Neighborhood Watch: Using CNNs to Predict Income Levels on StreetView Images

Ambika Acharya, Helen Fang, Shubha Raghvendra
{aacharya, hfang9, sraghven}@stanford.edu

MOTIVATION

The government spends millions of dollars every year attempting to collect income-level data throughout the United States. This process can not only be expensive, but time-consuming and often inaccurate. We are interested in predicting the income levels of different communities within a city based on images gathered by Google Street View cars. We hope these techniques can be leveraged to improve the information gathering process.

QUESTION

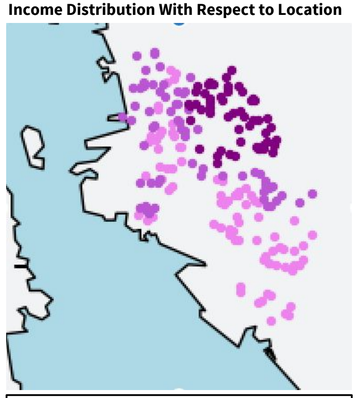
Given Google Street View images for a particular neighborhood in Oakland, California, can we predict the income index of that neighborhood? What parts of an image are most indicative of income and how do these differ across income levels?

DATA SET

The Google Street View dataset lists urls to images displayed in 6 rotations for a given location. The most recent census provided income data. We labeled images with income using the FCC's Census Block Conversions API. For our multi-class classification approach, we used ~40K images from Oakland, because there is more income variance relative to other Californian cities. The income classes are <75k, 75-150k, and >150k based on the income distribution data.

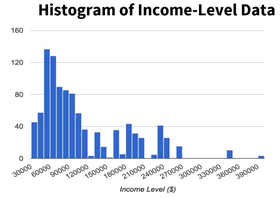
To the left is a sample Google Street View image found at 37.805045° latitude, -122.211325° longitude, and 60° rotation.

Income Distribution With Respect to Location



Histogram of Income-Level Data

The figure on the right shows the distribution of income levels from 1000 sampled images. The distribution skews towards the left with a wide spread of points above <150k.

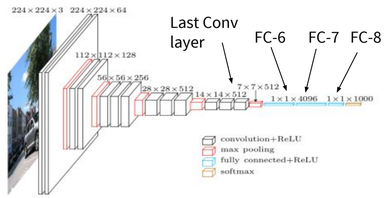


REFERENCES

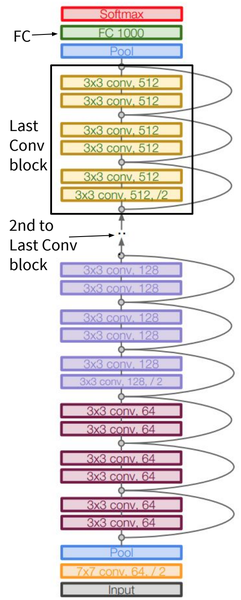
[1] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38, 2010.
[2] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of the us. *arXiv preprint arXiv:1702.06683*, 2017.
[3] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and L. Fei- Fei. Fine-grained car detection for visual census estimation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
[4] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

METHODS

VGG-16



ResNet-18

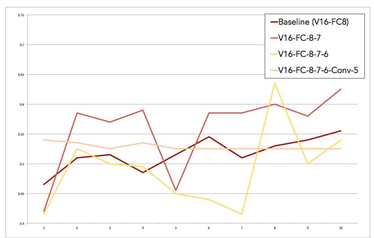


<u>VGG-16 Models</u>
Baseline: VGG-16 trained on ImageNet, FC-8 fine-tuned to 3 classes.
V16-FC-8-7: Baseline + FC-7 fine-tuned.
V16FC-8-7-6: Baseline + FC-7, FC-6 fine-tuned.
V16-FC-8-7-6-Conv-5: Baseline + all FC's and last Conv layer fine-tuned.

<u>ResNet-18 Models</u>
RN18-FC: ResNet-18 trained on ImageNet, FC fine-tuned to 3 classes.
RN18-FC-Conv4: RN18-FC + last Conv block fine-tuned.
RN18-FC-Conv4-3: Baseline + last 2 Conv blocks fine-tuned.

