

Improving Geolocalization with Social Media Captions: A Retrieval-Based Approach

Anton Ljungar
FOI Swedish Defence Research Agency
Stockholm, Sweden
anton.ljungar@foi.se

Sofia Reineck
Basalt AB
Enköping, Sweden
sofia.reineck@basalt.se

Edward Tjörnhammar
FOI Swedish Defence Research Agency
Stockholm, Sweden
edward@foi.se

Sebastian Öberg
FOI Swedish Defence Research Agency
Stockholm, Sweden
sebastian.oberg@foi.se

Abstract—The ability to predict a geographical location based on the content of a single image is a challenging problem requiring vast knowledge of the world. In recent years, vision transformers have been effectively utilized to automate this process and have achieved superhuman performance. However, previous work has been constrained to only look at images, ignoring the fact that many images posted on social media are accompanied by captions. In this work, we introduce GeoCIS, a retrieval-based multimodal geolocalization model capable of making predictions using images and text with a novel joint caption-image embedding vector. To achieve this, GeoCIS was trained on a dataset of 2.4 million caption-image pairs from the Russian social network VK, collected from the countries of Russia, Belarus, Ukraine, Moldova, Kazakhstan, and Georgia. We develop three variants of GeoCIS, each trained using one of three different CLIP-like foundation models. For each foundation model, we also train an ablated variant using only images to assess the contribution of captions to the overall performance. One of the foundation models is designed for Russian, the primary language used on VK, while the other two are designed for English, using automatically translated captions. This allowed us to evaluate how language localization, as opposed to translation, influences performance. When the GeoCIS variants were evaluated against their no-caption ablation, they achieved the best results on all distance metrics. Similar results were observed on a manually curated dataset where the captions did not contain any overt geographical clues such as country names, cities and similar references. In contrast, the effects of language localization were less conclusive. Our work thus clearly shows that social media captions improve the geolocalization performance of a multimodal model, even if the captions themselves do not contain clear geographical clues.

Index Terms—Geolocalization, Visual Place Recognition, Cross-Modal Retrieval, Social Media, Russia, CLIP, VK

I. INTRODUCTION

With the rise of social media, images are posted from all around the world publicly online daily. These images constitute a vast source of information and potential for open-source intelligence (OSINT), and the ability to perform geolocalization on these images can be of great importance. The problem of image geolocalization, however, is challenging and relies on subtle visual clues and general knowledge of the world. Over the past 15 years, it has been approached

in various ways, often achieving new state-of-the-art results. These methods include image retrieval [2], using convolutional neural networks (CNNs) and a hierarchical geocell division [3], using vision transformers [4] and more recently by using large language models (LLMs) [5]. However none of the suggested methods have tried to use the additional information that can be gained from looking at the accompanying caption of an image posted on social media. These captions can hold a lot of relevant information, including the exact location of the image, and as such, we propose that they be used when trying to geolocate images coming from social media. Thus, we introduce GeoCIS, a multimodal geolocalization model capable of making predictions using images and text. GeoCIS was trained on a dataset of 2.4 million caption-image pairs from the social network VK, collected from the countries of Russia, Belarus, Ukraine, Moldova, Kazakhstan, and Georgia. By building GeoCIS, we set out to answer the following question:

How can the geolocalization capability of a retrieval-based multimodal model be improved by including social media captions?

The remainder of this paper is organized as follows: Section II presents related work in geolocalization and classification of social media image-caption pairs, providing context for the methods used in the paper. Section III describes our proposed model and methodology. In Section IV, we present and discuss our results. Section V concludes the paper and outlines future research directions. Finally, Section VI briefly discusses the ethical considerations of our work.

II. RELATED WORKS

To understand how geolocalization of images can benefit from the use of social media captions, one needs to understand the field of image geolocalization and how it relates to social media captions.

A. Image Geolocalization

According to [1], one can sub-divide image geolocalization into approaches that only use ground images, only use aerial

images or cross-view approaches which use both types of images. In this paper, we are only interested in images coming from social media and they almost entirely consist of ground images. As such, we will be focusing our efforts into the field of ground image geolocalization. We further align ourselves with the conclusions of [5] and [6] that methods in ground image geolocalization are preferably divided into the categories: retrieval-based, classification-based and generation-based.

