CS4830 Big Data Laboratory

Jan – May 2021

Lab 8 - Assignment

ME17B158 - Omkar Nath

The DL.ipynb file uploaded on moodle uses a pre-trained mobilenet model to run inference on flowers dataset using Pyspark.

1. Modify the above code to run inference on CIFAR 10 dataset using Pyspark. (7)

The code has been modified to run on the CIFAR 10 dataset, and has been given in the attached file *"ME17B158_Lab8.ipynb".*

Screenshots of certain key sections of the Code has been attached below.

```
[7] import pathlib
    data_dir = tf.keras.utils.get_file(origin='http://pjreddie.com/media/files/cifar.tgz', fname='cifar', untar=True)
```

The modified file path for the data set

```
[8] images = spark.read.format("binaryFile").option("recursiveFileLookup", "true").option("pathGlobFilter", "*.png").load(data_dir)
```

Modified data type for reading files.

```
def extract_label(path_col):
  """Extract label from file path using built-in SQL functions."""
  return regexp_extract(path_col, "_/([^/]+)", 1)
```

Modified file path location.

```
[14] # Using mobilenetv2
     mobilenet_v2_udf = imagenet_model_udf(lambda: models.mobilenet_v2(pretrained=True))

     predictions = df.withColumn("prediction", mobilenet_v2_udf(col("content")))
     display(predictions.select(col("path"), col("prediction")).show(5, truncate = False))

     /content/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py:392: UserWarning: In Py
       "in the future releases. See SPARK-28264 for more details.", UserWarning)
     +----------------------------------------------------+----------------------------------------+
     |path                                                |prediction                              |
     +----------------------------------------------------+----------------------------------------+
     |file:/root/.keras/datasets/cifar/test/4672_frog.png |{n02128925, jaguar, 13.927441}          |
     |file:/root/.keras/datasets/cifar/test/8562_bird.png |{n07745940, strawberry, 9.118496}       |
     |file:/root/.keras/datasets/cifar/train/10327_frog.png|{n01630670, common_newt, 11.725792}    |
     |file:/root/.keras/datasets/cifar/train/23455_deer.png|{n02114712, red_wolf, 7.620327}        |
     |file:/root/.keras/datasets/cifar/train/38450_frog.png|{n02457408, three-toed_sloth, 9.893125}|
     +----------------------------------------------------+----------------------------------------+
     only showing top 5 rows

     None
```

Running mobilenet_v2 on the dataset, along with the results.

2. Try out a few different models pre-trained on Imagenet and report which one works better (calculating exact accuracy is difficult as the class names in imagenet and CIFAR 10 dataset don't exactly match, but still printing out the predictions for a few points and looking at the class names should give a hint). (3)

Zip the code file along with a pdf having the observations about different models.

Various pre-trained models are tried. Specifically, the models that have been tried are as listed in the below table:

mobilenet_v2
resnet18
alexnet
vgg16
googlenet
mobilenet_v3_large
mobilenet_v3_small
resnext50_32x4d
wide_resnet50_2

Upon analyzing the results obtained from the various models, it is found that two models are able to perform decently on the dataset, which are "**mobilenet_v3_large**" and "**resnext50_32x4d**", as both seem to be able to predict certain classes of animals correctly, such as frogs. The rest perform very poorly.

Between the two, mobilenet_v3_large seems to perform slightly better, as it is also able to predict certain classes like of deer.

Below are screenshots of the results obtained from the various models:

```
[14] # Using mobilenetv2
     mobilenet_v2_udf = imagenet_model_udf(lambda: models.mobilenet_v2(pretrained=True))

     predictions = df.withColumn("prediction", mobilenet_v2_udf(col("content")))
     display(predictions.select(col("path"), col("prediction")).show(5, truncate = False))

     /content/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py:392: UserWarning: In Py
       "in the future releases. See SPARK-28264 for more details.", UserWarning)
     +----------------------------------------------------+----------------------------------------+
     |path                                                |prediction                              |
     +----------------------------------------------------+----------------------------------------+
     |file:/root/.keras/datasets/cifar/test/4672_frog.png  |{n02128925, jaguar, 13.927441}          |
     |file:/root/.keras/datasets/cifar/test/8562_bird.png  |{n07745940, strawberry, 9.118496}       |
     |file:/root/.keras/datasets/cifar/train/10327_frog.png|{n01630670, common_newt, 11.725792}     |
     |file:/root/.keras/datasets/cifar/train/23455_deer.png|{n02114712, red_wolf, 7.620327}         |
     |file:/root/.keras/datasets/cifar/train/38450_frog.png|{n02457408, three-toed_sloth, 9.893125}|
     +----------------------------------------------------+----------------------------------------+
     only showing top 5 rows

     None
```

