What Would They Say? Predicting User’s Comments in Pinterest

J. C. Gomez, T. Tommasi, S. Zoghbi and M. F. Moens

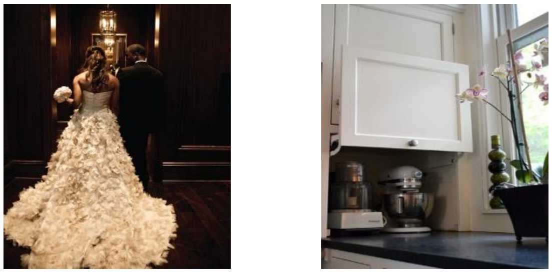
Abstract— **When we refer to an image that attracts our attention, it is natural to mention not only what is literally depicted in the image, but also the sentiments, thoughts and opinions that it invokes in ourselves. In this work we deviate from the standard mainstream tasks of associating tags or keywords to an image, or generating content image descriptions, and we introduce the novel task of automatically generate user comments for an image. We present a new dataset collected from the social media Pinterest and we propose a strategy based on building joint textual and visual user models, tailored to the specificity of the mentioned task. We conduct an extensive experimental analysis of our approach on both qualitative and quantitative terms, which allows assessing the value of the proposed approach and shows its encouraging results against several existing image-to-text methods.**

Keywords— **Multimodal Clustering, Pinterest, Social Media, User Generated Content, Deep-Learning Representation, Automatic Image Annotation.**

I. INTRODUCTION

THE USE of images in our everyday life has dramatically changed in the last years. On one side, we became prolific image providers thanks to the increased popularity of digital cameras and smart phones. On the other, we are now active image consumers: almost every website is enriched with pictures and when surfing the web the images are often the ones that mainly attract our attention. Social media like Facebook and Pinterest have contributed to boost both these tendencies, as confirmed by the fact that more than 300 million pictures have been uploaded per day since 2012 [1]. In turn, Pinterest allows its users to create boards of visual bookmarks called *pins*. There, each user can collect and save images found online to share content, plan trips, collect recipes, etc. The images collected by Pinterest users are commonly posted together with short textual comments. Unlike a standard tag annotation, or a visual content description, those comments may contain emotions, opinions or thoughts, together with a superficial description of the images content. Such comments can give clues on the user personal interests, way of thinking and language style (see Fig. 1). Although mining all this information may have a key role in several tasks like user modelling, sentiment analysis or personalized advertisement, up to our knowledge these user-specific comments have not been extensively studied yet.

In this work, we expect to make a first step in the analysis of such user-generated content, proposing three main contributions. (1) We introduce the novel task of predicting personalized users’ comments to their images; (2) we present a new dataset collected from Pinterest consisting of over 70,000 images accompanied with users’ comments; and (3) we perform an extensive experimental study on the task and proposing a method able to leverage over visual and textual pin similarities.



Love this train

Hinged cabinet. Love this to hide appliances

Figura 1. Ejemplo de dos imágenes en Pinterest con comentarios personales de los usuarios.

The rest of the paper is organized in the following way: in section 2 we briefly review the literature; in section 3 we describe the proposed task and the dataset used for experimentation; in section 4 we present and discuss our approach; in section 5 we show the experimental results; and in section 6 we conclude the paper with an overall discussion and possible future research directions.

II. RELATED WORK

There exist two major trends in the literature related to associating text to images. One focuses on assigning keywords (or tags) to an image, the other targets the more challenging task of providing a full description of the image. In both cases the final aim is to recognize the image content in terms of depicted objects [7] or scene [5]. Previous work relies mostly on content-based strategies that either predict the text from the images by training a model of their relationship [16, 17, 18, 19], or propagates tags through a k-nearest neighbor method [4, 15]. In these studies, the analysis is generally performed on data collections where the ground truth image annotations are well defined with a large consensus among of users or experts.

