Ministerie van Defensie

Master Thesis

Echancing Image Geolocalization with a Hybrid of Semantic and Vision Transformer Embeddings

---

by

Yesse Okkerman

(2737996)

First supervisor:	Britt van Leeuwen
Daily supervisor:	Ward Jansen
Second reader:	Sandjai Bhulai

September 16, 2024

Submitted in partial fulfillment of the requirements for the VU degree of Master of Science

Echancing Image Geolocalization with a Hybrid of Semantic and Vision Transformer Embeddings

Yesse Okkerman
Vrije Universiteit van Amsterdam
Amsterdam, The Netherlands
y.f.c.okkerman@student.vu.nl

ABSTRACT

This study investigates the enhancement of image geolocation by utilizing a hybrid approach that combines semantic and vision transformer (ViT) embeddings. The primary goal is to predict the geographic location of an image at the country level by leveraging different neural network (NN) architectures, including a baseline model and a more complex NN with three hidden layers. We implemented and tested models using embeddings derived from pretrained ViT and CLIP models, which were evaluated across various performance metrics to determine the most effective combination of model architecture and embeddings. The findings demonstrate that the NN with three hidden layers outperformed the baseline model in geolocation accuracy, particularly when using the combined embeddings from ViT and CLIP. This hybrid approach proved especially effective in enhancing the model's capability to predict geographic locations, as evidenced by superior performance on benchmark datasets measured by the "% @ km" distance metric. However, while the integration of multiple pretrained embeddings and increased model complexity led to significant improvements, the overall geolocation accuracy remains suboptimal, indicating the need for further optimization. This research underscores the potential of combining semantic and vision transformer embeddings in more complex neural networks for improved geolocation accuracy. The results suggest that while advancements have been made, future work is necessary to address the limitations observed and to fully realize the capabilities of these models in geolocating images with higher precision.

1 INTRODUCTION

In the modern era of rapid technological advancements, the ability to analyze and interpret visual data has become increasingly critical for national defense and security. The Ministry of Defence is continuously exploring innovative methods to enhance its operational capabilities, particularly in the realm of intelligence gathering and situational awareness. One such area of interest is the application of machine learning techniques to automate and improve the process of geolocation from images. The increasing amount of data generated today has further fueled the growth of interest in computer vision, as highlighted by Gil et al.[5].

One particularly interesting application of computer vision is determining the location depicted in an image or video. Humans can often infer a location by examining various aspects of the image, such as architecture, climate, and license plates. However, while human cognition is capable of such inferences based on experience, it requires vast cognitive capacity to have knowledge about every country on the planet. In contrast, computer models can be trained on extensive datasets containing diverse images from various regions, enabling them to efficiently and accurately make predictions based on visual data. By leveraging these large and varied datasets, models can learn to identify patterns and features that allow for accurate predictions across a wide range of scenarios demonstrated by Zhou et al. [21].

Embeddings are a crucial component in the process of geolocating images, as they provide a compact and meaningful representation of the visual features within an image, shown by Vaswani et al.[16]. These embeddings can be obtained from pretrained models, which are neural networks that have already been trained on large, diverse datasets, such as ImageNet [12]. Jacob Devlin et al.[3] stated that a pretrained model can be fine-tuned by adding an additional output layer for the specif task. Pretrained models are beneficial because they have learned to recognize a wide range of visual patterns and features, making them well-suited for various tasks, including geolocation. To obtain embeddings, an image is passed through a pretrained model, typically up to one of its later layers. Instead of using the model's final output layer, which is often designed for a specific task like classification, the output from an intermediate layer is extracted. This intermediate output serves as the embedding, encapsulating the essential features of the image in a high-dimensional vector. This process is a form of transfer learning, explained by Weiss et al.[18], where the pretrained model's knowledge, gained from a large dataset, is leveraged to extract meaningful representations from new data. These vectors, derived from the intermediate layers, capture important aspects such as texture, color, shapes, and spatial relationships that are critical for identifying the location depicted in the image. By using transfer learning in this way, the pretrained model's ability to recognize complex patterns is harnessed, even when the original task differs from the current one.

For the Ministry of Defence, this technological capability offers significant advantages. Automating the process of geolocating images not only enhances the efficiency of intelligence analysis but also reduces the cognitive burden on analysts, allowing them to focus on more complex tasks. The significant advancements made in the field of computer vision for geolocalization serve as a strong foundation for developing robust models capable of supporting defense operations. In this work, we propose a hybrid approach that integrates semantic embeddings from CLIP (Contrastive Language–Image Pre-training) with embeddings from a pretrained Vision Transformer (ViT). By leveraging the embeddings from both ViT and CLIP as features, we aim to address the classification challenge of predicting the country depicted in an image. Our method

builds upon the approaches of Pramanick et al. [9] and Haas et al. [6], combining the use of ViT with CLIP embeddings to capture useful information from the image. Instead of relying on semantic maps, we use CLIP embeddings to condense and summarize the visual content. These combined embeddings are then fed into a model designed to predict the location of previously unseen images. A key aspect of our approach is the development and refinement of this classification model to effectively interpret and distinguish complex patterns and features within the embeddings. The overall process involves several critical steps: image acquisition, preprocessing, feature extraction, detection/recognition, and decision-making.

This paper starts with a literature review on geolocalization using computer vision in Section 2. Section 3 covers the data used in the research, while Section 4 details the model descriptions. The experimental setup is presented in Section 5, followed by an analysis of the results in Section 6. Finally, Sections 7 and 8 offer a discussion and conclusion of the experiments conducted.

