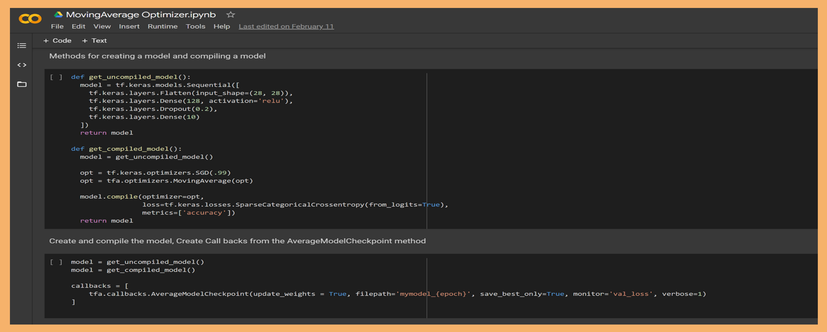
An explanation of the YouTube 8M Dataset

TensorFlow Datasets is a repository that provides standardized datasets ranging from video, text, audio and much more. This allows individuals to use these standardized datasets and use them to train their machine learning models. The YouTube 8M is a dataset that we worked on that labels millions of videos with machine generated annotations. These are mainly used for large-scale video understanding.

Tutorial for Addons

An explanation of the Moving Average Optimizer

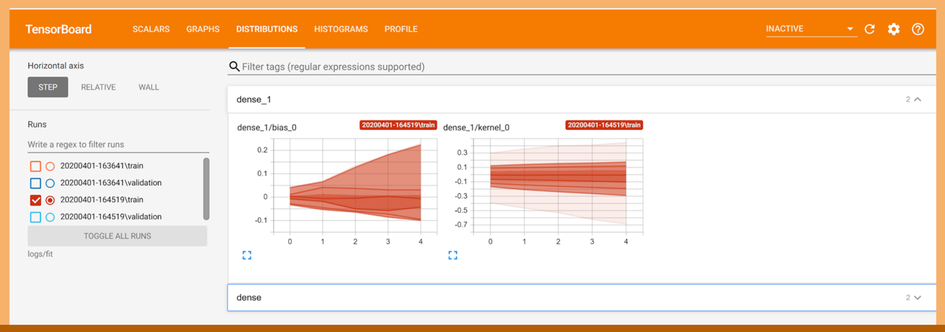
The moving average optimizer was made by the github user Squarick and required a tutorial to show new users how it worked. In essence, a moving average is a dynamic operation that takes the average of subsets of a full dataset. For the moving average optimizer, the moving average is taken for all the parameters of a tensor, and is stored in an average model checkpoint.



Tutorial for Tensorboard

An explanation of the distribution dashboard:

This dashboard displays the change in bias and weights on each tensor over the course of a training session. Each drop down menu represents a tensor, with the names reflecting what type they are. The graphs display the change in bias and weights. Each graph has 6 lines, these are to delineate the percentiles of the values being changed. The closer the six lines get to being evenly spaced, the closer the tensor is to being optimally trained



Smaller Issues

Links fixed:

10 broken links fixed in total, relative pathing used as per request to try and future proof these documents.

Missing Variables:

TensorFlow recently started updating their code from version 1 to version 2 and Bayesian_Gaussian_Mixture_Model.ipynb, file was getting an error

tfds.Split.All does not work:

In their documentation, they had a function that the user wanted to use but was not working. So after some research about why it was happening, the code was updated to clearify the problem.

Conditional Covarience

This issue was put on here to update their formula.

Results

We managed to fix seven bugs, with several more under review. TensorFlow's complex subject matter made working in its code very difficult, but we managed to make the best of this experience and learned alot from its robust community. Participating in the Tensorflow issue pages and chat boards has bolstered our communication abilities with people both above and below our level of technical expertise. Working from home, due to the recent pandemic, has also given us valuable exposure to Jira and the scrum method of software development.

Aknowledgements

Special Thanks:

- College of Charleston School of Sciences and Mathematics, and the Department of Computer Science
- Conchylicultor(TensorFlow Member)
- Squadrick (Tensorflow Member)
- Google