Geo-Localization Using Multimodal Large Language Models

Lorenzo Alvisi[1,2][0000−1111−2222−3333]

[1] IMT School for Advanced Studies, Lucca, Italy lorenzo.alvisi@imtlucca.it
[2] Institute of Informatics and Telematics, National Research Council, Pisa, Italy lorenzo.alvisi@iit.cnr.it

Abstract. Image-based geolocalization is a crucial task with many applications, ranging from environmental monitoring and disaster response to navigation through fighting disinformation; however, assigning a precise geographical location to an image based solely on visual content involves overcoming multiple significant challenges. The first one lies in the variety of landscapes ranging from urban areas to remote zones, such as deserts or forests, each with distinct visual characteristics. Additionally, the same place could differ in weather, seasons, lighting, and angles thus further complicating the task. Machine learning tackled this problem by trying to locate images in a specific region, or biome, of interest. A global approach has been attempted from just a handful of studies, such as [8] or [6]. Usually, the training of these models is a very challenging and expensive task, so since it is a common trend in computer science to use Large Language Models (LLMs from now on) to solve tasks that previously required specific training, we tried a new approach based on these new generative models. Moreover, after the advent of multimodal models, such as ChatGPT-4o [10] or LLava [7], it is possible to extend the analysis to images.

In this paper, our goal is to evaluate how these models perform in global geolocalization. We extracted 2,000 images from a pre-existing dataset [3] composed of 10,000 Google tool Street View images along the coordinates of the location where they were taken. These images were mainly taken in remote areas from different biomes, such as deserts, forests, and other similar environments, as shown in Figure 2, but also include some urban scenes. We limited the dataset to 2,000 images, similarly to [8], where the same number of images was utilized. To process the dataset we supplied both ChatGPT-4o and LLava with the prompt, shown in Figure 1, and one of the images in our data. The output of our models gave us three predictions containing both the approximate coordinates of the location and the state they were in, then we chose just the most accurate one. Each image is analyzed separately from the other to avoid context contamination between different iterations of the process. Due to computational performance and unstable outputs, we narrowed down the dataset used for LLava to 250 images.

Our results, especially for ChatGPT-4o are promising. We found that ChatGPT-4o correctly guessed the state approximately 76.5% of the time. This high success rate is reinforced by the fact that in the 23.5% of cases where the model guessed the state incorrectly, 60.5% of these

incorrect guesses were within 750 km of the correct state. This indicates that even in those cases where the state is mistaken, it often provides a geographically accurate prediction. Contrary LLava performed quite poorly on this task: we guessed correctly the state only in the 20.9% of the cases and only 24.5% of the incorrect guesses were within 750 km of the correct state. Due to the different methods used, we needed to adapt our metrics to compare our results to the state-of-the-art. The first metric followed the existing literature and used the distance between real and computed coordinates as a proxy for different levels of geographical accuracy (750 km is the radius threshold for the state), the second checked if the country of the input coordinates matches the state obtained as input and the third is the union of the two previous cases. As shown in Table 1, we obtained results similar to the state-of-the-art since our results are better than [6] in one metric out of the three, and better than [5] in two out of three.

Finally, it is also important to underline the effect that these results will have. Given the usage simplicity, its versatility, and its efficiency this type of geolocation can impact many different fields. It can easily increase situational awareness during crises, enabling emergency services, law enforcement, and professionals to identify where incidents are occurring to respond more effectively. For example, these models could be used as threat mitigation in scenarios, such as those described in [1], or during natural disasters [11]. Within realms like open-source intelligence (OSINT), this type of geolocation can help track and analyze threats, gather information, and improve the accuracy of analyzed data. Moreover, it can help verify the authenticity of information found on social media, as the usage of unaltered images in a new, but false context is one of the most prevalent methods to mislead the public. By cross-referencing the claimed location of an event with actual geolocation data, anyone can confirm whether the information is credible thus reducing the impact of out-of-context images. This result is especially helpful as online coordinated behaviour has already proven to spread disinformation during the 2019 UK election [9] and the 2020 United States election [12], and while the disinformation can easily spread on social media, such as telegram [2] and twitter [4]. However, there are notable drawbacks to consider, concerning potential violation of privacy. Geolocation data can be misused to harm individuals, revealing their private information and location without consent. This can lead to dangerous activities such as swatting, where, due to false reports the victim is raided by police, but also it can put individuals participating in witness protection programs at great risk.

