Land Classification Using Deep Neural Networks Applied to Satellite Imagery Combined with Ground-Level Images

Lus Ricardo Fonseca Agostinho Freixinho
Instituto Superior Técnico, Universidade de Lisboa

***Abstract*—With the exponential growth of georeferenced labeled and high-detailed imagery information of our world available on the web, combined with both computational power and the algorithms improvements from the last years, the land classification task becomes more reliable from the start, and thus it is possible to achieve sharper results when applying deep learning techniques to that imagery.**

Besides using exclusively high resolution data to train deep learning models, we can now apply it to terrain elevation models in order to have a more precise information about the visibility areas of a picture to better classify the land.

Mapping the multiple sources of human activity involved in land or minimizing natural hazards, are two of many applications of a Geographic Information System (GIS), and our geographic knowledge discovery benefits from having this kind of information.

This work advances a new study on the feasibility of applying deep neural networks and visibility analysis techniques to a combination of satellite imagery and ground-level images to classify land, either from nowadays (that contain satellite imagery) and also from past periods of time, when satellite imagery was not available but only ground-level photographs.

I. Introduction

Land classification models provide fundamental information in multiple spatial planning contexts. This specific computer vision task can be split in two main classifications, Land cover and Land use. Both are demanding classification tasks, aiming to describe the components that exist on the surface of the planet and the functional activities of a certain area respectively.

Multiple studies, some using deep learning techniques [1], have explored efficient techniques to produce these types of classification, using either aerial imagery or ground-level photographs, however the combination of both sources to create this kind of classification systems using convolutional neural networks is still a poorly explored area.

Now, more than ever, we have access to plenty land imagery data that has been provided, everyday during the past years, by millions of people that keep feeding the web's insatiable appetite for content by posting photos on social networks, images that were captured by satellites and, from now on, the aerial shots from drones. All these media resources have been used as contributors to expand our geographic knowledge discovery. With historic content, such as geotagged photographs from last century (and even the one before), appearing on the internet freely available it is now possible to expand our past geographic knowledge too, using these methods.

The amount of photographs posted every minute on the internet contribute to a valuable source of Volunteered Geographic Information (VGI), potentially containing timely geographic information, that can help to give another perspective of the land that aerial images cannot obtain, and assigning the citizen to an important role in land mapping tasks [2]. Working with the provided content from platforms containing millions of photographs, such as Flickr and others do, it is possible to train deep neural networks models.

After the ImageNet Large-Scale Visual Recognition Challenge 2012 [3], winning entry [4], the approach to problems related to image classification usually are solved by feeding a state-of-the-art deep Convolutional Neural Network (CNN) model with a large data source of information in order to retrieve a classification label. Multiple algorithms have been developed to interpret high-level properties of images and produce estimations, such estimations for natural beauty using CNNs trained with ground-level photographs [5].

With visibility analysis composed by viewshed techniques, that incorporate terrain elevation models and camera calibration methods, is now possible to assign photographs to specific geolocations autonomously, in order to generate geographic maps under a desired classification, which recently slowly started to be under sight of some research studies.

With these available resources and known technological mechanisms, now it is possible to create a system that uses geotagged photographs either from the past and the present (available in social networks and VGI) combined with satellite imagery (made accessible by reliable institutions that collect these ground truth information) to explore geographical knowledge either from the past or from the present.

In this case specially, this new novel approach, besides trying to map land coverage classification from the current years in order to compare with the ground truth of land coverage maps from the present, it is testing the feasibility to use photographs from the 19th and early 20th Century (combined to the trained model developed) to generate land coverage maps of periods of time when resources such as aerial photographs or satellite imagery were not a reality, which might help exploring this type of geographical knowledge of some epochs of our history that until now we did not have knowledge about. The code used for the development of

this approach can be found on my Github[1].

A. Objectives

This thesis introduces a novel approach for possible solution to the demanding classification task of land classification, that usually requires a lot of costly resources in order to be reliable, exploring a different technique.

Since, nowadays, there are a lot of images containing possible descriptors of land (either from the present and from the past) freely available on the internet, the key idea behind it, is to perform a geographically land classification, through an automatic system, that combines ground-level photographs and satellite imagery, applying viewshed techniques and deep neural networks.

In the end, the goal is to understand how well this model can classify a new geographic location using available imagery regarding that place and also if it is possible to classify land, using historical photographs, of past periods of time, in which there was no access to some resources that today we take for granted, such as satellite imagery.

B. Contributions

The main contribution of this M.Sc thesis is the implementation of a system that mechanizes, through deep learning, mapping land coverage classification, relative to periods of time in which satellite imagery or aerial photographs were not available to aid the land classification task, using ground-base photographs (from nowadays and historical) and satellite imagery from today .

II. Concepts and Related Work

A. Fundamental Concepts

1) Convolutional Neural Networks (CNN): can be interpreted as Neural Networks that share their parameters across the space, are built over a chain of individual layers, each of them achieving different results. This chain is characterized by each step receiving feature maps and sending to the next step a new feature map until the last step, the fully connected layer, that will produce a map to a class of probability for a certain given feature map as input.

A major component that defines this architecture are the convolutions. The objective of a convolutional network is to progressively apply convolutions in order to reduce the spatial dimension of the input while increasing the depth so that in the end we remain having only parameters that map to features of the initial image in order to apply a classifier afterwards.

2) Recurrent Neural Networks (RNN): are a powerful model for sequential data modeling and feature extraction that brings the learning persistence concept to neural networks. These networks are called recurrent due to the fact they have a memorybased on the information computed so far in the process. This memory expands from performing the same task for each element of the input sequence in loop, which will result in an output that was influenced by the previous computations of the process.