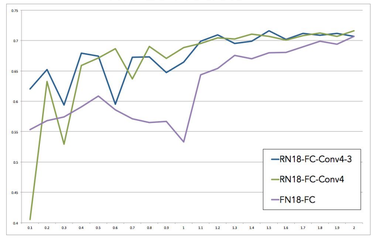
RESULTS

Validation Accuracies Across VGG Models



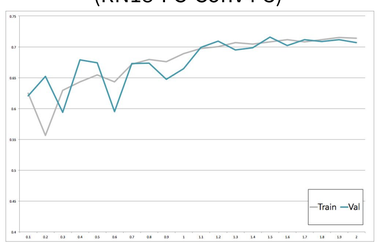
Validation accuracy for 4 different training architectures for the VGG across 10 epochs, trained on 40K images. V15-FC-8-7 was the highest-performing, and that ResNet architectures (shown on right) outperformed VGG architectures.

Validation Accuracies Across RN-18 Models



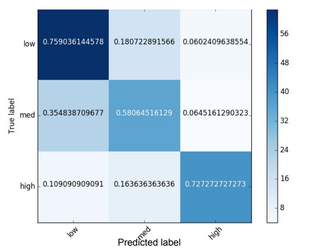
Validation accuracy for 4 different training architectures for the ResNet across 20 epochs (10 learning reinitialized params, 10 tuning all model params), trained on 40k images. Note that RN18-FC-Conv4-3 and RN18-FC-Conv4 were the highest-performing.

Validation versus Train Accuracy (RN18-FC-Conv4-3)



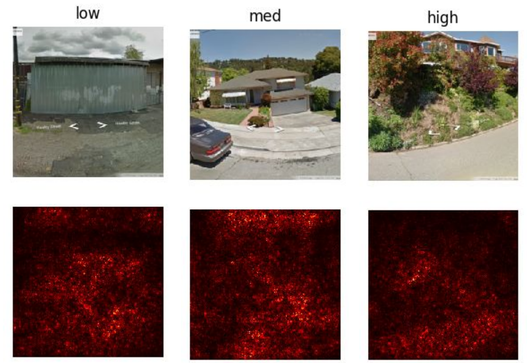
Validation versus train accuracy for our highest-performing model, in which we retrained the 3rd and 4th convolutional layers and the final fully connected layer. The key take-away from this figure is that our model did not overfit.

Normalized Confusion Matrix



The confusion matrix gives a visual representation for the percentage of labels that are predicted correctly and incorrect in each category. Per our discussion on the right hand side, RN18-FC-Conv4-3 is apt at correctly identifying low-income areas.

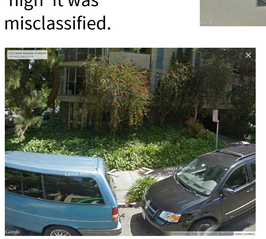
Saliency Maps of the Different Income Brackets



These maps are of images correctly classified by the ResNet FC-8 model on 1000 images. Images labeled 'high' tend to be identified by how much greenery the image has, 'low' images by their concrete and metal content, and 'med' a mix of both.

Visualization of classifications with respect to location

This image's true label is 'high' but was classified as 'med'. We believe because this image has less greenery than most from class 'high' it was misclassified.



This image's true label is 'low' but was classified as 'high'. We believe because this image has more greenery than concrete, unlike other images from 'low' it was misclassified.

DISCUSSION & FUTURE WORK

We found that classifying income brackets based solely on neighborhood image snapshots is a challenging task. From saliency maps we see that our best model, RN18-FC-Conv4-3, was able to best detect greenery in high income images, and concrete and metal structures in lower income areas. Additionally, we find that predicting medium income neighborhoods is the most challenging, likely because the characteristics of these images fall in the intersection of the other two classes. Compared to prior work, the best accuracies for similar tasks are ~70% as well, suggesting this task should be supplemented with additional data or image segmentation techniques. Additionally, grouping together sets of images in a neighborhood using pointer networks would better reflect mean characteristics of that area and could be a more meaningful representation. We also ask how results and saliency maps differ across various cities in California.