Parker Williamson
2/14/2018
Springboard Data Science Career Track

Capstone Project 1 – Clothing Categorization

Problem

Organizing dishes and clothes are some of the most time intensive tasks that are still done manually in the home. I hope to take a step in the direction of alleviating those tasks by creating an automated system for clothes categorization. New items of clothing must also be manually categorized when being added to online stores and clothing categorization can save time for that as well.

Client

Washer and dryer makers and indirectly home owners can benefit from some of this dynamic classification. It has the potential to move towards saving millions of people hours per week. As a result a folding machine should be made if the categorization is right a high percentage of the time. Another segment of customers which will be reached first is online clothes sales.

Dataset

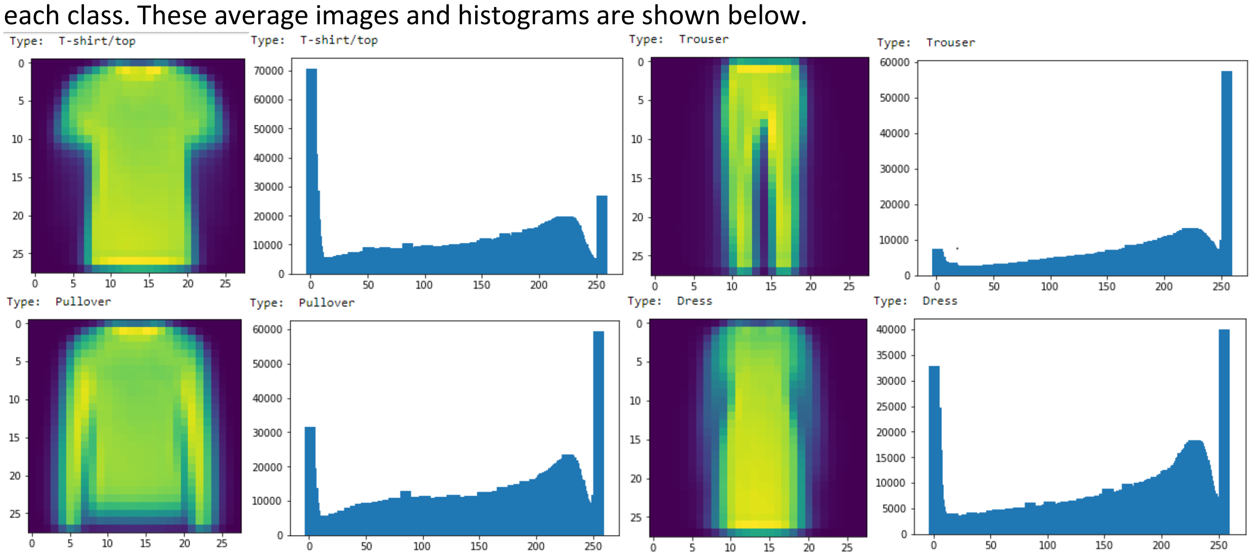
Fashion.mnist (https://www.kaggle.com/zalando-research/fashionmnist/data) will be the main data I use to explore image processing further and train on.

