Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-020-00429-0

REGULAR ARTICLE

Predicting brand confusion in imagery markets based on deep learning of visual advertisement content

Atsuho Nakayama[1] · **Daniel Baier**[2]

Received: 15 July 2019 / Revised: 4 August 2020 / Accepted: 8 November 2020

Abstract
In the consumer goods industry, unique brand positionings are assumed to be the road to success. They document product distinctiveness and so justify high prices. However, as products are getting more and more interchangeable, brand positionings must rely—at least partially—on supporting advertisements. Here, especially ads with visual content (e.g. photos, video clips) are able to connect brands with desirable emotions and values. Recently, besides TV, cinema, newspaper, also search engines, social networks, photo-, video-sharing platforms are used to spread such ads. In this paper, we demonstrate, how deep learning based on such ads can be used to predict uniqueness of brand positionings. A sample application to the German Pils beer market is used for demonstration.

Keywords Brand confusion · Brand positioning · Convolutional Neural Network (CNN) · Grad-CAM · VGG16

Mathematics Subject classification Primary 68T10; Secondary 90B60 · 62H30 · 62H35

1 Introduction

Today, in many consumer goods markets, the physico-chemical differences between competing products are diminishing, mainly due to standardization in product and production technology. So, e.g., Stiftung Warentest, a well-known German consumer organization, compared 1739 products in 72 nondurable consumer goods markets (e.g., beer, butter, cheese, cognac, parfume, shampoo) and found out that even the low priced

✉ Atsuho Nakayama
atsuho@tmu.ac.jp

[1] Tokyo Metropolitan University, 1-1 Minami-Ohsawa, Hachioji-shi 192-0397, Japan

[2] University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany

Published online: 19 November 2020

products were—on average—of the same ingredients, production quality, and taste than the high priced ones (Stiftung Warentest 2020). They received similar results when comparing durable products (e.g, notebooks, smartphones, TVs, washing machines) in selected markets (Stiftung Warentest 2020): Physico-chemical differences were moderate and low priced products were on average of the same quality as high priced ones.

For producers of high priced products, this development causes serious problems: Relying on (diminishing) physico-chemical differences is not enough to justify high prices. Instead, they have to connect their brands with emotions and values and so make them particularly desirable for their targeted consumers. Coca Cola and Pepsi Cola in the soft drink markets or Apple and Samsung in the smartphone and computer markets are examples for the necessity and success of these so-called positioning strategies (see, e.g., Dissanayake and Amarasuriya 2015). The same holds for the premium Pils beer (also called TV beer) market. As many breweries use the same product and production technology, physico-chemical differences are moderate, recognition even among loyal customers in blind tests (beer tasting without knowing the brand name) is poor (Almenberg et al. 2014). Positioning strategies based on ads, often with appealing visual content, are necessary. These marketing and promotion campaigns take more than 15% of the breweries' total costs, placing the brewing industry among the top branded consumer industries worldwide (Madsen 2017).

In the marketing literature such markets—where physico-chemical differences between products are moderate but nevertheless consumers distinguish them according to their brand positionings based on ads—are called imagery brand markets (see, e.g., Keon 1983, 1984; Böckenholt and Gaul 1985; Gaul and Baier 1994; Baier and Gaul 1998). Here, the main task for marketers consists in developing and spreading ads with appealing visual content (e.g. photos, video clips) that strenghten the own brand's positioning and in controlling that the own ads don't strenghten competing brands due to similar ad contents and/or derived brand positionings.

Several approaches have been proposed and applied to support these tasks. The basic approach is to perform so-called brand confusion experiments (see, e.g., Keon 1983, 1984; Böckenholt and Gaul 1985; Gaul and Baier 1994; Kroeber-Riel and Esch 2015): Consumers are confronted with print ads of competing brands (with masked brand names and slogans) and asked to guess the advertised brands. The resulting confusion matrix (indicating how often each brand was guessed when a print ad was shown) then allows to discuss brand confusion. Recently, based on image data analysis and classification (Baier et al. 2012), an alternative approach was proposed (Baier and Frost 2018): Similarities between print ads (with masked brand names and slogans) are calculated based on extracted low and high level image features (e.g. color and edge distributions, number of faces, MPEG features). It could be shown that based on these similarities a confusion matrix can be predicted that closely replicates the results of a brand confusion experiment.

In this paper, we extend this approach in (1) that we analyze larger numbers of ads with visual content (photos, video clips) and (2) that we use deep learning for the prediction of brand confusion. The first extension allows to analyze besides print ads in traditional (e.g., newspapers, magazines, on advertising pillars) and new channels (e.g., search engines, social media, photo sharing platforms) also the important video clips in traditional (e.g., TV, cinema) and new channels (e.g., search engines, social

media, video sharing platforms). The second extension takes into account that deep learning has evolved rapidly over recent years. Here, especially, convolutional neural networks (CNNs) have become the go-to algorithm for tasks were visual content is studied. More generally, CNNs work on all perceptual tasks (Chollet and Allaire 2018) and at major recent computer vision conferences, it was nearly impossible to find presentations that didn't involve CNNs in some form.

The proposed new approach can be summarized as follows: For a product category under study, we collect a training sample of ads with visual content (photos or video clips) where we know the advertised brand. Then, we mask the brand name and slogan in the images or video clips and train a CNN to predict the correct brand when an ad is presented. We argue that this ability to guess correctly is connected with the uniqueness of brand positionings and the prediction of other brands describes the brand confusion potential. As already mentioned, this approach is closely connected to the work by Baier and Frost (2018) but in contrast to the approach used there we apply a CNN and are also able to analyze large numbers of ads and video clips.

The rest of this paper is organized as follows. Section 2 reviews traditional approaches for predicting brand confusion in imagery markets including the work by Baier and Frost (2018). Then, Sect. 3 discusses the new approach for predicting brand confusion based on ads with visual content. Section 4 presents an application to the German beer market based on ads with photos, Sect. 5 an application based on video clips. Finally, Sect. 6 summarizes the paper and gives an outlook.

