User-generated Content Curation with Deep Convolutional Neural Networks

Ruben Tous*, Otto Wust†, Mauro Gomez†, Jonatan Poveda†, Marc Elena†
Jordi Torres*‡, Mouna Makni * and Eduard Ayguadé*‡

*Universitat Politècnica de Catalunya (UPC). Barcelona, Spain
†Adsmurai. Barcelona, Spain
‡Barcelona Supercomputing Center (BSC). Barcelona, Spain

Email: {rtous; mmakni, lcruz}@ac.upc.edu,{mauro, jonatan, otto
@adsmurai.com, {jordi.torres; eduard.ayguade}@bsc.es

***Abstract*—In this paper, we report a work consisting in using deep convolutional neural networks (CNNs) for curating and filtering photos posted by social media users (Instagram and Twitter). 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. 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.**

I. Introduction

Instagram users share more than 80 million photos per day, captured from all corners of the earth. Twitter users post more than 500 million tweets each day, from which a 7% contain images. A significant part of this visual user-generated content has potential value for digital marketing tasks. On the one hand, users' photos can 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.

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. The most part of current solutions rely on costly manual curation tasks over random samples. This way many contents are not even processed, and many valuable photos go unnoticed. We propose using deep convolutional neural networks (CNNs) 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.

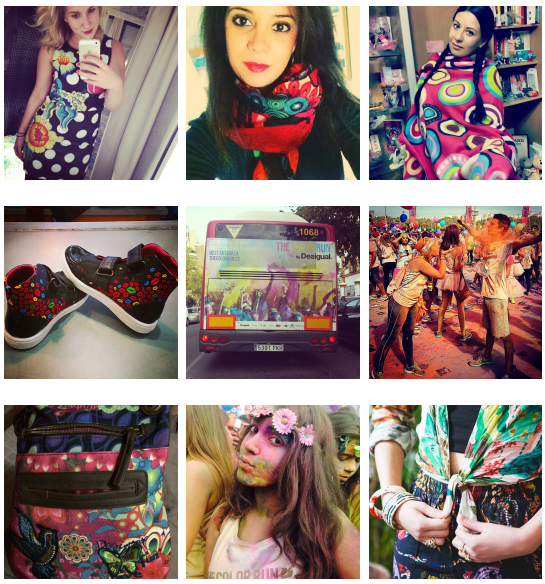

Figure 1. Example images posted by Instagram users and tagged with Desigual's promotional hashtags (e.g. #lavidaeschula)

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. As a final step, a human expert performs a final selection with the help of a search interface

 DOI 10.1109/BigData.2016.7840751

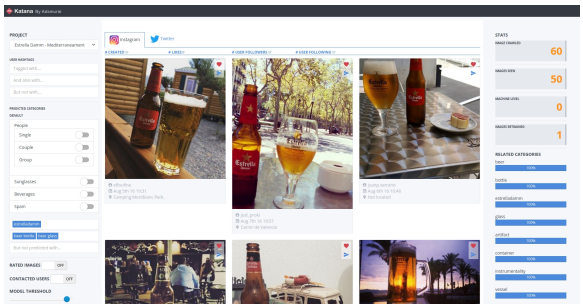

Figure 2. Screenshot of the user interface of the system, where users can navigate, search and select images from the database.

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.

II. Related Work

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 [1], [2] or [3], 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). As far as we know, this is the first work that addresses the automatic recognition of visual brand identity in images. In [4] 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 [5]. Regarding the annotation of images with generic object categories, we reuse Google's Inception-v3 model [6], trained for the ImageNet Large Scale Visual Recognition Challenge and 1000 object categories with a top-5 error rate of 15.3%. Regarding the recognition of visual brand identity, 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].

III. 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 amount 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 is 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 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 (Figure 2 shows a screenshot of the user interface). 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).

IV. Methodology

A. Image semantics recognition

During the acquisition stage, captured images are processed by multiple deep convolutional neural networks that try describing their visual content. These CNNs are trained with generic, manually-labelled, images. Among these CNNs, we reuse Google's Inception-v3 [6], trained

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 I
SOME OF THE 1000 ILSVRC TAGS AND THEIR EXPANDED WORDNET HYPERNYMS.

for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and 1000 object categories. As the 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 I 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 models, that include more than 100 new tags (Table II shows some of them). Notable examples are *spam* and *selfie*, tags that have proven to be very useful when searching these kind of images.

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.

B. 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

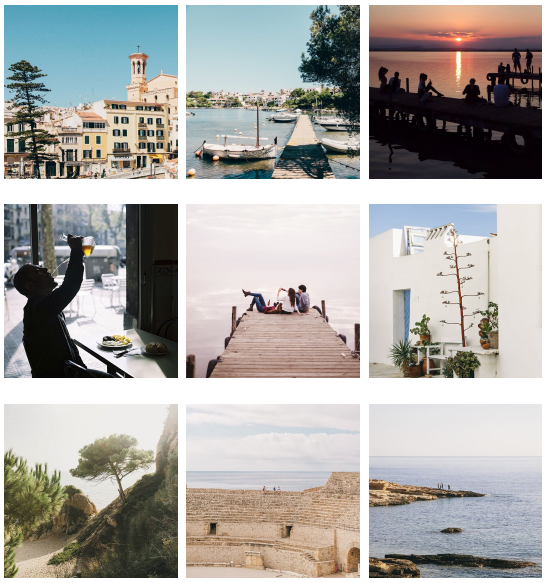

Figure 3. Example images showing the visual brand identity of the Estrella Damm beer brand, related to the mediterranean lifestyle

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.