The task of image annotation with users’ comments has started to emerge in computer vision and natural language literature only more recently [6, 14]. This task challenges the standard automatic annotation solutions in two main aspects: first, personal collections often contain a limited number of images, and second, the associated text reflects the user interests and style, being a combination of objective descriptions and subjective expressions. Existing methods overcome the first issue by leveraging over further external information like personal calendars or metadata about location and time [3, 12]. The second issue is generally addressed by mining the

J. C. Gomez, KU Leuven, Belgium, jcgcarranza@gmail.com
T. Tommasi, UNC Chapel Hill, NC, USA, ttommasi@cs.unc.edu
S. Zoghbi, KU Leuven, Belgium, susana.zoghbi@cs.kuleuven.be
M. F. Moens, KU Leuven, Belgium, sien.moens@cs.kuleuven.be

annotation history of the users, and focusing on word whose definitions suit better the specific person, e.g. using *kitty* instead of the more generic *cat* [14]. In this work, we push the limit of automatic annotation and we focus on generating sentences analogous to the comment that a user would post on social media to accompany an image.

III. PROBLEM SETUP AND DATASET

In Pinterest users post pins which are organized in boards. A pin is an <image,text> pair, and a board groups several pins under user-defined topics which can be specific (e.g. a famous person) or generic concepts (e.g. food or clothes). Be u an user and $\mathbf{P}^u_{i,j}=\left(\mathbf{g}^u_{i,j},\mathbf{x}^u_{i,j}\right)$ his/her pins collection, where $\mathbf{g}$ refers to the image and $\mathbf{x}$ to the associated text. The indexes j and i identify respectively the specific pin j=1,…, L^u_i and the board it belongs to i=1,…,M^u**.** M^u indicates the number of boards of user u, and L^u_i the number of pins inside board i. Finally n_u=$\sum_{m=1}^{M^u} L^u_m$ is the total number of pins of user u.

In this work, we collected randomly 70,200 pins belonging to 117 users by directly crawling the Pinterest website. We selected 3 boards per user, saving 200 pins per board for a total of 600 pins per user. All the images in the collection have an associated comment, which are in English and are of a variable length from one (12.33% of the comments) to some tens of words. Todas las imágenes de la colección tienen comentarios asociados, los cuales están en inglés y tienen una longitud variable de una (12.33% de los comentarios) a algunas decenas de palabras. Within the 10 most frequently used words (no stop words) in the comments from the collection, we have: *love*, *easy*, *great*, *cute* and *beautiful*, which shows that users try more to express an opinion or emotion rather than describing the image content. The size of the images also varies but they are all saved in JPG format. There are 4.2% of images that are shared by 2 or more users, which shows that there is a great diversity of content in the pins, including products (e.g. clothes and jewelry), interests (e.g. food and decoration), photographs (e.g. animals and landscapes) and more abstract content (e.g. paints and designs).

We preprocessed all the textual (comments) and visual (images) information from the pins in the following way: we cleaned each comment by removing special symbols (e.g. stars, hash tags, etc.), urls and one-letter words (with exception of *i* and *a*), to form each textual vector $\mathbf{x}^u_{i,j}$ as a string of words separated by spaces. We processed each image using a convolutional neural network [13] to obtain $\mathbf{g}^u_{i,j}$ as a vector of visual features. We used for this the DeCAF library [2], considering the activation values of the 4,096 neurons in the 7-th layer as the image features.

For the experiments we divided the collection of pins in a training and a test set. We selected at random 10 pins per board per user (30 pins per user) to form the test set (3,510 pins in total), the remaining 66,690 pins were used for training. During training, both the visual and the textual information are available. Durante el entrenamiento, tanto la información textual como la visual están disponibles. During the test phase, only the visual part $\mathbf{g}^u$ of the test pins for user u is available (without information of the board of origin); and our goal is to automatically generate the associated comments $\mathbf{x}^u$ for such images.

IV. METHOD

The original data organization of pins in boards provides already some hint on the comments that a user can post. For instance, a comment like "*amazing outfit*" fits well for a board on clothes, and "*that looks delicious*" for a board on food. Our proposal of associating comments to new images consist on taking advantage of this structure and define a method based on the combination of multimodal pin clustering [10] and text transfer. We describe below the two important steps and general scheme of our proposal is depicted in Fig. 2.

