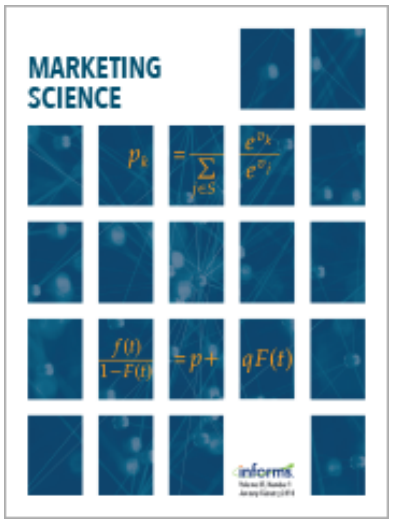


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Visual Listening In: Extracting Brand Image Portrayed on Social Media

Liu Liu, Daria Dzyabura, Natalie Mizik

To cite this article:
Liu Liu, Daria Dzyabura, Natalie Mizik (2020) Visual Listening In: Extracting Brand Image Portrayed on Social Media. Marketing Science 39(4):669-686. https://doi.org/10.1287/mksc.2020.1226

http://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Vol. 39, No. 4, July–August 2020, pp. 669–686
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Visual Listening In: Extracting Brand Image Portrayed on Social Media

Liu Liu,[a] Daria Dzyabura,[b] Natalie Mizik[c]

[a] Leeds School of Business, University of Colorado, Boulder, Colorado 80309; [b] New Economics School, Moscow 121353, Russia; [c] Foster School of Business, University of Washington, Seattle, Washington 98195

Contact: liu.liu-1@colorado.edu, https://orcid.org/0000-0002-2227-2336 (LL); ddzyabura@nes.ru, https://orcid.org/0000-0003-0729-2379 (DD); nmizik@uw.edu, https://orcid.org/0000-0002-7220-290X (NM)

Received: May 8, 2017
Revised: September 30, 2018; August 24, 2019; November 19, 2019
Accepted: November 24, 2019
Published Online in Articles in Advance: July 10, 2020

https://doi.org/10.1287/mksc.2020.1226

Abstract. Images are close to surpassing text as the medium of choice for online conversations. They convey rich information about the consumption experience, attitudes, and feelings of the user. In this paper, we propose a "visual listening in" approach (i.e., mining visual content posted by users) to measure how brands are portrayed on social media. We develop BrandImageNet, a multi-label deep convolutional neural network model, to predict the presence of perceptual brand attributes in the images consumers post online. We validate BrandImageNet model performance using human judges and find a high degree of agreement between our model and human evaluations of images. We apply the BrandImageNet model to brand-related images posted on social media to extract brand portrayal based on model predictions for 56 national brands in the apparel and beverages categories. We find a strong link between brand portrayal in consumer-created images and consumer brand perceptions collected through traditional survey tools. Firms can use the BrandImageNet model to automatically monitor their brand portrayal in real time and better understand consumer brand perceptions and attitudes toward their and competitors' brands.

History: K. Sudhir served as the senior editor and Olivier Toubia served as associate editor for this article.
Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mksc.2020.1226.

Keywords: social media • visual marketing • brand perceptions • computer vision • machine learning • deep learning • transfer learning • big data

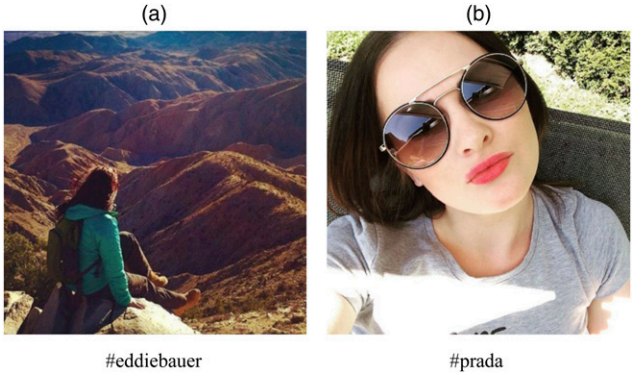
1. Introduction

A profound shift has recently occurred in how people consume information and in the origins of the information itself. Corporations and traditional media have lost their monopoly on information creation and communication channels. Much of the brand-related content is now created and spread through social media platforms (e.g., Twitter and Instagram), user discussion forums, blogs, and so on. Firms now operate in an environment in which customers are actively participating in sharing their brand perceptions and experiences on social media and co-creating brands and brand identities. With this broader, more egalitarian distribution model, monitoring how a brand is portrayed on social media is essential for effective brand management.

In this paper, we focus on consumer-created visual content. With the proliferation of camera phones, generous data plans, and image-based social media platforms, photo taking and sharing have become an important part of consumers' social lives (Diehl et al. 2016). Images are becoming an increasingly prevalent form of online conversations. Instagram has 1 billion monthly active users, and these users share an average of 95 million photos daily.[1] In these shared photos, consumers often tag brands. For example, a search of #*nike* on Instagram returns over 92 million photos tagged with the brand name Nike.[2] Through these images, consumers communicate about brands with each other.

Consumer-created brand images on social media are different from product images on retailers' websites. In the images posted on social media, consumers often link brands with usage context, feelings, and consumption experiences. For example, consider the images in panels (a) and (b) in Figure 1. The first image is hashtagged with *eddiebauer* and the second with *prada*. Panel (a) shows a person wearing Eddie Bauer and resting on a rock, facing mountains in the distance. Panel (b) shows the user taking a picture of herself wearing Prada sunglasses and bright red lipstick. The two images differ in terms of their content (mountainous landscape versus head shot) and their visual properties (color palate, contrast, etc.). They also differ in terms of the mood and experience consumers may associate with the two brands. The images created

Figure 1. Sample Images from Instagram Hashtagged with Brand Names



and shared by consumers offer a new avenue to study brand associations and consumers' brand perceptions. Firms need to monitor how consumers are portraying their brands on social media, or risk missing a large part of consumer brand conversations.

Academic researchers and practitioners have developed a set of tools for monitoring social media and listening in on consumers' online brand conversations. The focus so far has been on textual content. However, given that images are on their way to surpassing text as the medium of choice for online conversations, and the rich information about the consumption experience and feelings conveyed through images, monitoring visual content is important to get a more complete understanding of brand-related online conversations.

We introduce a "visual listening in" approach for monitoring visual brand content created and shared by consumers on social media. We develop and validate a model, BrandImageNet, to allow firms to monitor their brand portrayal on social media and evaluate it relative to competitors' and to the firm's preferred brand positioning. The BrandImageNet model maps images to specific perceptual attributes (e.g., glamorous, rugged) and allows us to measure how brands are portrayed along these attributes. For example, is Prada portrayed as a glamorous brand and is it portrayed as more glamorous than Eddie Bauer? We focus on identifying perceptual brand attributes from images, which is different from identifying functional attributes of the product (Gardner and Levy 1955, Park et al. 1986, Keller and Lehmann 2006). Positioning brands on intangible attributes is at the core of many marketing strategies because it allows firms to differentiate themselves from one another in categories with functionally similar products. In categories such as beverages and apparel (our focus categories in this paper) consisting of brands offering products with similar functionality, the feelings, perceptions, and attitudes consumers have toward the brand often make a big difference. The intangible brand attributes allow consumers to choose the brand that feels most appropriate to them and to express themselves through that brand.

We develop our BrandImageNet model using a deep-learning framework to predict whether a particular perceptual attribute is expressed in a given image. We focus on apparel and beverages, two product categories for which consumers frequently post and tag photos online, and explore four brand attributes that are highly relevant for brands in these two categories: glamorous, rugged, healthy, and fun. We selected these four attributes from Young and Rubicam's (Y&R) BAV (BrandAsset Valuator) consumer-brand-perception survey (Mizik and Jacobson 2008) to allow subsequent comparisons of our model-based findings with external survey-based findings. We selected these specific four brand attributes because they are meaningful differentiators in the apparel and beverage sectors (they have a higher range of scores across brands in these sectors according to the BAV survey).

One challenge with studying images is that image data are unstructured. To draw any inferences, a first and important step is to create a meaningful representation of the images. Typically, doing so would require creating a feature vector summarizing the image data. An emerging area of machine learning, called deep learning, uses "deep" neural networks to simultaneously learn both the model parameters and the feature representation needed for prediction tasks. It works with raw image data directly and automatically learns the representation of data with multiple levels of abstraction (LeCun et al. 2015). Our BrandImageNet is a multi-label convolutional neural network (ConvNet), fine-tuned from the Berkeley Vision and Learning Center (BVLC) Reference CaffeNet model on a set of Flickr images we collected for brand attribute identification. It achieves a high average out-of-sample accuracy of 90% and an average AUC (area under the receiver operating characteristic curve) of 0.85.

We apply our BrandImageNet model to consumer-created and firm-created brand images shared on Instagram. We collect a set of images hashtagged with brand names for a total of 56 brands in two product categories: apparel and beverages. Using the BrandImageNet model, we compute the probability of attribute presence in each brand image. Then, for each brand and each of the brand attributes we study, we compute an image-based brand image (IBBI) metric as the average probability of attribute presence across all consumer-generated images of this brand. A higher score indicates more images portray the brand as having a given brand attribute. We undertake

validation analyses and show that our model predictions have a high degree of agreement with human judges evaluating images and that the IBBI metric reflects consumer brand perceptions. Specifically, we find strong positive correlation between IBBI and brand perception scores from survey data (Y&R BAV survey and our online brand perceptions survey of active Instagram users). These results show that consumers' brand portrayal on social media reflects and is indicative of consumer brand perceptions. Consumer-created brand images contain valuable brand-related information, and our model is able to pick it up.

