Image Geolocation with Directed Feature Identification

Madeline Brumley
University of Washington
madbrum@uw.edu

Claire Gong
University of Washington
cgong16@uw.edu

Abstract

Image geolocalization, or the prediction of the exact geographic location depicted in a photo, is known to be a very challenging task for vision models. For instance, certain geographic locations can appear very different in photographs depending on the time of day or season, complicating the classification task. Specialized humans, however, can perform very well on this task. We identify directed recognition of key objects pictured in the image as a key strategy for the most capable human guessers. To explore the effect of this strategy as applied to deep learning models and investigate a more human-interpretable approach to this task over the current state-of-the-art, we introduce and evaluate a new architecture for the image geolocalization task that employs guided visual search through instance segmentation to identify key image features and feature-region classification on these image features. We also investigate the effect training on image features has on model performance compared to models trained on entire images and find that image feature extraction for geolocalization is promising, boosting performance significantly over baseline models.

1. Introduction

The online game "Geoguessr" has exploded in popularity in recent years, due in no small part to the novelty and unique difficulty of its core task: image geolocalization. Image geolocalization, or the prediction of the exact geographic location depicted in a photo, is a difficult task for the average person, requiring familiarity with the geographical characteristics of a wide range of different regions around the world and a nuanced understanding of diverse cultural and infrastructural idiosyncrasies to distinguish between geographically similar regions. Among the loyal followers this game has amassed are many highly capable guessers who have strengthened these understandings and are able to successfully determine locations of images within only a few kilometers of where they were taken.

Meanwhile, image geolocalization has proven to be a very challenging task for vision models. The appearance of a geographic location in a photograph is highly dependent on lighting, time of day, and seasonality, among other factors; where humans are able to identify important visual features despite these changing conditions, it is much more difficult to devise a scheme for vision models to recognize these features. The top-performing models on this task in recent years, TransLocator and StreetCLIP, both utilize visual transformer (ViT) architectures and creative preprocessing of input images to approximate the necessary feature recognition for geolocalization, generally performing well with classifying images by continent and country. However, these models still perform relatively poorly when making predictions on the regional level within a country, performing even worse for more granular predictions [4, 7].

We find that the strategies that successful human guessers use for this task are underutilized when developing geolocation models and note the most effective strategy used by successful human guessers: directed recognition of key objects pictured in the image. Identifying regional idiosyncrasies of objects appearing in the image, such as telephone poles, plant species, etc., allows professional guessers to achieve high success rates. While current models attempt this indirectly, attention mechanisms are not human-interpretable, and it is unclear if attending to specific parts of the image is indicative of their successful identification of the most important objects within the image. If the average human can accurately identify their location by 1) identifying the key semantic features likely to exist in their surroundings, 2) extracting these semantic features using visual information, and 3) comparing each feature with features of the same type previously seen in a variety of other locations, it seems reasonable that a model able to robustly carry out these subtasks would perform well on the image geolocalization task compared to models unable to do reliably do so.

We therefore propose a novel approach for image geolocalization. This approach constructs a model ensemble consisting of a masked-attention mask transformer for instance segmentation followed by an ensemble of vision transformers for feature classification. We first perform instance segmentation on an image input to robustly identify characteristic features of the location represented in the image. In this way, we mimic human guessers' identification of key semantic features in the image (vehicles, foliage, architectural features, infrastructural features, people, ect.) for closer inspection. This closer inspection, for each feature, is performed by a specialized vision transformer trained on feature-region pairs for features of the same type. With region classification done on each feature, the final geolocalization output is computed by majority vote across features.

As a more simplified approach, we also train a single vision transformer on all types of features to perform feature-region classification. We intuit that specialized feature-to-region classifiers would learn more fine details of their respective feature and could thus distinguish between their characteristics more robustly, improving performance on the task, but suspect that limited data may prevent these classifiers from fitting well enough to these features. Thus, we also feed all feature data to a single classifier to observe if extracting images features in advance and classifying over them at all improves on classifying over the entire image as current models do.

With this work, we hope to also encourage future work towards increasingly human-interpretable approaches to this task. By deliberately segmenting the image into features that are meaningful to humans, and then performing geolocalization for each feature, we are able to clearly segment the geolocalization task into subtasks, and can interpret where and why models tend to fail with more precision. Understanding why models perform as they do at each stage of the computation is crucial for identifying where they can be improved and enables more direct and productive work toward improvement on a given task. We aim for our work to be interpretable enough to encourage productive interpretation and extension of our approach.

