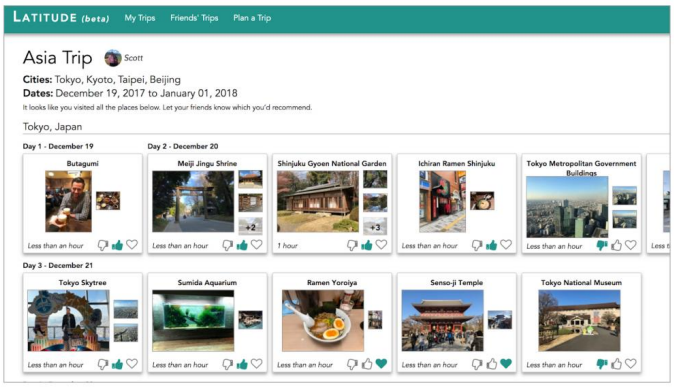
Identifying Travel Stops from a Photo Album (Localized Landmark Detection)

Scott Ings, MBA '18 (sings@stanford.edu)

Motivation

Sharing travel recommendations is harder than it should be. What if we could use your trip photos to reverse-engineer your itinerary for you?



Problem Definition

Goal: Build a location-specific landmark detection model

Input: Travel photos and a location (e.g. city)

Output: Label for any popular attraction identified in each photo

Experimental Dataset: Google labeled landmark detection Kaggle data (subset: 118 landmarks, 144k records)

Train	Dev	Test
100,978	21,640	21,642

Challenges

- Large number of total classes (restricting to local geographic area is huge improvement)
- Landmarks can have many very different visual components and can be quite broadly defined; photos usually include only a subset
- Lots of noise in photos not related to classes (e.g. people)
- Landmarks often have a lot of visual detail, which can be lost at lower resolutions
- Training end-to-end models can take a lot of processing time for this type of data

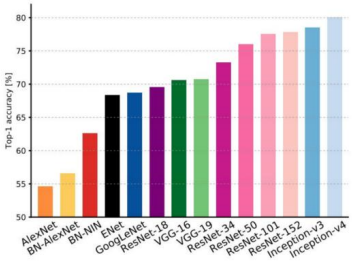
Approach

ImageNet Transfer Learning

Given the large overlap in application, use transfer learning on models shown to have success with the ImageNet classification problem:

- Inception v3 – High Accuracy
- MobileNet – Small Footprint

ImageNet Model Accuracy (Top 1)



Key Hyperparameters

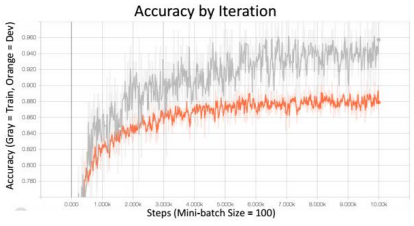
- Mini-batch size
- Resolution of images
- Size of training set
- Layers to retrain
- Image duplication

Results

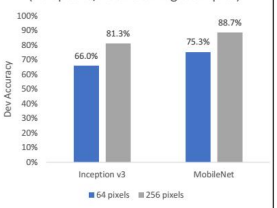
Best Model: MobileNet (100% size), 256px resolution

Test Accuracy: 85.8% (Train 99.0%, Dev 88.7%)

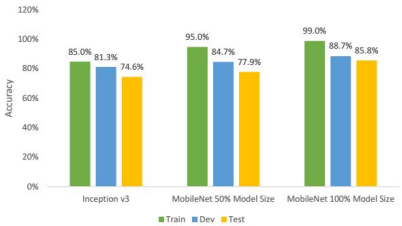
Accuracy by Iteration



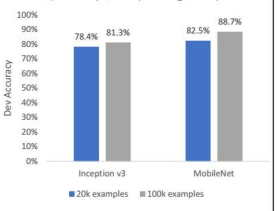
Impact of Image Resolution
(10 epochs, 100k training examples)



Comparing Architectures
(10 epochs, 256px image size, 100k training examples)



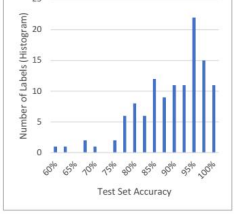
Impact of Training Set Size
(10k steps, 256px image size)



Discussion

- Somewhat surprisingly, the smaller MobileNet model outperformed Inception v3
- Error rate was relatively well distributed across landmarks
  - Worst examples had high-variance images (e.g. skyline or indoor/outdoor)
- Image resolution has a major impact on accuracy
- Training set size has smaller, but meaningful impact
- Mini-batch size had negligible impact
- Most promising exploration going forward:
  - Experiment with even higher resolutions
  - Gather more training data
  - Experiment with random distortions
  - Retrain additional model layers
  - Different models for indoor/outdoor?
  - Incorporate image timestamps?

Test Accuracy By Ground Truth Label



Error Example:
High-Variation in Landmark

Error Example:
Indoor and Outdoor Aspects