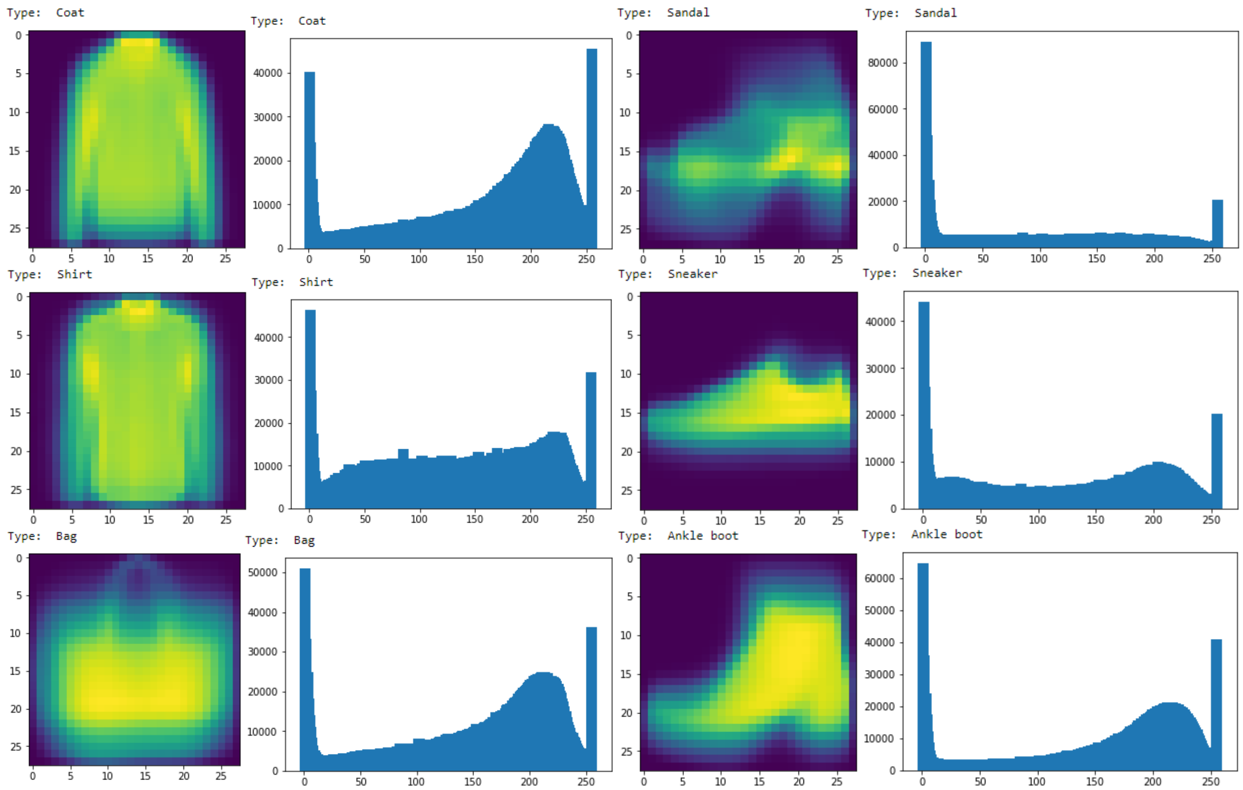
Alternative datasets

If time allow testing against photos taken and may have to improve it further by using another dataset such as Deep Fashion (http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html).

Findings

The clothing MNIST dataset has 10 different clothing types: T-Shirt, Trousers, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot. To get a general understanding of each type I took the average of each type by pixel and made one average image. I also took histograms of all the images for each class. These average images and histograms are shown below.





Printing out averages images and their histograms shows how similar the T-shirt/top category is to shirt and how similar pullovers are to coats. All the other categories have pretty distinct features compared to each other. Coats tend to be a little larger/brighter than pullovers and there's a little bit of a button line. Shirts sometimes have sleeves, where T-shirts do not and also have a little bit of a button line.

Preprocessing was the next part of the categorization that I explored. The technique was different depending on if you were running CNN or one of the other categorization techniques. For Multi-layer Perceptron (MLP), SVC and Logistic regression it was more beneficial to normalize and standardize to a 0-1 range. When using CNN standardizing and normalizing to a 0-255 range improves the results, the images were initially normalized to a 0-255 range. For CNN standardization improves results by ~1%.

In addition to analyzing the models, the image categories were explored as well. In order to compare how correlated the categories of clothing were, the means of the categories were compared. The most similar means would be most similar in value, so the means were ordered from smallest to largest and then a z-test was performed to determine if the different categories have the same means statistically similar means. The ordered means of the groups have no statistically significant correlation, because all of the p-scores are below .025 (the p-scores are show below). A p-score shows the results of the null hypothesis test determining where or not the means are correlated. It is not a probability, but you can see how far away the results are from being above .025.

```
p-score 0-1(Z): 0.0
p-score 1-2(Z): 0.0
p-score 2-3(Z): 0.0
p-score 3-4(Z): 0.0
p-score 4-5(Z): 0.0
p-score 5-6(Z): 0.0001
p-score 6-7(Z): 0.0
p-score 7-8(Z): 0.0
p-score 8-9(Z): 0.0
```

The predicted (using standardized MLP) versus actual results were compared and the results are shown below the diagonal large values on the diagonal clearly show that large majority were correctly classified. The rest of the values (not correct values) were analyzed using percentile to look for 95% outliers.

