Image Classification: CNN vs Random Forest

Shree Murthy

2024-02-19

Table of Contents

1. Problem Statement

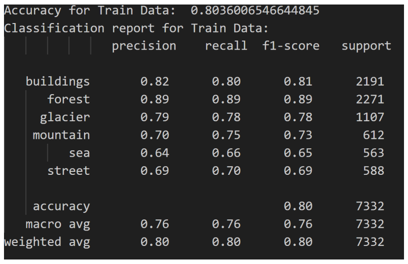
The Intel Image Classification dataset contains thousands of images of natural scenery around the world. The images are divided into 6 categories: buildings, forest, glacier, mountain, sea, and street. The goal of this project is to build a model that can accurately classify these images into their respective categories.

Image classification is an ideal way to learn about CNN-based models and understand how it can outperform traditional machine learning models like Random Forest. To explain, why we need to use a CNN model I will describe how the Random Forest Model performed and why it was not ideal for this task. The rest of this report will delve into how the Random Forest was built and how the data was preprocessed.

The Random Forest model performed poorly on the dataset. The model was unable to learn the complex relationships between the images and the classes.

1.1 Random Forest Results and Analysis

The Random Forest model was trained and tested on the seg_train and seg_test data. The model's performance was as follows:

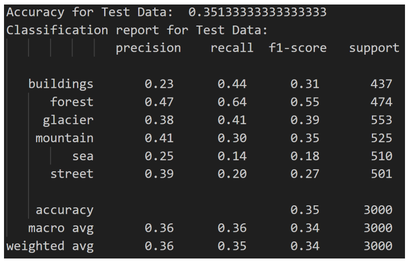


```
Accuracy for Train Data:  0.8036006546644845
Classification report for Train Data:
              precision    recall  f1-score   support

   buildings       0.82      0.80      0.81      2191
      forest       0.89      0.89      0.89      2271
     glacier       0.79      0.78      0.78      1107
    mountain       0.70      0.75      0.73       612
         sea       0.64      0.66      0.65       563
      street       0.69      0.70      0.69       588

    accuracy                           0.80      7332
   macro avg       0.76      0.76      0.76      7332
weighted avg       0.80      0.80      0.80      7332
```

Figure 1: Random Forest - Training Results



```
Accuracy for Test Data:  0.35133333333333333
Classification report for Test Data:
              precision    recall  f1-score   support

   buildings       0.23      0.44      0.31       437
      forest       0.47      0.64      0.55       474
     glacier       0.38      0.41      0.39       553
    mountain       0.41      0.30      0.35       525
         sea       0.25      0.14      0.18       510
      street       0.39      0.20      0.27       501

    accuracy                           0.35      3000
   macro avg       0.36      0.36      0.34      3000
weighted avg       0.36      0.35      0.34      3000
```

Figure 2: Random Forest - Testing Results

Based on these results, the model became extremely overfit. The model's training accuracy was nearly double the testing accuracy (80% and 35% respectively). Furthermore, the model was overfit because the F1-score, precision, and recall scores were all balanced and similar to each other. However, when the testing data was used, those scores were all over the place and heavily biased in favor of the forest class. This imbalance just underscores that the model was trained and was overfit. Running MAE, MSE, R2 on this model doesn't make sense because this is a classification task.

Also, the model not being able to discern when the sea was present is very concerning. The other classes may have confused the model and made it think that light, or dark, blue referred sky and not the sea. Below are some example outcomes of the test predictions:

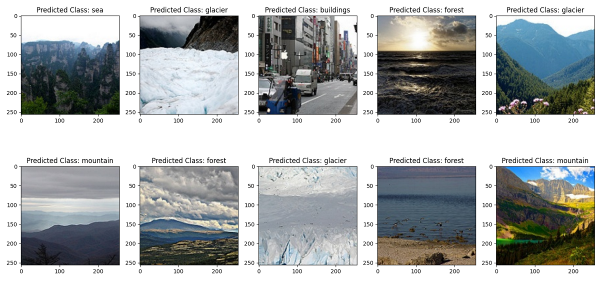


Figure 3: Random Forest - Test Predictions

This picture ties the numerical outcomes together. So many pictures are misclassified as forest. This idea is inline with the previous analysis that the model was biased towards classifying predictions as the forest class. The sea class definitely took the most impact from this bias. For a dataset that is supposed to be natural scenery and have a variety of pictures of different angles and areas, the model should be able to discern what class represents what class. The Random Forest was just not able to do that.

