



Deploying a Self-Supervised Learning Based Model to Search Events Across Space and Time

Iksha Gurung*, Muthukumaran Ramasubramanian, Rodrigo Almeida, Soumya Ranjan, Leo Thomas, Andrew Bollinger, Manil Maskey, Rahul Ramachandran, Sowmya Subramanian, Kathryn Berger, Lilliane Thomas, Heidi Mok, Tammo Feldmann, Vincent Sarago, Nick Ingalls, Sajjad Anwar

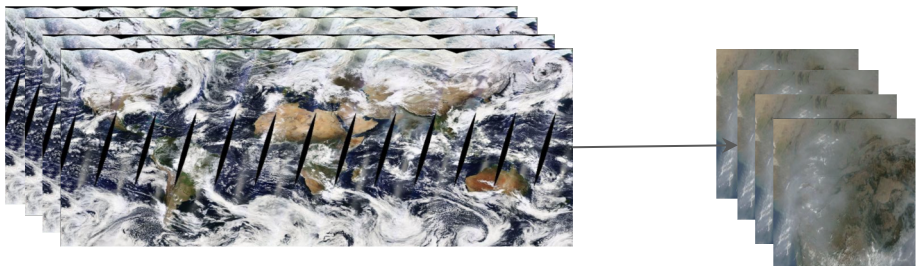
Motivation

- Scientific Study of natural events, phenomena, or disasters require examples which span across time and space.
- Machine Learning adaptation is on the rise, but there's a lack of labeled training datasets that could be used to train or validate the models.
- Best case scenario:
 - There's an event database that tracks events available through time and space.
 - Provides all data associated with the events.
- Real life scenario:
 - Some events are better tracked than others.
 - Scientists need to spend significant time identifying and gathering examples of events from different sources.

