Automated Detection of Street-Level Tobacco Advertising Displays

Charlie Moffett [1] , Prince Abunku [2] , Jianghao Zhu [3] , Isha Chaturvedi [4] , Guobing Chen [5], federica B bianco [6] , Greg Dobler [7]

Abstract

Tobacco marketing, restricted almost exclusively to the point-of-sale in recent years, has proven to be effective in getting more people to consume and fewer to quit cigarettes and smokeless tobacco products. The lack of empirical documentation linking product exposure to behavior, however, is a key obstacle to the adoption of additional restrictions on point-of-sale tobacco advertising. The goal of this project is to map point-of-sale tobacco marketing practices across New York City using automated detection of tobacco signage in street-level imaging data. Convolutional neural networks, which are particularly effective at detecting objects in images, were trained to identify and classify outdoor advertisements of cigarettes and smokeless tobacco. Previous analyses of visual data in public health research involving manual image coding are prohibitively costly and time-consuming. The importance and motivation of the project stems from the immediate and comprehensive effect of tobacco advertisements on its sales and consequently on public health. Detected advertisements derived from our model output provide a proof-of-concept for measuring exposure of at-risk communities to tobacco displays.

1 INTRODUCTION

1.1 Motivation

POST advertising is an embedded element of the urban landscape of New York City. Comprised of a variety of marketing practices, it includes signs on the insides and outsides of retail stores, and has a more immediate and comprehensive effect on tobacco sales than any other marketing channel [24]. There is substantial evidence of disparity in the way tobacco products are advertised at the point-of-sale depending on the community demographic profile of focus. Tobacco manufacturers have long targeted minority and lower income populations in particular, mainly through this sort of tailored advertising [33]. The goal of this project is to map POST marketing practices across New York City (NYC) using an automated method of detecting and classifying tobacco signage [5]. In comparing the POST landscape with socioeconomic characteristics at the neighborhood level, we also aim to explore marketing disparities and variable exposure of communities to tobacco advertisements.

1.2 Research Statement

The gap between how humans are able to perceive the world, both graphically and semantically, and how computers can be trained to automatically interpret imagery data, is rapidly closing [38; 5; 11]. The automated tobacco signage detection model employed in this project involves Faster R-CNN, a state-of-the-art convolutional neural network related to image recognition [32]. By efficiently discriminating between backgrounds and targets within an image, this model enables a detection algorithm to focus on areas that are more likely to contain tobacco signages. Development of a model that can not only identify signs, but also distinguish between tobacco ads and other types of signs (i.e. stop signs, cellular service provider ads), is critical to successful detection when applied to images that the model has not seen before. In this project, we seek to improve the existing Faster R-CNN model in its ability to identify signs and discriminate tobacco signages from other types of signs in NYC.

1.3 Literature Review

There are many pathways from examining urban environments to analysis of human health and policy-making. Writers for centuries have suggested that cities shape individual well-being, including German philosopher Friedrich Engels [14] and French sociologist Émile Durkheim [13]. In recent years, there has been a renewal of interest in geographic characteristics in the field of public health. However, understanding whether there are specific features of the urban context that are causally related to urban health is complex. First, while cities may share high-level characteristics [39], cities are markedly different from one another and can evolve drastically over time. Second, multiple interrelated and constituent factors are often important determinants of population health in cities [19]. For example, efforts to curb tobacco and alcohol usage with taxation might be stifled or even have the opposite effect depending on social norms at the local level [21].

While the application of technology to health-related social and behavioral research in and of cities is a longstanding practice, the emergence of real-time activity tracking and other remote sensing modalities has inspired new confidence in causal inference [1]. Big data methods are not only less costly than direct observation, but also enable coverage of larger spatial and temporal ranges. These expansive and increasingly sophisticated information stores could lead to more effective

interventions at the clinical level and in public health practice, and should provide the basis for more focused and effective health policy [4].