2 RELATED WORK

Image geolocalization involves analyzing visual data to determine the location where an image was captured, a task that remains a significant challenge in computer vision, particularly when applied at a global scale. The complexity of this problem is heightened by various factors such as variations in lighting, weather, seasons, time of day, climate, and viewing angles, all of which can drastically alter the visual characteristics of a scene. To address these challenges, various approaches have been developed over the years, evolving from traditional methods to more sophisticated deep learning models.

Early Approaches: CNNs in Geolocalization. Convolutional Neural Networks (CNNs) are deep learning models that are designed for processing and analyzing visual data. Unlike traditional NNs, CNNs excels in learning spatial feature hierarchies from images. They achieve this through a series of convolutional layers that apply filters to the input image, detecting features like edges, textures, and shapes at various levels of abstraction, demonstrated by Oshea et al.[8]. One of the pioneering steps in applying deep learning to geolocalization was the introduction of CNNs by Weyand et al.[19] through their PlaNet model. This model divided the Earth's surface into discrete geographical cells and trained a CNN to classify images into these cells. PlaNet demonstrated that CNNs could effectively handle geolocation tasks across various scales, from city level to continent level, significantly improving geolocation accuracy compared to earlier non-deep learning methods. Building on this foundation, Seo et al. [13] extended this approach with CPlaNet, which employed combinatorial partitioning and geoclass voting to enhance accuracy. This method effectively balanced the trade-off between cell size and the availability of training data, representing a significant advancement in geolocalization techniques.

Advances with Transformer Architectures. While CNNs have played a crucial role in advancing geolocalization, recent developments in transformer-based architectures have also significantly influenced the field. The Vision Transformer (ViT) architecture, introduced by Dosovitskiy et al. [4], expands the natural language processing transformer application to image recognition. Unlike traditional CNNs, ViT utilizes self-attention mechanisms, patch embeddings, and positional encoding to process images, demonstrating competitive performance without relying on convolutions. These innovations are beneficial for geolocalization because of the understanding of global context, the ViT can capture more complex and varied visual patterns. A visual representation of the architecture of the ViT is shown in Figure 1.

Incorporating Diverse Data Sources. In parallel with advances in model architecture, integrating diverse data sources has proven to be another effective strategy for enhancing geolocalization performance. For example, Zhang et al. [20] showed that combining ground-view images with aerial images could improve geolocalization accuracy. However, the reliance on cross-view data introduces challenges for global scalability, particularly in regions where such data is scarce. In scenarios where cross-view datasets are unavailable, it becomes imperative to maximize the extraction of information from ground-view images alone. Clark et al. [2] addressed this by proposing a transformer-based approach that uses hierarchical cross-attention to better understand and differentiate between geographic hierarchies, such as countries, states, and cities. This method significantly enhances geolocalization by learning visual features specific to each geographic level.

Geographic Context in Image Processing. Understanding and processing images within their geographic context is critical for accurate geolocalization, especially when predicting the country where an image was taken. Pramanick et al. [9] demonstrated that transforming an image into a semantic segmentation map can improve generalization across various scenes, particularly when images are captured under diverse conditions. Their TransLocator model utilizes multi-task learning to predict both geographic location and environmental context, thereby improving the model's efficiency and performance across different landscapes.

Integrating Text and Visual Data. Beyond solely visual data, incorporating textual information alongside images offers a more comprehensive understanding of geolocation. Haas et al. [6] explored this by utilizing CLIP [10], a model developed by OpenAI that processes and understands images and text in tandem. By training CLIP on a vast dataset of internet images, paired with textual descriptions and geographic tags, Haas et al. leveraged the model's ability to associate textual and visual information, thereby enhancing geolocalization capabilities, particularly in accurately identifying image locations.

Our Approach: Combining ViT and CLIP for Country-Level Geolocalization. Building on these foundational works, our research aims to combine the strengths of the Vision Transformer (ViT) and CLIP embeddings to predict the country in which an image was taken. By integrating the self-attention mechanisms and patch embeddings of ViT with the visual-textual understanding capabilities of CLIP, we seek to capture the rich and diverse information embedded in images. These embeddings are then processed through a neural network with multiple hidden layers, which, as demonstrated by Uzair and Jamil [15], significantly enhances the network's ability to handle complex data and improve prediction

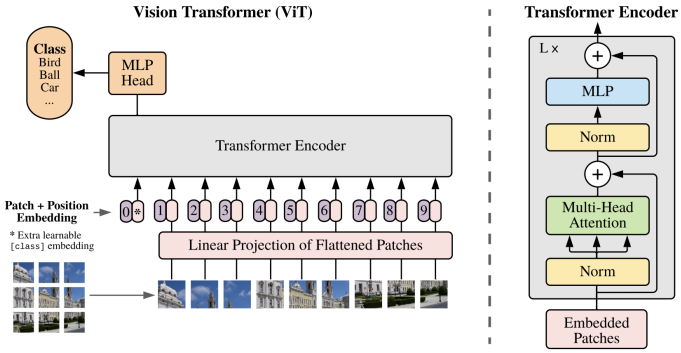

Figure 1: Vision Transformer architecture

accuracy. This hybrid model leverages the advanced feature extraction capabilities of ViT and CLIP, with the ultimate goal of achieving more reliable and accurate country-level geolocalization.

3 DATA

We utilized the dataset specifically curated for the MediaEval Placing Task 2016 [7], which is derived from the Yahoo Flickr Creative Commons 100 Million (YFCC100M)[14] dataset. This dataset includes 4.2 million geotagged images, each accompanied by geographic coordinates, making it an invaluable resource for developing models to predict an image's location based on visual features. The dataset encompasses a wide variety of visual content, ranging from well-known landmarks and landscapes to more ambiguous images with minimal geographical cues, such as photos of food and portraits. This diversity is crucial for advancing image recognition and spatial prediction technologies. The images' precise coordinates can be mapped to different levels of geographic categories—such as country, province, or city—depending on the classification task at hand. Figure 2 displays two images: one with distinctive architectural features and landscape elements unique to its location, making it easily identifiable if these characteristics have been previously encountered and recognized, while the other image is more ambiguous, lacking clear, recognizable features that could immediately indicate its location. Appendix A provides additional images illustrating examples of ambiguous cases, highlighting the challenges associated with this classification problem.

