

# 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. INTRODUCCIÓN

EL USO de imágenes digitales en nuestra vida diaria ha cambiado mucho en los últimos años. Por un lado nos hemos convertido en proveedores prolíficos de imágenes gracias a la creciente popularidad de las cámaras digitales y los teléfonos inteligentes. Por el otro, también nos hemos convertido en consumidores ávidos y activos: casi cualquier sitio web está enriquecido con fotos o imágenes, además de que cuando navegamos en la Web las imágenes son las que comúnmente atraen más nuestra atención. Las redes sociales como Facebook and Pinterest han contribuido a impulsar estas tendencias, como lo confirma el hecho de que más de 300 millones de fotos por día se han subido a Facebook desde 2012 [1]. A su vez, Pinterest permite a sus usuarios crear tableros de marcadores visuales llamados *pins*. En estos tableros cada usuario puede coleccionar y guardar las imágenes que se encuentra en línea con el propósito de compartir contenido, planear viajes, coleccionar recetas, etc. Las imágenes que coleccionan los usuarios en Pinterest son comúnmente publicadas en sus tableros junto con comentarios cortos en forma de texto. A diferencia de una anotación con etiquetas de imágenes, o de una descripción detallada del contenido visual de las mismas, los comentarios que los usuarios ponen en este sitio suelen expresar emociones, opiniones o pensamientos, en conjunto con una descripción superficial de lo que aparece en las imágenes. Tales comentarios puedan dar una pista sobre los intereses personales de los usuarios, su forma de pensar y su estilo de escribir o expresarse (ver Fig. 1). Aunque explorar toda esta información puede tener un rol clave en diversas

tareas como la creación de modelos de usuarios, análisis de sentimientos y publicidad personalizada, hasta donde sabemos este tipo de comentarios específicos aún no ha sido completamente estudiado.



*Love this train*



*Hinged cabinet. Love this to hide appliances*

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

En este trabajo queremos dar un paso más en el análisis del contenido generado por el usuario con tres contribuciones principales: 1) Introducimos una nueva tarea: predecir automáticamente los comentarios de los usuarios para sus imágenes; 2) presentamos una nueva colección de datos (consistente en más de 70,000 imágenes junto con los comentarios de los usuarios); y 3) realizamos un estudio experimental extensivo para la tarea.

El resto del artículo está organizado de la siguiente manera: en la sección 2 hacemos una breve revisión de la literatura; en la sección 3 describimos la tarea propuesta y la colección de datos utilizada para los experimentos; en la sección 4 presentamos y discutimos nuestro enfoque; en la sección 5 mostramos los resultados experimentales; y en la sección 6 concluimos el artículo con una discusión general y posibles líneas futuras de investigación.

## II. TRABAJO RELACIONADO

Existen dos líneas principales en la literatura relacionada con asociar automáticamente texto a imágenes. La primera se enfoca en asignar palabras clave (o etiquetas) a una imagen; la segunda en proveer una descripción completa de la imagen. En ambos casos el objetivo final es reconocer el contenido de la imagen en términos de los objetos [7] o de la escena [5] mostrados. Para ello, los trabajos previos utilizan principalmente estrategias basadas en contenido, las cuales predicen el texto para las imágenes ya sea al entrenar un modelo con la relación entre texto e imágenes [16, 17, 18, 19], o al propagar las etiquetas a través de un método de vecinos más cercanos [4, 15]. En estos estudios, el análisis se realiza generalmente en colecciones de datos donde las anotaciones para describir las imágenes están bien definidas gracias a un concurso entre usuarios o expertos.

