Auto-categorising Mobile Devices

Russell Foo

Original Challenge

National Data Science Challenge 2019 by Shopee to create an automatic solution to extract product information from large volumes of images and free text.

About the Data Set

160k rows with 14 columns

➔ **Rows**

Every row has an empty column

➔ **Columns**

Brand has the most info followed by Colour

➔ **Approach**

'Divide-and-conquer' approach, tackling each column individually; focusing on Brand and Color

itemid	0
title	0
image_path	0
Brand	5292
Color Family	75499
Phone Model	75821
Storage Capacity	97180
Memory RAM	103293
Warranty Period	112257
Operating System	115052
Features	120985
Phone Screen Size	127636
Camera	135740
Network Connections	136906

Problem Statement

Auto-categorising Mobile Product Information

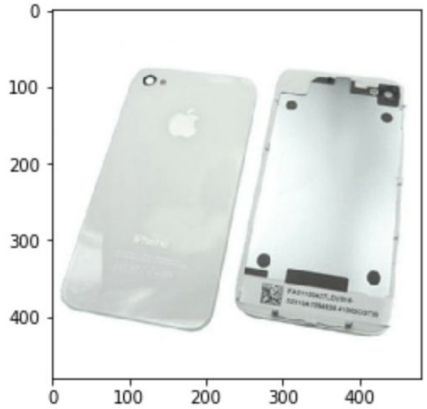
(Brand & Colour) from Free Texts & Images

Brand
Color Family

||

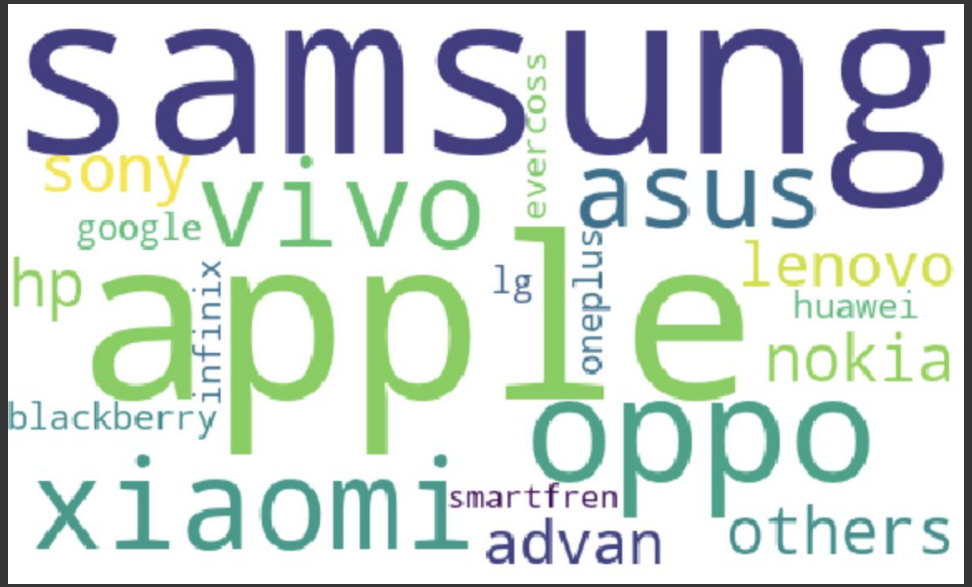
apple iphone 4s back glass spare part original...