Data properties

The dataset presents a significant challenge due to the uneven availability of data across different areas worldwide, shown in Table 1 and Figure 4. While it offers a rich collection of geolocated images from numerous cities, certain regions are underrepresented, creating gaps in the data. This imbalances complicates efforts to develop comprehensive models or analyses, as the uneven distribution can lead to biased results and hinder the generalizability of findings.

The task involves classifying images based on the country they represent, utilizing a dataset that encompasses 228 countries. To evaluate the model's performance, the centroids of these countries

Figure 2: Two example images of the YFFC100M dataset for illustration.

serve as reference points. For achieving more fine-grained predictions, the model also considers administrative boundaries within countries, focusing on their subdivisions. These boundaries are sourced from the GADM database[1], which provides detailed information on global administrative areas, including the boundaries of countries and their subdivisions. Figure 3 illustrates the boundaries of the Netherlands, both at the country level and within its subdivisions with the centroids of the sub-division.

Preprocces

Preprocessing of the images from the dataset involves several steps. All images contain the exact latitude and longitude and need to be assigned to one of the countries and their sub-division. The GADM database exists of the country, sub-division and the geometry of the sub-division, as can been seen in Figure 3. For both the models the images needs to be preprocessed to receive the embeddings from the models. For the ViT model the images are resized to a resolution of 224x224 pixels and normalized across the RGB channels with mean 0.5 and standard deviation 0.5. Then the images are presented to the model as a sequence of fixed-size patches, 16x16 pixels, which are linearly embedded. The last hidden state of the model will be used as embedding for the image. The CLIP model uses two separate models, one for processing text and one for processing images. The

[1] https://gadm.org/about.html

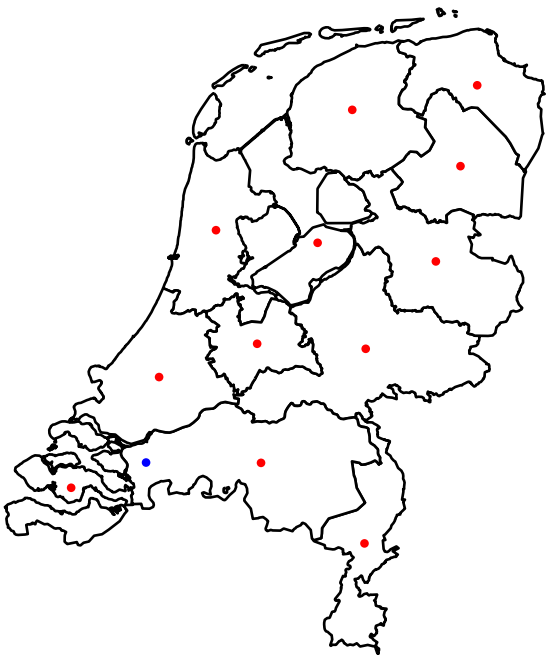

Figure 3: Boundaries of the Netherlands, the sub-division and the centroids. The red dots represent the centroids of a sub-division and the blue dot is a location of a photo.

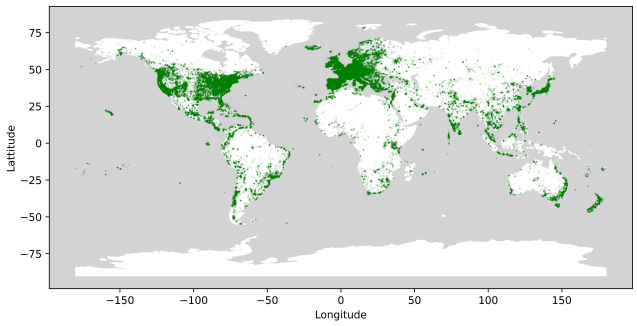

Figure 4: Global coverage of the dataset, with noticeable overrepresentation in certain countries.

model used for processing the images is a ViT. The embedding is a component that transforms an input image into a fixed-size vector in a high-dimensional space. This embedding captures features and semantics of the image.

Benchmark datasets

Benchmark datasets such as IM2GPS and IM2GPS3k have been crucial in image geolocation research. IM2GPS, one of the pioneering large-scale datasets for geolocation, includes images paired with GPS coordinates, serving as a fundamental resource for training and assessing models that estimate the location of an image based on its visual features. The dataset only contains 237 images. IM2GPS3k, an expanded version of the original IM2GPS, offers a broader and more diverse set of images from various geographic locations, facilitating more comprehensive testing of geolocation algorithms. In Appendix B examples images are displayed as illustration. These datasets are particularly valuable for benchmarking because they provide standardized evaluation criteria, enabling consistent comparison of model performance. In this research paper, we utilize both IM2GPS and IM2GPS3k as benchmarks to evaluate the effectiveness of our proposed geolocation model. By employing these datasets, we can assess how well our model generalizes across different geographic regions and environmental conditions, ensuring that our approach is both accurate and reliable.

4 MODEL DESCRIPTION

A baseline model and a neural network with multiple hidden layers were implemented and tested, utilizing embeddings from both the ViT and CLIP models. These models were evaluated across various metrics to identify the optimal combination of model architecture and embeddings for the geolocation of images. During training, the models were validated using a separate validation set to monitor performance and guide model selection.