2 Traditional approaches for predicting imagery brand confusion

2.1 Brand confusion experiments

The standard approach to control unique brand positionings in imagery brand markets consists in conducting a so-called brand confusion experiment (see, e.g., Keon 1983, 1984; Böckenholt and Gaul 1985; Gaul and Baier 1994; Kroeber-Riel and Esch 2015): A sample of consumers is exposed to a sample of ads (e.g. print ads or video clips with masked brand names and slogans) and asked to guess the advertised brand. Brands whose ads show a high percentage of correct guesses are assumed to have a strong brand positioning respectively a low brand confusion potential. Brands whose ads show a low percentage of correct guesses are assumed to have a weak brand positioning, the wrongly guessed brands further detail this brand's high confusion potential. Here, the risk is high that investments in ads support the other brands. The results of data collection are typically summarized in a so-called confusion matrix (ads or advertised brands as rows, guessed brands as columns) which contains for each ad or advertised brand in the rows the percentage of guessed brands in the columns and forms the basis for further analyses.

So, Keon (1983) presented the TRINODAL mapping approach that can be used to visualize the relative positioning of ads, brands, and consumers in a low-dimensional space. Multidimensional unfolding is applied to the brand confusion matrix in order to estimate the point coordinates. Euclidean distances between an ad point and all brand points inversely reflect the corresponding percentage or number of guesses when the ad was presented. Brands with unique positionings are located close to their ad points

and far away from other brand points. Brand points which are closely located together reflect a high confusion potential between them (Keon 1984). In the map, consumer points are integrated which reflect ideal brand positionings for the sample of consumers which are derived from their additionally collected preference evaluations in the brand confusion experiment. Using such mappings, unique brand positionings and brand confusion can be easily judged. The TRINODAL mapping approach was extended by several authors. So, e.g., Nishisato and Gaul (1990) proposed and applied dual scaling and its forced classification procedure, Zielman and Heiser (1993), Okada and Imaizumi (1997), and Chino (2002) developed multidimensional scaling of asymmetric similarity matrices (the confusion matrix) procedures.

As an alternative, Espejo and Gaul (1986) proposed the application of two-mode hierarchical clustering. They embedded the confusion matrix into a so-called grand matrix, a symmetric proximity matrix with respect to the superset of ads and brands (the two modes). The confusion matrix delivers the identical similarity values between brands and ads as well as ads and brands. The other similarity values (between ads and ads as well as brands and brands) are declared as missing. Hierarchical clustering algorithms that can deal with missing values derive two-mode clusters (with ads and brands) which can be used to discuss unique brand positionings (if a brand forms a cluster with its ads) and brand confusion (if two or more brands form a cluster with their ads and/or ads from other brands). Sample applications—also by other authors and with alternative clustering procedures—demonstrate the usefulness of this approach (see, e.g., Espejo and Gaul 1986; Gaul and Baier 1994; Mechelen et al. 2004; Rocci and Vichi 2008).

2.2 Predicting imagery brand confusion based on ad content similarities

Recently, an alternative to brand confusion experiments has been proposed (Baier et al. 2012; Baier and Frost 2018). The main assumption behind this approach is that ad similarity with respect to extracted low-level and high-level image features is a predictor for confusion among the advertised brands. Similar color distributions (e.g., high shares of blue- and green-colored pixels due to landscape motifs), similar edge distributions (e.g., many right angles in technical surroundings), and/or similar objects/subjects (e.g. buildings, cars, people) should decrease the probability that the correct brand is guessed when an ad is presented.

Baier and Frost (2018) showed in their comparison that this approach is promising. A brand confusion experiment was conducted. 446 consumers were confronted with 16 recent print ads for 16 German beer brands—for each brand one ad with brand name and slogan masked—and asked to guess the advertised brand. Using the IMADAC software (Baier et al. 2012; Baier and Frost 2018), altogether 90 featurewise distance matrices for the 16 print ads were extracted, and condensed to 20 principal components. Then, the consumers were divided into a training and a test sample and the confusion matrix for the training sample was used to estimate an assumed multinomial logit model that relates the componentwise distances between the print ad for the advertised brand and the print ads for the other brands to the guessing behavior. The trained prediction model was controlled by the confusion data from the test sample of consumers. The results were promising: By image feature extraction and similarity

calculation a predictive model could be formed that replicates the results of the brand confusion experiment.

However, recently, much progress has been made in visual content recognition based on deep learning. Here, especially, convolutional neural networks (CNNs) have become the go-to algorithm for many computer vision tasks. Consequently in the following, we discuss whether the usage of CNNs for comparing ads with visual content and predicting the uniqueness of the brand positioning and/or brand confusion is an alternative.

3 Deep learning approaches for predicting imagery brand confusion

3.1 Deep learning approaches for visual content recognition

For some time, deep learning approaches demonstrate their superiority above other machine learning algorithms when it comes to visual content recognition. Especially the ImageNet challenges (see, e.g., Krizhevsky et al. 2012; Russakovsky et al. 2015; Simonyan and Zisserman 2015; He et al. 2019) showed this superiority impressively: When the challenge is to predict the presence of objects of 1000 categories (e.g., cat, dog, car, ship, tree, flower) in millions of images (especially photos collected in the internet), then, a special form of deep learning approaches, the so-called convolutional neural network (CNN) clearly outperforms all other machine learning algorithms.

A CNN for object recognition in images is a regularized multilayer neural network, where the lowest (or first) layer has the two-dimensional arrays of image values (binary, gray, or color values) as input. Then, from layer to layer the values of near-by neurons of one layer are weighted according to a filtering system (convolved) and passed to selected neurons of the next higher layer. The restriction on selected connections of neurons follows the concept of receptive fields from visual biology but also reduces the number of parameters. The highest (or final) layer then has the category indicators as output (LeCun et al. 1990).

For calibrating a CNN with its—even with these restrictions—still large number of parameters, typically millions of images with known categories are needed ("ground truth"). A loss function calculates the difference between the predicted category indication on the basis of current parameter estimates and the true category indication for a training sample of images. The parameters of the CNN are learned such that this loss function is minimized, i.e., the difference is as small as possible, or below a selected threshold using back-propagation. After calibration of the parameters, the CNN is expected to correctly predict category indication also for images of a test sample but also for images where no ground truth is available (see, e.g., Chollet and Allaire 2018, for more details). How many layers contribute to a network is called the depth of the network. Modern CNNs often involve tens or even hundreds of successive layers and, typically, huge numbers of parameters (weights for the connections, thesholds) that have to be learned and require millions of images for calibration.