---

J. C. Gomez, KU Leuven, Belgium, jcgcaranza@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

La tarea de anotar imágenes con comentarios de usuarios ha comenzado a emerger más recientemente en las bibliografías de visión por computadora y procesamiento de lenguaje natural [6, 14]. Esta tarea desafía a la anotación automática estándar en dos aspectos: primero, las colecciones personales de los usuarios frecuentemente contienen sólo un número limitado de imágenes; y segundo, el texto asociado suele reflejar los intereses del usuario y su estilo, siendo una combinación de descripciones objetivas y expresiones subjetivas. Los métodos existentes superan la primera cuestión mediante el uso de información externa: tales como calendarios personales o meta datos sobre localización y fechas [3, 12]. La segunda cuestión es afrontada con la explotación del historial de anotaciones que el usuario ha hecho, enfocándose en el uso de palabras cuya definición se acomoda mejor al usuario, p.ej. utilizar la palabra *kitty* (gatito en inglés), en vez del concepto más genérico *cat* (gato en inglés) [14]. En este trabajo, intentamos mover el límite de anotación automática y nos enfocamos en la generación de oraciones similares a las que un usuario publicaría en las redes sociales para acompañar una imagen.

### III. DETALLES DEL PROBLEMA Y LA COLECCIÓN DE DATOS

En Pinterest los usuarios publican pins que están organizados en tableros. Un pin es un par  $\langle \text{imagen}, \text{texto} \rangle$ , y un tablero agrupa diversos pins sobre un tema definido por el usuario, el cual puede ser específico (p.ej. alguien famoso) o un concepto general (p.ej. comida o ropa). Sea  $u$  un usuario y  $\mathbf{P}_{i,j}^u = (\mathbf{g}_{i,j}^u, \mathbf{x}_{i,j}^u)$  su colección de pins, donde  $\mathbf{g}$  se refiere a la imagen y  $\mathbf{x}$  al texto. Los índices  $j$  y  $i$  identifican respectivamente al pin específico  $j=1, \dots, L_i^u$  y al tablero al cual pertenece  $i=1, \dots, M^u$ .  $M^u$  indica el número de tableros del usuario  $u$ , y  $L_i^u$  el número de pins dentro del tablero  $i$ . Finalmente  $n_u = \sum_{m=1}^{M^u} L_m^u$  es el número total de pins del usuario  $u$ .

Para este trabajo recolectamos al azar 70,200 pins pertenecientes a 117 usuarios mediante una exploración directa del sitio web de Pinterest. Seleccionamos 3 tableros por usuario y guardamos 200 pins por tablero, para un total de 600 pins por usuario. 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. Dentro de las 10 palabras (no vacías) más frecuentes usadas en la colección se encuentran: *love*, *easy*, *great*, *cute* y *beautiful*, lo que muestra que los usuarios intentan más expresar una opinión o sentimiento que describir lo que contiene la imagen. El tamaño de las imágenes también es variable, pero todas están almacenadas en formato JPG. Hay 4.2% de imágenes que son compartidas por 2 o más usuarios, lo que muestra que hay una gran diversidad de contenido en los pins, incluyendo productos (p.ej. ropa y joyería), intereses (p.ej. comida y decoraciones), fotografías (p.ej. animales y paisajes) y contenido más abstracto (p.ej. pinturas y diseños).

Toda la información textual (comentarios) y visual (imágenes) de los pins fueron procesadas de la siguiente forma: limpiamos cada comentario mediante la eliminación de símbolos especiales (asterisco, hashtag, etc.), urls y palabras de una sola letra (con las excepciones de  $i$  y  $a$ ), para formar cada vector textual  $\mathbf{x}_{i,j}^u$  como una cadena de palabras separadas por espa-

cios. Procesamos cada imagen utilizando una red neuronal convolucional [13] para obtener  $\mathbf{g}_{i,j}^u$  como un vector de características visuales. Para esto último utilizamos la librería DeCAF [2], considerando los valores de activación de las 4,096 neuronas en la séptima capa de la red como las características de la imagen.

