Understanding the Amazon Rainforest with Computer Vision.

Jessica Castillo
Universidad de los Andes
Bogotá D.C.
jl.castillo@uniandes.edu.co

Juan David García
Universidad de los Andes
Bogotá D.C.
jd.garcia20@uniandes.edu.co

Juan Francisco Suescún
Universidad de los Andes
Bogotá D.C.
jf.suescun10@uniandes.edu.co

Abstract

As home to 10% of the world species and being the worlds richest and most bio diverse biological reservoir, the Amazon rain forest should be one of the central efforts for natural conservation. Unfortunately, its massive land span makes small scale monitoring a difficult task. Multi label classification of satellite images from the 'Understanding the Amazon from Space' Kaggle competition was performed with a modified Inception V3 convolutional neural network. Experiments with backbone architecture, near infrared channels and alternative image representations were carried out resulting in a final score of 0.906 classification accuracy using the F_2 metric. This results opens the possibility of using this same data set for finer grain problems by providing an automated, high accuracy method of satellite image annotation for scientific, government and organizational use to provide near real time data about the state and future of the most bio diverse ecosystem in the planet.

1. Introduction

The Amazon rainforest constitutes the worlds richest and most biodiverse biological reservoir. Stretching for 6.7 million squared kilometers it is home to more than three million species of plants and animals, about 10% of the worlds species, and about one million human indigenous people. The vastness of the rainforest suggests that many of the plant and animal species that live in it are yet to be discovered or classified, making it a part of one of the last untouched places on earth. With 20% of all water in the planet passing through the Amazon, millions of animal and plant species, an enourmous capacity for stabilizing carbon emissions among many things, the Amazon rainforest should be one of the central efforts for natural conservation if humankind wishes to remain on the planet, as well as preventing a 6th massive extintion [1, 2].

Unfortunately during the past few years, human intervention in the Amazon has increased dramatically. Statistics from the National Institute for Space Research show an increase of 84% in deforestation and 30% in forest fires during the last year. Among the causes for the increase in the fires is the growth in deforestation for expansion of the agricultural frontier, legal and illegal mining projects, increase in the population size of areas that neighbour the rainforest, among others. In particular, one of the biggest challenges that arises in the monitoring of the Amazon is its sheer size and difficult access [3]. The massive span of the Amazon makes monitoring it a significant challenge but it is one that must be taken seriously to preserve the now fragile ecosystems, the millions of species that inhabit it and the varied universal cosmologies of the cultures that settle in it.

One of the possible ways to monitor such extensive areas is via satellite imagery. With new developments in satellite image resolution, the possibility of monitoring small activity in the Amazon has arisen, opening the doors to keep track of small scale logging and mining operations, expansion of agricultural and urban frontiers and monitoring forest fires. This technology has the potential of providing insights into where deforestation is happening, as well as, its causes and new solutions to tackle this problem. For this, Planet and SCOON developed a challenge that consists of correctly classifying multiclass imagery taken from space with the use of computer vision. Although the challenge itself is approximately 3 years old, the relevance and importance of this type of software development has only increased in this time. Appropriate multilabel classification

of satellite imagery could be the first step in a series of algorithms that can accurately monitor the changes affecting this biome.

2. Related work

There is a current interest in processing space imagery around the world as it constitutes one of the largest man made image datasets available. It has become of vital importance in areas such as conservation as it provides large scale surveillance possibilities. In the year 2006, Baker et al. [4] developed a method for classifying land coverage images which uses Gradient Boost Trees and SVM where they obtained 85% of accuracy in the CTA test dataset. There is a disadvantage about this method and it is because the model lack the ability for computers to learn features automatically [4].

By the decade of 2010s, deep learning methods were used by different approaches in order to classify and segment satellite images for many different type of problems. In [5], they used convolutional neuronal networks and a per-pixel classification methodology to categorize satellite images into groups of road, ground, water and vegetation. This method is similar as the one proposed by [6] but the core of this work is based on using street view data combined with satellite views to count the number of trees in a given area [6].

