Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 02, February-2025

Proposed System for Deducing Location from Image

Vineet Prashant, Prof. Archana Kadam,
Sanwad Rashinkar, Sakshi Raut, Pooja Sadgir
Department Of Computer Engineering, PCCOE, Pune

Abstract **– Geolocalizing images on a global scale is a complex challenge, especially when dealing with the diverse landscapes and environments found in India. While traditional methods based on retrieval systems are giving way to deep learning, the potential of transformer-based models in this area remains largely untapped.**
In this work, we present a deep learning framework specifically designed for geolocalizing static images across Indian locations. Our approach incorporates cutting-edge techniques such as semantic geocell creation, label smoothing with haversine distance, Voronoi tessellation, and contrastive pretraining. These innovations significantly improve the accuracy of location predictions, offering a fresh perspective on image-based geospatial analysis for India.
By tailoring our methods to the unique characteristics of Indian geography, we provide a robust and precise solution for image geolocalization, pushing the boundaries of what's possible in this field.

Keywords:
Image Geolocation, Semantic Geocells, Deep Learning Transformers, Contrastive Pretraining, ProtoNets, GeoGuessr.

I. INTRODUCTION

Image geolocation—identifying an image's location using only its visual content—has gained critical importance in areas such as autonomous navigation, social media analysis, and geographic mapping. Earlier methods relied on retrieval-based techniques with hand-crafted visual features, but these have since been surpassed by deep learning models, especially convolutional neural networks (CNNs).
In this paper, we present a transformer-based deep learning model designed to improve geolocation accuracy, with a particular emphasis on Indian locations. Our approach introduces semantic geocell creation and contrastive pretraining to enhance featureextraction, achieving greater precision compared to existing methods. By incorporating deep multi-task learning, the model further refines its location predictions, offering a significant step forward in image-based geospatial analysis.

II. LITERATURE REVIEW

The field of large-scale image geolocalization has witnessed transformative advancements, transitioning from traditional approaches to sophisticated deep learning frameworks. Hays and Efros (2008) introduced the IM2GPS model, marking the first large-scale geolocalization system based on hand-crafted features and a retrieval-based method. Their work relied on nearest-neighbor search within a large dataset of geo-tagged images, laying the foundation for future developments in this domain [1].
The shift to deep learning, as emphasized by Masone and Caputo (2021), marked a paradigm shift, transitioning from the limitations of hand-crafted features to the versatility of learned features extracted through deep neural networks [2].
Google's PlaNet model (Weyand et al., 2016) was a watershed moment, leveraging convolutional neural networks (CNNs) for geolocalization at a planetary scale. By training on millions of geotagged images, PlaNet automated the feature extraction process, achieving state-of-the-art results and setting a benchmark for global geolocation accuracy [3]. This model not only demonstrated the robustness of CNNs in capturing geographic patterns but also introduced the importance of leveraging massive datasets for training.
The advent of transformer architectures in computer vision further revolutionized geolocalization tasks. The Vision Transformer (ViT) (Kolesnikov et al., 2021) adapted

Multi-task transformers, originally developed for natural language processing (NLP), to image-based tasks. ViT processes images as sequences of patches, enabling the model to capture long-range dependencies and contextual relationships [4].
The shift to deep learning emerged as a crucial enhancement to geolocalization models. Ranjan et al. (2016) demonstrated this through the HyperFace framework, where auxiliary tasks such as pose estimation and landmark detection improved feature extraction for the primary task. Bingel and Søgaard (2017) further underscored that task selection plays a critical role, where complementary tasks significantly enhance model performance [5][6]. For geolocalization, tasks like scene classification and landmark recognition serve as valuable additions to improve location inference.
Recent advancements in geocell partitioning have introduced new methods for dividing the world into meaningful regions. De Fontnouvelle (2021) explored the use of rectangular geocells, offering a structured approach for partitioning the globe for classification tasks. This method demonstrated that the choice of partitioning impacts downstream model performance, highlighting the importance of geospatial clustering [7].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 02, February-2025

The introduction of StreetCLIP (Haas et al., 2023) marked a new era in geolocalization, integrating the transformer-based CLIP model with semantic geocell division. By leveraging advanced clustering techniques and Haversine smoothing, StreetCLIP achieved unparalleled accuracy, outperforming previous methods and even expert-level players of GeoGuessr. Its real-world applicability further underscores the robustness and precision of transformer-based models in geospatial tasks [8].