2. Related work

2.1. Image Geo-localization

Current state-of-the-art approaches to the image geolocalization task achieve generally high performance on planet-scale image geolocalization benchmarks such as IM2GPS and IM2GPS3k [5] on the continent- and country-level, but deflate in performance on the regional and city level. StreetCLIP [4] achieves SOTA accuracy on the continent-level for IM2GPS3k by pretraining OpenAI's vision transformer-based CLIP model on synthetic image captions geared toward the geolocalization task, transferring existing zero-shot capabilities from CLIP to the more specific image geolocalization domain. TransLocator maintains SOTA accuracy on the country-, region-, and city-level with a transformer-based architecture that synthesizes raw images with semantic segmentation maps of the images in a single pass through a vision transformer model, combining raw visual information with image features robust to appearance variations brought about by changes in seasonality or time of day [7]. Both models improve greatly on the previous SOTA by applying vision transformers to the task. TransLocator also notably benefits significantly from the introduction of semantic segmentation maps to the task, indicating that employing some form of feature extraction invariant to seasonal and time-dependent changes in appearance is necessary for high performance on this task.

We'd like to also mention GeoCLIP [1], a CLIP inspired image geolocator which uses GPS encoding for geolocalization. What is most notable about the model is that it is able to achieve competitive performance with only 20% of the 4.6 million Flickr geo-tagged dataset that many other SOTA models have utilized. Since the model is available for testing and already benchmarked on the Im2GPS dataset, we decide to make comparisons with our own model in the accuracy evaluations on our own dataset to give our model a point of reference with current SOTA models.

2.2. Guided Visual Search

Drawing on humans' innate capability for visual search, guided visual search algorithms for multimodal LLMs have emerged as a promising technique for improving model capabilities on visual search tasks. Rather than relying on model attention to identify critical regions of an image, guided visual search algorithms draw on multimodal LLMs' contextual understandings of where certain objects might be located in an environment to explicitly delineate important sections of an image over which the model may then continue searching for a target object. By restricting the domain of model attention to specific areas of an image, using the same reasoning as a human might, models show a significantly improved ability to identify small features in an image where they otherwise would fail [8]. In this way, identifying critical patches of an image by reasoning about the image as a human might and then computing over these patches shows promise as an effective strategy for improving vision model capabilities on other tasks where attending over subsections of an image is useful for humans.

3. Methodology

3.1. Model

We thus design a model ensemble to first extract image features and then perform feature-region classification on each feature for geolocalization. The feature extraction step uses an image segmentation model to identify individual objects in the image, including people, vehicles, infrastructural elements, and environmental features like foliage or the sky. We use Mask2Former [2], a pre-trained masked-attention mask transformer, for this subtask, whose pretraining data includes all classes of objects important for human geolocalizers. Once these features are extracted, they are each passed into respective expert vision transformers that specialize in specific feature types. This is effectively a classification task; we train each Vision Transformer on images of a particular type of feature labeled by region to produce an ensemble of feature-to-region classifiers. We specifically chose ViT-B-16 for this task due to its manageable size given our resources and better performance over larger architectures on less training data [3]. Classifications for each feature are collected, and the most commonly appearing region label is chosen as the overall label for the image. The result is a model that mimics human strategies for image geolocation and is thus more interpretable than current SOTA models.

Existing attention-based approaches in SOTA work see models attend to “important” regions of the image, but it is unclear exactly why these models attend to these regions. It is reasonable to conjecture that separating the steps of recognizing these identifying features through instance segmentation and then attending to the details of these features will improve performance: attention in current vision models presumably has to take care of both steps at the same time, while our approach dedicates more compute to each step separately. In doing so, our model may learn finer details of important image features and thus can differentiate between features of the same type across regions much better.

3.2. Datasets

We restrict the scope of our task to country-level classification. We do this for several reasons: we can explore a wide breadth of regions with different characteristics to test our architecture, available training data is most plentiful for this level of the task, and this domain is well-explored and thus has plenty of benchmarks on which we can evaluate our results.