However, in order to reduce the needed images with ground truth for training, the usage of pre-trained CNNs (e.g., AlexNet, GoogLeNet, VGG16) is state-of-the-art. So, He et al. (2019)—as many other authors—argue that feature representations (i.e.

architecture of the network, trained parameters of the lower layers) learned in another context can transfer useful information to a target task. They especially refer to using pre-trained CNNs from the ImageNet challenges (see, e.g., Krizhevsky et al. 2012; Russakovsky et al. 2015; Simonyan and Zisserman 2015) as a helpful starting point when solving related visual content recognition tasks.

Here, the work of Zhou et al. (2017) is a good reference for the task here: Zhou et al. (2017) were interested in CNNs for scene recognition and used pre-trained CNNs for object recognition as starting points. The main difference between the object recognition task (i.e., indicating whether a cat, dog, car, ship, tree, flower and so on is on the image as in the ImageNet challenges) and the scene recognition task (i.e., indicating categories like "teenage bedroom", "messy kitchen", "darkest forest-path", "green mountains", "sunny coast") is that scenes are not directly connected to one or more specific objects on the image but more or less describe the place or context of the image as a whole.

Following prior work of Xiao et al. (2010, 2016) on traditional approaches for scene recognition (feature extraction, modeling), Zhou et al. (2017) developed about 900 scene categories by completing phrases like "I am in …" or "Let's go to …" with words from the WordNet dictionary (Miller 1995). Concrete names of destinations (e.g. "New York", "Germany") but also too general terms (e.g. "workplace", "outdoors") were not allowed as scene categories. Moreover, scene categories also should use a substantive and an adjective. Then, using the so generated two-words scene descriptions, they collected millions of images using image search engines (Google Images, Bing Images, Flickr) and controlled the categorizations of the images using human scene recognition via Amazon Mechanical Turk. After further cleaning of categorized images (e.g. by merging confusable scene categories), more than 10 million images with indications of 434 scene categories were left and could be used to train a CNN. For this purpose, they started—as already mentioned—with pre-trained CNNs from the ImageNet challenge (e.g., AlexNet, GoogLeNet, VGG16), kept the architecture and trained parameters of the lower layers, but used their scene detection datasets for training the parameters of the higher layers. The performance of the so trained CNN was impressing, especially the VGG16 architecture (based on Simonyan and Zisserman 2015) showed a high accuracy even when only 50 images per scene category were used for training. VGG16 is the abbreviation for its developers, the "Visual Geometry Group" from Oxford, and the number of layers, 16.

3.2 Approaches for predicting imagery brand confusion based on ad contents

When developing deep learning approaches for judging the uniqueness of brand positionings, first, one would have to understand why visual content (e.g., photos and video clips) in ads are so important. As already discussed, the physico-chemical differences between competing products are low and the intended positioning must rely—at least partially—on campaigns that consistently connect the brand with desirable emotions and values by repeatedly showing similar scenes and objects in connection with the brand. In this context, usually, the so-called dual coding theory (Paivio 1971) is referred to. This theory assumes that humans store verbal and visual content in different parts

of their long-term memory and that the storage and retrieval of visual content is much more effective. In a long series of experiments, starting with Paivio and Csapo (1973), it could be demonstrated that positionings communicated by photos of medium complexity (e.g. objects, scenes) can be learned much easier by consumers than positionings communicated by slogans or other verbal content. Burns et al. (1993) summarize as reasons for this higher effectiveness of the visual content that the consumers more easily can store them and reload them into their working memory, that verbal content is better anchored in the consumers' experience base, and that they are experiential rather than discursive. Burns et al. (1993) also argue that these three reasons lead to a higher robustness and durability of the positioning effects compared to learning via verbal content (e.g. slogans, explanations).

Producers in imagery markets are well aware of these advantages and therefore use ads with visual content (e.g. photos or video clips) to communicate their positioning and to connect their brand with emotions and values (see, e.g., Kroeber-Riel and Esch 2015, p. 245ff.). So, e.g., since the mid 1990s, the German beer brand "Krombacher" positions their brand as "natural" and for this purpose uses—across all channels—ads where a blue lake surrounded by green trees is shown (see, e.g., Esch 2013, p. 574ff.). For German consumers, this often and consistently—across all channels—presented landscape motif is closely connected to the brand "Krombacher" and its intended positioning (Baier and Frost 2018). This even holds when the consumers are not able to read and/or hear the brand name and/or the brand slogan ("Eine Perle der Natur." which means "A pearl of nature.") during ad exposure.

Against this background—that the learning based on visual content is most important when evaluating positionings—the basic idea of our new deep learning approach for predicting unique positionings and/or brand confusion in an imagery market can be proposed with the following four steps:

1. Collect the names of all relevant brands in the imagery market.
2. Collect ads with visual content for these brands. In all ads, mask the brand name and the brand slogan (as usually done in brand confusion experiments, see, Baier and Frost 2018). Split the ads into a training sample and a test sample.
3. Calibrate a CNN to predict the advertised brand based on an ad using the training sample.
4. Use the CNN to predict the advertised brand using the test sample. Analyze the resulting confusion matrix (advertised brands times guessed/predicted brands) in the same way as in a brand confusion experiment.

