Image Geolocalization with Directed Feature Identification

Claire Gong, Madeline Brumley

Motivation

- **Image geolocalization** is known to be a challenging task for vision models, but human guessers can do very well
- Dataset construction for task is challenging: **hard to represent every region**
- Most high-performing models require **lots of data to train**
- Human guessers achieve success by identifying **characteristic features** in an image such as bollards or license plates
- Can models perform better and extract better representations from data by **attending to extracted features** like humans do?

Experimental Setup

- We fine-tune pretrained ViT-B-16 models on **raw images** from G50 and G3 datasets to use as baselines.
- We **construct feature-to-country datasets** from G50 and G3 using Facebook's Mask2Former model for instance segmentation.

Expert Feature Classifiers:

- We group feature classes from the G50 feature-to-country dataset into **4 main groups:**
  - Ground (road, terrain)
  - Vegetation (foliage)
  - Structures (buildings)
  - Vehicles (cars, trucks, etc.)
- We then split each group into test/train splits & fine-tune a ViT-B-16 model **for each group** for feature-to-country classification.
- We evaluate each model on their respective test splits by **top-1 percent labeling accuracy.**
- We then assemble the model ensemble with Mask2Former and evaluate it on **IM2GPS**, measuring top-1 percent country-label accuracy.

All-Feature Classifiers:

- We fine-tune 3 ViT-B-16 models on the entire feature-to-country dataset from G50, the entire feature-to-country dataset from G3, and a selection of 15k features from G50 equally distributed across **6 countries**.
- We evaluate on respective test splits and IM2GPS with top-1 percent accuracy.

References

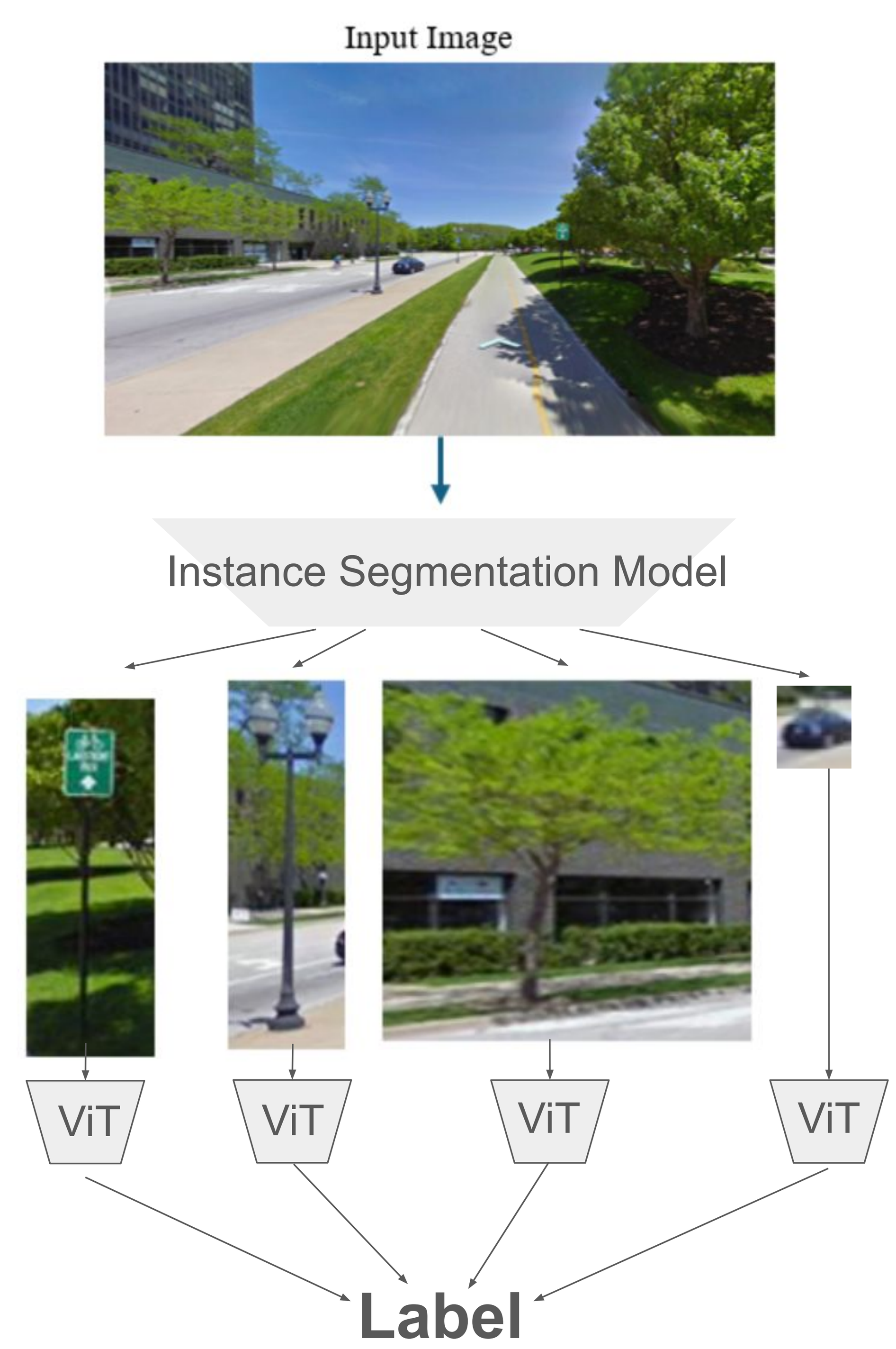
[1] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 3