Our model ensemble is trained on two datasets that will serve as input images: Geolocation - Geoguessr Images (50k) [6] (which we shorten to G50) and our own curated dataset of Google Street View Images (3k) (which we refer to as G3). We split these datasets into test and train sets and use them to train and evaluate our model. Geolocation - Geoguessr Images (50k) consists of 50 thousand web scraped Google Street View images from mygeoworld.com spanning 124 different countries. Each image is paired with its corresponding country label. Our original dataset, Street View Images (3k), consists of three thousand web scraped Google Street View images from randomstreetview.com of 55 different countries. Again, each image is paired with its corresponding country label.

Due to the nature of our model ensemble which requires feature-specific data to train each feature-to-country classifier, we also create three original feature-to-country datasets for the feature classifiers that we extract from the two previously mentioned image datasets using Mask2Former. Each of the feature datasets consist of a feature to country to image hierarchy where each image has both a feature and country label. First, from G50, we extract all features (crops of each important part of the image) from every image to create a dataset of 99,000 images. Second, from the extracted features dataset of 99,000, we take the extracted features of six countries, where each country has 2,500 feature images, resulting in a dataset of 15,000 images equally distributed across each country in hopes of creating a more balanced and unbiased dataset. Third, we extract from G3 to a dataset of 6,000 feature images.

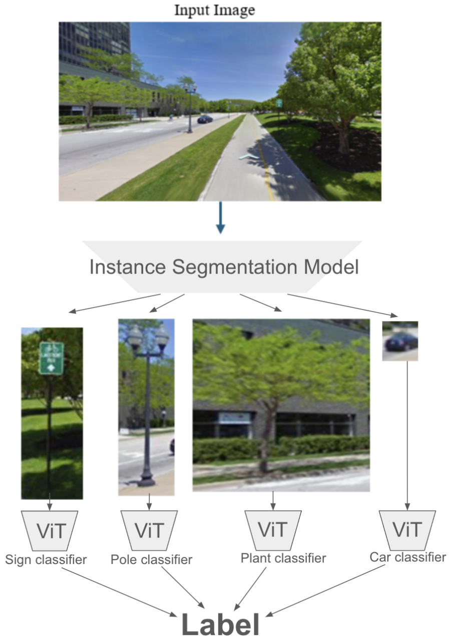

Figure 1. Model architecture

We use original test splits from the above datasets to evaluate feature classifiers, and evaluate the model ensemble on the IM2GPS benchmark dataset of 237 images of various pictures of locations, landmarks, plants and animals, and indoor facilities to give us a point of reference in comparison to current SOTA models. We choose the IM2GPS evaluation set because the majority of SOTA geolocalization models evaluate on this set, and we wish to compare our results with theirs. This set contains a mixture of city and country names which we convert to country labels. On both the feature-to-country test splits and IM2GPS, we evaluate models by measuring top-1 raw percent labeling accuracy.

3.3. Probability Output

We can mathematically define image geolocation to further explain our intuition for our approach to image geolocation with identified location-based objects. First, we start with defining the general purpose of image geolocation with I as the input image and C as the classified country.

$$P(I, C) \tag{1}$$

We can introduce an intermediate variable, F, which represents the features the model uses to identify the country. We elaborate on F later on.

$$P(I, F, C) \tag{2}$$

By conditional probability:

$$P(C|F, I)P(F|I)P(I) \tag{3}$$

In 3, we identify $P(F|I)$ as the first phase of our model where given an image, it identifies the probability of the features. $P(C|F, I)$ is the second phase of our model where given the features derived from the image, it identifies the probability of the associated country. $P(I)$ simply represents the image that is inputted into the model.

There are two types of variables for F. First, where F is an embedding. To our knowledge, this is where pretty much all of the geolocation models fall under. An embedding is defined by the model itself based on the features it deems as important, it is not explicit and as we've seen by the performance of related geolocation models, with large datasets and due to its end-to-end optimization (I to C), it gives the neural network the most flexibility to find a good embedding.

However, these embeddings are set in an arbitrary method based on the model itself and in situations where there may not be large datasets available, such as more remote areas or countries, the embeddings may find it difficult to be able to generalize features towards areas with few data. This is where we introduce our approach, where F is instead label pictures of specific features we have pointed out, such as the telephone polls, or specific road signs.

4. Experiments

Our experiments seek to evaluate the performance of geolocalization models using feature extraction versus geolocalization models not using feature extraction. We also seek to evaluate how classification over all features compares to classification over separate features, and hope to also observe how models perform when trained on small versus large datasets.