Retrieval-based methods use a database of images or GPS coordinates to search for the images or GPS coordinates that best match the query image. The first modern take on image geolocalization was the retrieval-based method called IM2GPS [2] in 2008, which used hand-crafted features to represent each image. More recently, retrieval-based methods are often incorporated as a part of a bigger algorithm, like in the case of PIGEON and PIGEOTTO [13], however in the GeoCLIP [7] paper they created a shared embedding space for GPS coordinates and images which was used to search for GPS coordinates matching the query image.

Classification-based methods instead turn the problem of image geolocalization into a classification problem. The problem is solved by partitioning the Earth's surface into smaller non-overlapping cells, or geocells, and then training a model to predict which geocell a query image belongs to. The center coordinate of the predicted geocell is often chosen as the resulting GPS coordinate. Classification-based methods became popular with the emergence of deep learning, with the PlaNet [3] paper being one of the first ones in 2016. The partitioning of the world in these first papers was achieved by creating geocells given by naive rectangular grids. However [8] and later [13] recognized that results could be improved by using semantic geocells, which are geocells divided by territorial and/or natural occurring borders.

Generation-based methods incorporate a large language model (LLM) as a part of the geolocalization pipeline. In the Img2Loc [9] paper from 2024, they introduced the idea of using an LLM for geolocalization by letting a retrieval component create a list of candidate images from a query image and then letting an LLM use this list to predict the correct coordinate of the query image. This idea was improved upon in the G3 [5] paper, which combined the techniques from Img2Loc with those from GeoCLIP. In the GeoRanker [6] paper, this was further improved by letting an LLM re-rank the list of candidate coordinate locations before the final prediction.

However, none of the aforementioned methods use social media captions to improve their ability to geolocalize.

B. Classifying image-caption pairs from social media

Although research on how to geolocalize social media posts using both image and caption text is limited, there is active research on classifying social media posts using both image and caption text from the CrisisMMD dataset [10]. The CrisisMMD dataset contains 16058 manually annotated social media posts from Twitter collected during different natural disasters. Each social media post has been labeled according

to three categories: informativeness, humanitarian, and damage severity. In [11], they try to correctly classify these images by finding features for images and text separately, by using CNNs [18] and skipgrams [17] respectively, and then combining these features in a neural network consisting of fully connected layers. In [12], they used a similar approach, but the features were instead coming from transformer-based models, and the fusion of the features was also transformer-based.

III. METHODOLOGY

GeoCIS is built by fine-tuning a foundation model based on CLIP [15] with a large amount of caption-image pairs from VK. This makes it possible for GeoCIS to, given a previously unseen caption-image pair, retrieve the most similar caption-image pair from the training dataset based on cosine similarity. The predicted coordinate will be the coordinate associated with the retrieved caption-image pair. A schematic of the architecture of GeoCIS as well as our data collection pipeline can be seen in Figure 1. In the following sections, we will go more into the details of our data collection and model pipeline.

A. Data Collection

VK is a Russian social media platform that is popular within the Commonwealth of Independent States (CIS).¹ VK was deemed an interesting platform to work with because of its open API and its many images with accompanying captions and GPS coordinates. To collect the data from VK, the CIS countries were all divided into semantic geocells by using the Database of Global Administrative Areas (GADM) [19]. The second administrative level was chosen for the geocell division, resulting in a total number of 3472 geocells. Given a time frame, a GPS coordinate and a radius, a search for all images from VK matching those constraints can be made. Such a search was made for each geocell, using the geocell's center coordinate together with a radius that would make the corresponding area cover the entire geocell, and each month from the 1st of February 2012 to the 1st of March 2024. By selecting only images with a corresponding caption, 5,474,626 caption-image pairs with corresponding GPS coordinates were downloaded. The collected data is visualized in a heat map in Figure 2.