+

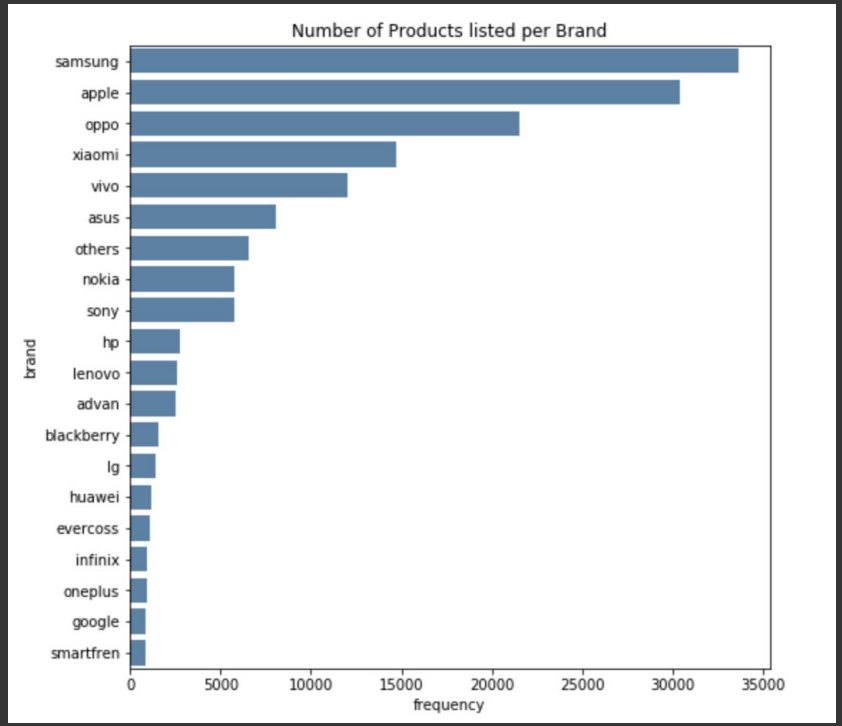


+

Mobile Brand



Handling the Data



brand	frequency
samsung	33667
apple	30399
oppo	21548
xiaomi	14679
vivo	11982
asus	8042
others	6584
nokia	5785
sony	5754
hp	2726
lenovo	2607
advan	2520
blackberry	1537
lg	1413
huawei	1186
evercoss	1091
infinix	907
oneplus	902
google	869
smartfren	840

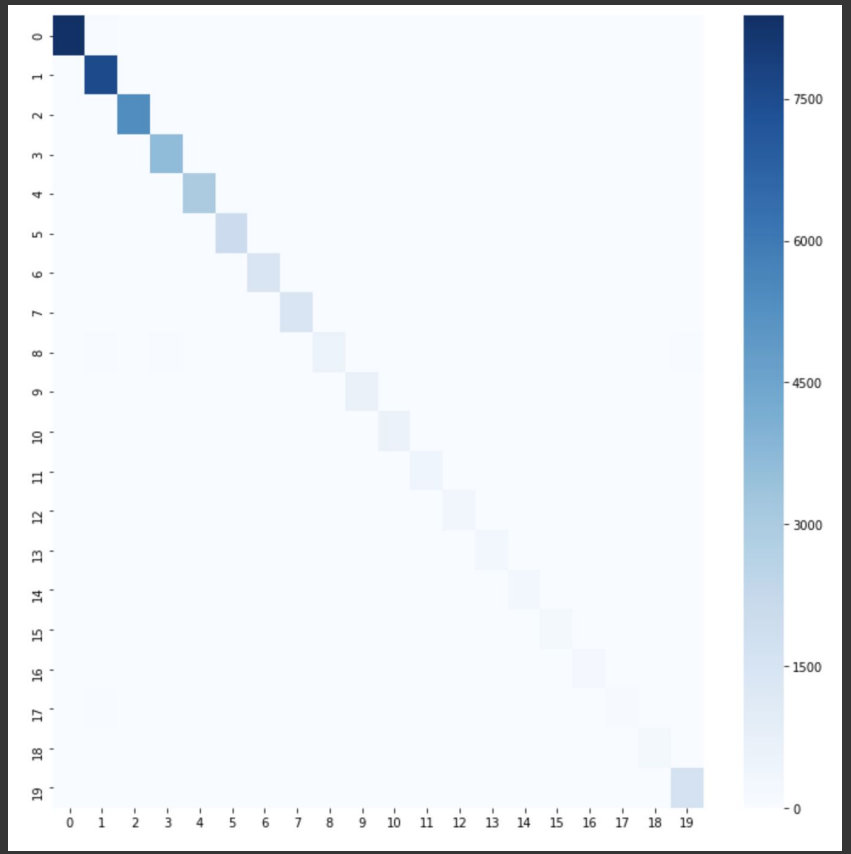
Logistic Regression

Model	Baseline Accuracy	Validation Accuracy
Logistic Regression	0.99125	0.98566
Logistics Regression with Custom Stop Words	0.99008	0.98584

```
['promo', 'ram', 'garansi', '64gb', 'redmi', 'wa', '32gb', '4g', '4gb', 'new', 'resmi', 'beli', 'pro', 'plus', 'ipa
d', 'note', 'murah', 'black', 'chat', 'wifi', 'gratis', 'original', 'android', 'garansi resmi', 'rom']
```

Logistic Regression

brand_val	brand
0	samsung
1	apple
2	oppo
3	xiaomi
4	vivo
5	asus
6	nokia
7	sony
8	hp
9	lenovo
10	advan
11	blackberry
12	lg
13	huawei
14	evercoss
15	infinix
16	oneplus
17	google
18	smartfren
19	others



Findings

Looking at data set and model again

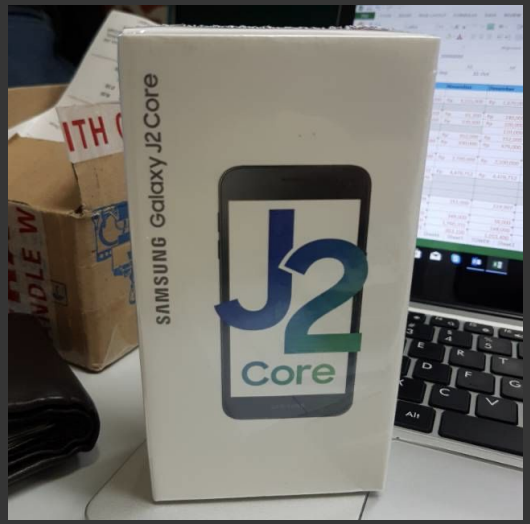
➔ **Custom Stop Words**

Improved accuracy of the model on unseen data probably due to the common specs of phones across the market (e.g. 4G, Wifi).