Step 1 and step 2 are quite similar to the usual steps when preparing a brand confusion experiment (see, e.g., Keon 1983, 1984; Böckenholt and Gaul 1985; Gaul and Baier 1994). However, a major difference consists in the possibility (and the need) not only to collect one ad with visual content for each brand since there is no need to present the ads to the consumers in an experiment (for guessing the advertised brand). Instead, the possibility (and the need) is to collect as many ads as possible across all channels (e.g., TV, cinema, magazine, newspaper, search engines, social networks, photo- and video-sharing platforms) for training a CNN. This possibility offers the advantage to judge the positionings across all channels and—also—to judge whether the campaigns are consistent over time. As Kocyigit and Ringle (2011) have shown in

their survey on detergent brands, brand confusion and its effects on satisfaction can be theoretically explained and measured using brand similarity and brand credibility (reverse coded) but also brand diversity, brand clarity (reverse coded) and brand continuity (reverse coded) items. In our later applications we try to collect all published ads with visual content for this purpose. To make photos and video clips comparable for the CNN, we extract photos from the video clips in a standardized fashion. As in a brand confusion experiment, however, the texts (brand names, slogans, additional info) in the ads have to be pixeled out (by reducing the resolution of the pictures or by using dedicated apps for this purpose).

Step 3 simulates the human learning of positionings by calibrating a CNN. As discussed in the previous subsection, pre-trained CNNs (e.g., AlexNet, GoogLeNet, VGG16) are state-of-the-art when tasks have to be performed that are closely related to object recognition and/or scene recognition. This is also the case when analyzing ads: So, e.g., when the consumers in the experiment from Baier and Frost (2018) or in other experiments (e.g., Kroeber-Riel and Esch 2015) were asked for the information they used to guess the advertised brand, most often they mentioned objects (e.g., bottles, glasses, buildings, young women, animals) and scenes (lake with trees, park with trees, beach with dunes, party) but also dominant colors (blue-green, golden, black/dark), which are closely related recognition tasks as in the ImageNet challenge for object recognition (see, e.g., Krizhevsky et al. 2012; Russakovsky et al. 2015; Simonyan and Zisserman 2015; He et al. 2019) or in Zhou et al. (2017) for scene recognition. However, in contrast to these tasks, the brand guessing task seems to be a combination of the other tasks. In our approach we solve this task also by using a pre-trained CNN using VGG16, the winner in the Zhou et al. 2017 comparison, and re-training based on the ads and their advertised brands.

In all our analyses we used a randomly initialized fully-connected network on top of the pre-trained CNN (VGG16). Since the large gradient updates triggered by the randomly initialized layers would be disruptive to the already-learned features (Chollet and Allaire 2018), we trained the top-level classifier first, and only then started fine-tuning the weights alongside it. Fine-tuning was performed with a very slow learning rate. Hence, the magnitude of the updates should stay very small. The previously learned features would have been broken if large updates were used. We typically used a stochastic gradient descent (SGD) optimizer rather than an adaptative learning rate optimizer such as RMSProp and Adam. The SGD optimizer includes support for momentum, learning rate decay, and Nesterov momentum. The RMSProp optimizer divided the learning rate for a weight by a running average of the magnitudes of recent gradients for that weight. This was the mini-batch version of using the sign of the gradient and is usually a good choice for recurrent neural networks. Adam is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. This was straightforward to implement, computationally efficient, had small memory requirements, was invariant to diagonal rescaling of the gradients, and well suited to problems that are large in terms of data and parameters. It is also appropriate for non-stationary objectives and problems with very noisy and sparse gradients. After instantiating the VGG base and loading its weights, we added our previously trained fully-connected classifier on top.

We then proceeded to freeze all layers up to the last block. Finally, we started training the entire network with a very slow learning rate.

After training this CNN on the training sample of ads (using a train -validation split for preventing overfitting), in step 4, the CNN was applied on the test sample which resulted—averaged across all ads of the test sample that belonged to a specific brand—to a confusion matrix similar to the results of a brand confusion experiment: The results are summarized in a matrix (advertised brands as rows, guessed brands as columns) which contains row by row for each advertised brand in each column the percentage of guesses in favor of a brand. Advertised brands where the percentage of guesses is near 100% for the brand itself are assumed to have a unique positioning, advertised brands where the percentages of guesses are shared among some brands have a high brand confusion potential.

4 Application to the German beer market based on ads with photos

Pils, a type of pale lager, is the most popular beer in Germany. 9.17 million Germans drink regularly (more than one time a week) this type of beer. The market is dominated by premium producers, e.g., "Krombacher" (4.42 million hectoliters Pils in 2017), "Bitburger" (3.18), "Veltins" (2.22), "Warsteiner" (1.92), "Hasseröder" (1.89), "Radeberger" (1.84), and "Beck's" (1.77) (Statista 2020).

As many breweries use the same product and production technology, physicochemical differences are moderate. So, e.g., in a current product comparison of 42 German breweries (including the premium producers) more than half of the brands received excellent grades, differences could only be found with respect to eco-friendly packaging, differences with respect to taste were not communicated since they could only be traced back to suboptimal storage and transport. The premium producers heavily rely on ads with visual contents to position their brands. Taste or quality is not communicated in these advertisements. Figure 1 shows print ads for 16 of these premium brands later used in the application, the same as used in Baier and Frost (2018) to understand which kind of information is given in these advertisements.

So, e.g., the market leader—"Krombacher"—connects the brand with a blue lake and green trees to position the brand as "natural", "Radeberger" demonstrates "exclusivity" by showing the Semper opera in Dresden, the location where the brewery was founded near-by, "Beck's" uses a green bottle on an ice bucket to demonstrate "freshness". The shown ads are only samples for ads with similar visual elements for this positioning purpose. Many other photos and video clips are spread by the producers across a variety of channels (e.g., TV, cinema, newspaper, magazine, internet) in order to strengthen their brand's positionings.

