Capstone project title: Sea lion population count in the western Aleutian Islands, Alaska based on the aerial images of the coast

Problem to solve:

One of the major focuses of restoring the endangered population of Sea lions in the Aleutian islands is regular population counts, that are performed annually. Changes in population helps to define the contributing factors, find dependencies. Thousands of photos are usually taken by scientists during the survey period.
As the majority of the photos are taken by the drones - there is a high chance that the photo will not contain any seals on it. After the photos are taken, the results of the counting is time sensitive, but now it takes several months for the scientists to process images. From my point of view, there are 2 major opportunities for the application of the machine learning:

- Define which photos contain useful information - at least one seal for counting - this may be done preferable at the board of the unoccupied aircraft system not to even store not meaningful data. This step will reduce storage and processing time for the counting, as less images will be processed. **But makes sense only if this gives any benefit in that. Can I tell that on this stage?**
- Count the seal population on the rest of the photos, defining number of:
    - adult males,
    - subadult males,
    - adult females,
    - juveniles
    - pups.

(Question): Should I approach this as 2 different steps or there won't be any difference between 2 models and just one model that counts empty images as 0 seals on them. May be makes sense to do the comparison in this project?

    -

As of now NOAA fisheries user crowdsourcing to perform the counting:
https://www.zooniverse.org/projects/sweenkl/steller-watch

Why this is an interesting problem:

-> Application of machine learning to this problem can practically help the endangered species program, increase the amount of cases of practical application of ML that will help advocating for ML products investments.
-> This project deliverable will be focused on the particular dataset, but the approach taken can be re-used for the other types of the aerial photography identification. Plan to have two different models - one model will be detecting yes/no of the sea lions presence. When the photos are "cleaned up" - second models will be counting the number of different types of the sea lions.

-

Data

This project will be using dataset provided by NOAA on Kaggle:
https://www.kaggle.com/c/noaa-fisheries-steller-sea-lion-population-count/data
Data archive contains 98,5 GB of photos, divided into:

- Train/*.jpg: a set of training images, with each filename corresponding to a train_id
- Train/train.csv: a list of ground-truth counts for each train_id
- TrainDotted/*.jpg: copies of the training images with markings showing where each animal is (see below)
- Test/*.jpg: a set of test images, with each filename corresponding to a test_id

Train set: 

Train.csv:

```
train_id,adult_males,subadult_males,adult_females,juveniles,pups
0,62,12,486,42,344
1,2,20,0,12,0
2,2,0,38,20,0
3,8,5,41,7,38
4,6,9,2,0,0
5,6,4,14,4,19
6,2,5,20,18,0
7,6,2,33,16,3
8,9,5,76,4,51
```

TrainDotted folder contains the same images as in the Train folder, but with colored dots placed over the animals. The color scheme of the dots is:

- red: adult males
- magenta: subadult males
- brown: adult females
- blue: juveniles
- green: pups

Example: 

Approach

- Supervised or unsupervised: Supervised. The training set is prepared by the scientists who do the counting every year.

- Classification or regression: Classification.

- Prediction: The model should be able to predict the amount of the 5 different types of the sea lions on the given image.

- Predictors: Aerial images of the coast. Sea lions silhouettes differ by color and shape from the non-animal objects on the photos.

I found other animals on the photo:

- Approach: May use the deep learning:

- Deep Residual Neural Network
- Deep Convolutional Neural networks

May be using Keras framework. Will need to use some approach for dimensionality reduction.

Next steps we can try to use Yolo or SSD for building stat-of-the art classifier.

- Bounding boxes can give us, count of the objects
- Bounding boxes -> gender for giving us the specific count per gender and adult/pups of the object.

Final deliverable

Application deployed as a web service with an API and a simple web app to interact with API.

Swati

- We may need a slide deck as an outline of the entire project.

Computational resources

Resources are usually significant concern for the image classification. Will be using local machine for the start of the development and training, using the subset of photos Kaggle provides for this challenge. Deep learning approach seems to be working good on GPU, so will probably use AWS: P2.8xlarge or P2.8xlarge - 1/8 GPU - 61/488 GiB (https://aws.amazon.com/ec2/instance-types/p2/) ?