Noname manuscript No.
(will be inserted by the editor)

Automated Curation of Brand-related Social Media Images with Deep Learning

Ruben Tous · Mauro Gomez · Jonatan Poveda · Leonel Cruz · Otto Wust · Mouna Makni · Eduard Ayguadé

Received: date / Accepted: date

Abstract This paper presents a work consisting in using deep convolutional neural networks (CNNs) to facilitate the curation of brand-related social media images. The final goal is to facilitate searching and discovering user-generated content (UGC) with potential value for digital marketing tasks. The images are captured in real time and automatically annotated with multiple CNNs. Some of the CNNs perform generic object recognition tasks while others perform what we call visual brand identity recognition. When appropriate, we also apply object detection, usually to discover images containing logos. We report experiments with 5 real brands in which more than 1 million real images were analyzed. In order to speed-up the training of custom CNNs we applied a transfer learning strategy. We examine the impact of different configurations and derive conclusions aiming to pave the way towards systematic and optimized methodologies for automatic UGC curation.

Keywords Social Media · Instagram · Twitter · User Generated Content · Deep Learning · Marketing

1 Introduction

Nowadays, there is a growing interest in exploiting the photos that users share on social networks such as Instagram or Twitter [3][18], a part of the so-called user-generated content (UGC). A significant part of these images has potential value for digital marketing tasks. On the one hand, users' photos can

Ruben Tous, Leonel Cruz and Mouna Makni
Universitat Politècnica de Catalunya (UPC). Barcelona, Spain
E-mail: rtous@ac.upc.edu

Mauro Gomez, Jonatan Poveda and Otto Wust
Adsmurai. Barcelona, Spain

Eduard Ayguadé
Barcelona Supercomputing Center (BSC). Barcelona, Spain

The final publication is available at Springer via http://dx.doi.org/10.1007/s11042-018-5910-z

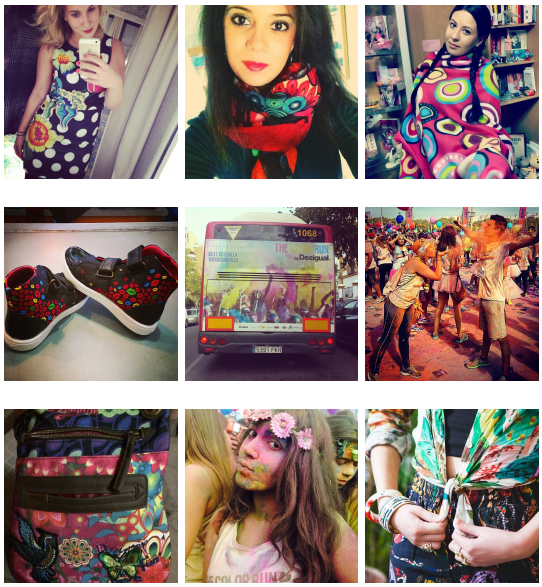

Fig. 1 Example images posted by Instagram users and tagged with Desigual's promotional hashtags (e.g. #lavidaeschula)

be analyzed to obtain knowledge about users behavior and opinions in general, or with respect to a certain products or brands. On the other hand, some users' photos can be of value themselves, as original and authentic content that can be used, upon users' permission, in the different brands' communication channels. This work is related to this second use case, searching, discovering and exploiting user-generated content (UGC) for digital marketing tasks, that has been traditionally addressed by the so-called *content curation* technologies.

Platforms for photo-centric UGC are proliferating rapidly nowadays (e.g. Olapic [12], Chute [2] and Curalate[5]), but discovering valuable images on social media streams is challenging. The potential bandwidth to analyze is huge and, while they help, user defined tags are scarce and noisy. A large part of current solutions relies on costly manual curation tasks over random samples. This way many contents are not even processed, and many valuable photos go unnoticed. Adoption of image recognition techniques in commercial UGC systems is currently very limited. In the best case, they provide generic classifiers whose categories and original training data were not specific to UGC. Often these classifiers limit to the categories of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), but the vast majority of Instagram/Twitter photos are people-centric (selfies, food, clothes, etc.) while ILSVRC is more generic (fauna, flora, etc.). An important particularity of UGC is the huge amount of *spam images*, i.e. images that, in the most usage scenarios, have no value neither as knowledge carriers nor as a exploitable content. The incapacity of detecting the multiple types of spam images limits

the usability and efficiency of existing solutions. Another difficulty of adopting image recognition techniques into UGC systems is the high computational cost of CNN-based image classifiers and object detectors. These systems need to process incoming streams of hundreds of images per second and a very volatile traffic. Any additional processing component need to be extremely efficient and scalable.

In this work, we propose an approach based on deep convolutional neural networks (CNNs) and transfer learning to minimize manual curation as much as possible and to make it more efficient. As a result, we increase the number of photos processed several orders of magnitude, we increase the quality of the resulting photos (as more photos are analyzed and only the best ones go through manual curation), we enable near real-time discovery and, last but not least, we drastically reduce the cost.