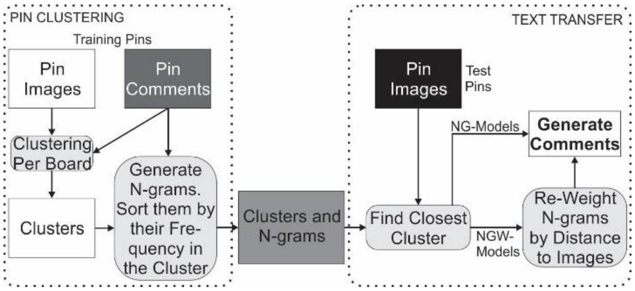


Figura 2. Descripción general de los métodos NG- y NGW- para la generación automática de comentarios para imágenes.

Pin clustering. In this phase we separately consider each board i of user u, and we cluster its training pins. Clustering the pins per board allows us to build specific clusters and a finer mapping between images and comments. In the following, we keep fixed the user and to simplify the notation we neglect the superscript u. The clustering process is multimodal, i.e. it uses the visual and textual information together. More in detail, we first evaluate the pairwise Euclidean distance among all the images and all the comments, obtaining two distance matrices D_g,D_x∈$\mathrm{R}^{L_i\times L_i}$ for images and comments respectively. Second, the contribution of each matrix is combined linearly by using a weighting parameter λ in the following way: D=λD_g+(1-λ)D_x. Finally, we apply a hierarchical clustering with average linkage over matrix D to create a final set of C_i clusters of pins per board. Each cluster c_i=1,…,C_i contains l_{c_i} pins that are similar among them.

Afterwards, for each cluster c_i we compute its visual centroid as the average visual feature vector of the cluster. We save each centroid together with two other metadata: the average number of words $\overline{w}_{c_i}$ in the comments inside cluster c_i and an n-gram language model. This model consists in extracting all posible n-grams k=1,…, K_{c_i} from the comments of cluster c_i, computing in addition the frequency $f_{c_i,k}$ of each n-gram inside the cluster and saving the indexes of the pins where each n-gram appears.

Text transfer. In this phase, given a test image, we first identify its closest cluster c* using the Euclidean distance regarding the visual centroids. Inside the cluster c* we calculate the distance of the test image to all the cluster images $d_{t,j}$, j=1,…,l_{c*} and we use the distances to weight the n-gram frequencies. We use in turn the weighted frequencies to build a comment for the test image. For that, we first sort in descending order the n-grams based on their weighted frequencies $\tilde{f}_{c*,k}=f_{c*,k}/d_{t,j^k}$, where d_{t,j^k} is the distance of the test image to the closest image whose comment contains the *k*-th n-gram; afterwards we iteratively create the comment for the test image using the sorted n-grams. The comment starts as an empty string, we concatenate to it an n-gram in each step, and we check the intermediate comment for redundancy and we remove repeated 1-2-…-n-grams. The process of aggregating new n-grams and cleaning the comment continues until the obtained sentence has a length equal or greater than $\overline{w}_{c*}$ (the average length of comments in group *c**). Finally, we eliminate conjunctions (*and*, *or*, *but*, *for*, *so*, etc.) from the end of the comment. Once cleaned, the final composed comment is transferred to the image.

We named our proposed method as NGW-CVT: *N-Gram Weighted transfer by Clustering over Visual and Textual Information.* As part of the described full procedure, we investigate specific cases and different variants. First of all, by tuning the parameter λ during clustering, we can trigger the importance of the visual and textual training cues. We have three models here:

NGW-CT – With λ=0, builds the clusters based solely on textual information of the pins.

NGW-CV – With λ= 1, builds the clusters based solely on visual information of the pins.

NGW-CVT – With λ= 0:5, builds the clusters using the same proportion of textual and visual information.

Secondly, when transferring the text to an image, we can rely directly on the unweighted n-grams. In this way, we obtain three models defined similarly as before:

NG-CT – With λ= 0, builds the clusters based solely on textual information.

NG-CV – With λ= 1, builds the clusters based solely on visual information.

NG-CVT – With λ= 0:5, builds the clusters using the same proportion of textual and visual information.

These models simplify the training and test procedures: during training it is possible to disregard the pin index for the n-grams, and during testing it is not necessary to compute the distances between the test image and the images inside the cluster. However, they have the disadvantage that the same comment will be transferred to all the test images that are assigned to a specific cluster.

