Image Classification using Pyspark

Franco Bueno Mattera
CS 777 O1

Project Description

- Compare Performance of Image classification algorithms on MLlib
  - Two MLlib Algorithms vs Transfer Learning
- Explore General feasibility of using Pyspark for Image Classification
  - Complexity
  - Challenges

Methodology

About the dataset:

- Arboles de Chile Dataset from Kaggle Website
- Over 6000 images of chilean trees (color images)
- 300 x 300 x 3
- 3 species selected:
  - Lithraea caustica
  - Peumus boldus
  - Ulmus americanas

Peumus Boldus

https://www.kaggle.com/code/mpwolke/rboles-en-chile/data

Methodology

Preparation of the dataset:

- **Transformations:**
  - Labeling images
  - Vectorization
  - Matching labels with order of images
  - Rotating Images to increase train size
  - Reducing image size to 64, 64, 3

Lithraea caustica

https://www.kaggle.com/code/mpwolke/rboles-en-chile/data

Methodology

Training The model:

- Multilayer perceptron classifier
  - Available on pyspark
  - Not very customizable
  - Not scalable to multiple layers

- Naive Bayes Classifier
  - Very Fast
  - Not resource intensive
  - Easy to use API

- Transfer Learning Vg-16
  - External Library
  - Not available on MLlib

Ulmus americana

https://www.kaggle.com/code/mpwolke/rboles-en-chile/data

Results

- Performance was not good
    - Same result for Pre-Trained Network

Algorithm	Precision	Recall	F1
MLP	*0.313*	0.3	0.305
NIB	*0.120*	0.133	0.125
VG-16	0.300	0.300	0.267

Discussions

- Similar Results:
  - MLP yielded better relative results
  - Surprising low performance for pre trained network
- One possible explanation for low performance is small dataset
- Another explanation is complexity of classifying complex pictures

https://www.kaggle.com/code/mpwolke/rboles-en-chile/data

Lessons Learned

- Complexity of preparing the data without the proper library API support
- Possible to perform GPU enhancement but most guides are for distributed spark
- Possibility of wrapping Transfer learning on UDF, more research is needed
    - Library was developed(sparkdl) but not maintained in years
- Highest Challenges:
    - Memory issued (Java Heap Space)
    - Computing Time
- It is better to direct efforts on integrating transfer learning into Pyspark

Thank you!