Deep Learning for Semantic Segmentation of Remote Sensing Imagery

William Chen, Nick Guo, Sherrie Wang
{wic006, nickguo, sherwang}@stanford.edu

Introduction
- Identifying the location of croplands would greatly benefit agricultural development, food security assessment, and poverty reduction.
- However, progress in creating crop maps is limited by a lack of segmentation data in regions of interest.
- We train neural networks on multi-task classification and use intermediate layers to segment images.

Dataset
- Landsat 8 satellite median composite for 2016.
- 4.5 degrees latitude by 8.0 degrees longitude.
- 500M pixels divided into 194k patches (50x50 px).
- Segmentation ground truth from USDA’s Cropland Data Layer (CDL).

Features
- Landsat 8 bands:
  1. Ultra blue
  2. Blue
  3. Green
  4. Red
- NIR and SWIR capture ground properties that are difficult to see in RGB. For this reason, they are effective for separating land cover types, and often play a key role in pixel-level supervised classification problems.

Models
- Baseline Model:
  - 4 conv/conv layers of increasing filters followed by a dense layer.
- Modified ResNet-50 Model:
  - 5 stages containing combinations of convolutional blocks and/or identity blocks, followed by a dense layer.
  - To obtain a last conv layer that is high res. (12x12) for use in segmentation, we set all strides to 1.

Multi-Task Learning
- To simulate conditions in data-poor settings, we re-frame the problem as a multi-task classification problem.
- Each segmentation is turned into a 5-dimensional binary label, corresponding to whether background, corn, soybean, forest, and grassland pixels respectively appear in the image.

Semantic Segmentation
- Following Zhou et al. [3], we calculate a class activation map (CAM) for each of the 5 classes and compare them to ground truth segmentation.
- Using (1) the last ResNet convolutional layer output F with k filters and (2) a dense layer weight matrix W, the CAM for class c is defined as:
  \[ CAM^c = \sum \mathbf{w}^c_k f_k(x, y) \]
- Taking the argmax over the 5 CAMs to obtain a segmentation map results in low average segmentation accuracy of 0.18.
- However, there is still some correspondence between each CAM and ground truth, as seen below.

Discussion
- Both a simple baseline network and a ResNet achieve high classification accuracy on the multi-task problem.
- The initial ResNet model we developed does not perform as well as our simpler baseline CNN, perhaps due to truncation of the ResNets later layers.
- High classification accuracy does not translate to high segmentation accuracy with our current strategy.
- Future work includes trying different ways to generate segmentation from CAMs & new architectures (e.g., U-Net).

References