Automated identification of waterpipes on Instagram: an application in feature extraction using Convolutional Neural Network and classification based on Support Vector Machine classifier

Youshan Zhang, M.S., 1 Jon-Patrick Allem, Ph.D., M.A., 2 Jennifer B. Unger, Ph.D. 2
Tess Boley Cruz, Ph.D. 2

1 Department of Computer Science, Lehigh University, Bethlehem, PA, USA,
2 Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

Address for correspondence: Jon-Patrick Allem, 2001 N. Soto Street, 3rd Floor Mail, Los Angeles, CA 90032. Phone: (858) 603-0812, E-Mail: allem@usc.edu

Word count: 2643 Tables: 0 Figures: 10

Keywords: Image classification, Feature extraction, support vector machine, convolution neural network, Instagram, social media.

Abstract

1
Background: Instagram (a popular image-based social media app) with millions of posts each day can be used to inform public health interventions and policies but current research relying on image-based data often relies on hand coding of images which is time consuming, costly, and may be subject to researcher bias. What is more, current best practices in automated image classification (e.g., support vector machine (SVM), Backpropagation (BP), and artificial neural network) are limited in their capacity to accurately distinguish between objects within images. Objective: The goal of this study is to demonstrate how convolution neural network (CNN) can be used to extract unique features within an image and how SVM can then be used to classify the image. Methods: Images of waterpipes or hookah (an emerging tobacco product possessing similar harms to that of cigarettes) were collected from Instagram and used in analyses (n=840). CNN was used to extract unique features from images identified to contain waterpipes. SVM classifier was built to distinguish between images with and without waterpipes. Methods for image classification were then compared to show how CNN + SVM classifier could improve accuracy. Results: As the number of the validated training images increased, the total number of extracted features increased. Additionally, as the number of features learned by the SVM classifier increased, the average level of accuracy increased. Overall, 99.5% of the 420 images classified were correctly identified as either hookah or non-hookah images. This level of accuracy was shown to be some improvement over earlier methods that used SVM, CNN or Bag of Features (BOF) alone. Conclusions: CNN extracts more features of the images allowing the SVM classifier to be better
informed which results in higher accuracy compared with methods that extract fewer features. Future research can use this method to reduce computational time in identifying millions of images. By taking images of waterpipes from Instagram, we place our methods in a context that can be utilized to later inform health researchers. Automated and scalable approaches to identifying waterpipes in images from Instagram may facilitate reaching out to Instagram users who may need tobacco-related education to reduce misconceptions about hookah use.

1 Introduction

Instagram (a popular image-based social media app) with millions of posts each day can be used to inform public health interventions and policies but current research relying on image-based data often relies on hand coding of images which is time consuming, costly, and may be subject to researcher bias. While automated image classification is useful for large-scale image classification (e.g., processing and assigning labels to millions of images), current best practices in automated image classification are limited in their capacity to accurately distinguish between objects within images. Automated image classification has been used in supervised, unsupervised, and hybrid approaches in classifying data[^1][^2][^3]. Supervised methods can be broken down to stages of training and testing. The training stage consists of training a classifier by images and its labels e.g., describing image content, such as person, dog, elephant, etc. The testing stage predicts the labels of the test images (in a new set of images) by the trained classifier. Unsupervised methods classify images
based on differences in spectral values (representing information from the images),
while hybrid methods combine both supervised and unsupervised methods.

Prior research has focused on ways to overcome the methodological
challenges of automated image classification such as low accuracy. For example,
Pettronin and colleagues improved the Fisher Kernel approach to extend the bag-of-
visual-words (BOV) for large-scale image classification using internet images from
ImageNet and Flickr, which increased precision from 47.9% to 58.3%, but did not
improve accuracy\cite{4}. Kesari and colleagues used the Backpropagation neural network
approach to classify large images with good accuracy (97.02%), but this approach
could not identify multiple categories of an image\cite{5}. To reduce the time and spatial
complexity of images, Simonyan and colleagues proposed two visualization
techniques using deep Convolutional Networks (ConvNets) to classify artificial
images. They combined understandable visualizations of ConvNets, that first
maximized the scores of images within different classes, with gradient-based
ConvNets visualization, that generated the saliency map of every image
(corresponding to one class), in order to use a deconvolution network to segment
objects in the images\cite{6}.

