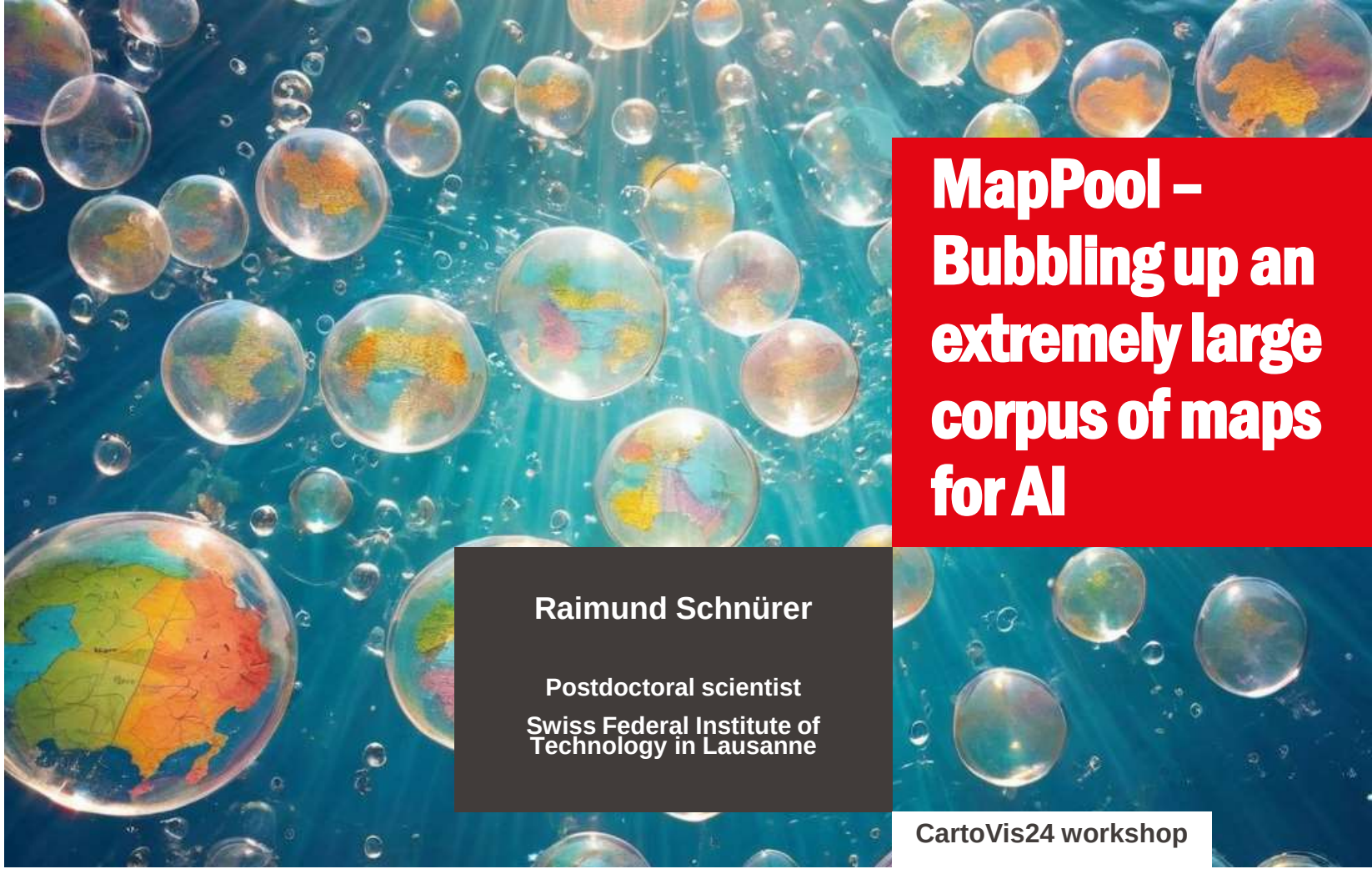


MapPool – Bubbling up an extremely large corpus of maps for AI

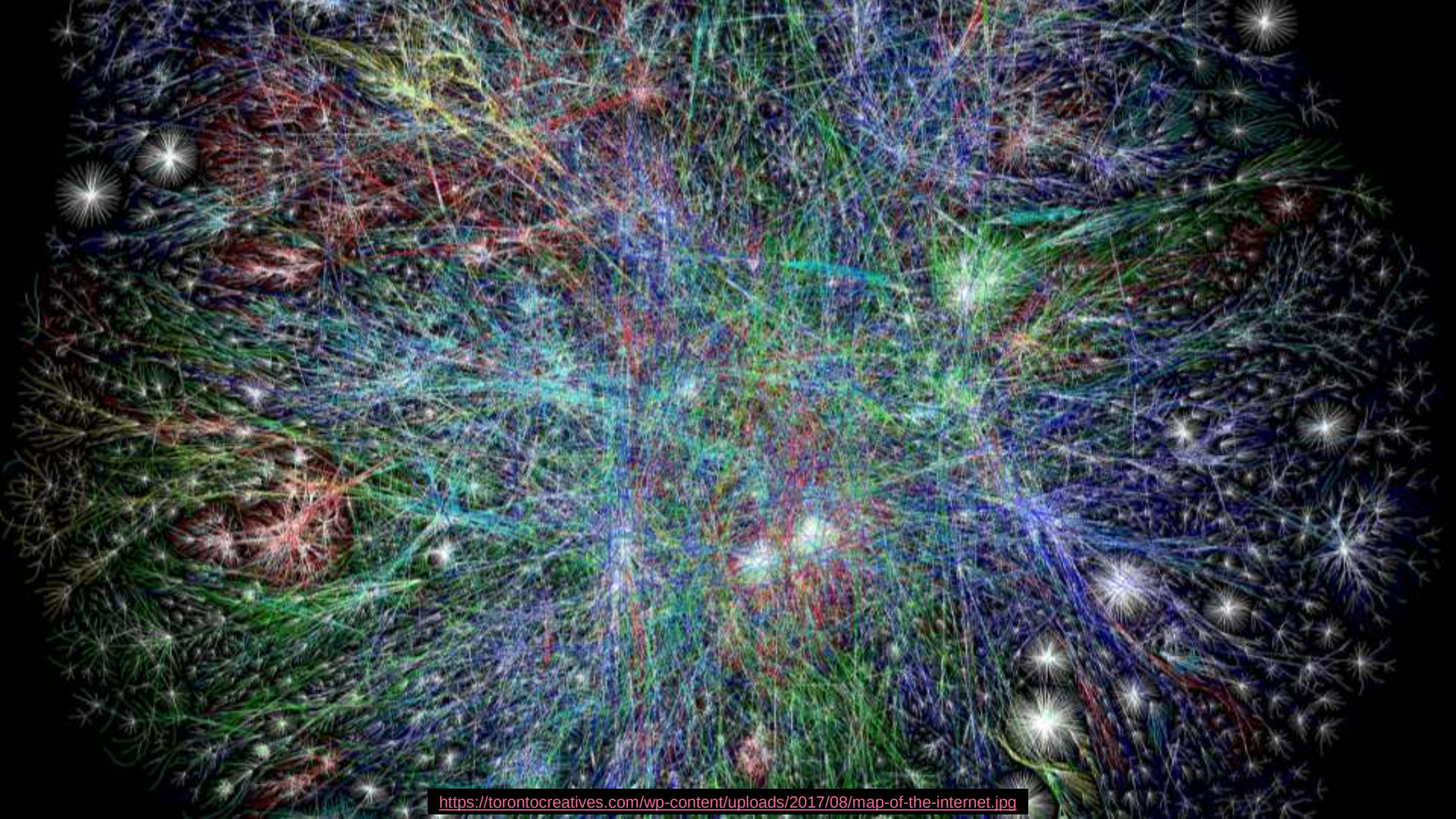
Raimund Schnürer

Postdoctoral scientist

Swiss Federal Institute of Technology in Lausanne

CartoVis24 workshop

Warsaw, 07.09.2024



https://torontocreatives.com/wp-content/uploads/2017/08/map-of-the-internet.jpg

COMMON CRAWL

The Data | Resources | Community | About | Search | Contact Us

Common Crawl maintains a free, open repository of web crawl data that can be used by anyone.

