Geo-Localization of Images

Predicting Locations from Images Using Regression-Based Models

Graduate

Tobias Rothlin

Introduction: Geo-Localization aims to determine the location on Earth where a particular image was captured. Using visual clues in the image, it should be possible to determine where a picture was taken. Not every picture contains the same amount of information, limiting accuracy. If this task is performed reliably, it presents a range of applications. Foremost among these is its use in Open Source Intelligence (OSINT), which analyzes social media posts and publicly accessible images and videos and determines or verifies the associated metadata.

Approach: This thesis proposes using a regression-based model to predict latitude and longitude directly. Unlike previous methods that rely on databases, multiple model passes, or classifications, this approach trains a regression model to predict GPS coordinates directly from images. A Vision Transformer (ViT), specifically a pre-trained CLIP model, is utilized for feature extraction. A custom "Location Head" consisting of two Transformer Encoder Layers is added to the model. This head refines the CLIP image embeddings for location prediction. The CLIP ViT is frozen during training. Finally, the model employs a regression head to predict latitude and longitude. Various configurations, including pre-training and task-specific adaptations, were tested and evaluated.

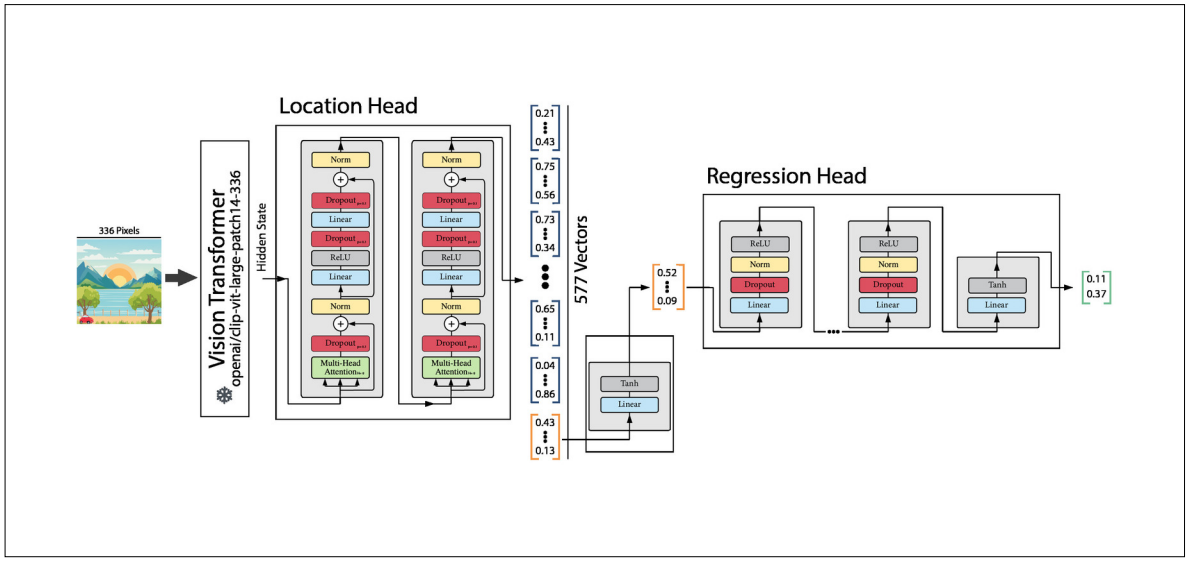
Result: The final model shows competitive performance with other transformer-based models for Geo-Localization even though the model architecture is significantly more straightforward, and only minimal supporting code is needed for the model to generate location predictions. The regression approach shows comparable or better performance for the country- and continent-level predictions, however requires further refinement and a larger training dataset to achieve good street- and city-level precision. The model performs very well on the Holdout test dataset, which was sampled from the same source as the training dataset; however, this performance drops significantly when the model is evaluated on out-of-distribution data (GWS15K, IM2GPS3K, IM2GPS).

Example from GWS15K Test Dataset
GWS15K Dataset Google Street View

Model Results compared to other Geo-Localisation models
Own presentment

Dataset	Model	Distance (% @ km) Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
GWS15K	StreetCLIP	-	-	-	-	-
	TransLocator	0.5	1.1	8.0	25.5	48.3
	GeoDecoder	**0.7**	1.5	8.7	26.9	50.5
	GeoCLIP	0.6	**3.1**	**16.9**	45.7	74.1
	Final Model	0.0	0.59	13.49	**46.14**	**81.87**
IM2GPS3K	StreetCLIP	-	22.4	37.4	61.3	80.4
	TransLocator	11.8	31.1	46.7	58.9	80.1
	GeoDecoder	12.8	33.5	45.9	61.0	76.1
	GeoCLIP	**14.11**	**34.47**	**50.65**	**69.67**	83.82
	Final Model	0.0	3.44	32.43	68.03	**84.95**
IM2GPS	StreetCLIP	-	28.3	45.1	74.7	88.2
	TransLocator	19.9	48.1	64.6	75.6	86.7
	GeoDecoder	**22.1**	**50.2**	**69.0**	**80.0**	89.1
	GeoCLIP	-	-	-	-	-
	Final Model	0.0	6.33	37.97	75.11	**93.67**
Holdout	**Final Model**	**0.18**	**32.08**	**88.08**	**96.86**	**99.34**

Regression Model Architecture
Own presentment



Advisor
Prof. Dr. Mitra Purandare

Co-Examiner
Dr. Cristiano Malossi, Rüschlikon, Zurich

Subject Area
Data Science