Para los experimentos dividimos los pins de la colección en un conjunto de entrenamiento y uno de prueba. Seleccionamos al azar 10 pins por tablero por usuario (30 pins por usuario) para formar el conjunto de prueba (3,510 pins en total), los restantes 66,690 pins se utilizaron como conjunto de entrenamiento. Durante el entrenamiento, tanto la información textual como la visual están disponibles. Durante la fase de prueba, únicamente la parte visual  $\mathbf{g}^u$  de los pins de prueba para el usuario  $u$  está disponible (sin información del tablero del que provienen); nuestra meta es entonces generar automáticamente los comentarios asociados  $\mathbf{x}^u$  para tales imágenes.

### IV. MÉTODO

La organización original de los pins por tableros provee de entrada ciertos indicios sobre los comentarios que un usuario puede publicar. Por ejemplo, un comentario del tipo “*amazing outfit*” queda bien en un tablero que trata sobre ropa, y algo como “*that looks delicious*” en un tablero sobre comida. Nuestra propuesta para asociar comentarios a nuevas imágenes consiste en aprovechar esta estructura de los datos y definir un método basado en la combinación de agrupar pins de forma multimodal [10] y transferir texto. Un esquema general de nuestra propuesta se muestra en la Fig. 2 y a continuación describimos los dos pasos importantes.

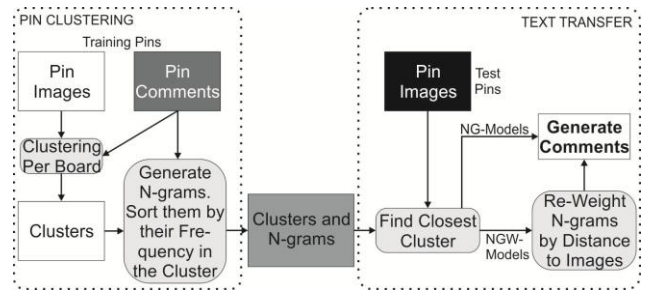


Figura 2. Descripción general de los métodos NG- y NGW- para la generación automática de comentarios para imágenes.

**Agrupamiento de pins.** En esta fase consideramos de forma separada cada tablero  $i$  del usuario  $u$ , y agrupamos sus pins de entrenamiento. El agrupar los pins de cada tablero nos permite crear grupos específicos y mapeos más finos entre imágenes y comentarios. En lo siguiente mantenemos fijo al usuario y para simplificar la notación quitamos el superíndice  $u$ . El proceso de agrupamiento es multimodal, es decir utiliza la información visual y textual en conjunto. Más específicamente: primero calculamos la distancia euclidiana entre todas las imágenes y todos los comentarios, obteniendo dos matrices de distancias  $D_g, D_x \in \mathbb{R}^{L_i \times L_i}$  para imágenes y comentarios respectivamente.

Segundo, la contribución de cada matriz se combina linealmente por medio de un parámetro  $\lambda$  de la siguiente forma:  $D = \lambda D_g + (1 - \lambda) D_x$ . Finalmente, realizamos un agrupamiento

jerárquico con ligamiento promedio sobre la matriz  $D$  para crear un conjunto final  $C_i$  de grupos de pins por tablero. Cada grupo  $c_i=1, \dots, C_i$  contiene  $l_{c_i}$  pins similares entre sí.

Enseguida, para cada grupo  $c_i$  calculamos el centroide visual como el promedio de los vectores de características visuales del grupo. Guardamos cada centroide junto con otros dos metadatos: el número promedio  $\bar{w}_{c_i}$  de palabras en los comentarios dentro del grupo  $c_i$  y un modelo de lenguaje por n-gramas. Este modelo consiste en extraer todos los posibles n-gramas  $k=1, \dots, K_{c_i}$  de los comentarios del grupo  $c_i$ , calculando además la frecuencia de cada n-grama  $f_{c_i,k}$  dentro del grupo y guardando los índices de los pins donde ese n-grama aparece.