https://commoncrawl.org/

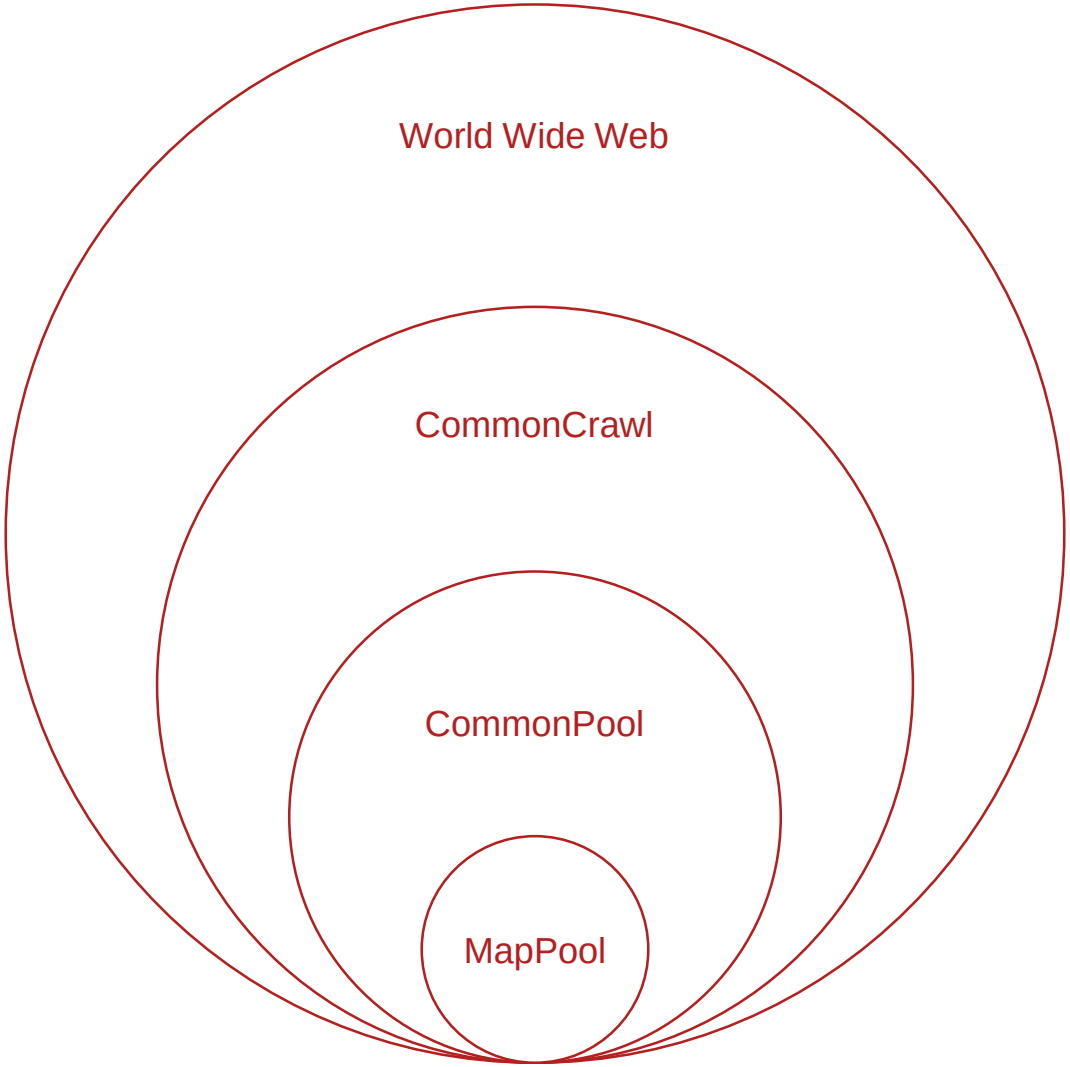
DataComp - CLIP

Welcome to DataComp, the machine learning benchmark where the models are fixed and the challenge is to find the best possible data!