The way we do this is automatically tagging the incoming images with multiple CNNs. Some of the CNNs perform generic object recognition tasks and annotate the images with tags that describe their semantics (e.g. "beach", "car", etc.). Other CNNs perform what we call visual brand identity (VBI) recognition. Given a brand, we train a model with images that it has used in its previous marketing campaigns and that are representative of the brand's visual identity. Given a campaign for a certain brand, we use the corresponding VBI CNN to automatically pre-select images that fit the visual identity of that brand. When appropriate, we also train and apply object detectors, usually to discover images containing logos. As a final step, a human expert performs a final selection with the help of a search interface that enables expressing conditions over the images' metadata (the original ones and the ones generated by the CNNs). The expert's actions are recorded and, once they reach a certain amount (a training batch), they are used to fine-tune the corresponding VBI CNN. We report experiments with 5 real brands in which more than 1 million real images were analyzed.

The contributions of this work are summarized as follows:

- We present a combined method to automatically annotate social media images with semantic metadata that facilitate their filtering in digital marketing tasks. The proposed method integrates state-of-the-art image recognition and object detection algorithms. Image recognition is used for tagging the images with generic semantic labels and also to signal if an image fits the visual identity of a given brand. Object detection is used for signaling if an image depicts the logo of a given brand.
- We introduce the concept of visual brand identity (VBI) recognition, consisting in determining if a given image expresses and reflects the culture and character of a certain brand.
- We propose a transfer learning approach to enable the training of custom, brand-specific, classifiers on-demand efficiently. This approach also mitigates the overfitting problem related to the reduced number of training samples provided for this type of classifiers.

- We evaluate the classification accuracy, classification time and model training time of the different algorithms involved. We provide evaluation results for representative subsets of the generic image recognition classifiers, the visual brand identity recognition classifiers and the logo detectors. We report a real experiment in which more than 1 million images where captured and classified (for 5 different brands) in real-time during one month.
- Finally, we examine the impact of different configurations and derive conclusions aiming to pave the way towards systematic and optimized methodologies for automatic UGC curation.

2 Related Work

2.1 Classification and search of brand-related images in social networks

The work presented in this paper is related to recent works attempting to facilitate the classification and search of images in social networks such as Instagram and Twitter. Some works, such as [11], [13], [17], [9] or [6], also apply scene-based and object-based image recognition techniques to enrich the metadata originally present in the images in order to facilitate their processing. All latest works rely on CNNs as an underlying technique. In our case, the applied image recognition techniques, while also relying in CNNs, are tuned for content curation for digital marketing tasks. This implies new problems, such as the need to recognize more abstract categories (e.g. "mediterranean") and the need to deal with smaller datasets (e.g. brand-based image datasets). Previous works such as [10], [21] and [8] also classify social media data paying special attention to brands and products. Regarding the annotation of images with generic object categories, we reuse Google's Inception-v3 model [16], trained for the ImageNet Large Scale Visual Recognition Challenge and 1000 object categories with a top-5 error rate of 15.3%. Regarding the training of classifiers for new categories, we solved the overfitting problem related to the usage of small training sets by applying a transfer learning approach the same way Berkeley researches do in [7].

2.2 Visual brand identity recognition

As far as we know, this is the first work that addresses the automatic recognition of visual brand identity in images. In [1] researchers from Georgia Tech and Yahoo labs identified a relationship between certain visual aspects (warmth, exposure, and contrast) and a photos engagement on Flicker and Instagram. Some years before, researchers from the University of Portsmouth, UK, analyzed how wavelength hues influenced users' perception and reaction [4]. In [20], authors perform personalized (for each individual viewer) image emotion classification in social networks. In our approach, emotions are just one of the aspects to consider, as brands pay also attention to other aspects (lifestyle,

values, etc.). Regarding the training of classifiers for the recognition of visual brand identity, we also applied a transfer learning approach to avoid overfitting as we need to deal with small training sets.

2.3 Logo detection

Regarding the detection of logos, in [19] authors propose a dense histogram type feature to classify logo and non-logo image patches from the Sina Weibo platform, a Chinese microblogging site. In [14], authors propose CNN-based approach able to predict bounding boxes and class probabilities in just one evaluation, without the need to apply a sliding window. This approach is extremely fast and authors claim being able to process images at 45 frames per second. In our case, due to need to generate the detectors in an on-demand basis, we opted for a solution with worse real-time performance but that doesn't need too many resources for training (data and computation power). We use a Tensorflow implementation of OverFeat [15], a method that integrates detection, recognition, and localization within a single CNN. It provides a low fall-out (false positive rate), an acceptable recall and low detection times.

3 Outline of the system

The system that we have developed processes images for one or more marketing campaigns. Each user (usually a brand's account manager) can operate multiple campaigns simultaneously. The functionality of the system can be divided into two different stages, data acquisition and data consumption.

