Waste Classifier Network

Final Project Report

Team members: Alexander Haynie, Carlos Prieto date : 05/01/20

https://github.com/charlypri/MLfinalProject

Project Topic

The main topic of our project is to generate an image classifier that estimates whether a picture depicts a recyclable object or an organic one. To do this we will design and train a neural network that will carry out this classification task.

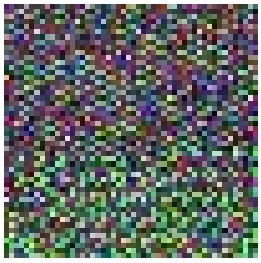
Data

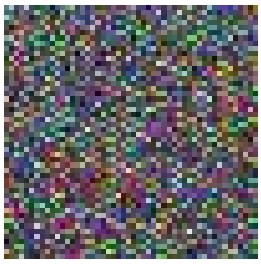
The dataset that we have used can be found at this link. The training set is formed by 12.567 pictures of "organic" elements and 10,000 pictures of "recyclable" elements, which makes this dataset really well-balanced and also big enough to make a really good classifier since the dataset is really broad. The test set is formed by 1401 pictures of "organic" elements and 1112 pictures of "recyclable" elements. This makes our data divided into 85% training images and 15% as test images all of them are RGB colored.

Data cleaning

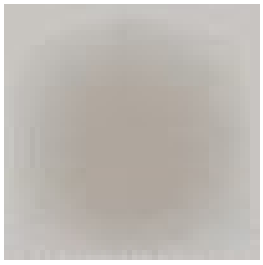
Since our data was formed by raw folders with pictures we had to load and store the information of those pictures into our variable X, and we also had to mark them with their label and then shuffle this data to form our training and testing sets. We also had to apply some data formatting techniques in order to be able to use it as we pleased. We had to find extra pictures for both training sets since there were some pictures in the index list that did not exist. What we did for this was simply duplicate a random picture that already existed and added it on that index. On the other hand we also had to transform our Y labels data set to one-hot encoded data in order to feed it to our network. The data was already balanced and used the same data types.

Exploratory Data Analysis (EDA)



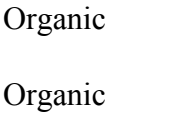


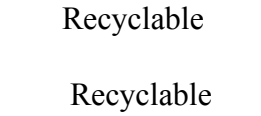
The pictures were all the same resolution and format. Some exploratory analysis we did is



calculate the mean image for both labels; “Recyclable” and “Organic” but since we could not process the images in high resolution due to the lack of computation power in our computers this information was not as good as it could have been. We hoped to identify with this analysis if we could find patterns and characteristics from the average image of each label. Here are the resulting mean images for organic and recyclable respectively:

Organic	Recyclable





Here we can see in using different encodings we obtain different images and with encoding “int8” we saw that the average image for organic has more colorful pixels whereas the recyclable mean image is pretty much grayscale.

Models

The model used throughout the whole machine learning task was Neural Networks since this is proven to be a really good model for image processing tasks, also We started by getting out data in the right format and size to feed it to our neural network. The original size of the pictures is 242, 208 pixels which was way too big to be processed by our computers so we started with an initial size of 50x50 pixels. We started with a train_test_split with a test size of .2 and a random state of 42.

We built a multilayer perceptron, We first used two Keras Conv2D Layers, this layer creates a convolution kernel that is wind with layers input which helps produce a tensor of outputs.After that we add a max pooling layer which is a sample-based discretization process. The objective is to down-sample an input representation in our case the images, reducing its dimensionality and allowing for assumptions to be made about features contained in the sub-regions binned.
Then we set 25% setting a fraction of inputs to zero in an effort to reduce overfitting and we flatten the tensor to one dimension. After we have done this we add two dense layers which are fully connected layers and the most common type of layer used on multi-layer perceptron models with another 50% dropout in between those two. We use a categorical cross entropy loss as the evaluation tool.

```
model = Sequential()
model.add(Conv2D(32, kernel_size=(3, 3),activation='relu',input_shape=(50, 50, 3)))#242, 208, 3 is original shape
model.add(Conv2D(64, (3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.25))
model.add(Flatten())
model.add(Dense(128, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(2, activation='softmax'))
model.compile(loss='categorical_crossentropy',optimizer='Adam',metrics=['accuracy'])
```

We have used different activation functions and different hyperparameters for the multiple layers that we have in our neural network but we did not obtain significant changes on the accuracy results.

6. Results and Analysis

This are the results from the training of our network on 5 epochs and 50x50 resolution:
loss: 0.3649 - accuracy: 0.8525 - val_loss: 0.4010 - val_accuracy: 0.8220

These results are pretty good, and show that our model is catching those intrinsic details that separate organic from recyclable. We have seen an improvement of 10 percent more increased accuracy from our initial results. After training a couple different models we got an 82% accuracy on the training set, and an 90% on the test set.

Conclusion

We have learned that the higher quality of imagery, the better accuracy we get. Our code helps determine if a particular item is organic or recyclable. This learning could be used in automated trash sorting either on a large or small scale. The major difficulties of using our code with trash sorting would be making sure different categories of trash were stacked on top of eachother. Many of the images we used to train and test with contained an item or multiples of that item and little else. For our code to be effective and improve in determining recycling or organic, the items would need to be spread apart enough for the image to isolate the item with enough quality.