SNA Project – Predicting Poverty in India by combining Multiple Nontraditional Big Data Sources using Machine Learning

Ujjwal Sehrawat

2017368

Introduction (Motivation)

- Poverty - a multidimensional phenomenon
- Traditional poverty metrics - SAE technique (census + survey)
- Big Data with ML algorithms can be used as proxy (better granularity, cheaper, timeseries)
- Data Integration - an additional step.

Literature Review

- Remote sensing - satellite land daytime & nighttime images.
- Telecommunications / mobile data - CDR.
- Social sensing data - POI (restaurants, malls, etc), social media user behavior.
- Incorporate Environmental, political, and cultural power inequalities.
- Other proxy data: Street-view images.
- Data Integration of these including ancillary indicators of poverty.

i

Dataset Description

- Granularity - district levels - approportioned.
- Traditional Variables for poverty estimation
  - Health
  - Education
  - Standard of living
  - ICRISAT macro data
  - Control - Caste, age, disability, religion, birth rate

- Non-traditional variables (our focus thus far):
  - Satellite - landsat daytime imagery
  - Satellite - OLS nighttime imagery
  - Satellite - MODIS NDVI Imagery
  - Social sensing data - POI (city to village - time taken to travel paths) - in the process of collection through Google Map API as it has associated costs. ($0.004 per road travel distance between two points)

Data Collection + Preprocessing

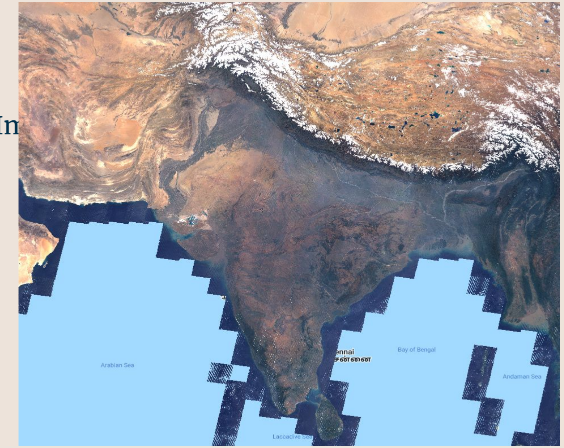
- Data collection of all traditional variables for available years. For now, we focus on year 2011 and the state of Jharkhand.

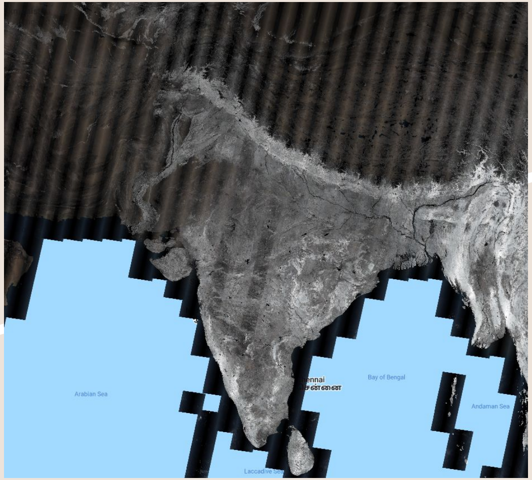
- Preprocessing satellite images before training the Visual Encoder CNN - Resnet18.

- Google Earth Engine Javascript API has been used to extract all satellite imagery data. Distance data between POIs is being extracted from Google Maps Distance Matrix API.