During data acquisition the system captures and annotates new images from social media with potential value for a given campaign, and indexes them into a database. During this stage, new images are captured in real-time, as they are published on the underlying sources (Instagram and Twitter). Descriptors of the images (including the URL pointing to the image content) are acquired using the APIs provided by these underlying sources. These APIs impose limits over the number of images that can be obtained during a certain period of time. So, processing the entire stream of images produced by a given API is not possible. APIs provide the possibility to subscribe to certain filters, such as tags or geolocation bounding boxes. These filters produce partial streams that may be overlapped. In order to capture images with potential value for a campaign, our system first needs information about geographical areas and/or hashtags that are related to the campaign (e.g. promotional hashtags such as Desigual's "#lavidaeschula" or Estrella Damm's "#mediterraneamente"). These data are used by the system to program a set of subscriptions to the underlying sources. Each subscription will produce a continuous stream of images that we call "channel". The throughput of the channels may be extremely volatile, requiring a proper scalability strategy. Once a new image is captured, it is processed by multiple deep convolutional neural networks that

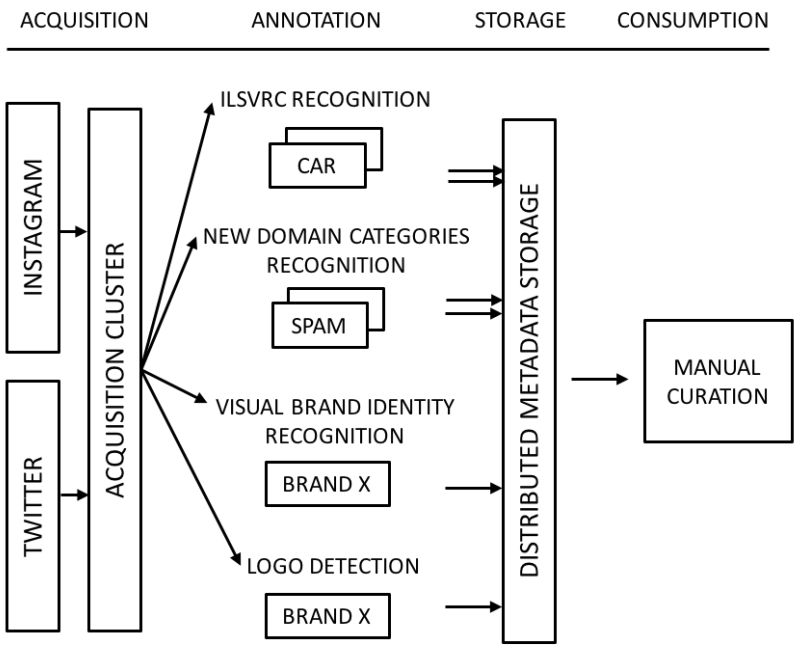

Fig. 2 Overall data flow of the sytem. Images are acquired and annotated in real-time. The resulting metadata are stored and queried by a manual curation graphical user interface.

automatically enrich the image's metadata with tags that describe their visual content (e.g. "selfie", "pizza", etc.) plus a score that measures how the image fits the visual identity of the brand.

During the data consumption stage users can navigate, search and select images from the database. Depending on the communication channel where an image is going to be used (paid ads, organic posts, images feeds, etc.) the user who post the image will be asked authorization. Both stages (acquisition and consumption) interact through the common images database, and they can occur concurrently (once images start feeding the database users can start using them). Figure 2 shows visually the overall data flow of the system.

4 Methodology

4.1 Image semantics recognition

During the acquisition stage, captured images are processed by a set of multi-class and binary image classifiers that try describing their visual content (for logo detection, explained below, we have applied a different approach). All the involved classifiers share the same architecture, Google's Inception-v3 [16], a

Original ILSVRC tag	4-depth WordNet hypernyms
Siberian husky	sled dog, working dog, dog, canine
beer bottle	bottle, vessel, container, instrumentality
red wine	wine, alcohol, beverage, food
consomme	soup, dish, nutriment, food
cowboy hat	ten-gallon hat, hat, headdress, clothing
burrito	dish, nutriment, food, substance
...	...

Table 1 Some of the 1000 ILSVRC tags and their expanded WordNet hypernyms.

deep convolutional neural network trained for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). One of the classifiers that we apply is Inception-v3 itself, which let us classify the images into 1000 different categories. As these categories are very specific terms from WordNet, we expanded ImageNet categories with their corresponding WordNet hypernyms reaching a total of more than 7,000 different tags (Table 1 shows some of them).

However, we have observed that many objects and scenes that typically appear in Instagram/Twitter images do not appear in ILSVRC. The vast majority of Instagram/Twitter photos are people-centric (selfies, food, clothes, etc.) while ILSVRC is more generic (fauna, flora, etc.). Also, even if an object or scene appears in ILSVRC often it is not part of the ILSVRC categories dictionary (i.e. WordNet). In order to provide a more comprehensive and practical set of tags, we have trained our own classifiers (more than 100, Table 3 shows some of them), retraining the last layer of Inception-v3. The most part of them are binary classifiers, that enable us to determine if an image should be annotated with a given tag (one for classifier). Notable examples are *spam* and *selfie*, tags that have proven to be very useful when searching this kind of images. The criteria of inclusion of new tags is currently heuristic, driven by the feedback of users interacting with the final curation interface. Figure 3 shows some example images from the spam dataset.