An exploration of the existing work done in Amazon Satellite Image Classification demonstrated that neural networks are still the go-to method to solve this problem. The architectures and methods between references vary substantially, with some authors using ResNet-50 architecture, VGG-16 networks and Inception-v3. Most of the efforts made by the authors consisted in fine-tuning existing architectures with variations in the activation functions and final linear classifiers. The most successful CNN for the Kaggle competition [7] reported an F_2 measure of 0.933 [8], which will help as a reference as to the minimal expected output of our own method.

In [9], researchers implement three different type of neuronal networks such as VGG, ResNet and DenseNet. The VGG model allows more non-linearities and the filters are quite smaller but more effective. In ResNet due to the residual connections, the model can train deeper without degrading the features which results in a higher accuracy compared to other architectures. Finally, DenseNet is made up of blocks that are connected to every other layers in the architecture which alleviates the problem of vanishing gradients [9].

Chen et.al [10] developed in their work a hybrid neural network to perform small object detection. In this method, the authors separates the feature maps of the last layer and the max-pooling layer to extract features at different scales. Once those features are obtained, they are used to train the network and predict the class for each one of the images. In [11] the authors proposed a fractional type convolution filtering. This technique shows an improvement in satellite image classification while performing filtering algorithms as a pre-classification stage of the method.

Finally, it is important to note that [9] used Data Augmentation techniques such as reflection, rotation and random cuts as well as corrections in image contrast as methods to improve the F_2 measures without directly intervening in the CNN itself.

3. Approach

3.1. Database

The dataset provided for the project consists of 40,479 training images and 61,191 test images. The 4 bands of the tif chips correspond to RGB + nIR , where the nIR is an extra infrared channel that could prove beneficial as an extra layer of information during the training of the model. According to the competition data explanation, each of the images correspond to a 947.2m x 947.2 m area of the rainforest which may contain one or many of the desired category labels.

The training data is accompanied by annotations for each image. Given as the problem constitutes a multilabel classification problem with 17 possible classes for each image. Each image can belong to a single class or multiple classes corresponding to the phenomena of interest happening in the Amazon rain forest. Example images and their corresponding labels taken from the dataset can be observed in the example images shown below.

Figure 1. Example images and their corresponding multicategory labels.

Apart from the categories presented above, there are some images where it is difficult to visualize the ground because of meteorological conditions, including cloud labels that can be cloudy, partly cloudy or hazy. An example of this can be seen on the first and third images on

the top row from left to right. Other noteworthy phenomena from the data set can be visualized in this figure, such as the difficulty in distinguishing 'roads + primary rain forest' from 'water + primary rain forest', the small spatial scale of features that correspond to tags such as 'selective logging' and finally, the prevalence of the 'primary' tag in the data set.

According do the data explanation, the most common categories are primary rain forest, agriculture, rivers, towns/cities, and roads. Less common categories include slash and burn, artesanal mining, conventional mining, logging and blooming.

3.1.1 Evaluation Metrics

In order to directly compare our results with the results obtained by other teams in the competition, we will have our evaluation metric be the F_2 score which is defined in terms of precision (P_i) and recall (R_i). With N test samples, X_i and $\breve{X}_i$ being the true label and the prediction respectively we obtain the F_2 score in the following way:

$$P_i = \frac{|X_i \bigcap \breve{X}_i|}{|\breve{X}_i|}$$

$$R_i = \frac{|X_i \bigcap \breve{X}_i|}{|X_i|}$$

$$F_2 = \frac{5}{N} \sum_{i=1}^{N} \frac{P_i R_i}{4P_i + Ri}$$

Or in a more compact way, this formula is equivalent to:

$$F_2 = \frac{(1+\beta^2)(P \cdot R)}{\beta^2(P+R)}$$

The use of the F_2 score is standard for this specific competition as it adapts better to a multi-label classification problem, giving more importance to the Recall of the algorithm than to its Precision. All the experiments carried out were evaluated in the official server from the competition and presented as late submissions. However, the train database was divided in order to have a subset for the F_2 scores of validation.