Existing methods of research regarding the impact of retail tobacco advertising on smoking behavior are difficult to scale up for various reasons. Studies demonstrating the correlation between tobacco retailer density and cigarette purchases highlight the significance of how the distribution of urban features impact social phenomena, but fail to document important variations in display characteristics that could be used to support additional regulation toward reducing the impact of displays [25]. Proximity to public schools are a known factor with respect to exterior displays of tobacco advertisements on the part of retailers, but specific details about the visual landscape that citizens experience are again omitted from analysis [23]. Both of these types are studies are at risk of geographical bias by focusing on individual communities, and are prohibitively labor-intensive when attempting to scale across larger areas despite employing digital data collection methods.

More recently, public health researchers have started using photography as a means of achieving scale and accessing visual information about remote locations [22]. Use of crowdsourcing methods, like those provided by Amazon's web-based marketplace Mechanical Turk (MTurk), has been effective in enabling time-efficient photograph annotation for detection and classification of tobacco ads in imagery, but leaves analysis susceptible to bias by anonymous MTurk workers. Furthermore, the cost of hiring MTurk workers to label images for areas larger than a mid-sized city may prove too costly for clinical studies of tobacco marketing.

Recent advances in neural networks show promise towards automating image analysis[12]. A neural network, is a computational model for recognizing patterns and objects[37]. Convolutional neural networks (CNNs) are a subtype of neural networks that are being widely used for image classification, as they provide higher accuracy, require less computational power and can learn appropriate features by themselves automatically. In classification, there is generally an image with a single object as the focus, and the task is to identify that image[28]. However, often time the images of interest are more complex, with multiple overlapping objects and heterogenous backgrounds, which makes it difficult for a CNN model to detect these objects. This issue has been addressed by region-based CNNs (R-CNNs) which take an image set and identifies the locations and classifications of the main objects in the image. This model is used for our project for detecting tobacco signages.

2 DATA

2.1 Street View Imagery

Our primary data source for the training and testing of our predictive model are images from Google Street View [10]. Addresses for each active retail tobacco vendor in New York City, taken from Health Data NY [27], were used with the Google Street View API to extract images of potential POST displays. Four images of 90 degree rotation angle each were taken from every known address of tobacco vendors in New York City in order to cover the panoramic 360 degree view available at each location. This process returned over 42,776 images.

To label the training set, we visually inspected the images one-by-one to mark the tobacco signages. Training the model requires thousands of annotated images so that the model can go through these images and discover the important features of tobacco signages. Labeling is simply determining whether an image contains the object we are looking for and indicating where the object is located in the image. This data is stored in an XML file, and these files can be generated with existing tools. We use the tool *labelImg* [36] for our labeling exercise, as it had been extensively vetted and saves the labels in the PASCAL (Visual Object Classes) VOC format [16]. This format of images is the standard input to the Faster RCNN model.

Street view images were manipulated using variations in color balance, contrast, and other image pre-processing techniques. This has several purposes. The primary reasons are to reduce overfitting and predict different sources of data. Secondarily, we sought to provide the automated detection model with generalized data so that systematic biases could be minimized.

2.2 Socioeconomic Data

To analyze youth exposure to tobacco retailers, we gathered Census Block Group geometries and youth population statistics from American FactFinder [17]. Locations of retail tobacco vendors with active licenses from the City of New York were borrowed from the dataset used to generate street view imagery. Because the original dataset for tobacco retailers contained point locations for all New York state, the dataset was filtered for only those retailers found within the administrative boundaries of New York City boroughs [30].

3 METHODOLOGY

3.1 POST Detection via RCNN

To automate the detection of POST displays, we are using a deep learning algorithm known as Faster RCNN [32] . The implementation of Faster RCNN is adopted from existing GitHub repository by endernewton [9]. In addition to preparing training data and validation data, three different anchor scale sets of [4,8,16], [8,16,32], and [4,8,16,32] were set for 3 training attempts with 70,000 iterations. Due to the limited number of tobacco advertisements detected in images from Google Street View in comparison to object detection standards, the [4, 8, 16] anchor scale for our training set provided the best mean average precision (mAP).

Fig. 1
Image retrieved from Google Street View in Brooklyn with angle set as 0 degree

Fig. 2
Image retrieved from Google Street View in Brooklyn with angle set as 90 degree

Fig. 3
Image retrieved from Google Street View in Brooklyn with angle set as 180 degree

Fig. 4
Image retrieved from Google Street View in Brooklyn with angle set as 270 degree

Figure 1. Images retrieved from Google Street View in Brooklyn at four 90 degree rotate angles.