4.2 Visual brand identity (VBI) recognition

Besides annotating the incoming images with tags that describe their semantics, we have also trained a set of CNNs that perform what we call visual brand identity (VBI) recognition. Nowadays, the main course in almost all branding initiatives is to develop a unique and consistent visual brand identity that expresses and reflects the brands culture and character. A VBI may involve the preference for some colors, lighting, themes, etc. Its easy to find examples of visual identities for iconic brands such as Coca-Cola, Levis or McDonalds.

Each VBI classifier is a binary image classifier based on Inception-v3 and fine-tuned with a dataset provided by a brand. The classifier learns to distinguish which images satisfy some visual patterns found in the brand's imagery. Initially, the classifier is trained with a dataset provided by the brand (e.g.

Fig. 3 Example images from different sub-categories of the spam dataset.

a set of images used in a previous marketing campaigns and that are representative of the brand's visual identity). Later, each time a certain amount of usage actions (selection of new images by the user) have been recorded (a training batch), the model weights are updated. The only differences between the VBI classifiers and the image classifiers described in the previous section are that they operate over a special kind of custom datasets and that they are constantly fine-tuned during a batch-based online learning stage. Figure 4 and Figure 5 show some example images used to train VBI classifiers for the Estrella Damm beer brand (related to the Mediterranean lifestyle) and Pepsi, respectively.

4.3 Models training

In order to reduce overfitting, improve accuracy and reduce training times we have chosen a Transfer Learning approach for training the classifiers (for both the image semantics recognition and the VBI recognition). The method consists on fine-tuning a deep architecture already trained with millions of images on a set of traditional object recognition tasks. We start by processing each one of our training images through all the layers of the Google's Inception-v3 model [16] except the last one. For each image, we save the values of the penultimate layer of Inception (called the image *bottleneck*). Once we have computed all the bottlenecks, we replace the final layer of Inception with a new one, defined over the categories of the model that we want to train (e.g. a binary spam-detector model with just two classes). Then we run some (around

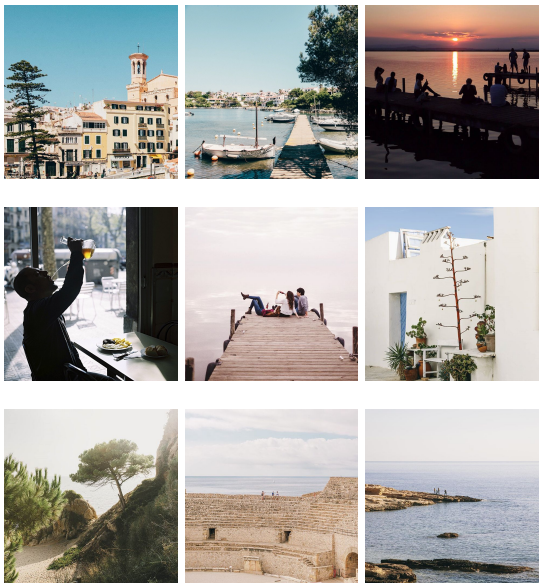

Fig. 4 Example images showing the visual brand identity of the Estrella Damm beer brand, related to the Mediterranean lifestyle

Fig. 5 Example images showing the visual brand identity of the Pepsi

4K) training steps over the network (feeding the bottlenecks directly into the final layer).

We have compared the obtained results with two other methods, a Bag of Words approach (BoW) and training our own lightweight CNNs. Regarding BoW, with the help of OpenCV we computed Opponent-SIFT descriptors from the images and clustered them to obtain a dictionary of k-dimensional visual words. With the help of the dictionary we transformed each image into a k-dimensional vector. With the obtained vectors, we trained an SVM classifier with an RBF kernel. We performed experiments with different configurations (different descriptors, different downscaling sizes, different kernels, etc.). Regarding the training of our own CNNs, we defined and trained, with the help of TensorFlow, a lightweight, 6-layer deep convolutional neural network (3 convolutional+relu layers, two fully connected layers and a softmax layer). We applied data augmentation and disabled Local Response Normalization.

4.4 Logo detection

There are some brands that are interested in images where their logo appears. In these cases, the previous image classification tasks are not enough and a custom object detector need to be trained. The requirements for these detectors are 1) that they have a low fall-out (false positive rate) and an acceptable recall, as the goal is to minimize manual curation as much as possible and missing some positives is not critical; 2) that they have a low detection time and 3) that they don't need too many resources for training (data and computation power), as the detectors need to be generated in an on-demand basis. As it provides a good trade-off of these requirements, we use a Tensorflow implementation of OverFeat [15]. Overfeat, is a method that uses a single CNN for integrating the tasks of detection, recognition, and localization (bounding box) of objects within an image. While Overfeat applies a sliding window approach, usually a computationally expensive technique, its particular implementation with a CNN avoids redundant computations at the first layers of the network (shared by the overlapping windows). We annotate each image with the information (bounding boxes) of all the occurrences of a given logo.

