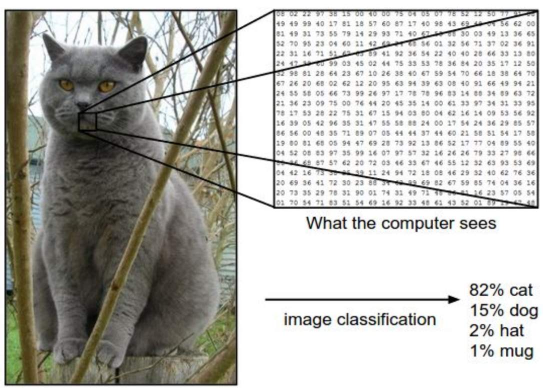
MultiLabel Image Classification
Final Project - ECE 285
Group-4

Ayush Jasuja, Manu Seth, Ranti Dev Sharma

Introduction

Image classification is a core computer vision task where given an image we design algorithms to assign one or more labels to images. There are both supervised and unsupervised algorithm to get good results but lately deep convolutional neural nets using supervised learning have gained edge over other methods as far as accuracy is concerned. Training these networks on the other hand require some good hardware support like GPUs and multiple hours of training.



As part of our final project we worked on an online multi-label image classification challenge organized by Planet on Kaggle. Kaggle is an online platform where companies post machine learning competitions for participants from all over the globe. The main idea of the competition is to understand satellite images of Amazon basin. Understand those images will help humans figure out the rate of deforestation, human encroachment and how different activities are affecting wildlife and biodiversity. Machine learning makes this process faster, allowing humans to respond quickly and adapt accordingly for a better future.
Our implementations mainly consisted of applying some state of the art deep convolutional neural networks like VGG16 and ResNet50. We also implemented self designed neural nets and got good accuracy with them. We used Google Cloud to train these networks on virtual machines with 4GB of gpu processing power. The details about data, methodology and network architecture, results and conclusions are described in the below sections.

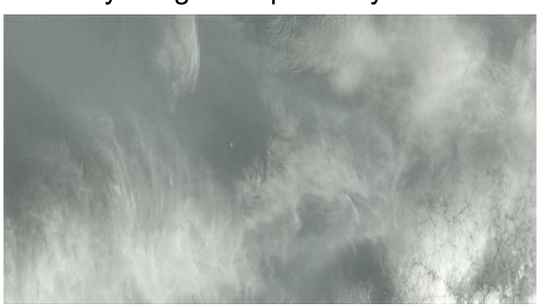
Data

As part of the competition we were provided a dataset of forty thousand processed jpg images.

These images are 256x256 rgb images. Every image was assigned one or more labels out of the seventeen overall labels. These labels could be primarily classified into three divisions - atmospheric conditions, common land cover/land use phenomena, and rare land cover/land use phenomena. All the labels were manually assigned by competition designers and as per their website it may contain some noisy(incorrect) labels.

We give some example and short description of multiple labels from the competition website:

1. **Cloud Cover Labels :** These labels classify the weather condition in the sky. Images are classified as clear, partly cloudy, cloudy, and haze. Cloudy images obscure the land view by ninety percent or more. Partly cloudy images only let us see patches of land cover. Clear images show no signs of cloud whatsoever and haze can be thought of as a blurry version of clear images. Following images show examples of partly cloudy, cloudy and hazy images respectively.



2. **Common Labels :** These are the frequent labels in the dataset. It contains labels like primary - amazon rainforest, water - rivers/lakes, habitation - human cities or homes, agriculture - commercial agriculture land, roads - for transportation, cultivation - shifting