J. Hays, A. Efros, et al. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2, 3, 4, 5, 7, 8

Rohan K. Geolocation - geoguessr images (50k). https://www.kaggle.com/datasets/ubitquitin/geolocation-geoguessr-images-50k/data, 2021. [Online; accessed 27-Feb-2024]. 3

Architecture

- **3-step model ensemble** for image-to-country classification



- At inference time, input images sent through **instance segmentation model** to extract important geographical signifiers
  - Road markings, vegetation, buildings, cars, ect.
- Each category of feature sent to a respective expert **Vision Transformer** fine-tuned for **image-to-country classification**
- Most frequently appearing country label across features chosen as final label

Results

Model	Accuracy
G50 Baseline	
G3 Baseline	0.3155
G50 Feature Classifier	0.5393
G50 6 Country Feature Classifier	0.8181
G3 Feature Classifier	0.5626
G50 Ground Classifier	0.4946
G50 Structure Classifier	0.3919
G50 Vegetation Classifier	0.3721
G50 Vehicle Classifier	0.1716
GeoCLIP	

Table 1. Results from ViT Test Accuracies

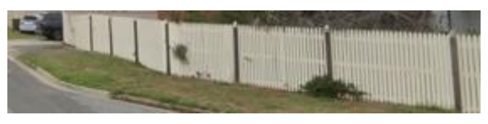
Model	Accuracy
G50 Feature Classifier	
G50 6 Country Feature classifier	0.2911
G3 Feature Classifier	0.0000
G50 Feature Ensemble Classifier	0.0302
G50 Ground Classifier	0.0251
G50 Structure Classifier	0.0101
G50 Vegetation Classifier	0.0754
G50 Vehicle Classifier	0.0000

Table 2. Model Evaluation Accuracies on Im2GPS

Datasets

- We use a mix of original and publicly available image-to-country pair datasets for training and evaluation:
  1. **(G50): Geolocation - Geoguessr Images**, 50k Google Street View images from 124 countries.
  2. **(G3): Original dataset of 3k Google Street View images** from 55 countries, equally distributed across countries
- We use **instance segmentation** to extract features from images in both datasets to create **feature-to-country datasets**.
  1. 99k features from G50, all countries
  2. 6k features from G3, all countries
  3. 15k features from G50, equally distributed across 6 countries
- We evaluate models on **original test sets** derived from above feature-to-country datasets, and on **IM2GPS** test set of geo-tagged Flickr images.

Sample image with label "Norway" from original Street View images (G3) dataset



Sample image with label "Australia fence" from G3 feature-to-country dataset

Discussion

- Performance varies between different expert feature classifiers, **dragging down overall ensemble model performance significantly**
- Likely due to differences in **extracted feature quality**.
- Feature extraction is **still promising**– performance of models trained on feature-to-country data on test splits is **markedly improved from baseline model performance**
- Distribution of data across countries in G50 **extremely unequal**, likely causing reductions in accuracy.
- IM2GPS images are very dissimilar to Street View training data, likely contributing to abysmal performance.

Future Work!

- Is directed feature extraction better or worse than random crop and scaling data augmentation?
- Is there a better way to extract features so that all features are the same quality?
- Are accuracy scores on IM2GPS higher on the continent-level?