**Transferencia de texto.** En esta fase, dada una imagen de prueba, primero identificamos el grupo  $c^*$  más cercano a la imagen, usando la distancia euclidiana estimada con respecto a los centroides visuales. Una vez en el grupo  $c^*$  calculamos las distancias entre la imagen de prueba y cada una de las imágenes del grupo  $d_{t,j}$ ,  $j=1, \dots, l_{c^*}$  y usamos estas distancias para pesar las frecuencias de los n-gramas. Las frecuencias pesadas se usan a su vez para construir el comentario para la imagen de prueba. Para ello primero ordenamos de forma descendente los n-gramas en base a su frecuencia pesada  $\tilde{f}_{c^*,k} = f_{c^*,k} / d_{t,j^k}$ , donde  $d_{t,j^k}$  es la distancia de la imagen de prueba a la imagen más cercana cuyo comentario contiene el  $k$ -ésimo n-grama; después creamos el comentario para la imagen de prueba iterativamente utilizando los n-gramas ordenados. El comentario inicia como una cadena de texto vacía, concatenamos un n-grama en cada paso de la iteración y revisamos el comentario intermedio por redundancia para eliminar 1-2, ..., n-gramas repetidos. El proceso de agregar nuevos n-gramas y limpiar el comentario continua hasta que éste tiene una longitud igual o mayor que  $\bar{w}_{c^*}$  (la longitud promedio de los comentarios en el grupo  $c^*$ ). Finalmente, eliminamos conjunciones (*and, or, but, for, so, etc.*) del final de la frase. Una vez limpio, transferimos el comentario a la imagen de prueba.

Llamamos a nuestro método propuesto como NGW-CVT: *N-Gram Weighted transfer by Clustering over Visual and Textual Information* (Transferencia de N-Gramas Pesados por medio del Agrupamiento de Información Textual y Visual). Dentro del proceso descrito anteriormente, investigamos casos específicos y diversas variantes. Primero, al modificar el parámetro  $\lambda$  durante el agrupamiento, podemos ajustar la importancia de la información visual o textual. Con esto creamos tres modelos:

**NGW-CT** – Con  $\lambda=0$ , utiliza únicamente la información textual de los pins para realizar el agrupamiento.

**NGW-CV** – Con  $\lambda=1$ , utiliza únicamente la información visual de los pins para realizar el agrupamiento.

**NGW-CVT** – Con  $\lambda=0.5$ , utiliza en la misma proporción la información visual y textual para realizar el agrupamiento.

Segundo, cuando transferimos el texto a una imagen, podemos utilizar directamente la frecuencia no pesada de los n-gramas. De esta forma obtenemos también tres modelos:

**NG-CT** – Con  $\lambda=0$ , utiliza únicamente la información textual de los pins para realizar el agrupamiento.

**NG-CV** – Con  $\lambda=1$ , utiliza únicamente la información visual de los pins para realizar el agrupamiento.

**NG-CVT** – Con  $\lambda=0.5$ , utiliza en la misma proporción la información visual y textual para realizar el agrupamiento.

Estos tres últimos modelos simplifican las fases de entrenamiento y prueba: durante el entrenamiento podemos omitir los índices de los pins de los n-gramas; durante la prueba no es necesario calcular las distancias entre las imágenes de prueba y las imágenes del grupo. Sin embargo, tienen la desventaja de que el mismo comentario será transferido a todas las imágenes de prueba que sean asignadas a un grupo específico.

## V. EXPERIMENTOS

En esta sección presentamos los detalles del análisis experimental con la colección de datos de Pinterest. Empezamos por definir los valores de los parámetros para el método propuesto y por describir los modelos de base para comparar (Sección V.A); y después presentamos y discutimos los resultados obtenidos (Sección V.B).