1.2 Overall

Thus, the inability of the Random Forest model to learn relationships and not process the information properly to classify images correctly is why I will be using a CNN model to classify the images. The model should perform and learn the relationships better than the Random Forest model because the pixels aren't flattened and the model reads the images in its totality.

2. Exploratory Data Analysis

The dataset contained roughly 24,000 images of natural scenery. However, I only used ~12,000 images for the scope of this project. I develop code on a virtual server used by other students. When copying over the files, I was unable to copy all the images. To copy the images I used the scp command. Nevertheless, ~12,000 images made for a solid sample set. The specific breakdowns of the train, test, and prediction sets are as follows:

- Train: 7335 images
- Test: 3000 images
- Prediction: 1785 images

After copying the data, I ran an eda.py file to visualize the classes and the images.

The first step was to visualize the train and test class distributions.

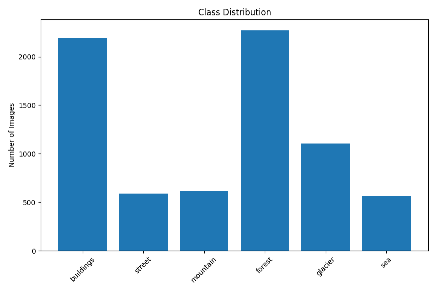


Figure 4: Train Class Distribution

This train class was imbalanced. The street, mountain, and sea classes had significantly fewer images than the other classes. The buildings and forest classes had the most images. This imbalance could lead to biased models. However, I chose not to address the imbalance issues because models need to be trained on a variety of sources. This imbalance could occur in the real world, albeit for a different situation, and the model must handle it and learn relationships.

The test class distribution is as follows:

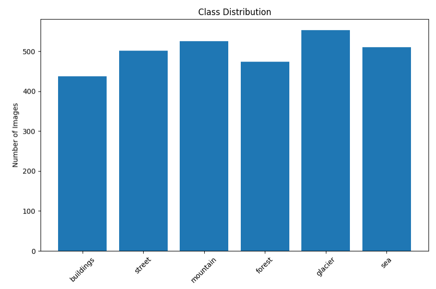


Figure 5: Test Class Distribution

The test class was not as imbalanced as the train class. The classes were more evenly distributed. While some classes had more than others, the difference was not large enough such that the model wouldn't have sufficient data to predict unseen data of a wide variety.

Next, I visualized the images of the train and test sets.

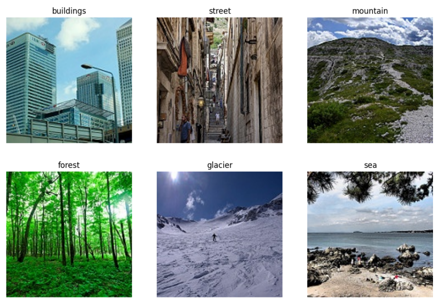


Figure 6: Train Images

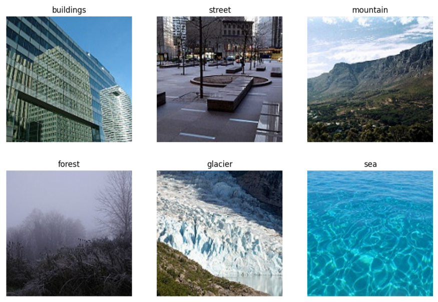


Figure 7: Test Images

The images were of high quality and had similar sizes. Based on these sample images, the model should be able to understand where everything is located and not worry about areas being darker and unrecognizable.

The final step I took was to analyze the image sizes.