V. EXPERIMENTS

In this section we provide the details of the experimental analysis on the Pinterest dataset. We start by defining the parameter settings of the proposed method and describing several reference baselines (section V.A). We then present and discuss the obtained results (section V.B).

A. Model Parameters and Baselines.

Our proposed method has two main parameters. One is the dimensionality *n* of the n-grams extracted from the training pins and then recombined to generate the test text. We fixed it to *n*=4, since larger n-grams appear difficult to combine, producing text that is complicated to read. The second parameter is the number of clusters for each board. We considered C_i^u=60, a value that produced good results on a validation set extracted from the training.

We assess the performance of our approach against four different baselines. The first two are fully based on textual information. Previous work indicated that the tagging history of the user provides sufficient information to predict future annotations, regardless of the specific image content [14]. Following this finding we consider two models:

FNG – This technique calculates the frequency of each n-gram over all the pins n_u of a user *u*, without taking into consideration neither the associated images, nor the board where the pin was saved. The same iterative merging procedure used in our NGW- method is applied by taking into consideration the average number of words of all the user's comments as upper limit for the length of the generated comment. The obtained comment is assigned to all the test images.

FPin – It can happen that multiple pins have exactly the same textual comment. In this approach we count how many times each comment is repeated and we assign the most frequent one to all the test images.

We test also two methods fully based on visual information. The first one keeps the focus on each specific user by relying on his/her training pins. The other exploits an external source of data:

CPin – In this method, after the initial clustering procedure per board, the test image is assigned to one of the closest cluster *c**. The visually closest training pin inside *c** is then used as the source of textual information: its whole comment is transferred to the new image without passing through the n-gram statistics of the cluster.

Im2Text – This method was introduced in [9] to automatically create image descriptions using as reference a collection of 1 million captioned photos from Flickr. In the presence of a query image, its similarity to all the Flickr photos is evaluated and the caption associated to the closest matching sample is transferred. The similarity is defined on the basis of global visual cues using the gist features [8] and the pixel values of the image thumbnails (images reduced to size 32x32). In work we use the code provided by the authors (http://vision.cs.stonybrook.edu/vicente/sbucaptions/). Since the representation used in this model captures only high level information, we test also a refined version of this method: after having obtained 20 candidate Flickr images per query, we select the closest one on the basis of the similarity evaluated with the DeCAF features computed for those images, we call this method Im2Text-DeCAF.

We evaluate all the models in two ways. The first one is an automatic evaluation using the BLEU score [11] between the generated comment for a test image and the user's ground truth comment. The BLEU score is a very conservative and strict measure that matches exactly complete words, as a consequence, it does not consider as valid generated comments that do not match exactly the original comment. Because of that, we conduct a second higher level quantitative analysis. For that, we passed 200 randomly selected pins to human evaluators. We used the Crowdflower (www.crowdflower.com) platform and the text of each pin is

TABLE I. RESULTS OF THE AUTOMATIC EVALUATIONS AND MANUAL OF THE COMMENTARIES GENERATED BY THE DIFFERENT MODELS AND THE ORIGINAL COMMENTARIES

	Automatic					Manual	
Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**Razón BLEU-3**	**Error Board**	**Readability**	**Relevance**
NG-CT	0.085^{A}	0.048^{A}	0.036^{A}	0.062	**0.200**	1.84	1.45
NG-CV	0.084^{AB}	0.046^{AB}	0.034^{AB}	0.066	0.215	1.76	1.40
NG-CVT	0.089^{ABC}	0.049^{AB}	0.036^{AB}	0.067	0.217	1.82	**1.64**
NGW-CT	0.095^{BCD}	0.058^{C}	0.044^{C}	0.076	**0.200**	1.91	1.46
NGW-CV	0.099^{DE}	0.060^{CD}	0.046^{CD}	**0.079**	0.215	1.60	1.46
NGW-CVT	0.102^{EF}	0.062^{CD}	0.047^{CD}	**0.079**	0.217	1.86	1.47
FNG	0.031	0.009^{E}	0.005^{E}	0.015	-	1.54	0.71
FPin	0.040^{G}	0.016^{E}	0.007^{E}	0.017	-	1.88	1.09
CPin	**0.115**^{EF}	**0.076**	**0.058**	0.077	0.215	1.80	1.48
Im2Text	0.017^{GH}	0.001^{F}	0.000^{F}	0.000	-	1.90	0.91
Im2Text-DeCAF	0.018^{GH}	0.001^{F}	0.000^{F}	0.000	-	**2.17**	0.76
Original	-	-	-	-	-	2.22	2.07