### A. Parámetros de los modelos y modelos de referencia

Los modelos propuestos tienen dos parámetros principales. El primero es el tamaño  $n$  de los n-gramas a extraer de los pins de entrenamiento. Fijamos este parámetro experimentalmente a  $n=4$ , puesto que n-gramas más grandes son difíciles de combinar y tienden a producir comentarios complicados de entender. El segundo parámetro es el número de grupos a crear para cada tablero. En este caso consideramos  $C_i^u=60$ , un valor que produjo buenos resultados en una conjunto de pins de validación extraídos del conjunto de entrenamiento.

Para efectos de comparación, evaluamos el desempeño de nuestros modelos con respecto al de cuatro modelos de referencia. Dos de ellos utilizan solo la información textual de los pins. Trabajos previos indican que el historial de anotaciones de un usuario provee información suficiente para predecir anotaciones futuras, independientemente del contenido de la imagen [14]. Siguiendo esta idea consideramos dos modelos:

**FNG** – Este método calcula la frecuencia de cada n-grama dentro de todos los pins  $n_u$  del usuario  $u$ , sin considerar ni la imagen asociada ni el tablero al que pertenece el pin. Después aplica el mismo procedimiento para mezclar n-gramas de los modelos NGW-, considerando el promedio de palabras de todos los comentarios del usuario como longitud máxima permitida para el comentario a generar. El comentario obtenido se asigna a todas las imágenes de prueba.

**FPin** – Puede ocurrir que múltiples pins tengan exactamente el mismo contenido textual. Este método cuenta cuántas veces se repite cada comentario y asigna el más frecuente a todas las imágenes de prueba.

Los otros dos modelos de referencia están basados totalmente en la información visual. El primero se enfoca en cada usuario en específico y en sus pins de entrenamiento. El segundo explota una fuente de información externa:

**CPin** – Este método, después del agrupamiento inicial por tablero, asigna la imagen de prueba al grupo más cercano  $c^*$ . Dentro de  $c^*$  busca la imagen que sea más parecida a la imagen de prueba y el comentario asociado a la imagen encontrada es transferido completo a la nueva imagen sin utilizar las estadísticas de los n-gramas del grupo.

TABLA I. RESULTADOS DE LAS EVALUACIONES AUTOMÁTICA Y MANUAL DE LOS COMENTARIOS GENERADOS POR LOS DIFERENTES MODELOS Y DE LOS COMENTARIOS ORIGINALES.

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

**Im2Text** – Este método fue introducido en [9] para crear automáticamente descripciones de imágenes utilizando como referencia una colección de 1 millón de fotos etiquetadas de Flickr. Para una imagen de prueba, el método calcula la similitud de esta imagen con todas las fotos de Flickr, encuentra la más parecida y transfiere las etiquetas de esta foto a la imagen de prueba. La similitud se calcula en base a pistas visuales que utilizan las características gist [8] y los valores de los píxeles de imágenes miniaturas (reducidas a un tamaño de 32x32). En este trabajo usamos el código proporcionado por los autores (<http://vision.cs.stonybrook.edu/vicente/sbucaptions/>). Adicionalmente, puesto que la representación utilizada en este modelo es de alto nivel, también probamos una versión refinada del mismo: después de haber obtenido 20 imágenes candidatas de Flickr por imagen de prueba, seleccionamos la más parecida basándonos en la similitud evaluada con las características DeCAF calculadas para el conjunto de imágenes candidatas; llamamos a este método Im2Text-DeCAF.