ViT

The initial embedding used for the geolocalization task is derived from the Vision Transformer (ViT). Vision Transformers begin by splitting the image into patches, where each patch is assigned a learnable embedding. This embedding serves as the input to the Transformer encoder, representing the image. To retain positional information, a position embedding is added to each patch, enabling the model to grasp the spatial relationships between patches, much like the approach used in natural language processing (NLP). The Transformer encoder is a block consisting of alternating multi-headed self-attention (MSA) and multilayer perceptron (MLP), as described by Vaswani [16] and shown in Figure 1. This model relies on the self-attention mechanism, involving three key components: Queries, Keys, and Values. In ViTs, the MSA mechanism is applied in the same way as in the original Transformer model. The patch embeddings, along with their positional encodings, are linearly transformed to generate the queries (Q), keys (K), and values (V):

$$Q = XW^Q, K = XW^K, V = XW^V \tag{1}$$

Where X represents the input embeddings, and W^Q, W^K, W^V are the weight matrices. The MSA mechanism computes the attention scores, which determine the relationship between different patches in the image. The formula is given in equation 2, and is the dot product of the queries and keys, scaled by the square root of the key dimension, d_k and multiplied by the values matrix V.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

The softmax, shown in Equation 3, turns the similarity scores between the query and keys into a set of attention weights.

Table 1: The first table displays the number of photos per country, and the second table provides the number of photos per sub-division within the YFCC100M dataset. Both tables highlight the imbalances in the distribution of images across countries and sub-divisions.

Country	Number of observations
United State of America	10.633.534
Great Britian	341.392
Spain	196.374
France	163.572
...	...
Marshall Islands	8
Nauru	6
United States Minor Outlying Islands	1
Wallis and Futuna	1
Mean	17.605
Median	847

Sub-division	Number of observations
USA.5_1	341.392
GBR.1_1	278.834
USA.33_1	196.374
FRA.8_1	163.572
...	...
NGA.24_1	8
TUR.15_1	6
VNM.31_1	1
MYT.9_1	1
Mean	1.225
Median	53

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for} \quad i = 1, \ldots, K \tag{3}$$

This produces a weighted sum of the values based on the computed attention scores, this allows the model while processing the patches to focus on different parts of the image. The encoder consists of multiple layers of the MSA mechanism and a feed-forward neural network. The MSA mechanism allows the model to process the images in parallel while focusing on different parts of the image. The output of the MSA mechanism is transformed by a non-linear transformation in the feed-forward network. Finally, the output of the Transformer encoder is passed through a fully connected layer for the classification task, with the last layer being the classification token (CLS). This describes the architecture of a Vision Transformer, which leverages the MSA mechanism to capture dependencies across an image.

As describe using a pretrained model for obtaining the embeddings save time. In this paper we use the pretrained model from Google that is trained on the ImageNet-21k dataset[11], encompassing approximately 14 million images across 21,000 classes, it has learned to capture intricate and diverse visual features. This extensive pretraining enables the model to perform exceptionally well in computer vision tasks, such as image classification, object detection, and image segmentation. The feature representation from the pretrained model contains a wide variety of features that are useful for geolocalization. These embeddings capture high-level visual patters and structures. Furthermore, its robust and generalized feature extraction capabilities make it an ideal candidate for fine-tuning on specific datasets, allowing it to adapt to specialized applications while leveraging the comprehensive knowledge gained from its initial training. On the other hand is the use of a pretrained ViT a downside because of the model might focus on general visual patterns rather than subtle and distinctive visual cues that are essential for distinguishing between countries. Additionally, there may be a domain mismatch, as the features learned during pretraining may not align with the features needed for geolocalization.

ViT semantic segmentation

To enhance the network's understanding of a scene in an image, RGB images are converted into their corresponding semantic segmentation maps, see Figure 5. In this semantic segmentation process, the conditions of the scene, including daytime and weather, are generalized. HRNet[17] is used to generate these semantic maps, where pixels in the original RGB image are assigned to predefined object classes. The ViT then transforms the image produced by HRNet into an embedding, focusing more on the general details of the image.

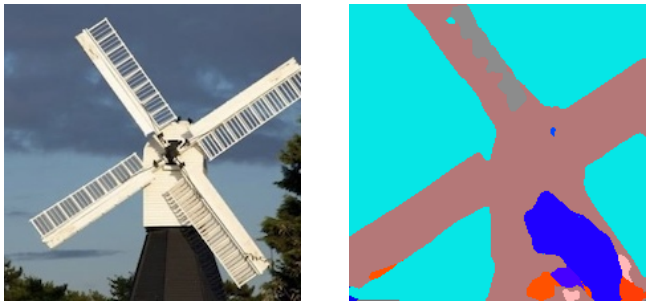

Figure 5: A RGB image with the semantic segmenation mapping.

CLIP

OpenAI created the CLIP (Contrastive Language–Image Pretraining) [10] model to bridge the gap between natural language understanding and computer vision. The concept laid the foundation for CLIP. The model is trained to understand both images and text by learning from paired image and text data. The main components of the model are the text encoder and image encoder. The image encoder uses a ViT, while the text encoder processes text with a transformer-based model similar to GPT. The goal of the contrastive learning objective is to bring the embeddings of matching image-text pairs closer together while pushing apart the embeddings of

non-matching pairs, as illustrated in Figure 6. By training CLIP to align visual and textual information within a shared embedding space using contrastive learning on large and diverse datasets, the model gains the ability to understand both visual and semantic content effectively.

