POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Transformer-based Geolocation of Indoor Images with Segmentation-based Visual Explanations

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Imanuel Rozenberg

Advisor: Prof. Mark James Carman

Academic year: 2020-2021

1. Introduction

The need to geolocate images of indoor scenes arises mainly from the need to help police forensic investigations with a tool capable of identifying, or at least narrowing the research when dealing with human trafficking material. Internet technologies such as P2P networks or Instant Messaging applications allow to freely exchange in an encrypted manner harmful and sensitive material. Computer forensics experts often come in possession of such material by direct retrieval from a criminal's device or by reports of the Internet Watch Foundation. Unfortunately, in many cases it is very hard to identify the location of the committed crimes in order to intervene and prosecute. An additional problem for which it might be useful to geolocate interior images is related to fake news, or even fake ads on websites as Airbnb and Booking, with images from places different from the one being referenced. With the advent of deep learning the ability to solve image classification problems improved drastically, thus allowing also to explore the capability of Neural Networks to perform content-based indoor geolocation. We argue that there are many decorative and structural patterns that are strictly correlated to the geographical location and that it is possible to learn these features which are helpful to distinguish one location from another. The aim of this work is to develop a tool to facilitate geolocating indoor pictures, providing a list of predictions at different granularity levels, such as: Continent, Country and City. In addition, to make more useful and interpretable the predictions of our model, we also employed some explanatory techniques capable of highlighting discriminating information in the geolocation process.

2. Background

Most of the image geolocation models focus on outdoor images as iconic monuments, different architectural and urban styles, but also landscapes scenery, may be immediately recognizable, while a more generic interior scene may be extremely difficult to locate. With the advent of deep learning, most works approach the geolocation problem as a classification problem, dividing the world map into a set of cells of different size according to the desired granularity level, each corresponding to a different class. Two of the most important state-of-the-art architectures for geolocation are: PlaNet [9] and Müller-Budack et al. [6]. The first model, based on an Inception-based architecture, using S2 Google's algorithm generates the labels by partitioning

the surface of earth into thousands of multi-scale geographic cells.
The Novelty introduced by MüllerBudack et al. approach consists of incorporating hierarchical knowledge at different spatial resolutions, plus extracting the depicted scene which the geolocator could potentially benefit from. In fact, in an outdoor geolocation task when trying to identify the correct city, urban scenes with people, sidewalks and buildings could be very informative, while at country level we deal with a wider variety of scenary like mountains, flora and beaches. Stylianou et al. in [7] were the pioneers in the attempt to geolocate indoor images. They created a huge dataset of hotel images in order to geolocate hotel rooms across the world by trying to predict both the hotel instance and the hotel chain. We decided to adopt the geolocation by classification approach, focusing on the indoor geolocation task.

3. Research Questions

In this thesis we address the following questions:

1. Is it possible to develop a geolocation model, which could be useful in applications such as law enforcement, capable of geolocating indoor scenes at different levels of granularity with a meaningful level of accuracy?
2. Is it possible to provide meaningful and useful explanations for the predictions given by the deployed deep learning model?
3. What are the most appropriate backbone and dataset to adopt for the feature extractor of an indoor geolocation model?

4. Approach and Model architecture

In 2021 a silent revolution started to happen in the image classification world, with the ViT[2] which adopted a pure transformer for computer vision tasks, showing that the reliance on CNNs is not necessary. After a series of experiments which will be presented in *Chapter 6*, we decided to use the Swin Transformer [4] as image encoder. The Swin Transformer is a vision Transformer, developed by Microsoft Research Asia, which performs local self attention over non overlapping, shifting windows and then builds hierarchical feature maps by merging image patches in deeper layers. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. We then connect to our image encoder a Multi-level Classifier which consists of a fully connected network made for multiple parallel classifications at different levels: world sub-region, country and city level. Moreover, we also add a block with a Neural Network (Deeplabv3+ [1]) for multi-class semantic segmentation capable of identifying and segmenting numerous objects present in interior scenes. The segmentation block turns to be very helpful especially when exploited to partition the image to be analysed with SHAP [5] algorithm, in order to produce visual explanations of the predicted outcome. The whole framework is shown in Fig. 1. This approach improves the performances of SHAP, in our framework, in terms of computational cost and interpretability, being able to assign Shapley values directly to furniture and structural elements with semantic value like: window, floor or ceiling. A deeper detailed description of the results obtained by our approach with SHAP with various visual experiments is presented in Chapter 6.