B. Data Filtration

Not all images that were downloaded were suitable for geolocalization. For instance, a large amount of caption-image pairs could be described as classified ads, similar to posts found on the Facebook Marketplace² platform. Hence, we filtered the images by classifying them into one of twenty different categories using the CLIP model

¹The CIS is an intergovernmental organization formed after the dissolution of the Soviet Union, with the countries of Armenia, Azerbaijan, Belarus, Kazakhstan, Kyrgyzstan, Moldova, Russia, Tajikistan, Ukraine and Uzbekistan as either founder or member states. Ukraine and Georgia ended their participation in the CIS after the Russo-Ukrainian War and the Russo-Georgian War, respectively.

²<https://www.facebook.com/marketplace/>

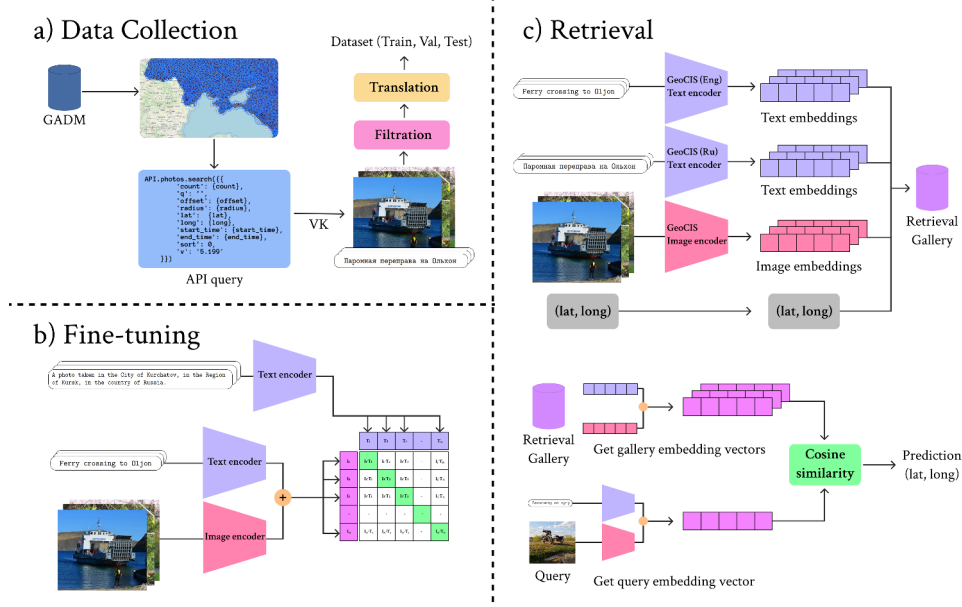


Figure 1. Overview of the GeoCIS model and data collection pipeline.

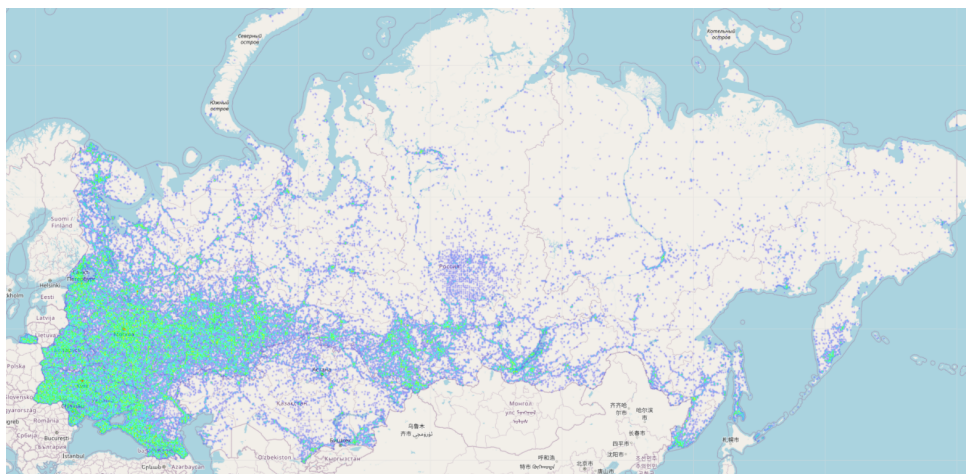


Figure 2. A heat map showing the distribution of the collected 5,474,626 caption-image pairs downloaded from VK.

clip-vit-base-patch32³. Seven of these categories led to the image being filtered out, and the other thirteen kept the image. After the filtering, 2,564,096 caption-image pairs remained. The number of images in each category is shown in Figure 3.