For evaluating unique brand positionings and brand confusion in this market with our proposed approach, we collected for the 16 brands in Fig. 1 (step 1 of our approach) recent ads with visual content (photos, video clips). Here, the AdVision Digital archive of advertisements (AdVisionDigital 2020) was very helpful. In this database, the ad campaigns of most German brands including the print ads, TV spots, cinema ads, out-of-home displays, info screens, radio spots, internet banners used since 1999 are made available for researchers and marketeers. Within this database, many ads with

Fig. 1 Sample ads with visual content (not masked) for 16 German Pils beer brands (one image for each brand). The advertised brands are column by column, top to bottom: "Beck's", "Berliner Kindl", "Berliner Pilsner", "Bitburger" (first column), Hasseröder", "Holsten", "Jever", "Karlsberg", "König Pilsener", "Krombacher", "Licher", "Lübzer", "Radeberger", "Veltins", "Warsteiner", "Wernesgrüner" (fourth column)

visual content (photos) could be found for each of the 16 brands that were used in campaigns during the last five years. For most brands about 50 ads with visual contents were available, some brands even had up to 80-90 of them. However, since some ads showed similar content (e.g. when a campaign was used in a consumer magazine and a out-of-home poster campaign), an evenly distributed selection was made with a similar number of ads for all brands, resulting in 817 ads in total (see Fig. 1 with one ad per brand). The number of ads is small but comparable to the scene recognition task by Zhou et al. (2017). Also, it should be mentioned that these numbers are huge compared to the usual number of ads used in a brand confusion experiment (e.g., one ad per brand in Baier and Frost 2018).

The found ads showed a considerable variation: So, e.g., the "Krombacher" photos consistently showed the lake and the green trees—at least in parts of the background—

but also, e.g., many different bottle and glass variations. The "Radeberger" photos often showed the opera building and a glass with beer or a bottle, but also, e.g., other buildings or buildings from the inside (opera, dancing), however often in similar color settings. The "Beck's" photos often showed the green bottles, groups of persons drinking beer, or ships with green sails (from former campaigns, still used).

For calibrating the CNN (step 3) the 817 images were splitted randomly in 80% to be used as a training sample (n=653) and 20% as a test sample (n=164). R calling the Python packages Keras and TensorFlow was applied for training the CNN (Chollet and Allaire 2018). We used a VGG16 pre-trained CNN with a fine-tuned final block alongside the top-level classifier. All images were rescaled by 1/255 and resized to 64x64 pixels. The small resolution of the images was selected to ensure that the texts on the images couldn't be read by the CNN. So, a separate pixeling out or masking of the brand names and slogans seemed not to be necessary. The number of epochs and batch size were set to 40 and 20, respectively. One epoch means that an entire dataset is passed forward and backwards through the neural network once. Training data were divided into a number of batches. In order to mitigate overfitting, data augmentation and a 50% drop-out rate were implemented. Data augmentation means that the training images are automatically modified by random transformations (rotations, height shifts, width shifts, zooming, flipping) when used as input during the batches and epochs so that during the learning process the CNN never sees the same picture twice. Dropout randomly sets output features to 0 and so additionally prevents the CNN from just modeling the training data (overfitting). A validation split of the training sample with 70% of the images for calibration ($n = 465$) and 30% for validation ($n = 188$) supports these mitigation issues.

Figure 2 shows the development of loss (normalized deviations between predicted and true brand indications) and accuracy (percentage of correct brand predictions) during the training process (number of epoches performed). As can be easily seen, the CNN is able to predict / guess the advertised brand in many cases: In the calibration subsample of the training data ($n = 465$) the accuracy is 92.05%, in the validation sample ($n = 188$) the accuracy is 87.23%. Please note, that the validation split especially was used to determine the number of epochs to mitigate overfitting. The (hold out) test sample ($n = 164$) supported this selection: There, also an accuracy of 90.8% could be observed. Please note that the training sample (including the validation split) used data augmentation whereas the test sample used the original images (and therefore has an improved accuracy compared to the validation sample).

Also, in order to check whether the CNN is really able to predict the brand according to relevant image content, the Gradient weighted Class Activation Map (Grad-CAM) was used. Grad-CAM was proposed by Selvaraju et al. (2017) to identify the regions of the image, where the trained model receives the main contributions to its predictions. The idea is to use the class-specific gradient information flowing into the final layer of a CNN to produce a coarse localization map of the important regions and is a generalization of the Class Activation Map (CAM) by Zhou et al. (2016). Zhou et al. (2016) revisited the global average pooling layer and shed light on how it explicitly enables the CNN to achieve remarkable localization, despite being trained on image-level labels. Unlike CAM, Grad-CAM requires no re-training and is broadly applicable to any CNN-based architecture.

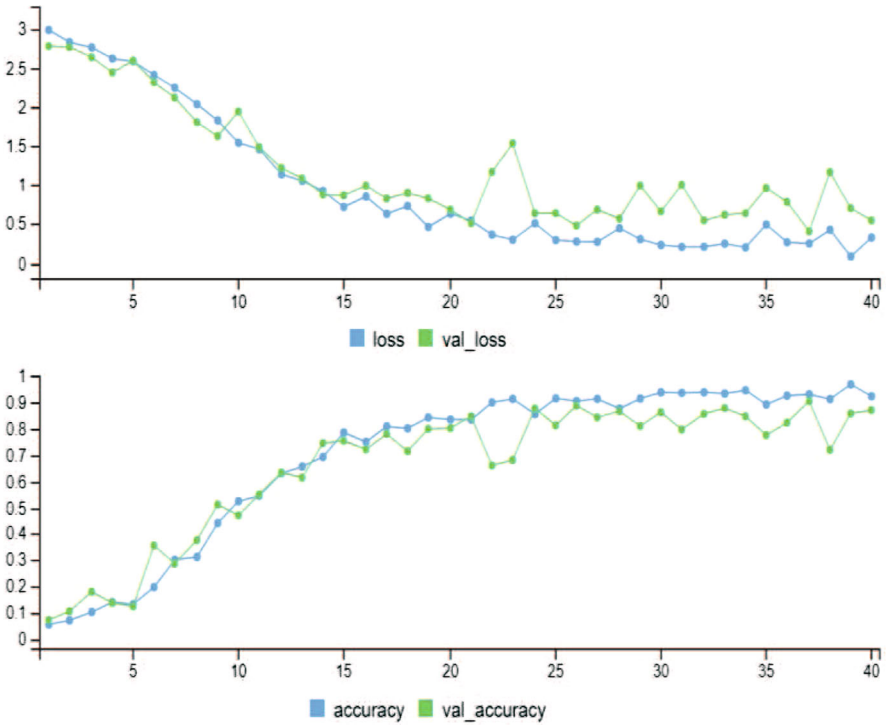

Fig. 2 Development of loss (normalized deviations between predicted and true brand indications) and accuracy (percentage of correct brand predictions) during the training process (number of epoches performed, i.e. complete runs through the training data) according to our analysis of the 16 beer brands based on 465 images for training (blue line) and 188 images for validation (green line). The loss function measures 164 images for test—without data augmentation—resulted in an accuracy of 90.8%

