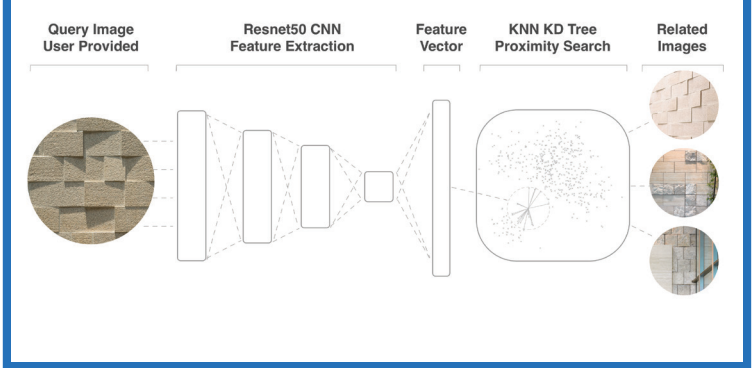
The Case: KPF

KPF

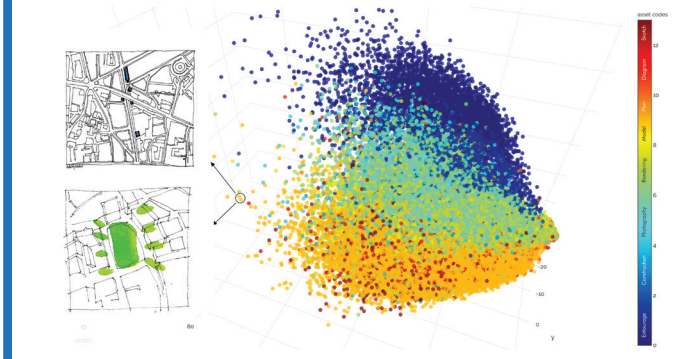
Founded in 1976 and headquartered in New York City, with additional offices in London, Shanghai, Hong Kong, Seoul, Abu Dhabi, San Francisco, Singapore, and Berlin, Kohn Pedersen Fox Associates (KPF) is a leading global architecture firm providing architecture, interior, programming, and master plan services for both public and private sector clients.

Systems Configurations

KPF relies on a machine learning (ML) pipeline to perform visual searches on their vast image data store in the event that a tag search is not working well, i.e., assets not properly categorized or incorrectly tagged. The KPF ML pipeline ran on a CPU-only Linux server based on TensorFlow ResNet50 for image feature extraction and sci-kit learn KNN for proximity search. The challenge is that the process, although functional, does not perform as quickly and efficiently as necessary. In order to accelerate the process, enable further refinements, and simply improve overall pipeline performance, KPF required a hardware solution purpose-built for their specific data science workflow.



Visual Search - Solution Architecture Diagram



Visual asset clusters in the PCA-reduced 3d space

Testing a Possible Solution

KPF elected to test a BOXX Data Science Workstation (DSWS), a Linux-based system purpose-built for deep learning and equipped with a Rapids ML library and NVIDIA Quadro RTX 8000 GPU. The complete product specifications are as follows:

BOXX Data Science Workstation
GPU: NVIDIA Quadro RTX 8000
Processor(s): Dual Intel Xeon SP Silver 4210 2.2Ghz 13.75MB cache, 9.6 UPI (Ten-Core)
System Memory: 128GB DDR4-2933 MHz ECC REG (8 - 16GB DIMMS)
Optical Drive: 20X Dual Layer DVD±RW Writer
M.2 Storage: 1.0TB SSD M.2 PCIe Drive
Operating System: Ubuntu Desktop 18.04
Software Applications: NVIDIA CUDAX-AI Data Science Workstation Software Stack

One of the key advantages of the RTX8000 GPU is its large memory capacity, which obviously accommodates larger batches of data, but also allows it to process the data at a much faster rate. Because of this, KPF data scientists could run more experiments and then compare and evaluate the results of each one, selecting those that yielded the best outcomes. Most importantly, the BOXX test system arrived with a complete, pre-installed, GPU-accelerated data science software stack offering significant advantages:

- Simple plug-and-play with no set-up time necessary.
- Eliminated the need for data scientists to spend time performing IT tasks, i.e., building data science software stack by hand.
- Featured the popular data science software packages Docker and Python, Pandas, Sci-kit Learn, numpy, numba.
- Included GPU-accelerated libraries TensorFlow, PyTorch, NVIDIA Rapids, and XGBoost, which provide significant data science workload acceleration.

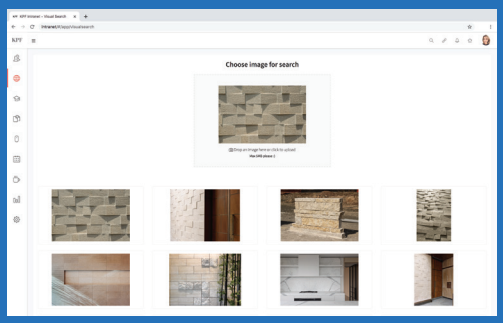
Benchmarking

Loading dataset | In order to perform a benchmark study, KPF refactored the codebase of the ML pipeline to take advantage of the Rapids ML cuDF, a GPU-enabled dataframe. That change contributed to a solid improvement in the data read time. Loading the 2GB dataset (which comprises over 100k visual assets translated to ~250M datapoints) accelerated from 64 seconds (using scikit-learn Pandas) to 3 seconds using cuDF.

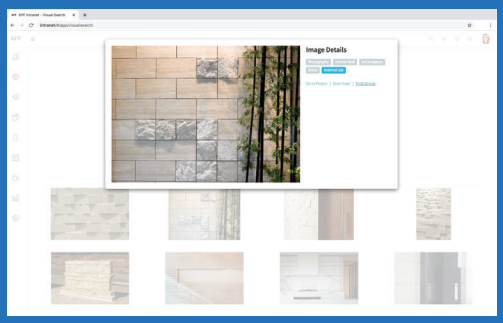
Training | In order to perform the proximity searches within the current ML pipeline, the dataset needs to be fitted into KNN tree using scikit-learn library. This process could take from 35-40 seconds up to a minute. To take advantage of faster GPU computational speed, we looked into replacing the model build with scikit-learn KNN into Rapids cuML, NearestNeigbors model. Thanks to that change, this step in the pipeline was reduced to approximately 10-11 seconds.

Searching (inference) | The last part of the ML pipeline, performing proximity searches, also improved with cuML model. The process now takes ~300 milliseconds. We noticed that the new cuML model performs best if the input data is converted to Numpy array instead of cuDF.

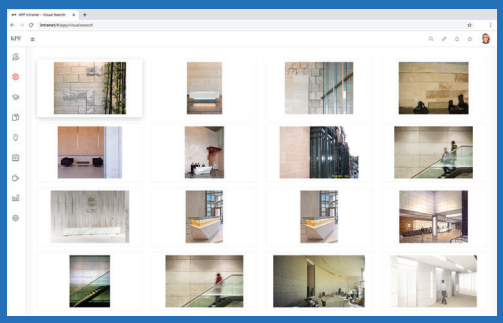
Visual Search - User Interface



Step 1



Step 2



Step 3

Next Steps

The dataset we use in this pipeline (KPF's visual asset library) is built through feature extraction using the Resnet50 based model (2.6sec). We believe that we can accelerate this process by taking advantage of the GPU-enabled TensorFlow built into the BOXX DSWS.

We would also like to test PCA or TSNA for dimensionality reduction in order to provide interactive visualization of all the assets we have in our catalog. We use Plotly's Dasher for 3D visualization and scikit-learn PCA. We are incorporating the techniques that will render the results faster and allow smoother interaction.