Our study makes both a methodological and substantive contribution to the marketing literature. We propose a new approach to measure consumers' brand perceptions from the voluble stream of public consumer-generated brand imagery. We validate the model and demonstrate its performance through tests against human judgments and surveys of consumer brand perceptions. This study is one of the first to use a deep-learning approach in marketing. Deep-learning methods are becoming more popular and have significantly improved performance of many computer vision tasks (e.g., object detection); however, to the best of our knowledge, no extant work on predicting perceptual brand attributes from images exists. This study is also one of the first in marketing to systematically study consumer-created imagery at a large scale.

Substantively, our study establishes a link between consumer-created images and brand perceptions. Past research has examined how firms use imagery, such as logos and image advertisements, to position brands and how visual stimuli influence consumer behavior (Wedel and Pieters 2007). We show that visual content generated by consumers reflects brand perceptions. We present evidence showing that consumer-created brand images contain valuable brand information. Our approach goes beyond the practice of simply counting the frequency of brand logo presence in the images. Many firms have started to look into consumer-created visual content on social media, but most efforts are focused on brand logo detection and tracking brand mentions. We show that deeper insights can be extracted from consumer-created images. Firms can use our BrandImageNet model to monitor their brand portrayal and examine the effectiveness of their positioning strategies. Because our BrandImageNet model can be easily extended to include other attributes and is highly scalable, brands can use it to track relevant brand perceptions automatically and in real time from images posted on social media.

The rest of the paper is organized as follows. First, we discuss related literature in both marketing and computer science. Then, we introduce our BrandImageNet model and discuss its architecture, training process, and the data we use to train it. Next, we apply our model to brand-related images posted on Instagram to measure consumers' brand portrayal and demonstrate its performance using human judges and survey instruments. We conclude with a discussion of limitations and directions for future research.

2. Related Work

This paper is related to four streams of research in marketing and computer science: user-generated content (UGC), visual marketing, computer vision, and deep-learning applications in marketing. In this section, we review how we contribute to all four.

Mining UGC for online marketing intelligence is becoming increasingly popular in both marketing and computer science research, as well as among practitioners. In the marketing literature, Netzer et al. (2012) propose an approach to help firms understand market structure and monitor the topics discussed in relation to their brand by mining brand associations from consumer-generated content on forums. Culotta and Cutler (2016) measure consumers' brand perceptions by mining the brands' social connections through social networks on Twitter. Several papers investigate the relationship between consumer-generated online content (e.g., product reviews and ratings) and sales (e.g., Chevalier and Mayzlin 2006, Liu 2006, Archak et al. 2011). The focus so far has been on text. Only recently have marketing researchers begun to consider visual content (Zhang and Luo 2019, Zhang et al. 2018, Pavlov and Mizik 2019). We propose leveraging the ever-increasing flow of consumer-generated images available online for marketing research. We contribute to this literature by advancing a new approach to measuring consumers' brand perceptions, using images that consumers post on social media.

Images constitute a large part of consumer online conversations. With the exclusive focus on text content, a significant portion of online brand conversations is not being "heard." Indeed, many consumers, particularly millennials, prefer visual-based online communication, such as Snapchat or Instagram, to text. Further, for some consumer categories (e.g., beverages, apparel), visual content is more easily available than text content. Consider, for example, the beverages category. Few people would write a review of bottled water or orange juice, but many people tag Fiji Water, Tropicana, and Minute Maid on Instagram. Finally, text mining of UGC tends to recover mostly functional product attributes. Netzer et al. (2012), for example, recover topics in a car discussion forum related to functional attributes such as powertrain warranty, good mileage, suspension noise, and emergency brakes. Similarly, Timoshenko and Hauser (2019) recover functional needs for oral care

products, such as "an oral-care product that does not affect my sense of taste" and "an oral care product that is quiet." Tirunillai and Tellis (2014) recover attributes such as ease of use, reliability, portability, safety, and comfort. We demonstrate that recovering nonfunctional dimensions of brand image from visual content is possible. Images may convey information about the usage situation, the setting, mood, and feelings associated with the product. Indeed, although a consumer is unlikely to write that he/she is feeling rugged wearing Levi's jeans, a photograph tagged with Levi's may convey ruggedness.

Our work also contributes to the literature on visual marketing, which has studied how consumers perceive different visual stimuli. Firms have been using visual stimuli to shape consumer brand perceptions through brand logos, advertisements, retail-store decorations, and, more recently, social media. Raghubir and Greenleaf (2006) examine how geometry, specifically, the ratio of the sides of a rectangular product or package, affect consumers' purchase intentions and preferences. Zhang et al. (2009) model the effects of feature-ad design characteristics (e.g., ads' surface size and number of supporting colors) on sales of the feature product and investigate the mediating role of attention. Wedel and Pieters (2014) investigate how color schemes affect rapid gist perception of an ad, product category, and brand recognition under brief and blurred exposure conditions. They find that the color of the central diagnostic object in the ad plays a key role in protecting gist perception. Xiao and Ding (2014) find that faces in print ads affect ad effectiveness. Our analysis contributes to this literature by demonstrating that visual content can convey perceptual brand attributes. We find that brand images created and shared by consumers reflect their brand perceptions.

Mining visual UGC requires different tools than mining text UGC. We draw on image-classification methods developed in the computer vision and deep learning literature. Until recently, much of this literature has focused on object identification and recognition. The field of computer vision was originally developed for engineering applications, such as automatic product inspection, autonomous vehicle navigation, surveillance, and medical image processing to aid diagnostics. More recently, researchers have begun developing tools for predicting image aesthetics, image style or genre, image interestingness (e.g., Jiang et al. 2013, Karayev et al. 2014), and visual sentiment (Giannakopoulos et al. 2015). Our work is closely related to this stream of research on predicting abstract constructs from images. We study a new problem: measuring perceptual brand attributes from images. We solve the problem using deep convolutional neural network architecture, which is a type of deep neural network (LeCun et al. 2015). Deep-learning methods are well suited for marketing problems, because of their ability to handle unstructured data, scalability, and superior predictive performance. As marketing researchers begin to adopt machine-learning methods (e.g., Zhang and Luo 2019, Gabel et al. 2019, Liu et al. 2019), ours is among the first applications of deep-learning methods to marketing problems.

3. Machine-Learning-Based Identification of Brand Attributes in Images

To measure how brands are portrayed along particular attributes on social media, we need a scalable method to determine whether a particular brand attribute is expressed in a given image: Does the image look rugged? Fun? Glamorous? Healthy? In fact, an image can depict multiple brand attributes. An image can, for example, look both rugged and fun. This attribute identification problem can be formulated as a multi-label image-classification task—that is, classifying images into multiple brand attributes—where each image may depict more than one attribute simultaneously. To solve this problem, we need (1) an annotated data set of images labeled with brand attributes to use for training, and (2) an algorithm that learns the underlying function mapping images to brand attributes. In this section, we describe how we created a data set of labeled images and developed our multi-label convolutional neural network model, BrandImageNet, for identifying perceptual attributes in an image. Later, we apply our brand attribute identification method to brands in the apparel and beverages categories.

3.1. Data

Training a multi-label image-classification model of brand attributes requires an annotated data set consisting of images labeled with respect to whether they express each attribute. Unlike object-detection tasks, which are typically trained on large public data sets of labeled images, no existing data set is annotated with perceptual brand attributes. Thus, we created one.

We gathered an annotated set of images from Flickr, an online photo-sharing website. Users share photos on Flickr, label their uploaded photos with titles and descriptions, and provide free-form tags. Flickr has been used as a data source in previous visual and social network research (e.g., Dhar et al. 2011, McAuley and Leskovec 2012, Zhang et al. 2012). Flickr lends itself well to gathering an annotated set, because unlike some other social media platforms (e.g., Instagram), it provides a search engine that returns the most relevant images for a keyword. The search is based on text labels provided by users, image content, and clickstream data (Stadlen 2015). An image ranked at the top for a particular query has

often been validated by tens of thousands of users who clicked on the image, reflecting a large population consensus regarding a strong association between the image and the query term. Because recognizing brand attributes from images is a more subjective task than recognizing objects from images, using annotations based on a large population consensus is important to decrease noise in the labels data.[3]

For each attribute, we queried the attribute term on the Flickr search engine and collected about 2,000 images in the top search results. We use these images as our positive instances (i.e., images that express the brand attribute). We also needed to collect negative instances for each attribute consisting of images that do not express that brand attribute (images that are not rugged, not glamorous, etc.). We need both the positive and negative images for the algorithm to learn how to separate them. To gather negative images, we queried the antonyms of our attributes in the Flickr search engine (drab for glamorous, gentle for rugged, unhealthy for healthy, dull for fun) and again collected about 2,000 images in the top search results. To capture a wider spectrum of the negative cases, we also used the images collected for the other attributes and their antonyms as negative examples of a given brand attribute, if they were not already included in the positive image set of that brand attribute. For example, for the attribute healthy, we used "healthy" as the query to collect positive instances. For negative instances, we used images that were returned for the query "unhealthy" and images that were returned for all other terms (glamorous, rugged, fun, drab, gentle, dull) but were not in the "healthy" set. The entire annotated set has a total of 16,360 images. Each image is associated with four labels for the four brand attributes that we study. Figure 2 shows one sample image returned by Flickr for each queried brand attribute and its antonym.