Figure 3 shows the results obtained from Grad-CAM when relating the predicted class (brand) to a lower CNN layer (in our case the block3_conv3 layer with a resolution of 16 × 16 due to an image input of 64 × 64). The left images show the ads in resolution 64 × 64. Each column contains the first five ads to one brand (here, from left to right: "Berliner Kindl", "Berliner Pilsner", "Krombacher", "Lübzer", "Radeberger"). It can be easily seen that reading texts in this resolution is difficult, therefore no additional pixeling out of the texts was done. The right images show the same ads overlaid by the heatmap generated by Grad-CAM. White areas have highest activation followed by yellow and green areas. The "seen" regions when predicting "Berliner Kindl" are the large faces and the glass, when predicting "Krombacher" its the background with the lake and the trees, when predicting "Lübzer" the lighthouse and/or the similar glass, when predicting "Radeberger" the bottle and the glass. Of course, the regional activation is not the sole contributor to the class prediction, also e.g. color distributions are important, but it seems to be clear that the brand prediction are based on "meaningful" contents.

This ability to predict now can be used to calculate confusion matrices: For the ads in the test sample, the brand guess probabilities are calculated and aggregated across the

Fig. 3 The first five images (ads) for five brands (column by column) in 64x64 resolution on the left and—in the same order—with superimposed class activation map heatmap according to Grad-CAM on the right. White areas have highest activation followed by yellow and green areas

advertised brands. Table 1 shows the resulting confusion matrix. Here, the diagonal values reflects the percentage of "correct" predictions by the CNN which—as we already now—is 90.8% across all ads. Especially high (99 or 100%) is this percentage for brands like "Berliner Pilsner", "Hasseröder", "Holsten", "Krombacher", "Licher", and "Veltins", low, e.g., for brands like "Beck's", "Biburger", "Berliner Kindl", and "Warsteiner". A closer look into these ads reflects the problem (as partially discussed above): These brands use ads with high content variations or which have close content relations to ads of other brands. So, e.g., "Berliner Kindl" and "Holsten" use similar layouts (grey colors, large faces, brown glass with beer). However, since "Holsten" is more consistent in this direction (similar content across all ads) its positioning in this direction is stronger and "Berliner Kindl" ads are assumed to advertise for "Holsten". A similar argument holds for "Beck's", a brand that changed its ad content recently and therefore has many variations in the training and test sample. In the next section we will check whether we can find similar results when analyzing video clips.

5 Application to the German beer market based on video clips

Besides the already analyzed print ads and out-of-home-displays, we now analyze the collected video clips from the AdVision Digital archive of advertisements (AdVisionDigital 2020). For each brand we selected one video clip and used the app DiaShow11 to extract about 500 frames as images from each clip. Similar as in the previous section we divided the sample of 7974 images into a training sample (4471 for training, 1907 for validation/mitigating overfitting) and a test sample (1596 images). We used the same CNN based on VGG16 and the same resolution for image input (64x64) for calibration and prediction.

Figure 4 shows, respectively, the loss and accuracy according to our analysis of the beer brand images extracted as frames from video clips. Again, the blue line

Table 1 Predicting brand confusion based on the calibrated CNN: Predictions are made using the 164 test images (ads)

Advertised brand		Predicted brand															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Beck's	800	3	2	0	0	94	1	1	2	0	75	0	0	17	3	2
2	Berliner Kindl	0	821	0	0	0	92	0	0	0	0	0	85	0	0	0	0
3	Berliner Pilsner	0	0	999	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Bitburger	0	0	0	803	0	0	0	0	80	0	0	0	2	0	2	112
5	Hasseröder	0	0	0	0	1000	0	0	0	0	0	0	0	0	0	0	0
6	Holsten	0	0	0	0	0	1000	0	0	0	0	0	0	0	0	0	0
7	Jever	1	0	0	0	0	1	889	0	93	12	0	1	0	0	1	1
8	Karlsberg	0	0	0	0	0	0	0	909	0	0	1	0	0	89	0	0
9	König Pilsener	1	4	1	47	1	2	4	0	926	0	0	3	3	0	5	3
10	Krombacher	0	0	0	0	0	0	0	0	0	1000	0	0	0	0	0	0
11	Licher	0	0	0	0	0	0	0	0	0	0	1000	0	0	0	0	0
12	Lübzer	3	1	1	1	0	2	26	0	18	1	0	945	0	0	0	1
13	Radeberger	1	2	4	3	1	6	14	5	283	4	33	19	398	3	214	9
14	Veltins	0	0	0	0	0	0	0	0	0	0	0	0	0	1000	0	0
15	Warsteiner	0	0	0	0	0	165	0	0	1	0	0	0	0	0	834	0
16	Wernesgrüner	0	0	0	1	0	0	111	0	0	0	0	0	0	0	0	888

Each row contains the predicted probability that the brand in the column is guessed when an ad for the brand in the row is shown (averaged values across all ads for the same brand). Probabilities are given as per mille values (have to be divided by 1000)