These earlier approaches have moved automated image classification forward,
however there are still a number of significant limitations to overcome\cite{7,8,9}. For
example, the large number of images that need to be extracted to train a model require
great computational power. Additionally, the BOV method cannot localize the objects
within an image and cannot use visual word positions (e.g., if a cup was in an image, BOV could not find its position)\textsuperscript{[10][11]}. SVM has limited generalization ability and it cannot solve a nonlinear problem. CNN, on the other hand, can improve the generalization of the algorithm, and can solve nonlinear problems\textsuperscript{[12][13][14]}. While CNN has high accuracy, to get better results, the parameters should be fine-tuned (e.g., input image size, patch size, and the number of convolution layers), and network performance is hard to optimize\textsuperscript{[15][16]}.

The purpose of this study is to determine whether combining CNN and SVM can achieve higher accuracy in image classification compared to CNN or SVM alone. To achieve this, data from Instagram containing images of waterpipes also known as hookah (an emerging tobacco product possessing similar harms to that of cigarettes) were examined. By identifying waterpipes in images from Instagram we can identify Instagram users who may need tobacco-related education to reduce misconceptions about hookah use. Additionally, by taking data from Instagram, we place our methods in a context that can be utilized to later inform researchers in the health domain who wish to analyze social media to inform education and policy\textsuperscript{[17][18][19][20][21][22][23][24]}.

2 Methods

2.1 Data acquisition

Data used in this study comprised posts on Instagram between February 19, 2016 and May 19, 2016 in the United States that included the hashtag #hookah. Details on data collection are described elsewhere\textsuperscript{[23]}. Matlab was used to classify
images into two categories (images containing a waterpipe (hookah) and those without).

2.2 Convolution neural network

Image features comprised of 25 layers were extracted using AlexNet (a well-trained convolution neural network software). Among these 25 layers, there is an input and output layer, seven Rectified Linear Units (ReLU) layers, two normalization layers, three pooling layers, two dropout layers, one softmax layer and eight learnable weights layers which contain five convolutional layers and three fully connected layers (refer to [25] for more details). The input layer is comprised of 227×227-pixel images. The ReLU layer reduces the number of epochs to achieve the training error rate greater than traditional tanh units. The normalization layer increases generalization and reduces the error rate. The pooling layers summarize the outputs of adjacent pooling units \[26\][26]. The dropout layer efficiently decreases the test errors\[27\], and both dropout layer and the softmax layer reduce the over-fitting phenomenon, while the output layer is the categories of images. To extract the features, we fine-tuned the network that removed the last four layers of the original 25 layers, since all of the layers are not suitable for extracting the features. The layers at the beginning of the network can only detect the edges of the images, so we used the results of the fully connected layers to extract features.

2.3 Support Vector Machine

6
SVM is a supervised learning model with algorithms that analyze data for classification, that has been used to predict the categories of objects in images\cite{28}\cite{29}. Our proposed method goes beyond earlier research as the input (feature vectors) is based on the results of the convolution neural network, which can boost accuracy and save time (see \cite{7} for more details about SVM). We trained the SVM classifier using the feature vectors, and then based on the classifier, the categories of images were predicted.

### 2.4 Analytical approach

First, we classified images into two categories: hookah and non-hookah images and labeled accordingly. Figure 1 shows the classification scheme e.g., input images dimension is 227×227×3 pixels. The output of CNN is the 4096×1×1 features maps of two image classes, which are used to train the SVM classifier, then the classifier is used to predict the categories of test images. The hookah images contain a waterpipe and the non-hookah images do not contain a waterpipe (Figure 2). Next, we divided image sets into training and test images; the training images were used to extract and learn the features (N=420, randomly selected), while the test images were used to calculate the accuracy of the method (N=420, randomly selected). To extract features of the images, the dimension of the input images was made uniform e.g., the image size was 227×227. We loaded the pre-trained CNN by utilizing AlexNet\cite{26}, which has been trained by more than one million images. As discussed above, the AlexNet has been fine-tuned in our method e.g., we removed the last four layers of
the AlexNet and used the data of the final fully connected layer. Based on the data of the last fully connected layer, we computed the feature of the training and test images based on CNN. Then, the class labels were extracted from the training and test images sets.