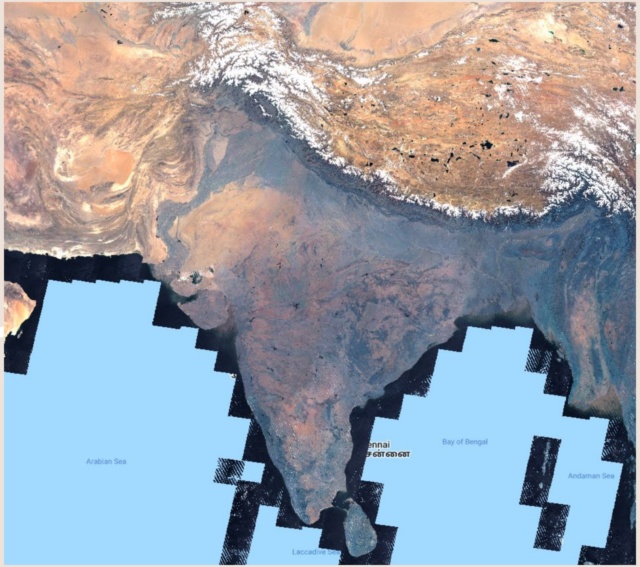
- Images checked and processed for-
  - Cloudiness -> masked pixels
  - Composite images - averaged image for a year
  - Resolution comparability between landsat and OLS.
  - Shapefiles for districts -> 3D arrays exported per district. [R G B color bands]

Normal Im



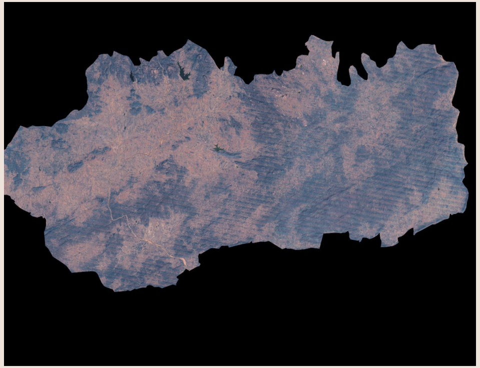


Panchromatic Image

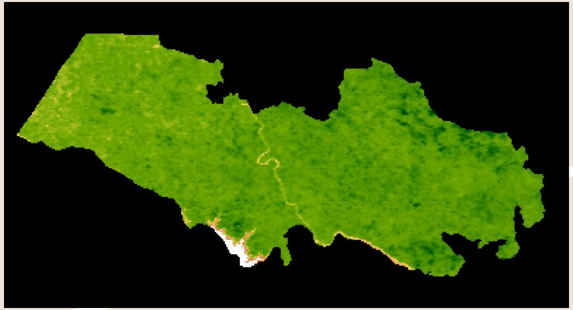


Pansharpened Cloudfree Image

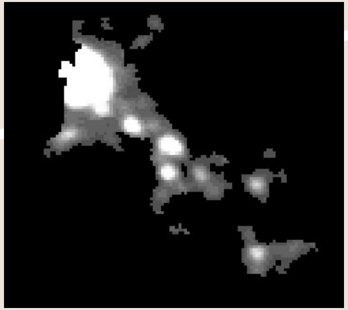
Satellite Image Inputs - District-level



Simdega - A district in Jharkhand (Landsat 2011 composite cloud-free, pansharpened, 15 m resolution RGB Image)