[1]https://github.com/luisfreixinho/Thesis-Land-Coverage-Classification

3) Viewshed Analysis: An analysis that relies on the use of elevation models. Tries to discover which points, within a DEM of a specific terrain, are visible given a single (or multiple) viewpoint/observer point.

B. Deep Learning Methods for Image Classification

Convolutional Neural Networks have been used in image classification tasks for almost 40 years now. The first successful application of a convolutional network, LeNet architecture [6], used mainly for character recognition tasks.

In 2012, comes up a strong reappearance of this topic promoted by the fast-paced evolution of computational power, improved algorithms and the availability of larger datatsets that led to AlexNet [4] (Supervision), which was a huge success when applied to large-scale imagery, by outperforming all previous non deep learning based models by a significant margin. This architecture is based on a convolutional layer followed by pooling layer and a normalization and again one more convolutional and pooling layers and normalization. After this normalization follows a few more conv layers, a pooling layer, and then several fully connected layers, making a total of five convolutional layers and three fully connected layers.

AlexNet has been used a lot from then until a couple years ago , although multiple efforts have been made to enhance the SuperVisions architecture and achieve convolutional networks state-of-the-art results on classification tasks. One example of these tries, addressed the depth aspect on ConvNets by fixing all the other parameters and adjusting the depth by adding more convolutional layers, up to 19 weight layers (VGGNet [7]), in order to understand how would it affect the neural network's accuracy and performance in the large-scale image recognition setting, by presenting upfront two different key details - much deeper networks with much smaller filters.

The novel approach here was to go from the 8 typical AlexNet layers to 16 or 19 layers combined with the usage of very small, but stacked, filters on convolution layers with the objective of taking advantage of the large receptive field they produce plus the increase of non-linearity and the decrease of the amount of parameters to learn, that a single 7x7 filter would not offer. The Pooling layers perform 2x2 max-pooling without padding and 2 of stride.

Three fully-connected layers follow the stack of convolutional layers, being the first two composed by 4096 channels each, and the last one with 1000 class channels. With this simple structure all the way through the network, these VGG models achieved 7.3% top five error on the ImageNet challenge at the year they were presented.

More recently appeared the MobileNets [8], a new model built over the suggestion of a factorization of standard 3x3 convolutions, into one depth-wise 3x3 convolution, followed by a point-wise convolution, which is was proved more efficient. The model considers hyperparameters that choose higher accuracy or performance. This choices reflects in several advantages of these model in comparison to other state-of-the-art models, such as reduced network size, number of

parameters, more fast in performance and small, low-latency convolutional neural network. It is also considered interesting for mobile applications.

By applying these models as the ones briefly described before it is now possible performing image classification tasks such as estimating the scenicness of a photography [5], by using modified versions of GoogLeNet CNN initialized with weights from another CNN pre-trained with content related to this task of classification landscape pictures.

It is now possible too apply deep learning techniques as a procedure to retrieve modifications of original images to improve the results of possible posterior computer vision tasks, as e.g detection, object classification or segmentation tasks, by colorizing gray scale images in a indistinguishably way from real color photos through per-pixel classification problems [9] [10], or even by image-to-image high quality translation between multiple domains [11].

C. Terrain Classification with Georeferenced Multimedia Contents

Classifying land based on georeferenced media results from outputting an objective label to a portion of land using imagery that belongs to it, which can be challenging, due to the implicit subjectivity that such type of content can transmit. Another challenge results from the exponential growth of the available and noisy media content online. Previously, geographic knowledge discovery tasks, such as land cover mapping, were performed mostly with limited and exclusive content like topographic maps, satellite imagery and other overhead imagery e.g. the national land cover database, which were resource-expensive to collect due to the detailed information they had to contain.

Some works rely on the analysis of the usefulness of using Volunteered Geographic Information (VGI) [12], such as Flickr, regarding this terrain classification task, and consider that there is potential using geotagged photographs in land cover usage analysis despite the existent informality in these kind of online data repositories.

In some cases it is explored usability of social media data (Sina Weibo) about Points of Interest (POI) in terms of land validation [13] by introducing a two step validation framework that uses modified decision trees. In the end is performed an artificial surface validation using common pixel comparison and confusion matrix methods to validate the accuracy of the classified data in comparison to the ground truth of Beijing.

About using deep learning to classify terrain, there are advances regarding mapping land use of a campus [1], using Google Street View (GSV) imagery (due to the regular updates this dataset gets) as a way to produce updated land classification maps with lower costs with the help of a Siamese Network that combines the feature vectors resulting from feeding CNN with multiple picture from that dataset [14].

Using both satellite imagery and geotagged ground-level photographs as input, there are works suggesting a end-to-end novel DNN architecture (using VGG16) to estimate geospatial functions [15]. The architecture implemented, for each ground-level image, (G_i, l_i) that belongs to a li location, retrieves the features using a CNN (in this case VGG-16 initialized with weights for Place), $f_g(Gi)$, and for each pixel location of the aerial image there is an interpolation using NadarayaWatson kernel regression regarding the nearest ground-level images of that location, Gl, in order to accomplish a dense feature map of HxWx51 dimension that characterize both appearance and distributional information of those ground-level images. For the aerial images, using the convolution layers $\{1_{1-2}, 2_{1-2}, 3_{1-3}, 4_{1-3}, 5_{1-3}\}$ from the VGG-16 CNN model and reducing the dimensionality of the outputted feature maps, they fuse the ground-level feature map to them by applying an average pooling with a kernel size of 6x6 and a stride of 2, reducing that same feature map, and then concatenating the channels dimension with the aerial image feature map at the layer 3_3, considered by the authors, a layer that offers good trade-off between computational cost and expressiveness.