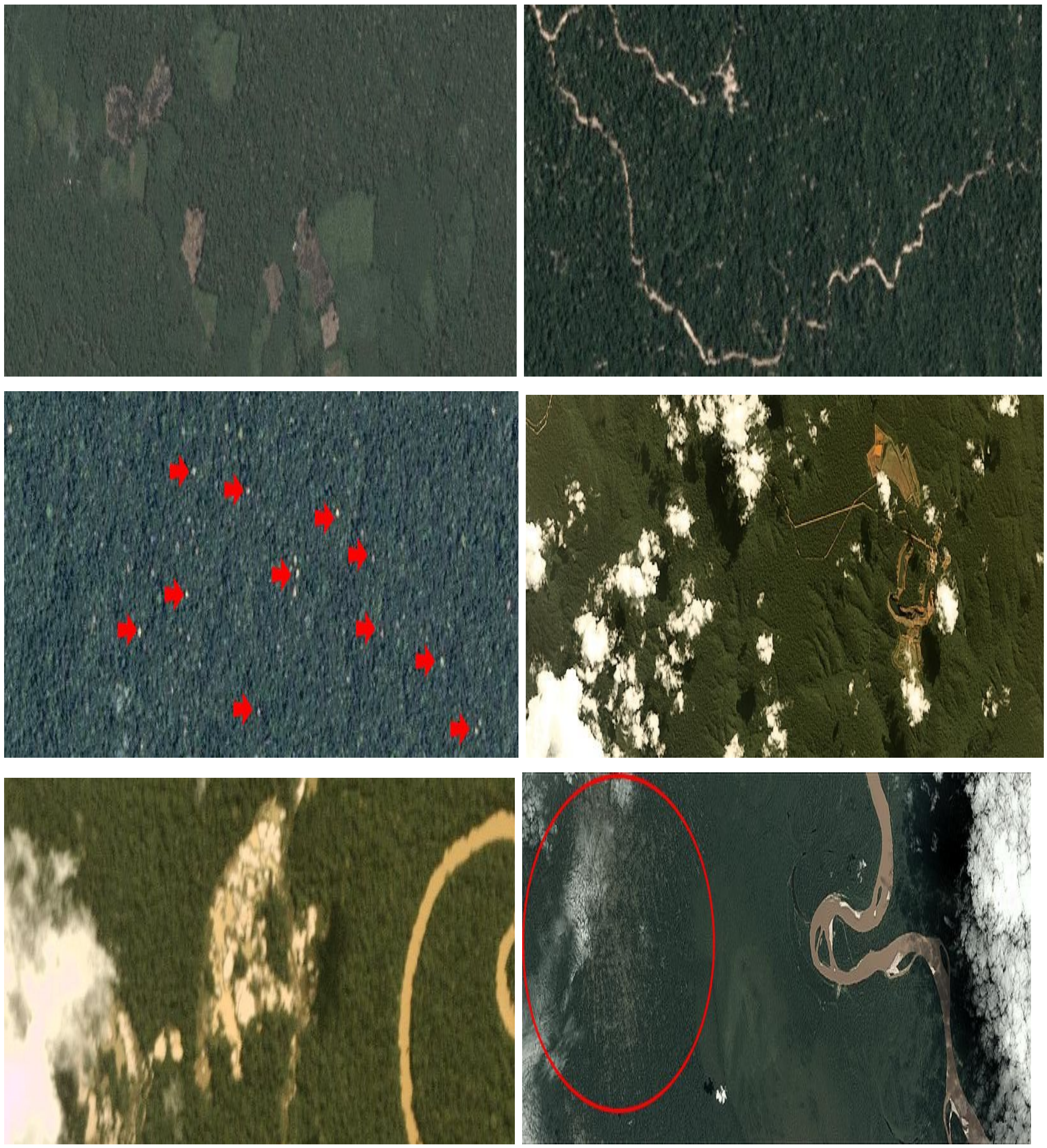
cultivation practiced by rural communities and bare ground - naturally occurring tree free areas. Following images show examples of these images in order:





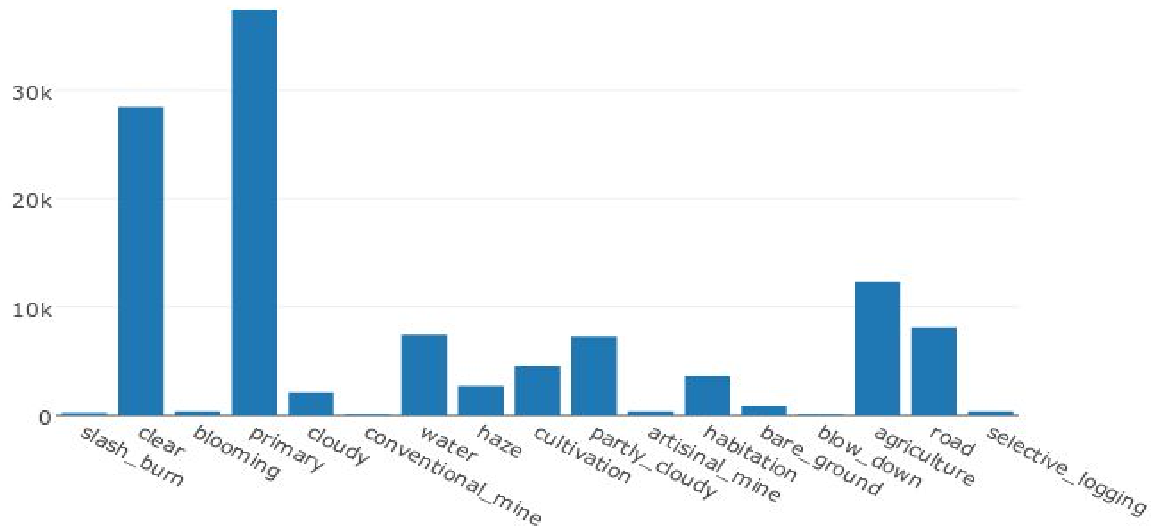


3. **Less Frequent Labels :** These labels are present in less than one percent of the data and are outliers in the sense that the machine learning architectures fail to predict them correctly given their frequency as compared to more frequent labels. It contain labels like Slash (\m/) and burn**,** Selective Logging, Blooming, Conventional Mining, Artisinal Mining - small scale mining and Blow Down.

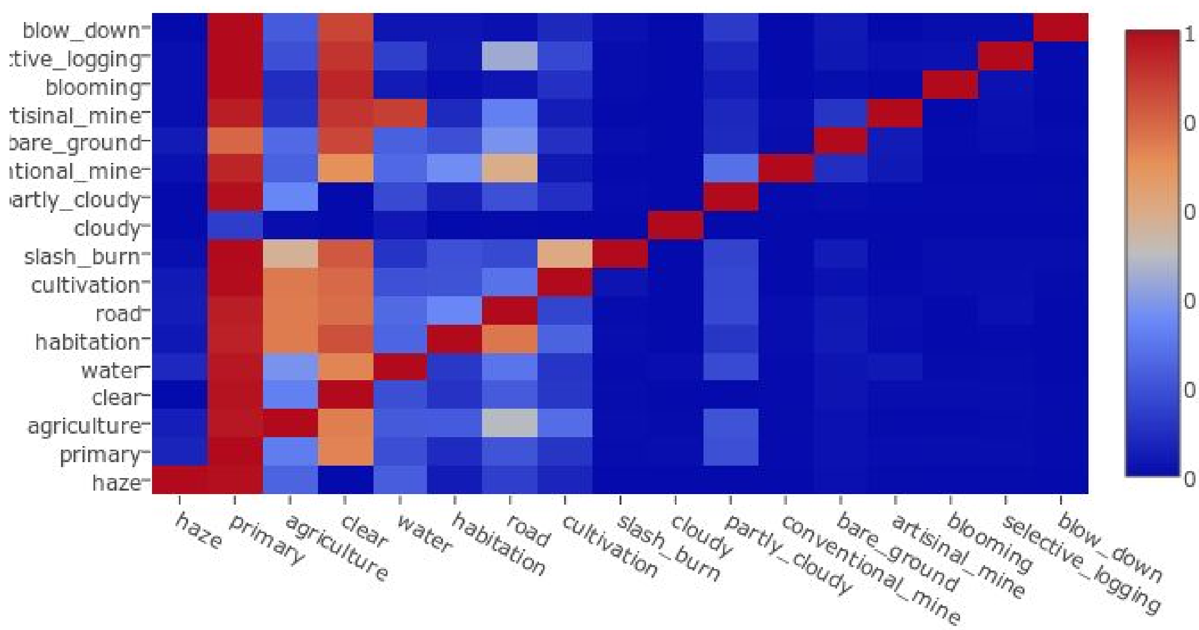


Label Statistics

Below is a histogram to show the frequency of multiple labels. We can see that the labels are not uniformly present across all forty thousand images.