checked by three independent people, using two criteria: readability and relevance. The readability measures whether a sentence is understandable for a human, while the relevance measures to which extent the sentence is meaningful for the associated image. Each criterion had four levels: not (0), low (1), medium (2) and high (3). For each pin, Crowdflower returns the label with the highest confidence based on the input from the judges. In addition, we also evaluate the readability and relevance of the ground truth comments to have a better perspective on the results obtained with our models.

All the experiments to generate the comments were run on a Linux PC with a 3.4 Ghz Core i7 processor and with 16GB in RAM. The implementation of the methods was done in Java and MatLab.

B. Results

Table 1 presents the results of the automatic and manual evaluation of the proposed and baseline methods, and the manual evaluation of the original commentaries. The first five columns in the table show the results for the BLEU score, and the error percentage in board assignation. BLEU-n represents the BLEU score considering n-grams of size n= 1, 2, 3 when calculated and using an exact match between words. The values shown are averages of the BLEU scores for all the test pins. The BLEU-3 ratio represents the percentage of times a model obtains values greater than 0 for BLEU-3. For both this ratio and BLEU-n, higher values are better, with 1 as the maximum. With the values of the BLEU-n scores we conducted pair-wise Wilcoxon signed rank tests, in order to determine statistical significance in the differences among models. We used a significance level of α=0.05 and a Bonferroni correction by the number of comparisons (α/55). Superscripts indicate the groups of values that are no significantly different. The Board Error represents the percentage of incorrect board assignments (the lower the better, with zero as the minimum). The last two columns in the table are for the results of the manual inspection by independent human evaluators in Crowdflower of 200 pins and their generate/original comments. These results are expressed as the mean of readability and relevance of the comments (with the 3 as the maximum). The best results for each measure are marked in bold.

By analyzing the first four columns in Table 1, we observe that model CPin is the best performing approach in terms of BLEU scores, although the BLEU-1 values for this model are not significantly different to the values of the models NGW-CV and NGW-CVT. The good performance of CPin is due to the fact that when transferring a complete pre-defined comment of a user, there is a higher probability of capturing text parts that appear also in the ground truth comment. Regarding our approach we notice that all the NG- models exhibit statistically similar BLEU values, with NG-CVT scoring slightly higher. The same behavior statistically similar is observed for the group of NGW- models, with NGW-CVT presenting the best performance over both NG- and NGW-. These values indicate that our models, just by relying on sentence parts, are capable of capturing the semantic and style of the users' comments. The multimodal combination of visual and textual information during clustering has a positive effect that can be seeing in the results of the models –CVT, nevertheless the difference regarding the models –CV and –CT does not seem to be significant. Nonetheless, we believe that when testing the models with more pins, the significance would be higher. Similarly, re-weighting the n-grams based on local distances also has a positive impact, which could be seen in the better performance of the models NGW-.

The Board Error in the sixth column of the table indicates a percentage of the incorrect board assignments always lower than 22%. This is positive, since correctly identifying the board of a pin restricts the local neighborhood of the test image, helping to identify the topic of the pin and generate comments related with this topic.

To better put the BLEU score in context, we indicate in the fifth column of the table the ratio between the number of test images where BLEU-3 is higher than zero and the total number of test images. We observe that the ratio is extremely low, with less than 10% of the original test images' comments being replicated in the generated comments, providing evidence on the difficulty of the task. In addition, we notice that CPin has lower ratio than NGW-CV and NGW-CVT models, but at the same time has a higher score for BLEU-3 than the other two models. This indicates that CPin generates high BLEU-3 scores on few samples (it generates comments that better match with the original comments), while NGW-CV and NGW-CVT models generate BLEU-3 scores greater than zero more often, but with low values. Part of the good performance of CPin with the BLEU score is due to the fact that, by transferring a pre-defined comment, it allows for more flexibility in length of the generated text length than our NG- and

NGW- models which rely on a fixed length (average number of words per cluster). Since the BLEU measure tends to promote short sentences, our models are penalized more often.