X – Predicted

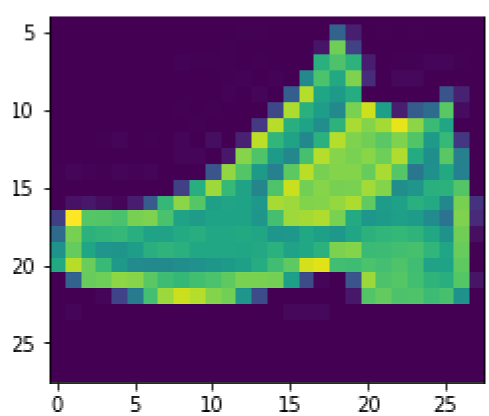
Y – Actual (top is 0)

```
[[ 853.    0.   10.   15.    1.    1.  114.    1.    4.    1.]
 [   4.  982.    1.    6.    2.    0.    3.    0.    1.    1.]
 [  23.    0.  663.    9.  201.    1.   98.    0.    4.    1.]
 [  48.   12.    8.  878.   21.    0.   31.    0.    1.    1.]
 [   3.    2.   36.   24.  866.    0.   66.    0.    3.    0.]
 [   0.    0.    0.    0.    0.  895.    0.   57.    3.   45.]
 [ 152.    1.   38.   17.   94.    1.  688.    0.    9.    0.]
 [   0.    0.    0.    0.    0.    6.    0.  967.    1.   26.]
 [  11.    0.    4.    1.    4.    2.   12.    8.  956.    2.]
 [   0.    0.    0.    0.    0.    5.    0.   44.    0.  951.]]
```

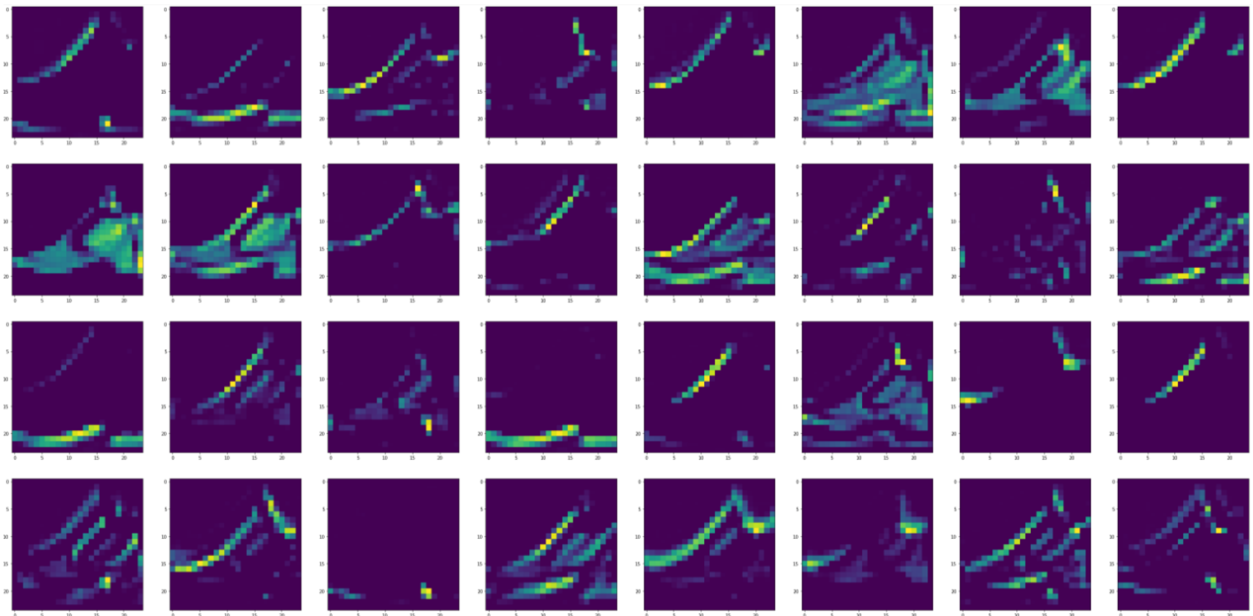
There were three values that stood out, and they were: T-shirts/tops were commonly classified as shirts, Pullovers were commonly classified as coats and shirts were commonly classified as T-shirts/tops. Those are some of the categories that would be hard for a human to classify as well as you can see when looking at the average image of those types. The bias towards classifying pullovers as coats was an interesting feature, because it would make sense for the confusion to go both ways but pullovers may have more than on distinct category one of which overlaps with coat. Some more investigation could be done, but sticking with those categories it may be difficult to untangle that correlation.

Finally there are visualizations of the convolutional neural network (CNN) intermediate layers. The first layer breaks the image into many different pieces and the second layer abstracts them to focus more on the general form. This is of the first CNN architecture and layout. Results may be improved by changing it, but these results are representative of the filtration and abstraction that the neural network does on every image.

