# Summary of progress on image-geolocation model

This document outlines the current planned pipeline for the image-geolocation model part of the Multi-Modal Multi-Lingual Location Extraction Model (Multi-LM). The textual geoparsing model is separate to this and is not considered in this document.

## Proposed pipeline

Our proposed pipeline consists of two main models. First, a broad-location model is used to estimate the location in which an image was taken, to within some broad region. Once the broad location is estimated, we use a second model which uses features from within the image, such as text and points of interest, along with the estimate of the first model, to provide a more precise location estimate.

A diagram of a model

Description automatically generated

Figure 1: Proposed pipeline for image location estimation using broad location and precise location models.

## Broad location model

Initial experiments have indicated that modern convolutional neural networks perform well on simple geolocation tasks. Figure 2 shows the confusion matrix for an EfficientNet (Tan & Le, 2019) model trained to predict the country associated with a Google Street View image, from a pool of 10 potential countries. While this is a promising indication of the ability of CNNs to extract locationally important features, it is not scalable to the Multi-LM project and has biases due to country sizes and data availability.

We propose a query-based model, in which an image, , is compared to a set of test co-ordinates, , and the model aims to estimate some function of the distance between the test co-ordinates and the true location of the image, . The model structure is outlined in figure 3.

A graph with blue squares and white text

Description automatically generated

Figure 2: Confusion matrix for the EfficientNetv2 model, trained to predict the country in which a Google Street View image was taken, from a list of 10 potential countries.

A diagram of a network

Description automatically generatedAt inference, we provide the model with a single image and a set of standard test coordinates, which are representative a wide range of potential locations. We can then test more co-ordinates around those which the model thinks are ‘closest’ to the true co-ordinates, building up a heat map indicating the broad location of the image.

Figure 3: structure for the query based broad location estimation model, in which a transformation, , of the distance between the true location, , and a test location, , for an image,

This approach relies on the model being able to run quickly at inference, as many test co-ordinates will be needed to properly estimate the location of the image. However, it means we are not relying on a categorical approach, which would require either large amounts of data for each region, or will introduce regional biases.

## Precise location model

Once the broad location of the image is known, we can use other features in the image to get a more precise location estimate. Initially, we have been focussing on using textual information in images to do this, however later work will look at point-of-interest information, and features which are projectable onto a map, such as road and building layouts.

### text extraction

Text extraction consists of two steps – text detection and text recognition. We leverage the Yolo v7 model (Wang, Bochkovskiy, & Liao, 2022), fine-tuned on the Coco-text v2 dataset (Veit, Matera, Neumann, Matas, & Belongie, 2016), to do text detection, and a TrOCR model (Li, Lv, Chen, & Cui, 2022) to convert the identified text into string data. Figure 4 shows a typical output from the text detection model.

A storefronts of a store

Description automatically generated

Figure 4: Text detection using Yolo v7 model.

The model is able to extract Latin alphabet text well, including text across various scales. It is currently unable to detect text written in non-Latin characters due to limitations in the training data, however we are currently looking for a dataset which will help to address this. The model also struggles with text which has been rotated, and frequently splits phrases into separate words. This can be a problem with street names, which often contain multiple words.

Once our text extraction model which is working well, we will build a pipeline which can use textual information within the image, along with some prior knowledge of the initial location estimate, to better geolocate the image. This will work alongside a point-of-interest based model, which is yet to be developed.

## Datasets

We are currently using Google Street View as a main data source for this work, as it can provide consistent images which are highly localisable and cover a wide variety of places. Once we have a proof-of-concept model working with this data, we will aim to acquire geotagged image data from social media sources such as Flickr, or from online news.

Currently, we are only using images which are taken in urban or semi-urban areas, as these tend to have more identifiable features, especially for the precise geolocation model. Further, we are only considering images which are taken outside, as interior images are too difficult to geolocate.

## Timeline

|  |  |
| --- | --- |
| **Task** | **Estimated completion date** |
| Collection of geotagged Google Streetview dataset | **Early May 2024** |
| Broad location model - development, testing and tuning | **June 2024** |
| Precise model – text extraction | **June 2024** |
| Precise model – points of interest | **July 2024** |
| Integration of broad and precise model | **August 2024** |
| Integration with geoparsing model | **August 2024** |

The expected timeline for the rest of the project is below. We are currently recruiting for two full-time members of staff to help with this project.

02/05/2024

Joseph Shingleton

Geospatial Data Science Group

University of Glasgow

# Bibliography

Li, M., Lv, T., Chen, J., & Cui, L. (2022). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. *ArXiv*.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ArXiv*.

Veit, A., Matera, T., Neumann, L., Matas, J., & Belongie, S. (2016). COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. *ArXiv*.

Wang, C.-L., Bochkovskiy, A., & Liao, H.-Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object. *ArXiv*.