Introduction

Our project consisted of a modified version of Google's Landmark Recognition challenge. We trained a VGG16 model on a large dataset to achieve a validation accuracy of above 65%.

Data/Features

We used Google's labelled Landmark training set of 50 categories and 2500 images. We then split these into training, validation, and test sets. All images were reduced to 256x256 size. We used zoom, horizontal flips, and rotations for image augmentation. We trained on these images and the augmented images.

Example data image of the Berlin Cathedral.

Separable Convolutional 2D

The first model was composed of Separable Convolutional 2D layers, max pooling layers, batch normalization layers, dropout layers to prevent overfitting, and finally a softmax predictive layer. Separable convolutions is first a depthwise spatial convolution (acting on each input channel separately) followed by a pointwise convolution which mixes together the resulting output channels.

Residual Neural Network

We used a prebuilt model of a residual neural network from Keras. We initialized the weights on ImageNet, and then added a 128-unit dense layer and a Softmax output layer.

Xception

We used a prebuilt Xception model, and made the last convolutional block trainable.

Model Accuracies (50 classes)

<table>
<thead>
<tr>
<th>Model</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Separable Convolution 2D</td>
<td>0.95</td>
<td>0.73</td>
</tr>
<tr>
<td>VGG 16</td>
<td>0.93</td>
<td>0.99</td>
</tr>
<tr>
<td>ResNet 50</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>Xception</td>
<td>0.96</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Discussion and Future Work

With additional computation resources, the VGG-16 model could be trained with more epochs and a dataset of more classes to gauge how practically applicable our model is. While attempts at training ResNet-50 on 50 or more classes resulted in overfitting, additional computation power could also allow flexibility in tuning hyperparameters. To classify non-landmark images, DELF2 can be used since inliers vary for correctly and incorrectly classified images. DELF can extract features in an image which can be compared to the features of an image of the class that the non-landmark image was classified to. Comparing the number of matched between images of the class and those not in it can create a threshold for whether or not an image classified by our model is actually in the class. This problem requires greater computation power.

Results and Visualizations

The maximum activation image is composed of the image that would maximize the filter output activations of each filter layer in the model.

Blocking out key features of the structure in an implementation of occlusion sensitivity, such as the dome of the Berlin Cathedral, drastically reduced the accuracy of the model, showing that our model is looking for such features. The saliency maps visualize which regions of the image would cause the most change to the output, if they were changed. The class activation maps visualize the attention over the penultimate convolutional layer in respect to the input. Those areas which are most important are represented by a higher “heat map” designation.

Adil Nygaard (adiln@stanford.edu)
Aamnah Khalid (aamnah@stanford.edu)
Uzair Navid Iftikhar (unavid@stanford.edu)