Prior competitions in machine learning have focused on finding the best model, with a fixed set of training and test data. However, many recent advances (CLIP, DALL-E, Stable Diffusion, or Flamingo) are due in part to large multimodal datasets. DataComp centers the role that data plays by fixing the training code, and encouraging researchers to innovate by proposing new training sets.

We provide an experimental testbed centered around a new candidate pool of 12.8 billion image-text pairs from Common Crawl. Participants in our benchmark design new filtering techniques or curate new data sources and then evaluate them by running our standardized CLIP training code followed by an evaluation on 38 downstream datasets. Our benchmark consists of multiple scales, with four candidate pool sizes and associated compute budgets ranging from 12.8 million to 12.8 billion. This multi-scale design facilitates the study of scaling trends and makes the benchmark accessible to researchers with varying resources. More information can be found on this blog post.

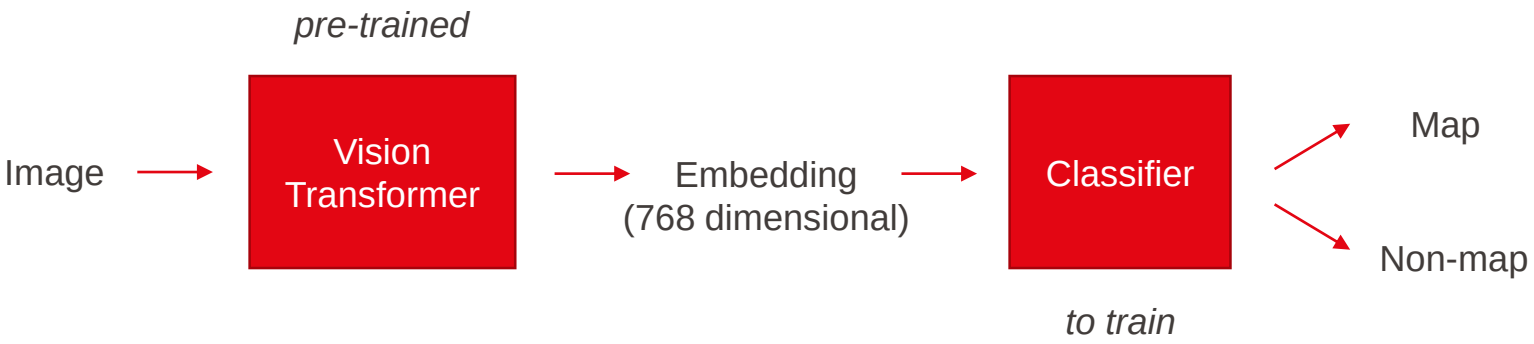
https://www.datacomp.ai/dcclip/


World Wide Web
CommonCrawl
CommonPool
MapPool

Related work

	Goel et al. (2011) Harvesting maps on the web	**Schnürer et al. (2021)** Detection of Pictorial Map Objects with Convolutional Neural Networks	**Li & Xiao (2023)** Computational Cartographic Recognition: Identifying Maps, Geographic Regions, and Projections from Images Using Machine Learning
Data (Training:Test)	4,000 maps 4,000 non-maps (50:50)	3,100 maps 3,100 non-maps (60:40)	500 maps 500 non-maps (80:20)
Methods	SVM, Waterfilling & kNN	CNNs	SVM, MLP, CNNs
Accuracy	77%	96.7%	100%

Method


pre-trained
Image
Vision Transformer
Embedding (768 dimensional)
Classifier
Map
Non-map
to train

