Dog Breed Image Recognition (Using Cloud Services)

DATS6450_10 - Cloud Computing Group Project

Luis Ahumada, Jesse Borg, Sarah Gates

Fall 2020

The George Washington University

INTRODUCTION

Cloud services and applications are becoming increasingly popular, with many companies electing to transition to these services. Cloud Computing is becoming an important aspect for businesses and organizations for some the following reasons:

- Cost Effective
- Competitiveness in a Digital Society
- Scalability
- Data Back-Ups and Recovery

This has led companies like Amazon to dedicate a section of their company specifically to providing cloud services. In 2006, Amazon Web Services (AWS) began offering IT infrastructure services to businesses in the form of web services. AWS has become one of the biggest cloud providers thanks to its wide range of services offered, high reliability, and adaptability. In 2019, AWS alone had a total revenue of $280.5 billion, which shows just how popular its services have grown to be.

This project will leverage the wide variety of services provided by AWS and apply them to a previous project to convert it from computer based to cloud based.

The project was a machine learning project which aimed to classify different dog breeds from a number of images. The set of images was split into a training and testing set so that the program can learn which images belong to certain breeds. It will then try to classify the testing set and the quality of the program is indicated by the success rate of classification. For this project, instead of storing the images on a laptop and using downloaded programs to run the algorithms, the images will be stored online and Python will be run on an IDE (PyCharm) using AWS.

FEATURES

For this project, a number of AWS services will be used to successfully classify dog breeds based on different labelled images.

Amazon Boto3
Boto3 will be used to manage the AWS EC2 and S3 from PyCharm using Python.

Amazon S3
A bucket will be created using Amazon S3 which every member of the group will have access to. All of the unclassified images will be stored in a single folder, with the goal being for our program to use these images and classify the breeds based on the model.

Amazon EC2
An EC2 instance will be initialized which will be used as a gateway to access PyCharm through an Ubuntu virtual machine. This eliminates the need to have a downloaded program which runs python as PyCharm is perfectly capable of running python. This will run the program which will classify the dog breeds and store them in different folders in an S3 bucket.

DATA

The data from this project was obtained from a predefined dataset which is included in python and is called ‘ Stanford Dogs Dataset’. This dataset was useful as it contained enough images to be able to train the program adequately, and its wide variety of dog breeds means that it will challenge the algorithm.

The description of the dataset is as follows:

- Images of 120 breeds of dogs from around the world
    - Number of categories: 120
    - Number of images: 20,580
    - Annotations: Class labels, Bounding boxes

Out of the 20,580 images, 12,000 will be used for training (58%) and 8,580 will be used for testing (42%).

EXPECTED OUTCOMES

Based on the information given and from the research done, we expect the following outcomes from our project:

- Fine-grained image categorization of dog breeds based on the testing set from the “Stanford Dogs” dataset.
- Storage of all the images in an S3 bucket.
- A running EC2 instance which will be used to access PyCharm on an Ubuntu virtual machine which will be used to run python and execute the program.

The success of our project will be determined by the success rate of image classification by the program and whether or not each member can easily recreate the program.

PROJECT ARCHITECTURE

The project will have three main parts; data-preprocessing, model implementation and validation.