Figure 2. Image retrieved from Google Street View in Brooklyn with Newport-branded tobacco ad tagged with bounding box (tagged in green color here).

Our main metric for evaluating performance is mAP. This measure is our precision for each class in our data that we are detecting. The mAP value can change based on choices of our parameters such as number of iterations, anchor size, and classification threshold. Anchors are region boxes that are used to contain objects in an image. Multiple anchors are generated for an image and ranked according to how likely they contain a single object. The number of iterations reflects the number of times our model trains on a dataset. The higher the number of iterations, the more opportunities our model has to reduce the error between its output and the correct classification of a sign.

3.2 Study of spatial and social components

To visualize POST advertising density as a function of geographic location and number of nearby youth, we built an interactive web mapping application using an open source stack of QGIS [35] for data cleaning and handling, and Mapbox GL JS [34] to embed the responsive map online. Catchment areas radiating from the center of each Census Block Group (CBG) captured the number of detected tobacco advertisements within an acceptable walking distance of 400 meters [40]. This ad display count was tempered with

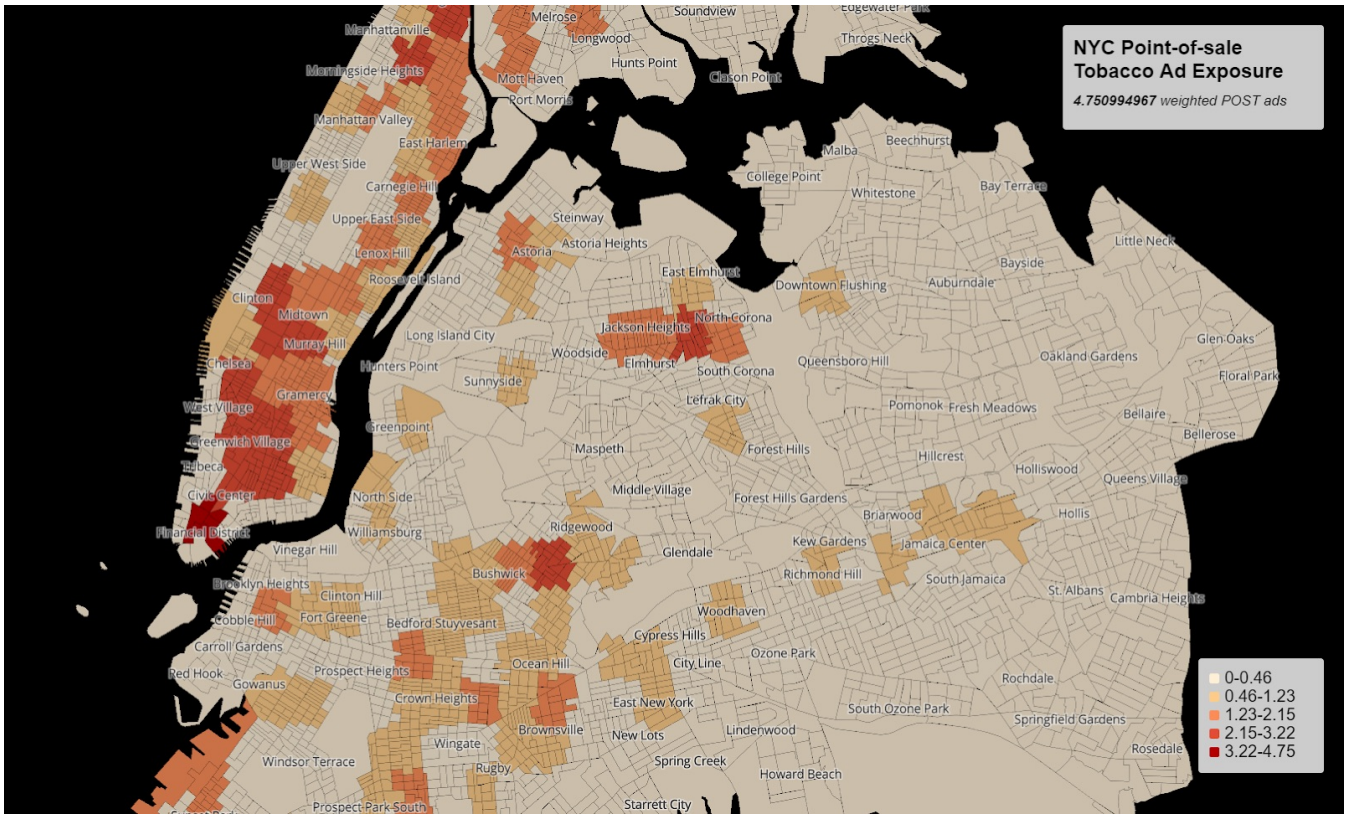

Figure 3. POST display density weighted by youth population as a function of geographic position. This choropleth map provides insight as to how youth exposure to tobacco advertising varies across New York City.

a weighted youth average by summing up all of the ad displays detected and youth population products for each CBG and dividing by the total number of youth in New York City to prevent high youth density areas from artificially carrying more significance than other exposure zones.

Choropleths are a popular choice for visualizing data distributions as thematic maps. Our choropleth shades CBGs in New York City based on the weighted display counts described above. The values in the data distribution are grouped into five classes using the Jenks natural breaks method, which seeks to optimize the arrangement of clustered values by minimizing the variance within each class and maximizing variance between classes. The map demonstrates in a particularly dramatic manner how youth exposure to tobacco advertising varies across NYC.

4 RESULTS

From Table 1, we can see improvement of 6% in mAP by comparing Model A with Model B . In this project, all tobacco advertisements, regardless of brand or message, have been grouped as a single class. Both Model A and Model B were trained and tested with MD17 dataset. The only difference between model A and model B are the anchor scales. Model C was adopted from Model B by continuing training and testing using the PANIC18 dataset. PANIC18 was also derived from Google Street View but annotated and manipulated according to the procedure described in this project. We found that PANIC18 has more variety in terms of tobacco promotions and products, including electronic cigarettes, while MD17 focused almost exclusively on traditional cigarette brands like Marlboro and Newport. Since PANIC18 is a smaller dataset, Model C fell short of learning the features of less common products.