To optimize the SVM classifier, we automatically optimized hyperparameters (such as learning rate, the number of layers in CNN, mini-batch size) of the waterpipe features vector, and based on the optimized results, we arrived at an optimized SVM classifier (Figure 1)[30][31]. We then assessed the performance of the SVM classifier by using the test images and increased the number of images to improve accuracy (the number of images is increased from 42 to 420) (Figures 2 & 3). Based on the trained classifier, we predicted the classes of new images.

Figure 1 shows the scheme of our method. Input images dimension is 227×227×3. The output of CNN was 4096×1×1 features maps of two image classes. These features were trained by the SVM classifier, and the trained classifier was later used to predict the categories of test images.
3 Results:

Results demonstrated that hookah features could be extracted by CNN, with image categories classified by the SVM, maintaining a high level of accuracy (highest 99.5%). Figure 4 shows the features that were extracted from the first convolution.
layer; this layer can only detect the edges and blobs, while more features were extracted from the remaining convolution layers. Figure 5 shows the feature vectors of the 420 training images, with range [-20 20]; the majority of feature vectors are located between -10 and 10. Figure 6 is the histogram of the feature vectors. The maximum number of features was between -2 and 2. This interval [-2 2] reflected the most important features of the hookah images. Figure 7 shows the relationship between the function evaluations and the minimum objective. The minimum objective and the estimated minimum objective are similar, however there are differences across certain function evaluations. The maximum number of error is less than 0.001, and this error can be accepted. Based on the optimized SVM classifier, we evaluated the performance of our method by the test images.

Figure 4 shows the extracted features of the first layer using the CNN. The original hookah image is on the left. The feature images (right) contains a montage of 96 images, which can reflect the processing of extracting features.
Figure 5 shows the features of the total image sets (420 images). The x-axis is the image features vector with 4096 total feature vectors. The y-axis is the range of the features with the range between -20 and 20.

Figure 6 shows the histogram of the features. The interval of [-2 2] contains the maximum number of features.
Figure 7 shows the relationship between the function evaluations and the minimum objective. When the function evaluation was 25, the error between minimum objective, and the estimated minimum objective was highest.

3.1 The test image classification results

Figure 8 shows the relationship between the percentage of validated images (e.g., the training images, excluding the test images) and the average level of accuracy of the method. As the number of the validated training images increased, the total number of extracted features increased. Additionally, as the number of features learned by the SVM classifier increased, the average level of accuracy increased. The number of validated images was equal to the percentage × the number of the training set, e.g., if 10%, then the validated images=10% × 420=42.
Figure 8 shows the line graph of the accuracy of the classifier with different number of validated images. From the chart, with the increasing of the percentage of validated images, some of the accuracies boosted significantly, for example, from the 80%, the accuracies increase faster than previous percentages, demonstrating that more training images are beneficial to predicting results.

Overall, 99.5% of the 420 images classified were correctly identified as either hookah or non-hookah images (Figure 9). In the first row, no non-hookah images were incorrectly classified as hookah images. In the second row, there were two hookah images incorrectly classified as non-hookah images, representing 3.4% of all the data. In the first row, 100% of hookah images were correctly classified. In the second row, 99.1% were correctly classified as hookah images. In first column, 99% were correctly classified as hookah images and 0.1% were correctly classified as non-hookah images. In the second column, out of 210 non-hookah images, 100% were correctly
classified as non-hookah images and no images were incorrectly classified as hookah images.