We first construct feature-to-country datasets from G50 and G3 by performing instance segmentation on the data with Mask2Former. Segmentation maps are converted to bounding boxes, and the segment of the image bounded is saved with both its corresponding feature label generated by the model and its country label from the originating dataset.

For each experiment, each dataset used is split with a 7:1.5:1.5 ratio into train, validation, and test sets and used accordingly. All classifiers referenced are pre-trained ViT-B-16 image classification models, and all models are fine-tuned for 11 epochs with a learning rate of 2e-5. We tested epochs of 3, 5, 11, 12, and 30 and found that 11 epochs were adequate for models to fit to the train data without too significant computational cost.

4.1. Universal Feature-to-Country Classification

Our first, more straightforward experiment hopes to observe the effect of performing geolocalization over image features instead of overall images. To measure this, we fine-tune a pre-trained ViT-B-16 model on unsegmented images directly from the G3 dataset to serve as a baseline and fine-tune another ViT-B-16 model on the feature-to-country G3 dataset described above. We evaluate the models on their respective test splits as well as IM2GPS, omitting test images representing regions the models were not trained to classify.

We perform the same procedure for the G50 dataset. However, images are not distributed equally across countries in this set, so we additionally select 15 thousand images from G50, equally distributed across 6 countries, and perform the same procedure over this set as well. Doing this ensured an equal, unbiased distribution of data, as well as ample enough data to learn decent representations.

4.2. Specific Feature-to-Country Classification

Here, we evaluate our proposed model ensemble discussed in Section 3.1. To begin constructing our ensemble, we decide to train 4 expert feature-to-country classifiers, separated by 4 feature categories: ground, vegetation, structures, and vehicles. These categories group together feature labels produced during the image segmentation process; the vehicles category, for example, encompasses "car," "bus," and "truck" features extracted by the segmentation model. We choose these categories of features to group together related, expressive features observed in our feature-to-country datasets. We then train each classifier over only their category of feature, selecting, for instance, only vegetation feature data from our feature-to-country dataset to train the vegetation expert model on. We train these models only on feature-to-country data from the G50 dataset.

Due to sparsity in the dataset, some countries lack any data corresponding to a given feature. In these cases, in order to process our datasets consistently and easily without having to get rid of entire country classes, we add a single 32x32 random noise image to these folders. We expected that performance for these countries would be poor regardless, and that a few additions of noise would not throw off the data distribution too significantly.

Once these models are trained, we evaluate them on feature-specific test splits following the same procedure as all other evaluations. We also evaluate them on IM2GPS. Finally, we ensemble them together with Mask2Former and evaluate the ensemble model performance.

4.3. Dataset Evaluation

We also evaluate GeoCLIP on an original test split of our G3 dataset. We suspect that the distributions of IM2GPS and our Street View datasets are very dissimilar and thus wish to observe how a model trained on IM2GPS generalizes to our data.

5. Results

We find widely ranging accuracies across models when evaluated on original test-splits, and almost universally poor performance on IM2GPS for each model we fine-tuned. Models trained on more well-distributed data, and greater quantities of data, generally show better performance. The G50 6 Country Feature Classifier notably achieves 81.81% test accuracy on its test split and achieves better than random chance when evaluated on IM2GPS test images over its 6 countries.

Model	Top-1 Country Accuracy
G3 Baseline	0.3155
G3 Feature Classifier	0.5626
G50 Feature Classifier	0.5393
G50 6 Country Feature Classifier	0.8181
G50 Ground Classifier	0.4946
G50 Structure Classifier	0.3919
G50 Vegetation Classifier	0.3721
G50 Vehicle Classifier	0.1716
GeoCLIP	0.1835

Table 1. Results from Original Test Split Accuracies

Model	Top-1 Country Accuracy
G3 Feature Classifier	0.0000
G50 Feature Classifier	0.0005
G50 6 Country Feature classifier	0.2911
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 Evalulation Accuracies on IM2GPS

Due to constraints on compute, we were unable to train and evaluate baseline models on the G50 dataset and the G50 6-country subset.

6. Discussion

6.1. Models

From our results, we observe that models we trained on more data performed significantly better than models trained on smaller amounts of data, which is to be expected. Feature classifiers trained on upwards of 15-20 thousand images, like the ground and vegetation feature classifiers, achieve higher accuracy than classifiers trained on only 5 thousand images such as the vehicle feature classifier. The best performing models on a given domain overall are also the classifiers trained on the most data: the combined feature classifier over 6 countries achieves 81% test accuracy on its respective test set, which is certainly partially attributable to its significantly smaller classification domain but still demonstrative of how feature classifiers can certainly achieve high accuracy when trained on unbiased, higher quality data in large amounts.