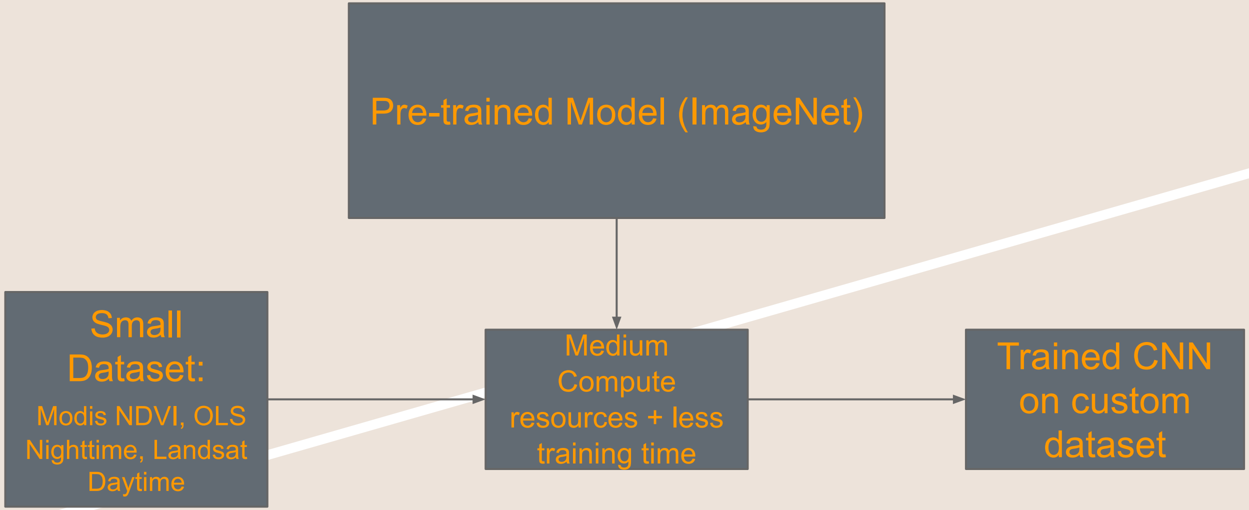


Jamtara - A district in Jharkhand (MODIS 2011 composite cloud-free, pansharpened, 30 m resolution NDVI band Image)