C. Data Curation

The resulting dataset was split into several smaller parts. The split was based on the owner IDs of the corresponding social media posts. Each caption-image pair on VK comes with a unique owner ID, indicating which user is the owner of the caption-image pair. 95% of the owner IDs, and their

corresponding caption-image pairs, were used to create a training dataset called VK2M, resulting in 2,444,288 caption-image pairs. A validation dataset was created called VK60kVal containing 58,880 caption-image pairs from 2.5% of the owner IDs. A test dataset was created called VK60kTest containing 60,928 caption-image pairs, again from 2.5% of the owner IDs.

To prevent the same image from appearing in both the training dataset and the other datasets, because different user could possibly post the same images, a second filtration was applied. Images were again processed by the CLIP model clip-vit-base-patch32, and if an image from VK58kVal or VK60kTest was found to have a cosine similarity

³<https://huggingface.co/openai/clip-vit-base-patch32>

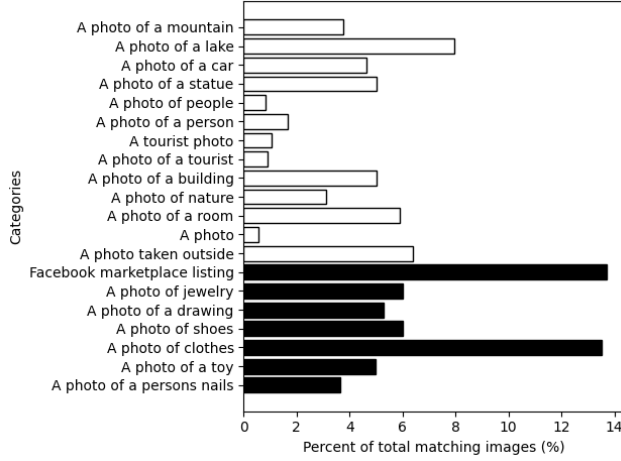


Figure 3. A plot showing the percentage of images from the VK dataset that match each of the twenty filtration categories. Images matching a black category are filtered out, and images matching a white category are kept.

of 0.999999 or higher with a image from VK2M, it was removed from VK60kVal or VK60kTest respectively. This resulted in VK60kVal having 57,222 caption-image pairs and VK60kTest having 59,134 caption-image pairs in the end.

Finally, a test dataset, VK1280Test, was created, containing 1,280 caption-image pairs. VK1280Test was created from a subset of 1,897 caption-image pairs from VK60kTest, where the posts were manually curated not to contain any easily searchable terms in their captions, such as country names, city names, famous buildings, restaurants, tourist attractions, parks, rivers, mountains and more. See Table I for a summary of the different datasets.

Table I
THE DIFFERENT DATASETS AND THEIR SIZES

Dataset	caption-image pairs
VK2M	2,444,288
VK60kVal	57,222
VK60kTest	59,134
VK1280Test	1,280

Since we wanted to fine-tune two English variants of GeoCIS on the data, and the captions were predominantly Russian, we had to translate all the captions for the English model. Translation was performed using the MADLAD-400 translation model MADLAD-400-3B-MT⁴ [21] using a maximum sequence length of 200 tokens for the input and 100 tokens for the output. All captions were translated from various original languages to English.

To fine-tune our models, we needed synthetic captions that explicitly contained information about the geographic location where the images were taken. Thus, for each image and each

language, an associated synthetic caption was created based on the geocell of that image. The synthetic captions were created using the names of the top three administrative divisions, as well as their corresponding cell type. In the English case, the cell types were “Country,” “Region,” “Republic,” “Territory,” “City,” “District,” and so on. We limited it to the top two administrative divisions for Moldova due to a lack of a more granular division in the GADM dataset, which is where we got all the English place names and their geographic information. Place names in Russian were instead sourced from GeoNames [20]. An English example with its corresponding Russian version is shown below:

A photo taken in the City of Kurchatov, in the Region of Kursk, in the country of Russia.