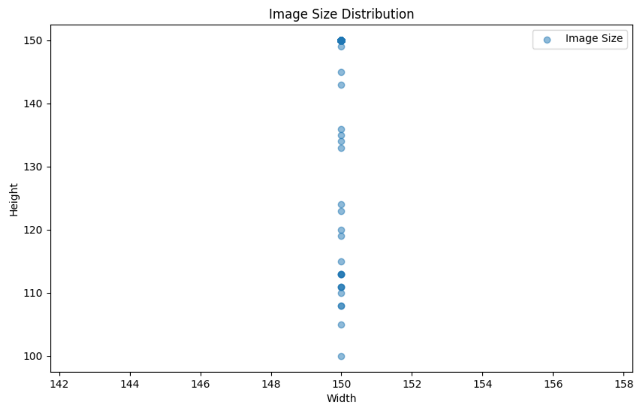


Figure 8: Train Image Sizes

Both, the train and test had the same image size distribution; thus, I am only including the Train distribution. All the images had a width of 150 and height that ranged from 100 to 150. While these are good sizes to train with, I augmented the images to a 256x256 size for two reasons. The first was to ensure all the images are the same size. The second is because when scaling an image to a larger size, the pictures can become more unclear and tougher to analyze. This is done purposefully to ensure that the model is trained on grainer data and can handle unclear images.

Finally, for the prediction data, the data is completely unlabeled and there are no class distributions. The images are of the same size as the train and test images. From this analysis of the prediction data, I will only use this to predict the model and not for other analysis. The goal of this dataset will be to see how the CNN model and Random Forest model perform on unseen, unlabeled data.

Overall, the images are high quality and are a good size. While, there are imbalances within the train data, I will be keeping this imbalance to see how the models will handle it. The dataset is perfect for the scope of this project. If I didn't have issues with the copying, I envision some of the imbalances may have been slightly allieviated. Nevertheless, I'm content with with the dataset. After all the exploratory data analaysis, the next step is to preprocess the data and proceed with the model building.

3. Methods

This section will be split it into two subsections. The first will delve into data preprocessing and the second will delve into model building. This section aims to detail my steps to ensure I'm creating the best model possible and to be consistent between the two models.

3.1 Data Preprocessing

On a high level, both models will resize the image to 256x256; however, there are slight differences within each one due to different operations available within the models.

3.1.1 Random Forest Model

The Random Forest model's data preprocessing will be as follows:

1. Resize the image to 256x256
2. Flatten the image to a single row
3. Store the class label and flattened image into an array
4. Convert into np.array and split into Train and Test sets
5. Pickle the data to reduce randomness, leakage, and better protect the data

The Random Forest model is a simple model that cannot take in multi-dimensional data. Thus, the image must be flattened into a single row. This results in every pixel being treated as an independent feature. While, this may not be ideal for image classification, it is a great way to provide a baseline that can be compared to the CNN model and inform us if the CNN model is worth using for your task.

During the preprocessing the label and image are stored into separate arrays that are then converted to an np.array. This is step enables the model to read the data and understand the class labels. Pickling the data is done to reduce the number of times the images are preprocessed; since we aren't using batches of images to be preprocessed, the preprocessing might take a long time. Pickling the data will ensure that the data is preprocessed once and can be used multiple times.

3.1.2 CNN Model

The CNN model's data preprocessing will be as follows:

1. Use ImageDataGenerator to augment the images (rotation and flip were applied to the train and validation sets)
2. Resize the image to 256x256
3. Batch size is set to 64
4. Images are preprocessed using the preprocess_input function from the VGG16 model

While the above 4 steps showcase the general steps there are more tweaks done within the preprocessing steps. The way the data is provided there are three directories (seg_train, seg_test, and seg_pred). seg_train and seg_test are meant to be used for model training. The seg_pred data is meant to be unlabeled data that is used to test how well the model predicts unseen data (i.e there is no class label; therefore, the we cannot derive specific metrics). I want to ensure that I can gather the model's metrics. Thus, I took the seg_train data and ran a train test split to create a new validation set. Thus, the model will be trained on this new validation set and the remaining seg_train data. This leaves the seg_test set, which is labeled and unseen by the model, to assess the model's performance and gather metrics.

For the CNN model the information isn't pickled because the ImageDataGenerator enables us to batch the images and preprocess them faster.