Example

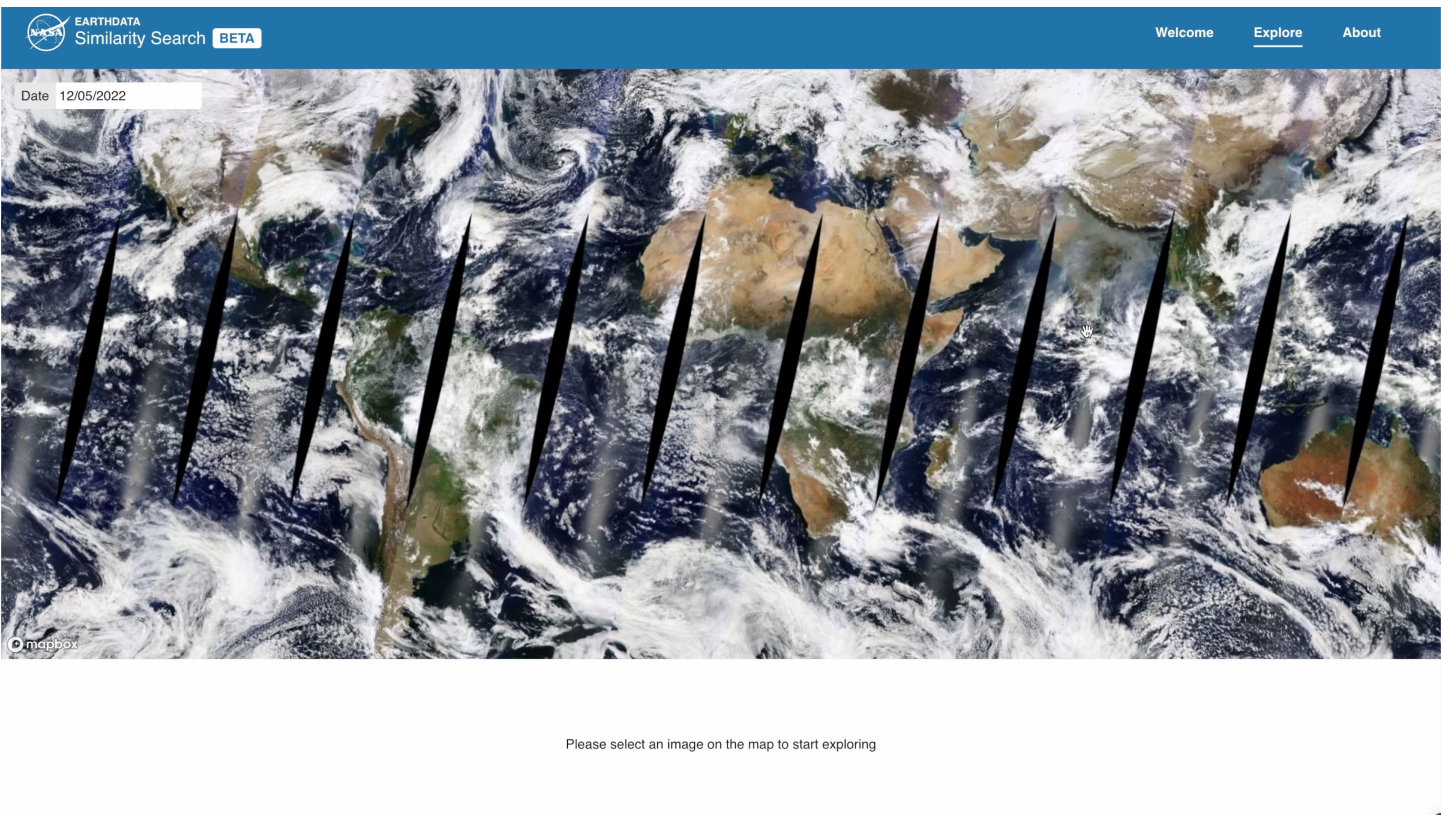


Identify smoke event, catalog, download



*Identify other days when smoke occurred,
catalog, download, **repeat!***

Demo



Background

- SpaceML introduced a self-supervised learning (SSL) based model to the Earth Science domain in 2020/2021.
- The model was used to create embeddings and indices for optical satellite images (primarily Moderate Resolution Imaging Spectroradiometer, MODIS, True Color).
- It only indexed 1 year of data.
- Used Web Map Tile Service to download tiles of MODIS True Color.

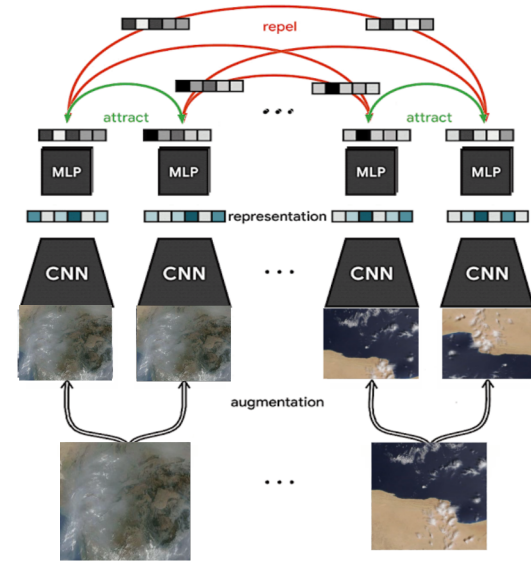
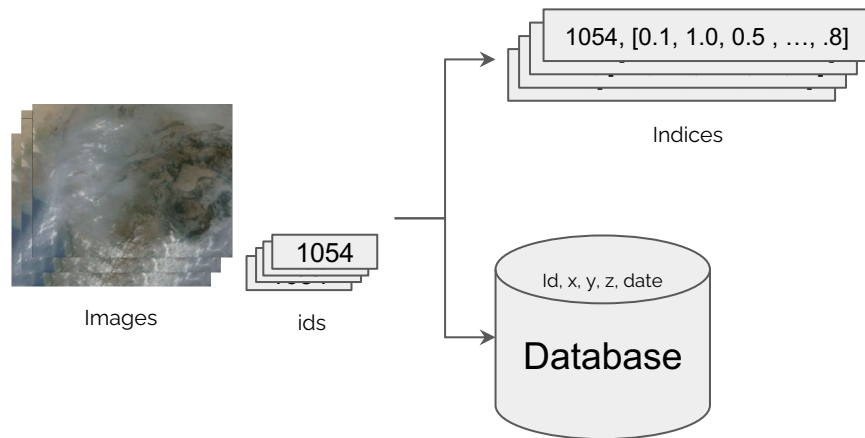


Fig: Self-supervised learning with satellite images

1 year of data is not enough for good Scientific Study or training datasets.

