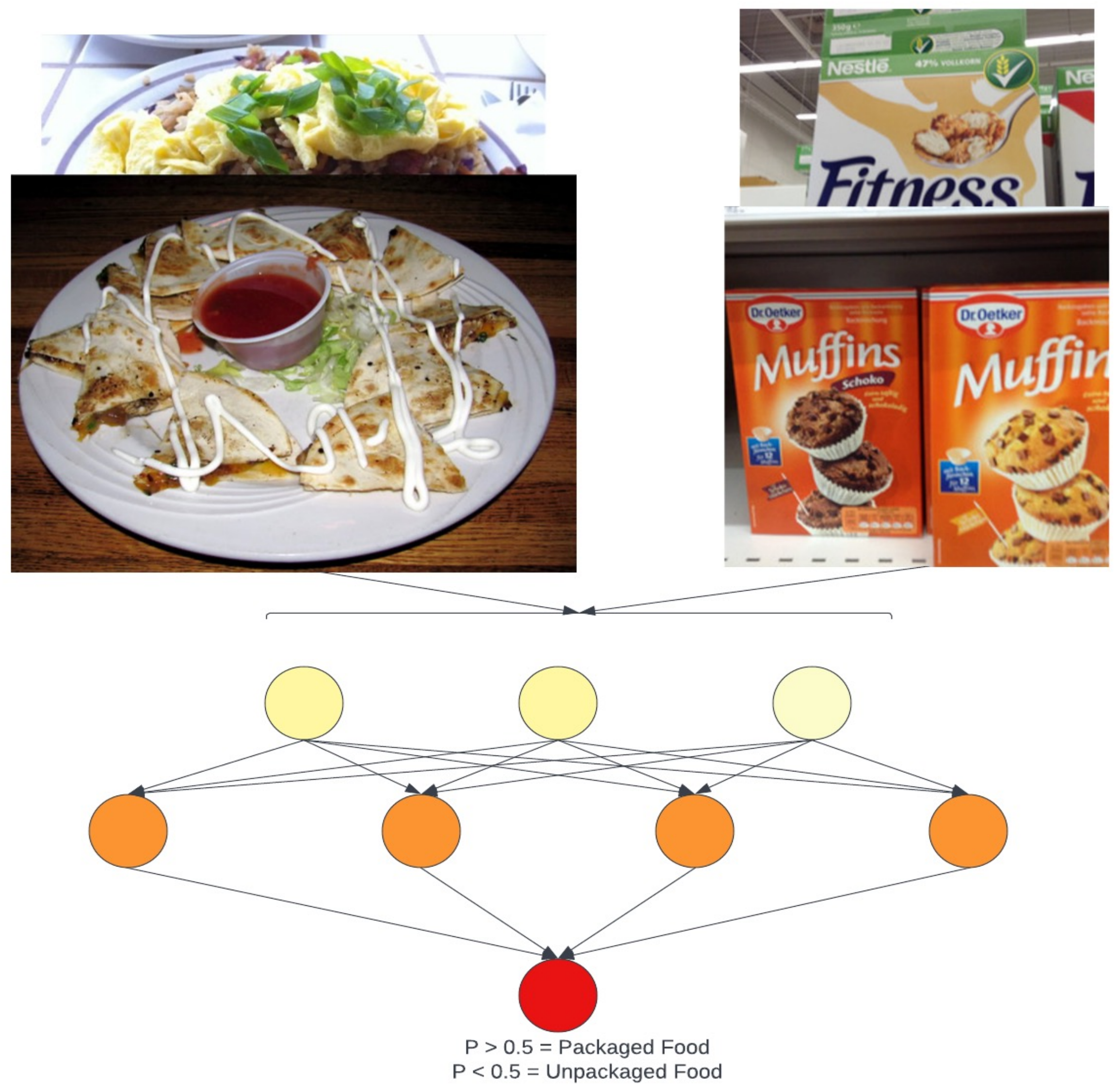
A Deep Learning approach to classifying images of packaged versus unpackaged food

Shubham Bipin Kumar, Himanshu Hansaria, Abhinav Sinha, School of Informatics and Computing, Indiana University

1. Motivation

- The paper[1] explains In case of most of the dietary assessment problem using computer vision users end up feeding the pictures of both the cooked food as well as packaged food as they are the part of their diet.
- At the same time photos show only packaging of the food and not its content, which is actual food .
- This gives us motivation to first address the problem of identifying unpackaged food and packaged food and further explore the possibility of identifying the contents of the packaged food using textual or other properties.
- At the present we were not able to find any model that tries to address this problem.