3.2 Model Building

3.2.1 Random Forest Model

The Random Forest model was built using the scikit library. Random Forests are a simple model that can be used for classification tasks and depend on multiple decision trees to determine the best class given the input data. The model was built using 10 estimators (i.e. 10 decision trees). The model was fitted with the seg_train data and model performance was assessed with the seg_test data. I also used the model to predict the seg_pred data to see how well it will do on those picturesand produce some outputs that will be analyzed in the results section.

3.2.2 CNN Model

The CNN model was built using the VGG16 model. VGG16 is a pre-trained model that was trained on the ImageNet dataset. The model was built using the Keras library. The layers of the model were frozen and I added a few layers to the model. The layers were as follows:

```
# Add custom layers
x = Flatten()(base_model.output)
x = Dense(1024, activation="relu")(x)
x = Dropout(0.5)(x)
x = BatchNormalization()(x)
x = Dense(512, activation="relu")(x)
x = BatchNormalization()(x)
x = Dense(256, activation="relu")(x)
x = BatchNormalization()(x)
output = Dense(num_classes, activation="softmax")(x)
```

The dense layers are activated using the ReLu function and the output layer is activated using the softmax function. Relu is used to ensure that the model can learn complex relationships and the softmax function is used to ensure that the model can output a probability distribution. The softmax activation is ideal for multi-class classification tasks because the probability distribution can be used to determine the best class (highest probability = best class).

There are 3 dense layers to ensure the model can learn about the complex relationships in the data. The dropout layer is added to prevent overfitting and ensure the model doesn't learn weird features that are specific to the training set. The batch normalization layers are added to ensure the outputs are normalized and smoothed out. This better improves the relationships learned and reduce overfitting.

The model was compiled using the Adam optimizer and the categorical crossentropy loss function. The model was trained using the new train and validation data that was derived from the seg_test dataset. The model was trained for 10 epochs and the batch size was set to 64. Early Stopping was also used to prevent overfitting. The patience was set to 3 (i.e if the model's validation loss did not improve for 3 epochs, the model would stop training). The model's performance was assessed using the seg_test data. The model was also used to predict the seg_pred data to see how well it will do on those pictures and produce some outputs that will be analyzed in the results section.

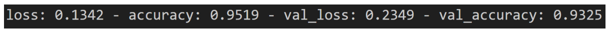
Once the model is done training, the model is saved to be used for future predictions.

4. Results

The Random Forest results were already discussed in the Problem Statement section. This section will focus soley on the CNN model.

4.1 CNN Model

The CNN model was trained and tested on the train test split conducted on the seg_train data. The model's performance was much better than the Random Forest model. The model's performance was as follows:



```
loss: 0.1342 - accuracy: 0.9519 - val_loss: 0.2349 - val_accuracy: 0.9325
```

Figure 9: CNN Model - Training Results

This model performed much better than the Random Forest model. The model's training and testing accuracy and loss were very close together. This is an extremely good sign because the model isn't too overfit. The performance of the training model bodes well for unseen data. The second graph of the loss shows that the model's val_loss is greater than the train_loss by the time the epochs are completed. While this indicates some overfitting it is not as severe as the Random Forest Model. Also, both loss values are within 0.15 of each other which isn't a large difference. Coupled with strong accuracy scores for the two sets, the model is deemed to be trained well with minimal overfitting.

For the unseen data the model performed as follows:

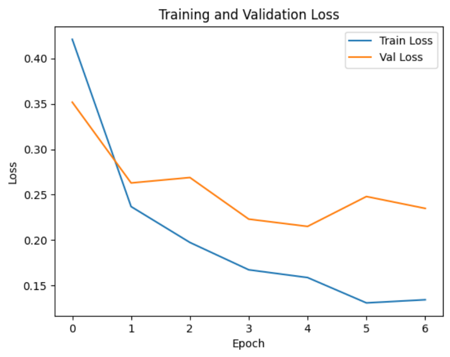
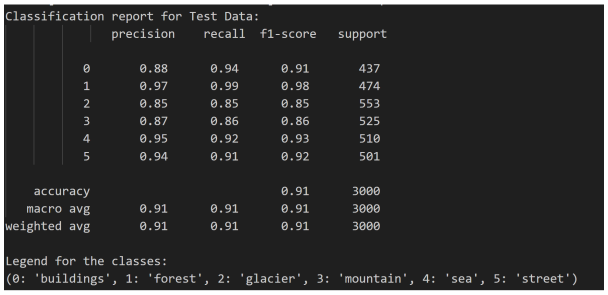


