MSc in Business Analytics

Big Data Content Analytics

Project Neural Networks

Authors

Aloukou Eirini BAPT1702
Lymperi Ourania-Anna BAFT1711

BUSINESS ANALYTICS
Master of Science

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF BUSINESS

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

Contents

1. Description

Every minute, the world loses an area of forest the size of 48 football fields. Deforestation in the Amazon Basin accounts for the largest share, contributing to reduced biodiversity, habitat loss, climate change, and other devastating effects. But better data about the location of deforestation and human encroachment on forests can help governments and local stakeholders respond more quickly and effectively.

In this project, we are challenged to perform a deep learning analysis with satellite image classification, in order to label image chips with atmospheric conditions and various classes of land cover/land use. Resulting algorithms will help the global community to better understand where, how, and why deforestation happens all over the world - and ultimately how to respond. The technical target of the project is to train a model that would recognize the two most important types objects in satellite data in order to track the human footprint in the Amazon rainforest. This could also assist to differentiate between natural and human causes of forest loss.

The regulation of practices that affect the environment has been a relatively recent development in the United States, but it is a good example of government intervention in the economy for a social purpose. Since the collective rise in consciousness about the health of the environment, such government intervention in business has become a hot topic not only in the United States political arena but across the globe.

Therefore, this is very interesting and important project, since the protection of the environment means protecting human health. It is also a business-wise project since the idea of environmental measures is hot topic and its importance will grow through the years. Today, environmental protection policy remains at the forefront of political discussion and at the top of the current administration's agenda particularly as it relates to clean energy and climate change.

The business workflow of the project is represented in the following figure:

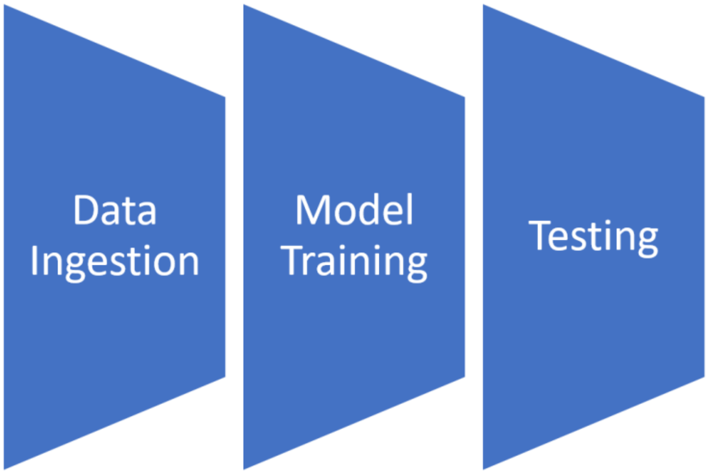

Figure 1: Business Workflow of the project

2. Mission

The Amazon is the world’s largest rainforest and the largest river basin on the planet. There’s a reason the Amazon was the place that inspired scientists to coin the term “biodiversity”. The region is home to 10 percent of all plant and animal species known on Earth. There are approximately 40,000 species of plants and more than 400 mammals. Birds add almost 1,300, and the insects reach millions. Of plant species found in this region, 75% are unique to the Amazon, and there are 3,000 species of fish, the largest number of freshwater fish species in the world. In addition to its unparalleled diversity of life, the Amazon plays an essential role in helping to control the entire planet’s atmospheric carbon levels. The Amazon Basin stores approximately 100 billion metric tons of carbon, more than ten times the annual global emissions from fossil fuels.

While it covers 2.6 million square miles across nine countries — Brazil, Bolivia, Peru, Colombia, Ecuador, Venezuela, Guyana, Suriname, and French Guiana — about 60 percent of the Amazon Basin is in Brazil.

Today, the Amazon is facing a multitude of threats as a result of unsustainable economic development; 20% of the Amazon biome has already been lost and the trend will worsen if gone unchecked.

Amazon is the **biggest deforestation front** in the world and interventions are urgently needed to prevent a large-scale, irreversible ecological disaster.