Evaluamos todos los modelos de dos formas. La primera es una evaluación automática calculando la medida BLEU [11] entre un comentario generado y el comentario real. BLEU es una medida muy estricta que evalúa la coincidencia exacta de palabras, por lo cual no considera la posibilidad de tener un comentario generado que sea válido para una imagen sin que éste coincida exactamente con el comentario original. Debido a esto, aplicamos una segunda evaluación de más alto nivel. Para ello, seleccionamos 200 pins al azar y los subimos a la plataforma Crowdfunder ([www.crowdfunder.com](http://www.crowdfunder.com)), donde el comentario de cada pin es revisado y evaluado por tres personas de forma independiente utilizando dos criterios: legibilidad y relevancia. La legibilidad mide si un comentario es entendible para un humano, mientras que la relevancia indica en qué medida un comentario es válido para la imagen asociada. Cada criterio tiene cuatro niveles: no (0), bajo (1), medio (2) y alto (3). Para cada pin, Crowdfunder devuelve el nivel con la confianza más alta basado en la entrada de los evaluadores. Adicionalmente, también evaluamos la legibilidad y relevancia de los comentarios originales con respecto a sus imágenes; esto con la intención de tener una mejor perspectiva de los resultados obtenidos con nuestros modelos.

Todos los experimentos fueron realizados usando una computadora PC con Linux, un procesador Core i7 a 3.4 GHz y 16 GB de RAM. La implementación de los métodos fue hecha en Java y MatLab.

## B. Resultados

La Tabla 1 presenta los resultados de las evaluaciones automática y manual de los modelos propuestos y de referencia, además de la evaluación manual de los comentarios originales. Las cinco primeras columnas en la tabla muestran los resultados para la medida BLEU y el porcentaje de errores de asignación de tablero. BLEU-n representa la medida BLEU considerando n-gramas de tamaño n=1, 2, 3 cuando se calcula y utilizando una comparación exacta entre palabras. Los valores presentados son promedios sobre todos los valores BLEU para los pins de prueba. La Razón BLEU-3 representa el porcentaje de veces que un modelo obtiene valores mayores a cero para BLEU-3. Tanto para esta razón como para BLEU-n, valores más altos son mejores, siendo 1 el valor máximo. Para los valores de las medidas BLEU-n realizamos pruebas de los rangos con signo de Wilcoxon por parejas para saber si las diferencias entre modelos son significativas. Utilizamos un nivel de significancia de  $\alpha=0.05$  con una corrección de Bonferroni por el número de comparaciones ( $\alpha/55$ ). Las letras en los superíndices indican los grupos de valores que no son significativamente diferentes. El Error de Tablero representa el porcentaje de asignaciones incorrectas de tablero (entre más pequeño mejor, siendo 0 el valor mínimo). Las últimas dos columnas de la tabla corresponden a la evaluación manual de 200 pins y sus comentarios generados/originales por evaluadores humanos en Crowdfunder. Estos resultados están expresados como la media de legibilidad y relevancia de los comentarios (siendo 3 el valor máximo). Los mejores resultados para cada medida están marcados en negritas.

Analizando las cuatro primeras columnas de la Tabla 1 observamos que el modelo CPin es el que tiene mejor desempeño en términos de las medidas BLEU, aunque el valor BLEU-1 para este modelo no es significativamente diferente que los valores de los modelos NGW-CV y NGW-CVT. El buen desempeño de CPin se debe al hecho de que al transferir un comentario completo de un usuario, es más alta la probabilidad de capturar partes del texto que también aparezcan en el comentario original. Respecto a nuestro enfoque, vemos que todos los modelos NG- muestran valores de BLEU estadísticamente similares, con NG-CVT obteniendo valores un poco más altos. El mismo comportamiento estadísticamente similar se observa en los modelos NGW-, con NGW-CVT presentando el mejor desempeño de todos los modelos, tanto NG- como

NGW-. Estos valores indican que nuestros métodos, al hacer uso de porciones de oraciones, son capaces de capturar la semántica y el estilo de los comentarios de los usuarios. La combinación multimodal de información visual y textual durante el agrupamiento tiene un efecto positivo que puede verse en los resultados de los modelos -CVT, sin embargo la diferencia respecto a los modelos -CV y -CT parece no ser significativa. No obstante, creemos que al probar los modelos con más pins, la significancia sería más alta. Igualmente, el repesado de los n-gramas basado en distancias locales también tiene un impacto positivo, como puede ser visto en el mejor desempeño de los modelos NGW-.