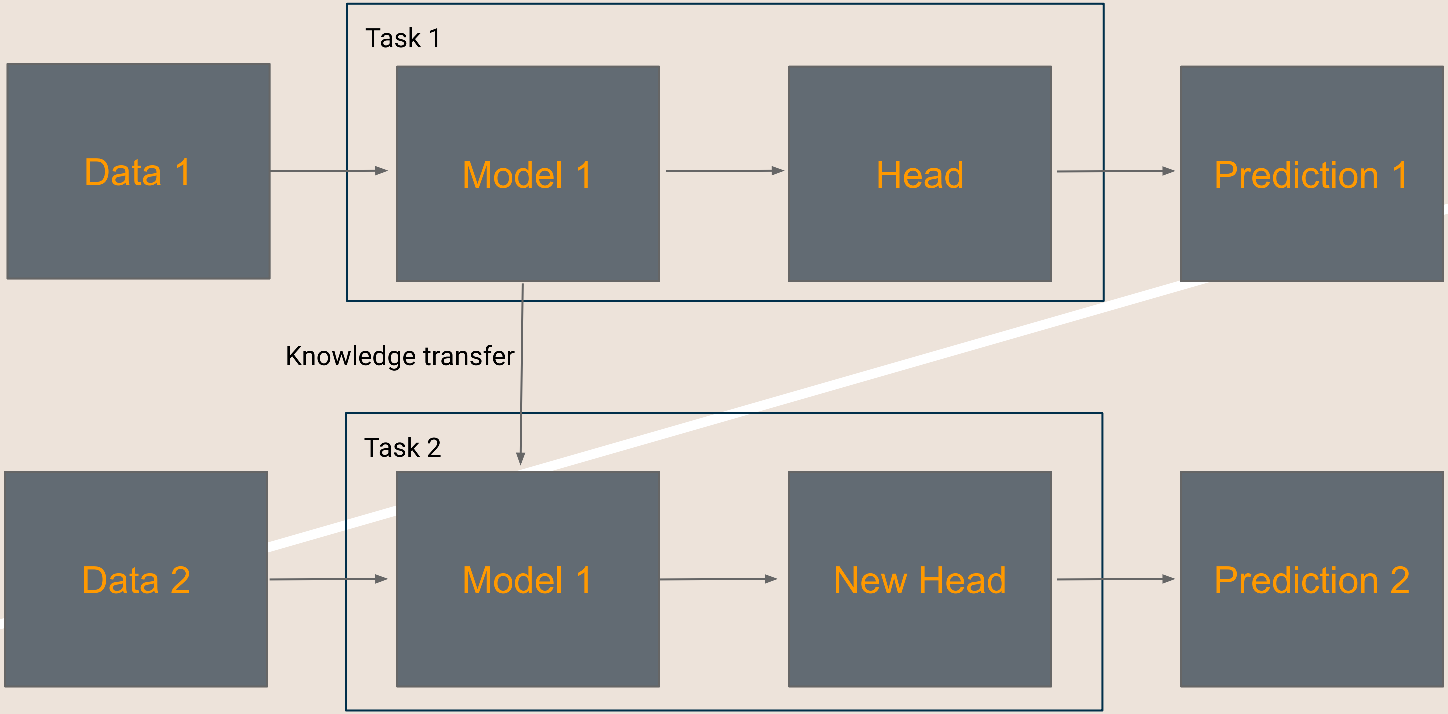


Dhanbad - A district in Jharkhand (OLS Nighttime 2011 composite cloud-free, 1 km resolution “stable lights” band Image)

Unsupervised Learning

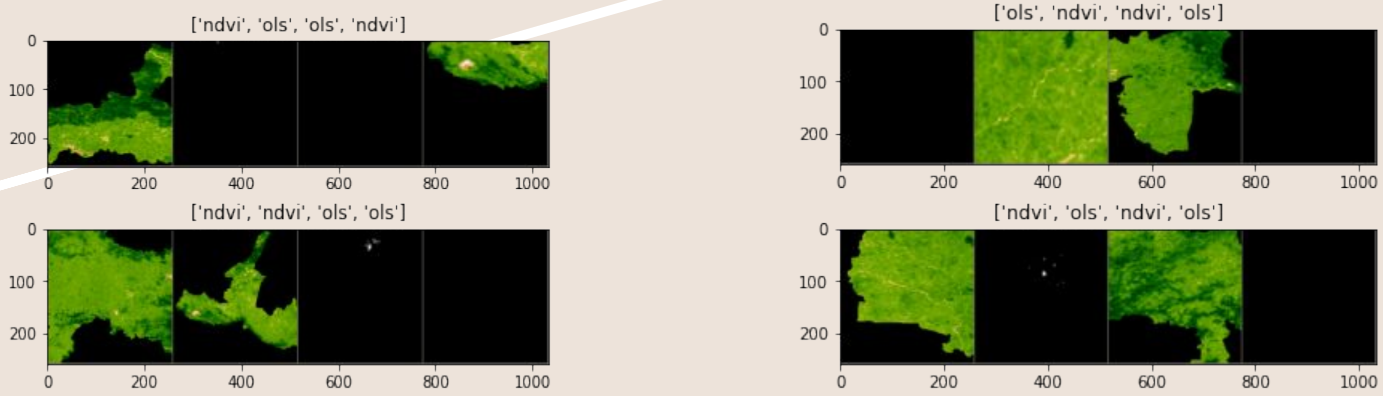


Transfer Learning Approach



Analysis and Results

- So far collected and preprocessed data on OLS Nighttime, Landsat Daytime, MODIS NDVI, Road distance between Village and nearby city (in progress).
- Transfer learning technique (Jean et al, 2014) used for satellite image data.
    - Visual Encoder CNN - ResNet18 (pretrained on ImageNet) - identifies simple image features.
    - We remove the final (classifier) layer as we need early layers only since our dataset is small - then we fine tune.
    - CNN trained with MODIS NDVI data and OLS Nighttime data to detect more complex image features that can predict the latter based on former.
    - Accuracy - an average accuracy of 90% in the below task.



Going Forward

- So far we have done the convnet model training with OLS nighttime and NDVI, and learned the numerical features.
- The next step is to regress real-world poverty estimates with these features (ridge regression) to build a predictive model.
- The same will be done between Landsat and nighttime and other more granular metrics (road distance map)
- Data Integrated through Information Infusion process used in Deep Learning.
- Check feasibility at each stage and finally after data integration.