2. Challenges

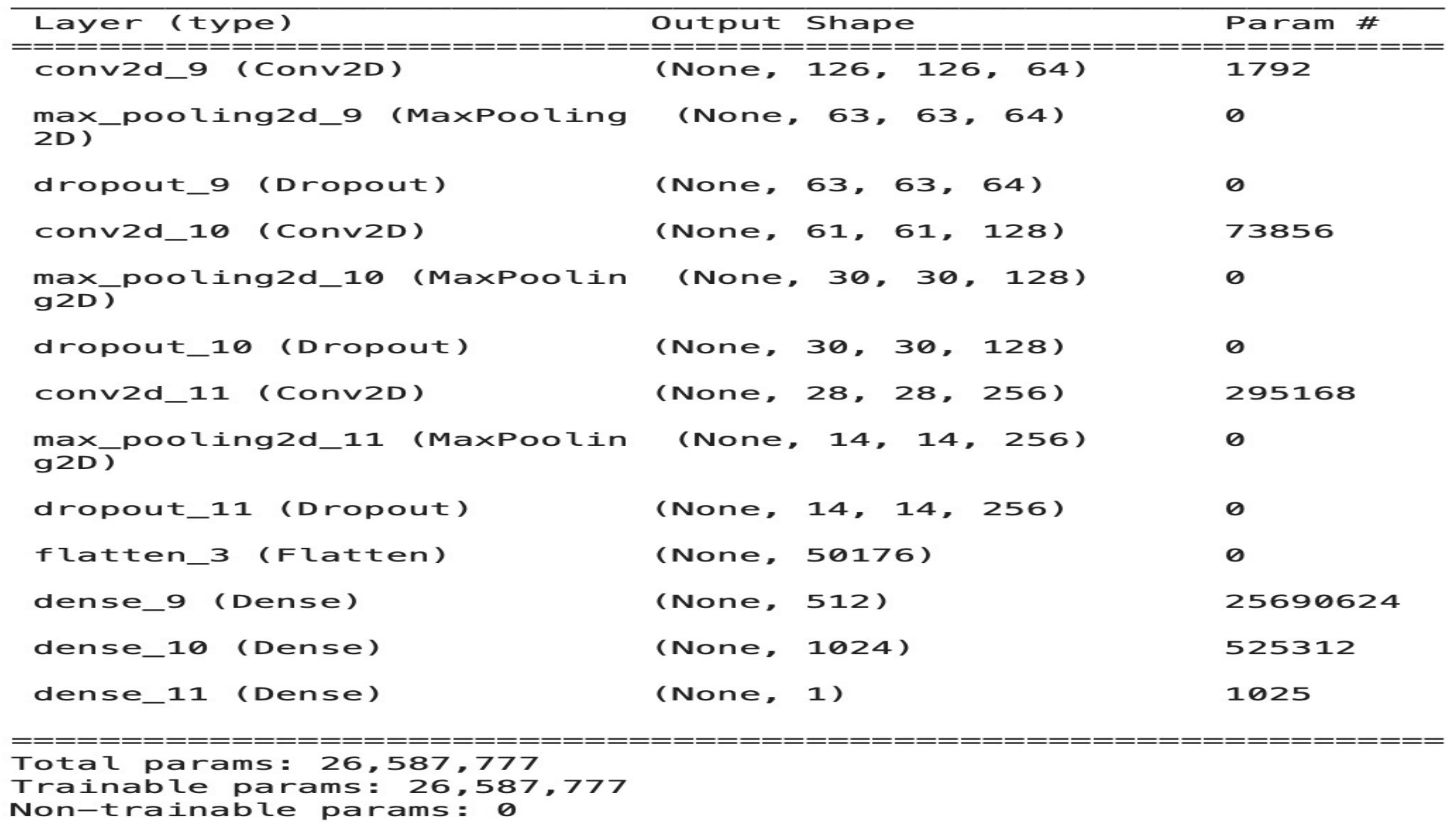
- .Real life photos quality which might be blurred has a major impact on the accuracy of the model.
- Transparent packaging is another issue faced as the model finds it very difficult to classify it as as packaged or unpackaged food owing to the fact that it looks similar to the packaged food.
- Another bigger challenge was confusion created by the presence of non-food item or humans along with food which did not yield good result.
- We spent some time on sorting out different image formats in the dataset and converting them to a single format.