To estimate the geospatial function, $F(l(p))$ (being $l(p)$ described as location of a pixel), both feature maps, previously detailed, are re-sized to HxW using bilinear interpolation and the hypercolumn features from the conv layers $\{1_2, 2_2, 3_3, 4_3, 5_3\}$ plus the ground-level feature map are extracted, resulting in a 1043 of length hypercolumn feature. This hypercolumn feature is finally passed to a small multilayer perceptron, consisting in 3 layers with size of 512,512 and K (number of outputs for the task) and intermediate layers using a leaky ReLU activation function, will estimate the geospatial function.

III. Mapping Land Coverage with Ground Level Imagery

With the goal in mind of developing a novel model able to, based on the combination of geo-referenced volunteered ground-level photographs and aerial imagery, produce geographic maps depending on land classification characteristics, this land classification task can be divided in four main tasks that can be briefly described as follow:(1) collection and data filtering; (2) perform visibility analysis techniques that explore terrain elevation models; (3) create a land classification model with a large-scale geo-tagged photo collection; (4) production of the land classification maps.

A. Process

The steps of the process are the following:

1) Considering a discrete raster land classification map belonging to a certain geographic location produced based on a satellite image, for each regularly-spaced pixel that exists there are collected the ten most near pictures taken from the center of it. It is associated then to each pixel's center coordinate a dictionary containing the land classification value and an array of those ten images with their respective information, such as the image itself, date it was taken and the precise distance to the pixel. To find these ten nearest images for each pixel of the raster it is calculated the distance, given the

law of haversine, between the centroid of the pixel and the coordinate of each photograph belonging to a group of hundred nearest photos.

2) Alternatively to just calculate the 10 nearest photos, in this step is tested the added component of visibility analysis. Having, for each pixel of the raster the information of which pixels are visible from that point, it is possible to take in consideration the visibility limitations of the photographs when calculating the 10 nearby images for each pixel on the previous step.
3) Having the information regarding the land classification of each pixel and its associated 10 nearby pictures, a deep learning model is trained. This model is a recurrent neural networks, more in concrete, a GRU, which is fed with sequences of images' feature map (coming from using a CNN pre-trained with ImageNet), sorted by their distance to each pixel in order to associate a sequence of images to a land classification value.
   Once the model is trained, it is possible to produce land classification maps of other regions from different epochs.
4) On this last step, to produce a land classification map of a new region based on the trained model relative to step 3, it is needed to gather information for this new location. This information is collected the same way as step 1 and step 2, although, now it is relative to a new geographical area and there is no data regarding that zone classification beforehand. Having associated for each pixel in the raster a sequence of nearby images, these pixels are classified using the model that was previously trained and, using GDAL, once again, it is possible to rebuild the new raster with land classification.

B. Visibility Analysis

Independently on the location that a photography is taken, it cannot capture the whole geographic points around within an infinite radius. On this project, this technique is applied to calculate the viewshed area, with the help of a DEM, that represents the elevation of the geographical area considered. Since the land classification maps produced in the end of the whole process have a lower scale, meaning that a pixel on those maps cover a wider area than a pixel on the DEM, due to efficiency reason to calculate viewshed for each pixel, the resolution of the DEM was scaled. During this scaling process to produce the new pixels with the new height value, it is considered the average height of all the pixels of that area.

After upscaling the DEM, all the pixels containing a non relevant land classification are not considered for the viewshed analysis. By non relevant it is meant pixels without a valid land coverage classification or classified as open waters.

Once all the relevant observer points are found, the next step is responsible to produce the viewshed for each one of this observer points in order to understand which pixels and corresponding coordinates are visible. Doing this, the information which associates a coordinate to a list of seen coordinates becomes available in order to filter out the photographs that cannot capture that area. Figure 1 is visual representation of this viewshed process for one of the observer points.

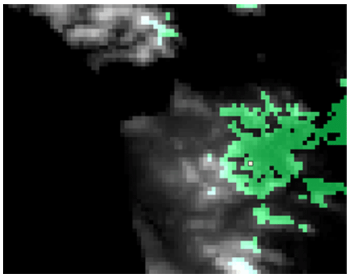


Fig. 1: Viewshed Analysis. In this image, the yellow dot represents the observer point while the green area represents the visible area of that observer point

To gather the ten nearby images of a pixel regarding the previous process, it was implemented, with the help of GDAL library, a searching function so that for each pixel, retrieves its viewshed raster file and considers, using the k-nearest neighbors algorithm from Scikit Learn, which are the closest 10 pictures that are a visible point in the raster, by verifying to which pixel in the raster the picture belongs to and then if it belongs to a pixel with the visible value.

C. Deep Learning Applied to Land Classification

This implemented novel land classification model is a combination of two models. Firstly when receiving a list of the sequence of relevant images for each pixel of a raster, it is needed to extract high-level features from those same images. For this first task, it was implemented a CNN architecture available on the high-level python Neural Network library, Keras and it was treated as a transfer learning. This CNN architecture is the MobileNet [8] which is pre-trained with ImageNet.

Once implemented the MobileNet to use the Imagenet weights it will be possible to extract the most important features of the images that exist linked to a pixel of raster. For that, all considered images are cropped to a size of 224x224, which is one of the admittable input shapes for this model. Instead of including the fully-connected layer at the top of the MobilenNet, the output feature map, with a size of 7x7x1024, will be passed to a RNN model, in this case, a Gated Recurrent Unit (GRU). The choice of a GRU is due to the fact that we are modeling data as a sequence of images' features. Each feature map, coming from the CNN, is related to an imagem present in a sequence (of 10 images). Each image is closer (or at the same distance) than the next one to the point they are all associated with.