3.2. Final Layer Adjustment

The pre-trained models used in the research were all designed for single label image classification. To adjust these to our multi-label problem, the final classification layer was modified to a layer containing 17 sigmoid classifiers, one for each class, turning each category into a binary cross entropy activation problem. The activation of each sigmoid classifier beyond its threshold value would confer the image with the corresponding category label, thus providing a multi-category image output.

In order to determine the appropriate threshold, the validation was performed with different thresholds values from 0 to 1, to determine which was the one that maximized the results of the F_2-score in the validation database. Thus, the threshold found was 0.23, as shown in the figure 2.

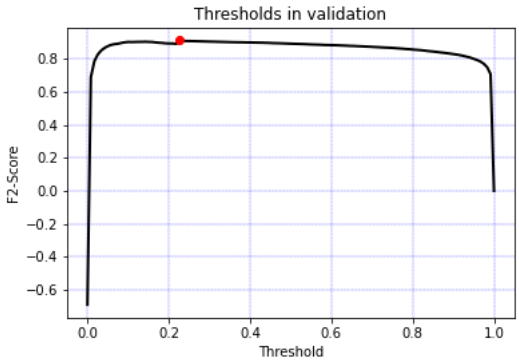

Figure 2. F_2-scores for different threshold values in the validation dataset.

3.3. Infrared Channel

To make use of the full data set information provided for the challenge, experiments were conducted by adding the infrared channel as a fourth input channel to the convolutional neuronal network. The use of this channel also allowed us to calculate different auxiliary index channels that could provide useful in the overall training of the network. We calculated the NDVI (Normalized Difference Vegetation Index) as well as the NDWI (Normalized Difference Water Index) for each training image and used them separately as an additional fourth channel to train the network and evaluate their individual impact on the overall network performance.

NDVI and NDWI channels are calculated as following:

$$NDVI = \frac{NIR - R}{NIR + R}$$

$$NDWI = \frac{G - NIR}{G + NIR}$$

NDVI and NDWI are both used in precision farming and biomass estimation with satellite imagery. As can be seen in 4, the calculated NDVI index channel and NDWI channel provide different information to both RGB and NIR channels indicating a possible benefit of its use in network training.

Apart from these two single channels, two other three-channel representations of the infrared channel were calculated using combinations of the NIR channel and the red,

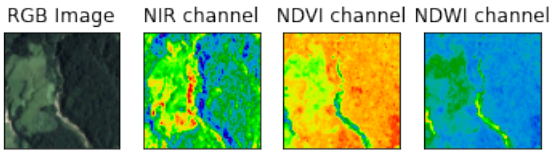

Figure 3. Training image and its corresponding NIR channel, NDVI channel and NDWI channel.

green and blue channels from the RGB image, as an attempt to change image representation to improve final results.

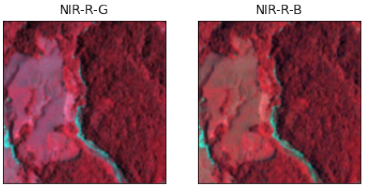

Figure 4. NIR-R-G and NIR-R-B three-channel combinations to create extended representation of the training images.

The effect of the different infrared representations (RGB, RGB + NIR, RGB + NDVI, NIR-R-G, NIR-R-B) were evaluated on the ResNet18 architecture to establish their effect on the overall result of the network.

3.4. Architecture

To begin determining the best network architecture for this problem, we created a very simple architecture called AmazonSimpleNet, shown in figure 5, to establish a baseline. Different base architectures were used with pre-trained weights based on the ImageNet dataset. Once the architecture was adjusted to fulfill the requirements of a multi-label classification problem as explained in the section above, each neural network was run and the corresponding F_2-score was computed as the performance metric in the official competition server. Overall, nine different architectures were experimented with varying the base model (Resnet18, Resnet101,InceptionV3) as well as the channels that were provided as input (RGB, RGB + NIR, RGB + NDVI, NIR-R-G, NIR-R-B). Finally, for the most promising results an extended training was run increasing the amount of training epochs.