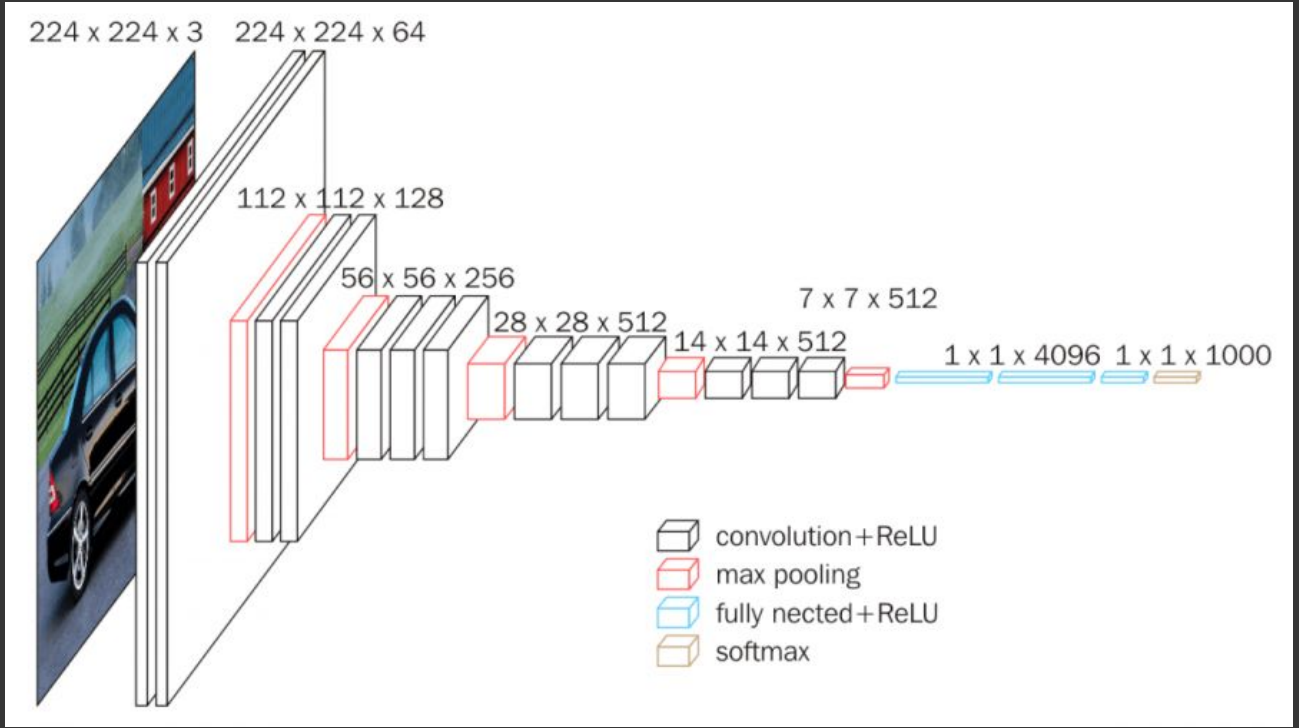
➔ **More Stop Words?**

Free texts included non-english terms which need to be picked up by domain expert.

Mobile Colour



VGG16


224 x 224 x 3
224 x 224 x 64
112 x 112 x 128
56 x 56 x 256
28 x 28 x 512
14 x 14 x 512
7 x 7 x 512
1 x 1 x 4096
1 x 1 x 1000
convolution+ReLU
max pooling
fully nected+ReLU
softmax