It is important to clarify that, in our work, logo detection and visual brand identity (VBI) recognition are independent components. VBI recognition analyzes how an image fits a given brands culture and character and it is useful to discover valuable images for certain brand's communication channels (e.g. ads). In the case of the Estrella Damm beer brand, for instance, the company was interested in images transmitting the Mediterranean lifestyle to be used as ads. The selected images for that campaign did not depict the company's products. On the other hand, logo detection serves to discover images of consumers interacting with a specific product, and it is useful, for instance, to select images for an ecommerce site (e.g. the ecommerce site of Bershka employs Instagram images).

4.5 Datasets

Table 2 shows the details of the datasets used in this work for which results are provided in the next section. We worked with three different groups of datasets, the ones for training new generic image recognition classifiers or NGR (e.g. "selfie" and "spam"), the ones for visual brand identity recognition classifiers or VBI (e.g. "Pepsi" and "FC Barcelona") and the ones for logo detection or LOGO (e.g. "Estrella Damm logo"). We only provide details about a representative subset of the classifiers/detectors actually trained. Regarding NGR, we finally trained more than 100 new classifiers (here we show details about 6 of them). Regarding VBI, we trained more than 20 classifiers (here we show details about 6 of them). Regarding LOGO, it is a new feature in which we are currently working and here we only provide results for the Estrella Damm logo.

Regarding datasets acquisition for NGR, the most part of the new models required to acquire training images that are not part of any public images dataset. In order to solve this problem, we combined images both from Instagram and the WWW. Instagram photos were obtained from the Instagram API, filtered with user defined tags and manually purged. As user defined tags are very noisy this method proved to be inefficient and very time-consuming. In order to facilitate the generation of more ground truth annotations and a larger training dataset we also obtained images from Google Images through the Custom Google Search API. This method, which allowed to automatically annotate a bigger set of images, turned out to be very useful as almost all the retrieved images showed the desired category (e.g. "handbag") and minimum manual purge was required. The resulting dataset contains more than 50K images distributed in 100 different categories.

Regarding datasets acquisition for VBI, the brands are responsible of providing a curated dataset of images that satisfy their VBI criteria. In our prototype, our main source of (positive) training images are the brand's Instagram profiles. With the help of the Instagram API we have been able to collect training sets with sizes varying from several hundred (e.g Ecooltra) to several thousand (e.g. Desigual). As negatives we use images from multiple NGR classifiers' datasets with low probability of semantic overlapping. While the performance of a VBI classifier depends initially on the quality of the starting dataset, it improves as the system is used.

The data acquisition for LOGO is the only one which required a totally manual process. First, we captured Instagram images annotated with the Estrella Damm promotional hashtag. Then we manually annotated the bounding boxes of all the logo occurrences with a tool that we developed for this task.

It's worth mentioning that we didn't use a public dataset such as BrandSocial-Net because the goal of the work was to evaluate the viability of the whole approach on a real scenario. This included to experience with the generation of training datasets for custom (brand-specific) classifiers on demand.

tag	type	#positives	#negatives
selfie	NGR	295	8,959
group_selfie	NGR	98	8,884
spam	NGR	319	8,979
burguer	NGR	474	8,845
nails	NGR	434	8,866
sushi	NGR	571	8,920
Pepsi	VBI	680	8,422
FC Barcelona	VBI	1,023	8,366
Estrella Damm	VBI	663	8,363
Desigual	VBI	1,381	8,862
Catalunya Experience	VBI	89	8,809
Estrella Damm logo	LOGO	322	1,681

Table 2 Summary of training datasets (NGR=New generic recognition, VBI=Visual brand identification, LOGO=Logo detection).

4.6 Sofware and hardware setup

The data acquisition system was implemented with Java and served as a set of RESTful APIs. The scalable image metadata database was implemented with Elasticsearch. The image recognition service was implemented with Python and TensorFlow, and also implemented as a set of RESTful APIs. Once in production, the system is running over a cluster of 6 Amazon EC2 t2.large instances (dual core 3.3 GHz Intel Xeon processor and 8 GB of memory). The CNNs were trained over a high-end server with a quadcore Intel i7-3820 at 3.6 GHz with 64 GB of DDR3 RAM memory, and 4 NVIDIA Tesla K40 GPU cards with 12 GB of GDDR5 each.