A. COMMON FINDINGS FROM LITERATURE REVIEW

Imagenet and geo-tagged datasets: Large- scale datasets such as ImageNet and platforms like Google Street View play a pivotal role in geolocalization research. These datasets provide millions of labeled images, essential for pretraining and fine-tuning models. Notably, PlaNet and StreetCLIP utilized massive datasets of geotagged images to train their deep learning frameworks, underscoring the necessity of diverse and large-scale data for robust geolocation [3][8].

Benchmark datasets for image geolocalization: Early datasets such as IM2GPS (Hays & Efros, 2008) set benchmarks for model evaluation. PlaNet introduced a custom dataset with millions of geotagged images, raising the standards for geolocalization performance evaluation. Additionally, the inclusion of panoramic datasets, such as those used by StreetCLIP, highlights the evolution of benchmark datasets in representing real-world conditions [1][3].

B, COMMON EVALUATION PARAMETERS

Accuracy Of Geolocalization : Accuracy remains the most widely used metric, measuring how closely predicted geographic coordinates align with the true image location. PlaNet demonstrated substantial improvements over earlier methods, establishing a baseline for evaluating geolocalization accuracy at scale. Accuracy metrics, particularly for planetary-scale tasks, reflect the effectiveness of a model in capturing complex spatial relationships [3].

Distance-Based Metrics: Distance-based metrics, such as the Haversine formula, are critical for calculating the precision of geolocalization outputs. Haversine smoothing, introduced in StreetCLIP, refines cluster transitions, ensuring geospatial continuity. This approach not only enhances model precision but also mitigates artifacts caused by boundary discontinuities, further improving real-world applicability [8].

III. PROPOSED METHODOLOGY

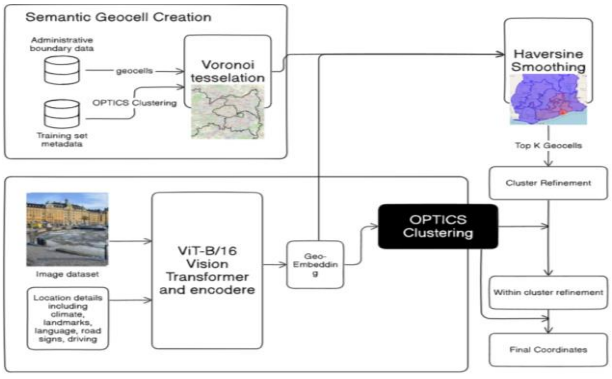

Fig. A. System diagram of proposed system

This proposed system introduces a structured framework for the vectorization of images and embedding location data. The system combines advanced clustering techniques, transformer-based visual encoding, and geospatial refinement to enhance geolocation accuracy and precision.

A. SEMANTIC GEOCELL CREATION

The first phase involves the systematic generation of geocells using administrative boundary data. Maps of cities, towns, and regions provide the baseline for defining geocell boundaries, ensuring alignment with real-world divisions.

Administrative Boundary Data: Using maps and metadata ensures that geocells correspond to meaningful geographical entities, offering a logical foundation for spatial analysis. By integrating administrative data, the system establishes clear boundaries for geospatial classification.

Optics Clustering: The OPTICS (Ordering Points To Identify the Clustering Structure) algorithm is applied within these boundaries to form density-based clusters. This method is particularly suited for spatial datasets, as it handles varying densities and complex clustering structures effectively, overcoming the limitations of traditional k-means or hierarchical clustering.

Voronoi Tessellation*:* Voronoi tessellation further divides the clustered space into regions around each geocell. These tessellations create unique boundaries for each cluster, enabling a structured representation of geographically separated regions. This step enhances the granularity of the spatial division, facilitating better regional analyses.

B. LOCATION DATA AND VISUAL ENCODING

This phase integrates image data with location-specific metadata, transforming visual information into meaningful geospatial embeddings.

Image Dataset and Location Metadata: A dataset of images, enriched with metadata such as landmarks, climate, road signs, and language, forms the input. Metadata provides additional context crucial for differentiating locations with similar visual patterns but distinct geographical attributes.

Vision Transformer (ViT-B/16): The Vision Transformer (ViT-B/16), known for its ability to process images as patch sequences, is employed to extract high-dimensional visual embeddings. Unlike traditional CNNs, ViTs capture long-range dependencies, offering a more holistic representation of spatial patterns and features.