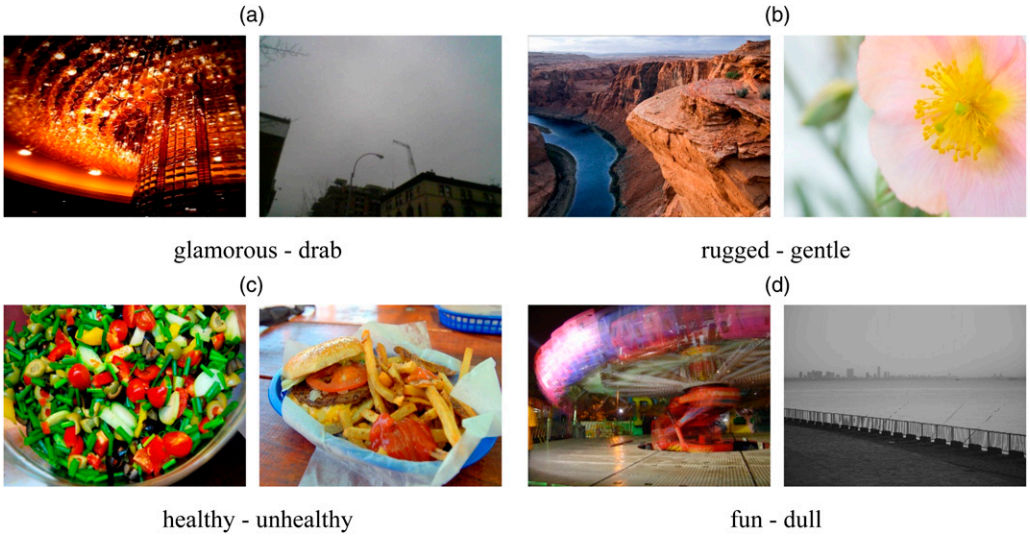
3.2. BrandImageNet Model

Next, we built and trained a multi-label image-classification model that maps images to perceptual brand attributes.

Image classification is challenging because the raw input data are difficult to capture quantitatively—more difficult than other unstructured data, such as text. Text consists of words, which can be grouped by topic or classified into positive or negative valence. Images, on the other hand, are composed of pixels, and a single pixel in isolation does not lend itself to meaningful interpretation. Therefore, any modeling of image data requires creation of a meaningful representation of the images. It requires creating a set of features, or predictor variables, through some complex transformations of the raw pixel data. The process of defining features is a critical step in achieving good predictive performance. However, manual feature engineering is hard, time-consuming, and requires great domain knowledge, and there is no systematic approach that would guarantee finding features that best characterize the underlying problem. Thus, we adopt a deep-learning approach (LeCun et al. 2015).

Deep learning does not require human input to manually engineer image features to represent raw image data. It is a representation learning method, meaning it automatically optimizes the feature-extraction step for the prediction problem. It learns the features and trains the classifier over these features simultaneously, by minimizing classification error. A deep neural network does so by applying several "layers" of simple nonlinear transformations of the raw data. The result, or output, of each transformation serves as the

Figure 2. Sample Images of Brand Attributes and Their Antonyms in Flickr Data Set



input into the next transformation layer. Through multiple layers of such transformations, the network extracts higher- and higher-level representations of the data, allowing the final layer to easily classify the data. For example, working with image data, lower layers of a deep-learning model may extract edges and textures, whereas higher layers detect motifs, object parts, and complete objects (Goodfellow et al. 2016). The final layer maps the resulting features onto the target variables with a classification function.

3.2.1. A Multi-Label Convolutional Neural Network. We set up a multi-label convolutional neural network to identify multiple brand attributes from images. ConvNet lends itself well to this problem. It is a state-of-the-art method most commonly used for analyzing visual imagery. It has demonstrated an outstanding performance in the computer vision field for tasks such as object detection and classification, as well as in other fields (e.g., medical image analysis). ConvNets vary in number of layers and each layer's type, parameters, and hyperparameters. A typical ConvNet involves four types of layers: convolutional, activation, pooling, and fully connected layers.

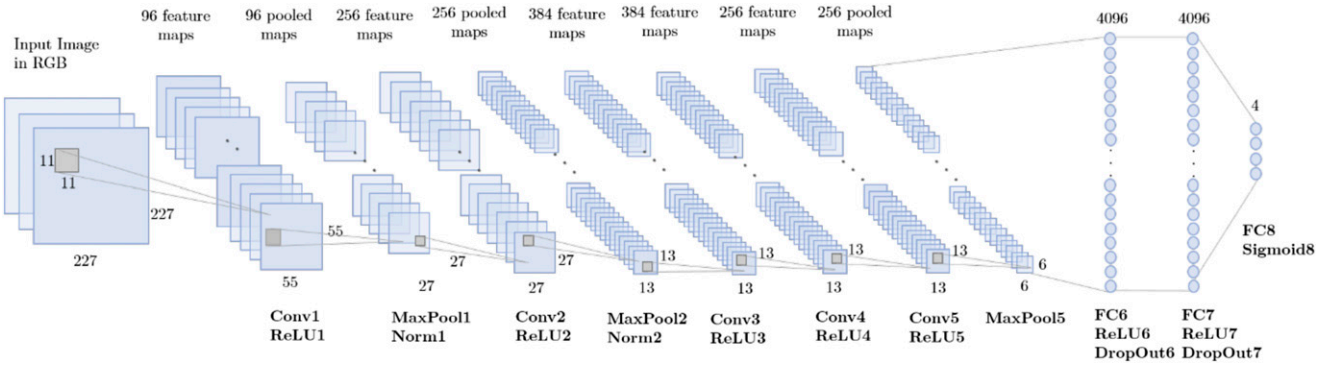
The architecture of BrandImageNet, our multi-label ConvNet, is depicted in Figure 3. The input into the network is an image, which is 227 pixels wide and 227 pixels tall. Each pixel is represented in the RGB system by three integers ranging from 0 (black) to 255 (white), reflecting the contribution of each of the three (red, green, blue) color channels.[4] The image is represented by a three-dimensional matrix and a total of 154,587 ($227 \times 227 \times 3$) numbers. The output of the network is an array of k numbers, with each number representing the probability of the image expressing a particular brand attribute.[5] In our case, $k = 4$, and the output contains the probabilities of the image being glamorous, rugged, healthy, and fun. The brand-attribute labels are not mutually exclusive (i.e., an image can be classified as positive on multiple attributes).

Our multi-label ConvNet model maps the input, an image (the original matrix of pixel values), to the output, the probability scores, through eight trainable layers (five convolutional layers and three fully connected layers) and 15 nontrainable layers (ReLU activation, max pooling, normalization, dropout, and sigmoid layers). A trainable layer contains model parameters to be learned and hyperparameters to be set. The nontrainable layers only contain hyperparameters to be set. We describe each type of layer in detail.

The architecture of our network is based on and is modified from the Berkeley Vision and Learning Center (BVLC) Reference CaffeNet model, which is an implementation, with minor variations, of the widely used AlexNet (Krizhevsky et al. 2012).[6] Both AlexNet and BVLC Reference CaffeNet models are designed to classify a given image as one of 1,000 different objects (e.g., a tree, a ship, etc.) and have been trained on the ImageNet data set of 1.2 million images (Deng et al. 2009). One key difference between our model and the underlying BVLC Reference CaffeNet model is in the final layer of the network: ours is a four-sigmoid function rather than a 1,000-way softmax function, because ours is a multi-label classification problem, whereas the BVLC Reference CaffeNet is a 1,000-class object classification model in which each image can belong to one class only.

3.2.1.1. Convolutional Layer. Convolutional layers form the core of ConvNet architecture. Each convolutional layer extracts spatial local features from input data (an image for the first layer and the output of previous layer for convolutional layers 2–5) to generate multiple feature maps as output of convolution operation.

Figure 3. Architecture of Multi-Label BrandImageNet Model



Notes. Conv stands for convolutional layer. FC stands for fully connected layer. The core architecture is based on and is modified from the BVLC Reference CaffeNet model. Our hyperparameters in Conv1–Conv5 and FC6–FC7 are set to the values specified in and the parameters to be estimated are initialized with the learned parameter estimates of the BVLC Reference CaffeNet model. Parameters in the final FC8 layer are initialized with random values.

Specifically, each convolutional layer convolves the input data with a set of kernels. A kernel is often interpreted as a filter: A kernel filters the input to extract certain kinds of features. Each kernel is convolved with the input data by sliding across the width and height of the input data and computing dot products between the kernel and local regions of the input. The output of this convolution is a two-dimensional feature map, where each unit in the feature map is the result of a dot product between the kernel and a specific local region of the input that the unit connects to. Convolutional layer l with d^l different kernels outputs d^l feature maps, where each feature map represents a certain type of local feature extracted. For example, the first convolutional layer in our network has 96 kernels and outputs 96 two-dimensional feature maps.

A convolutional layer l is defined by several hyperparameters: number of kernels d^l, size of each kernel's weight matrix (k^l, k^l, d^{l-1}), padding p^l, and stride s^l.[7] The padding p^l determines the number of zeros to pad around the border of the input before the convolution operation. The stride s^l determines the number of pixels the kernel slides, horizontally or vertically, between applications of the kernel to the input image during convolution operation. For example, $s^l = 2$ means that the kernel will be applied to every other pixel. We set these hyperparameters to values specified in the BVLC Reference CaffeNet model. The weight matrices $\{\boldsymbol{K}_1^l, \ldots, \boldsymbol{K}_{d^l}^l\}$ and biases $\{b_1^l, \ldots, b_{d^l}^l\}$ of the d^l kernels are the parameters of the convolutional layer l we estimate.