5 Results

The combined method that we propose integrates multiple state-of-the-art image recognition and object detection algorithms. As mentioned previously, the goal of the work is not to re-evaluate the performance of these algorithms but to analyze their suitability among other alternatives and to asses the viability of the whole approach on a real scenario. Therefore, following we provide representative results for the three different groups of models that we trained (image recognition classifiers or NGR, visual brand identity recognition classifiers or VBI and logo detectors or LOGO). The difference between NGR and VBI classifiers is the way the training data is obtained, but the underlying method is the same. In order to evaluate the suitability of the chosen method (Transfer Learning) we compared its classification accuracy, classification time and training time with two alternative methods (BoW+RBF-SVM and lightweight CNN). We also provide some results of the chosen method when applied to VBI classifiers. Regarding LOGO, we apply an overlap crite-

tag	BoW	CNN	CNN-TL
selfie	72%	87%	93%
group_selfie	76%	88%	95%
spam	69%	78%	91%
burguer	81%	89%	95%
nails	83%	92%	97%
sushi	86%	93%	96%

Table 3 Training results of some of the 100 new models that we have trained for image semantics recognition.

rion of 0.5 and evaluate the accuracy, recall and fall out of the chosen method in the case of the Estrella Damm logo.

5.1 Classification accuracy

Table 3 shows some representative results obtained for the image semantics recognition part. The results show that the classic Bag of Words approach (BoW) provides the worst accuracy but it is the fastest to train and predict and the one with a smallest memory footprint. The approach consisting in defining and training our own deep convolutional neural network provides accuracy improvements of more than 10% with respect to BoW. However, with our small training sets this method implies a strong overfitting, as pointed in [7]. Training times in a high-end server with 4 NVIDIA Tesla K40 GPU cards are between 3 and 5 hours. One advantage of this method (with respect of the Transfer Learning approach that we finally chose) is that models have a small memory footprint. Another advantage is that predictions are faster (as the network is significantly simpler). Disadvantages of this method (with respect to Transfer Learning) are overfitting, significantly higher training times and (about 5%) lower accuracies. Finally, the transfer learning approach improves accuracies (with respect to training our own lightweight CNN) about 5%, reduces overfitting and reduces training times to less than 2 hours (significantly less if some images are reused as the bottlenecks need to be obtained just once). One significant disadvantage (specially when many models have to be served simultaneously) is that models have a big memory footprint. Another disadvantage is that prediction times are higher.

Because of its advantages in terms of accuracy, reduced overfitting and training times, we finally chose the Transfer Learning approach for training the classifiers. Its ability to work with small datasets is especially suited the visual brand identity recognition task. Table 4 shows the VBI classification results that we obtained for 5 real brands with the Transfer Learning approach.

Regarding logo detection, we trained a detector for the Estrella Damm beer brand with 2K manually annotated 320x320 Instagram images. The resulting detector provides a 97% accuracy, a 68% recall and a 99.4% fall out (only 6 of each 1000 negatives escape to the filter). Figure 6 shows some example images containing one or more occurrences of the logo of Estrella Damm. A typical

Brand name	training	accy.
Pepsi	5,922s	87%
FC Barcelona	6,141s	96%
Estrella Damm	5,789s	93%
Desigual	6,310s	95%
Catalunya Experience	5,624s	76%

Table 4 Training setup and results for 5 visual brand identity classifiers.

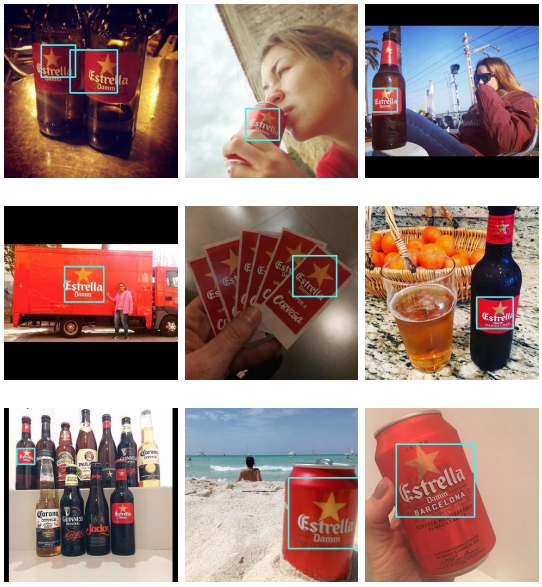

Fig. 6 Example images containing one or more ocurrences of the Estrella Damm logo.

Estrella Damm campaign generates more than 200K images, the most part of which do not contain the logo. The detector losses some images with the logo (as it has a relatively low recall), but discards a huge number of true negatives, close to 200K images.

5.2 Classification time

Classification time is critical as the system needs to process in real-time a huge and volatile amount of incoming images. Each image needs to be classified by multiple models. Figure 7 shows a decomposition of the average classification time of one low-resolution (320x320) Instagram image by one model. The values are just indicative as download times are context-dependent. The main component of the classification time is the bottleneck computation (0.4 seconds). This computation, along with the downloading of the image (0.25 seconds), need to be done just once, as the bottlenecks are the same for all the