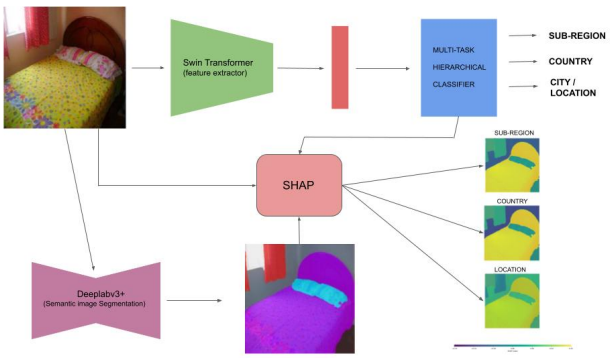

Figure 1: Overall architecture of our Hierarchical Indoor Geolocation model.

5. Dataset

We adopted two different dataset to train our model: the Hotel50k dataset presented by Stylianou et al. and the Airbnb dataset; in addition we created a small dataset for testing the inference performance of our model by scraping from different web sites of real estate agencies. The Hotel50K dataset consists of 1.3 mln images of hotel indoors belonging to more than 45k ho-

tels. The images are collected from two different websites: Expedia and TraffickCam. Expedia is a travel website showing images of hotels, while the latter is a website displaying images uploaded by users, usually the images are low quality and with occasional clutter obstructing the scene. This dataset is particularly useful for law enforcement in fighting human trafficking. The Airbnb dataset is the result of a collection of datasets provided by Inside Airbnb web-page, related to 100 different locations, which vary from big, densely populated cities (e.g. New York, Rome ...) or regions like (e.g. Barossa Valley, Western Australia ...). In addition, in order to use only interior images, a scene classifier was used to filter the Airbnb dataset. The last dataset is a collection of images that we gathered from data provided by many different real estate agency web-sites, in order to test the model inference capabilities on our demo. This dataset contains pictures coming from 15 different countries which we select according to different types of power plugs in the respective countries.

6. Experimental Results and Evaluations

This section is dedicated to the different experiments and evaluations that we did during the development of this Master Thesis. We first introduce our framework setup and then we present the experiments.

6.1. Setup

All the experiments were trained via Cross-validation. In order to make the experiments reproducible we set a random seed (13) for all the experiments. Both datasets (Hotel50k and Airbnb) were split into 90% training and 10% testing. Finally the training set was further split into 80% training and 20% validation. We initially run our model on a random data generator, but later, because of class imbalances, we decided to adopt a country conditioned sampling. We decided to adopt batches of 16 images due to the dimension of the dataset and memory constraints. The models optimize a Sparse Categorical Crossentropy via the Adam optimizer. We used Early Stopping condition to avoid overfitting, setting the patience to 3. After many experiments we decided to set the Droput to 0.1 and Learning Rate to 1e-4. In order to compare and evaluate the performances of different models we used three different metrics: the loss function, prediction accuracy and Improvment over Majority class (IoMc) for each level (sub-region, country and city). The IoMc measures the improvement of our model with respect to a classifier that always assigns to any image the most frequent sub-region, country and city in the dataset, and it is defined as follows:
The IoMc is defined as follows:

$$IoMc = \frac{Accuracy_{valid}}{Prob_{Majority\_class}} \tag{1}$$

6.2. Experiments

We present here the different experiments carried out to choose the best setting for solving the indoor geolocation problem. The entire framework was developed using Pytorch Lightning.
Experiment A - Backbone architecture. We compared different state-of-the art CNN models (VGG-19, EfficientNetB0) with some of the most recent vision Transformers as the Swin Transformer over the same classification task. The result shows that the Swin Transformer is the most suitable backbone for our model, showing the best performances among the tested networks in different settings as Single vs Multi-task Learning. Once adopted the Swin Transformer as model backbone we tested different configurations according to the different datasets (Hotel50k and Airbnb) and to the limited memory resources.
Experiment B - Pretraining on ImageNet22k vs ADE20k. Usually most of the model backbones used for Image classification are pretrained on ImageNet1k or ImageNet22k, since in our task we are geolocating images according to interior scenes we thought that it could be interesting to use as a backbone the Swin Transformer pretrained on the ADE20k [10] which is a dataset, meant for segmentation, of annotated images covering many scene categories as interior of houses.
Experiment C - Modifying the number of fully connected layers. The Hotel50K dataset presents many more labels for the city category (10458) than the Airbnb dataset (100 different cities). We have noticed that while we reach meaningful accuracy values on other categories: sub-region and country, in order to improve pre-

dictions at city level we need to modify our baseline classifier, using more parameters for the city predictor, while keeping the other classifier unchanged. We have also proven that adding an ulterior head for the prediction of the Hotel chain boosts the performances of our model on all other tasks. In Tab 2 and Tab 1 we present the results, in terms of accuracy, achieved by our model on both Hotel50K and Airbnb dataset.

