worldNET | SPARSE UNDISCRETIZED DEEP CONVOLUTIONAL INTERPOLATION FOR IMAGE GEOPOSITIONING

Woody Hulse, Mindy Kim, Nuo-Wen Lei

Abstract

For nearly two decades, a popular topic of interest for computer vision researchers has been the prediction of geographic location from image data. Recent developments in Deep Convolutional Neural Networks (DCNNs) have accelerated this development, with state-of-the-art models showing upwards of 70% accuracy in general location prediction. All of these attempts, involve a straightforward process—create a feature extraction model, and then classify overall location based on those features and a discretization of the world map. Cutting edge models also require massive amounts of training data (~100M training samples) in order to accurately make these predictions.

We propose a training-efficient location predictor for sparse geographic data by continuous latitude and longitude coordinates. By taking a continuous approach, we can derive meaningful nonlinear interpolations of feature data and are able to create a more explainable mapping of global geographic features. We show that our model, worldNET, can deliver comparable performance (~55% continent accuracy) for seen locations and triple guessing performance on unseen locations (~21% continent accuracy).

Introduction

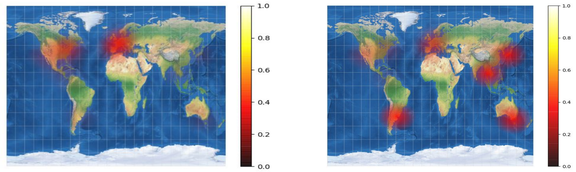
Inspired by the recent popularity of the game GeoGuessr, we propose a model that can infer continuous GPS location by recognizing features in an image. Specifically, we propose three approaches, two similar to previous works (nearest neighbors, transfer learning) and another approach that includes an interpolation head paired with a stored memory of training coordinates without need for discretization of the data at any point. We examine the effect of scarcity of training data and sparsity in the location distribution on the efficacy of each of these models.. Our dataset consists of a small sample size (2300) of only a few truly unique locations (23). A model with this sample efficiency in theory better understands the data, and in larger-scale applications may be more successful than the current state-of-the-art.

Methodology

Preprocessing

We use data from the GSV-Cities dataset, which includes over 560,000 geotagged images from 23 cities. We normalize each coordinate to 0-1 and resize each image to 300 x 400 resolution. We de-bias the data by algorithmically reducing the eurocentric concentration of training data.

Before and After Dataset De-Biasing



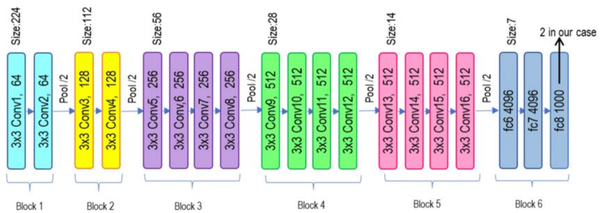
Loss Function

We compute the Mean Haversine distance loss to compensate for spatial distortions near the poles.

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Simple VGG

We first take an approach commonly used for image-to-GPS tasks–using a pretrained model and attaching a head. Instead of a classification head, we consider the idea of attaching a head with a final output layer of 2 nodes.

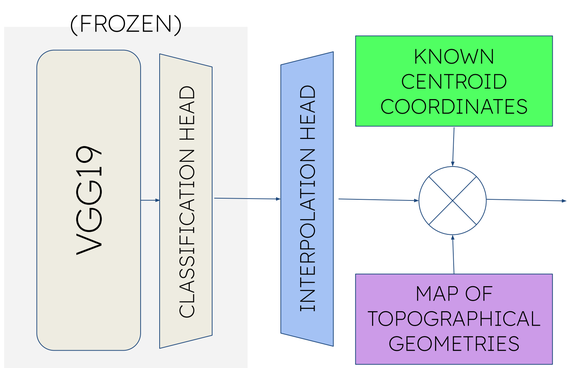


Nearest Neighbors

Our second approach was similar to a k-nearest neighbors (K-NN) idea of taking the vgg feature descriptors of the training images to find k closest descriptors to each testing image and utilize a mean shift clustering algorithm to calculate the weighted mean of the longitude and latitude coordinates.

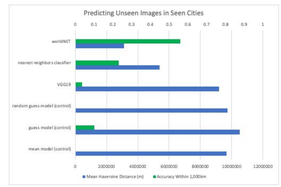
worldNET

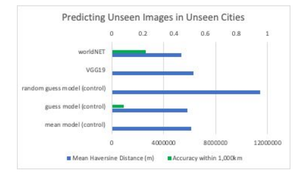
worldNET first clusters locations in the training data. With those centroids, we train a VGG-based classifier model to predict a latent probability distribution over those centroids. After training this classifier, we freeze the weights and append another head which predicts the proportion of each known cities' influence on the predicted image location, outputting the product of those predicted interpolations and known centroids. These are then mapped on a land geometry map to prevent predictions in the ocean.



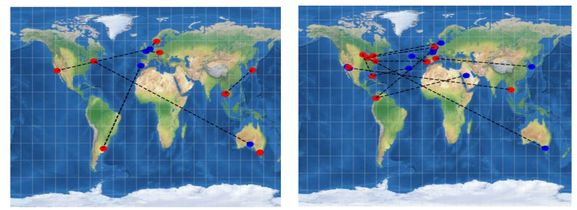
Results

We show that, compared with conventional approaches, worldNET performs significantly better in both distance predictions and accuracy. We see reasonable prediction even when we predict over cities which are foreign to the training dataset, and few mistakes for known cities. Conventional approaches notably fail–they can only converge on the mean of the training data





Prediction of Unseen Images in Seen vs. Unseen Cities



Conclusion

The VGG implementation follows the basic feed forward architecture of a DCNN is successful in a classification version of the task, however, due to a lack of confidence in spatial reasoning, always converges to repeatedly predicting the mean of the dataset.

Our continuous model, worldNET, takes advantage of the sparsity of geographic locations in our dataset and uses the prediction of a probability distribution of city centroids as a proxy to predict gps coordinates. These results show great promise for the spatial reasoning capacity of a worldNET-like model. It should be notes, though, that this model is very dependent on a reasonable distribution of training data–it will simply be unable to predict coordinates outside of the range that it sees in training. Future studies can improve the interpolation head of the model to mitigate this.