The convolutional layer l outputs $\boldsymbol{Y}^l$, consisting of d^l two-dimensional feature maps of size $H^l \times W^l$, where $H^l = (H^{l-1} - k^l + 2p^l)/s^l + 1$ and $W^l = (W^{l-1} - k^l + 2p^l)/s^l + 1$. The ith feature map, denoted $\boldsymbol{Y}_i^l$, is the result of the convolution of the ith kernel and input from previous layer $\boldsymbol{Y}^{l-1}$, defined as

$$\boldsymbol{Y}_i^l = \sum_{j=1}^{d^{l-1}} \boldsymbol{Y}_j^{l-1} * \boldsymbol{K}_{ij}^l + b_i^l,$$

where $\boldsymbol{K}_i^l$ is the weight matrix and b_i^l is the bias of the ith kernel, j indexes the depths of the input data and kernel matrix $\boldsymbol{K}_i^l$, and * denotes convolution. The unit at position (m, n) of the ith feature map, denoted $\boldsymbol{Y}_i^l(m, n)$, is computed as

$$\boldsymbol{Y}_i^l(m,n) = \sum_{j=1}^{d^{l-1}} \sum_{q=1}^{k^l} \sum_{r=1}^{k^l} \boldsymbol{Y}_j^{l-1}(q + (m-1) \times s^l - p^l, r + (n-1) \times s^l - p^l) \cdot \boldsymbol{K}_{ij}^l(q,r) + b_i^l.$$

For example, the size of the kernels in the first convolutional layer is $(11, 11, 3)$. A unit in a feature map is a dot product of a kernel and an $11 \times 11 \times 3$ region in the input image it connects to. Convolutional layers utilize parameter sharing (i.e., units in the same output feature map share the same kernel weights and biases).

3.2.1.2. ReLU Activation Layer. An activation layer takes the input from the previous layer and outputs activation maps by applying an element-wise nonlinear activation function over the input volume. A ReLU activation layer uses a rectified linear unit as its activation function, defined as

$$ReLU(x) = \max(0, x).$$

The activation function is an element-wise operation. Thus, the dimensions of the output and the input are the same. In our model, all five convolutional layers and the first two fully connected layers are followed by a ReLU layer. A ReLU layer does not change data size.

3.2.1.3. Max Pooling Layer. A pooling layer, sometimes called a subsampling layer, is often inserted between convolutional layers to progressively reduce the spatial size (height and width) of the data representation. It helps reduce the number of model parameters to be learned and control overfitting. It also helps make the extracted feature maps robust to small image distortions.

A max pooling layer applies a max pooling function to down-sample the input data. Similar to convolutional layers, it has hyperparameters defined by the size of the max filter (k^l, k^l) and stride s^l. For each depth slice of the input data, it slides spatially and computes a max over local neighborhoods of size (k^l, k^l) and disposes of the rest, thus down-sampling the input data. For example, a max pooling layer with a max filter of size (2, 2) and a stride of 2 will down-sample an input of size (H, W, D) to an output of size (H/2, W/2, D). Note that as it operates independently on every depth slice of the input, the depth of the output is the same as that of the input. In our model, the output of the first, second, and fifth ReLU layers passes a max pooling layer.

3.2.1.4. Fully Connected Layer. Following the convolutional layers are three fully connected layers. Unlike convolutional layers where each unit is connected to a local region of data from the previous layer, units in a fully connected layer are connected to the entire previous layer. The unit i in a fully connected layer l is defined as the dot product of the entire output from the previous layer $\boldsymbol{Y}^{l-1}$ and a unique weight matrix $\boldsymbol{W}_i$, plus a bias term b_i:

$$y_i^l = \boldsymbol{Y}^{l-1} \cdot \boldsymbol{W}_i + b_i.$$

Each unit in a fully connected layer is associated with a different weight matrix and bias offset. For example, the final fully connected layer FC8 contains four units. Each of these four units has a different bias term and weight vector (it is a vector because the size of the input from the previous layer is (4,096, 1)).

The resulting dot products from fully connected layers FC6 and FC7 are passed through a ReLU layer to add nonlinearity. The output of the fully connected layer FC8 is fed into the final Sigmoid8 layer for final multi-label classification.

3.2.1.5. Sigmoid Layer. The final Sigmoid layer is an activation layer. It takes the input from the previous layer and applies a sigmoid function to each element of the input to produce an output between 0 and 1:

$$Sigmoid(x) = \frac{1}{1 + e^{-x}}.$$

In our network, it takes the four output units from FC8 and produces final predicted probabilities for our four brand attributes.

3.2.1.6. Other Layers. We have also added normalization layers (to limit the unbounded activation from ReLU activation layers) and dropout layers (to reduce the number of model parameters) after fully connected layers FC6 and FC7 for regularization purposes to reduce overfitting, as in Krizhevsky et al. (2012). The appendix offers full details of our network architecture and hyperparameters of each layer.

3.2.2. Model Training with Fine-Tuning. Convolutional layers and fully connected layers are trainable layers with model parameters to be learned. These parameters are the kernel matrices and biases $\{\boldsymbol{K}, \boldsymbol{B}^c\}$ in the convolutional layers and the weight matrices and biases $\{\boldsymbol{W}, \boldsymbol{B}^f\}$ in the fully connected layers. We train the model by minimizing a cross-entropy loss function using stochastic gradient descent:

$$L\left(\boldsymbol{K}, \boldsymbol{B}^c, \boldsymbol{W}, \boldsymbol{B}^f\right) = -\frac{1}{N}\sum_n \sum_{a \in \boldsymbol{A}} \left[y_n^a \log\left(Prob\left(y_n^a = 1\right)\right) + \left(1 - y_n^a\right)\log\left(1 - Prob\left(y_n^a = 1\right)\right)\right],$$

where N is the number of training images, $\boldsymbol{A} = \{glamorous, rugged, healthy, fun\}$, y_n^a is the label of the image n for attribute a (1 for positive training examples and 0 for negative training examples), and $Prob(y_n^a = 1)$ is the probability of the nth image expressing attribute a.

A challenge with training a ConvNet model is that it requires very large sets of training data, because both the kernel and the weight matrices have high dimensionality and can have millions of parameters. Training the model from scratch (i.e., initializing all the parameters with random numbers) with a small data set like ours can result in overfitting.

We adopt a fine-tuning approach to avoid the overfitting problem. That is, instead of initializing model parameters with random numbers, we initialize model parameters using learned parameters from the BVLC Reference CaffeNet model.[8] The fine-tuning approach is a case of transfer learning (Bengio et al. 2011, Bengio 2012): knowledge learned in one domain is transferred to help model learning in another domain. It is similar to using an informed prior in Bayesian estimation. Previous work has shown that fine-tuning a pretrained network can significantly increase model performance and avoid overfitting (Donahue et al. 2014, Girshick et al. 2014, Yosinski et al. 2014).

We fine-tune our BrandImageNet model on our data as follows. We initialize parameters of the five convolutional layers and the first two fully connected layers FC6 and FC7 with parameters learned from the BVLC Reference CaffeNet model. We fine-tune the parameters with a small learning rate of 0.0001 over 10,000 iterations. We initialize parameters of the final fully connected layer (FC8) with random values and train it from scratch on our data set with a larger learning rate of 0.001. We use a higher learning rate for the final layer to allow the parameters in the final layer to change quickly with our data. We use a smaller learning rate for earlier layers to preserve the parameters learned from the BVLC Reference CaffeNet model and to transfer their extracted features to our application.

We use the Caffe deep-learning framework (Jia et al. 2014) to fine-tune our BrandImageNet model on a single K80 graphics processing unit node in a university's high-performance cluster. We use 80% of the images for model training, 10% of the images as a validation set to choose the best model snapshot, and the remaining 10% as a holdout sample for evaluation of model performance. We train the model for 10,000 iterations on the training set and choose the iteration at which the model has the best performance on the validation set to select our best model snapshot. We get the best model at iteration 2,800.

3.3. Model Evaluation

We evaluate the performance of our multi-label BrandImageNet model using its per-attribute prediction accuracy (with a 0.5 threshold) and AUC (Bradley 1997). The prediction accuracy is computed as the number of correctly predicted instances divided by the total number of instances. We set the predicted image label for the attribute a presence equal to 1, if the predicted probability of attribute a presence is greater than 0.5 and set it to 0 otherwise.

The AUC is the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (i.e., assign a higher probability of being positive to the positive than to the negative). It evaluates how well a classifier ranks its predictions and separates two classes. The higher the AUC, the better the classifier at separating positive from negative instances. Unlike our per-attribute prediction accuracy measure which assesses model performance at a single classification threshold (0.5), AUC is a measure of aggregated model performance across all possible classification thresholds.

Table 1 presents our model performance for the four perceptual brand attributes we study. Our multi-label ConvNet model achieves an average accuracy of 90% and an average AUC of 0.85, demonstrating good predictive performance.

4. Brand Portrayal on Social Media

We have trained BrandImageNet, a multi-label ConvNet model, to ascertain the presence of four perceptual attributes in an image. We wish to apply this model to brand images posted on social media to understand consumer brand portrayal. In this section, we first show that our model trained on Flickr images can be applied to and performs well on Instagram data. Then, we show that brand portrayal in the consumer posts on social media reflects consumers' brand perceptions. Finally, we present an illustration: a case study of underwear brands.

4.1. Instagram Brand Image Data

Instagram is an image-based social media platform and a popular communication medium. It was launched in 2010 and has one billion active monthly users who, on a daily basis, share an average of 95 million photos. Instagram users often hashtag brands, creating a collection of brand-related images.

We collected data for 56 large national brands covered in the Y&R BAV survey (Mizik and Jacobson 2008) in two product categories for which consumers frequently post photos: apparel and beverages. For each brand, we used Instagram's application programming interface to obtain consumer-created photograph posts hashtagged with the brand name. We filtered out spam, resale, and photos posted by the official account of the brand, and retained about 2,000 photographs for each brand. The final consumer-generated-image data set contains 114,367 photographs. All data were collected between May and October 2016.

Table 1. Out-of-Sample Predictive Performance for Multi-Label BrandImageNet Model on Flicker Hold-Out Data

	Accuracy	AUC
Glamorous	88.6%	0.846
Rugged	91.3%	0.853
Healthy	89.9%	0.859
Fun	89.4%	0.827

We also collected firm-created brand images from the brands' official Instagram account pages, giving us a data set containing 67,863 images. Three beverage brands included in our consumer image data are not represented in this set because they did not have an official account on Instagram at the time of data collection.