Фотография сделана в городе Курчатов, в Курской области, в стране Россия.

A summary of the data collection pipeline can be viewed in Figure 1 a).

D. Fine-tuning

Three different variants of GeoCIS were developed based on three different foundation models: the original CLIP model clip-vit-base-patch14-336⁵, StreetCLIP⁶ [14] and the RuCLIP model ruclip-vit-large-patch14-336⁷ [16]. StreetCLIP is a fine-tuned version of the same CLIP model, further fine-tuned on 1.1M Google Street View⁸ images specifically to be better at geolocalization. RuCLIP is a foundation model based on CLIP that, unlike CLIP, is trained on solely Russian text.

Instead of fine-tuning our GeoCIS variants by matching images with geolocation-indicative texts, like in [13], we fine-tuned them by matching caption-image pairs with geolocation-indicative texts. In order to achieve this we used a novel joint caption-image vector, where we averaged the output embeddings from the text encoder and image encoder of the CLIP-based foundation model. We trained our three foundation models, CLIP, StreetCLIP and RuCLIP, to associate this joint caption-image vector with the corresponding synthetic caption by fine-tuning with a cosine similarity-based contrastive loss. This part of the construction can be viewed in Figure 1 b).

We trained each foundation model with the Adam optimizer for four epochs on the dataset VK2M with a learning rate of $5e^{-7}$, weight decay $1e^{-3}$ and a batch size of 32.

E. Retrieval

For each fine-tuned model, we created a retrieval dataset containing image and caption embeddings generated from all caption-image pairs in VK2M, along with their corresponding geographic coordinates. To predict the location of a previously unseen image-caption pair, the model first generates

⁵<https://huggingface.co/openai/clip-vit-large-patch14-336>

⁶<https://huggingface.co/geolocal/StreetCLIP>

⁷<https://huggingface.co/ai-forever/ruclip-vit-large-patch14-336>

⁸<https://www.google.com/streetview/>

⁴<https://huggingface.co/google/madlad400-3b-mt>

Table II
PERFORMANCE OF DIFFERENT GEOCIS VARIANTS ON VK60kTEST

Foundation Model	Finetuned	Caption	Median error (km)	Distance (% @ km)				
				Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
CLIP	no	no	980.47	8.12	11.16	16.93	40.21	84.22
		yes	939.57	10.20	14.77	20.73	41.91	84.30
	yes	no	809.09	10.38	14.58	22.34	47.39	86.20
		yes	667.16	12.61	20.60	29.47	53.25	88.10
StreetCLIP	no	no	960.54	8.02	11.04	17.30	40.94	84.38
		yes	927.02	10.18	15.00	21.12	42.38	84.25
	yes	no	808.03	10.49	14.72	22.42	47.45	86.34
		yes	670.86	12.75	20.95	29.83	53.23	87.82
RuCLIP	no	no	1036.69	6.28	8.53	14.18	37.45	83.12
		yes	972.07	8.45	12.43	18.51	40.51	84.81
	yes	no	931.83	7.41	10.65	17.45	41.97	84.68
		yes	682.22	11.61	21.00	30.06	52.66	87.91

Table III
PERFORMANCE OF DIFFERENT GEOCIS VARIANTS ON VK1280TEST,
A DATASET WITH CAPTIONS NOT CONTAINING ANY OVERT GEOGRAPHICAL CLUES.