Geo-Embedding Generation: Outputs from the Vision Transformer are fused with geolocation metadata to create geo-embeddings. These embeddings combine visual features with geospatial attributes, resulting in a robust representation suitable for diverse environments and geographical complexities.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 02, February-2025

C. OPTICS CLUSTERING ON GEO-EMBEDDINGS

The geo-embeddings undergo a secondary round of OPTICS clustering to refine the groups further. This step combines visual and geographical contexts, grouping geocells that exhibit both spatial and visual similarities. By leveraging the rich feature representations from the geo-embeddings, this stage ensures clusters are not only geographically coherent but also visually aligned.

D. HAVERSINE SMOOTHING

Refinement of clusters through Haversine smoothing ensures greater geographical precision:

E. TOP K GEOCELLS

Using the Haversine formula, distances between geocells are calculated, and the most significant clusters are selected as the "Top K" geocells. These clusters represent the most critical locations in terms of geospatial importance.

F. CLUSTER REFINEMENT:

Haversine smoothing adjusts cluster boundaries to respect real-world geographical continuity. This ensures that clusters reflect actual spatial relationships, avoiding artifacts introduced by earlier clustering steps.

G. FINAL REFINEMENT AND OUTPUT

The final stage focuses on improving the granularity and precision of the geolocation output: The system fine-tunes individual clusters, analyzing subsets of data for improved resolution. This step ensures that the clusters accurately capture fine-grained geographical details.

The refined output is a set of highly accurate coordinates, enriched with visual and spatial features. These coordinates address gaps in existing geolocation datasets and enable precise location predictions.

IV. ALGORITHMS

Following is a description of the important algorithms required.

A. CLUSTERING:

Clustering is a way of grouping similar things together. In geolocation, it's used to divide large areas into smaller regions to make analyzing and predicting locations easier. For instance, methods like K-Means divide an area into clusters based on how close data points are to each other, while other techniques like OPTICS focus on identifying dense groups of points, even if the clusters have irregular shapes. Sometimes, clustering is also done based on meaningful features like vegetation, road types, or building styles, which helps in making location predictions more accurate. By organizing data this way, clustering simplifies complex information and allows systems to handle large amounts of data more effectively. It's like breaking a big map into smaller sections to focus on one part at a time.

In geolocation systems, clustering is like creating a map with zones, where each zone represents similar characteristics, making it easier to understand and use the data.

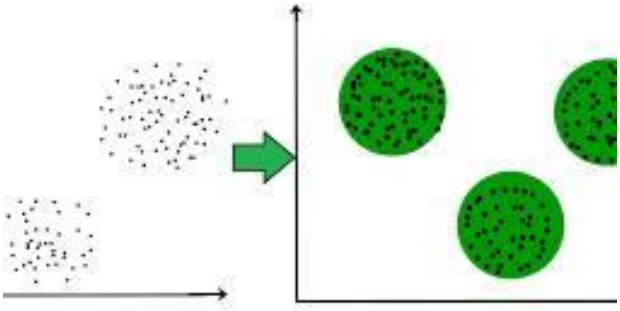

Fig.B. Clustering [13]

B. VISION TRANSFORMER (VIT):

The Vision Transformer (ViT) is a neural network model that works differently from traditional methods like CNNs. Instead of analyzing the entire image at once, it splits the image into small patches and processes them as sequences. By using a self-attention mechanism, ViT identifies relationships between these patches, allowing it to capture details from the entire image, not just local areas. This makes it especially useful in tasks like cross-view geo-localization, where images from different perspectives, like aerial and street views, need to be compared. ViT stands out because it includes position embeddings to understand spatial layouts, can model relationships across the whole image from the beginning, and can remove unnecessary parts of the image to focus on the important ones. However, it requires large amounts of data for training and can be computationally demanding. Techniques like adaptive regularization and selective cropping have been introduced to make it more efficient and applicable to real-world tasks. [7]

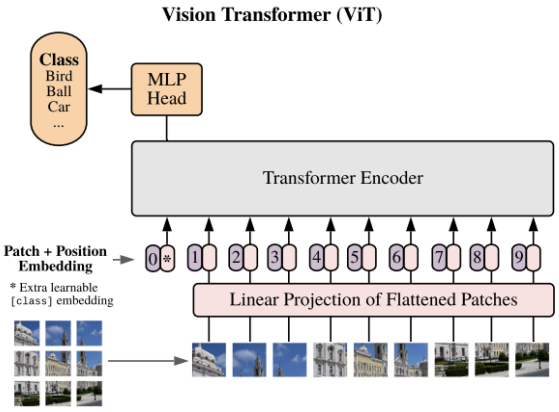

Fig.C. Vision Transformer(ViT)[15]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 02, February-2025