4.2. Model Application to Instagram Image Data

We use our multi-label BrandImageNet model to compute the probabilities that a given brand image expresses the four brand attributes we study. Then, we compute the average probability of attribute presence across all consumer-generated images for each brand and each attribute. We use these average probabilities as our metric of how a brand is portrayed by consumers. The higher this score, the more visual content portrays the brand as having a given attribute.

That is, let $\{\boldsymbol{X}_1^{(b)}, \ldots, \boldsymbol{X}_{N^{(b)}}^{(b)}\}$ be $N^{(b)}$ images hashtagged with brand name b. We apply our BrandImageNet model to each image n of brand b to compute the probability that a perceptual attribute a is expressed in this image n: $\Pr(y_n^{(b)}(a) = 1|\boldsymbol{X}_n^{(b)}), n = 1, \ldots, N^{(b)}$, $a \in$ {glamorous, rugged, healthy, fun}. We compute the average of these probabilities and use it as our image-based brand image (IBBI) metric. IBBI for brand b, attribute a is defined as follows:

$$IBBI_{ba} = \frac{\sum_{n=1}^{N^{(b)}} \Pr\left(y_n^{(b)}(a) = 1 \middle| \boldsymbol{X}_n^{(b)}\right)}{N^{(b)}}.$$

We compute IBBI for each of the four attributes, for the entire set of 56 brands. We also compute IBBI for each brand, using firm-created brand images to measure firms' brand portrayal. The full list of our estimates of both consumers' and firms' brand portrayal is provided in Tables OA1–OA4 in the online appendix. Consider, for example, Prada and Eddie Bauer brands. Based on consumer-created brand images, $\text{IBBI}_{\{\text{Prada, glamorous}\}}$ is 0.22, whereas $\text{IBBI}_{\{\text{Eddie Bauer, glamorous}\}}$ is 0.13. This difference is significant (T-stat = 13.5, $p < 0.001$): consumers portray Prada as more glamorous than Eddie Bauer on Instagram. On the ruggedness attribute, $\text{IBBI}_{\{\text{Eddie Bauer,rugged}\}}$ is 0.18, which is significantly higher (T-stat = 14.3, $p <$ 0.001) than Prada's $\text{IBBI}_{\{\text{Prada, rugged}\}}$ of 0.08. The estimates we obtain match our intuition and align with the data pattern in the BAV survey. In the next section, we present a more formal evaluation of model and IBBI metric performance.

4.3. Model Validation on Instagram Image Data

We trained our model on Flickr data because doing so is cost-effective and provides reliable labels for brand attributes. However, to study consumer portrayal of brands on social media, we intend to use Instagram data, which might potentially be different from the Flickr data. Thus, we undertake a thorough validation of our BrandImageNet model application to Instagram images. First, we conduct small-sample validation tests of model performance using human judges and show that our model is applicable and performs well on Instagram data. Then, we consider our IBBI metric against two surveys of consumer perceptions (BAV, a large-scale survey of a nationally representative sample of U.S. adults, and the survey we conducted with active Instagram users) and show that the IBBI metric of brand portrayal on social media reflects consumer brand perceptions.

4.3.1. Assessing Model Performance Against Human Judges. First, we examine the degree of agreement between our model and human judges' identification of brand attributes in Instagram images. For each of our four brand attributes (glamorous, rugged, healthy, fun), we select a set of Instagram images classified by our multi-label BrandImageNet model as (1) "attribute is present" (50 images with the highest predicted probability), (2) "attribute is not present" (50 images with the lowest predicted probability), and (3) "ambiguous" (50 images with around 50% predicted probability of reflecting a given attribute). We investigate these three groups of images to evaluate our model performance at different ranges of probability prediction.[9] We present these images (total number of images = 600, 150 images per attribute) to U.S.-based Level 3 judges on Figure Eight network and ask them to indicate whether the image is glamorous? Rugged? Healthy? Fun? Figure Eight is an on-demand data label and annotation platform specifically designed for collecting training data for machine-learning tasks.[10] We use Level 3 judges—described by the platform as the "smallest group of most experienced, highest accuracy contributors"—to assure quality. We collect 20 independent judgments for each attribute and each image. Each judge is asked to evaluate up to a maximum of 50 images and is compensated based on the number of total evaluations they submitted.

Table 2, column 1, presents the AUC of our model using the majority opinion of human judges as the "ground truth." Column 2 presents a measure of agreement between the model and human judges' labeling of images. We compute agreement as the percentage of images for which the majority of human judges evaluating an image assigns this image the same label as our model (1 = attribute is present, 0 = attribute not present). To offer a relevant benchmark for the agreement measure, we also calculate and report in column 3 this measure for each individual human judge versus the majority vote of human judges. The average percentage of images for which an individual judge agrees with the majority of judges is very close and, on average, slightly lower than the BrandImageNet model versus humans' agreement numbers in column 2. Importantly, the Instagram data reveal the difficulty of identifying brand attributes from images and the heterogeneity in human evaluations of images: we have only 70 (of 600 studied) images with 100% agreement across 20 human judgments.

Table 3 presents our measure of agreement between the model and human judges for the three types of images we selected (attribute present/ambiguous/attribute not present). We find high levels of agreement for unambiguous images: agreement between the model and human judgments is notably higher in groups 1 and 3, where the presence or absence of an attribute is more pronounced. For the ambiguous images in group 2, where the model-based probability score for an attribute presence is close to 50%,

Table 2. Aggregate BrandImageNet Model Performance According to Human-Based Image Labels

	AUC: Model vs. the majority vote of human judges	Agreement: Model vs. the majority vote of human judges	Agreement: A single human judge vs. the majority vote of human judges, average
Glamorous	0.93	83%	85%
Rugged	0.96	85%	83%
Healthy	0.91	78%	80%
Fun	0.94	84%	80%
Average	0.94	83%	82%

Notes. Data are based on 600 images (150 images for each attribute) and 12,000 total judgments (20 judgments for each image and attribute). The model-based label is equal to 1 if the model-based probability estimate for attribute presence is greater than 50%, and 0 otherwise. The human-based label for an image is equal to 1 if the majority of the judges indicate attribute presence, and 0 otherwise. Agreement is the percentage of images for which the majority of human judges evaluating an image assign this image the same label as our model. The total cost of data collection is $288.96, with an average cost per judgment of $0.024.

Table 3. Agreement Between BrandImageNet Model and Human Labels by Image Group

	Group 1: "Attribute present" (high predicted probability)	Group 2: "Ambiguous images" (~50% predicted probability)	Group 3: "Attribute absent" (low predicted probability)
Glamorous	96%	52%	100%
Rugged	100%	56%	100%
Healthy	84%	52%	98%
Fun	94%	58%	100%

Notes. Data are based on 600 images (150 images for each attribute, 50 images per attribute in each group) and 12,000 total judgments (20 judgments for each image and attribute). Agreement is computed as the percentage of images for which the majority of human judges evaluating an image assign this image the same label as the BrandImageNet model. The model-based label is equal to 1 if the model-based probability estimate for attribute presence is greater than 50%, and 0 otherwise.

we see lower agreement between model-based and human-based labels.

Table 4 presents an alternative metric to evaluate agreement between the model and human judgments. It shows the percentage of human judgments indicating the presence of an attribute in an image, by image group. As expected, we see a monotonic decline in the positive identification of attribute presence by human judges as we go from high to low model-predicted likelihood of attribute presence.

4.3.2. Accessing IBBI Metric Against Survey-Based Brand Perceptions Data. To gain further insights, we compare the model-predicted IBBI scores from consumer-generated and firm-created brand images on Instagram with the data from a large-scale consumer survey of brand perception conducted by the Y&R BAV among U.S. adults.

The BAV survey asks a large, nationally representative sample of respondents to indicate whether they perceive a given brand as possessing a particular attribute. The BAV data are summarized as a percentage of consumers who responded affirmatively to the question of whether a particular brand represents a given attribute. We use BAV data for glamorous, rugged, fun, and healthy from the first quarter of 2016 to match our image data collection timeframe.

Table 5 presents Pearson correlations for the brand metrics computed from consumer-created brand images, firm-created brand images, and the BAV survey. We see a high degree of correspondence between the consumer and firm image-based IBBI measures. Most, but not all, correlations are high and highly significant. We also find a high degree of correspondence between consumer image-based IBBI and BAV measures and firm image-based IBBI and BAV measures for the attributes that are more relevant for the category. For apparel, we observe higher correlations for the glamorous and rugged attributes. For beverages, we see high correlations for healthy.

Table 5 shows high and significant correlations between IBBI measures and BAV survey data. However, a significant difference likely exists between the population of BAV respondents and the population of active Instagram users posting images online. To gain further insights, we replicate our correlation analyses using the data from a survey we conducted among active Instagram users. If our IBBI metrics of brand portrayal reflect consumer brand perceptions, then we should see a higher degree of correspondence between IBBI and brand perceptions of Instagram users.

Indeed, the BAV survey data come from a nationally representative sample of U.S. adults. Consumer-created images on Instagram are posted by active Instagram users, who tend to be relatively young and tech savvy. Moreover, many consumer brand images depict personal consumption experience. For example, in a pretest that we ran, we find that at least one of three human evaluators whom we asked to review the images flagged 81.5% of the consumer-generated Prada photos as depicting personal consumption experience, whereas only about 3.5% of the BAV survey respondents report using the brand regularly and 86.2% report that they have never used it.[11]

Thus, we conducted an online survey to collect brand perceptions data from a random sample of

Table 4. Percentage of Human Judgments Indicating the Presence of an Attribute in Instagram Images by Group

	Group 1: "Attribute present" (high predicted probability)	Group 2: "Ambiguous images" (~50% predicted probability)	Group 3: "Attribute absent" (low predicted probability)
Glamorous	83%	34%	8%
Rugged	79%	39%	6%
Healthy	76%	24%	23%
Fun	80%	37%	14%

Note. Data are based on 600 images (150 images for each attribute, 50 images per attribute in each group) and 12,000 total judgments (20 judgments for each image and attribute).