El Error de Tablero en la sexta columna de la tabla indica un porcentaje de asignaciones incorrectas de tablero siempre menor a 22%. Esto es positivo ya que identificar correctamente el tablero al que pertenece un pin restringe el vecindario local de la imagen de prueba, ayudando a identificar el tópico del pin y a generar comentarios relacionados con este tópico.

Para poner los valores BLEU en contexto, en la quinta columna de la tabla mostramos la razón entre el número de imágenes de prueba para las que se generaron comentarios con un BLEU-3 mayor que cero y el número total de imágenes de prueba. Se observa que esta razón es muy baja, con menos del 10% de los comentarios originales de las imágenes de prueba siendo replicados en los comentarios generados, lo cual muestra la difícil que resulta la tarea propuesta. Adicionalmente, vemos que el modelo CPin tiene una razón más baja que los modelos NGW-CV y NGW-CVT, pero también tiene un valor para BLEU-3 que es más alto que los otros dos modelos. Lo anterior indica que CPin genera valores altos para BLEU-3 pero en pocas imágenes de prueba (esto es, genera comentarios que coinciden en mayor proporción con los comentarios originales), mientras que los modelos NGW-CV y NGW-CVT generan valores más bajos para BLEU-3 pero en muchas más imágenes. Parte del buen desempeño de CPin con la medida BLEU se debe a que al transferir un comentario predefinido, esto permite más flexibilidad en la longitud del texto generado que con respecto a los modelos NG- y NGW-, los cuales están basados en una longitud fija (el promedio de palabras por grupo). Puesto que la medida BLEU tiende a preferir oraciones cortas, nuestros modelos son penalizados más a menudo.

Las últimas dos columnas de la Tabla 1 reportan los resultados de la evaluación en Crowdfower por legibilidad y relevancia de los comentarios generados y los originales. En la tabla observamos que Im2Text-DeCAF produjo los mejores resultados en legibilidad, seguido de NGW-CT. Con ello, es claro que nuestro enfoque es el más adecuado para generar comentarios legibles cuando no es posible acceder a una fuente externa de información, tal como la colección de datos filtrados de Flickr utilizada en Im2Text, la cual contiene imágenes con descripciones que reflejan directamente el contenido visual. El mejor resultado para relevancia es obtenido por NG-CVT seguido de CPin. La importancia del proceso de agrupamiento es de nuevo evidente cuando se compara los resultados para relevancia de FNG, FPin, Im2Text y Im2Text-DeCAF con el resultado de nuestros modelos NG- y NGW-. El agrupamiento permite la transferencia de porciones de oraciones (n-gramas) entre imágenes que comparten similitudes textuales y visuales. Este intercambio de contenido enri-

quece el comentario generado e incrementa la probabilidad de obtener una mejor relevancia respecto a la imagen. Por otro lado, los valores de legibilidad y relevancia asignados por los evaluadores a los comentarios originales ponen todos los otros resultados en perspectiva. Al observar los resultados en la tabla, es claro que los comentarios originales también pueden contener ruido (son ilegibles) y a veces son difíciles de asociar con la imagen a la que pertenecen (no parecen ser relevantes). Esto muestra de nuevo lo complicado y desafiante que es la tarea propuesta.

Para la evaluación cualitativa de los resultados obtenidos, reportamos en la Tabla 2 algunos de los comentarios generados automáticamente por los diferentes modelos y su comparación con los comentarios originales. En estos ejemplos observamos que algunas oraciones producirían valores parciales para BLEU (p.ej. los comentarios de NG-CV, NG-CVT y NGW-CV para las imágenes 2 y 3), mientras que algunos producirían valores de 0 para BLEU aunque sean comentarios válidos/relevantes (p.ej. los comentarios de NG-CV, NG-CVT y NGW-CVT para las imágenes 1 y 4). El efecto de asignar incorrectamente un tablero para una imagen de prueba puede verse en el comentario generado por NG-CT para la imagen 5, donde este comentario está fuera del tópico de la imagen. También en la tabla observamos que FNG produce resultados con ruido, mientras que FPin produce comentarios cortos y generales. Esto era de esperarse puesto que tales modelos no utilizan la información visual y por lo tanto tienden a mezclar los tópicos de los usuarios. Por otro lado, CPin también produce comentarios cortos, pero éstos tienden a estar más relacionados con las imágenes. Finalmente, los modelos Im2Text y Im2Text-DeCAF, los cuales utilizan fuentes de información externa, producen comentarios bien estructurados, sin embargo, éstos suelen no captar el sentido de las imágenes.