Understanding and encoding an image into an embedding creates a space with a richer representation of the image. Similar to how Clark et al.[2] utilized additional information through a semantic segmentation map, we use CLIP to extract an embedding space that summarizes the image. CLIP utilizes a contrastive learning objective where it tries to maximize the similarity between the correct image-text pairs. This approach encourages the model to learn distinctive features that are effective for differentiating between different classes. By doing so, CLIP becomes proficient at identifying and distinguishing various objects and concepts within images. For the visual representations of the images, the CLIP ViT-L/14 model is used. This model model was pretrained on a dataset of 400 million images and captions. We employ CLIP because it excels in generalized zero-shot learning scenarios, making it ideal for image geolocalization of both familiar and unfamiliar. The multimodel understanding from CLIP is based on joint vision-language training. The large dataset of image-text pairs allows the model to have a multimodel representation that captures the relationship between images and their textual description. This enables CLIP to understand images in the context of language, which can help in gelocalization by linking visual cues to contextual descriptions.

CLIP is trained on a large dataset, not only for the task of understanding geographic aspects. This results that CLIP can capture general visual patterns and broad concepts and may lack to understanding geographic nuances that are crucial for the task of geolocalization.

Processing embeddings

Utilizing a ViT and CLIP embeddings as features for a neural network can significantly enhance the task of classifying countries based on images. The ViT model leverages self-attention mechanisms to capture intricate patterns and dependencies within image data, enabling it to produce high-quality embeddings that encapsulate complex visual features. When combined with CLIP embeddings, which are trained on vast datasets of text-image pairs to understand and represent multimodal information, the resulting feature set becomes even more robust. These embeddings encode rich semantic information that reflects both the visual and contextual nuances of the images. By feeding these comprehensive embeddings into a neural network, the model can effectively learn to distinguish between subtle country-specific visual cues, such as architectural styles, landscapes, cultural artifacts, and more. This approach not only improves classification accuracy but also enhances the model's ability to generalize across diverse and complex datasets.

Baseline

Processing the embeddings with a simple neural network serves in this paper as a baseline model. This model is used in Haas et al.[6] and Pramanick et al.[9] for classifying images on their embeddings obtained from the different pretrained models. This model consists of a fully connected layer between the embeddings and the target classes for classification. This linear layer can only model linear relationships between the features in the embedding and the output classes. Non-linear relationships might not be captured effectively without additional layers or non-linear activations. Using a single linear layer is computationally efficient and can perform well if the embeddings are already well-separated for different classes. However, it might struggle with more complex datasets where classes are not linearly separable in the embedding space. However, this baseline provides a starting point, allowing for the addition of more complex architectures in subsequent models to enhance performance and capture the detailed nuances within the data.

NN with hidden layers

The embeddings derived by pretrained models capture rich and complex information. Leveraging these embeddings as input for a NN requires a network with sufficient complexity to effectively capture and interpret these features. Pramanick et al. [9] and Haas et al. [6] utilized a simple neural network to process input embeddings, supplementing their models with auxiliary data. In contrast, our approach employs a more complex network architecture designed to process these embeddings. Our model features three hidden layers with sizes of 1024, 512, and 256 neurons, respectively, and concludes with a softmax function in the final layer to facilitate multiclass classification. The three hidden layer structure with a larger number of neurons at the first layer and narrowing it down to less neurons is a common choice. The first layer captures the wide range of features from the input embeddings. The model refines the features by processing the data to a layer with fewer neurons. This results in a compacter meaningful representation of the data. The decreasing layer sizes enable a smooth transition from the high-dimensional input space to the lower-dimensional output space, helping to create a structured mapping from inputs to outputs.

Performance evaluation

In this paper we use the predicted country error relative to the correct country as the primary metric. This approach contrasts with previous research, which typically segments the world into fixed-size geocells and uses the midpoint of each geocell for evaluation. In addition to assessing the model's accuracy, we evaluate whether the true target country appears within the top 5 and 10 predictions. The model assigns a probability to each country, and we count the instances where the true target country is included in these top 5 or 10 predictions. Consistent with prior research on image geolocalization, the "% @ km" statistic is analyzed for the performance of the models and embeddings. For a given dataset, the "% @ km" statistic measures the percentage of predictions that fall within a specified kilometer-based distance from the actual location. Similar to previous work, we assess this metric at three distance thresholds: 200 km (region-level), 750 km (country-level), and 2,500 km (continent-level). The 1 km (approximately street-level accuracy) and 25 km (city-level) thresholds are not used in this thesis, the classification task is conducted on country level. To calculate the distance between two points on Earth, the haversine formula is used. The formula for the Haversine distance d is calculated as follows:

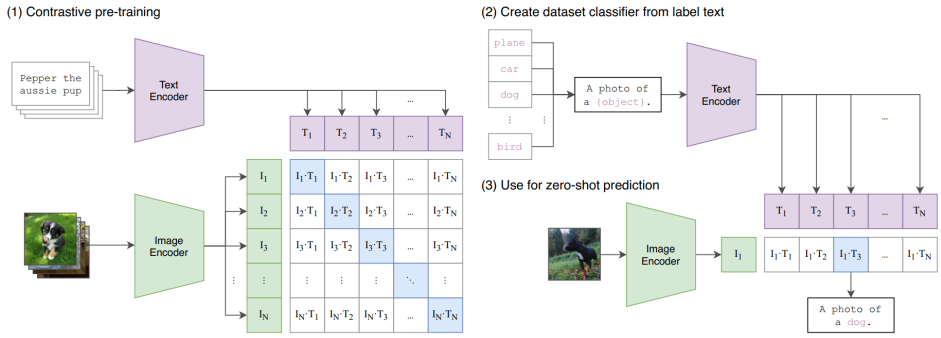

Figure 6: CLIP architecture

Table 2: Comparison of the different models against the different input embeddings on the validation dataset.