Using mobilenet_v2

```
[15] # Using resnet18
     resnet18_udf = imagenet_model_udf(lambda: models.resnet18(pretrained=True))

     predictions = df.withColumn("prediction", resnet18_udf(col("content")))
     display(predictions.select(col("path"), col("prediction")).show(5, truncate = False))

     /content/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py:392: UserWarning: In
       "in the future releases. See SPARK-28264 for more details.", UserWarning)
     +-----------------------------------------------------+-----------------------------------+
     |path                                                 |prediction                         |
     +-----------------------------------------------------+-----------------------------------+
     |file:/root/.keras/datasets/cifar/test/4672_frog.png  |{n02130308, cheetah, 11.560522}    |
     |file:/root/.keras/datasets/cifar/test/8562_bird.png  |{n01443537, goldfish, 10.951969}   |
     |file:/root/.keras/datasets/cifar/train/10327_frog.png|{n01744401, rock_python, 7.9980874}|
     |file:/root/.keras/datasets/cifar/train/23455_deer.png|{n02114712, red_wolf, 7.421795}    |
     |file:/root/.keras/datasets/cifar/train/38450_frog.png|{n01871265, tusker, 8.802372}      |
     +-----------------------------------------------------+-----------------------------------+
     only showing top 5 rows

     None
```

Using resnet_18

```
[16] # Using alexnet
     alexnet_udf = imagenet_model_udf(lambda: models.alexnet(pretrained=True))

     predictions = df.withColumn("prediction", alexnet_udf(col("content")))
     display(predictions.select(col("path"), col("prediction")).show(5, truncate = False))

     /content/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py:392: UserWarning:
       "in the future releases. See SPARK-28264 for more details.", UserWarning)
     +-----------------------------------------------------+-----------------------------------+
     |path                                                 |prediction                         |
     +-----------------------------------------------------+-----------------------------------+
     |file:/root/.keras/datasets/cifar/test/4672_frog.png  |{n02128925, jaguar, 12.137699}     |
     |file:/root/.keras/datasets/cifar/test/8562_bird.png  |{n01443537, goldfish, 7.694492}    |
     |file:/root/.keras/datasets/cifar/train/10327_frog.png|{n01644900, tailed_frog, 7.7939105}|
     |file:/root/.keras/datasets/cifar/train/23455_deer.png|{n02356798, fox_squirrel, 8.7287}  |
     |file:/root/.keras/datasets/cifar/train/38450_frog.png|{n02487347, macaque, 7.625961}     |
     +-----------------------------------------------------+-----------------------------------+
     only showing top 5 rows

     None
```

Using alexnet

```
[17] # Using vgg16
     vgg16_udf = imagenet_model_udf(lambda: models.vgg16(pretrained=True))

     predictions = df.withColumn("prediction", vgg16_udf(col("content")))
     display(predictions.select(col("path"), col("prediction")).show(5, truncate = False))

     /content/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py:392: UserWarning: In Pyt
       "in the future releases. See SPARK-28264 for more details.", UserWarning)
     +-----------------------------------------------------+-----------------------------------+
     |path                                                 |prediction                         |
     +-----------------------------------------------------+-----------------------------------+
     |file:/root/.keras/datasets/cifar/test/4672_frog.png  |{n13037406, gyromitra, 11.249056}  |
     |file:/root/.keras/datasets/cifar/test/8562_bird.png  |{n02002724, black_stork, 7.4124026}|
     |file:/root/.keras/datasets/cifar/train/10327_frog.png|{n02128925, jaguar, 10.290902}     |
     |file:/root/.keras/datasets/cifar/train/23455_deer.png|{n02099601, golden_retriever, 7.6250567}|
     |file:/root/.keras/datasets/cifar/train/38450_frog.png|{n02457408, three-toed_sloth, 8.371567}|
     +-----------------------------------------------------+-----------------------------------+
     only showing top 5 rows

     None
```

Using vgg16

```
[18] # Using googlenet
     googlenet_udf = imagenet_model_udf(lambda: models.googlenet(pretrained=True))

     predictions = df.withColumn("prediction", googlenet_udf(col("content")))
     display(predictions.select(col("path"), col("prediction")).show(5, truncate = False))

     /content/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py:392: UserWarning:
       "in the future releases. See SPARK-28264 for more details.", UserWarning)
     +--------------------------------------------------+-----------------------------------+
     |path                                              |prediction                         |
     +--------------------------------------------------+-----------------------------------+
     |file:/root/.keras/datasets/cifar/test/4672_frog.png  |{n02128925, jaguar, 8.538764}      |
     |file:/root/.keras/datasets/cifar/test/8562_bird.png  |{n02606052, rock_beauty, 5.672103}|
     |file:/root/.keras/datasets/cifar/train/10327_frog.png|{n02128385, leopard, 6.2645426}    |
     |file:/root/.keras/datasets/cifar/train/23455_deer.png|{n03016953, chiffonier, 5.7657547}|
     |file:/root/.keras/datasets/cifar/train/38450_frog.png|{n07730033, cardoon, 3.9538639}    |
     +--------------------------------------------------+-----------------------------------+
     only showing top 5 rows

     None
```