## VI. CONCLUSIONES






En este trabajo presentamos un primer esfuerzo en la dirección de generar automáticamente comentarios sobre imágenes, los cuales puedan contener opiniones subjetivas, emociones y descripciones del contenido de las imágenes; todo expresado con estilos de escritura propios de usuarios específicos.

Para afrontar esta tarea, utilizamos una colección de datos de la red social Pinterest y presentamos un enfoque basado en combinar de forma conjunta el contenido visual y textual de los pins publicados por los usuarios de la red, aprovechando la organización nativa de los datos por tópicos de interés para los usuarios, para separar los pins en grupos, y luego utilizar un modelo de lenguaje basado en n-gramas para extraer porciones de contenido textual que son específicos de cada grupo y de cada usuario. Durante la fase de prueba, una nueva imagen se asigna a uno de estos grupos utilizando únicamente su información visual y después se genera un comentario al combinar los n-gramas asociados con el grupo asignado. Presentamos varios modelos que usan la misma estrategia pero que varían en la forma en que utilizan los datos visuales y textuales.

Los resultados obtenidos con nuestro enfoque son realmente alentadores. Los modelos fueron capaces de producir resultados razonables para la medida BLEU (que es muy estricta), y buenos resultados (según evaluadores humanos) para la legibilidad y relevancia de los comentarios generados. Esto abre



TABLA II. EJEMPLOS DE IMÁGENES, LOS COMENTARIOS GENERADOS POR LOS DIFERENTES MODELOS Y LOS COMENTARIOS ORIGINALES PUBLICADOS POR LOS USUARIOS.

	1	2	3	4	5
					
Modelo	Comentarios				
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!

varias posibilidades, puesto que entender el estilo de escritura de los usuarios puede ayudar en otras tareas, como la identificación de usuarios específicos o la publicidad personalizada.

Otras líneas de investigación interesantes para explorar son: primero, modelar la tarea propuesta como un problema de aprendizaje supervisado estructurado. La idea en ese caso sería predecir un conjunto de oraciones a partir de un conjunto de características visuales, y al final combinar las oraciones para formar un comentario. En este caso, sería necesario contar con un conjunto de entrenamiento más grande para poder encontrar de forma correcta las asociaciones entre las características visuales y el texto. Segundo, incluir fuentes externas de información (como el modelo Im2Text), o seleccionar y combinar los datos de varios usuarios con intereses y estilos similares, lo cual puede ayudar a enriquecer el contenido final de los comentarios. Tercero, utilizar plantillas de lenguaje que impongan una estructura gramatical con la finalidad de generar comentarios que sean más legibles. Finalmente, sería interesante explorar cómo hacer que la recomendación de productos sea más personal y relevante para los usuarios en redes sociales. Especulamos que emular el estilo de escritura de clientes potenciales permitiría enganchar mejor a la audiencia gracias a que podemos “hablar su mismo idioma”.

## AGRADECIMIENTOS

La investigación para este artículo fue parcialmente financiada por el proyecto PARIS (IWT-SBO-Nr. 110067).

## REFERENCIAS

- [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 cross-entropy,” in 2011 ACM International Conference on Multimedia (ACMMM), 2011, pp. 233–242.

- [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.



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).



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.