Some Numbers to Consider

- Total number of years: **22**
- Web Map Tile service (WMTS) *zoom level*: **9**
- Dimension of a *single tile/image*: **256 X 256 X 3**
- Total tiles for *1 day*: **35 K**
- Total tiles for *1 year*: **12 M+**
- Total tiles for *22 year*: **278 M**
- Number of *images per index*: **~1 M**
- Index Size for *1 year*: **1.6 GB**
- Index Size for *22 Years*: **35.2 GB**



Note: Numbers here are for TERRA MODIS True Color

Fig: Images, Ids, and Indices

Architecture Discussion I

- We are using Facebook AI Similarity Search (FAISS) for Artificial Nearest Neighbour calculation
 - Larger indices distance calculation is handled on disk
 - Reduces memory requirement
- Embeddings are generated in ECS clusters using a service "Coiled" for cluster orchestration and management.
- All Embeddings and indices are stored in a S3 bucket.
- All metadata of images are stored in a Amazon Web Services (AWS) Relational Database (RDS), postgres.

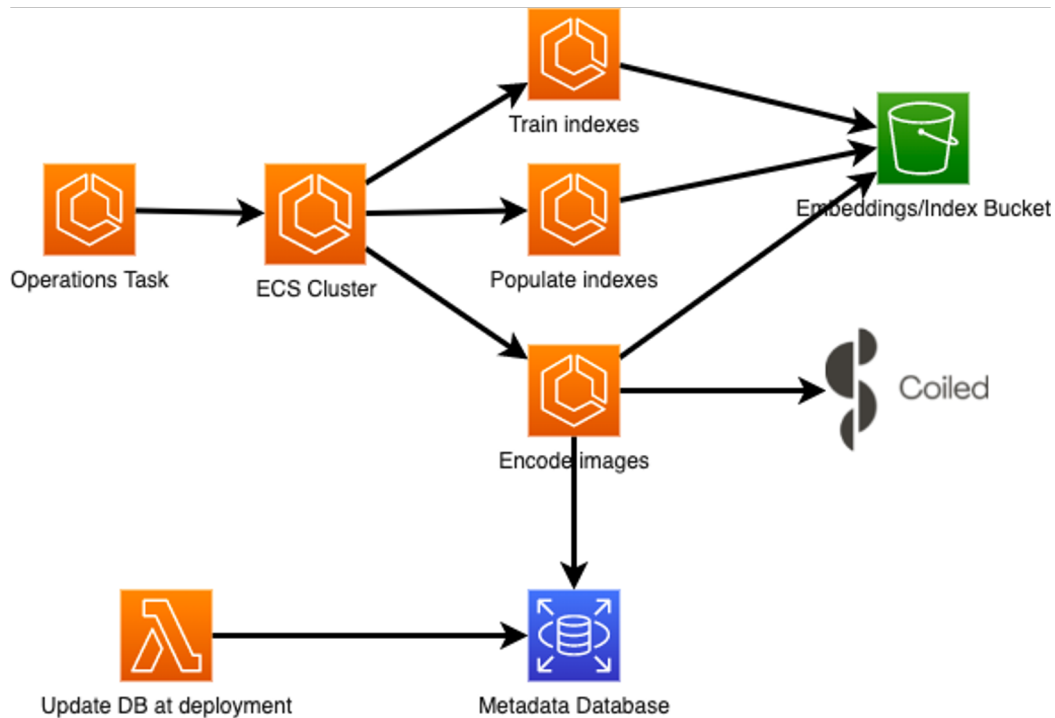


Fig: Embeddings creation process

Architecture Discussion II

- Cloud native deployment with FAISS for neighborhood search.
- Elastic Clusters are used to serve the indices for scalability.
- Auto-scaling is enabled to handle higher request load.
- Similar results are cached using Redis.



Fig: Architecture of serving indices

Lessons Learned

- Similarity Search for satellite images works!
- Migration of Proof of Concept to Production needs a large amount of considerations before it can be made available for public usage.
- The same approach can be applied to other datasets and bands.

Resources

- Similarity Search Portal: <https://earthdata.nasa.gov/dashboard/labs/similarity-search>
- FAISS: <https://github.com/facebookresearch/faiss>
- SpaceML's Similarity Search: <https://github.com/spaceml-org/Image-Similarity-Search>

Thank you!

Iksha Gurung, ig0004@uah.edu