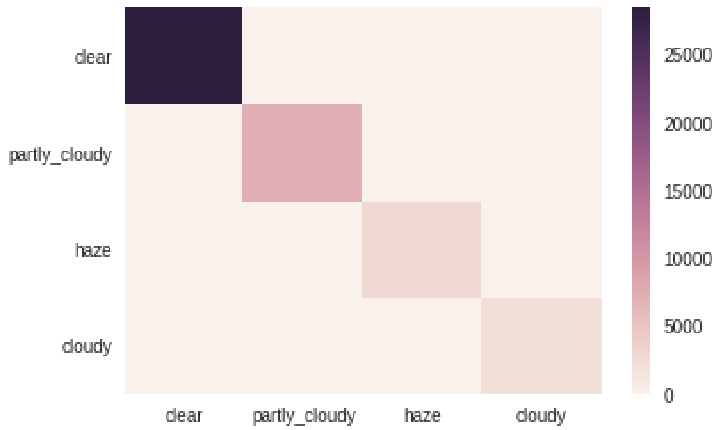


We also analysed the co-occurrence matrix of labels together. Below is an heatmap for the same.

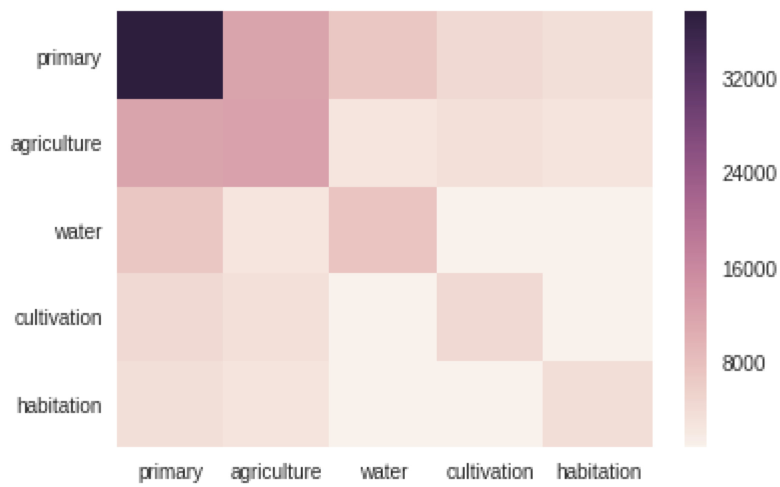


We can see that cloudy images do not overlap with any other labels (and rarely with primary). Further, analysing the subcategories of labels revealed more interesting patterns:

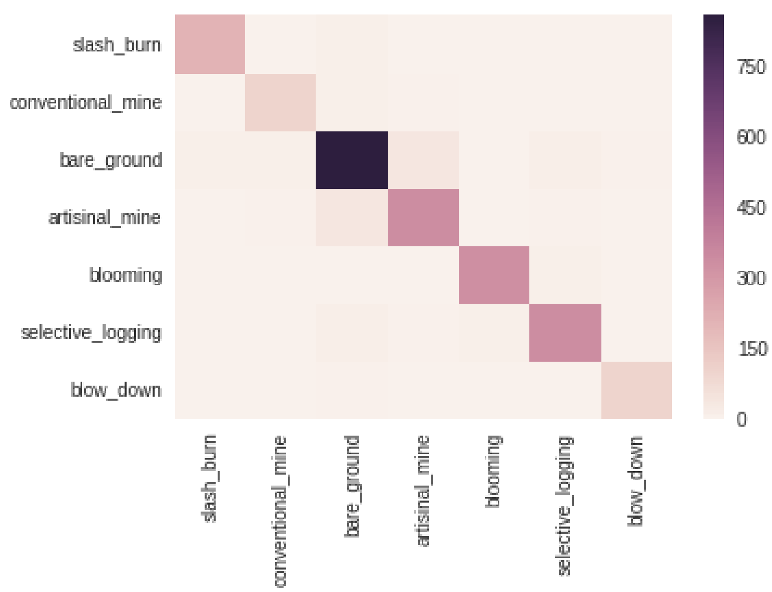
1. No two weather labels occur together.