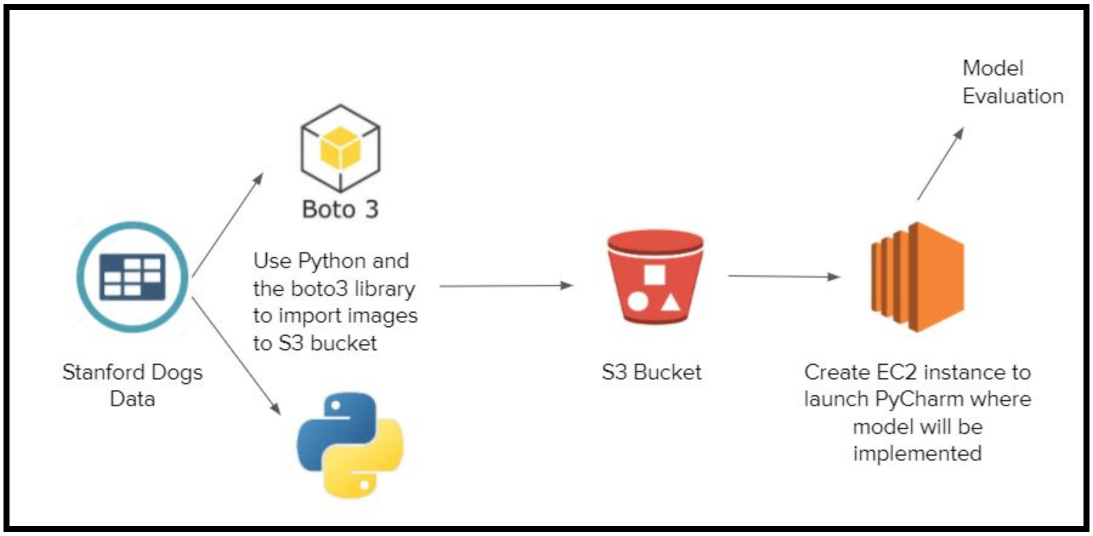
Data-Preprocessing
To make sure the program works well and the expected outcomes are obtained, the dataset will be explored to ensure the data is clean, there are no missing values and ensure it can be easily manipulated. Once this has been confirmed, the entire dataset will be stored in a single folder within an S3 bucket, in preparation for the classification.

Model Implementation
The script from a previous machine learning project will be used to classify the dog breeds based on the images. Before this project is carried out, the code was tested to make sure it is working as expected.
Then, an EC2 instance will be set up to access Python from SSH and run the model.
Finally, the model will be used to categorize dog breeds.

Validation
The model will be validated by calculating the accuracy and loss of the breed classification and whether or not the model classified the correct breeds from the testing set.

PROJECT IMPLEMENTATION

The data flow for this project can be found below:


Model Evaluation
Boto 3
Use Python and the boto3 library to import images to S3 bucket
Stanford Dogs Data
S3 Bucket
Create EC2 instance to launch PyCharm where model will be implemented

The image shows how our model will be implemented, including what features from AWS will be used.

The Stanford dogs dataset will be inspected, and once satisfied, the images will all be uploaded into one folder. Only one S3 bucket will be used as it is much easier to keep all of the information in one bucket and removes a lot of work.

An EC2 instance will be initialized on AWS, which will be used to launch PyCharm on an Ubuntu virtual machine using SSH. From Pycharm, the model will be run on python and will attempt to classify dog breeds based on the images provided. The script will be modified so that it creates a master folder with all the images initially, and then creates other folders based on the breed classifications, which will be stored in their corresponding folders.

Finally, the success of the model will be determined based on a few factors. The main measure of success will be the model's capability to successfully categorize the different breeds. Another measure of success is how easy it is to implement this project using AWS as opposed to working locally. It's usability between different members of the group will also be looked at to determine whether or not using AWS makes it more efficient and easier to collaborate.

RESULTS

The model successfully classified the test set (8,580 images) with an accuracy of 0.8480 and a validation loss of 0.8236.

The images were successfully uploaded to our S3 bucket using boto3.

CONCLUSION

Deploying the project in AWS tools allowed us to successfully implement a project from start to finish using the variety of cloud services available. We were able to take a developed model and build the pipeline using tools like Boto3, S3, and EC2 with strong results. Since the dataset was curated, the next steps we could explore would be working with raw data that needs more cleaning processes. There are many ways that this project could be expanded, and we feel accomplished that the bulk of the machine learning processes are already completed.

REFERENCES

Dataset Reference

Primary:
Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao and Li Fei-Fei. Novel dataset for Fine-Grained Image Categorization. First Workshop on Fine-Grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. [pdf] [poster] [BibTex]

Secondary:
J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009. [pdf] [BibTex]