	Top-1	Top-5	Top-10
Region Accuracy	0.705	0.966	0.998
Country Accuracy	0.583	0.880	0.953
Location Accuracy	0.389	0.695	0.814

Table 1: Classification accuracy on the Airbnb dataset, after 21 epochs, at sub-region, country and location level. Top-k indicates that correct location was predicted amongst the first k results.

	Top-1	Top-5	Top-10
Region Accuracy	0.691	0.922	0.979
Country Accuracy	0.597	0.814	0.890
City Accuracy	0.213	0.364	0.423
Chain Accuracy	0.696	0.864	0.922

Table 2: Classification accuracy on the Hotel50k dataset, after 37 epochs, at sub-region, country and city level. Top-k indicates that correct location was predicted amongst the first k results.

Experiment D: Comparing Integrated Gradients and Grad-CAM based Explanations. In our first attempt to generate useful visual explanations highlighting the features which mostly impacted on the predictions of our model we decided to investigate Integrated Gradients [8]: a model agnostic, class agnostic gradient based method. Integrated gradients is a method which computes the multiplication of the gradient with respect to the inputs and their derivatives in order to produce visual explanations at pixel level. We applied this technique on the three different granularity predictors, see Fig. 2, of our model, producing different saliency maps for each level of granularity allowing to understand also which are most informative for each corresponding geographical level of granularity.

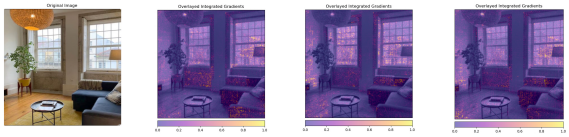

Figure 2: The picture on the left is taken from the Airbnb dataset, it shows a living room of an airbnb located in Porto, (Portugal). On its right the attributions given respectively at sub-region, country and city level by Integrated Gradients method.

We then decided to make a comparison between the attribution produced by Integrated Gradients and the heatmaps produced by Grad-CAM, for each geographical granularity level. Grad-CAM is one of the most commonly used methods for visualizing which kind of information a CNN model is using when making a prediction. One of the main advantages, with respect to Integrated Gradients, is that Grad-CAM is a class-specific method which computes the gradients of the target output with respect to a given layer, thus producing different heatmaps according to the targeted class. Grad-CAM was originally thought only for CNN: this means that, in order to produce meaningful heatmaps on the Swin Transformer it requires some adaptations. The main idea is to treat the output of the last attention layer before the classification head as the designed feature map in case of a CNN. An example showing the heatmaps produced by Grad-CAM is shown in Fig. 3.

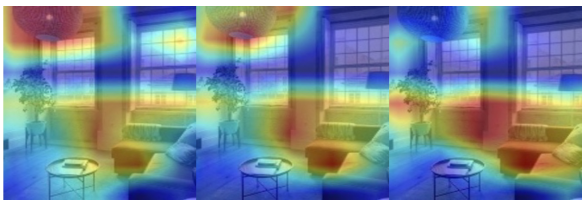

Figure 3: The analysed picture is taken from the Airbnb dataset, it shows a living room located in Porto, (Portugal). Heatmaps highlight most influential areas at sub-region, country and city level.

Experiment E: Comparing Segmentation Methods for SHAP-based Explanations. When producing visual explanations of the predictions of our model, we would like to blame single objects or elements instead of coarse areas. For this reason we investigated an approach

which exploits the semantic segmentation of indoor scenes produced by our model and combines it with SHAP [5]. SHAP is an algorithm capable of extracting both: **intrinsic content** of the image (some parts of the image are more important than others) and **discriminative information** (some parts of the image are useful for distinguishing the classes), fairly distributing the prediction among the learned features. SHAP belongs to the methods based on coalition game theory Shapley value, which have a solid theoretical justification although they are computationally expensive. When used to interpret predictions of image classification models, usually the procedure consists in first partitioning the analysed image into a fixed set of superpixels, and then run SHAP algorithm on the superpixels instead of single pixels, for efficiency reasons. The explanations are thus given with respect to the superpixel partition. Unfortunately when dividing an indoor scene image into superpixels we don't have a direct correspondence between superpixels and single objects and elements present in the image. For this reason we decided to use the segmentation produced by our model instead of random superpixels. This approach leads to more interpretable visual explanations. In Fig. 4 we present the results of the algorithm on the top-3 most probable sub-regions. Running SHAP on top-3 most probable sub-regions, countries and cities shows which elements had the greatest influence on the prediction of each class. In this way, in case the correct prediction is in the top-3 predicted classes we are able to see which elements have contributed the most to that prediction, and at the same time also to point out the ones that may have led our model to incorrect predictions.