Deforestation rates in the Amazon have declined over the last decade, but continue at an alarming rate. Brazil is responsible for half of the deforestation in the Amazon, but deforestation in the Andean Amazon countries – namely Bolivia and Peru – is increasing. Deforestation is concentrated in particular in 25 “sub-fronts” that span across multiple countries.

The areas showing the greatest deforestation rates are those that have more roads. The strong correlation between the location of deforestation fronts and the presence of existing roads or projections of new roads suggests that in the near future there will be isolated deforestation fronts becoming connected along major infrastructure development routes. This is a fact that will be confirmed in this project using data techniques.

For most of human history, deforestation in the Amazon was primarily the product of human settlement and development of the land, subsistence farmers who cut down trees to produce crops for their families and local consumption. Prior to the early 1960s, access to the forest's interior was highly restricted, and the forest remained basically intact. Farms established during the 1960s were based on crop cultivation and the slash and burn method. However, the colonists were unable to manage their fields and the crops because of the loss of soil fertility and weed invasion. The soils in the Amazon are productive for just a short period of time, so farmers are constantly moving to new areas and clearing more land.

But in the later part of the 20th century, that began to change, with an increasing proportion of deforestation driven by industrial activities and large-scale agriculture. By the 2000s more than three-quarters of forest clearing in the Amazon was for cattle-ranching. These farming practices led to deforestation and caused extensive environmental damage. Deforestation is considerable, and areas cleared of forest are visible to the naked eye from outer space.

The result of this shift is forests in the Amazon were cleared faster than ever before in the late 1970s through the mid-2000s. Vast areas of rainforest were felled for cattle pasture and soy farms, drowned for dams, dug up for minerals, and bulldozed for towns and colonization projects. At the same time, the proliferation of roads opened previously inaccessible forests to settlement by poor farmers, illegal logging, and land speculators.

Why Save the Amazon?

For one, the Amazon is on the frontlines of the fight against global warming.

Currently, the Amazon is a carbon sink, meaning it stores carbon dioxide and prevents it from entering the atmosphere and fueling climate change. Deforestation, on the other hand, releases that carbon into the air, making global warming worse. Because of this, deforestation accounts for about 10 to 15 percent of global greenhouse gas emissions. Losing the Amazon means more carbon emissions and a warmer world.

No matter how far from the region we live, the Amazon plays an important role in all of our lives, and we all play a role in protecting the homes of thousands of people and some of the world’s rarest wildlife.

Previous work

Multi-label satellite image classification has been a task of interest ever since the first multi-spectral remote sensing imagery became available (to civilians) in the early 1970s. The overall approach has remained conceptually the same - record satellite observations, derive a feature vector for the image, run a classification algorithm, produce classification labels.

Planet, designer and builder of the world’s largest constellation of Earth-imaging satellites will soon be collecting daily imagery of the entire land surface of the earth at 3-5 meter resolution. While considerable research has been devoted to tracking changes in forests, it typically depends on coarse-resolution imagery from Landsat (30-meter pixels) or MODIS (250 meter pixels). This limits its effectiveness in areas where small-scale deforestation or forest degradation dominate.

Furthermore, these existing methods generally cannot differentiate between human causes of forest loss and natural causes. Higher resolution imagery has already been shown to be exceptionally good at this, but robust methods have not yet been developed for Planet imagery.

Our approach

In this project, we will try to identify the deforested areas using deep learning techniques and to understand the size of the problem and its causes. Given the dynamic labeling – with training images having between one and seventeen different labels – we will experiment with various multi-class, multi-label classifiers using state-of-the-art deep learning approaches borrowing from existing image classification model architectures.

The authorities could also reforest the harmed areas and propose measures for the protection of the ones of high risk. For that reason, we used only the necessary data provided in order to have the most accurate results. The following algorithms will help the global community to apply them in every forest area worldwide, that has environmental interest or suffers from deforestation.

3. Data

The data used for this project was derived from Kaggle website and that is images from Planet's full-frame analytic scene products using our 4-band satellites in sun-synchronous orbit (SSO) and International Space Station (ISS) orbit. The set of images for this assignment use the GeoTiff format and each contain four bands of data: red, green, blue, and near infrared (RBG-NIR). The specific spectral response of the satellites can be found in the Planet documentation. Each of these channels is in 16-bit digital number format and meets the specification of the Planet four band analytic ortho scene product.