4. Results

4.1. Baseline

Our baseline experiment using our highly simple AmazonSimple net yielded a F_2 of 0.81137 in the validation dataset, and a final result of 0.86607 in the test dataset. This was a surprising result as the score is quite high for such a basic network with no residual connections based only in RGB channels. This gave us a base metric to improve and a

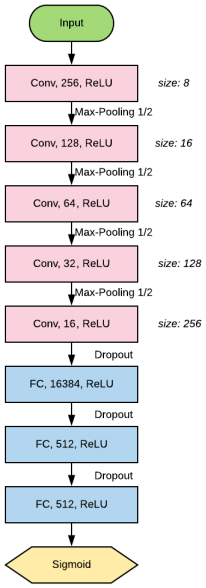

Figure 5. AmazonSimpleNet architecture.

starting point for the rest of the experiments.

4.2. Infrared Channel

The list of experiments with the infrared channel, NDVI, NDWI and NIR-color channel combinations are presented in table 1. The RGB + NDWI experiment was unable to run to completion as the gradient exploded. The experiment results were compared to the F_2 result obtained using only the main RGB channels with ResNet18. It is important to note that none of the combinations using the infrared channels improved the overall performance of the network and they implied a heavier computational cost.

Table 1. Experiments using different infrared channel combination as additional information and main channel substitution. F_2 score calculated over the test dataset, and the best result is highlighted in bold.

Backbone	Representation Channels	F2 Score
ResNet18	**RGB**	**0.88836**
ResNet18	RGB + NIR	0.61550
ResNet18	RGB + NDVI	0.78410
ResNet18	RGB + NDWI	N/A
ResNet18	NIR-R-G	0.87581
ResNet18	NIR-R-B	0.87836

Although a thorough search for improvement using a large scope of possible ways to add NIR channel information was conducted, no benefit was found of adding fourth-channel information or changing the original RGB image representation.

4.3. Architecture

Results for the different backbones used, the number of training epochs and the final F_2 score obtained from the

online test server are presented below.

4.3.1 20 Epochs

The results for the three main backbones with only RGB channel input are shown in the following table. As can be seen highlighted in the table, Inception V3 provided the best F_2 score on the validation servers.

Table 2. Results for ResNet18, ResNet101 and Inception V3 using 20 training epochs. Best results are highlighted in bold.

Backbone	Epochs	F2 Val Score	F2 Test Score
ResNet18	20	0.37738	0.88836
ResNet101	20	0.87463	0.84604
Inception V3	**20**	**0.90907**	**0.90601**

In addition, the train and validation losses of the architecture with the best results are shown in figure 6. As observed, the training of this experiment is appropriate, both losses have a correct decrease and there are no signs of overfitting or underfitting of the model.

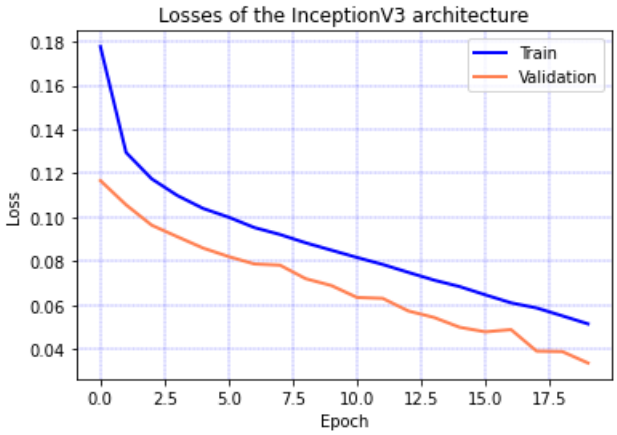

Figure 6. Losses obtained during the training of the best model with Inception V3 as a backbone.

4.3.2 40 Epochs

The final experiments carried out included increasing the number of training epochs as a way to reduce possible model underfitting during the training phase of the model. The number of epochs was duplicated and the same input channels were kept in order to directly evaluate the effect of increasing the number of epochs.