The last two columns of Table 1 report the results of the Crowdflower evaluation for legibility and relevance for the generated and ground truth comments. In the table we observe that Im2Text-DeCAF produced the highest results in legibility, followed by NGW-CT. With this, it is clear that our approach is the most suitable to generate legible comments when is not possible to acces an external information source, such as the Flickr's filtered dataset used in Im2Text, which contains images with descriptions that directly reflect the visual content. The best result in terms of relevance is achieved by NG-CVT, followed by CPin. The importance of the clustering process is again evident here when comparing the results for relevance of FNG, FPin, Im2Text and Im2Text-DeCAF with the result of our NG- and NGW- models. Clustering allows to transfer portions of sentences (n-grams) between images that share visual and textual similarities. This content exchange enriches the generated comment and increases the chances of getting a higher relevance regarding the image. On the other hand, values of readability and relevance assigned by the evaluators to the original comments, put all the other results in perspective. By observing the results in the table, it becomes clear that the original comments can also be noisy (not readable) and sometimes be difficult to associate with the image they belong to (not relevant). This shows again how complex and challenging the task is.

For the qualitative evaluation of the obtained results, we repot in Table 2 some comments generated automatically by the different models and we compare them with the original comments. In these examples we observe that some sentences would produce a partial value for the BLEU scores (i.e. comments of NG-CV, NG-CVT and NGW-CV for images 2 and 3), while some of them would produce values of 0 for the BLEU, even if they are valid/relevant comments (i.e. comments by NG-CV, NG-CVT and NGW-CVT for images 1 and 4). The effect of incorrectly assigning the board to a test image can be seen in the comment generated by NG-CT for image 5, where the comment is off the topic. Also, in the table we observe that FNG produces noisy results, whilst FPin returns short and general comments. This was expected since such models do not use the visual information, and thus tend to mix users' topics. On the other hand, CPin also produce short comments, but these tend to be more related to the images. Finally, Im2Text and Im2Text-DeCAF, which rely on external sources, produces well-structured comments, however, they usually do not capture the meaning of the images.

VI. CONCLUSIONS

In this work we introduced a first effort in the direction of automatically generating image comments that may contain subjective opinions and emotions together with descriptions of visual content, all expressed with specific users' personal writing styles.

In order to tackle this task we used a dataset from the social network Pinterest, and we presented an approach based on jointly combining the visual and textual content from pins published by the users; taking advantage of the natural organization of pins in users' relevant topics in order to cluster the pins, and then using a n-gram language model to extract portions of textual content specific for each user and each cluster. During the test phase, a new image is assigned to one of these cluster using only its visual information and then a new comment is generated by combining the n-grams associated with the assigned cluster. We introduced a diversity of models that use the same strategy but vary in the way they utilize textual and visual data.

The results obtained by our method are truly encouraging. The models were able to produce reasonable results in terms of the BLEU score, which is a very strict measure, and good results (according to human evaluators) in terms of readability and relevance of the generated comments. This opens various possibilities, since the understanding of the user's writing style could help several other tasks, like user identification or personalized advertisement. Further research directions include first the modelling of the task as a structured supervised learning problem. The idea would be to predict a set of sentences given the visual features, and in a final stage to combine the sentences to form a comment. In this case a larger training dataset would be needed to correctly find associations between visual features and text. Second, the inclusion of external sources (similar to the Im2Text model), or the selection and combination of data from multiple similar users, could help to enrich the content of the final comments. Third, the use of templates may further help generating more readable comments by imposing a grammatical structure. Finally, it would be interesting to explore how product recommendations may become more personal and relevant. We speculate that by emulating the writing style of potential costumers it would be possible to engage the audience by "speaking their same language".

ACKNOWLEDGEMENTS

Research for this article was partially funded by the PARIS (IWT-SBO-Nr. 110067) project.

REFERENCES

[1] doc U.S. SEC. 2011. http://www.sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm.

[2] 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," arXiv preprint arXiv:1310.1531, 2013.

[3] A. C. Gallagher, J. Luo, C. G. Neustaedter, T. Chen, and L. Cao, "Image annotation using personal calendars as context," in 2008 ACM International Conference on Multimedia (ACMMM), 2008, pp. 681-684.