From Figure 4, the model detected all four traditional cigarette brands signs in one image with high probabilities. Numbers next to bounding boxes are the probabilities of such objects being tobacco advertisements. Similarly, the traditional “Newport” and “Marlboro” brand signs were detected as well. Such detections are typical examples that were found from our model output. We used a classification probability of 0.7 for the model.

From Figure 5, we can see that our model incorrectly detected street signs, blurry boxes, and traffic cones which could be impacted by the sign and similarities show in the bottom row. They share features like color combination, shapes,

and sizes. Other false detection examples include trash bins, windows, and glass doors.

Figure 4. True positive detection examples from Model C output. The numbers next to bounding boxes are the probabilities of such objects being tobacco signs. Model C detected popular cigarette brands with near perfection.

Figure 5. False positive detection examples from the model output. The model incorrectly detected street sign, blurry boxes, and traffic cones as tobacco signs, likely because they share features like color, shape, and size.

	Model A	Model B	Model C
Anchor Scale	[8, 16, 32]	[4, 8, 16]	[4, 8, 16]
Iterations	70K	70K	70K
Dataset	MD17	MD17	PANIC18
mAP	0.56	0.62	0.13
Pretrained Model	Res101	Res101	Model B

Table 1. Detection performance of three models: Model A, Model B, Model C

5 CONCLUSIONS

Our main goal of this project was to automate the process of detecting tobacco ads in an urban environment. Automation reduces the amount of resources it takes to collect data on where signs are located. Improving on the accuracy of our model would allow us to also perform a longitudinal study as new images are captured for the same locations in order to analyze how the marketing practices of tobacco companies changes over time in the city, particularly in response to new regulations. As we continue our work, there are several methods we plan to employ to improve the precision of our detection algorithm. Collecting high resolution images, using the Google Street View premium service for example, would support human recognition of additional ad during the labeling process. Furthermore, a larger original training dataset would support our model in identifying unique characteristics of tobacco ads for use in classifying ads from yet-to-be-seen images of storefronts. We collected over 40,000 images but only had time to visually inspect about half of them. Finally, higher confidence in our detection of cigarette ads in the city using automated methods would be encouraging for detecting other urban features as well. Similar studies could be undertaken to determine the impacts of alcohol advertising or fast food displays, and provide another demonstration of how emerging technologies might be used beyond commercial applications to help improve our communities.