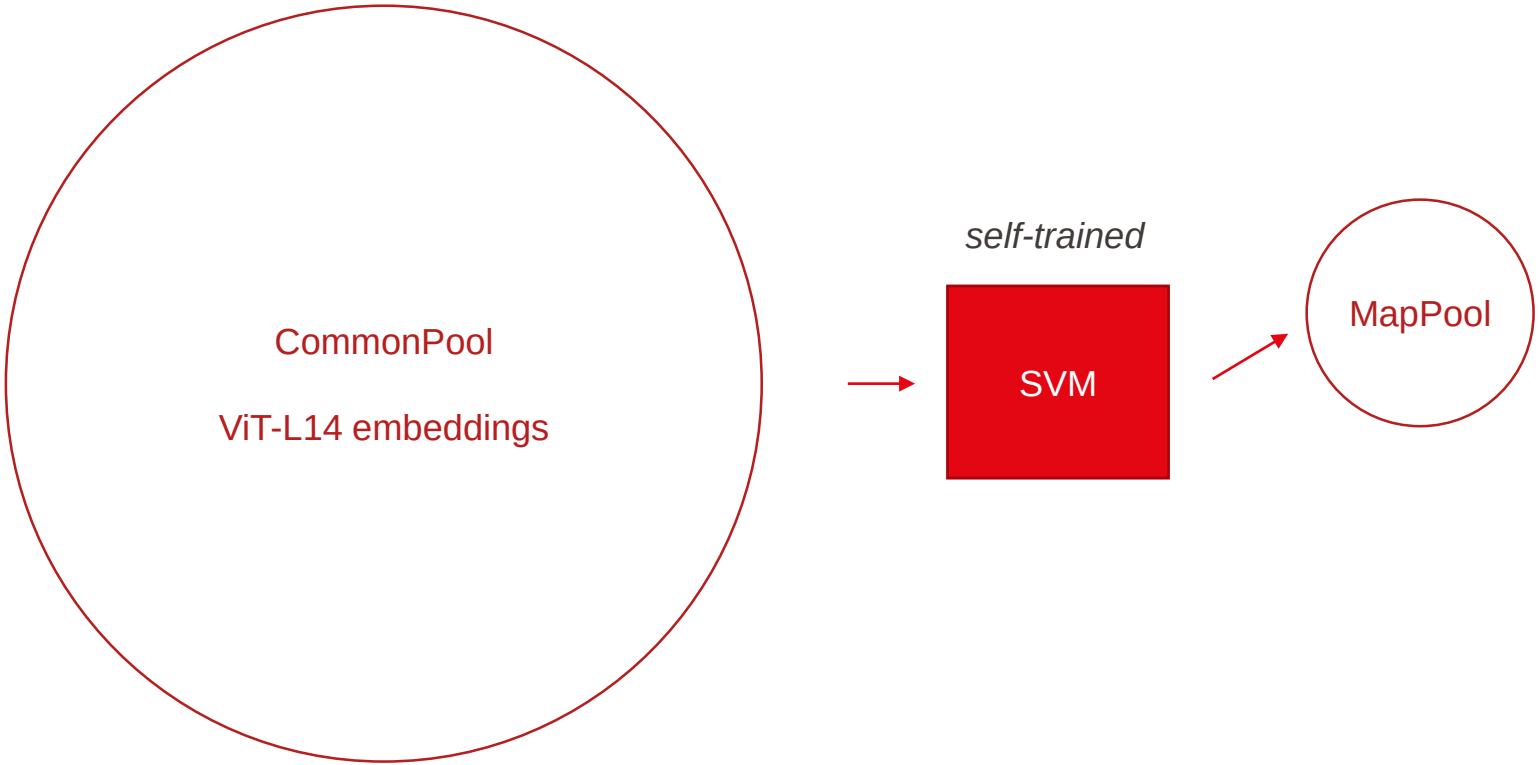
Results

Raimund Schnürer

Vision Transformer + Classifier	Accuracy
ViT-L/14 + L2 distance to averaged embeddings	96.7%
ViT-L/14 + Logistic Regression	97.9%
ViT-L/14 + Multilayer Perceptron	98.2%
ViT-L/14 + Support Vector Machine	**98.5%**

Baseline CNNs [Schnürer et al. 2021]	96.7%

Inference



Numbers

- Training of the classifier: <1h
  - ➢ 1 SVM model (15MB)