The imagery has a ground-sample distance (GSD) of 3.7m and an orthorectified pixel size of 3m. The data comes from Planet's Flock 2 satellites in both sun-synchronous and ISS orbits and was collected between January 1, 2016 and February 1, 2017. All of the scenes come from the Amazon basin which includes Brazil, Peru, Uruguay, Colombia, Venezuela, Guyana, Bolivia, and Ecuador.

The data is available in both tif and jpg format. The JPG chips were produced for reference and practice, by the Planet visual product processor. The jpg images were represented as vectors of numbers, RBG values for each pixel. Those values range anywhere between 0 and 255. They refer to the same scene content.

The given data contained labeled training and testing images of the Amazon river basin taken from Planet Labs satellites, as well as one csv file with the corresponding categories/labels. Each JPG image is on average **15KB**, and the TIFs **538KB**. Therefore, we used only the jpg images for this problem solving assignment, because tif images were consuming more memory space.

The chips are labeled reasonably well. The labels fit into three categories: atmospheric conditions, common land cover/use occurrences, and rare land cover/use occurrences. Sometimes multiple labels occur in a single image, with some exceptions.

The training set consists of **40479** labeled images and we used the 10% of this for validation. The dimension of images, schema of our data, was originally (256, 256, 3). Each image that used for training has several labels, 17 in total. Some of the satellite images available are the following:

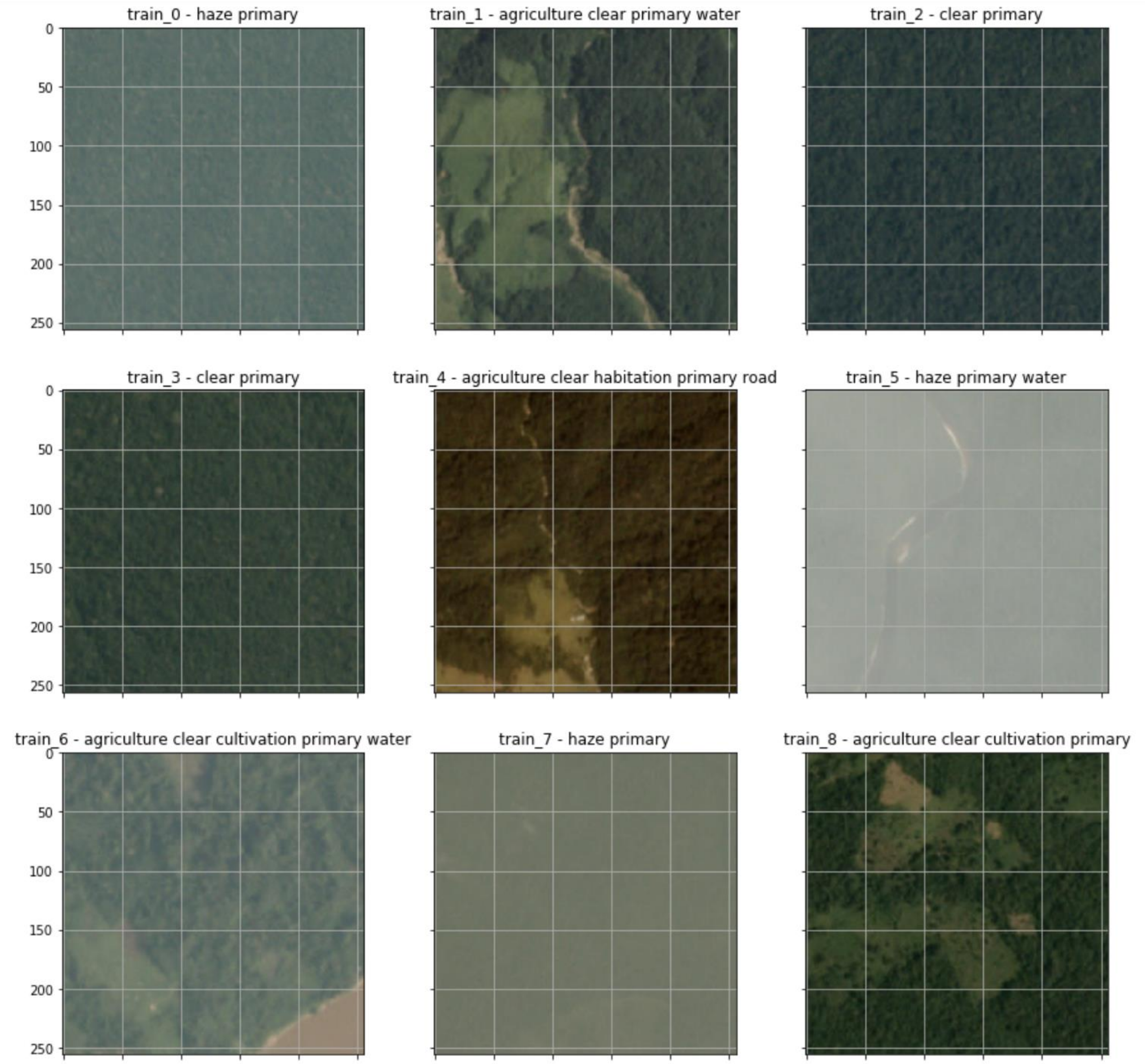
train_0 - haze primary
train_1 - agriculture clear primary water
train_2 - clear primary
train_3 - clear primary
train_4 - agriculture clear habitation primary road
train_5 - haze primary water
train_6 - agriculture clear cultivation primary water
train_7 - haze primary
train_8 - agriculture clear cultivation primary
0
50
100
150
200
250

