International Cartographic Association
Association Cartographique Internationale

MapPool – Bubbling up an extremely large corpus of maps for AI

Raimund Schnürer [a,*]

[a] *Digital Humanities Laboratory, Swiss Federal Institute of Technology in Lausanne, raimund.schnurer@epfl.ch*

* Corresponding author

Keywords: machine learning, dataset, map classification, foundation model

Abstract:

MapPool is a dataset of 75 million potential maps and textual captions. It has been derived from CommonPool, a dataset consisting of 12 billion text-image pairs from the Internet. The images have been encoded by a vision transformer and classified into maps and non-maps by a support vector machine. This approach outperforms previous models and yields a validation accuracy of 98.5%. The MapPool dataset may help to train data-intensive architectures in order to establish vision and language foundation models specialized in maps. The analysis of the dataset and the exploration of the embedding space offers a large potential for future work. It is accessible via https://geoai.icaci.org/mappool/

Introduction

State-of-the-art machine learning architectures like transformers or diffusion models are trained on huge amounts of data. As of the beginning of 2024, the largest publicly available dataset is CommonPool with 12 billion text-image pairs (Gadre et al., 2023), which originate from CommonCrawl[1], an even larger collection of website contents. Datasets including only map images do not exist in these magnitudes. There are digital libraries with historical maps[2]; however, those usually contain only a few contemporary maps. Besides this, retrieving maps via text search engines is cumbersome since some map images may have missing or incorrect labels as well as the term 'map' also has non-cartographic (e.g., mathematical) meanings and differs across languages.

Methods

MapPool has been created by classifying the image embeddings included in CommonPool, which have been generated by two pre-trained vision transformers (ViTs). The model (i.e., L/14) with more parameters and outputting 768-dimensional embeddings has been considered since it has achieved higher classification accuracies (Radford et al., 2021). Previous map classifiers have been established by applying a waterfilling and k-nearest neighbour algorithm (Goel et al., 2011) as well as convolutional neural networks (Li & Xiao, 2023; Schnürer et al., 2021). In this work, different map classifiers (Table 1) from scikit-learn[3] have been trained on the image embeddings of 1,860 maps and 1,860 non-maps, and have been evaluated on 1,240 maps and 1,240 non-maps. The images originate from Pinterest (Schnürer et al., 2021) and were encoded by a ViT-L/14 for this experiment. Only simple classification models have been considered due to their efficiency and as meaningful embeddings have already been created by the vision transformer.

Model	**Accuracy [%]**
Xception / InceptionResNetV2	*96.7*
ViT-L/14 + L2 distance to averaged embeddings	96.7
ViT-L/14 + Logistic Regression	97.9
ViT-L/14 + Multilayer Perceptron (3x256 units)	98.2
ViT-L/14 + Support Vector Machine (polynomial, degree 3)	98.5

Table 1. Validation accuracy of different approaches distinguishing maps and non-maps. Two baseline models by Schnürer et al. (2021) are listed in the first row (in italics).

[1] https://commoncrawl.org/
[2] http://maphistory.info/imagelarge.html
[3] https://scikit-learn.org/ (+ https://intel.github.io/scikit-learn-intelex/)

Results

75 million images have been identified as potential maps by filtering the image embeddings of CommonPool with a support vector machine. The images are located at 1.8 million different Internet domains. 500,000 image embeddings could be classified within 10 seconds. Downloading, classifying the whole dataset, and uploading the results took about 50 hours with 10 CPUs, 120GB RAM, and 500MB/s of network traffic on average.

As the Internet is constantly changing, 48 million of the original images are still downloadable (Figure 1). 6TB of space are required to store them in their original formats and 100GB of space are needed when creating 128x128px thumbnails in the WebP format with 60% quality. Downloading the images with the img2dataset[4] library took 40 hours with 24 CPUs, 30GB RAM, and 40MB/s of network traffic on average.

Figure 1. The first 1,000 images of MapPool

Discussion

A qualitative inspection of the detected maps looks promising; however, it is not known what the actual accuracy is. Examples of some false positives are flags and charts. The false negative rate is hard to estimate due to the high number of non-maps among the CommonPool images. Mixtures between natural images and maps (e.g., a map printed on a bag, a map in a park) have not been further examined. The underlying map definition is explained in Schnürer et al. (2021).

Textual embeddings have not been considered in the separation process so far. The training dataset for the map classifier has a large visual variety, such as pictorial maps and 3D maps as well as sketches and paintings. However, the textual descriptions may be too biased since the training dataset originates only from one source.

Similarly to the CommonPool dataset, only links to the images are stored for MapPool due to copyright protection. When using the dataset, copyright and ethical regulations for AI, which are continuously adjusted, need to be respected.

Outlook

A detailed analysis of the content and metadata of maps in MapPool, potentially resulting in a search engine, is the subject of future work. Additionally, the visual and textual embedding space may be explored to refine the map classifier and to detect duplicates among the images. It can be examined whether training with map-only images leads to better results for cartographic tasks, for instance generating maps based on textual prompts, than with a mixture of maps and other images.

Acknowledgements

The author would like to thank Emanuela Boroș for the support in using the high-computing cluster at EPFL.

References

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., … Schmidt, L. (2023). DataComp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, *36*, 27092–27112.

Goel, A., Michelson, M., & Knoblock, C. A. (2011). Harvesting maps on the web. *International Journal on Document Analysis and Recognition (IJDAR)*, *14*(4), 349–372. https://doi.org/10.1007/s10032-010-0136-2

Li, J., & Xiao, N. (2023). Computational Cartographic Recognition: Identifying Maps, Geographic Regions, and Projections from Images Using Machine Learning. *Annals of the American Association of Geographers*, *113*(5), 1243–1267. https://doi.org/10.1080/24694452.2023.2166010

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., … Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Schnürer, R., Sieber, R., Schmid-Lanter, J., Öztireli, A. C., & Hurni, L. (2021). Detection of Pictorial Map Objects with Convolutional Neural Networks. *The Cartographic Journal*, *58*(1), 50–68. https://doi.org/10.1080/00087041.2020.1738112

[4] https://github.com/rom1504/img2dataset