Figure 9 shows the confusion matrix of the test images (column 1 and 2 are the hookah and non-hookah categories, respectively, column 3 is the accuracy of classified results). The first two green squares show the number of the test images and the percentage of the correct image classifications. For example, there were 208 images correctly classified as hookah, and this number accounted for 49.5% of all test images (420). Similarly, 210 images were properly classified as non-hookah, and this accounted for 50% of all test images.

3.2 Compare with other methods

We compared our method with CNN, SVM, BOF (Bag-of-Features, see [32][33] for more details). Figure 10 shows how the accuracy of various models can be improved as a function of the size of the training data. Our method had the highest accuracy (99.5%), compared to the other models.
4 Discussion

This study showed that the use of CNN to extract features and SVM to classify images results in higher accuracy in automated image classification compared to CNN or SVM alone. Compared to earlier work using CNN, SVM, and BOV, our method improves accuracy when the number of training images is increased with accuracy approaching 100% (99.5%). This illustrates that our method is suitable for distinct images like waterpipes.

By identifying waterpipes in images from Instagram we can identify Instagram users who may need tobacco-related education to reduce misconceptions about hookah use. Instagram may be used to bolster the reach and delivery of health information that communicates the risk of hookah use\textsuperscript{[34][35][36][37]}. Data from social media (e.g., Instagram, Twitter, and Facebook) can provide public health and policy
experts with useful information on what people care about and what occupies their time.\textsuperscript{17}\textsuperscript{21} Earlier research used Instagram images to capture and describe the context in which individuals use, and are marketed tobacco products.\textsuperscript{22}\textsuperscript{23} For example, analysis of Instagram data on electronic cigarettes demonstrated that a majority of images were either individuals showing their favorite combinations of products (e.g., type of electronic cigarette device and flavored juice), or people performing tricks with the products (e.g., blowing a large aerosol cloud in competition with others),\textsuperscript{24} demonstrating how and why people use this tobacco product. In a separate study, researchers found that Instagram was used by restaurants, bars, and nightclubs to cross-promote hookah and alcohol specials to appeal to potential customers.\textsuperscript{23}

These earlier studies provided timely information from a novel data source, however their methods relied upon hand coding of images—a process requiring time, expertise and sample sizes small enough to reasonably code by hand, ultimately limiting the scope of the work. The findings from the current study showed how automated image classification can be used to overcome such limitations. Additionally, the methods from the current study can help researchers in tobacco control identify what proportion of viewers on a social media site are interested in certain products such methods may be crucial to document the changing tobacco landscape.

The findings from this study should be considered with several limitations in mind, including the long computational time (it takes hours to train the network) during the training stage and the iterative process that is required to arrive at a well-
trained classifier. To eliminate the problem of overfitting, we used ReLU, softmax, dropout layers in the CNN, and utilized several different training datasets (the number of datasets is different which increased from 42 to 420, in figure 10). In the future, researchers should try to enlarge the sets of training images to extract specific features of an image, which may achieve higher accuracy with less computation power. The methods developed in this study were only applied in the context of images from Instagram that focused on waterpipes and should be applied in more categories and other contexts in the future.

5 Conclusion

Findings demonstrated that by combining CNN and SVM to classify images resulted in 99.5% accuracy in image classification, which is an improvement over earlier method using SVM, CNN or BOV alone. CNN extracts more features of the images allowing the SVM classifier to be better informed which results in higher accuracy compared with methods that extract fewer features. Future research can use our method to reduce computational time in identifying objects in images.

Acknowledgments

Research reported in this publication was supported by Grant # P50CA180905 from the National Cancer Institute and the FDA Center for Tobacco Products (CTP). The NIH or FDA had no role in study design, collection, analysis, and interpretation of data, writing the report, and the decision to submit the report for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or FDA.
Competing Interests

None declared

Author Contributions
YZ and JPA conceived of the study and analyzed the data. YZ and JPA drafted the initial manuscript. JBU and TBC received funding for the study. JBU and TBC revised the manuscript for important intellectual content. All authors have approved the final manuscript.

Reference:


