Cloud-Based Deep Learning on AWS Open Data Registry: Automatic Building and Road Extraction from Satellite and LiDAR

Yunzhi Shi
Amazon Web Services
Austin, TX USA
shiyunzh@amazon.com

Xin Chen
Amazon Web Services
Chicago, IL USA
xcaa@amazon.com

Tianyu Zhang
Amazon Web Services
Austin, TX USA
ttizha@amazon.com

ABSTRACT

There is a large amount of public open data hosted in the AWS Open Data Registry. The datasets range from genomics to climate to transportation information. They are well structured and easily accessible. However, there are few examples of how to leverage the datasets in machine learning (ML) model development in the cloud.

We create this tutorial by developing Jupyter notebooks to train and test deep learning models using AWS SageMaker to extract building footprints and road networks from satellite imagery and LiDAR data in the registry. The notebooks reproduce winning algorithms from the SpaceNet challenges. In addition to the SpaceNet satellite images, we compare and combine USGS 3D Elevation Program (3DEP) LiDAR data to extract the same and our results outperform some of the top winning teams'.

We will share the notebooks and provide hands-on step-by-step instructions for running ML services on AWS to extract features from large scale geospatial data in the cloud. Through the tutorial, the audience will be able to train the building and road extraction models on AWS, apply the models to other regions where satellite or LiDAR data are available, and experiment with new ideas to improve the performances. The audience will experience the benefits of cloud computing and storage first-hand.

1 INTRODUCTION

Sharing data and computing in the cloud allows data users to focus on data analysis rather than data access. The AWS Open Data Registry is a service that helps users discover and share public open datasets in the cloud. Everyone can analyze the open data and build services on top of it using a broad range of computing and analytics products such as Amazon EC2, Amazon Athena, AWS Lambda, Amazon SageMaker, and Amazon EMR. This helps the community develop new cloud-native techniques, algorithms, formats, and tools.

AWS Open Data Registry includes several large-scale geospatial datasets. For example, SpaceNet was launched in August 2016 as an open innovation project offering a repository of freely available imagery with co-registered map features. The SpaceNet partners also launched a series of competitions to encourage improvement of remote sensing ML algorithms. Another dataset is USGS 3DEP LiDAR data. Its goal is to complete acquisition of nationwide LiDAR to provide the first ever national baseline of consistent high-resolution topographic elevation data.

We author Jupyter notebooks of automatic building and road extraction using deep learning techniques through Amazon SageMaker. We reproduce winning algorithms from SpaceNet challenges, and combine both SpaceNet satellite image and USGS LiDAR data to train and evaluate model performances. We demonstrate the model accuracy improvement by introducing LiDAR data. We will share the notebooks through a via GitHub repository and provide hands-on interactive instructions.

The target audiences are both academics, industry data scientists and ML practitioners interested in learning to use ML services on AWS and getting hands-on experience of running large scale feature extraction from geospatial datasets.

2 DATASETS

2.1 SpaceNet Dataset

SpaceNet data is a large corpus of labeled satellite imagery published by the project partners and hosted on AWS. The project also launched a series of competitions ranging from automatic building extraction, road extraction, to recently published multi-temporal urban development analysis. The dataset covers 11 areas of interest (AOIs), including Rio de Janeiro, Las Vegas, and Paris, etc. For Las Vegas, the AOI used in our tutorial, the images in this AOI cover 216km^2 area with 151367 building polygon labels and 3685km road labels.

2.2 USGS 3DEP LiDAR Dataset

SpatialAPI'20, November 3–6, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8164-2/20/11...$15.00
https://doi.org/10.1145/3423452.3430693

The USGS 3DEP LiDAR dataset is presented in two formats. The first is a public repository in Entwine Point Tiles format, which is a lossless, full resolution, streamable octree structure based on LASzip encoding; this format is suitable for online visualization. The other is in LAZ (compressed LAS) format.

3 TUTORIAL TIMELINE

3.1 Part 1: Introduction to AWS ML Services (20min)

To provide some background we will begin with an introduction to these AWS ML services: Amazon SageMaker, a fully managed ML service that provides users the ability to build, train, and deploy ML models quickly; Amazon SageMaker Ground Truth, a data labeling service that makes it easy to build highly accurate training dataset. Amazon SageMaker Notebook Instance, a ML compute instance running the Jupyter notebook app, offering a ML development environment that allows users to prepare and process data, write code to train, deploy and validate models. Amazon SageMaker also provides several images of built-in ML algorithms that make the training process much smoother and simpler; In the training job, Amazon SageMaker Hyperparameter Tuning helps to tune the hyperparameters and find the best version of a model automatically. After training, Amazon SageMaker can deploy the trained model into production with a single click so that it can start generating predictions for real-time or batch data input and monitor the performance of model.