Because our model deserves improvement, the socioeconomic analysis map stands as a proof of concept. Upon strengthening the confidence of our detections, we plan to remeasure the exposure of children to tobacco advertising and expand our analysis to include scrutinization as to how companies advertise to different socioeconomic groups. As a result, we think it's quite useful as an exploratory analysis and begs the question about how the tobacco industry might be targeting various parts of NYC differently (prices, brands, products, marketing campaigns). Furthermore, it will be insightful in future work to overlay this data with school locations and likely trajectories to examine exposure risk specifically for school children on their way to, from, and around school. Big Tobacco clearly has a sense of where their target demographics reside and spend their time, but the breakdown probably isn't as straightforward as a spatial distribution of population density or median income.

REFERENCES

[1]Deepti Adlakha. Quantifying the Modern City: Emerging Technologies and Big Data for Active Living Research. *Frontiers in Public Health*, 5, may 2017. doi: 10.3389/fpubh.2017.00105. URL https://doi.org/10.3389%2Ffpubh.2017.00105.

[2]R. L. Andrews and G. R. Franke. The Determinants of Cigarette Consumption: A Meta-Analysis. *Journal of Public Policy and Marketing*, 10:81-100, 1991.

[3]J. J. Arnett and G. Terhanian. Adolescents' responses to cigarette advertisements: links between exposure liking, and the appeal of smoking. *Tobacco Control*, 7(2):129–133, jun 1998. doi: 10.1136/tc.7.2.129. URL https://doi.org/10.1136%2Ftc.7.2.129.

[4]Jeremiah A. Barondess. Health Through the Urban Lens. *Journal of Urban Health*, 85(5):787–801, jul 2008. doi: 10.1007/s11524-008-9300-0. URL https://doi.org/10.1007%2Fs11524-008-9300-0.

[5]Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2010. doi: 10.1109/cvpr.2010.5539992. URL https://doi.org/10.1109%2Fcvpr.2010.5539992.

[6]O B J Carter, B W Mills, and R J Donovan. The effect of retail cigarette pack displays on unplanned purchases: results from immediate postpurchase interviews. *Tobacco Control*, 18(3):218–221, mar 2009. doi: 10.1136/tc.2008.027870. URL https://doi.org/10.1136%2Ftc.2008.027870.

[7]2010 Census. 2010 Census Block Groups Polygons. URL {http://data.beta.nyc/dataset/2010-census-block-groups-polygons}dateaccessed={11July2018}.

[8]Public Health Law Center. Master Settlement Agreement. Technical report, 2018. URL http://www.publichealthlawcenter.org/topics/tobacco-control/tobacco-control-litigation/master-settlement-agreement.

[9]Xinlei Chen and Abhinav Gupta. An Implementation of Faster RCNN with Study for Region Sampling. *arXiv preprint*, arXiv:1702.02138, 2017.

[10]Wikipedia contributors. Google Street View.

[11]Ron Dekel. Human perception in computer vision. *ICLR*, 2017. URL https://arxiv.org/abs/1701.04674. Accessed on Mon, April 30, 2018.

[12]Deep Learning Object Detection. Deep Learning Object Detection Methods for Ecological Camera Trap Data. https://arxiv.org/pdf/1803.10842.pdf. URL https://arxiv.org/pdf/1803.10842.pdf. Accessed on Mon, July 23, 2018.

[13]Emile Durkheim. *Suicide : a study in sociology*. The Free Press, 1897.

[14]Friederich Engels. *The Condition of the Working Class in England*. Otto Wigand, Leipzig, 1845.