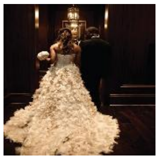
[4] M. M. Kalayeh, H. Idrees, and M. Shah, "Nmf-knn: Image annotation using weighted multi-view non-negative matrix factorization," in 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 184–191.

[5] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating image descriptions," in 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1601–1608.

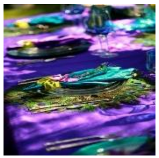
[6] X. Li, E. Gavves, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Personalizing automated image annotation using crossentropy," in 2011 ACM International Conference on Multimedia (ACMMM), 2011, pp. 233–242.

TABLE II. SAMPLE IMAGES, COMMENTS GENERATED BY DIFFERENT MODELS AND ORIGINAL COMMENTS PUBLISHED BY USERS.

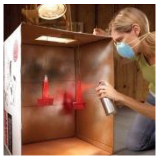
1



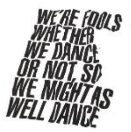
2



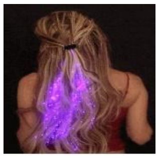
3



4



5



Model	**Comments**				
NG-CT	Repin by my wedding princess charlene of monaco beautiful gown wow love wedding dress lace pronovias so pretty	Cool	Tried and true decorating rules how to paint laminate furniture pretty much the best	I am in repair i'm dalai lama love hope	A very unusual amethyst geode iphone case cute purple house pansies
NG-CV	Whoa that's a dress alright	Jim zuckerman fruits stunning color splash rainbow of color cool	Entire page of mess free painting tips	Thank you inspiration at its best	Zuhair murad couture 2011 fausto sarli charles michael riley silk princess purple fairy
NG-CVT	So pretty	Jim zuckerman magic night aurora borealis love this stunning color splash rainbow of color cool	Entire page of mess free painting tips	Words to live by thank you true love	Zuhair murad couture 2011 fausto sarli charles michael riley princess purple fairy purple reign
NGW-CT	Repin by my wedding young and elegant wedding wow love wedding dress beautiful gown pronovias lace so pretty	Cool	Pretty much the best website ever knockoff diys of retail decor anthropologie	Quotes if you're afraid to success repeat each morning dalai lama hope love	Ruby falls by nikki pike tanzanite cute purple house pansies
NGW-CV	Whoa that's a dress alright i'd beautiful gown	Jim zuckerman rainbow of color splash fruits stunning	Entire page of mess free painting tips	Thank you quotes if you're afraid true	Zuhair murad couture 2011 elie saab
NGW-CVT	So pretty	Jim zuckerman magic night aurora borealis love this fruits tunning cool	Entire page of mess free painting tips	Words to live by thank you true love	Zuhair murad couture 2011 elie saab
FNG	In love with	Balanced rock in the garden	An amazing organization site i could spend hours on this is the primer	I am in repair	Balanced rock in the garden
FPin	Purple	Purple	Tried and true decorating rules	Dude	Purple
CPin	So pretty	Stunning	Entire page of mess free painting tips	Repeat each morning	Elie saab
Im2Text	No sleep, 27 hours on a bus, 5 hours on a train and 4 hours walking around prague in the sun. Lovely	Zinnias in rock wall	One of the many intricately carved wooden doors in lamu	Colorful wall tiles marking the street names in santo domingo	The dog in front of the main door of the cathedral of aparecida
Im2Text-DeCAF	Aisha in gold and white dress	The ocean spills over the rocks filling the colorful pools of life below.taken at golden point in palos verdes, california.	The pig which has it's own house in our village ha ha	Colorful wall tiles marking the street names in santo domingo	Lipsy the cat spent a lot of the holiday being carried around in a box by daisy
Original	Love this train	Color Splash! Pretty & Vibrant	Entire page of mess-free painting tips!	Dance!	Awesome hair style for a wild party!

[7] A. Makadia, V. Pavlovic, and S. Kumar, “Baselines for image annotation,” International Journal of Computer Vision, vol. 90, no. 1, pp. 88–105, 2010.

[8] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” International Journal of Computer Vision, vol. 42, no. 3, pp. 145–175, 2001.