Figure 2: Labeled satellite images

For the purposes of our analysis, we used Python, focusing on a specific library specializing in deep learning, Keras and Tensorflow. Keras is a high level API which is built on top of Tensorflow. The general technique or family of algorithms we will be using in neural networks, with notable example being CNN.

From the next diagram, we can see statistical data such that the most popular labels of these images are primary (rainforest), clear, agriculture and road and the least common labels are selective logging, artisinal mining, blooming, slash and burn, blow down and conventional mining. Further worth noting is that the distribution of labels was quite unbalanced as evidenced in Figure 3:

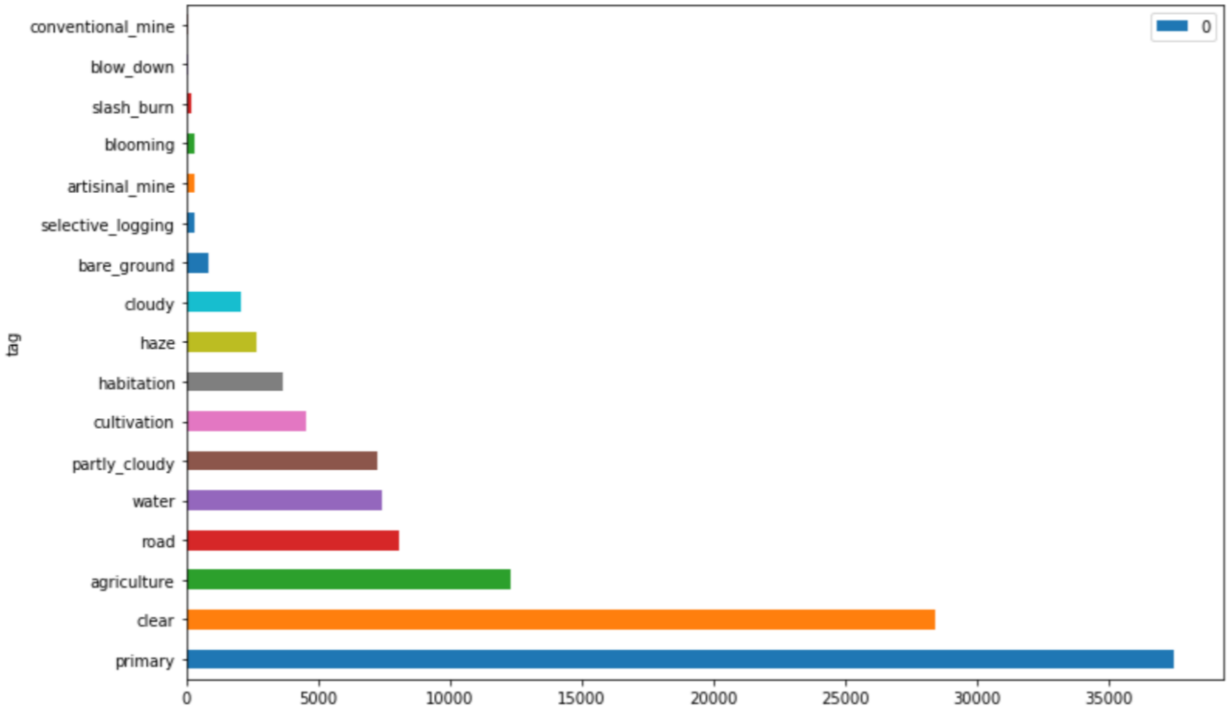

Figure 3: Frequency plot providing all the categories of the data

In order to proceed with our analysis, we need to apply some preprocessing methods to our data. First of all, we reduced the number of different labels to the two of the most frequent and important ones for our purpose to analyze, agriculture and road. We consider these as the most important, since they are the strongest indication of a deforested area (human causes of forest loss). This decision will also protect us from high degree of noise, since the road images look a lot alike to the river images.

Agriculture: Commercial agriculture, while an important industry, is also a major driver of deforestation in the Amazon. For the purposes of this dataset, agriculture is considered to be any land cleared of trees that is being used for agriculture or range land.

Road: Roads are important for transportation in the Amazon but they also serve as drivers of deforestation. In particular, "fishbone" deforestation often follows new road construction, while smaller logging roads drive selective logging operations. For our data, all types of roads are labeled with a single "road" label. Some rivers look very similar to smaller logging roads, and consequently there may be some noise in this label.