[15]N. Evans, A. Farkas, E. Gilpin, C. Berry, and J. P. Pierce. Influence of Tobacco Marketing and Exposure to Smokers on Adolescent Susceptibility to Smoking. *JNCI Journal of the National Cancer Institute*, 87(20):1538–1545, oct 1995. doi: 10.1093/jnci/87.20.1538. URL https://doi.org/10.1093%2Fjnci%2F87.20.1538.

[16]M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007, 2007. URL valuehttp://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.htmlhere.

[17]American FactFinder. American FactFinder. URL {https://factfinder.census.gov/}dataaccessed={11July2018}.

[18]FTC. Cigarette Report for 2016. *Federal Trade Commission*, Retrieved from https://www.ftc.gov/reports/federal-trade-commission-cigarette-report-2016-federal-trade-commission-smokeless-tobacco, 2018.

[19]Sandro Galea and David Vlahov. URBAN HEALTH: Evidence Challenges, and Directions. *Annual Review of Public Health*, 26(1):341–365, apr 2005. doi: 10.1146/annurev.publhealth.26.021304.144708. URL https://doi.org/10.1146%2Fannurev.publhealth.26.021304.144708.

[20]Donald Gifford. *Suing the Tobacco and Lead Pigment Industries*. University of Michigan Press, 2010. doi: 10.3998/mpub.291047. URL https://doi.org/10.3998%2Fmpub.291047.

[21]Michael Grossman. Health Benefits of Increases in Alcohol and Cigarette Taxes. *Addiction*, 84(10):1193–1204, oct 1989. doi: 10.1111/j.1360-0443.1989.tb00715.x. URL https://doi.org/10.1111%2Fj.1360-0443.1989.tb00715.x.

[22]V. Ilakkuvan et al J. Cantrell, O. Ganz. Implementation of a Multimodal Mobile System for Point-of-Sale Surveillance: Lessons Learned From Case Studies in Washington, DC, and New York City. *Public Health and Surveillance*, 1(2):e20, 2015.

[23]T. R. Kirchner, A. C. Villanti, J. Cantrell, A. Anesetti-Rothermel, O. Ganz, K. P. Conway, D. M. Vallone, and D. B. Abrams. Tobacco retail outlet advertising practices and proximity to schools parks and public housing affect Synar underage sales violations in Washington, DC. *Tobacco Control*, 24 (e1):e52–e58, feb 2014. doi: 10.1136/tobaccocontrol-2013-051239. URL https://doi.org/10.1136%2Ftobaccocontrol-2013-051239.

[24]Bonnie RJ Lynch BS. *Growing up Tobacco Free: Preventing Nicotine Addiction in Children and Youths*. Washington (DC): National Academies Press (US), 1994. URL https://www.ncbi.nlm.nih.gov/books/NBK236761/. Accessed on Mon, April 30,

2018.

[25]D. Eadie M. Stead and A. M. MacKintosh. Young people's exposure to point-of-sale tobacco products and promotions. *Public Health*, 136, 2016.

[26]Donald P. Mullally. The Fairness Doctrine: Benefits and Costs. *Public Opinion Quarterly*, 33(4):577, 1969. doi: 10.1086/267746. URL https://doi.org/10.1086%2F267746.

[27]Health Data NY. Active Tobacco Retailer Map. *https://health.data.ny.gov/Health/Active-Tobacco-Retailer-Map/88k2-euek/data*, 2018.

[28]A Brief History of CNNs in Image Segmentation.

[29]U.S. Department of Health and Human Services. The Health Consequences of Smoking - 50 Years of Progress. A Report of the Surgeon General. *Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health*, 2014.

[30]NYC OpenData. Borough Boundaries data retrieved = 23 June 2018. URL https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm/data.

[31]William H. Redmond. Effects of Sales Promotion on Smoking among U.S. Ninth Graders. *Preventive Medicine*, 28(3):243–250, mar 1999. doi: 10.1006/pmed.1998.0410. URL https://doi.org/10.1006%2Fpmed.1998.0410.

[32]Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, jun 2017. doi: 10.1109/tpami.2016.2577031. URL https://doi.org/10.1109%2Ftpami.2016.2577031.