[9] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2text: Describing images using 1 million captioned photographs,” in 2011 Advances in Neural Information Processing Systems (NIPS), 2011, pp. 1143–1151.

[10] H. Ordonez, J. C. Corrales, and C. Cobos, “Business processes retrieval based on multimodal search and lingo clustering algorithm,” Latin America Transactions, IEEE, vol. 13, no. 3, pp. 769-776, 2015

[11] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in 2002 Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 311–318.

[12] K. Ramnath, S. Baker, L. Vanderwende, M. El-Saban, S. Sinha, A. Kannan, N. Hassan, M. Galley, Y. Yang, D. Ramanan, A. Bergamo, and L. Torresani, “AutoCaption: Automatic caption generation for personal photos,” in 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), 2014, pp. 1050–1057.

[13] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: An astounding baseline for recognition,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 806-813.

[14] N. Sawant, R. D. J. Li, and J. Z. Wang, “Quest for relevant tags using local interaction networks and visual content,” in 2010 International Conference on Multimedia Information Retrieval (ICMR), 2010, pp. 231–240.

[15] J. Tang, R. Hong, S. Yan, T. S. Chua, G. J. Qi, and R. Jain, “Image annotation by knn-sparse graph-based label propagation over noisily tagged web images,” ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 2, pp. 1–14, 2011.

[16] J. Weston, S. Bengio, and N. Usunier, “Large scale image annotation: learning to rank with joint word-image embeddings,” Machine Learning, vol. 81, no. 1, pp. 21–35, 2010.

[17] S. Zoghbi, G. Heyman, J. C. Gomez and M. F. Moens, “Cross-modal attribute recognition in fashion,” in 2015 NIPS Multimodal Machine Learning Workshop (MMML), 2015.

[18] S. Zoghbi, G. Heyman, J. C. Gomez and M. F. Moens, “Fashion meets computer vision and NLP at e-commerce search", International Journal of Computer and Electrical Engineering, vol. 8, no. 1, pp. 31–43, 2016.

[19] S. Zoghbi, G. Heyman, J. C. Gomez and M. F. Moens, “Cross-Modal Fashion Search,” in 2016 International Conference on MultiMedia Modeling (MMM), 2016, pp. 367–373

Juan Carlos Gomez, received a M.Sc. degree in Astrophysics and Ph.D. degree in Computer Science from INAOE, Mexico, in 2002 and 2007 respectively. He was a postdoctoral researcher at KU Leuven, Belgium, from 2008 to 2009 and from 2011 to 2015, and at ITESM, Mexico, during 2010. He is in the process of being appointed as a full professor at the Department of Electronics at University of Guanajuato, Mexico. His research interests are machine learning, data mining, evolutionary computation and information retrieval, areas where he has published several peer-reviewed papers. He is a member of the National Researcher System (SNI level 1) in Mexico.

Tatiana Tommasi, is a research assistant at the University of North Carolina at Chapel Hill (USA). Her research interests include machine learning and computer vision with a focus on knowledge transfer and object categorization using multimodal information. She completed her Ph.D. in Electrical Engineering at the Ecole Polytechnique Federale de Lausanne (EPFL, Switzerland) in 2013 and was a postdoc at KU Leuven (Belgium) from 2013 to 2015. She is (co)author of more than 20 peer-reviewed papers.

Susana Zoghbi, is a Ph.D. student in Computer Science at the KU Leuven. She obtained a M.Sc. degree from the University of British Columbia in 2011. Her research interests lie at the boundary of computer vision and natural language processing, and include deep learning, topic modeling and graphical models. More information can be found at http://people.cs.kuleuven.be/~susana.zoghbi/.

Marie-Francine (Sien) Moens, is a full professor at the Department of Computer Science at KU Leuven, Belgium. She holds a M.Sc. and a Ph.D. degree in Computer Science from this university. She is head of the Language Intelligence and Information Retrieval (LIIR) research group. Her main interests are in the domain of automated content recognition in text and multimedia data and its application in information extraction and retrieval using statistical machine learning, and exploiting insights from linguistic and cognitive theories. She is currently a member of the Council of the Industrial Research Fund of KU Leuven and is the scientific manager of the EU COST action iV&L Net (The European Network on Integrating Vision and Language).