AI2CYBER - GeoSpy

Phase 3: Evaluation & Error Analysis

Dataset Analysis

Before discussing the model's performance it is important to examine the model's training and test set distributions. From CLIP's paper we find that it is trained on the WebImageText dataset which contains a wide variety of image-text pairs containing, but not limited to, urban/rural scenes, items, patterns, portraits, etc.

CLIP's paper does not specifically address the model's performance on geo-localization, therefore we will define our own geo-location task and metrics. Since the assignment does not provide a specific test set, we will utilize Google StreetView. StreetView's image library spans more than 100 countries, contains over 220 billion samples, and presents diverse scenarios, including rural and urban scenery. All images are easily accessible through the API, high resolution, captured under consistent lighting/weather conditions, and directly geolocatable for our ground-truthing requirements. Also, in CLIP and WebImageText's papers there is no mention of StreetView images being contained in the training datasets, so we can reasonably assume the images to be unseen to the model.

Model Performance

CLIP can be used for image geolocalization by comparing an image's visual features with textual descriptions of locations. For example, CLIP can recognize the visual features of the Eiffel Tower and compare them to textual descriptions of Paris. If the visual features of the image are similar to the textual descriptions of Paris, CLIP can infer that the image was likely taken there.

However, CLIP is not perfect for image geolocalization as it is more accurate when used with images that have clear visual features that are associated with a specific location. For instance, CLIP is more likely to accurately geolocate an image of the Eiffel Tower than an image of a generic street in Paris.

To evaluate CLIP's performance, we used three classification methods:

- Continent level
- Country level (classify only within the predicted continent)
- Country level (classify across all countries)

Our test set consisted of 33 Street View images, roughly equally distributed across all continents, representing both urban and rural locations. Some locations were chosen randomly, others as edge cases near country borders, and some for their distinctive visual features. This distribution likely skews the evaluation metrics, as it doesn't meet the independent and identically distributed (I.I.D.) assumption relative to the training set. Ideally, the distribution of the user query set should be known to make the metrics relevant to a specific problem.

The metrics for the three classification methods on our own test set are shown below:

	F1 Score	Precision	Recall
Continents	**0.655**	0.820	0.676
Countries on Continent	0.333	0.360	0.333
Countries Direct	0.456	0.480	0.456

Considering the rapid prototyping process, the zero-shot approach, and the ambiguous features of the compiled test set, we could say that the model performance was non-trivial, although there is much room for improvement. The country-level prediction method based on the predicted continent performed poorly, likely due to the compounding error from the two-stage prediction process.

Error Analysis

As previously mentioned, the test set was manually curated, with many locations specifically selected for their ambiguity. Locations with distinctive vegetation, architectural styles, etc., performed well at the continental level. However, similar features across neighboring countries, or even within the same continent, led to misclassifications at the country level. For example, samples from Greece, Albania, Israel, and Turkey—despite not all being on the same continent—exhibit considerable visual similarity. Conversely, geographically distant locations were classified much more accurately.

The potential impact of factors like lighting, weather, and camera settings on model performance remains unanalyzed.

Finally, an attempt was made to handle exception classes (e.g., Interior, Portrait, Blur, Low Light, Micro, Macro, Abstract, Closeup, Satellite) to identify and exclude such images from standard classification. While implemented, this function has not been formally evaluated, and initial testing suggests it significantly degrades model performance.

Potential Improvements

During the design, development, and evaluation of the AI system, we identified several potential improvements for future iterations.

A key next step would be to more clearly define the specific use case and develop a representative test set. This would allow us to establish appropriate Key Performance Indicators (KPIs) and guide subsequent development efforts.

If the defined requirements necessitate a higher-performing system, fine-tuning ViT could be explored. This would involve significant computational resources, as well as substantial effort in data acquisition, processing, and model training. Fine-tuning CLIP would likely be even more resource-intensive due to the complexity of its training dataset.

The performance of CLIP, which matches text and image embeddings, can be influenced by the specific wording of class labels. We could investigate whether variations in labels (e.g., "USA" vs. "United States") impact prediction accuracy.

Inspired by techniques used in GeoGuessr, incorporating more granular visual cues could enhance location precision. These cues might include flags, license plates, or road signs. A zero-shot object detector (e.g., DINOv2) could be used to identify these items within an image, and the results could then be cross-referenced with a visual database. Furthermore, traffic or advertising signs could be processed by an OCR model to extract textual information, such as language or place names, to further refine location identification.