Model	Embeddings	Accuracy (%)	Mean Distance (km)	Top 5 (%)	Top 10 (%)
Baseline	CLIP	3.5	6205	6.9	29.3
	CLIP & ViT	22.0	4935	32.3	33.0
	ViT & seg ViT	3.1	6112	4.8	5.2
NN with hidden layers	CLIP	8.2	5107	10.2	28.1
	CLIP & ViT	26.4	4890	41.5	47.0
	ViT & seg ViT	3.1	6015	5.1	6.2

$$d = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \quad (4)$$

Where $\Delta\phi$ is the difference in latitude, $\Delta\lambda$ is the difference in longitude and ϕ_1, ϕ_2 are the latitude of centered point of area and the latitude of the image. As shown in Figure 3, where the red points represent the centered coordinates of the provinces and the blue point represents the coordinates of an image, the distance is calculated using the formula given in Equation 4.

5 EXPERIMENTAL DESIGN

To identify the best model for the geolocalization task, we conducted multiple tests on different models and evaluated them using various metrics. As discussed in Section 3 a general dataset of images is used to train the models. All the images are fed into the pretrained models explained in Section 4. The embeddings, CLIP, ViT and CLIP combined and ViT embeddings of the segmentation images and normal images are used as input for the different models.

(1) Baseline: a linear model between the input and output
(2) NN with hidden layers: a linear model with three hidden layers of size 1024, 512, 256

In this multiclass classification problem, the last layer of the models uses a softmax function, described in Equation 3. The softmax function converts the raw outputs of a model into a probability distribution across multiple classes. This allows the model to output the probabilities of each potential geographic location, indicating the likelihood that the image belongs to each location. The softmax function is used in conjunction with the cross-entropy loss function. The cross-entropy loss is designed to handle multi-class classification problems by comparing the predicted probability distribution over the classes with the actual distribution.

For the training of the model the dataset is split in a training dataset and a validation dataset, split 80/20. During the training a Stochastic Gradient Descent (SGD), Amari [1] optimizer iteratively updates the model's weights to minimize the Cross Entropy Loss. The network is fed with the different combinations of embeddings through its layers to generate predictions. These predictions are compared to the actual targets using the loss function. The gradient of the loss function with respect to each weight is computed through backpropagation, adjusting each weight to reduce the loss. SGD updates the weight by moving them in the direction opposite of the gradient, scaled by the learning rate. During the training mini-batches of size 128 are used to update the weights. After each epoch the model is used to predict on the validation set for the accuracy of the model. As long as the model is learning from seeing the data again this process continues iteratively over the epochs until the model converges.

For calculating the loss, the cross entropy is used. This criterion computes the cross entropy loss between the input normalized probabilities and target. The normalized input probabilities are the output of equation 3. This criterion is useful when the training dataset is unbalanced. The objective is to minimize the loss, meaning the model's predicted probability distribution should be as close as possible to the true distribution.

As described in Section 3, the data is imbalanced. Haas et al.[6] and Pramanick et al. [9] approaches the problem as a classification task by dividing the Earth into geographical cells (geocells), ensuring that each geocell contains a similar amount of data. For the problem we are solving it is not possible dividing the earth in more geocells. To get a finer prediction the sub-divisions can be used but borders for countries are fixed.

To address class imbalances during training, weights are applied using the inverse frequency method, where the weight for each class is defined as $w_i = \frac{1}{f_i}$. Here, f_i represents the frequency of class i. This approach assigns higher weights to less frequent classes and lower weights to more common ones, ensuring that the model pays more attention to underrepresented classes. Without these weights, the model is likely to predict only the class with the highest frequency, neglecting the minority classes. Since data imbalances often reflect real-world scenarios, using inverse frequency weights is an effective technique that minimizes the need for extensive data modification. The model employs a scheduler for the learning rate. At the start of the training process, the model needs to take larger steps to find a minimum in the optimization landscape. Once the model has located a minimum and progress stagnates, reducing the learning rate becomes beneficial. By dynamically adjusting the learning rate based on the training performance, this approach enhances the model's ability to continue learning effectively, leading to improved overall performance and generalization.

6 RESULTS

This section presents the results of testing two different models using three different combinations of embeddings for the geolocalization task. The baseline model that is a a linear layer between the input embeddings and the output classification layer and a NN with three hidden layers. The results are analyzed and discussed, providing an overview of the findings.

Evaluation and metrics

Table 2 presents a comprehensive comparison of two distinct models and their corresponding embeddings evaluated on a validation set across various performance metrics. The key metrics include accuracy, mean distance, and the frequency with which the true prediction ranks within the top 5 and top 10 most likely countries. Accuracy provides a straightforward measure of how often the models correctly identify the target country. The mean distance metric evaluates the average geographical deviation between the predicted and actual countries, offering insight into the precision of the predictions. The top 5 and top 10 metrics further assess the models by determining how often the correct country appears within the top 5 or 10 predictions, respectively. These metrics together provide a nuanced understanding of each model's strengths and weaknesses, enabling a clear comparison of their performance in capturing the true country with high confidence.

Baseline

Based on Parmanick et al. [9] and Haas et al. [6] the baseline model is a linear layer between the input and the output. Haas et al. uses auxiliary geographic, demographic and climate data to create the embeddings from the CLIP model. Parmanick uses the extra information from the semantic segmentation images from the pretrained ViT. The baseline will process the embeddings that are used in the other papers as the would do.

CLIP:. The baseline model with the CLIP input has an accuracy of 3.5% on the validation set. The true target country is among the top 5 highest probabilities 6.9% of the time, and within the top 10 probabilities 29.3% of the time.

CLIP + ViT:. The ViT embedding is added to the CLIP embedding, the feature space expands to 1536 dimensions: 512 from the CLIP embedding and 1024 from the ViT embedding. This combination significantly improves performance. On the validation set, the model achieves an accuracy of 22%, with the true target country appearing in the top 5 predictions 32.3% of the time and in the top 10 predictions 33.0% of the time.