Figure 10: CNN Model - Training vs Validation Loss



```
Classification report for Test Data:
              precision    recall  f1-score   support

           0       0.88      0.94      0.91       437
           1       0.97      0.99      0.98       474
           2       0.85      0.85      0.85       553
           3       0.87      0.86      0.86       525
           4       0.95      0.92      0.93       510
           5       0.94      0.91      0.92       501

    accuracy                           0.91      3000
   macro avg       0.91      0.91      0.91      3000
weighted avg       0.91      0.91      0.91      3000

Legend for the classes:
(0: 'buildings', 1: 'forest', 2: 'glacier', 3: 'mountain', 4: 'sea', 5: 'street')
```

Figure 11: CNN Model - Testing Results

The classification report showcases a consistent prediction trend across all classes. The F1-score, precision, and recall scores are balanced and extremely high. This indicates that the model is confidently predicting unseen data correctly and effectively. The model falters slightly when classifying glaciers and mountains. This does make sense becasuse glaciers and mountains can look similar and glaciers can be found on mountains. Nevertheless, the model classified these classes with 85%+ accuracy which is still strong and better than the testing accruacy of the Random Forest model.

The confusion matrix showcases the model's ability to classify the classes effectively. The matrix's main diagonal is dark and the off-diagonal values are light. Thus, the model is apt at classification properly. While the confusion matrix doesn't provide much new information, one point that stood out was that model sometimes misclassifies the following classes as each other:

- Buildings and Sea
- Mountain and Forest
- Mountain and Glacier

This could be attributed to the images containing similar features. For example, a picture of sea could contain a building and a picture of a mountain could contain a glacier. While it is

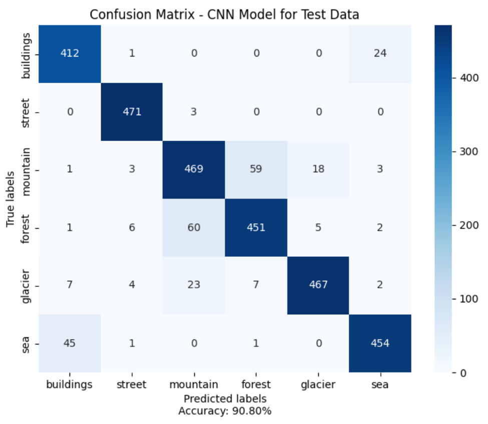


Figure 12: CNN Model - Confusion Matrix for Test Data

classifying something present in the picture, it ended up not matching the label. Nevertheless, the model is still able to classify these classes effectively and the misclassifications are not too severe.

Based on the model's performance I expected it to classify unseen, unlabeled data effectively. Thus, below are 10 images that the model predicted:

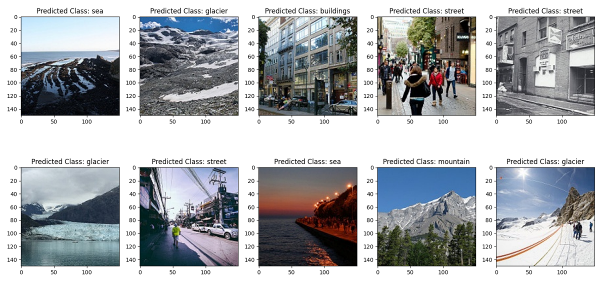


Figure 13: CNN Model - Test Predictions

This set of images showcases the model's ability to classify unseen data really well. Every picture looks like it was classified correctly. Since there are no ground truth labels, I cannot provide concrete accuracy values. Based on the eye test, the model classified the images correctly based on the information present.