2. Land labels overlap quite a lot (especially primary and agriculture)

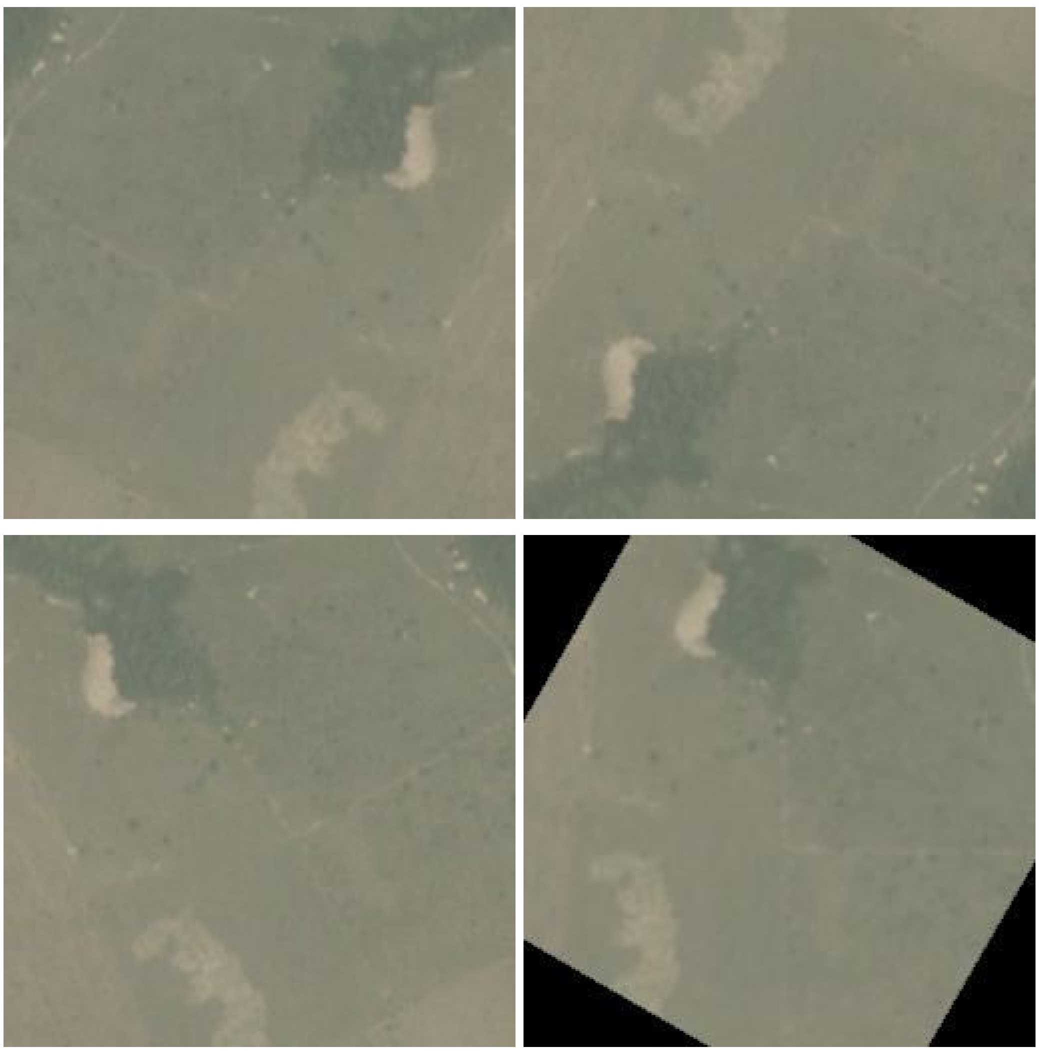


3. Rare labels don’t overlap much.



Data Augmentation

As seen in the statistics above, not all labels are present with uniform frequency. Some labels like primary occur in almost ninety percent of the images whereas others like blooming not even occur in one percent of the data. Therefore to make the learning more robust we decided to use image transformations on les frequent labels and used a tool imgaug. We applied transformations such as rotation and flipping across horizontal and vertical axis. Note that we only selected a subset of data which do not contain the most frequent labels to apply these operations because otherwise the data skew would have still remained given that labels like primary and clear are present in almost most of the images. Below is an example of series of transformations applied : flip horizontal, vertical and rotate by thirty degrees. After applying these transformations the dataset size increased by twenty percent, which is forty eight thousand images.



Methods

To test the model convergence we divided the dataset into two groups - training and validation sets. We reduced the image dimensions to 128x128 because of GPU size constraints. The criteria for measuring the performance of the network is F2-score. It is defined as follows for this particular competition:

p : precision, *r :* recall

$$(1+\beta^2)\frac{pr}{\beta^2 p + r} \text{ where } p = \frac{tp}{tp+fp},\ r = \frac{tp}{tp+fn},\ \beta = 2.$$

We used following models:

VGG16: This is a pretrained CNN provided by keras and it gave state of art results on imagenet classification challenge. It has 16 layers and about 60 millions training parameters. This network converged after 61 epochs with a f2-score of 0.70 on the validation set. On the test data it gives a precision score of about 0.84.

ResNet50: ResNet is the CNN created by Microsoft team which ILSRVC 2015 competition on image classification and performs better than humans on ImageNet dataset. The pre-trained weights of this network are also provided by keras and we learn the parameters on top of them. This network's performance is not as good as that of VGG16 probably because its has 50 layers and the backpropagation almost fails to do proper weight updates of the initial layers. Using high resolution images like 256x256 could give better results. The f2-precision score on the validation set was 0.63 which is lower than that of VGG16 and hence we didn't run it on the test set.

Simple CNN:

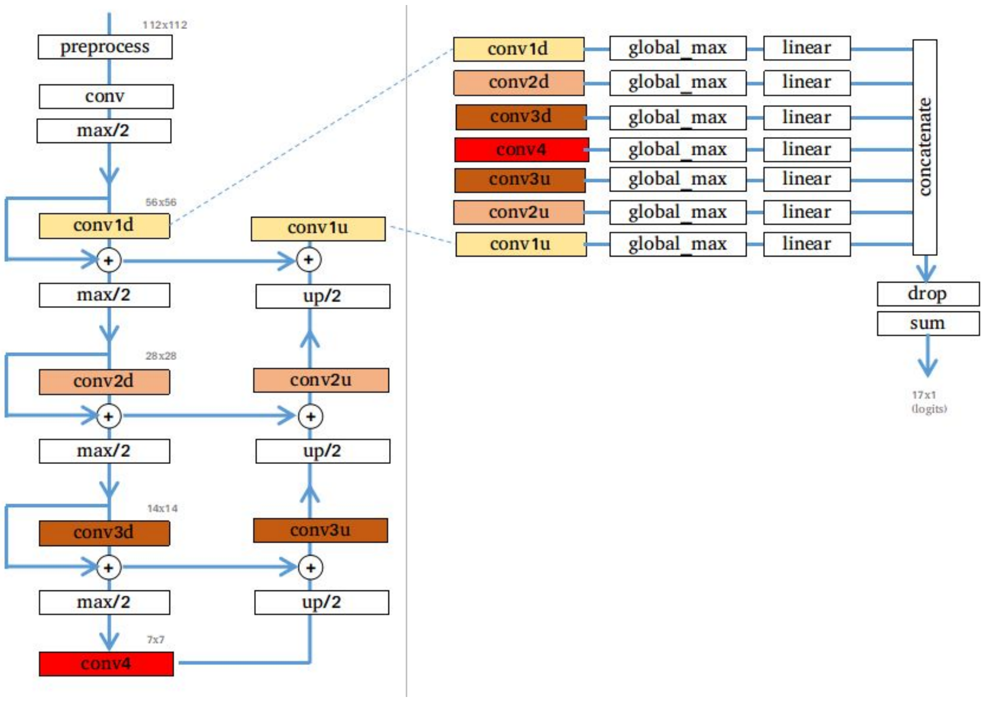
This is a small Neural Net we created to start our implementation. It contains just eight layers and takes just 4 epochs to train to converge. It gives a f2-precision score of 0.72 on the training set but performs miserably on the test set with a precision score of just 0.30. We added drop out layers in the network to avoid overfitting but the performance didn't improve much. Below is the architecture details(keras implementation) of the network.

```
model = Sequential()
model.add(Conv2D(32, kernel_size=(3, 3), activation='relu', input_shape=(256, 256, 3)))
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(17, activation='sigmoid'))

G4-Network

We also tried a network from the competition discussion section with the below architecture. Training on this network converged after 100 epochs with a prediction score of .74 on the validation set and gave our best score on the test set of 0.86.



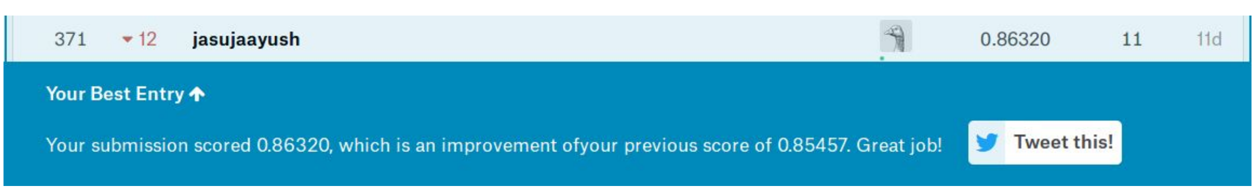
Post Processing

We did following post processing on results before submitting the results on kaggle:

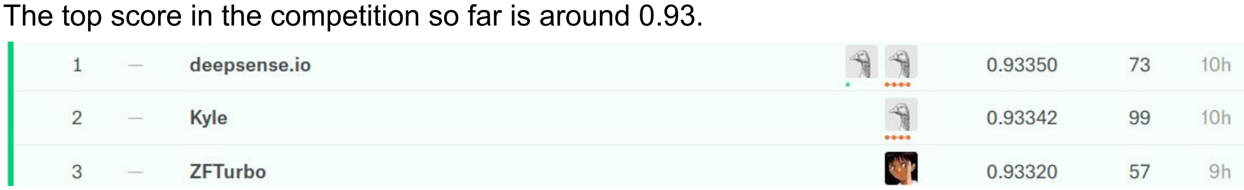
1. If multiple weather labels were predicted then we take the one with most probability because co-occurrence of different weather labels is zero.
2. If after the above step, weath label predicted was ‘cloudy, then we removed the other labels based on data analysis.

Results

We got F2-precision score of 0.863 based on G4-model. Below is the picture from the leaderboard of our score.



The top score in the competition so far is around 0.93.



References

[1] Planet - https://www.planet.com/company/careers/
[2] Cloud - http://cloud.google.com/
[3] CNN - https://en.wikipedia.org/wiki/Convolutional_neural_network
[4] Image augmentation - https://codebox.net/pages/image-augmentation-with-python
[5] Keras - https://keras.io/
[6] VGG16- https://keras.io/applications/#vgg16
[7] ResNet50 - https://keras.io/applications/#resnet50
[8] F1-score - https://en.wikipedia.org/wiki/F1_score