3.CNN Model Design

- As we understand from the literature available on the image classification and identification as well our own experimentation with logistic regression and SVM models, we came to conclusion the deep learning model perform the best .
- We started with a baseline model with3 blocks of sequential conv2d–maxpool layers with 64, 128 and 256 channels respectively. Conv2d kernel size: 3x3, MaxPool: 2x2. No. of channels 64, 128, 256 respectively. This is followed by 2 fully connected layers with 128 and 256 neurons resp. It gave us test accuracy of 80.22 %
- For our best model, we used 3 blocks of sequential conv2d–maxpool layers with 64, 128 and 256 channels respectively. Conv2d kernel size: 3x3, MaxPool: 2x2. No. of channels 64, 128, 256 .
- This is followed by 2 fully connected layers with 512 and 1024 neurons resp.
- We further added dropout of 0.5 for all the 3 convolutional layer (having experimented with 0.3 ,0.4)
- Our model is able to identify the packaged vs unpacked food by identifying the shapes , textual properties in an image which can be observed from the feature maps.

4. Experiments

- We have used the food 5k and Freiburg datasets which contains variety of cooked as well as packaged food.
- We went ahead and curated the data from different format to a single format for our use.
- We further developed a script to classify the data into training , validation and evaluation.
- Our model explained above had 1792 param at conv_2D layer and reached upto 25690624 params in the dense_3 layer.
- We trained our model for 5 epochs each with 94 steps.
- We tried different dropouts at different layers and achieved an accuracy of 96.50 %.
- Identified mis-classified images and using feature map observed which areas were highlighted in packaged and unpackaged food .



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_9 (Conv2D)            (None, 126, 126, 64)      1792

max_pooling2d_9 (MaxPooling   (None, 63, 63, 64)       0
2D)

dropout_9 (Dropout)          (None, 63, 63, 64)        0

conv2d_10 (Conv2D)           (None, 61, 61, 128)       73856

max_pooling2d_10 (MaxPoolin   (None, 30, 30, 128)      0
g2D)

dropout_10 (Dropout)         (None, 30, 30, 128)       0

conv2d_11 (Conv2D)           (None, 28, 28, 256)       295168

max_pooling2d_11 (MaxPoolin   (None, 14, 14, 256)      0
g2D)