Foundation Model	Finetuned	Caption	Median error (km)	Distance (% @ km)				
				Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
CLIP	no	no	1105.88	5.31	7.27	11.56	33.67	84.22
		yes	1070.52	5.08	6.88	10.70	35.16	82.27
	yes	no	949.38	6.88	9.45	16.33	41.41	84.38
		yes	878.87	7.11	10.08	17.50	43.83	84.77
StreetCLIP	no	no	1088.69	5.08	6.95	11.56	35.16	84.38
		yes	1098.96	5.16	6.72	10.16	33.44	82.27
	yes	no	921.67	6.64	9.06	15.39	41.80	84.84
		yes	887.99	7.11	10.55	16.80	43.05	85.16
RuCLIP	no	no	1113.44	4.61	5.78	10.62	33.91	84.06
		yes	1149.08	3.75	5.55	9.30	32.11	82.89
	yes	no	1014.37	4.77	6.72	11.72	36.95	85.62
		yes	963.51	3.98	6.56	13.44	39.22	85.62

the query’s joint caption-image embedding. It then searches the retrieval dataset for the embedding most similar to the query embedding, using cosine similarity. The geographic coordinates associated with the most similar embedding are returned as the predicted location. This retrieval process is illustrated in Figure 1 c).

F. Evaluation

The primary evaluation metric used in this paper is the *Percentage at Kilometer (% @ KM)* metric. This measures the percentage of guesses that are within some distance thresholds of the actual coordinates. The conventional thresholds of 1, 25, 200, 750, and 2,500 kilometers represent guesses on street-level, city-level, region-level, country-level, and continent-level. We also measure the median distance error in km as a secondary metric.

IV. RESULTS AND DISCUSSION

We evaluated all three fine-tuned GeoCIS variants on the two different datasets VK60kTest and VK1280Test (see Section III-C). For each GeoCIS variant, we further constructed an ablated variant, where we removed any usage of captions by the foundation model. In the ablated variants, we replaced any use of the joint caption-image embedding vector with a normal image embedding vector, both during training and at inference time. For the two English GeoCIS variants, we used the translated versions of our datasets (see Section III-C) when training. At inference the Russian captions were automatically translated by the MADLAD-400 translation model (see Section III-C) before being sent to the English variants. Lastly, we also evaluated all the foundation models without applying any fine-tuning. Similarly to how we did in the fine-tuned case, we evaluated both with and without using captions on both VK60kTest and VK1280Test.

The results for VK60kTest are summarized in Table II. The

results for VK1280Test are summarized in Table III.

A. VK60KTest Results

All GeoCIS variants outperformed their corresponding no-caption ablation variant on VK60KTest, achieving the best results on all six distance metrics. Further the fine-tuned variants outperformed their corresponding non-fine-tuned variants. These results make it clear that social media captions improve the geolocalization performance of a retrieval-based multimodal model, and indicate that it is best to include the captions both during training and inference for optimal performance.

However, the performance differences between foundation models are less definitive. Non-fine-tuned RuCLIP is the worst-performing model, and RuCLIP performs the worst in almost all cases. The easiest way to explain RuCLIP's generally poor performance is by noting that RuCLIP has been trained on less data than CLIP and StreetCLIP. 240M image-text pairs was used to train RuCLIP [16], while 400M was used for CLIP [15] and a total of 400M + 1.1M image-text pairs was used for StreetCLIP [14]. Another possible explanation to RuCLIP's poor performance could be that the GeoCIS variant using RuCLIP is exposed to the original captions, whereas the variants using CLIP and StreetCLIP are exposed to captions that are translated to English. Given that the CIS region is linguistically diverse, the RuCLIP variant thus may encounter captions in multiple languages, despite being trained exclusively on Russian text. The exception to RuCLIP's poor performance is the case with the fine-tuned RuCLIP using captions on the distance metrics (% @ 25KM) and (% @ 200KM), where it instead performs best of all models. This result is more complicated to explain, since RuCLIP performs so poorly otherwise, but some of the non-translated captions may contain local language use that gives hints as to which city and region the person is from. The theory is that this information gets lost in translation. As we will see, however, the results on the VK1280Test dataset cast doubt on this explanation.

The difference between the performance of StreetCLIP vs CLIP is minimal, and this can again be explained by the fact that they have been trained on almost the same amount of data. One would have hoped that StreetCLIP would outperform CLIP, given its specific training for geolocalization, but this does not appear to be the case in our setting. It is possible that the Google Street View images, which StreetCLIP has been trained on, are too dissimilar to images posted on social media.