4.2 Overall

Overall, the CNN model worked extremely well on the training, validation and test data as well as the unlabeled/unseen data. The model is definitely better than the Random Forest model. The model was able to learn the complex relationships between the images and the classes and classify the images effectively. The model's performance on the unseen data is a good sign that the model can be used to classify images in the future.

5. Discussion

In conclusion, the two models preformed drastically different. The random forest model was extremly overfit and biased to certain classes. The Random Forest model performed poorly on the dataset, whereas the CNN model performed very well. With respect to the problem, the stats of the Random Forest isn't something we want to pursue long-term. It makes no sense to use a model that is overfit and biased. Models need to strive for balance when trying to conduct image classification or any other task. The goal of this project was to use a CNN model to classify images and prove why it was important, the stats analyzed in the report underscore this goal.

For hyperparamter tuning, I experimented with the following:

```
- Number of estimators
- Epochs (and Early Stopping Patience)
- Dropout
- Dense layer nodes
- Batch size
- Image size
```

For the Random Forest, I initially used 100 estimators and then reduced it to 10. When it was 100, the model reach ~99% train accuracy and ~42% test accuracy. This was not a large enough jump from the aforementioned results to justify the time and computational resources required to run the model with 100 estimators. Reducing the estimators to 10 was a good sweet spot to ensure that the model learned enough about the images and minimzed the time/complexity of the model.

For the CNN Epochs, I started with 100 epochs with no early stopping. The result of this was that the train and val loss diverged very quickly and the model was too overfit. Essentially, the model was just losing information after a certain point. During this iterations, I noticed that the model performed well until 10 epochs before it diverged. The accuracy also dropped off for the validation data after the 10th epoch. With this in mind I dropped the epochs to 10. This was a good start; however, I noticed that the loss still diverged a fair amount around the 9th/10th epoch. Thus, I added early stopping with a patience of 3. While I could increase this, it doesn't make sense when working with 10 epochs to have a patience of > 3. The higher the patience the better off just running the model flat out with 10 epochs and no early stopping. Adding the early stopping ensured the model didn't overfit and the accuracy and loss didn't diverge quickly in the wrong directions.

The dropout and dense layer nodes were experimented with to ensure that the model didn't overfit and learned the complex relationships in the data. The dropout was initially set to 0.2 and one dense layer set to 512. The model still performed well however the accuracy and loss weren't too close to each other even though they were high values associated with the train and val's respective accuracy and loss. I increased the dropout to 0.5 and added dense layer nodes starting from 1024 and decreased to 512 and 256. This was a good balance to ensure the model didn't overfit and learned the complex relationships in the data. Adding and modifying the extra layers impacted the model's generalizations.

For the size of my dataset, a batch size of 64 provided a good balance of images per batch and reduced the computational resources required. I could've stuck with the default of 32; however, with 12,000 images, the model would've taken a long time to train. The batch size of 64 was a good balance to ensure the model was trained effectively and efficiently. The image size was increased to 256x256. This was done to ensure that the model was trained on grainer data and could handle unclear images. Also, it ensured all images were the same height and width to avoid biases.

5.1 Future Work

In the future I would want to experiment with creating my own CNN model from scratch without the transfer learning to see how well it would perform. It would reduce some computational resources because I wouldn't be loading in the large VGG16 that was trained on a wide variety of images and tasks. My own model would be catered to my task. It would be an interesting comparison to see if it was worth using the VGG16 model or if I could've just used my own model.

Also, I would want to experiment with using an SVM, KNN, or basic Decision Trees to see if another, simpler, model could do better than the Random Forest. The Random Forest was a good baseline to compare the CNN model to; however, it would be interesting to see if another model could do better.

Finally, I would want to experiment with the data augmentation and see if I could've done more to ensure the model was trained on a wide variety of images. I only tested with the flip and rotation operations. It would be intersting to see what other opertaions I could do and how that impacts the training process. I believe the augmentation plays a key role in the model's training, so it would be intersting to see if adding more improves or hinders the model's effectiveness.

6. Works Cited

- Confusion Matrix Visualization
- Image Augmentation and ImageDataGenerator

- EDA Ideas for Image Classification