Keywords: Large Language Models · Open Source Intelligence · Geolocalization.

Prompt

I will give you some images and you will try to locate them. I am aware that some predictions could not be precise so don't worry about blurriness, accuracy or absence of landmarks. You will follow the following rules, and you won't break them for any reason.
Rule 1: You will ALWAYS give me 3 predictions based on the given image. If you are not sure or able to locate them with precision or the image is blurred or in low resolution, still give me your best picks.
Rule 2: For each of the 3 predictions you MUST return me a python readable JSON. Each JSON will have the following keywords: national_state and coordinates. The coordinates must be formatted as a list of latitude and longitude, the national_state must be the nation name, and each field must be non-null.

Fig. 1: The prompt we used for both ChatGPT and LLava

Fig. 2: Dataset: Example of the images in our dataset

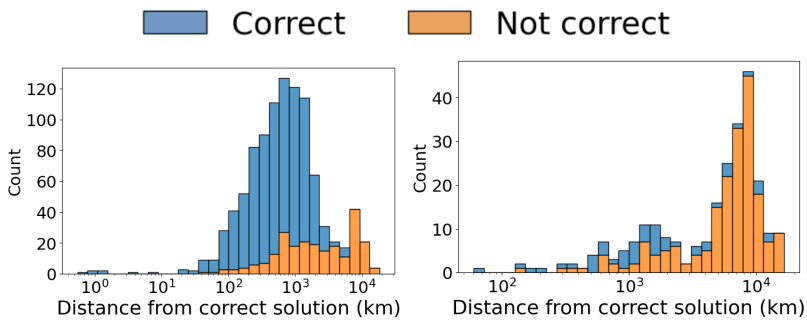

Fig. 3: Result: Distribution of the distances between the localization of the correct solution and the localization of the answer given by chatGpt (left) and LLava (right).

References

1. Alvisi, L., Bianchi, J., Tibidò, S., Zucca, M.V.: Weaponizing disinformation against critical infrastructures (2024)

	Country	**750 km**	**Country or 750 km**
ChatGPT-4o [this work]	76.5 %	55.2 %	**82.4** %
LLava [this work]	20.9 %	8.3 %	24.6 %
PIGEOTTO [6]	**82.3** %	**82.3** %	82.3 %
StreetCLIP [5]	74.7 %	74.7 %	74.7 %

Table 1: Comparison between the accuracy of our prompt (ChatGPT-4o and LLava) and the accuracy of other models (Piegeotto and StreetCLIP).

2. Alvisi, L., Tardelli, S., Tesconi, M.: Unraveling the italian and english telegram conspiracy spheres through message forwarding. arXiv preprint arXiv:2404.18602 (2024)
3. Chambaz, P.: Google street view: A curated dataset of google street view images (2022), `https://www.kaggle.com/datasets/paulchambaz/google-street-view`, accessed: 2024-06-10
4. Gambini, M., Tardelli, S., Tesconi, M.: The anatomy of conspiracy theorists: Unveiling traits using a comprehensive twitter dataset. Computer Communications 217, 25–40 (2024)
5. Haas, L., Alberti, S., Skreta, M.: Learning generalized zero-shot learners for open-domain image geolocalization (2023)
6. Haas, L., Skreta, M., Alberti, S., Finn, C.: Pigeon: Predicting image geolocations (2024)
7. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), `https://llava-vl.github.io/blog/2024-01-30-llava-next/`
8. Muller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation estimation of photos using a hierarchical model and scene classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 563–579 (2018)
9. Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., Tesconi, M.: Coordinated behavior on social media in 2019 uk general election. In: Proceedings of the international AAAI conference on web and social media. vol. 15, pp. 443–454 (2021)
10. OpenAI: Chatgpt: Generative pre-trained transformer (gpt-4o). `https://www.openai.com/chatgpt` (2024), accessed: 2024-06-12
11. Suwaileh, R., Elsayed, T., Imran, M.: Role of geolocation prediction in disaster management. In: International Handbook of Disaster Research, pp. 1–31. Springer (2023)
12. Tardelli, S., Nizzoli, L., Avvenuti, M., Cresci, S., Tesconi, M.: Multifaceted online coordinated behavior in the 2020 us presidential election. EPJ Data Science 13(1), 33 (2024)