Table 3. Results for ResNet18, ResNet101 and Inception V3 using 40 training epochs.Best results are highlighted in bold.

Backbone	Epochs	F2 Val Score	F2 Test Score
ResNet18	40	**0.89974**	**0.89848**
ResNet101	40	0.89021	0.89468
Inception V3	40	0.89589	0.88549

5. Discussion

Satellite imagery has the potential of becoming an important information provider to evaluate land use over time in the Amazon basin. Due to the extent of the datasets that are necessary to monitor such diverse phenomena, a quick automatic way to annotate the dataset is necessary to optimize resources. An algorithm like the one we used has the potential of becoming that primary filter through which new data is channeled before moving to more specific problems.

Perhaps the greatest source of error for the overall performance of the network consisted of the massive class imbalance in the training and test data sets. Even though the main distribution of the classes was similar in the training and validating sets as is shown in figure 7, the training of the network could be highly biased towards the most frequent categories such as primary, which occurred in the majority of the images. Future work could include Data Augmentation techniques in order to evenly distribute the label imbalance during the training, avoiding general overfitting towards the most frequent categories.

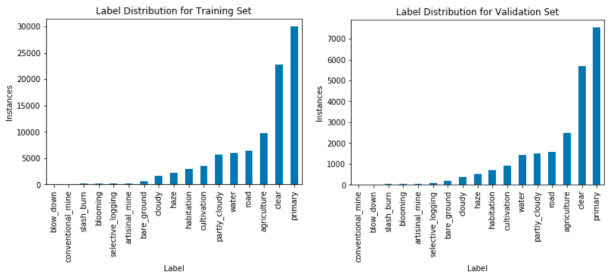

Figure 7. Histograms of categories in the Training and Validation datasets. Although the class distribution is similar, a massive class imbalance can be seen in both sets.

There are other queues from the data that could be used in future work as an extra layer of information to increase the accuracy of the model. This includes data insights such weather labels *(cloudy, partly cloudy, clear and haze)* being disjoint from each other, as well as labels that frequently appear together such as *Habitation* and *Agriculture*. Using this information as well as other mutually exclusive or associated categories could provide a richer network architecture to carry out the final classification problem. Apart from this, several discussions on the Kaggle platform have mentioned the presence of corrupt TIFF files which do not inhibit the model from working but can significantly reduce performance, specially when including the NIR channels.

Concerning the experiments made with the infrared

channels it was obtained that the best combination was NIR-R-B, however with respect to the baseline no significant improvement was obtained as far as the F_2 score is concerned. This could be the result of NIR transformation creating information loss or detail loss in some of the images from the training data set. It is interesting to note that none of the combinations that included adding a fourth channel of information resulted in a higher F_2 final score, considering each image was given a full extra channel of information.

One of the possible explanations that could explain this behavior is information redundancy, where the NIR, NDVI and NDWI channels did not provide additional information to the one found in the R,G and B channels. An example to illustrate this could be adding an extra color channel to the MNIST database, where the second channel provides majority of the information contained in the first channel and gives no extra queues for the network to correctly classify an image. However, the most likely cause for performance reduction could be due to the mentioned corrupted files in the training set.The appearance of the corrupted TIFF files represents a huge loss for the training possibilities, and perhaps in a correctly created data set the 4th channel information could greatly enrich the network and increase performance.

Regarding the results obtained when changing the model architecture, the best result evaluated in the challenge server was the Inception V3 model using RGB images during 20 epochs with an *F2* score of 0.90601. There are some reasons why this model has the best performance compared to the other two ResNet models. The first reason is that the InceptionV3 model has less trainable parameters 6.4 million [12], compared to the ResNet18 which has 11 million and ResNet101 with 44 million trainable parameters [13]. This leads to a much faster training of the network compared to the other two.