C. VORONOI TESSELLATION:

Voronoi tessellation is highlighted as a critical step in the Semantic Geocell Division Algorithm, which is designed to optimize geocells for image geolocalization tasks. This algorithm aims to subdivide geocells into smaller, semantically meaningful regions based on the geographic distribution of training samples. The process begins with a clustering step using the OPTICS (Ordering Points To Identify the Clustering Structure) algorithm, which identifies dense clusters of points within a geocell. If a significant cluster is found, Voronoi tessellation is applied to partition the geocell. This method creates polygonal regions around each point in the cluster, assigning the training samples to the nearest polygon. These polygons then define new geocells, which are balanced in size and retain semantic features such as urban and rural divisions or natural environmental contexts like vegetation and terrain. Importantly, this approach respects existing administrative boundaries and merges or splits regions adaptively to ensure a minimum number of training samples per geocell. This iterative refinement process continues until no further splits are needed. The resulting geocells are more granular and aligned with both geographic and semantic features, improving the accuracy and efficiency of geolocalization models. By combining clustering and Voronoi tessellation, this method enhances the interpretability and performance of geocell-based geolocalization systems.

Fig.D. Voronoi Tessellation [14]

D. SEMANTIC GEOCELL PARTITIONING:

Semantic geocell partitioning is an advanced method for dividing the Earth's surface into meaningful geospatial regions, optimized for tasks like geolocalization. Unlike traditional approaches that use uniform rectangular partitions, this method leverages semantic and geographic information to create geocells aligned with natural and administrative boundaries. The process begins with an initial setup based on detailed administrative regions, such as districts or states, which serve as the base geocells. Adjacent geocells are then merged iteratively to ensure each contains a minimum number of data points necessary for model training, prioritizing merges within the same administrative area or country. To refine these geocells further, the method employs the OPTICS clustering algorithm to identify dense clusters within a geocell, followed by Voronoi tessellation to divide it into smaller, semantically meaningful partitions. These partitions are validated to maintain balance in training data distribution while preserving semantic coherence. This adaptive partitioning captures distinctions like urban versus rural areas or natural landmarks, improving geolocalization accuracy by aligning model predictions with real-world features. While computationally intensive, semantic geocell partitioning ensures robust, semantically relevant divisions that enhance the effectiveness of machine learning models in spatial prediction tasks.

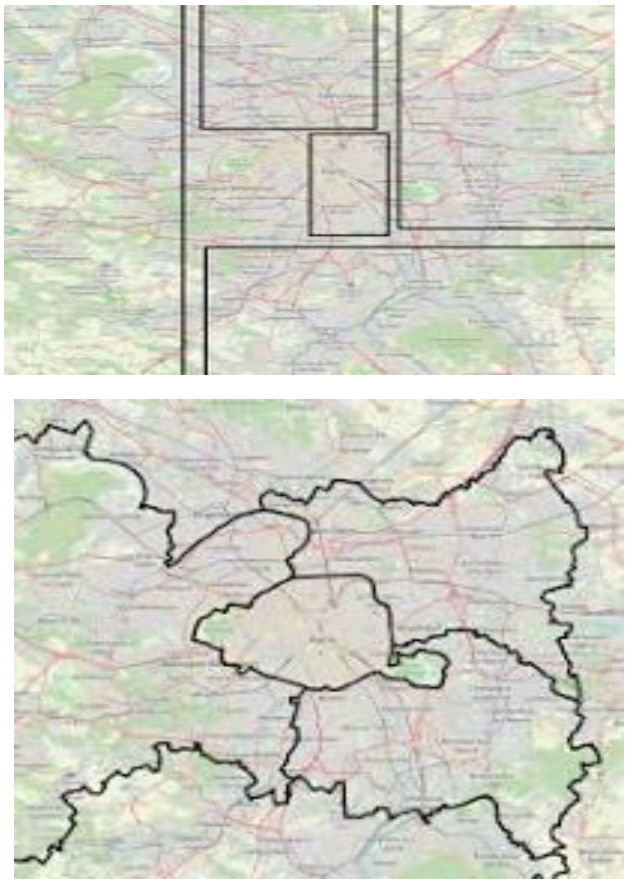

Fig.E. Semantic Geocell Partitioning[16]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 02, February-2025

