First steps of the project

Title: Image classification using tensorflow and TPUs

Members: Ragnar Kadai

Repsitory: https://github.com/RagnarKadai/A5

Business understanding

Background

Image classification or the function to categories images into groups is used in a variety of ways in today's technological landscape. For example it is used in the field of medicine and healthcare, security systems for object classification and the most prevalent example is in the use of autonomous vehicles to help it distinguish what is happening in the real world.

Business Goals

The goal is to use Tensorflow and Tensor Processing Units (TPUs) to help in the image classification. The project wants to use machine learning technologies to improve image classification accuracy.

Business success criteria

The success criteria for this project are to boost the sorting accuracy by at least 7% when using TensorFlow and TPUs.

Inventory of resources

A weekly amount of TPU usage is available on the website Kaggle.com, Lectures on the course and internet will also be available. There are multiple image classification datasets also available on Kaggle with an option to find more.

Requirements, assumptions, and constraints

Project code completion by 11.12.2023

Acceptable outcome: An image classification model with 88% accuracy.

Possible constraints: Time limitations due to academic commitments. Limited daily use for Kaggle TPUs and limitations on data found.

Terminology

Image classification: Categorizing images into classes/categories that are defined beforehand.

TensorFlow: An open-source machine learning framework developed by Google.

TPU: Tensor Processing Unit, a hardware accelerator designed by Google for machine learning tasks.

Accuracy: Measurement on how free of errors the model's predictions are.

Model: The algorithm used for prediction.

Noise: Difference between the model's prediction and the actual outcome

Data: Values collected through record keeping, polling, observing, or measuring. Used for analysis and decision making.

Costs and benefits

Costs: There are no evident costs associated with this project.

Benefits: Improved accuracy in image classification with potential for real application.

Data-mining goals

1. Preprocess the data to be suitable for TensorFlow and TPUs.
2. Identify the most efficient classification model.
3. Develop an algorithm for image classification using TensorFlow and TPUs
4. Create a project presentation poster by the deadline.

Data-mining success criteria

- Model accuracy: Achieve the minimum of 88% accuracy for image classification.
- Convert categorical features into a numerical format successfully.
- Find the best classifier for this task.
- Complete the project by the deadline.

In summary, the project is aiming to utilize TensorFlow and TPUs to help in image classification. When achieving this data-mining goal and success criteria, there will be a lot of progress made in the realm of image classification.

Data understanding

Outline Data Requirements

For the success of the image classification project, the primary requirement for data includes a well labeled dataset that contains images representative of the classes that are aimed at classifying. The dataset should cover a lot of different images, so the model understands how to classify a broad spectrum of images.

Verify Data Availability

The main datasets are on Kaggle, which is a popular platform for machine learning datasets.

Define Selection Criteria

Datasets with clear labels, high quality images and a good number of samples to train are the datasets that are good. Additionally different backgrounds, lighting and angles are good for the models application in the real world.

Describing Data

The current dataset comprises of a collection of .tfrec files put into different folders sorted by size jpeg size and then split into train, test and val folders. Each image is labeled and the data is well organized, since it is for Kaggle.

Verifying Data Quality

To verify the quality of the dataset it is essential to do a systemic check on it in case there are any missing or corrupted files. This is crucial to maintain integrity of the dataset, so the training of the model goes correctly. An absolute necessity is the checking of label accuracy. Finally understanding the variability of the images is important to understand where the model might be affected.

Project plan (initial)

- Gathering data from Kaggle or different sources should take about 2 hours.
- Preprocessing the data, making it usable for usage on TPUs on Kaggle should take about 2 hours.
- Splitting the data into testing and training should take an hour.
- Experimenting with various classification models should take about 8 hours or more.
- Training the model and developing the machine learning algorithm will take the most time and probably around 14 to 20 hours.
- Going over errors made and fixing them should take from half an hour to about 5 hours.
- Making the poster should take about 2 to 4 hours. After that the repo with the code and poster will be submitted