Using googlenet

```
[19] # Using mobilenet_v3_large
     mobilenet_v3_large_udf = imagenet_model_udf(lambda: models.mobilenet_v3_large(pretrained=True)

     predictions = df.withColumn("prediction", mobilenet_v3_large_udf(col("content")))
     display(predictions.select(col("path"), col("prediction")).show(5, truncate = False))

     /content/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py:392: UserWarning: In
       "in the future releases. See SPARK-28264 for more details.", UserWarning)
     +--------------------------------------------------+-----------------------------------+
     |path                                              |prediction                         |
     +--------------------------------------------------+-----------------------------------+
     |file:/root/.keras/datasets/cifar/test/4672_frog.png  |{n01644900, tailed_frog, 8.5415745}|
     |file:/root/.keras/datasets/cifar/test/8562_bird.png  |{n01818515, macaw, 7.593056}       |
     |file:/root/.keras/datasets/cifar/train/10327_frog.png|{n01644900, tailed_frog, 8.550882} |
     |file:/root/.keras/datasets/cifar/train/23455_deer.png|{n02423022, gazelle, 7.046612}     |
     |file:/root/.keras/datasets/cifar/train/38450_frog.png|{n13037406, gyromitra, 8.226663}   |
     +--------------------------------------------------+-----------------------------------+
     only showing top 5 rows

     None
```

Using mobilenet_v3_large

```
[20] # Using mobilenet_v3_small
     mobilenet_v3_small_udf = imagenet_model_udf(lambda: models.mobilenet_v3_small(pretrained=True

     predictions = df.withColumn("prediction", mobilenet_v3_small_udf(col("content")))
     display(predictions.select(col("path"), col("prediction")).show(5, truncate = False))

     /content/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py:392: UserWarning: In
       "in the future releases. See SPARK-28264 for more details.", UserWarning)
     +--------------------------------------------------+-----------------------------------+
     |path                                              |prediction                         |
     +--------------------------------------------------+-----------------------------------+
     |file:/root/.keras/datasets/cifar/test/4672_frog.png  |{n02130308, cheetah, 8.274256}     |
     |file:/root/.keras/datasets/cifar/test/8562_bird.png  |{n02002724, black_stork, 6.1072235}|
     |file:/root/.keras/datasets/cifar/train/10327_frog.png|{n01630670, common_newt, 8.539051} |
     |file:/root/.keras/datasets/cifar/train/23455_deer.png|{n02114712, red_wolf, 10.001152}   |
     |file:/root/.keras/datasets/cifar/train/38450_frog.png|{n13037406, gyromitra, 7.852109}   |
     +--------------------------------------------------+-----------------------------------+
     only showing top 5 rows

     None
```

Using mobilenet_v3_small

```
[21] # Using resnext50_32x4d
     resnext50_32x4d_udf = imagenet_model_udf(lambda: models.resnext50_32x4d(pretrained=True))

     predictions = df.withColumn("prediction", resnext50_32x4d_udf(col("content")))
     display(predictions.select(col("path"), col("prediction")).show(5, truncate = False))
```

```
/content/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py:392: UserWarning:
  "in the future releases. See SPARK-28264 for more details.", UserWarning)
+--------------------------------------------------+-----------------------------------+
|path                                              |prediction                         |
+--------------------------------------------------+-----------------------------------+
|file:/root/.keras/datasets/cifar/test/4672_frog.png  |{n02128925, jaguar, 15.668928}     |
|file:/root/.keras/datasets/cifar/test/8562_bird.png  |{n02002724, black_stork, 9.183935} |
|file:/root/.keras/datasets/cifar/train/10327_frog.png|{n01644900, tailed_frog, 7.7040114}|
|file:/root/.keras/datasets/cifar/train/23455_deer.png|{n02090379, redbone, 8.180999}     |
|file:/root/.keras/datasets/cifar/train/38450_frog.png|{n01644900, tailed_frog, 12.080752}|
+--------------------------------------------------+-----------------------------------+
only showing top 5 rows

None
```

Using resnext50_32x4d

```
[22] # Using wide_resnet50_2
     wide_resnet50_2_udf = imagenet_model_udf(lambda: models.wide_resnet50_2(pretrained=True))

     predictions = df.withColumn("prediction", wide_resnet50_2_udf(col("content")))
     display(predictions.select(col("path"), col("prediction")).show(5, truncate = False))
```

```
/content/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py:392: UserWarning: In P
  "in the future releases. See SPARK-28264 for more details.", UserWarning)
+--------------------------------------------------+-----------------------------------+
|path                                              |prediction                         |
+--------------------------------------------------+-----------------------------------+
|file:/root/.keras/datasets/cifar/test/4672_frog.png  |{n02128925, jaguar, 13.496078}     |
|file:/root/.keras/datasets/cifar/test/8562_bird.png  |{n07714990, broccoli, 7.3184166}   |
|file:/root/.keras/datasets/cifar/train/10327_frog.png|{n01630670, common_newt, 8.448229} |
|file:/root/.keras/datasets/cifar/train/23455_deer.png|{n02129604, tiger, 9.687185}       |
|file:/root/.keras/datasets/cifar/train/38450_frog.png|{n01688243, frilled_lizard, 9.081802}|
+--------------------------------------------------+-----------------------------------+
only showing top 5 rows

None
```

wide_resnet50_2