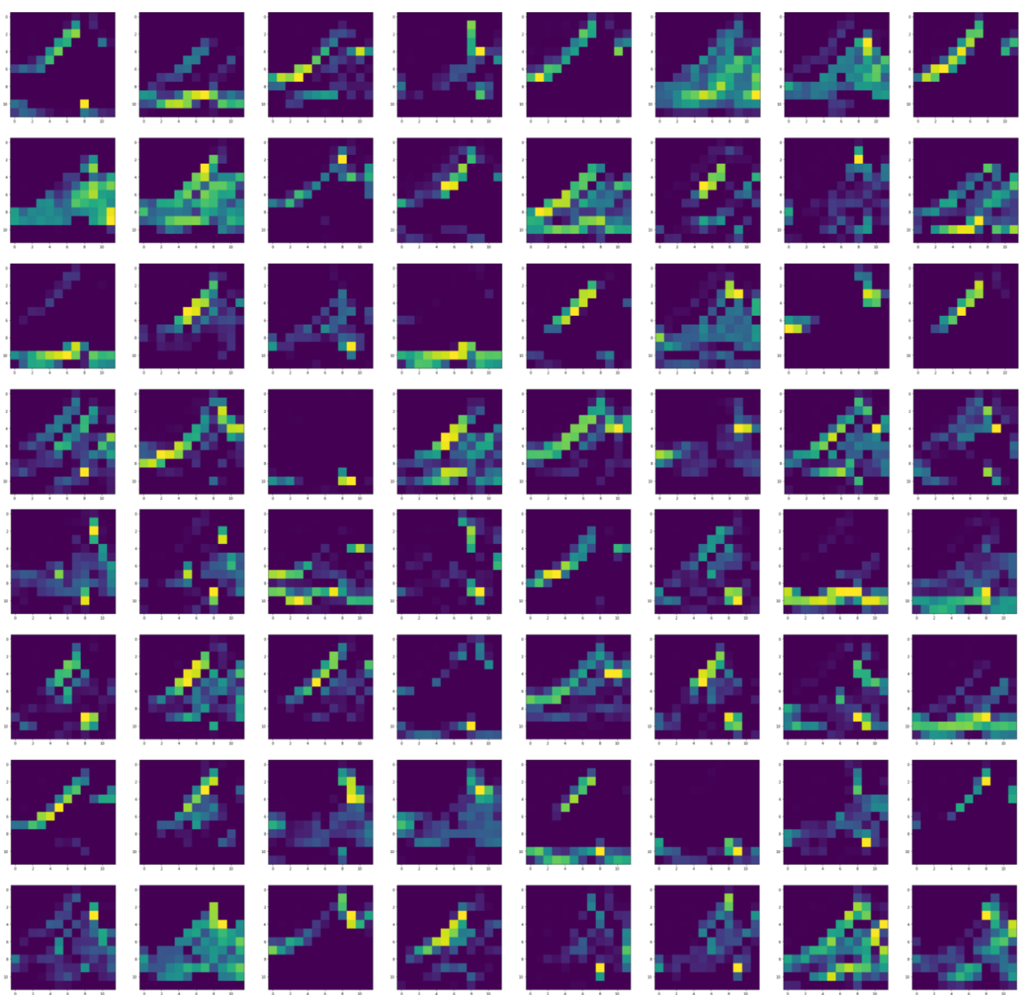
Original image



PHASE 1



PHASE 2



Further tweaking of the neural network is could improve results, but there are many ways that images can be separated and identified through the visual properties show in this document.

The results of Multi-layer Perceptron (MLP), SVC and Logistic regression were also compared with each other across different classification methods. MLP was the most efficient with Normalization as the image preprocessing. The size of hidden layers on the MLP

MLP (size of hidden layers 784-100-100):

- With no preprocessing – 87.65% accuracy
- Normalized(0-1) – 90.37% accuracy
- Standardized(0-1) – 90.09% accuracy

SVC:

- With no preprocessing – 70.67% accuracy
- Normalized(0-1) – 85.57% accuracy
- Standardized(0-1) – 81.87%

Logistic Regression:

- With no preprocessing – supposed to be normalized
- Normalized(0-1) – 84.45% accuracy
- Standardized(0-1) – 85.19% accuracy

CNN:

- Normalized(0-255) – 92.25% accuracy
- Standardized – **92.32% accuracy**

From my tests I learned that Normalization and standardization are roughly equivalent, the initial data is normalized to a range of 0-255. CNN preforms best with an input range of 0-255 and the rest with 0-1. Surprisingly the standardization of the logistic regression model is more effective despite sklearn stating that normalization it important for it. MLP is the most effective of the non-CNN classifier.