ViT + seg ViT:. The embeddings derived from the RGB image and the semantic segmentation mapping from the ViT achieve the lowest scores in accuracy, top 5, and top 10 metrics. This suggests that the baseline model using the ViT and semantic ViT inputs is performing the worst.

NN with hidden layers

The model we present contains of three hidden layers, of size 1024, 512 and 256.

CLIP:. The more complex model can handle the CLIP features better than the baseline model. The accuracy increases to 8.2%, with the top 5 predictions rising to 10.2% and the top 10 predictions to 28.1%.

CLIP + ViT:. The combined embedding, when used as input for the model with three hidden layers, delivers the best overall performance. It achieves an accuracy of 26.4% on the validation set. While its mean distance isn't the lowest, it outperforms other approaches across all other metrics.

ViT + seg ViT:. Using a deeper neural network does not enhance the ability to classify countries based on the embeddings obtained from the ViT and semantic ViT.

Table 6 shows that the best-performing model on the benchmark datasets is the deep neural network model using CLIP and ViT embeddings as input. This model outperforms simpler architectures, shown in Table 2 demonstrating the effectiveness of leveraging complex, pre-trained embeddings in conjunction with a deeper network to enhance geolocalization accuracy. The results indicate that the combination of these embeddings captures more nuanced visual information, leading to improved performance across diverse geographic regions in the dataset.

Benchmark datasets

Table 2 demonstrates that the best-performing model is the neural network with three hidden layers. This model was applied to the benchmark datasets IM2GPS and IM2GPS3k, using the embeddings obtained from ViT and CLIP, with the results presented in Table 6. The performance was evaluated using the "% @ km" distance metric, as detailed in Section 4. Notably, the combination of CLIP

Table 3: The distance "% @ km" of the NN with hidden layers on the benchmark datasets.

Benchmark	Model	Distance "% @ km"		
		Region	Country	Continent
		200 km	**750 km**	**2500 km**
IM2GPS	NN with hidden layers + CLIP	4.4	18.3	56.2
	NN with hidden layers + CLIP + ViT	8.4	23.8	56.6
	NN with hidden layers + ViT + seg ViT	2.7	12.9	33.7
IM2GPS3k	NN with hidden layers + CLIP	2.5	15.4	48.2
	NN with hidden layers + CLIP + ViT	4.4	20.3	52.2
	NN with hidden layers + ViT + seg ViT	2.7	10.4	30.9

and ViT embeddings achieved the highest performance, with a "% @ km" accuracy of 23.8%.

7 DISCUSSION

This section we will interpret the results of the models and the embeddings in detail, examining the effectiveness of each embedding method and classification approach. The embeddings were designed to encode visual features as well as contextual elements. These embeddings were then input into a baseline model and a NN with hidden layers. We will analyze the challenges encountered, such as dataset imbalance and computational complexity. Furthermore, this section will address the limitations of the current methodologies, providing insights into potential improvements and suggesting directions for future research. We aim to contribute to the broader understanding of image-based location prediction and highlight the implications of our results within this field of computer science.

Major findings

Model complexity. The findings of this study offer several important insights into the effectiveness of different models and embeddings for geolocalization tasks. Firstly, the neural network (NN) with three hidden layers demonstrated superior performance compared to the baseline model across two of the three tested embeddings. In contrast, the embeddings derived from the semantic segmentation mapping, as proposed by Pramanick et al. [9], yielded the poorest results. This suggests that increasing the complexity of the model architecture can significantly enhance its capability to learn from and generalize geolocalized image data.

Embeddings. Secondly, while the embeddings generated from CLIP models did not individually lead to substantial improvements in classification accuracy, the most striking result was observed when CLIP embeddings were combined with ViT embeddings. This combination produced the highest performance, particularly excelling on the benchmark datasets according to the "% @ km" distance metric. This highlights the combination's superior ability to accurately predict geographic locations from images.

The study confirms that incorporating both CLIP and ViT embeddings within a model equipped with hidden layers allows the model to effectively capture and utilize diverse information crucial for accurate country-level classification. These results indicate that leveraging multiple pretrained embeddings alongside a more complex model architecture can lead to significant enhancements in geolocalization accuracy.

Limitations

Challenges in classifying images based on national boundaries. One significant limitation of this study lies in the inherent challenge of classifying images based on national boundaries. Countries are political and cultural constructs whose borders often do not align with distinct geographic or environmental features, making it difficult for models to accurately differentiate images from different nations. This challenge is particularly pronounced in regions where neighboring countries share similar landscapes, climates, and architectural styles.

Additionally, the diversity within a single country—spanning urban and rural areas, varying topographies, and different climate zones—further complicates classification. For instance, images from rural areas of one country may closely resemble those from a neighboring country rather than urban areas within the same nation. While subdividing countries into smaller regions could partially address this issue, this approach was not employed in this study due to the focus on predicting at the national level and the increased complexity that regional classification would entail.

Impact of uneven image distribution across countries. Moreover, the uneven distribution of images across countries exacerbates this problem. Some countries, especially those with major tourist attractions or high population densities, are overrepresented in image datasets, while others, particularly less populated or economically developed nations, are underrepresented. This imbalance can lead to biased models that perform well in predicting locations within overrepresented countries but poorly in those that are underrepresented. To address these challenges, an alternative approach could involve dividing the Earth into equal geographic areas, each containing the same number of images. This method would mitigate the issue of imbalanced representation by ensuring that each area contributes equally to the training process. By focusing on geographic rather than political boundaries, the model can better capture the natural variations in landscapes and environments that

are more relevant to location prediction. Additionally, dividing the Earth into equal areas allows for a more consistent and fair comparison across regions, as each area would be treated equally, regardless of its political or cultural significance. This approach could lead to more accurate and generalizable models that perform better across diverse landscapes, ultimately improving the robustness of image location prediction systems. However, this method also comes with its own challenges, such as defining the optimal size and shape of these geographic areas and dealing with the inherent diversity within each area. Balancing these factors is crucial to ensuring the effectiveness of this alternative approach.