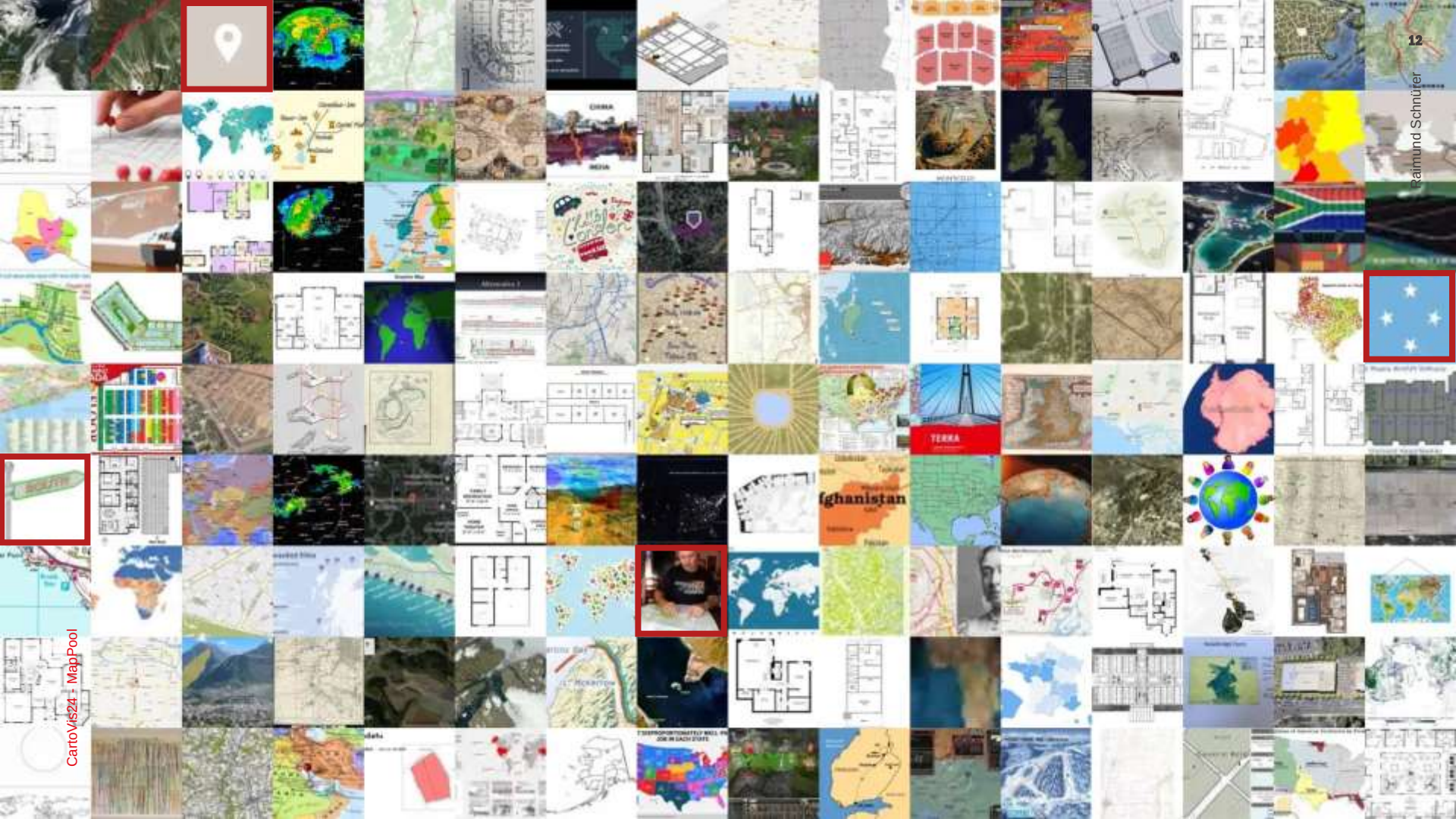
- Download, classification, upload of 13B embeddings: 50h
  - ➢ 75M map image embeddings (242GB)

- Download and downscaling of the map images: 40h
  - ➢ 48M map image thumbnails (100GB)

- Indexing image and text embeddings: 7h
  - ➢ 2 Facebook AI similarity search indices (8.4GB)


You see 14,440 map thumbnails
(i.e., 0.03% of the dataset)

Raimund Schnürer



MapPool model

The model predicts whether an image is a map or not. It takes about 30 seconds since it runs on a CPU (it is much faster on a GPU). Although the validation accuracy of the model is 98.5%, some outputs may not be correct. In this case, feel free to contact me.