Figure 2 represents the classification model implemented. The network is then trained as a whole using adam optimization and categorical focal loss function. The last one tries to put more training emphasis data harder to classify mostly due to having less representative elements of that class on the dataset

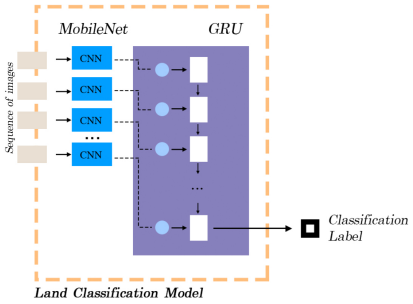


Fig. 2: Representation of the classification model implemented

Class id	Class Name	NY'05	SF'05	NY'38	SF'38
1	Water	6267	2111	6113	2120
2	Urban/Developed	19620	1760	14980	1568
3	Forrest/Agriculture/Grass	1716	615	6506	789
4	Others	63	41	83	40
Total	Sum of all classes	27682	4527	27682	4527

TABLE I: Distribution of the Land Coverage classifications among the raster pixels of both San Francisco and New York in 1938 and 2005, after grouping classes

Once the model is trained, it can be loaded and used to predict the classification of newer sequences of images or even produce land classification maps, by combining the results provided by the model and creating a new raster using GDAL.

D. Overview

As a visual overview, it can be seen on Figure 3, the whole process developed and implemented on this thesis. To note that the step 4 represents the production of a land classification map considering a new input of sequences of images from another region. This means that the model can be trained with photos obtained from a VGI. Then that data is going through all the previous process to generate the dataset "10 most nearby pictures", regarding the coordinates of each pixel on a raster and that dataset (e.g. Flickr photos). Once trained, the model can predict relatable information, such as photos from other data source (e.g. OldSF) and then, after those photos going also through the whole process previously presented (in order to associate them to pixels on a raster) it is possible to map land based on their information.

IV. Results

The goal of the following experiments is to understand the feasibility of mapping land coverage from nowadays and the past by exploring deep learning methods to ground level geo-referenced historical photos and recent photos in combination with satellite imagery.

A. Datasets and Data Processing

To evaluate the whole model implemented, it was decided to test it over two geographical regions of the United States. This happens because we are testing how it is possible to apply deep learning techniques combining ground-level photos and satellite imagery, and also possible applications of it, namely applying it to classify land coverage from previous years using historical photos and historical land classification data, and there is information regarding those topics associated with the United States, more precisely with San Francisco and New York.

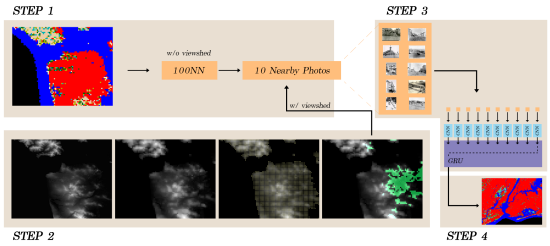


Fig. 3: Representation of the whole process implemented

Being San Francisco (SF) and New York (NY) the two geographical areas of study of the following experiments, in order to have a baseline regarding the land classification we are using datasets provided by USGS EROS researchers, the first one being the Modeled Historical Land Use and Land Cover for the Conterminous United States: 1938-1992 dataset, and the second one the Conterminous United States Land Cover Projections - 1992 to 2100 dataset. Both contain information relatively to land coverage (LC). From the first one it is used the dataset relative to 1938, which had Land Coverage(LC) information that was directly input by researchers and from the second one, the dataset relative to 2005, which was produced considering the National Land Cover Database, USGS Land Cover Trends, and US Department of Agriculture's Census of Agriculture to recreate the LC classification data of that time.

Both these datasets are rasters, with a pixel resolution of 250meter, that contain LCLU information, divided in 17 classes, from the whole US, so it was considered for these evaluation exclusively the areas within the bounding boxes relative to SF and NY, defined respectively by (37.860909, -123.055561, 37.686071, -122.365444), containing 255x65 pixels, and (40.912455, -74.25243, 40.500425, -73.712143), containing 199x152 pixels. Since there is barely information in relation to some of the 17 classes within these location, there were created some group classes. Firstly the original Class 0, 3, 4 and 5 are not considered, due to being Nan or purposely not used. The new Class 1 "belongs to water", Class 2 belongs to "Urban/Developed Area", Class 3 "Agriculture/Grass/Forest" (comprehending the classes 8 until 16 from the original dataset) and finally Class 4, "Others" (containing data relative to classes 6,7 and 17 from the original dataset). On Table I it is possible to view the distribution for this modified dataset with the relative rasters representation on Figure 4.

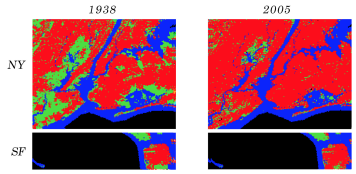


Fig. 4: USGS EROS Rasters visualization from SF and NY areas from 1938 and 2005

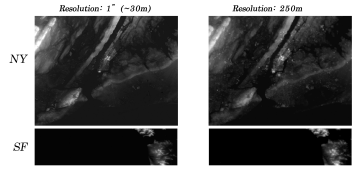


Fig. 5: Digital Surface Model withing the SF and NY boundaries, before and after the upscale

V. Evaluation

Regarding the visibility task, instead of using a DEM, we used a DSM, more specifically the version 2.1 of the ALOS Global Digital Surface Model (DSM), ALOS World 3D-30m (AW3D30) dataset, released by JAXA, containing information in a DSM about the height above sea level divided in units of area of 1 degree latitude and longitude.