VGG16 - Transfer Learning

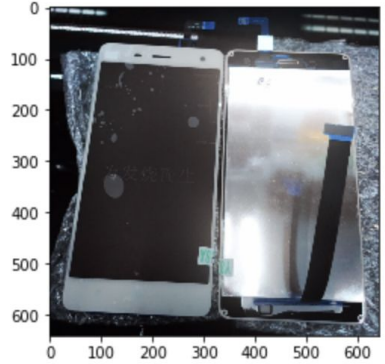
In [33]:
```
history = model.fit(X_train,
                    y_train,
                    batch_size=256,
                    validation_data=(X_test, y_test),
                    epochs=10,
                    verbose=1)
```

```
Train on 15000 samples, validate on 5000 samples
Epoch 1/10
15000/15000 [==============================] - 3040s 203ms/step - loss: 4.6317 - accuracy: 0.2871 - val_loss: 3.3489
- val_accuracy: 0.3370
Epoch 2/10
15000/15000 [==============================] - 1207s 80ms/step - loss: 2.8446 - accuracy: 0.3869 - val_loss: 2.6471 -
val_accuracy: 0.3946
Epoch 3/10
15000/15000 [==============================] - 1202s 80ms/step - loss: 2.3156 - accuracy: 0.4231 - val_loss: 2.3163 -
val_accuracy: 0.4010
Epoch 4/10
15000/15000 [==============================] - 1524s 102ms/step - loss: 2.0176 - accuracy: 0.4543 - val_loss: 2.0945
- val_accuracy: 0.4538
Epoch 5/10
15000/15000 [==============================] - 1231s 82ms/step - loss: 1.8228 - accuracy: 0.4747 - val_loss: 1.9324 -
val_accuracy: 0.4600
Epoch 6/10
15000/15000 [==============================] - 1256s 84ms/step - loss: 1.6783 - accuracy: 0.4923 - val_loss: 1.8451 -
val_accuracy: 0.4628
Epoch 7/10
15000/15000 [==============================] - 1239s 83ms/step - loss: 1.5840 - accuracy: 0.5075 - val_loss: 1.7487 -
val_accuracy: 0.4780
Epoch 8/10
15000/15000 [==============================] - 1202s 80ms/step - loss: 1.5034 - accuracy: 0.5203 - val_loss: 1.7136 -
val_accuracy: 0.4670
Epoch 9/10
15000/15000 [==============================] - 1245s 83ms/step - loss: 1.4479 - accuracy: 0.5307 - val_loss: 1.6576 -
val_accuracy: 0.5012
Epoch 10/10
15000/15000 [==============================] - 1286s 86ms/step - loss: 1.4141 - accuracy: 0.5389 - val_loss: 1.6059 -
val_accuracy: 0.4946
```

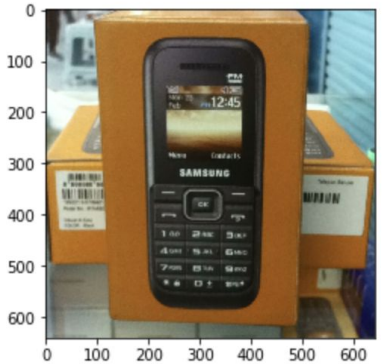
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
global_average_pooling2d_1 (	(None, 512)	0
dense_1 (Dense)	(None, 8)	4104

```
Total params: 14,718,792
Trainable params: 4,104
Non-trainable params: 14,714,688
```

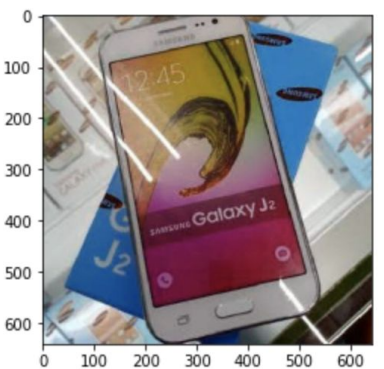
VGG16 - Predictions



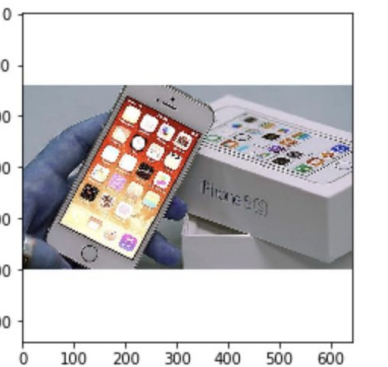
```
   new_color_val  color
2              2  white
```




```
   new_color_val  color
0              0  black
```




```
   new_color_val color
1              1  gold
```



```
   new_color_val  color
0              0  black
```

Findings

Looking at data set and model again

➔ **White = Gold?**

Many white phones were predicted as gold likely due to the packaging of said phones (e.g. Apple rose gold).

➔ **'Bad' Images**

Some of the are uploaded with their packaging and others have uploaded advertisement posters showing multiple phones.

Conclusion

For the auto-categorising feature to be successful,

- **Setting Expectations**
  Setting some guidelines for customers for follow when posting products (e.g. English medium, pictures with less noise).

- **Human Integration**
  In both models, the smaller brands and rarer colours have been clustered under ‘Others’. We will still need human resources to accurately label the devices that get sorted into others.

Thank you for the attention!