We also rescaled our images to smaller dimension (128, 128) using skimage algorithm and gray scaling, since we are not concerned about the color of the images, just its shape, i.e. shape of road, agricultural areas in squares etc. However, this was not always beneficial to the overall result, in most cases even worsening the classification accuracy, i.e. the forest is green and a deforested area is brown.

Another pre-processing method was the normalization of the data, in order to reduce the range of potential colors in each image. Therefore, we divided each 0-256 value by 255. This serves to reduce the range of values for each image to [0, 1].

Finally, the schema of our data, after the above preprocessing methods, is (128, 128, 1).

4. Methodology

The algorithm implemented for this study was Convolutional Neural Networks (CNN). The general methodology used could be briefly summarized to preprocessing the data, described extensively above, applying CNN algorithm (checking its parameters and accuracy results), analyzing the results and extracting important conclusions.

The CNN algorithm consists of 4 primary stages: convolution, non-linearity, pooling and classification. During convolution, features are extracted from the input image by using a sliding filter (smaller dimensions than the original) and getting a dot product (result of matrix multiplication) called a feature map. Different filters can be applied and feature maps of various sizes can be produced. After the convolution process, a non-linear function is again applied to each pixel. In the pooling step, the size of the feature map is reduced even further by applying a sliding window filter to it, extracting the maximum (or sometimes average) of all values present in the window each time. The last pooling layer is the input of the fully connected layer, the final stage of the process.

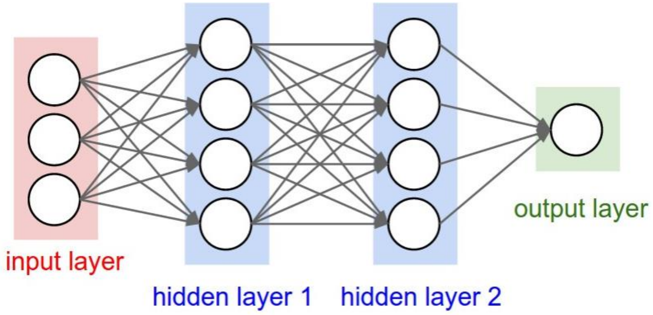

Figure 4: 3-layer Neural Network

The procedure followed in this approach, including any parameters and transformations, is presented below:

- The network is designed as follows:
  - Conv ReLu Layer with 64 3x3 filters,
  - Conv ReLu with 64 filters,
  - 2x2 Max Pool, Dropout,

- 0.25 Dropout after max pooling,
- Fully Connected ReLu layer,
- 0.5 Dropout after first dense layer
- Softmax activation

A figure of the network topology is presented in the Figure 4. The NN uses binary cross-entropy loss with the adam optimizer. The above hyper-parameters, ReLu, softmax, dense and max pooling were the most efficient to train our model. We also had 4 parameters (conv_1, conv_2, dense_1, dense_2) with their corresponding weights.

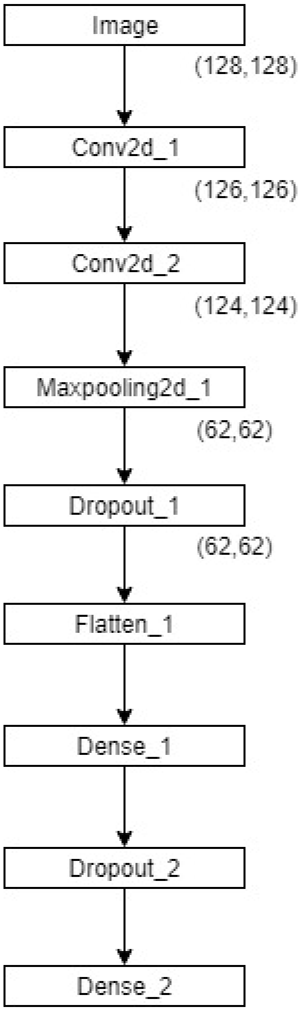

Figure 5: The architecture of the model

- Several small experiments were used to improve the performance of the network topology:
  - Both 128x128 downscaled images and 64x64 images were used.
  - Batch normalization did not seem to improve the score.
  - Using lower levels of dropout was less successful.
  - The most successful dropout levels were 0.25 after max pooling and 0.5 after the first dense layer.

  - A reduced number of filters in the convolution layers seemed to perform more poorly.
  - Using large numbers for epochs was less successful.
  - Changed the one of the labels with the most popular, primary, and the model accuracy was launched to approximately 95%.
- As a result of running the above described neural network on a split of 7,486 train / 832 validation samples over 5 epochs, an accuracy of 0.7812 was achieved.

	Accuracy	Loss
train	0.7459	0.5785
validation	0.7812	0.5256

Table 1: Results of the final model

For the implementation above, the batch size of the data was 128.

It should be noted that human accuracy for this particular dataset was 98%. The winner of this competition of the Kaggle achieved 0.93 accuracy in his model. However, we consider our implementation to be quite satisfying.

The fact that the accuracy could not get higher than this percentage, could be explained by the small amount of data remaining after the extraction of 15 categories and the high quality of the images. That could also be confirmed by the huge improvement of the model accuracy when the most common label replaced the existing label, increasing significantly the number of training data and causing overfitting.

5. Results

Figure 6 is the loss function of our model and, essentially, represents Table 1. More specifically, the red line is the train accuracy, the orange line is the validation accuracy and the blue line is the train loss while the green line is the validation loss. We could observe that the loss values are pretty high, and even when we tried to reduce the accuracy in order to improve the loss, it did not get better than this.

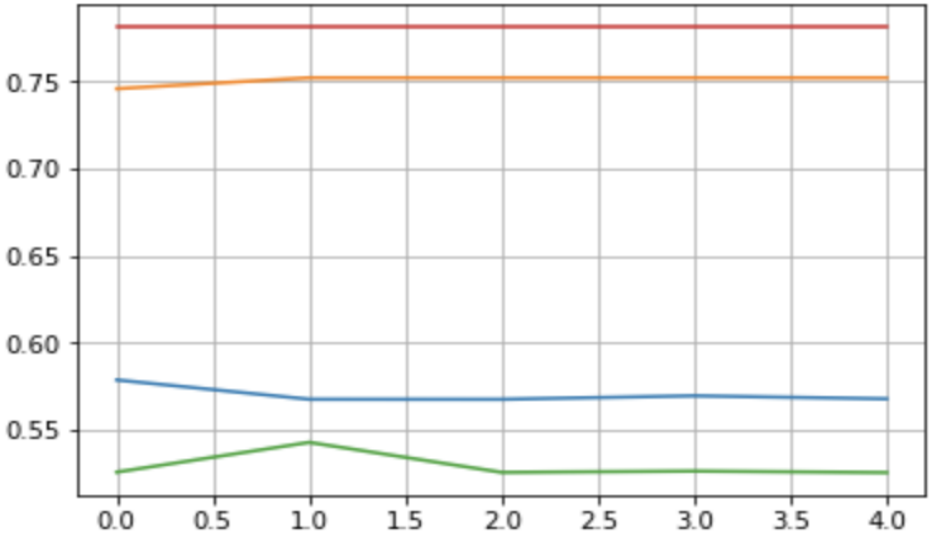

Figure 6: Loss function

The training time that the final model took, was one hour and a half. Compared to other cases, it is a small timing for a training model, that happened because of the small dataset we ended up with after reducing the categories.

Technical characteristics of the machine used for the assignment:

MacBook Pro

Processor 2,3 GHz Intel Core i5

RAM 8 GB 2133 MHz LPDDR3

SSD Storage

6. Members/Roles

Our team consists of two members: Aloukou Eirini as Data Scientist and Lymperi Orianna as Business Analyst. Eirini is a statistician, graduated from Statistics and Insurance Science of University of Piraeus and Orianna is an Environmental Engineer, graduated from Environmental Engineering School of Technical University of Crete.

For the purposes of the assignment, our team needed to conduct a lot of research. The two members contributed equally to bibliography research, finding similar cases and code fragments to face any difficulties and taking important data processing decisions. Therefore, the team members did not have distinct responsibilities during the implementation of the project, but we could say that Eirini was more upon the construction of the technical code and Orianna on the analysis of the results.

7. Bibliography

- Material and laboratory code from Big Data Content Analytics lectures
- A Quick introduction to neural networks: https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/
- An Intuitive Explanation of Convolutional Neural Networks: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/
- An intuitive guide to Convolutional Neural Networks: https://medium.freecodecamp.org/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050
- Amazon deforestation WWF: http://wwf.panda.org/our_work/forests/deforestation_fronts/deforestation_in_the_amazon/

8. Time Plan

The original time plan for the implementation of the project is the following:

	1st week	2nd week	3rd week	4th week
Bibliography research	🟦			
Data organization	🟦			
Model construction		🟦	🟦	
Results				🟦
Report				🟦

Original Time Plan

However, the time plan was changed as we deepened to the implementation of the project. The decision for the cleaning and understanding of the data took us more time than expected, as well as the model construction and results, since we had lack of programming skills. The actual time plan is presented next:

	1st week	2nd week	3rd week	4th week	5th week
Bibliography research	🟦				
Data organization	🟦	🟦			
Model construction		🟦	🟦	🟦	
Results				🟦	🟦
Report					🟦

Actual Time Plan

9. Contact Person

Team representative → Name: Orianna Lymperi
Phone: 6978694156
E-mail: ori.anna.93@hotmail.com

10. Discussion

In this project, we tried to tackle the challenge of understanding one subset of satellite images – those capturing images of the Amazon rainforest – with the particular goal of aiding in characterization and quantification of the deforestation of this area. A moderately successful approach was shown. A deep convolutional neural network trained from scratch was able to achieve an accuracy of 78% on the dataset. The model is far from perfect, but it was interesting to explore what did and did not work on this type of dataset. We have learned that the number of layers is very important, as are the thresholds in the final layer. And, certainly the input size can make a large difference on training time (as can use of GPU code). Additional work could also test on the tif set. However, it's likely that the labels were done on the jpg set, so it's unclear if this would help. Further work could include using VGGnet or Resnet transfer learning, tuning hyperparameters further, and exploring other types of networks. It's quite possible that it would be easier to achieve a better score when using a network like this that is pretrained on millions of regular images.