Table 5. Correlation Analyses of Model-Predicted IBBI from Consumer- and Firm-Created Images on Instagram and the BAV Survey-Based Measures of Brand Perceptions

		Consumer IBBI vs. firm IBBI	Consumer IBBI vs. BAV	Firm IBBI vs. BAV
Apparel	Glamorous	0.7838***	0.5519***	0.6100***
	Rugged	0.9122***	0.5467**	0.5035**
	Healthy	0.4680**	0.1794	0.3225*
	Fun	0.6061***	0.3583*	0.2883$^{\pm}$
Beverage	Glamorous	0.5518**	0.4568**	0.6582***
	Rugged	0.8259***	0.3596*	0.4708*
	Healthy	0.7370***	0.6976***	0.4766**
	Fun	0.3775*	0.1791	0.2584

$^{\pm}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$ (Pearson correlation).

Instagram users. We followed the data-collection protocol similar to that used by BAV. We have a total of 176 self-reported Instagram users responding to the survey of apparel brands and 175 self-reported Instagram users responding to the survey of beverage brands. Similar to BAV, we removed respondents who indicated no familiarity with a specific brand. As such, the number of respondents differs by brand and is lower for lesser-known brands. We used Figure Eight to collect survey data and have an average of 148 and a minimum of 87 responses per brand.

Table 6 presents correlations between our IBBI metrics and the survey of Instagram users. We find notably higher correlations of firm- and consumer-image-based IBBI with the Instagram users' survey data than with the BAV survey data (shaded cells in Table 6 mark notable improvement over respective Table 5 results), and the average increase in correlations of the survey data for the Instagram users with the consumer-based IBBI is somewhat greater than the increase with the firm-based IBBI. We observe higher correlations for glamorous, rugged, and fun in the apparel category. For beverages, we observe higher and more significant correlation for glamorous, rugged, and healthy. These results, based on a survey of a more relevant population of Instagram users, suggest that our IBBI measure indeed reflects consumer brand perceptions. Although we cannot fully eliminate the population differences (a random sample of Instagram users we obtained may be different from the Instagram users who hashtag brands, and people who hashtag Victoria's Secret may be different from people who hashtag Eddie Bauer), these results further support the point that our model is able to identify brand perceptions depicted in the brand imagery.

In summary, these results provide further support for the notion that IBBI measure reflects consumer brand perceptions.

4.4. A Case Study of the Underwear Brands: Firm and Consumers' Brand Portrayal on Social Media

Firms engage in branding through advertising, social media, and product packaging to create a particular brand image in consumers' minds. The imagery from a brand's official account reflects the firm's positioning efforts and the desired perceptions and associations the firm wants consumers to have.

Table 6. Correlations Between Model-Predicted IBBI from Consumer- and Firm-Created Images on Instagram and Survey-Based Measures of Brand Perceptions from Instagram Users

		BAV vs. Instagram user survey	Consumer IBBI vs. Instagram user survey	Firm IBBI vs. Instagram user survey
Apparel	Glamorous	0.9503***	0.5824***	0.6325***
	Rugged	0.9338***	0.6831***	0.6630***
	Healthy	0.8600***	0.0842	0.1941
	Fun	0.6486***	0.5672***	0.4914**
Beverage	Glamorous	0.9238***	0.5001**	0.5743**
	Rugged	0.5485***	0.7899***	0.6645***
	Healthy	0.9482***	0.7127***	0.5350**
	Fun	0.8714***	0.2130	0.3648*

Notes. The shaded cells represent instances of improvement over the correlations between consumer- and firm-image-based IBBI and the BAV data reported in Table 5. Average number of respondents per brand is 148 (min = 87, max = 175). Cost of the survey data collection is $547.20.

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$ (Pearson correlation).

The BrandImageNet model allows examination of consumers' and firms' portrayal of brands on social media. By comparing perceptual attributes expressed in the official brand images with those in consumer-created images, one may be able to examine whether firm portrayal matches consumers' portrayal of the brands and make inferences on whether the firms' brand-positioning objectives have been achieved.

Although we find high and significant correlations between IBBI scores from consumer- and firm-created images across all attributes that we study, we do observe few notable differences for some brands and attributes. Figure 4 presents a chart of IBBI scores for glamorous in the apparel category. Consider, for example, Victoria's Secret, Joe Boxer, and Hanes brands. All three brands come from the same apparel subcategory: underwear. Victoria's Secret has the highest IBBI for glamorous for firm images and a slightly lower IBBI for consumer images. Joe Boxer has a similar medium level of IBBI for both firm and consumer images. Hanes has a very large differential between the firm-based and consumer-based IBBI (firm images are much more glamorous). These numbers suggest consumer perceptions of Joe Boxer are consistent with brand positioning. Consumer perceptions of Victoria's Secret fall just short of the company's positioning on glamorous. Consumer perceptions reflected in their portrayal of Hanes, however, are highly inconsistent with how the company portrays the brand.

We have validated these patterns using human judges to rule out the possibility that they are simply an artifact of model-prediction error. We selected a random set of 500 consumer- and 500 firm-created images for each brand. For Joe Boxer and Hanes, we use all available firm-created images because less than 500 firm images are available on Instagram for these brands. A total of 2,297 images are included in this test. Ten Level 3 judges on the Figure Eight platform evaluated each of the 2,297 images. Table 7 presents a summary of our findings showing consistency between our model and human judgments.

Consistent with our model predictions, human judges perceive firm and consumer images for Victoria's Secret as highly glamorous—more glamorous than the images of Joe Boxer and Hanes. Importantly, both the model and the judges identify firm images as more glamorous than consumer images. Joe Boxer's firm images are judged as medium on glamor for both firm and consumer images, and although the slight positive differential is not significant in model predictions,

Figure 4. IBBI Scores for Glamorous in the Apparel Category

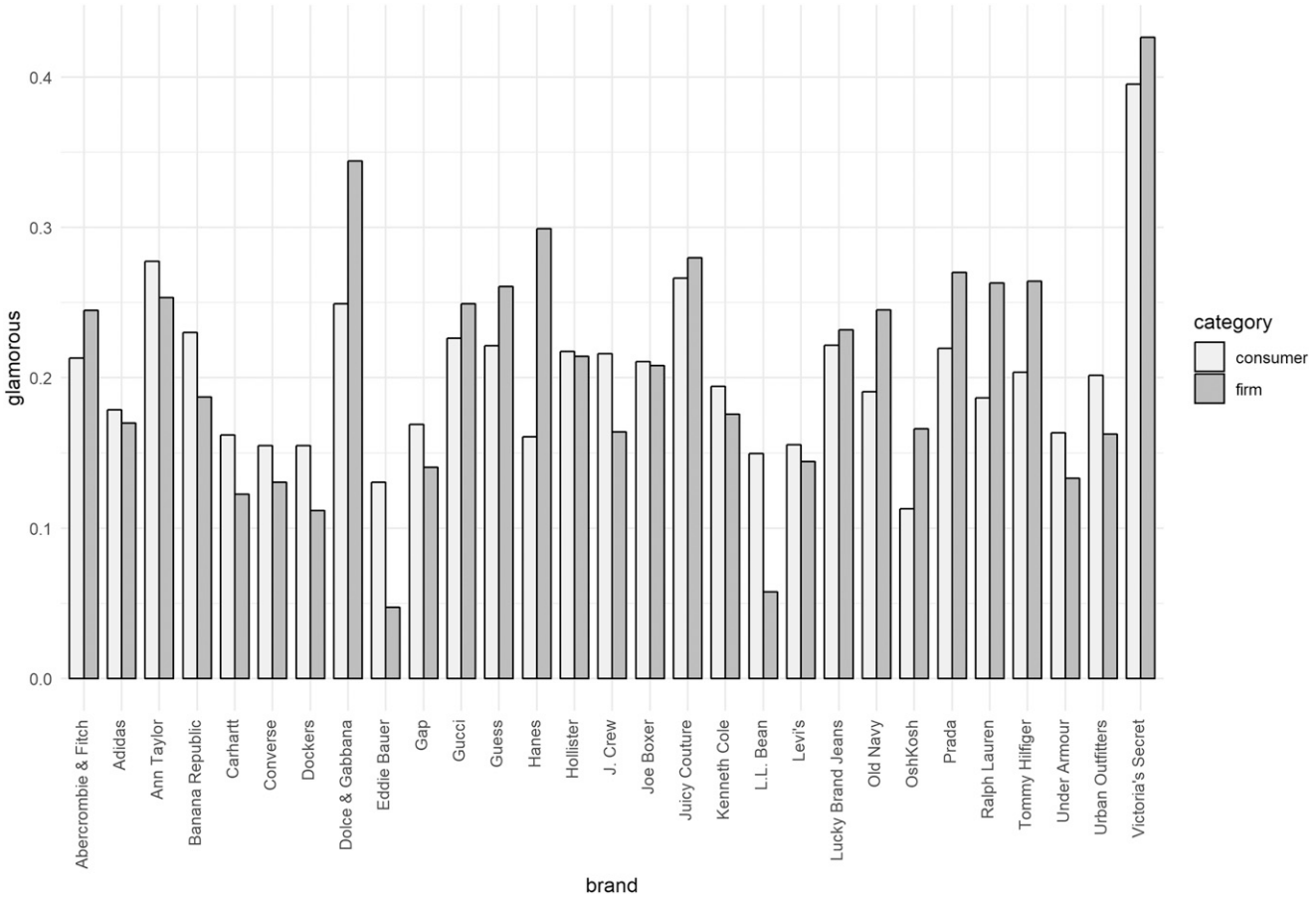


Table 7. Victoria's Secret, Joe Boxer, and Hanes Case Study

Panel A: IBBI scores: BrandImageNet model and human judges

		BrandImageNet[a]	Human judges[b]
Victoria's Secret	Consumer images (N = 500)	0.39	0.47
	Firm images (N = 500)	0.43	0.63
Joe Boxer	Consumer images (N = 500)	0.20	0.11
	Firm images (N = 134)	0.21	0.21
Hanes	Consumer images (N = 500)	0.16	0.12
	Firm images (N = 163)	0.30	0.36