Once again, since the evaluation is within the area of SF and NY, the dataset was clipped to those areas. Since the resolution of the DSM was different from the LCLU raster, I decided to upscale the first one to fit the second one's resolution by calculating how many pixels from the DSM were need to fit one pixel at the scale from the LCLU raster, and calculating the raster. A visual representation of this upscaling for each region can be seen on the Figure 5.

For the evaluation of the land classification model based on georeferenced photographs and satellite imagery, we used three datasets, that can fall into two distinct categories, being the first category characterized by containing photographs from recent years and the second one by containing historical photos.

For the first category I used Yahoo Flickr Creative Commons 100 Million dataset, developed by [16]. Regarding the locations of the experiments, this dataset has 792.349 georeferenced pictures taken in San Francisco and 1.144.239 taken in New York.

The second category contains two datasets. The OldNYC and OldSF, which contain 39.516 and 13.257 historical (from 1850-2000) georeferenced photographs for the NY area and SF area respectively.

An example of images that are in these 3 datasets can be seen on Figure 6.

A. Visibility Analysis Tests

The evaluation done was respectively to the changes the visibility analysis would affect while generating the dataset of the "10 most nearby pictures for each pixel". Just to complement the information regarding the generation of this dataset, the data that it holds, can be represented as a list of all pixels/coordinates (from USGS EROS Rasters) which have associated one of four land classification labels, a set/sequence of the 10 most nearby photos id of the chosen imagery dataset and their respective distance to the centroid of that pixel.

This process was tested over the SF area due to the computational process that implies producing a viewshed analysis for each pixel of DSM or DEM. Since SF dataset, has six times less LC categorized pixels (4527) than NY (27682), but still has 4527 to be processed, made it being the chosen region.

In order to calculate the viewshed for 1355 pixels of the SF DSM, it was used an algorithm that uses a viewshed plugin for QGIS. With the settings of observer height = 1.65m, the target height 10m and radius 20km, the viewshed analysis was performed for a portion of SF, producing a binary information (0= not visibile, 1= visibile) in relation to the pixels that were visible (or not) from 1355 tested pixels.

The viewshed is applied as a filter after calculating the 100 Nearest Neighbours pixels containing photographs to a considered pixel, as advanced before, to decide, from that group of 100 pixels, where are located the 10 nearest and visible photos relatively to those 1355 tested pixels. On Figure 7 it is possible to visualize and example about how those 10 photos are displaced in relation to the viewshed of chosen pixel that is being evaluated).

The first calculation revealed that, for those 1355 tested pixels and their respective 100NN pixels containing photos from Flickr (which gives a total of 135500 gathered pixels) 115019 pixels were visible containing photographs and 20481 were not. Using the the SF OldSF dataset, the resulting values were 60999 visible pixels containing photographs and 74501 not visible (Table II). Once having the 100NN pixels filtered down to just contain visible pixels, it was calculated the 10 most nearby pictures regarding OldSF and Flickr dataset within SF.

Afterwards it was tested how applying this viewshed technique would differentiate the choice of the 10 most nearby pictures. For this it was compared the calculation of the 10

Fig. 6: Example of images from OldNYC, OldSF and two from Fickr

TABLE II: Viewshed Analysis applied to the 100NN pixels for each SF dataset

	Dataset	
	Flickr SF	SF OldSF
Visible Pixels Within the 100NN	115.019	60.999
Not Visible Pixels Within the 100NN	20.481	74.501
Total	135500	135500

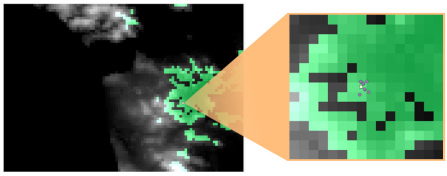


Fig. 7: Viewshed calculation for one pixel and the respective 10 most nearby photos locations (represented as purple)

TABLE III: Impact of the visibility analysis done over the 1355 pixels over the generated dataset, 10 most nearby pictures for SF

	Dataset	
	Flickr SF	SF OldSF
Pixels keeping the same 10 most nearby pictures	1220	765
Pixels changing the same 10 most nearby pictures	135	590
Total	1355	1355
Pictures keeping on 10 most nearby pictures sets	13053	9073
New Pictures on 10 most nearby pictures sets	497	4477
Total	13550	13550

most nearby pictures without viewshed techniques involved with the new one, using visibility analysis. From the 1355 pixels tested, using the OldSF dataset, 765 kept having the same 10 most nearby pictures while 590 suffered a change, which means that at least one different photograph replaced a photo from the nearby 10. When using the Flickr dataset, 1220 kept the same 10 most nearby pictures and only 135 changed at least one of them.

Since it was considered a change if one of the 10 most near pictures was replaced (regarding the tested pixels), it was tabulated how many pictures were replaced and how many were kept being considered one of those 10 pictures for each one of the tested pixels (13550 in total). For the OldSF, 9073 photos kept being the same on the 10 most nearby pictures, while 4477 possibly replaced a "not visibile" picture, while for the Flickr dataset, the values were 13053 kept being the same and 497 were pictures that were actually not visible from the its associated pixel. Both these two last results are displayed on the Table III.

As it can be seen from the results displayed on the Table II and Table III, when applying the visibility analysis during the computation of the 10 most nearby photos of SF using the historic/OldSF dataset there are more modifications regarding the set's content in comparison when using the Flickr dataset. This happens due to the different distributions of the average distance of the 10 most nearby photos on both datasets. For SF, (region tested) with Flickr dataset, the 10 most nearby images of a pixel are around 94% of the times less than 1km away, while using the OldSF, this value is around 47% (Figure 8). This information, combined to the fact that, usually, the closer a pixel x, in a DSM or DEM, is to an observer point pixel o, when applying a viewshed from that observer point, more probability the pixel $\mathbf{x}$ has to be visible from $\mathbf{o}$, explain these results.

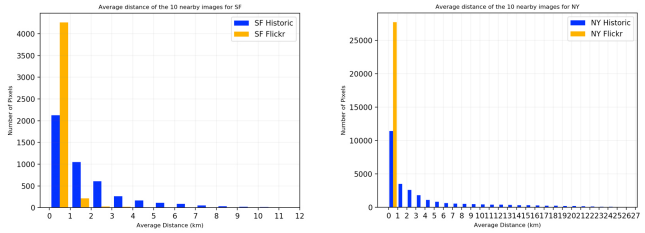


Fig. 8: Average Distance of the 10 most nearby photos from every pixel for each imagery dataset using each location

B. Land Classification Model Evaluation

Considering these generated datasets of the "10 most nearby (historical or flickr) photos" for each region, the land classification model and posterior land maps generation are evaluated in this section, mainly in terms of precision, recall, F1 score per class, and accuracy.

Either the training phase and the tests were conducted under the following computational conditions: CPU Intel Core i7 6700 CPU (3.4 GHz); NVIDIA GeForce GTX 980 GPU (4GB); 16GB of RAM.

In these carried out tests, two models were pre-trained. One using the information relative to the "10 most nearby historical photos" (which as said before, can be seen as a list of pixels, their coordinates, their LC class value and information related to its 10 most nearby pictures) and the other one using the generated dataset of the "10 most nearby flickr photos". Both San Francisco area, which means that the training samples for these tests are composed by pixels/coordinates, land coverage classification and photos from SF .

These models were trained using batches of 3 pixels. For each pixel, and due to memory problems, they were chosen the 5 most nearby pictures (instead of 10) to be inputted, sorted by the distance to the pixel, on a MobileNet to retrieve its feature maps from the last layer, and followed by feeding a GRU with those feature maps and trained according to the LC value that the sequence was label. The training was set up with batches of 3 sequences of 5 images due to memory limitations.

The models were trained during 30 epochs and there were considered data augmentation techniques, such as flipping horizontally the images, in order generate more possible sequences. For each one of these models, there were preformed three tests, using the following input/test data (referent to a geographically location that the model did not now about) to classify:

- Generated dataset of the 5 most nearby historical photos of New York;
- Generated dataset of the 5 most nearby historical photos colorized of New York;
- Generated dataset of the 5 most nearby flickr photos of New York.

Relatively to the tests, when using the NY historical photos colorization, it being is used an adaptation of the deep-koalarization, implemented by Nuno Ramanlal, in order to

TABLE IV: Model Pre-trained with historical photographs from OldSF. (P=Precision, R=Recall, F1= F1-Score, Acc=Accuracy)

	Testing Datasets								
	NY OldNYC			NY ColordOldNYC			NY Flickr		
C	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
1	0.20	0.25	0.22	0.19	0.15	0.17	0.20	0.43	0.27
2	0.54	0.54	0.54	0.54	0.75	0.63	0.69	0.41	0.51
3	0.21	0.14	0.17	0.18	0.05	0.07	0.07	0.09	0.08
4	0.00	0.02	0.01	0.02	0.04	0.02	0.01	0.05	0.01
Acc		0.383			0.45			0.39	

colorize the historical photos and posterior generation of the 5 most nearby historical photos colorized of NY dataset.

1) Results for the model pre-trained with historical photos from SF (1938): This section presents the results obtained by pre-trainning the classification model with sequences of 5 historical photos of SF, from OldSF, in order to predict/map land classification of NY in three ways: using sequences of 5 images regarding historical photos, colorized historical photos and also Flickr photos.

Table IV reports the obtained experimental results (in decimal values) regarding the accuracy of the prediction from each test and also the precision, recall and F1-score associated to the each class (of the defined 4) prediction from each test.

The results obtained, from each one the three tests, demonstrate that the classification model obtained the best results classifying sequences of images with class 2 (Urban/Developed) on the correct pixel, achieving the highest values for the precision, recall and F1-score. This can be explaining because of the unbalance of the classes presented on the training phase. As seen on Table I, the amount of pixels labeled as Urban/Developed in comparison to the other three classes is way bigger. Not even using the data augmentation process or the usage of the focal-loss loss function in order to consider more weight for the least used classification labels, made the other classes obtain a better values according to this metrics.

Figure 9 shows how the classifier distributed the land coverage labels among the pixels of the rasters. When trying to predict the land coverage map using historical photos non-colorized, it is possible to see a similar distribution of pixels falling in each of the 4 classes comparing to the baseline. When using the colorized ones, there was an increase of amount of pixels classified as developed. This resulted in a higher accuracy of the model since there were more pixels being accurately labeled with the class 2, which is the class that will cause more impact in the results due to the sample number it has associated. This, made this land map being the most accurate even though its discrepancy of pixels associated to each class compared to the baseline. When using the NY flickr photos to predict the land map, the model had problems differentiating between the class 1 and 2, most likely due to the different content the flickr and historical photos contain on themselves and their location, which resulted in bad class