3.1 Part 2: Walking Through Example Notebooks (70min)

We will walk through the main tutorial content available in GitHub repository Link. We will execute the notebook commands in real time on Amazon SageMaker and supplement with slide presentations in between. There are two notebooks in the tutorial resources as follows:

Building extraction. The 1st and 2nd SpaceNet challenge aimed to extract building footprints from the satellite images in various AOIs. The 4th SpaceNet challenge posed a similar task with more challenging off-nadir (i.e. oblique angle) imagery. We reproduce a winning algorithm and evaluate its performance for both RGB images and LiDAR data. Fig. 1 shows examples of input imagery (RGB + LiDAR), predicted building masks by models trained with both RGB and LiDAR data, and ground truth building masks.

Road extraction. The 3rd SpaceNet challenge aimed to extract road networks from the satellite images, while the 5th SpaceNet challenge added predicting road speed along with the road network extraction in order to minimize travel time and plan optimal routing. Similar to building extraction, we will reproduce a top winning algorithm, train different models with either RGB images, LiDAR attributes, or both, and evaluate their performance. Fig. 2 shows examples of input image (RGB + LiDAR), predicted road masks by models trained with both RGB and LiDAR data, and ground truth road masks.

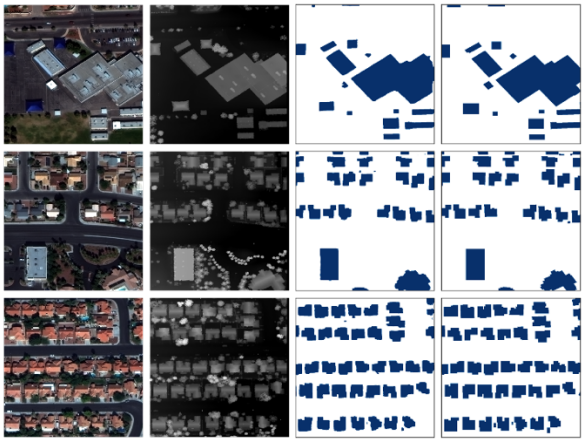

Figure 1: Examples of building extraction inputs and outputs. Columns from left to right: RGB image, LiDAR elevation image, model prediction trained with both RGB and LiDAR data, and ground truth building footprint mask.

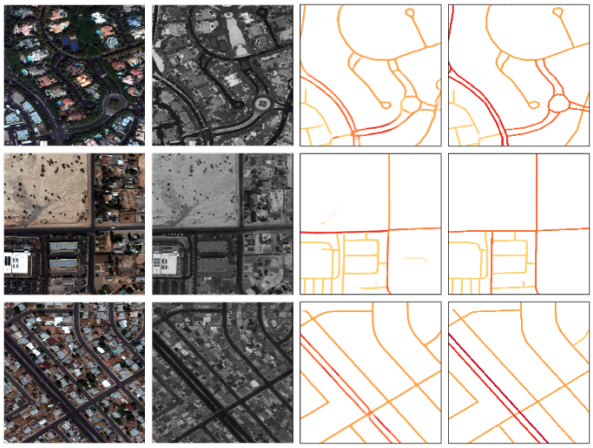

Figure 2: Examples of road extraction inputs and outputs. Columns from left to right: RGB image, LiDAR reflectivity intensity image, model prediction trained with both RGB and LiDAR data, and ground truth road mask.

5 SUMMARY

In this tutorial we present reproductions of SpaceNet winning algorithms by developing ML models on Amazon SageMaker instances to automatically extract building and road from large scale geospatial data in AWS Open Data Registry. In addition to RGB satellite imagery, we process USGS 3DEP LiDAR data and incorporate the LiDAR attributes in those models. Using dataset in Las Vegas AOI, we show LiDAR data can be used to train a model that performs the same task with similar accuracy, and outperforms RGB models when combined with RGB imagery. The audience acquires first-hand experience in cloud-based ML on AWS that is scalable, cost effective and easy to innovate.