Panel B: IBBI score differentials between firm and consumer images: Difference (T-stat, p-value)

	BrandImageNet	Human judges
Victoria's Secret: firm versus consumer images	0.04 (2.25, 0.02)	0.15 (7.92, <0.001)
Joe Boxer: firm versus consumer images	0.004 (0.24, 0.81)	0.10 (5.86, <0.001)
Hanes: firm versus consumer images	0.14 (7.61, <0.001)	0.24 (13.98, <0.001)

Panel C: IBBI score differentials between brands: Difference (T-stat, p-value)

		BrandImageNet	Human judges
Victoria's Secret versus Joe Boxer	Consumer images	0.18 (12.14, <0.001)	0.36 (22.67, <0.001)
	Firm images	0.22 (8.11, <0.001)	0.42 (15.75, <0.001)
Victoria's Secret versus Hanes	Consumer images	0.22 (15.01, <0.001)	0.35 (21.58, <0.001)
	Firm images	0.13 (4.98, <0.001)	0.26 (10.43, <0.001)
Joe Boxer versus Hanes	Consumer images	0.04 (3.71, <0.001)	−0.02 (−1.68, 0.09)
	Firm images	−0.09 (-3.26, 0.001)	−0.16 (−5.53, <0.001)

Notes. We collected human judgments using the procedure outlined in Section 4.3.1. To contain the costs, we used 10 U.S.-based Level 3 judges to evaluate each image. We obtain an AUC of 0.81 using majority-based human labels for each image as the "ground truth." Agreement between human and model labels is 79%. Average agreement between each individual human judge and the majority of human judges is 83%. Data collection cost is $573.60.

[a]The IBBI scores here are based on a random sample of 500 images and, as such, are different from the IBBI scores reported in Figure 4 and Table OA1 in the online appendix.

[b]We compute human-based IBBI as follows. For each image, we compute the probability that it expresses glamor as the number of judges who label it as glamorous divided by the number of judges evaluating the image (10 judges). Then, we compute the average of these numbers for a brand and use it as the human-based IBBI.

the judges perceive firm images as significantly more glamorous. Both the model and the human judges perceive firm images of Hanes as much more glamorous than consumer images of Hanes. Our model prediction scores are consistent with the human perceptions of the images. The observed differentials identified by our model in the consumer-created and firm-created brand images for a given brand and across brands are also present in human evaluations.

This case study considers differences in the depiction of brand attributes in consumer and firm images at one time point. Managers, however, can use the BrandImageNet model and our proposed IBBI metric to monitor and track changes in consumers' brand portrayal to detect changes in consumers' depiction of their brand. Firms can also use this approach to track how fast consumers pick up new brand positioning or how they respond to a brand crisis.

Overall, our empirical analyses suggest that the visual content that consumers post on social media reveals their brand perceptions. The method we propose can extract this information and help marketing managers monitor and identify changes in brand perceptions over time, identify discrepancies between intended brand positioning and consumers' perceptions, and better understand their brand positioning in the context of a competitive landscape.

5. Sensitivity Analyses

We undertook several sensitivity analyses to select the best model specification for our task of extracting perceptual brand attributes. For example, we considered different options for initializing our BrandImageNet model parameters for fine-tuning. We tested a set of models initialized with learned parameters from Karayev's Fine-Tuned Flickr-Style ConvNet model for image style recognition.[12] Karayev's ConvNet is a multiclass classification model, which itself was fine-tuned on 80,000 Flickr images from the BVLC Caffe Reference model. Karayev's ConvNet could potentially be a

better model to fine-tune for our application, because an image style recognition task is presumably more relevant to our perceptual-attribute-identification task than ConvNet models developed for object identification. We do not find that to be the case. We find that initializing with the original BVLC Caffe Reference model parameters generates a model with the best performance on the validation set. It also has the best out-of-sample fit on our Flickr data and generates the highest correlations with the BAV data and with our brand perceptions survey of active Instagram users.

Further, we took steps to identify a learning rate for the best model performance on our validation data. Although the differences in model performance across alternative learning-rate specifications are somewhat minor, we find that a lower learning rate of 0.0001, rather than the BVLC's recommended learning rate of 0.001, generates the best overall model performance in both validation and hold-out samples (results available upon request).

We also evaluated our deep-learning-based BrandImageNet model performance against alternative machine-learning approaches. We tested the BrandImageNet model against a popular Support Vector Machines (SVM) approach (see the online appendix for details of the SVM application). SVM is not a deep-learning method. It requires the researcher to first define and extract a set of image features. We selected a set of typical features popular in computer vision research related to color, shape, and texture and trained a multi-label SVM classifier on these features to predict the presence of perceptual brand attributes using our sample of labeled Flickr images. Similar to our approach to training the BrandImageNet model, we trained the SVM classifier on 80% of the Flicker images, used 10% as a validation set to choose model hyperparameters, and kept the remaining 10% as a holdout sample to evaluate model performance. Our SVM classifier achieves an average accuracy of 87% and an AUC of 0.72 across the four perceptual attributes that we train it to identify (see Table OA6 in the online appendix). The SVM performance is lower than that of our BrandImageNet model.

6. Discussion

The rapidly growing volume of visual brand-related content posted on social media is a new valuable information source for marketers and brand managers to monitor, track, and better understand brand performance. As text-mining approaches have gained popularity in leveraging UGC for brand monitoring, image-mining approaches are still relatively new and underutilized. Out study takes a step toward understanding and utilizing consumer-created images on social media. It bridges the image-mining and machine-learning literature with the branding literature by proposing an approach to online brand monitoring and market intelligence gathering from consumer-generated images.

The proposed approach enables managers to monitor how their brands are portrayed on image-based social platforms. We find that the brand-portrayal metrics derived from consumer-created brand images are strongly correlated with survey-based metrics of consumer brand perceptions. That is, consumers' portrayal of brands on social media contains valuable information about brands: it reflects consumer brand perceptions.

Firms can use our BrandImageNet model to understand consumers' brand perceptions, monitor brand performance and the success of repositioning efforts, and identify gaps in positioning strategies. Firms can use our model to screen consumer-created brand images to select some to feature in their marketing communications. Firms can also apply the BrandImageNet model to the visual content created and shared by a consumer (rather than a brand) to better understand this individual consumer's personality and identity (the focal components of the beliefs about self). Marketing theory postulates that better fit between consumer identity and brand identity is important for brand acceptance and success in the marketplace. That is, matching a consumer's identity attributes with fitting firm-generated imagery or ad copy can increase click-through rates, engagement, liking, and the likelihood of purchase. The methods described in the manuscript can be used to analyze an individual consumer's public visual data stream to extract specific attributes (e.g., healthy, glamorous, etc.) represented in the imagery the consumer creates and/or posts on social media platforms. The learned insights can be leveraged for better ad targeting by matching consumer identity characteristics with ad creatives. This application can be useful for both search engines serving ads and for the advertisers.

7. Limitations and Directions for Future Research

Our study has several limitations. One important limitation is that our approach does not allow to draw inferences about the specific image properties and content elements driving consumers' perceptions of whether the image conveys a particular perceptual attribute. Perceptions may be shaped by the focal content elements of the image (persons, products, objects in the image), the background of the image, or the design elements of the image (color scheme, texture, etc.). Understanding how these different elements shape the overall perceptions of the image is interesting and important. But because our BrandImageNet model utilizes a deep-learning approach, which simultaneously identifies the features and links them to

perceptual attributes in the images, we are not able to answer these questions in the current study. While ConvNet models have better predictive performance than traditional methods with feature engineering (e.g., SVM), they do not lend themselves to interpretability that may be feasible with supervised machine learning methods where researchers explicitly define image features. As such, we leave these questions for future research. Future research investigating these questions may help improve advertising copy design and social media campaigns.

Future research can also explore what drives the differences in the depiction of brands in consumer and firm images posted on social media. We find that consumers' portrayal of brands on social media reflects their brand perceptions (as measured with a traditional external survey tool). As such, the differences between the firm and consumer brand portrayal could simply reflect the differences in consumers' actual brand perceptions and the firms' intended brand positioning. There might, however, also be other reasons driving these differences. For example, differences can also be driven by the type of images that consumers are able to produce compared with firms (quality, artistry), and by the type of images consumers are willing to share online (e.g., what is socially appropriate, desirable, or consistent with the individual consumer's self-image and self-signaling objectives versus how the consumers actually perceive the brand). Consider, for example, Victoria's Secret brand. A customer might perceive Victoria's Secret brand as extremely glamorous but might not be able to create or be willing to post images on her/his Instagram account as glamorous and revealing as the images on the Victoria's Secret's Instagram account.

Another limitation of the presented approach is that we predefine which perceptual brand attributes we study. Future research in this area can explore unsupervised machine-learning methods to help uncover brand attributes that consumers care about but that brand managers have not yet recognized and considered. Visual content is very amenable to unsupervised machine-learning methods. Similar to text-mining applications for extracting topics and uncovering consumer needs (Timoshenko and Hauser 2019), visual content mining can be used to uncover new product-usage situations, identify individual personality traits of the content creator or a consumer group, and help design products and communication strategies to better fit and address the needs of the individual or consumer segment.