The second reason is the way that Inception tackles the problem of the variability of the size of the salient feature. Lager kernels are preferred for more global features that are distributed over a large area of the image, on the other hand, smaller kernels provide good results in detecting area-specific features that are distributed across the image frame [12]. For effective recognition of such a variable-sized feature, like those in our database, we need kernels of different sizes that can extract richer semantic information from larger and small details of the images. That is what is performed in InceptionV3. Instead of going deeper in terms of the number of layers, like in ResNet's architectures, it goes wider. Multiple kernels of different sizes are implemented within the same layer which results in a better performance of the network [12].

On the other hand, when carrying out the experiments by increasing the number of epochs it was observed that the performance of the best model of ResNet101 and InceptionV3 did not increase compared to the models trained during 20 epochs. It is evident that for the development of our experiments we did not present experiments changing other hyperparameters of the model such as the optimizer, loss etc since in the literature and in the challenge discussion forums it was found that the best methods are those which implement different types of transformations to the database images and those which implement the state of the art architectures as far as the image classification task is concerned.

Our overall performance was comparable to the top scores from the competition leaderboard, the top scoring performance being 0.934. It is important to note that the best performing algorithm used a series of high performing CNNs (11) trained simultaneously and made the final classification decision based on a weighted mean calculation of the output of each of the CNN's [14]. For future work a multi network solution that includes the final network that we present could be implemented in order to tackle complex classification problems such as this one in order to improve prediction capability.

6. Conclusion

We created a high performing CNN to predict categories on a multi label classification problem with Satellite Images from the Amazon Rain forest using Inception V3 as a backbone and only the RGB channels as input. Satellite image multi label classification is an important first step that opens the door to different finer grain problems such as estimating deforestation and mining areas, detection of new illegal mining activities and agricultural expansion as well as monitoring general biological phenomena in a monthly or yearly basis. It is a way to provide an initial baseline of control for such a vast, wild and unprotected area by governments and organizations. It is an entry to much more complex algorithms that provide scientists with vital information about the impact of economic and land use government policies, anthropologists with insights about the future of the ancestral cultures that inhabit still unexplored areas of the jungle like the Serranía del Chiribiquete in Colombia, biologists with new comprehensions of the cycles that affect biodiversity and humankind with real time data about the state and future of the most bio diverse ecosystem in the planet.

7. Credits

We hereby certify that all members of the team (Jessica Castillo, Juan Francisco Suescun and Juan David García)

participated in equal parts in the development of the algorithm, experimentation and results phase, article writing and video production.

References

[1] T. E. of Encyclopaedia Britannica, "Amazon rainforest region, south america," Oct. 2019.

[2] A. A. Foundation, "Water," 2018. Data retrieved from https://amazonaid.org/water/.

[3] E. Reuters, "Amazon deforestation could speed up in 2020: expert," 2020. Original article on https://reut.rs/2W68NoS.

[4] C. Baker, R. Lawrence, C. Montagne, and D. Patten, "Mapping wetlands and riparian areas using landsat ETM imagery and decision-tree-based models," *Wetlands*, vol. 26, pp. 465–474, June 2006.

[5] M. Längkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, "Classification and segmentation of satellite orthoimagery using convolutional neural networks," *Remote Sensing*, vol. 8, p. 329, 2016.

[6] J. D. Wegner, S. Branson, D. Hall, K. Schindler, and P. Perona, "Cataloging public objects using aerial and street-level images — urban trees," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6014–6023, 2016.

[7] Kaggle, "Planet: Understanding the amazon from space," 2020. Data retrieved from https://www.kaggle.com/.

[8] S. Chandak, V. Chitters, and S. Honnungar, "Understanding the amazon rainforest from space using cnns," 2017.

[9] Y. Chen, F. Dong, and C. Ruan, "Understanding the amazon from space," 2017.

[10] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *Geoscience and Remote Sensing Letters, IEEE*, vol. 11, pp. 1797–1801, 10 2014.

[11] C. Quintano and E. Cuesta, "Improving satellite image classification by using fractional type convolution filtering," *International Journal of Applied Earth Observation and Geoinformation*, vol. 12, pp. 298–301, 08 2010.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[14] K. Team, "Planet: Understanding the amazon from space, 1st place winner's interview," 2019.