6.2. Data

It's important to note the biases and limitations of each dataset that will influence our results. First, note that despite our goal of having our model to be able to recognize all countries, our model versions are only able to guess between 124 or 55 different countries so we are limited by this scope.

With Geolocation - Geoguessr Images (50k), due to it being webscraped off of mygeoworld.com, the images are not given completely at random, resulting in a non-uniform distribution of countries, with each country containing a number of images between 1 and 12,000. Therefore, this results in huge biases with the countries that have the most images compared to the countries that have very few. Countries like the United States or the United Kingdom, for example, have well over 10,000 images in the dataset, while countries like China or Egypt have under 100. With Street View Images (3k), we make sure to have each country take in an equal amount of images, so this type of bias is removed.

After looking at our results, we can see that the G3 feature classifier model trained better on the 3,000 image dataset than the 50,000 image dataset by 2.33% which we believe is due to less dataset bias and less scope of countries to classify from.

After looking at ViT test accuracy results, we can take note on how the G3 Classifier and the G50 6 Country Feature Classifier have the highest accuracies. Even though the G3 Classifier is trained on less data than the G50 Feature Classifier, we believe that due to an evenly distributed country dataset, it was able to perform better than the G50 Feature Classifier which, while containing more data, is significantly more biased. Additionally, the G50 6 Country Feature Classifier brings in a very high accuracy of 81.81%. Its dataset not only is evenly distributed like G3, but it has more data per country. This leads to promising results for future work with acquiring more data that is evenly distributed.

6.3. Feature Extraction

Due to the fact that we were unable to evaluate baseline models on unsegmented images for both the G50 and G50 6-country datasets, the conclusions we can draw from our results are limited. However, these preliminary results seem to encourage further investigation, as performing feature extraction and training on extracted features is indicated by our results on the G3 dataset to improve performance over classifiers trained on default images alone. The classifier trained on features extracted from the 55 country dataset improves on the classifier trained on only raw images from the dataset by 25%. Further investigation is certainly necessary to observe this effect. If evidence continues to support the idea that feature extraction helps improve classification accuracy overall, it fits within the intuition that training on more data (several features are extracted per image in the dataset) yields better results. The method we employ for feature extraction also bears similarities to established methods of data augmentation like random cropping and scaling that bolster model performance with more input data, which is a possible reason this effect is observed.

We also observe that inconsistencies in the quality of extracted features has drastic results on classification performance. The vehicle classification model for example performs significantly worse on its test set compared to the ground classification model. We believe that this occurs because the extracted vehicle features are significantly noisier and lower quality than ground feature data. The ground feature dataset consists of much larger, more consistently sized images; more content is being fed into the model with each image input. However, the vehicle feature dataset is much noisier, with some decently sized representations of various vehicles but by and large consisting of tiny patches only a fraction of the size of the image they were extracted from, thus containing much less useful information. Other feature data for features like traffic lights and street signs were similarly poorly distributed and noisy. This indicates that a more robust method of feature extraction and data collection may be necessary, perhaps involving the filtering out of smaller, less expressive data.

The G3 Feature Classifier compared to the G3 Baseline has a notable improvement in country accuracy. Additionally, it is clear that the G50 Ground Classifier has better performance on the test data than the G50 Vehicle Classifier by 32.3%. We find that both improved models feature mimic patching data augmentation which could lead to better results.

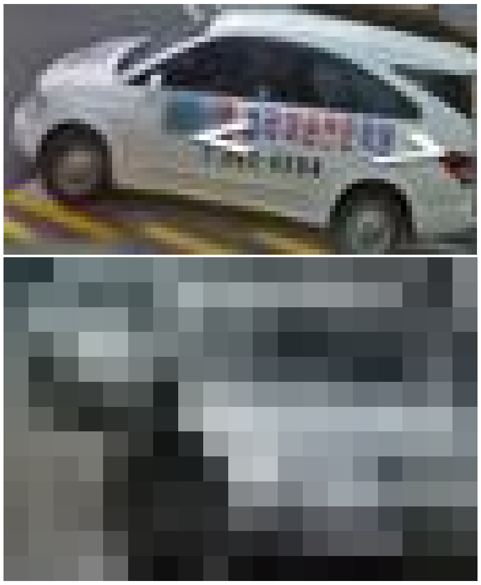

Figure 2. Two samples of extracted features labeled ”car.” Note the bottom image has been scaled up for visibility.