8. Conclusion

Visual content is a ubiquitous part of modern life. It affects consumers' beliefs, preferences, and decision-making at multiple stages. As such, incorporating visual content analyses into marketing models and management processes is imperative. The method presented in this paper provides evidence that capturing and analyzing rich image data generated by consumers is possible and can generate valuable insights for marketers. We hope that our research provides a foundation and encourages more future research in this emerging area.

Appendix. BrandImageNet Model Architecture and Hyperparameters

Layer	Size	Kernel size	Stride	Pad
Input image	227 × 227 × 3			
Convolution 1	55 × 55 × 96	11 × 11 × 3	4	0
ReLU1	55 × 55 × 96			
MaxPooling1	27 × 27 × 96	3 × 3	2	0
Norm1	27 × 27 × 96			
Convolution2	27 × 27 × 256	5 × 5 × 96	1	2
ReLU2	27 × 27 × 256			
MaxPooling2	13 × 13 × 256	3 × 3	2	0
Norm2	13 × 13 × 256			
Convolution3	13 × 13 × 384	3 × 3 × 256	1	1
ReLU3	13 × 13 × 384			
Convolution4	13 × 13 × 384	3 × 3 × 384	1	1
ReLU4	13 × 13 × 384			
Convolution5	13 × 13 × 256	3 × 3 × 384	1	1
ReLU5	13 × 13 × 256			
MaxPooling5	6 × 6 × 256	3 × 3	2	0
FC6	4,096			
ReLU6	4,096			
Dropout6	4,096	dropout rate = 0.5		
FC7	4,096			
ReLU7	4,096			
Dropout7	4,096	dropout rate = 0.5		
FC8	4			
Sigmoid8	4			

Endnotes

[1] See https://instagram-press.com/our-story/ (accessed August 15, 2019) and https://www.socialpilot.co/blog/social-media-statistics (accessed August 15, 2019).

[2] See https://www.instagram.com/explore/tags/nike/ (accessed August 15, 2019).

[3] We considered collecting a training set of Instagram images based on hashtags as an alternative. However, a hashtag is too noisy to serve as an effective supervisory label, because users apply hashtags that are irrelevant or leave out relevant hashtags. As a result, training models using such weakly supervised data usually require extremely large data sizes (Mahajan et al. 2018). Using MTurk to label the images is also not practical for our purposes. The costs of securing a sufficiently large sample of images and human judges on Mturk to get reliable labels is prohibitive. Flickr, on the other hand, is cost-effective and is also more reliable. In Section 4, we present evidence validating the appropriateness of our BrandImageNet model (trained on Flickr) for application to Instagram data.

[4] RGB constructs all possible colors as a combination of the red, green, and blue channels.

[5] Regular lowercase letters denote scalars, bold lowercase letters denote vectors, and bold uppercase letters denote matrices and sets.

[6] See https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet (accessed August 15, 2019).

[7] Note that d^{l-1} is equal to the depth (the third dimension) of the input data from the previous layer.

[8] As we discuss in Section 5, we undertook multiple tests of model fine-tuning (e.g., varying the fine-tuning learning rate and using different values to initialize parameters) to train our BrandImageNet model for best performance. We found that initializing with BVLC Reference CaffeNet parameters and using a 0.0001 learning rate in Conv1–Conv5 and FC6–FC7 and a 0.001 learning rate in FC8 generates the best model according to model performance on the validation set. Thus, we focus on this specification in our main discussion. This specification also generates best performance in Flickr out-of-sample tests and in Instagram tests against human judges and consumer-perceptions surveys.

[9] Figure OA1 in the online appendix reports the distribution of the predicted probabilities of attribute presence across all Instagram photos for each attribute.

[10] See https://www.figure-eight.com/ (accessed August 15, 2019).

[11] We ran a Figure Eight pretest task in which human judges evaluated whether an image depicted personal consumption experience or not. Each image was evaluated by three judges. Full results available upon request.

[12] See https://github.com/BVLC/caffe/tree/master/models/finetune_flickr_style (accessed August 15, 2019).

References

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.

Bengio Y (2012) Deep learning of representations for unsupervised and transfer learning. Guyon I, Dror G, Lemaire V, Taylor G, Silver D, eds. *Proc. ICML Workshop Unsupervised Transfer Learn.* (Bellevue, WA), 17–36.

Bengio Y, Bergeron A, Boulanger-Lewandowski N, Breuel T, Chherawala Y, Cisse M, Erhan D, et al (2011) Deep learners benefit more from out-of-distribution examples. Gordon G, Dunson D, Miroslav D, eds. *Proc. 14th Internat. Conf. Artificial Intelligence Statist., (Fort Lauderdale, FL)*, 164–172.

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7):1145–1159.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.

Culotta A, Cutler J (2016) Mining brand perceptions from Twitter social networks. *Marketing Sci.* 35(3):343–362.

Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. *2009 Conf. Comput. Vision Pattern Recognition* (IEEE, New York), 248–255.

Dhar S, Ordonez V, Berg TL (2011) High level describable attributes for predicting aesthetics and interestingness. *2011 Conf. Comput. Vision Pattern Recognition* (IEEE, New York), 1657–1664.

Diehl K, Zauberman G, Barasch A (2016) How taking photos increases enjoyment of experiences. *J. Personality Soc. Psych.* 111(2): 119–140.

Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) DeCAF: A deep convolutional activation feature for generic visual recognition. *Proc. Machine Learn. Res.* 32: 647–655.

Gabel S, Guhl D, Klapper D (2019) P2v-map: Mapping market structures for large retail assortments. *J. Marketing Res.* 56(4):4557–4580.

Gardner BB, Levy SJ (1955) The product and the brand. *Harvard Bus. Rev.* 33(2):33–39.

Giannakopoulos T, Papakostas M, Perantonis S, Karkaletsis V (2015) Visual sentiment analysis for brand monitoring enhancement. *2015 9th Internat. Sympos. Image Signal Processing Anal. (ISPA)* (IEEE, New York), 1–6.

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 Conf. Comput. Vision and Pattern Recognition* (IEEE, New York), 580–587.

Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning* (MIT Press, Cambridge, MA).

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. *Proc. 22nd ACM Internat. Conf. Multimedia* (ACM, New York), 675–678.

Jiang YG, Wang Y, Feng R, Xue X, Zheng Y, Yang H (2013) Understanding and predicting interestingness of videos. *27th AAAI Conf. Artificial Intelligence* (ACM, New York), 1113–1119.

Karayev S, Trentacoste M, Han H, Agarwala A, Darrell T, Hertzmann A, Winnemoeller H (2014) Recognizing image style. Valstar M, French A, Pridmore T, eds. *Proc. British Machine Vision Conf.* (BMVA Press, Durham, UK).

Keller KL, Lehmann DR (2006) Brands and branding: Research findings and future priorities. *Marketing Sci.* 25(6):740–759.

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Processing Systems* 25:1097–1105.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553): 436–444.

Liu X, Lee D, Srinivasan K (2019) Large scale cross category analysis of consumer review content on sales conversion leveraging deep learning. *J. Marketing Res.* 56(6):918–943.

Liu Y (2006) Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing* 70(3):74–89.

Mahajan D, Girshick R, Ramanathan V, He K, Paluri M, Li Y, Bharambe A, van der Maaten L (2018) Exploring the limits of weakly supervised pretraining. *Proc. Eur. Conf. Comput. Vision (ECCV)* (Springer Science and Business Media, Berlin), 181–196.

McAuley J, Leskovec J (2012) Image labeling on a network: Using social-network metadata for image classification. Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C, eds. *Eur. Conf. Comput. Vision* (Springer, Berlin), 828–841.

Mizik N, Jacobson R (2008) The financial value impact of perceptual brand attributes. *J. Marketing Res.* 45(1):15–32.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Sci.* 31(3):521–543.

Park CW, Jaworski BJ, Maclnnis DJ (1986) Strategic brand concept-image management. *J. Marketing* 50:135–145.

Pavlov E, Mizik N (2019) Increasing consumer engagement with firm-generated social media content: The role of images and words. Working paper, University of Washington, Seattle.

Raghubir P, Greenleaf EA (2006) Ratios in proportion: What should the shape of the package be? *J. Marketing* 70(2):95–107.

Stadlen A (2015) Find every photo with Flickr's new unified search experience. Accessed May 7, 2015, https://blog.flickr.net/en/2015/05/07/flickr-unified-search/.

Timoshenko A, Hauser JR (2019) Identifying customer needs from user-generated content. *Marketing Sci.* 38(1):1–20.

Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *J. Marketing Res.* 51(4):463–479.

Wedel M, Pieters R, eds. (2007) *Visual Marketing* (Lawrence Erlbaum Associates, New York).

Wedel M, Pieters R (2014) The buffer effect: The role of color when advertising exposures are brief and blurred. *Marketing Sci.* 34(1): 134–143.

Xiao L, Ding M (2014) Just the faces: Exploring the effects of facial features in print advertising. *Marketing Sci.* 33(3):338–352.

Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? *Adv. Neural Inform. Processing Systems* 27:3320–3328.

Zhang H, Korayem M, Crandall DJ, LeBuhn G (2012) Mining photo-sharing websites to study ecological phenomena. *Proc. 21st Internat. Conf. World Wide Web* (ACM, New York), 749–758.

Zhang J, Wedel M, Pieters R (2009) Sales effects of attention to feature advertisements: A Bayesian mediation analysis. *J. Marketing Res.* 46(5):669–681.

Zhang M, Luo L (2019) Can user-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp. Working paper, University of Southern California, Los Angeles.

Zhang S, Lee D, Singh PV, Srinivasan K (2018) How much is an image worth? Airbnb property demand estimation leveraging large scale image analytics. Working paper, David S. Tepper Business School, Carnegie Mellon University.