[33]Andrew B. Seidenberg, Robert W. Caughey, Vaughan W. Rees, and Gregory N. Connolly. Storefront Cigarette Advertising Differs by Community Demographic Profile. *American Journal of Health Promotion*, 24(6):e26–e31, jul 2010. doi: 10.4278/ajhp.090618-quan-196. URL https://doi.org/10.4278%2Fajhp.090618-quan-196.

[34]Mapbox Development Team. Mapbox GL JS. https://github.com/mapbox/mapbox-gl-js, 2018.

[35]QGIS Development Team. QGIS Geographic Information System. *Open Source Geospatial Foundation Project*, http://qgis.osgeo.org, 2018.

[36]Tzutalin. LabelImg. *Git code*, 2015. URL https://github.com/tzutalin/labelImg.

[37]Image Classification using Deep Neural Networks. Image Classification using Deep Neural Networks — A beginner friendly approach using TensorFlow. https://medium.com/@tifa2up/image-classification-using-deep-neural-networks-a-beginner-friendly-approach-using-tensorflow-94b0a090ccd4. URL https://medium.com/@tifa2up/image-classification-using-deep-neural-networks-a-beginner-friendly-approach-using-tensorflow-94b0a090ccd4. Accessed on Mon, July 23, 2018.

[38]Xiaogang Wang and Xiaoou Tang. A unified framework for subspace face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1222–1228, sep 2004. doi: 10.1109/tpami.2004.57. URL https://doi.org/10.1109%2Ftpami.2004.57.

[39]Geoff West. *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies*. Penguin Press, 2017.

[40]Yong Yang and Ana V. Diez-Roux. Walking Distance by Trip Purpose and Population Subgroups. *American Journal of Preventive Medicine*, 43(1):11–19, jul 2012. doi: 10.1016/j.amepre.2012.03.015. URL https://doi.org/10.1016%2Fj.amepre.2012.03.015.

SUPPLEMENTARY MATERIAL

1 MOTIVATION EXPANDED

Since the advent of modern advertising in the 1920s, tobacco marketers have worked to convert young audiences into smokers by inundating mainstream media channels, like movies and magazines, with their persuasive messaging [S20]. As the health risks associated with smoking became increasingly evident, however, marketing restrictions were enacted on the tobacco industry to reduce their influence on cigarette purchasing behavior. In 1971, "Big Tobacco" was forced to shift their marketing efforts from broadcast to print media when a ban on cigarette advertising for television and radio went into effect in the United States [S26]. The Master Settlement Agreement of 1998 later placed further restrictions on cigarette advertising, including practices that specifically targeted individuals under the age of 18 [S8]. Since then, nearly every tobacco marketing dollar has been funneled to reaching existing and prospective tobacco customers at one of the last places available for influencing purchasing behavior: point-of-sale tobacco (POST) vendors.

In it's most recent Cigarette Report, the Federal Trade Commission found that tobacco marketing expenditure in the United States now amounts to about $1 million per hour [S18]. The rise in combined annual budgets, from $8.30 billion in 2015 to $8.71 billion in 2016, was driven primarily by payments to cigarette retailers and wholesalers to reduce the price of cigarettes that end consumer pay. As evidenced by this vast marketing spend, existing and potential tobacco consumers are vulnerable to the influence of advertising at the point-of-sale. The association between POST exposure and harmful smoking behavior, especially with respect to school-age populations, is well documented [S2; S15; S3; S31; S6]. According to the U.S. Department of Health & Human Services, nearly 90% of adult smokers had their first cigarettes before the age of 18 [S29].

2 SOCIOECONOMIC ANALYSIS

Active retail tobacco vendor density in New York City, weighted by youth population, is plotted as a function of position. We used a choropleth map to visualize the data at the Census block group level using a Jenks classification, which clusters the weighted densities by minimizing the variance within each class and maximizing variance between classes. The map demonstrates for the viewer how youth exposure to tobacco retail varies across New York City. Data cleaning, handling, and spatial analysis were conducted using QGIS [S35], and visualization was facilitated using Mapbox Studio and Mapbox GL JS [S34].

Polygons of the 2010 Census Block Groups for New York City were gathered from data.BetaNYC [S7] and joined with statistics on population under 18 years of age from United States Bureau's American FactFinder [S17] using a Table Join operation. Using the Calculate Geometry function, new fields were added to the Census Block Groups feature dataset containing latitude and longitude coordinates of the Census Block Group centroids. Circular catchment areas with a radius of 400 meters, which is generally agreed upon as the acceptable walking distance in the United States [S40], were drawn around each centroid using the Buffer operation. Point locations of active retail tobacco vendors in New York State were added to the map document from Health Data NY [S27] A Points in Polygon Analysis was then conducted in order to tabulate the number of tobacco retailers contained in each catchment area.

Youth population counts were folded into the retailer density measure using a weighted average. The weighted average is generated by summed the products of retailer counts and youth population counts for each Census Block Group, divided by the sum of youth population counts in each Census Block Group across all of New York City. A constant value 'F' generated by this calculation can be expressed by the following equation, where R represents the number of retailers, Y represents number of youth, and b the Census Block Groups:

$$F = \frac{\Sigma_b \, R_b \cdot Y_b}{\Sigma_b \, Y_b}$$

3 GOOGLE STREET VIEW IMAGE EXTRACTION

In an effort to construct a dataset of tobacco display exposure areas, we collected images from all the known vendors of tobacco products in NYC. Four images are taken for each location in order to make sure every direction is covered. The known vendors come from a CSV file provided by Health Data NY [S27]. The Google Street View API [S10] allows for coordinates or addresses to be provided and returns an image for a given location. Some cleaning of the file is necessary. Any row that does not have New York City listed in the 'City' field are removed. We then create an address field to retrieve the images. Most of the coordinates provided in the CSV were incorrect, so we needed to use the addresses to retrieve the appropriate iamges. The address is created by combining the columns of the CSV that contain the street, city, and zipcode. We then enter the field into the Google Street View API along with a heading of 0, 90, 180, or 270, which is the direction the image will be retrieved from. There were 10,694 tobacco vendors in the dataset, resulting in a total number of images of 42,776 after gathering four images for each location.

4 IMAGE MANIPULATIONS

Once labeled the images, we manipulate the images to change the color balance, contrast and add noise to the images. In the future, images could come from other sources such as street cameras or photos taken by project participants. In conducting these manipulations, our hopes were that our model would be able to handle to find features in the data other than color and

position and handle a wider variety of images. Color balance is simply the change of intensity of the colors in an image.

To change color balance, we drew three random numbers from a Gaussian distribution and multiplied the RGB values by that amount. Color contrast is the difference between colors. We took a random uniform distribution of 5 numbers from .4 to 1 and 5 more from 1 to 1/.4. We then squared all the pixels in the image by each value and normalized it by multiplying by 255 which is the maximum pixel amount and dividing by the maximum pixel value in the image. Noise was added to the images by using an approach called "salt and pepper". This is when an image has sparsely occurring pixels in the data that is either white or black. Five salt and pepper images, five color balance images, and 10 contrast images where created for each image in our dataset. This gave us a total of 2,280 images. To make sure the images weren't too similar to one another, we removed images that had a similarity score of .95 or higher. The similarity metric used is known as the structural similarity index (SSIM). SSIM compares the luminance, contrast, and structures of images. This reduced the total number to 1,201.

5 RCNNS CONTINUED

The threshold is the probability of an image to be a tobacco advertisement and whether we classifiy it as such. The initial data we have for training is from a previous NYU CUSP student research group, which we refer to as MD17. In total, there are 300 different images with tobacco advertisements. 300 images were augmented to generate 9,331 total images by manipulating image characteristics such as color balance, contrast levels, and random noise. Different attempts of selecting training set and validation set were made. For example, a random sample of 70% of the 9,331 images as training and 30% as validation has been deemed inappropriate since there was considerably high overlap between training and validation set after image augmentation . This was established with cosine similarity and hashing by drawing parameters from a random distribution. Considering the limited initial data we had, choosing a pre-trained model is crucial, and is also a typical research approach which takes advantages of previously established work. ResNet101[S32], a neural network with 101 residual layers, was used.