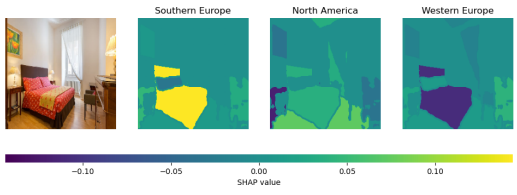

Figure 4: SHAP values on top-3 predictions at sub-region level of an Airbnb located in Rome, using the segmentation provided by our model.

Analysis A: Model Calibration.
This analysis was conducted exclusively on the model trained on the Airbnb dataset, to verify its adequacy before its deployment, as we will see in Chapter 7, in a web demonstration. Interpreting the probability scores of a deep learning model for multi-class classification as the probability that the corresponding class was detected is correct only if the model is calibrated. For this reason in order to interpret the prediction scores of our model as probabilities of class detection we need to check that our model is calibrated. The metrics and techniques that we adopted, **Expected Calibration Error** and **Reliability diagrams**, to visualize the calibration of the model were introduced by Guo Chuan et al. in [3]. After having proven that our model is well calibrated, see Fig. 5, we also inspected the **confusion matrix** for the three classification tasks (sub-region, country and location) in order to check if the model is biased toward a specific sub-region, country or location.

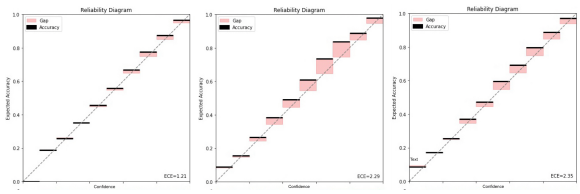

Figure 5: Reliability diagrams respectively for sub-region, country and location classifiers, showing the average confidence for each bin, as well as the accuracy of the examples within each bin.

7. Qualitative Analysis with a Demonstration Application.

In order to test the model on new user-uploaded images we build a demo using Gradio which is an interface that can wrap almost any Python developed machine or deep learning pretrained model with an easy-to-use user interface. The demo also incorporates a visual explanation of the prediction using a Shapley-based Interpreter. We claim that a qualitative evaluation through user study would be a reasonable tool to obtain an additional qualitative evaluation of our model. In this demonstration website, we decided to deploy our model for indoor geolocation trained on the Airbnb dataset, displaying the top-10 country predictions as in Fig. 6.
Thanks to this demo, we tested the inference capabilities of our model on many images com-

ing from different real estate web pages, but also photos taken inside student, workers and family houses. This allowed us to verify surprising generalisation capabilities of our model, especially on data from European interiors.

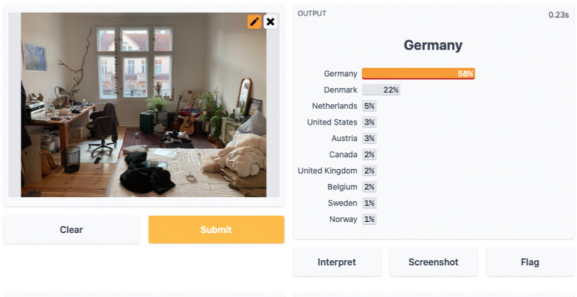

Figure 6: Top-10 predicted countries on a photo taken inside of a bedroom and a kitchen in Berlin.

8. Conclusions and Future Research Directions

In this final chapter we discuss the conclusion we have come to, through this thesis work, regarding the questions that we posed in *Chapter 3*. In particular, we focused on the indoor geolocation problem exploiting a set of deep learning techniques to correctly geolocate indoor's scene images, at different level of granularity, and give visual explanations about the predictions. The results achieved by our model, both Hotel50k and Airbnb dataset, are quite surprising, reaching a rather high level of accuracy not only at sub-region and country level but also at city level. Moreover we were also able to produce meaningful visual explanations which help to better understand the predictions and to add informative elements to our knowledge for the indoor geolocation task. To conclude, we suggest some possible directions for future developments of the project.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021.

[5] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[6] Eric Müller-Budack, Kader Pustu-Iren and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. *ECCV (12)*, 11216 of *Lecture Notes in Computer Science* :575–592, September 2018.

[7] Abby Stylianou, Hong Xuan, Maya Shende, Jonathan Brandt, Richard Souvenir, and Robert Pless. Hotels-50k: A global hotel recognition dataset. *AAAI, The AAAI Conference on Artificial Intelligence*, 2019.

[8] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017.

[9] Ilya Kostrikov Tobias Weyand and James Philbin. Planet - photo geolocation with convolutional neural networks. *ECCV, European Conference on Computer Vision*, 2016.

[10] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *CoRR*, abs/1608.05442, 2016.