B. VK1280Test Results

On the VK1280Test dataset we observe that, with few exceptions, all model variants performed worse. This suggests that the CLIP-based foundation models are effectively leveraging the geographical clues provided in the captions in VK60KTest, that are absent in VK1280Test.

The fine-tuned variants of GeoCIS, using either CLIP or StreetCLIP as the foundation model, again outperform their

no-caption ablation variant on all distance metrics on the VK1280Test dataset. However, the improvements are less pronounced than those observed on the VK60KTest dataset. In contrast, with RuCLIP as the foundation model, improvements were observed in only three out of six metrics. It is possible that this difference in results could again be explained by the amount of total training data used by the different foundation models (see Section IV-A). If this holds, this tell us that social media captions improve the geolocalization performance of a retrieval-based multimodal model even if the caption itself does not contain clear geographical clues, provided that the foundation model has been properly trained.

For the non-fine-tuned variants we observe that using captions at inference time often leads to decreased performance. This was not the case for the VK60KTest dataset. This outcome is reasonable however as without fine-tuning, and with no geographic clues to help them, the models have limited guidance on how to interpret social media captions.

V. CONCLUSION

In this work, we wanted to answer the following research question:

How can the geolocalization capability of a retrieval-based multimodal model be improved by including social media captions?

After training our GeoCIS variants on 2.4 million caption-image pairs from the Russian social network VK, both with corresponding captions and without, using our novel joint input vector, we can now say that the geolocalization capability effectively improves when captions are used, both by using them during fine-tuning and by using them at inference. The best performance is achieved when using them at both fine-tuning and inference time. We also see that the geolocalization clues found by our GeoCIS variants in the captions are not only trivial, such as city names and tourist attractions. Instead, our models can utilize information that does not contain obvious geographical clues in order to enhance their predictions.

A. Future Work

To better understand the impact of localization on retrieval-based multimodal models it would be valuable to replicate the experiments presented in this article across other languages and regions. ChineseCLIP [22] is a foundation model tailored to Chinese that would make a promising candidate to study in this regard, provided that a sufficiently large and appropriate Chinese dataset can be identified. ChineseCLIP uses pretrained weights from larger models, including CLIP, and train these further with a novel 200M Chinese dataset. This makes it probable that ChineseCLIP's performance would not be limited by the quantity and quality of the pretraining data, unlike what appears to be the case with RuCLIP.

There has been a lot of progress in the field of contrastive image-text embedding models, the class to which all foundation models discussed in this article belong. Among the most promising models are SigLIP [23] and its successor, SigLIP 2

[24], which introduces a novel loss based on the sigmoid function. An interesting avenue for future research would be to investigate how the high-performance of SigLIP 2 translates into geolocalization capabilities when trained on social media caption-image pairs.

Lastly, it would be interesting to see how the geolocalization performance of retrieval-based multimodal models scales with the size of the retrieval database. In our case, the retrieval database consists of 2.4M caption-image pairs. One potential research direction would be to experiment with progressively smaller databases to evaluate how retrieval performance is affected. If one observes diminishing returns in performance, smaller and more manageable models may be preferable in certain contexts.

VI. ETHICS

In the introduction, we talked about how the geolocalization of social media images has the potential to benefit society by being used for open-source intelligence or in the verification of online information. Alongside these benefits however, come several potential risks, one of them being privacy. To respect individual's privacy rights, we have chosen not to share the data or model weights used in creating our model. By describing the process of our data collection, a third party can verify the results herein, which leaves our scientific integrity intact.