Given a brand, we train a CNN with images that it has used in its previous marketing campaigns and that are representative of the brand's visual identity. We use the this CNN to mark user images that could fit the visual identity of the brand. Our main source of 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 hundreds (e.g Ecooltra) to several thousands (e.g. Desigual).

C. 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 trough all the layers of the Google's Inception-v3 model [6] 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 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.

D. 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.

V. Results

A. Classification accuracy

Table II 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).

tag	#positives	#total images	BoW	CNN	CNN-TL
selfie	295	9,254	72%	87%	93%
group_selfie	98	8,982	76%	88%	95%
spam	319	9,298	69%	78%	91%
burguer	474	9,319	81%	89%	95%
nails	434	9,300	83%	92%	97%
sushi	571	9,491	86%	93%	96%

Table II
TRAINING SETUP AND RESULTS OF SOME OF THE 100 NEW MODELS THAT WE HAVE TRAINED FOR IMAGE SEMANTICS RECOGNITION.

Brand name	#positives	#total images	training	accy.
Pepsi	680	9,102	5,922s	87%
FC Barcelona	1023	9,389	6,141s	96%
Estrella Damm	663	9,026	5,789s	93%
Desigual	1381	10,243	6,310s	95%
Catalunya Experience	89	8,897	5,624s	76%

Table III
TRAINING SETUP AND RESULTS FOR 5 VISUAL BRAND IDENTITY CLASSIFIERS.

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 specially suited the visual brand identity recognition task. Table III shows the VBI classification results that we obtained for 5 real brands with the Transfer Learning approach.

B. 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 4 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. This computation, along with the downloading of the image, need to be done just once, as the bottlenecks are the same for all the models and they can be fed directly into the final layer. As the time to process the final layer is very small, including more models does not imply a significant cost in terms of time, being the memory the only limitation in practice. On average we need 0.85 seconds to classify one image. As this is an embarrassingly parallel problem the throughput of the system can scale linearly adding more computational resources. We run as many classifications in parallel as possible, depending on the available cpu and memory.

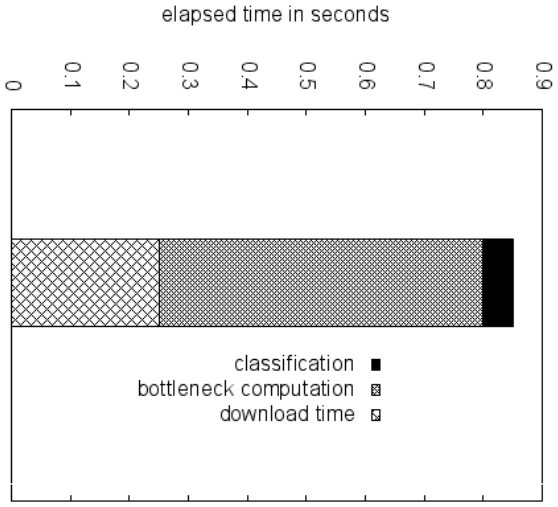

Figure 4. Classification time for a single 320x320 image and one model.

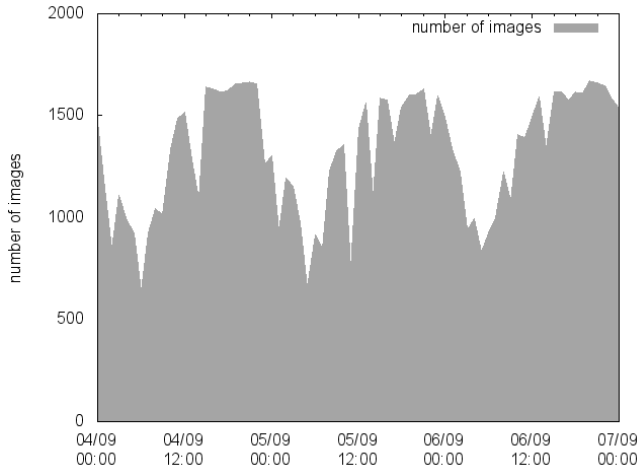

Figure 5. Slice of the time series showing the amount of images acquired per hour during September 2016.

Figure 5 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 simultaeously, 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 Baelaric Islands for the Estrella Damm campaign). Each captured image was classified with the corresponding VBI model, inception, and 5-10 of our own semantics 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.

VI. 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 0.85 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.

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] M. Park, H. Li, and J. Kim, "HARRISON: A benchmark on hashtag recommendation for real-world images in social networks," *CoRR*, vol. abs/1605.05054, 2016.

[2] R. Tous, J. Torres, and E. Ayguad, "Multimedia big data computing for in-depth event analysis." in *BigMM*. IEEE, 2015, pp. 144–147. [Online]. Available: http://dblp.uni-trier.de/db/conf/bigmm/bigmm2015.html#TousTA15

[3] E. Denton, J. Weston, M. Paluri, L. Bourdev, and R. Fergus, "User conditional hashtag prediction for images," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: ACM, 2015, pp. 1731–1740. [Online]. Available: http://doi.acm.org/10.1145/2783258.2788576

[4] S. Bakhshi, D. A. Shamma, L. Kennedy, and E. Gilbert, "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*, 2015, pp. 12–21. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10573

[5] T. Clarke and A. Costall, "The emotional connotations of color: A qualitative investigation," *Color Research & Application*, vol. 33, no. 5, pp. 406–410, 2008. [Online]. Available: http://dx.doi.org/10.1002/col.20435

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015. [Online]. Available: http://arxiv.org/abs/1512.00567

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "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*, 2014, pp. 647–655. [Online]. Available: http://jmlr.org/proceedings/papers/v32/donahue14.html