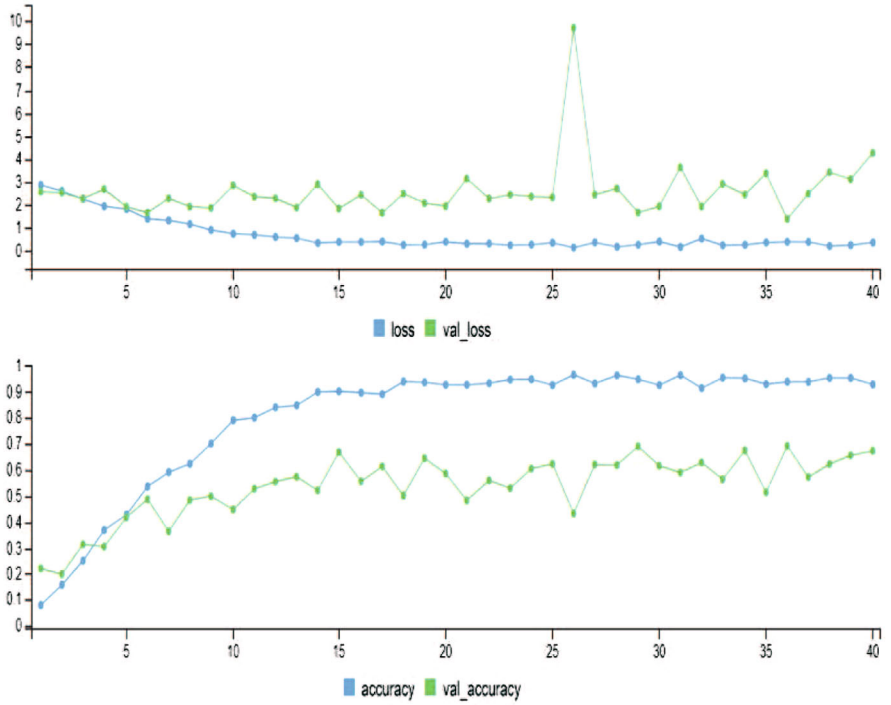

Fig. 4 Development of loss (normalized deviations between predicted and true brand indications) and accuracy (percentage of correct brand predictions) during the training process (number of epoches, i.e. runs through the data) according to our analysis of the 16 beer brands based on 4471 frames for training (blue line) and 1907 for validation (green line). 1596 frames for test—without data augmentation—resulted in an accuracy of 91.0%

shows the results obtained from the calibration subsample of the training data (92.9% accuracy) and the green line shows the results obtained from the validation subsample of the training data (67.7% accuracy). Please note that the training data used the same data augmentation and drop-out procedure as in the previous section to prevent from overfitting. The accuracy with respect to the test data (1596 images not used up to this point) was 91.0%. It should be mentioned that the prediction here is much more difficult than in the previous section since the frames from the video clips are drawn equidistantly which means that they also contain the introduction of the clip and parts where the typical elements of an ad (landscape, objects, colors) are not present. Also, well-known brands like "Krombacher" only use short fractions of their video clips to show their typical elements (in the video clip here at the end).

Nevertheless, also in this case it is possible to calculate confusion matrices with respect to the test data as given in Table 2. Again, one can easily see that the correct brands are predicted with high probabilities. One can easily detect similarities between the confusion matrices in Tables 1 and 2. "Beck's" has a weak positioning (only 80.9% resp. 86.9% guesses of the "Beck's" when visual content for "Beck's" is shown) and there is a high confusion in direction of the brand "Holsten". Also in the video clip,

Table 2 Predicting brand confusion based on the calibrated CNN: Predictions are made using the 1596 extracted test frames

Advertised brand		Predicted brand															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Beck's	869	8	3	0	1	45	16	4	1	1	0	18	1	5	26	3
2	Berliner Kindl	0	998	0	0	0	0	0	0	0	0	0	0	0	0	2	0
3	Berliner Pilsner	0	0	899	0	0	0	0	101	0	0	0	0	0	0	0	0
4	Bitburger	93	4	5	735	59	8	2	3	15	2	1	21	4	5	5	37
5	Hasseröder	0	0	84	0	916	0	0	0	0	0	0	0	0	0	0	0
6	Holsten	0	0	0	0	0	990	0	0	0	0	0	10	0	0	0	0
7	Jever	0	0	0	0	0	0	1000	0	0	0	0	0	0	0	0	0
8	Karlsberg	6	0	0	0	0	0	1	927	0	10	50	0	0	0	0	5
9	König Pilsener	1	0	10	0	1	0	0	0	949	0	0	35	0	1	1	1
10	Krombacher	0	0	0	0	51	0	0	0	0	939	0	0	0	1	9	0
11	Licher	0	0	0	0	0	0	0	0	0	0	1000	0	0	0	0	0
12	Lübzer	1	0	0	0	0	0	29	0	0	0	0	970	0	0	0	0
13	Radeberger	34	2	3	25	6	6	1	1	25	1	0	61	771	5	12	47
14	Veltins	0	14	0	0	0	31	0	0	0	0	0	0	0	955	0	0
15	Warsteiner	1	0	0	0	0	0	0	0	2	0	0	3	0	1	992	0
16	Wernesgrüner	6	144	5	7	10	14	1	1	60	1	0	10	3	15	13	712

Each row contains the predicted probability that the brand in the column is guessed when a frame for the brand in the row is shown (averaged values across all ads for the same brand). Probabilities are given as per mille values (have to be divided by 1000)

"Beck's" has a high variation of visual content which explains this confusion very well. The same holds, e.g., for the brands "Bitburger", "Karlsberg", "Radeberger", and "Wernesgrüner": As well as for the ads with photos as for the frames extracted from the video clips, the CNN has problems to predict the correct brand. A closer look here also shows the overall high variation of visual content along the video clips. This even holds for "Krombacher", a brand which was easily to be guessed by the CNN when analyzing the images: With respect to the extracted frames only with 93.9% probability the correct brand was predicted. As already mentioned, the short parts of the video clip where the "typical" "Krombacher" content was shown is the reason. This is an interesting reduction of accuracy compared to the confusion matrix in Table 1. But also differences in the other direction exist: "Berliner Kindl" has a much more consistent video clip compared to its variation in the ads with photos (that can be easily validated when showing the video clip): The frames extracted from the video clip led in 99.8% of the guesses to the correct brand, whereas this was only the case in 82.1% of the cases when an ad with photo was used as input.

6 Conclusion and outlook

In imagery markets, ads with visual content are very important to connect a brand with desirable emotions and values. However, when the ads of competing brands have similar visual contents, the ad investment can be wasted due to brand confusion. The traditional approach to control the confusion potential—the brand confusion experiment—has many pitfalls: So, e.g., only few ads can be taken into account during data collection, and—typically—only print ads are analyzed.