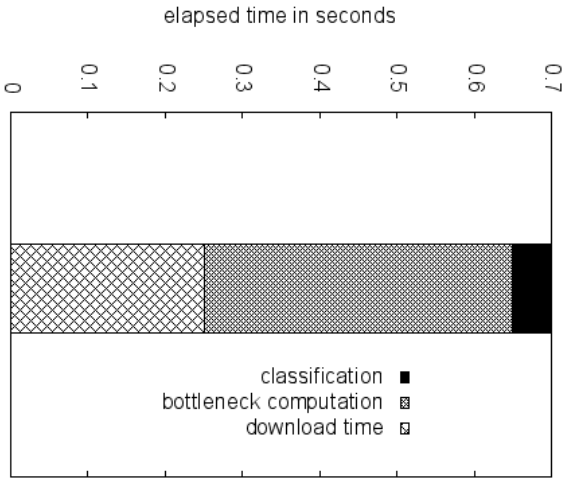

Fig. 7 Classification time for a single 320x320 image and one model.

models and they can be fed directly into the final layer. As the time to process the final layer is very small (0.05 seconds), including more models does not imply a significant cost, neither in terms of time, being the memory the only limitation in practice. On average, we need 0.7 seconds to classify one image.

Regarding logo detection, on average this stage adds around 0.8 seconds to the annotation of a single image (increasing the total time to around 1.5 seconds/image). Our current implementation does not support reusing intermediate layers among multiple logo detectors, so this time gets multiplied if more than one logo detector has to be applied to the same image. Even if just one logo detector is applied, a 0.8 seconds increase is a big penalty. For this reason, we have opted to only apply the logo detector when certain conditions are met. These conditions are expressed in terms of the other classifiers (e.g. not applying the logo detector if the image has been classified as spam). These conditions are configurable in a per-campaign basis.

The overall annotation process is an embarrassingly parallel problem and the throughput of the system can scale linearly adding more computational resources. We run as many annotations in parallel as possible, depending on the available cpu and memory.

Figure 8 shows a slice of a time series for the amount of images acquired and indexed during September 2016. In that period, the system was running an average of 5 campaigns simultaneously, including, but not limited to, Estrella Damm, Desigual, Catalunya Experience, Ecooltra and Shakn. Acquisition was done mainly based on hashtags (around 5 per brand), but some geolocation-based filters were also used (e.g. some specific beaches from the Balearic Islands for the Estrella Damm campaign). Each captured image was classified with the corresponding VBI model, inception, and 5-10 of our own semantics

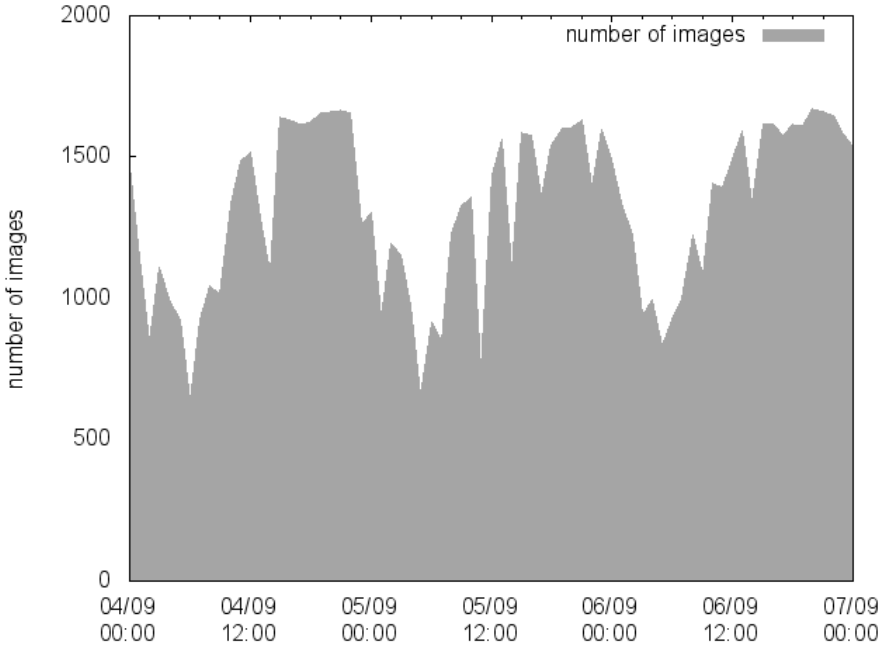

Fig. 8 Slice of the time series showing the amount of images acquired per hour during September 2016.

recognition models. More than 1 million images were captured, classified and indexed during one month, providing, on average, more than 200K images for each campaign.

6 Conclusions

The research work presented in this paper analyzes the usage of deep convolutional neural networks (CNNs) for curating and filtering user generated content (UGC) for digital marketing tasks. We have built a system that captures images from Instagram and Twitter in real-time, and processes them by multiple CNNs that automatically enrich their metadata with tags that describe their visual content and also how they fit the visual identity of a brand. As far as we know, this is the first work that addresses the automatic recognition of visual brand identity in UGC. We have compared the results of three different methods (BoW+RBF-SVM, lightweight CNN and Transfer Learning) and we conclude that the Transfer Learning approach is the one that better suits this domain (best accuracy, less overfitting with small datasets, and low training times). With this method, we have trained VBI classifiers for more than 10 real brands and more than 100 classifiers for generic description of social