Finding the best classification architecture. One limitation of this study is that it does not explore the full range of possible neural network architectures to identify the optimal design for the classification task. Due to constraints in time and computational resources, the focus was placed on demonstrating that even a basic architecture could yield significant improvements over existing methods. While this approach validates the effectiveness of simpler models, it leaves open the possibility that more complex architectures could further enhance performance. However, the study's findings suggest that substantial gains can still be achieved with relatively straightforward models, underscoring the potential of such approaches in resource-constrained settings.

8 CONCLUSION

In this study, we explored the effectiveness of different embeddings derived from pretrained models, specifically CLIP and ViT, in conjunction with both a baseline model and a NN with hidden layers for geolocalization tasks. The results revealed that the NN with three hidden layers outperformed the baseline model, particularly when using the combined embeddings from CLIP and ViT. This combination achieved the highest performance, especially on the "% @ km" distance metric, indicating its superior ability to accurately predict geographic locations. However, it's important to note that despite these improvements, the overall performance was not as strong as expected, suggesting that the model's learning process is still limited and not fully efficient for the task of geolocalization. This indicates that while the integration of multiple pretrained embeddings and increased model complexity can enhance accuracy, there remains significant room for improvement in optimizing the model to better handle the complexities of geolocalization.

Future research

While this study has made some progress in understanding the use of embeddings and model architectures for geolocalization tasks, several areas require further exploration and refinement.

Customize datasets for geolocalization. Future research might benefit from developing or selecting datasets more specifically tailored to the geolocalization task. However, curating such datasets may be challenging, as it requires extensive effort to ensure that the data is both relevant and balanced. There is also the risk that focusing too narrowly on a specific domain could limit the model's generalizability to other contexts.

Challenges in optimizing neural network architectures. Although different neural network architectures could potentially improve performance, finding the optimal architecture is a complex and resource-intensive process. There is no guarantee that alternative architectures will significantly outperform those tested in this study. Additionally, the fine-tuning of architectures for specific datasets could lead to overfitting, reducing the model's ability to generalize to new data.

Complexities of advanced embedding fusion techniques. Investigating advanced embedding fusion techniques may offer some improvement, but this approach also comes with risks. The integration of multiple embeddings can introduce additional complexity and computational overhead, with uncertain benefits. Moreover, finding the right method to combine embeddings effectively might prove to be a difficult task, with the potential for diminishing returns.

Exploring fine-tuning of pretrained models. Since the pretrained models are not specifically trained on a labeled dataset for the task of geolocalization, it would be interesting to explore the fine-tuning of models like ViT and CLIP. By fine-tuning these models on a dataset that includes a wide range of images from various geographic regions, it may be possible to significantly improve their classification accuracy.

In conclusion, this research underscores the potential of combining semantic and vision transformer embeddings in more complex neural networks for improved geolocation accuracy. The study demonstrates that even a basic neural network architecture can yield significant improvements, validating the approach despite not fully exploring more complex architectures due to time and resource constraints. The results suggest that while advancements have been made, future work is necessary to address the observed limitations and to fully realize the capabilities of these models in geolocating images with higher precision. Continued exploration of optimal architectures and fine-tuning techniques will be crucial in advancing the field further.

ACKNOWLEDGMENT

I would like to thank Britt and Ward for their time and patience throughout this process. Their support was essential in helping me complete this work.

REFERENCES

[1] Amari, S.-i. Backpropagation and stochastic gradient descent method. 185–196.

[2] Clark, B., Kerrigan, A., Kulkarni, P. P., Cepeda, V. V., and Shah, M. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes.

[3] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale.

[5] Gil, D., and Song, I.-Y. Modeling and management of big data: Challenges and opportunities. 96–99.

[6] Haas, L., Skreta, M., Alberti, S., and Finn, C. PIGEON: Predicting image geolocations. Publisher: [object Object] Version Number: 4.

[7] Larson, M., Soleymani, M., Gravier, G., Ionescu, B., and Jones, G. J. The benchmarking initiative for multimedia evaluation: MediaEval 2016. 93–96.

[8] O'Shea, K., and Nash, R. An introduction to convolutional neural networks.

[9] Pramanick, S., Nowara, E. M., Gleason, J., Castillo, C. D., and Chellappa, R. Where in the world is this image? transformer-based geo-localization in the wild.

[10] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. Publisher: [object Object] Version Number: 1.

[11] Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. ImageNet-21k pretraining for the masses.

[12] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge.

[13] Seo, P. H., Weyand, T., Sim, J., and Han, B. CPlaNet: Enhancing image geolocalization by combinatorial partitioning of maps.

[14] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. YFCC100m: The new data in multimedia research. 64–73.

[15] Uzair, M., and Jamil, N. Effects of hidden layers on the efficiency of neural networks. In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, IEEE, pp. 1–6.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need.

[17] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., and Xiao, B. Deep high-resolution representation learning for visual recognition. 3349–3364.

[18] Weiss, K., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. 9.

[19] Weyand, T., Kostrikov, I., and Philbin, J. PlaNet - photo geolocation with convolutional neural networks. vol. 9912. pp. 37–55.

[20] Zhang, X., Sultani, W., and Wshah, S. Cross-view image sequence geolocalization. Version Number: 2.

[21] Zhou, L., Pan, S., Wang, J., and Vasilakos, A. V. Machine learning on big data: Opportunities and challenges. 350–361.

APPENDIX

A Example images of the YFCC100M dataset







B Example images of the IM2GPS3k dataset