While the G50 Vehicle Classifier, is trained on the same method as G50 Ground Classifier, it suffers from the feature extraction of smaller objects.

6.4. Evaluation

Although we received sub-optimal accuracies on the IM2GPS benchmark evaluation data, there are still some good results outside of the country accuracy tests. For instance, we anecodtally noticed that although the models didn’t guess the correct countries most of the time, there was a general sense of them predicting the correct continent. For instance

Additionally, it is important to understand not only is there a limitation of our datasets due to non-uniform data distributions of country and feature images as explained earlier, but also that there is a limitation of the Im2GPS benchmark evaluation data.

We found that the content of the images are generally out of distribution for the data we trained our models on. For instance, famous landmark and close up shots of the plant and animal species take a big portion of the IM2GPS test set. Normally these famous landmarks have something unique about their architecture, with features that are not as commonly used in regular architectures throughout the country. Im2GPS’s close up shots of plants and animals normally have the background blurred and show species that can exist on multiple countries and continents, making it very difficult for our model to determine exactly one country.

Notably, GeoCLIP, which was trained on IM2GPS, achieved only 18% accuracy on our original test sets from G50 Street View data, despite performing well on the

Im2GPS3k test set, a larger test split from the IM2GPS dataset. This indicates that poor evaluation performance of our models on the IM2GPS test set is not entirely indicative of their failure to learn good representations at all, but rather that the representations learned do not generalize well to domains outside of Street View images.

As our model focuses more on the general sense of each country or region, instead of unique, rare features that don't represent the whole of a country, it makes sense that it doesn't perform as well on the Im2GPS dataset country-wise, and that it would generally perform a lot better continent-wise in its current state.

7. Future Work

The main takeaway from our experiments is that guided feature recognition does in fact help the embedding process for our geolocation image classification model which is clear from the big gain in accuracy points with the G50, G50 6 Country, and G3 Feature Classifiers from the G3 Baseline model. Therefore, there is definitely a promising opportunity to attempt guided feature recognition on different tasks.

Based on our understanding of how the model was limited by its dataset due to non-uniform country and feature distribution and its ability to demonstrate promising results when given access to uniform country distributed data with more size through the G3 and G50 6 Countries datasets, we believe that there is great value in creating bigger datasets that maintain an even distribution of labels. One specific cause of the uneven distributions of data are the discrepancies in the sizes of the objects in the input images and therefore, unequal image quality of the feature data. The smaller an object is in the picture, the poorer its feature quality and vice versa. One idea to address this calls for the help of diffusion, where one could take the poorer quality feature images and create images of higher resolution.

We also noticed that there are similarities when it comes to direct feature extraction and random cropping and scaling data augmentation. It seems worthwhile to determine if our directed feature recognition is really a reiteration of this data augmentation method, or if these methods produce different effects on model performance. Intuition leads us to hypothesize that directed feature recognition could be better because it provides more information that is deemed more important when classifying the image, and propose further experimentation on this.

As mentioned when examining the Im2GPS evaluation results, with the ancedotal observation that there is a general consensus on continent when comparing the output classified country for the feature images of the input image, it might be worthwhile to analyze our model on a continent-wise scale to better understand the generalizations it makes.

We also hope to perform more rigorous analysis on the effect of dataset size on transformer geolocalization performance when trained and evaluated on full images compared to performance when trained and evaluated on image features. We hypothesize that classification on extracted image features will tend to prevail over classification on full images when training from smaller datasets, but will eventually be overtaken by classification on full images when training on larger datasets. Our intuition for this is that though transformers training and evaluating on full images will attend to parts of the image in a somewhat arbitrary/uninterpretable way, they will also learn connections between features they deem important which will eventually improve their performance once they learn rich enough features on larger datasets.

References

[1] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization, 2023. 2

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 3

[4] Lukas Haas, Silas Alberti, and Michal Skreta. Learning generalized zero-shot learners for open-domain image geolocalization, 2023. 1, 2

[5] James Hays and Alexei A. Efros. im2gps: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

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

[7] Shraman Pramanick, Ewa M. Nowara, Joshua Gleason, Carlos D. Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild, 2022. 1, 2

[8] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023. 2