media images. We have employed a ground truth of more than 50K images. Each model can be trained in less than 2 hours and the most part of resulting accuracies are always above 90%. We also process the images with Google's Inception-v3 and expand its 1000 WordNet categories with their corresponding hypernyms to obtain a dictionary of more than 7,000 tags. On average, we need 0.7 seconds to classify each image. As the bottleneck computation and image download consume the most part of this time, applying more models sequentially has just a sub-linear impact on the elapsed time. In practice, we run as many classifications in parallel as possible, depending on the available cpu and memory. During a experiment conducted on September 2016, the system captured, classified and indexed more than 1 million images related to 5 different brands. Discovering valuable images among them is finally done by using the provided search interface and applying the different filters (over the VBI tags, expanded inception tags, our own semantics tags, and any metadata provided by Instagram/Twitter). With respect of traditional curation methods, our approach minimizes human visual inspection, increases the number of photos processed several orders of magnitude, increases the quality of the resulting photos, enables near real-time discovery and reduces the cost drastically. There are some issues that need to be addressed, however. Our logo detection implementation has a high detection time and is currently single-label. Directions for future research include the evaluation of alternative logo detection algorithms. On the other hand, regarding image semantics recognition, the criteria of inclusion of new tags is currently heuristic, and should be addressed. Besides, as the number of possible tags becomes very large, it becomes more difficult to design a human curation interface able to take profit from them.

Acknowledgements This work is partially supported by the Spanish Ministry of Economy and Competitivity under contract TIN2015-65316-P and by the SGR programme (2014-SGR-1051) of the Catalan Government.

References

1. Bakhshi, S., Shamma, D.A., Kennedy, L., Gilbert, E.: Why we filter our photos and how it impacts engagement. In: Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015, pp. 12–21 (2015)
2. Chute. enterprise ugc. Http://www.getchute.com/ (Accessed June 6, 2017)
3. Clark, M., Black, H.G., Judson, K.: Brand community integration and satisfaction with social media sites: a comparative study. Journal of Research in Interactive Marketing **11**(1), 39–55 (2017)
4. Clarke, T., Costall, A.: The emotional connotations of color: A qualitative investigation. Color Research & Application **33**(5), 406–410 (2008)
5. Curalate. Https://www.curalate.com/ (Accessed June 6, 2017)
6. Denton, E., Weston, J., Paluri, M., Bourdev, L., Fergus, R.: User conditional hashtag prediction for images. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15, pp. 1731–1740. ACM (2015)

7. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pp. 647–655 (2014)
8. Gao, Y., Wang, F., Luan, H., Chua, T.: Brand data gathering from live social media streams. In: Proceedings of the International Conference on Multimedia Retrieval, ICMR 2014, Glasgow, United Kingdom - April 01 - 04, 2014, p. 169 (2014)
9. Gao, Y., Zhao, S., Yang, Y., Chua, T.: Multimedia social event detection in microblog. In: MultiMedia Modeling - 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part I, pp. 269–281 (2015)
10. Gao, Y., Zhen, Y., Li, H., Chua, T.: Filtering of brand-related microblogs using social-smooth multiview embedding. IEEE Trans. Multimedia **18**(10), 2115–2126 (2016)
11. Nguyen, D.T., Alam, F., Ofli, F., Imran, M.: Automatic image filtering on social networks using deep learning and perceptual hashing during crises. CoRR **abs/1704.02602** (2017). URL http://arxiv.org/abs/1704.02602
12. Olapic. earned content platform. Http://www.olapic.com/ (Accessed June 6, 2017)
13. Park, M., Li, H., Kim, J.: HARRISON: A benchmark on hashtag recommendation for real-world images in social networks. CoRR **abs/1605.05054** (2016)
14. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 779–788 (2016)
15. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., Lecun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. In: Proceedings of the International Conference on Learning Representations (ICLR2014), CBLS, April 2014 (2014)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 2818–2826 (2016)
17. Tous, R., Torres, J., Ayguad, E.: Multimedia big data computing for in-depth event analysis. In: Proceedings of the 2015 IEEE International Conference on Multimedia Big Data (BigMM), April 20-22, 2015, Beijing, China, pp. 144–147. IEEE (2015)
18. Tous, R., Wüst, O., Gomez, M., Poveda, J., Elena, M., Torres, J., Makni, M., Ayguadé, E.: User-generated content curation with deep convolutional neural networks. In: Proceedings of the 2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016, pp. 2535–2540 (2016)
19. Wang, F., Qi, S., Gao, G., Zhao, S., Wang, X.: Logo information recognition in large-scale social media data. Multimedia Syst. **22**(1), 63–73 (2016)
20. Zhao, S., Yao, H., Gao, Y., Ji, R., Xie, W., Jiang, X., Chua, T.: Predicting personalized emotion perceptions of social images. In: Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016, pp. 1385–1394 (2016)
21. Zhao, S., Yao, H., Zhao, S., Jiang, X., Jiang, X.: Multi-modal microblog classification via multi-task learning. Multimedia Tools Appl. **75**(15), 8921–8938 (2016)