Here, a new approach is proposed: ads with visual content (e.g., photos and video clips) can be collected across all channels and analyzed. A CNN has to be trained basing on a pre-trained CNN (e.g., VGG16 from the ImageNet challenge) to check whether the brand currently has a unique positioning or exhibits brand confusion potential. The main idea behind this new approach is to check whether the CNN can be trained to predict the correct brand. If this is possible, the ads may lead to a unique positioning. If this is not possible, the potential for brand confusion is high.

In our application to the German premium Pils beer market, it could be shown, that photos and video clips can be collected and analyzed with the new approach. The calibrated CNN showed overall a good predictive accuracy but was also able to demonstrate that some ads and brands have—as expected—potential for brand confusion.

Of course, the analysis must be extended in the future. Other imagery markets should be analyzed and also alternative pre-tuned CNNs would be helpful. Also, it would be nice to additionally analyze photos and video clips posted by consumers not only photos and video clips posted by the producer or to make cross-validations, e.g., between CNNs trained on images and CNNs trained on frames extracted from video clips. Also, at the moment, video clips are only analyzed frame by frame.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number JP20K01963, JP16K00052. Also, we wish to thank the two reviewers and the editors for their valuable hints for improvement.

References

AdVisionDigital (2020) AdZyklopädie: Deutschlands größtes Online-Archiv für Werbung. Tech. rep., Hamburg, Germany, https://v2.adzyklopaedie.com

Almenberg J, Dreber A, Goldstein R (2014) Hide the label, hide the difference? Tech. rep., American Association of Wine Economists, Working Paper no. 165, New York, https://www.wine-economics.org/list-of-aawe-working-papers/

Baier D, Frost S (2018) Relating brand confusion to ad similarities and brand strengths through image data analysis and classification. Adv Data Anal Classif 12(1):155–171

Baier D, Gaul W (1998) Optimal product positioning based on paired comparison data. J Econ 89(1):365–392

Baier D, Daniel I, Frost S, Naundorf R (2012) Image data analysis and classification in marketing. Adv Data Anal Classif 6(4):253–276

Böckenholt I, Gaul W (1985) Zur mehrdimensionalen Analyse von Bildinformationen in Anzeigen für Imagery-Produkte. Vierteljahreshefte für Mediaplanung 4:20–29

Burns AC, Biswas A, Babin LA (1993) The operation of visual imagery as a mediator of advertising effects. J Advert 22(2):71–85

Chino N (2002) Complex space models for the analysis of asymmetry. In: Nishisato S, Baba Y, Bozdogan H, Kanefuji K (eds) Measurement and multivariate analysis. Springer Japan, Tokyo, pp 107–114

Chollet F, Allaire JJ (2018) Deep learning with R, 1st edn. Manning Publications Co., USA

Dissanayake R, Amarasuriya T (2015) Role of brand identity in developing global brands: a literature based review on case comparison between apple iphone vs samsung smartphone brands. Res J Bus Manag pp 430–440

Esch FR (2013) Moderne Markenführung: Grundlagen - Innovative Ansätze - Praktische Umsetzungen. Springer, Heidelberg

Espejo E, Gaul W (1986) Two-mode hierarchical clustering as an instrument for marketing research. In: Gaul W, Schader M (eds) Classification as a Tool of Research. Elsevier, Amsterdam, pp 121–128

Gaul W, Baier D (1994) Marktforschung und marketing-management: computerbasierte Entscheidungsunterstützung; Buch mit Diskette, 2nd edn. Oldenbourg, München

He K, Girshick R, Dollár P (2019) Rethinking imagenet pre-training. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4918–4927

Keon JW (1983) Product positioning: Trinodal mapping of brand images, ad images, and consumer preference. J Mark Res 20(4):380–392

Keon JW (1984) Copy testing ads for imagery products. J Adv Res 23(6):41–48

Kocyigit O, Ringle CM (2011) The impact of brand confusion on sustainable brand satisfaction and private label proneness: A subtle decay of brand equity. J Brand Manag 19(3):195–212

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

Kroeber-Riel W, Esch FR (2015) Strategie und Technik der Werbung: Verhaltenswissenschaftliche und neurowissenschaftliche Erkenntnisse, 5th edn. Kohlhammer, Stuttgart

LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, Jackel LD (1990) Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems, pp 396–404

Madsen ES (2017) Branding and Performance in the Global Beer Market. Economics Working Papers 2017-11, Department of economics and business economics, Aarhus University

Mechelen I, Bock HH, De Boeck P (2004) Two-mode clustering methods: A structured overview. Stat Methods Med Res 13:363–94

Miller GA (1995) Wordnet: a lexical database for english. Commun ACM 38(11):39–41

Nishisato S, Gaul W (1990) An approach to marketing data analysis: the forced classification procedure of dual scaling. J Mark Res 27(3):354–360

Okada A, Imaizumi T (1997) Asymmetric multidimensional scaling of two-mode three-way proximities. J Classif 14(2):195–224

Paivio A (1971) Imagery and verbal processes. Holt, Rinehart, and Winston. New York

Paivio A, Csapo K (1973) Picture superiority in free recall: Imagery or dual coding? Cogn Psychol 5(2):176–206

Rocci R, Vichi M (2008) Two-mode multi-partitioning. Comput Stat Data Anal 52:1984–2003

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115(3):211–252

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, D B (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 618–626

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arxiv: 1409.1556 accessed 26 May 2019

Statista (2020) Brauwirtschaft in Deutschland. Tech. rep., Hamburg, https://de.statista.com/themen/1490/brauereien-in-deutschland/

Stiftung Warentest (2020) Die besten Marken und Discounterprodukte. https://www.test.de/thema/marken-und-discounter

Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, pp 3485–3492

Xiao J, Ehinger KA, Hays J, Torralba A, Oliva A (2016) Sun database: exploring a large collection of scene categories. Int J Comput Vis 119(1):3–22

Zhou B, Khosla A, Oliva A L A, Torralba A (2016) Learning deep features for discriminative localization. In: IEEE conference on computer vision and pattern recognition (CVPR)

Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2017) Places: a 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1452–1464

Zielman B, Heiser W (1993) Analysis of asymmetry by a slide vector. Psychometrika 58:101–114