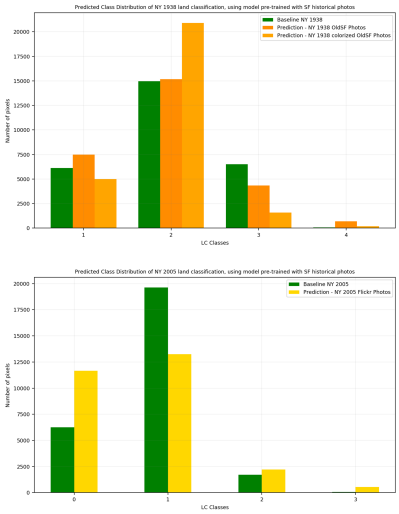


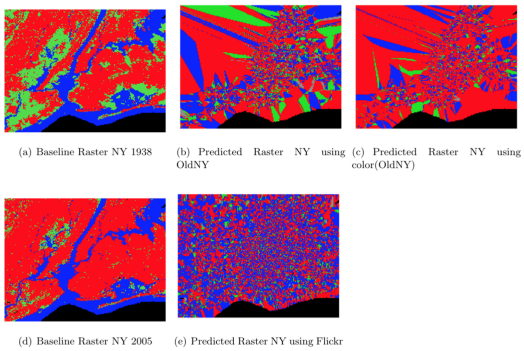
Fig. 9: Land classification distribution comparison between the Real NY 1938 (top) and NY 2005 (bootom) data and the predicted data, using the model pre-trained with SF historical photos

distribution of this prediction. flickr photos contain pictures of objects interiors and pictures of the sea and rivers, which are new to a model trained mostly with black and white/yellow photos of buildings and streets.

Nevertheless, based on these results, the whole model generated the land classification rasters presented on Figure 10. As it can be seen, the precision of the pixel classification follows the values presented on the Table IV.

2) Results for the model pre-trained with Flickr photos from SF (2005): Here are revealed the experimental results gathered by pre-trainning the classification model with sequences of 5 nowadays photos of SF, from flickr, in order to predict the land classification of NY using sequences of 5 images (historical photos, colorized historical photos and Flickr photos).

The predicted outputted classification land maps of NY are shown on Figure 12 and it is visible the predominance of



(a) Baseline Raster NY 1938 (b) Predicted Raster NY using OldNY (c) Predicted Raster NY using color(OldNY)

(d) Baseline Raster NY 2005 (e) Predicted Raster NY using Flickr

Fig. 10: Land Classification Maps generated with the developed model, regarding the pre-train using OldSF geotagged ground-level photographs

TABLE V: Model Pre-trained with flickr photographs from SF. (P=Precision, R=Recall, F1= F1-Score, Acc=Accuracy)

	Testing Datasets								
	NY OldNYC			NY ColorOldNYC			NY Flickr		
C	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
1	0.26	0.17	0.20	0.24	0.15	0.19	0.23	0.19	0.21
2	0.55	0.83	0.66	0.55	0.85	0.67	0.70	0.70	0.70
3	0.25	0.03	0.05	0.25	0.02	0.03	0.07	0.1	0.08
4	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.02	0.00
Acc		0.49			0.50			0.55	

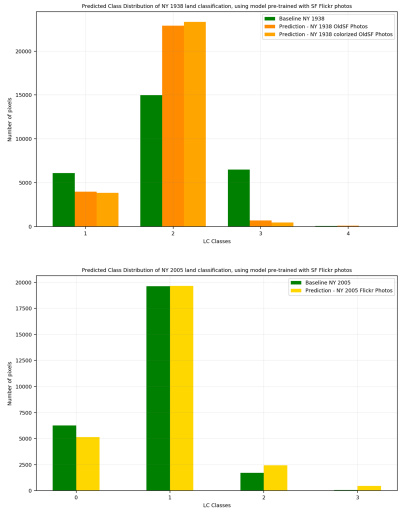


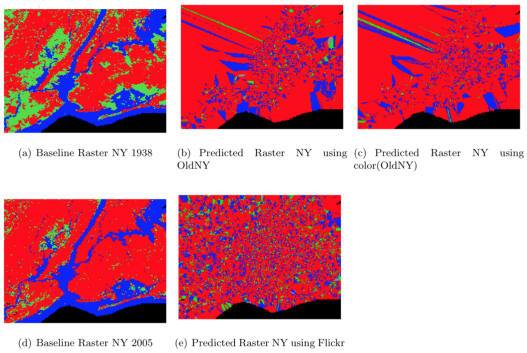
Fig. 11: Land classification distribution comparison between the Real NY 1938 (top) and NY 2005 (bootom) data and the predicted data, using the model pre-trained with SF flickr photos

the colours blue and red, associated with the classes 1 and 2 respectively.

As seen on the previous experimental test, the same happens with the results displayed on Table V obtained from testing over the pre-trained model with flickr images of SF. The class 2 is again the class achieving the best scores (reaching a F1 score achieving 70% when classifying land coverage using colorized historical photos), comparing with the other classes, most likely the because of excessive amount of training elements associated with that label.

Although the class 2 has achieved best results, now it is possible to see that in some cases, class 3 or 4, obtained no true positives when predicting correctly a pixel with the real classifying label. The difference between the dataset of the "5 most nearby flickr photos of SF" and the used on the previous tests, is that it considers the land coverage raster equivalent to SF 2005, which contains an even greater unbalance of weight classes, since 83.67% pixels considered on the generated "5 most nearby flickr photos of SF" dataset fall into either class 1 or 2, due to the urbanization increase over the last century.

When predicting the land coverage maps of NY 1938 using historical photos, the model achieved slightly less accurate and



(a) Baseline Raster NY 1938 (b) Predicted Raster NY using OldNY (c) Predicted Raster NY using color(OldNY)

(d) Baseline Raster NY 2005 (e) Predicted Raster NY using Flickr

Fig. 12: Land Classification Maps generated with the developed model, regarding the pre-train using nowadays geotagged ground-level photographs from Flickr

precise than when using flickr photos, that achieved 55% of accuracy. This happens due to characteristics and of each type of picture in each dataset. Even colorizing historical photos, the colorization is not the same as the recent photos that exist in flickr, although it increase 1% of the accuracy of the model but it is not as precise as when using the flickr dataset to predict.