E. LABEL SMOOTHING:

Label smoothing is a regularization technique used in machine learning to improve the generalization ability of classification models. Instead of assigning a hard one-hot label (e.g., [0, 1, 0] for a three-class problem), label smoothing distributes a small fraction of the label probability to the incorrect classes, resulting in a "smoothed" target distribution. For instance, a label of class 2 might be represented as [0.1, 0.8, 0.1] rather than [0, 1, 0]. This technique is particularly useful in mitigating issues like overfitting and overconfidence in model predictions. In the context of geolocalization, label smoothing becomes essential due to the complex nature of the classification task, where geocells represent geographic regions. The smoothing incorporates information about the distances between geocells by assigning higher probabilities to adjacent or semantically similar geocells. For example, when predicting the location of an image, the true geocell is given a high probability, but nearby geocells also receive smaller, nonzero probabilities based on their proximity to the true location. This approach acknowledges the continuity of geographic space, reducing sharp decision boundaries and encouraging the model to consider plausible alternatives.

The formula for label smoothing in geocell classification involves computing probabilities for each geocell based on the Haversine distance between the geocell's centroid and the true location. A temperature parameter controls the influence of distance on the probability distribution. By applying label smoothing, the model is trained more efficiently, as it learns shared characteristics of adjacent geocells, improving performance in regions with sparse training data and enhancing generalization across varied geospatial distributions.

V. CONCLUSION

The research done on image-based geolocation presents an opportunity with deep learning models, especially with the transformer variants, such as CLIP. They are likely to further improve image location prediction to much higher precision than the currently developed models, which are represented by PIGEON that may achieve close-to-perfect country-level precision while improving localization in proximity. Some techniques include semantic geocell creation, ProtoNet refinement, and contrastive pretraining, which might be key to overcoming such challenges. Nevertheless, challenges including management of visually similar environments, computational requirements, and gaps in data availability will continue to face such advancements. Explorations into these methods continue with the goal of developing models for use in practical applications, especially in underrepresented regions and complex environments. The proposed improvements may lead to the development of a robust system that can deliver high precision across various scenarios.

ACKNOWLEDGMENTS

We express our sincere thanks to our Technical Seminar-I Guide Prof. Archana Kadam for her encouragement and support throughout our seminar, especially for the useful suggestions given during the course of the seminar and having laid down the foundation for the success of this work. We would also like to thank our Research & Innovation coordinator Prof. Dr. Reena Kharat and Technical Seminar Coordinator Prof. Bodireddy Mahalakshmi for their assistance, genuine support and guidance from early stages of the seminar and during the entire course of this seminar work. We would like to thank Prof. Dr. Sonali D Patil, Head of Computer Engineering Department for her unwavering support during the entire course of this seminar work.

REFERENCES

[1] Berton, G., Masone, C., and Caputo, B. “Rethinking Visual Geo-localization for Large-Scale Applications, 2022a.” URL https://arxiv.org/abs/2204.02287.
[2] Hays, J. and Efros, A. A. “IM2GPS: estimating geographic information from a single image.” In Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2008.
[3] L. Haas, M. Skreta, S. Alberti and Chelsea Finn: “PIGEON: Predicting Image Geolocations”, 2024 URL https://arxiv.org/abs/2307.05845
[4] Weyand, T., Kostrikov, I., and Philbin, J. “PlaNet- Photo Geolocation with Convolutional Neural Networks”. In Computer Vision– ECCV 2016, pp. 37–55. Springer International Publishing, 2016. doi: 10.1007/ 978-3-319-46484-8 3.
[5] Ranjan, R., Patel, V. M., & Chellappa, R. (2016). HyperFace: A multi-task learning framework for face detection and other tasks.
[6] Bingel, J., & Søgaard, A. (2017). Impact of auxiliary tasks in multi-task learning on main task performance.
[7] Kolesnikov, A., Dosovitskiy, A., et al. (2021). ViT: Adapting transformers from NLP to vision tasks, including geolocation. ICLR.
[8] de Fontnouvelle, P. (2021). Geocell partitioning for geolocalization tasks
[9] Masone, C., & Caputo, B. (2021). Survey of deep learning-based visual place recognition and image geolocalization. ACM Computing Surveys.
[10] Haas, L., Skreta, M., Alberti, S., & Finn, C. (2023). PIGEON: Planet-scale image geolocalization using CLIP and refined geocell division
[11] Clark, A., Deng, Y., & Cheung, B. (2022). Transformer-based model for global geo-localization using geographic hierarchies. ECCV.
[12] Tanner, J., et al. (2022). Intra/inter-city geo-localization in urban environments using deep learning. ICCV. [13] Vo, N. N., Jacobs, N., & Hays, J. (2017). Combin