dropout_11 (Dropout)         (None, 14, 14, 256)       0

flatten_3 (Flatten)          (None, 50176)             0

dense_9 (Dense)              (None, 512)               25690624

dense_10 (Dense)             (None, 1024)              525312

dense_11 (Dense)             (None, 1)                 1025

=================================================================
Total params: 26,587,777
Trainable params: 26,587,777
Non-trainable params: 0
_________________________________________________________________
```

5. Results

- Our model worked exceedingly well and achieved a maximum accuracy of 96.50 % on the food 5K and Freiburg dataset.
- We are yet to test our model on the more real world images taken from mobile phone and are not curated.
- We have listed out few of the correctly classified and incorrectly classified examples here.
- Feature maps shows the the properties model used in classification process.

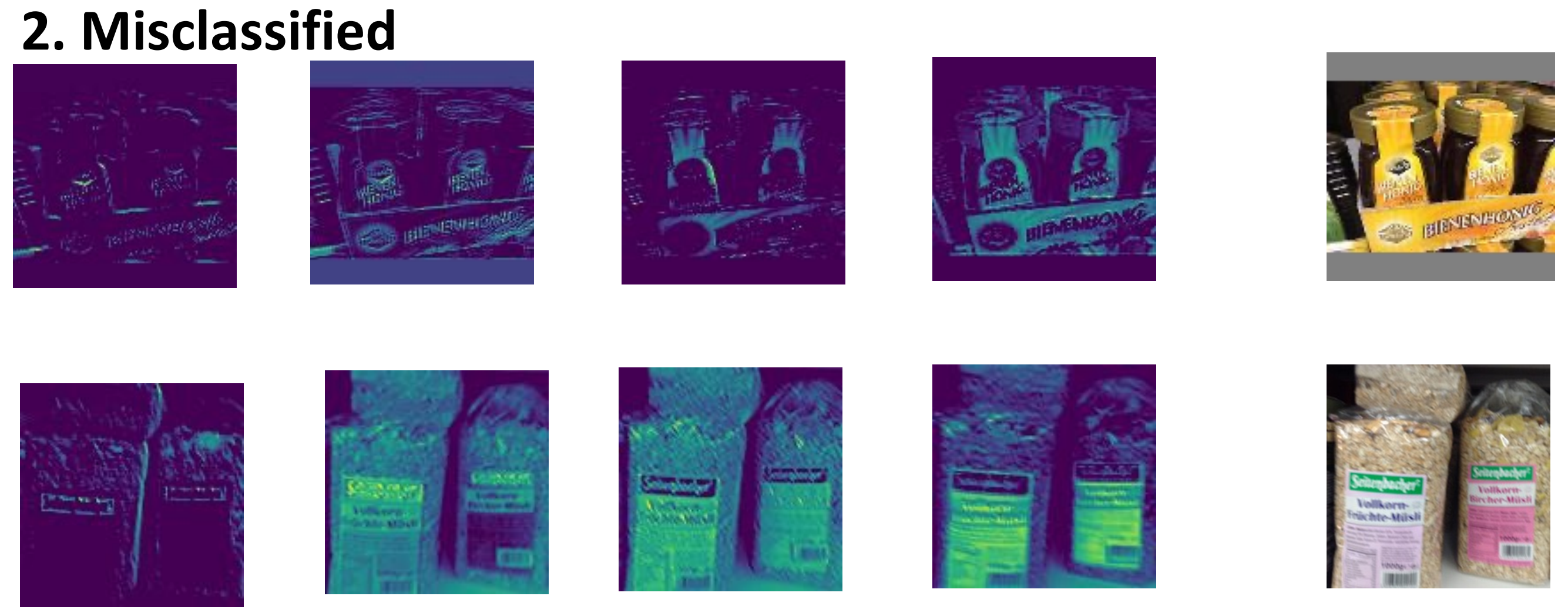
Correctly classified

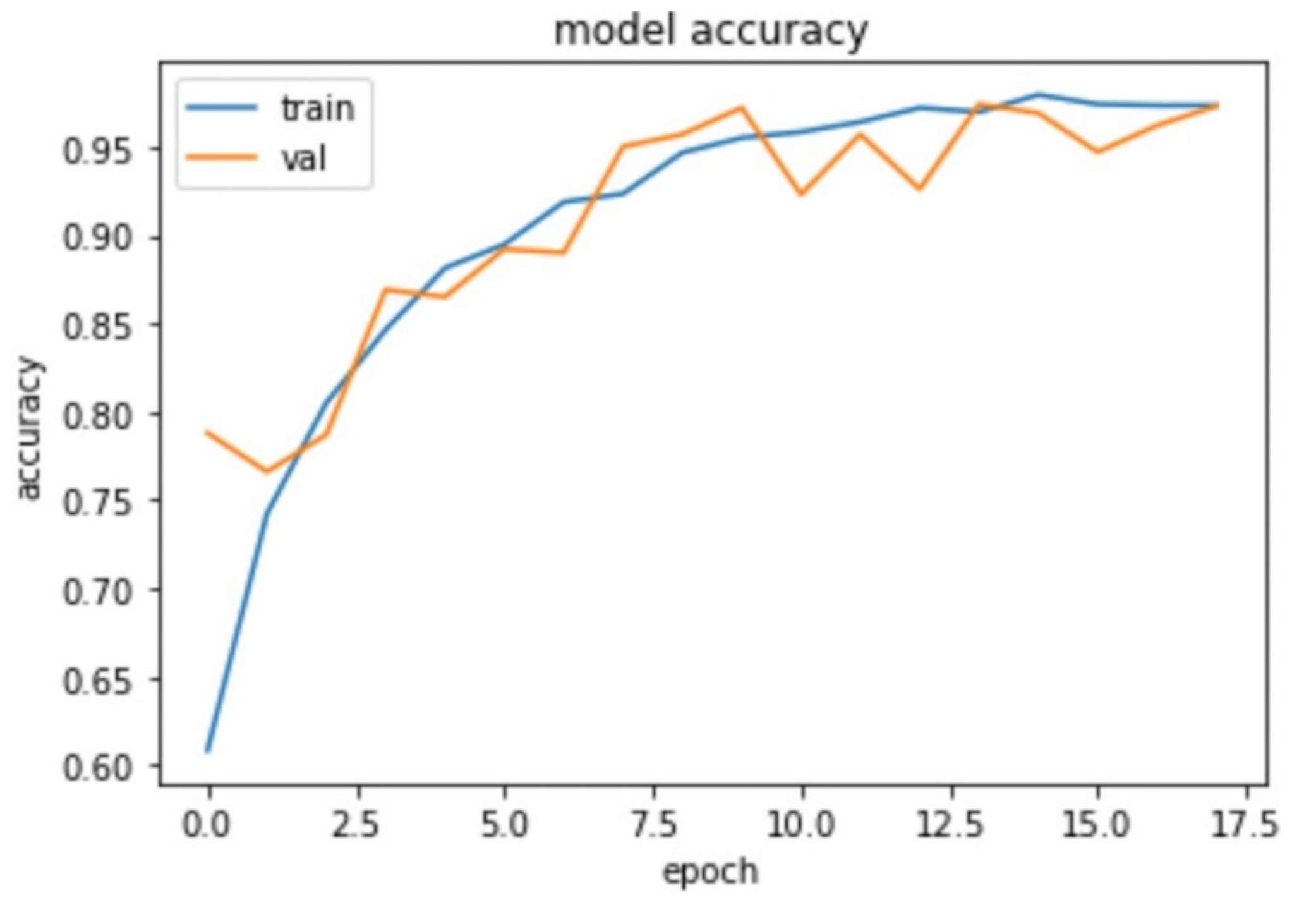
Incorrectly classified

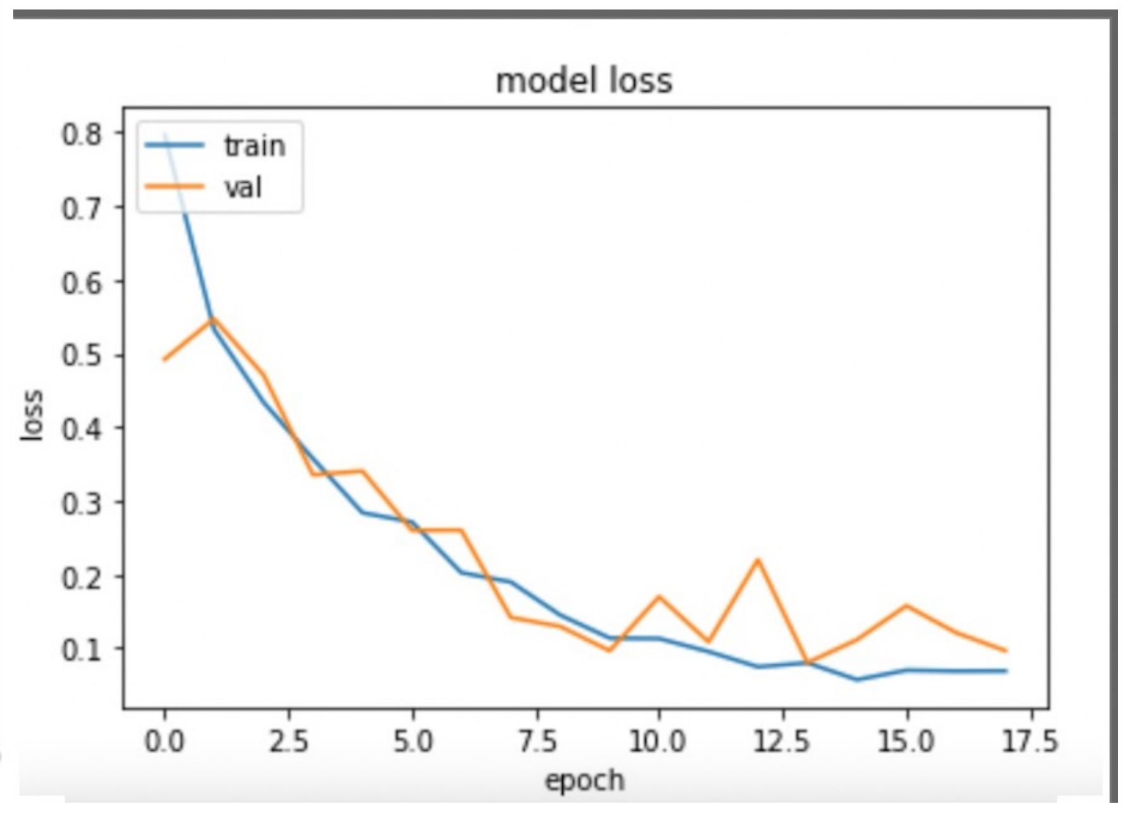
Feature Maps

1. Correct Classified

2. Misclassified







6. Future Work

- We will work on implementing the stretch goal of the project i.e identifying textual features from the images.
- We will be using OCR , barcode and other properties of the packaged images to correctly identify the contains of the package.

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7840289/