REFERENCES

- [1] D. Wilson, X. Zhang, W. Sultani, and S. Wshah, "Image and Object Geolocalization," *Int. J. Comput. Vision*, vol. 132, no. 4, pp. 1350–1392, Nov. 2023, doi: 10.1007/s11263-023-01942-3.
- [2] J. Hays and A. A. Efros, "IM2GPS: estimating geographic information from a single image," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2008, pp. 1–8. doi: 10.1109/CVPR.2008.4587784.
- [3] T. Weyand, I. Kostrikov, and J. Philbin, "PlaNet — photo geolocation with convolutional neural networks," in *Computer vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 37–55.
- [4] S. Pramanick, E. M. Nowara, J. Gleason, C. D. Castillo, and R. Chellappa, "Where in the world is this image? Transformer-based geolocalization in the wild," in *Computer vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 196–215.
- [5] P. Jia et al., "G3: An effective and adaptive framework for worldwide geolocalization using large multi-modality models," in *Advances in neural information processing systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., Curran Associates, Inc., 2024, pp. 53198–53221.
- [6] P. Jia, S. Park, S. Gao, X. Zhao, and Y. Li, "GeoRanker: Distance-Aware Ranking for Worldwide Image Geolocalization," May 19, 2025, doi: 10.48550/arXiv.2505.13731.
- [7] V. Vivanco Cepeda, G. K. Nayak, and M. Shah, "GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geolocalization," in *Advances in neural information processing systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 8690–8701. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/1b57aaddf85ab01a2445a79c9edc1f4b-Paper-Conference.pdf
- [8] J. Theiner, E. Müller-Budack, and R. Ewerth, "Interpretable Semantic Photo Geolocation," presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 750–760. Accessed: Jun. 12, 2025. [Online]. Available: https://openaccess.thecvf.com/content/WACV2022/html/Theiner_Interpretable_Semantic_Photo_Geolocation_WACV_2022_paper.html
- [9] Z. Zhou et al., "Img2Loc: Revisiting Image Geolocalization using Multi-modality Foundation Models and Image-based Retrieval-Augmented Generation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR '24. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 2749–2754. doi: 10.1145/3626772.3657673.
- [10] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, Art. no. 1, Jun. 2018, doi: 10.1609/icwsm.v12i1.14983.
- [11] Z. Zou, H. Gan, Q. Huang, T. Cai, and K. Cao, "Disaster Image Classification by Fusing Multimodal Social Media Data," *ISPRS International Journal of Geo-Information*, vol. 10, no. 10, Art. no. 10, Oct. 2021, doi: 10.3390/ijgi10100636.
- [12] Y. Wang, "Multimodal Event Classification for Social Media Based on Text-Image-Caption Assisted Alignment," *IEEE Access*, vol. 12, pp. 166730–166741, 2024, doi: 10.1109/ACCESS.2024.3496001.
- [13] L. Haas, M. Skreta, S. Alberti, and C. Finn, "PIGEON: Predicting Image Geolocations," in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, Jun. 2024, pp. 12893–12902. doi: 10.1109/CVPR52733.2024.01225.
- [14] L. Haas, S. Alberti, and M. Skreta, "Learning Generalized Zero-Shot Learners for Open-Domain Image Geolocalization," Feb. 01, 2023, arXiv: arXiv:2302.00275. doi: 10.48550/arXiv.2302.00275.
- [15] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 8748–8763. Accessed: Mar. 12, 2025. [Online]. Available: <https://proceedings.mlr.press/v139/radford21a.html>
- [16] A. Shonenkov et al., "RuCLIP – new models and experiments: a technical report," Feb. 22, 2022, arXiv: arXiv:2202.10784. doi: 10.48550/arXiv.2202.10784.
- [17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017, doi: 10.1162/tac1_a_00051.
- [18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 3rd International Conference on Learning Representations (ICLR 2015), 2015, Accessed: Mar. 13, 2025. [Online]. Available: <https://ora.ox.ac.uk/objects/uuid:60713f18-a6d1-4d97-8f45-b60ad8aebbbe>
- [19] "GADM," Accessed: Jun. 13, 2025. [Online]. Available: <https://gadm.org/index.html>
- [20] "GeoNames," Accessed: Jun. 13, 2025. [Online]. Available: <https://www.geonames.org/>
- [21] S. Kudugunta et al., "MADLAD-400: A Multilingual And Document-Level Large Audited Dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 67284–67296, Dec. 2023.
- [22] A. Yang et al., "Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese," May 23, 2023, doi: 10.48550/arXiv.2211.01335.
- [23] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11975–11986.
- [24] M. Tschannen et al., "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," Feb. 20, 2025, doi: 10.48550/arXiv.2502.14786.