Input image

output

Map

Clear Submit

More information: MapPool - Bubbling up an extremely large corpus of maps for AI

Keywords: map identification, map recognition, map classification

MapPool model

The model predicts whether an image is a map or not. It takes about 30 seconds since it runs on a CPU (it is much faster on a GPU). Although the validation accuracy of the model is 98.5%, some outputs may not be correct. In this case, feel free to contact me.

Input image

output

No map

Clear

Submit

More information: MapPool - Bubbling up an extremely large corpus of maps for AI

Keywords: map identification, map recognition, map classification

Backend url:
/backend
Index:
mappool

switzerland

[Clip retrieval] works by converting the text query to a CLIP embedding , then using that embedding to query a knn index of clip image embedddings

Display captions ☑
Display full captions ☐
Display similarities ☐
Safe mode ☑
Remove violence ☑
Hide duplicate urls ☑
Hide (near) duplicate images ☑
Enable aesthetic scoring ☐
Aesthetic score
Aesthetic weight
0.5
Search over image
Search with multilingual clip ☐

This UI may contain results with nudity and is best used by adults. The images are under their own copyright.

Are you seeing near duplicates ? KNN search are good at spotting those, especially so in large datasets.


Dieser artikel behandelt den spielgeld zum ausdruc...


スイスの地理 - Wikipedia：九州山脈 地図：すべての講義


3d view of Lessach Oberdorf


Kanton Basel-Stadt Landkarte


Kanton Luzern Schweiz


EPS Illustration.


Der Kanton Tessin, die Sonnenstube der Schweiz


Map of Swizerland, Zurich highlighted


Karte von Bever


Karte von Aire-la-Ville


carte-suisse_edited.jpg


switzerland map of black contour curves of vector ...


PLZ 2017 Schweiz


Fribourg


Spa Strandbad Klosters in Davos Klosters: Position...


Map jura Vector Clipart Royalty Free. 39 Map jura ...


Karte von Begnins


img


PLZ 2 Schweiz


Karte von Thayngen


Map of switzerland continous line vector


Map of Switzerland


PLZ 1010 Schweiz


Cantons Of


Kochi Prefecture


Kotlíková dotace Plzeňský kraj 2020


Schweiz clipart #12,

Backend url:
/backend
Index:
mappool

Clip retrieval works by converting the text query to a CLIP embedding , then using that embedding to query a knn index of clip image embedddings

Display captions
Display full captions
Display similarities
Safe mode
Remove violence
Hide duplicate urls
Hide (near) duplicate images
Enable aesthetic scoring
Aesthetic score
Aesthetic weight
0.5
Search over image
Search with multilingual clip

This UI may contain results with nudity and is best used by adults. The images are under their own copyright.

Are you seeing near duplicates ? KNN search are good at spotting those, especially so in large datasets.


mapa3.png


PAKIET: Mapy


Wydawnictwo Dwie Siostry


карта Польши


Heidrich Andrzej Płynie Wisła płynie


Trasa Misia Filusia


Tent London (70)


mapy2


Tent London (70)


Let's face it, paper maps aren't as interesting no...


Zeitmagazin partnersuche


Для увеличения необходимо кликнуть на


Scienna Mapa Polski Wydawnictwo Dwie Siostry Map A...


Fototapeta Lithuania Vintage Map


Polonia, mappa illustrata


Picture of: Political And Administrative


Mapster Zestawienia Map


pasaulio_atlasas_knygos_ki


Kultur Wissensfrage: Was bekommt man, wenn man in ...


_big
Miasta biorące udział w plebiscycie 'Superniasta'

Resources

- Image URLs, text captions, embeddings are publicly available: https://huggingface.co/datasets/sraimund/MapPool
- The map classifier is publicly available: https://huggingface.co/spaces/sraimund/MapPool
- Training data, thumbnails, indices, website code are available on request: raimund.schnurer@epfl.ch

Outlook

- Explore the dataset and embedding space
- Examine the usefulness for establishing map foundation models

- Improve the training dataset
- Include textual embeddings in the classifier

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

Thank you for your attention!

Short paper:
https://infoscience.epfl.ch/handle/20.500.14299/240495

Raimund Schnürer

https://schnuerer.dev

https://geoai.icaci.org