The resulting distribution of the predicted classes within the NY applying respectively historical, colorized historical and flickr images to the deep learning model developed can be found on Figure 11, and it shows clearly the unbalanced classification of the previous 2 classes mentioned, even though it has miss classified correctly other pixels (false positives).

When looking at the predictions done with the historical photos, the class distribution is much more unbalanced than when using the flickr, which follows the baseline distribution of its associated land map.

This model, since was trained with recent photos (Flickr San Francisco), shows how a reproduction of land classification from the past, as tested when predicting using photos of NY from OldSF and their colorized versions, can be done, even though could not precisely predict each pixel from the raster.

VI. Conclusion

As the common accessible devices that are able to take pictures, such as smartphones or digital camera, keep evolving and improving their image quality by allowing us to capture the world around us in a greater detail and resolution, combining to the fact that social networks, like Flickr, Instagram and others, are keep getting flooded with these pictures (which some time they can georeferenced), and the emergence of some projects that disclose geotagged imagery datasets from the past, it can be interesting to understand how it can ease and reduce costs of certain tasks that usually can be only accomplished using expensive resources, including money, time, equipment or even human workload, such as Land Classification.

This master thesis investigates how feasible is to apply deep learning methods, such as convolutional neural networks

and recurrent neural networks, to a combination of freely available ground-level photographs from different VGI (e.g. Flickr, OldNY, OldSF) with satellite imagery (providing terrain information) to classify land, either from actual years and specially from the past, in which satellite imagery to was not available but still there were ground-level image records.

As it can be seen on the evaluation Section, the results regarding how the model, that combines satelitte imagery and ground-level photographs, can precisely classify a pixel are not the best, mainly due to the unbalance of the trained samples. One of the problem has to do with the pollution that the datasets contain.

Although it was shown that using photographs from one region to train the model to classify other using its photos is possible to achieve some reasonable results regarding the land coverage class distribution and accuracy within the tested region.

In the end, this work concluded that this novel approach that considers both satellite imagery, visibility analysis and deep learning methods, treating a pixel as a sequence of nearby pictures to it, is still a demanding task, although it is possible to achieve some reasonable results even using VGI datasets without its content filtered, deep learning models that are used on mobiles (MobileNet). With some improvements either regarding computational problems or in terms of data filtering in relation to training samples it is possible to accomplish good results while predicting land classification from the present or the past.

A. Achievements

The biggest achievement of this thesis was the implementation of a novel approach, that considers deep learning methods and combines two different types of images, Ground-level photographs (obtained from VGI and datasets such as OldSF) and satellite imagery to classify land coverage. Specially the possibility of the model to classify land from the past, when it was not possible to visualize the terrain from satellites, is the biggest achievement of this work.

It also considers techniques, such as visibility analysis that could improve the results regarding this classification task.

B. Future Work

Regarding the results obtained, there are several possible ways of improvement that can possibly raise the accuracy of the applied models in classifying precisely each pixel within a land classification raster. As future work, the model can be improved and continued by addressing the following points:

- On the initial phase of data processing, the photos should be submitted over a filtering process, in order to consider only photos revealing land classification features (specially regarding the flickr dataset that contains images about everything);
- Build the processed dataset of nearby photos for each pixel in the raster considering also a temporal distance, in order to precisely attribute to a sequence of images from a given year the exact classification of that year ;
- Train and test the model under a more computational powerful machine, allowing a bigger sequence of nearby images to be considered for each pixel;
- Even though it requires more memory, consider DEM and Land Classification rasters with smaller scale, in order to precisely collect the most nearby photographs that truly correspond to a pixel.

REFERENCES

[1] Y. Zhu and S. Newsam, "Land use classification using convolutional neural networks applied to ground-level images," in *In Proceedings of the SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2015.

[2] G. Foody, L. See, S. Fritz, P. Mooney, C. Fonte, V. Antoniou, G. Foody, L. See, and S. Fritz, *Mapping and Citizen Sensor*. Ubiquity Press, sep 2017.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *In International Journal of Computer Vision*, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] S. Workman, R. Souvenir, and N. Jacobs, "Understanding and mapping natural beauty," in *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *In IEEE*, 1998.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017. [Online]. Available: http://arxiv.org/abs/1704.04861

[9] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *In Proceedings of the European Conference on Computer Vision*. Springer, 2016.

[10] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning Representations for Automatic Colorization," 2012. [Online]. Available: https://arxiv.org/pdf/1603.06668v2.pdf

[11] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *arXiv preprint arXiv:1711.09020*, 2017.

[12] V. Antoniou, C. C. Fonte, L. See, J. Estima, J. J. Arsanjani, F. Lupia, M. Minghini, G. Foody, and S. Fritz, "Investigating the feasibility of geo-tagged photographs as sources of land cover input data," *ISPRS International Journal of Geo-Information*, 2016.

[13] H. Xing, Y. Meng, D. Hou, F. Cao, and H. Xu, "Exploring point-of-interest data from social media for artificial surface validation with decision trees," *In International Journal of Remote Sensing*, 2017.

[14] S. Srivastava, J. E. Vargas Muñoz, S. Lobry, and D. Tuia, "International Journal of Geographical Information Science Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data," 2018. [Online]. Available: http://www.tandfonline.com/action/journalInformation?journalCode=tgis20

[15] S. Workman, M. Zhai, D. J. Crandall, and N. Jacobs, "A unified model for near and remote sensing," in *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, 2016.