

# Proyecto Final: Dame una imagen y te digo que pasa

Catherine Díaz Umaña 200055670

Sorelys Sandoval Díaz 200055799

*Electrics and Electronics Engineering Department, Universidad del Norte  
Barranquilla, Colombia*

## I. INTRODUCCIÓN

Los generadores de leyendas para imágenes son una aplicación de los sistemas de aprendizaje de máquinas y procesamiento de imágenes. Usualmente estos sistemas consisten de una red neuronal (un conjunto de neuronas artificiales que imitan la forma como el cerebro humano resuelve problemas), toma una imagen y genera una descripción simple de la misma. La primera parte de la red es convolucional y esta se encarga de identificar los objetos en la imagen. Esto es seguido de una red recurrente, la cual recibe la salida de la primera red para producir las frases dependiendo de ello. El sistema debe ser entrenado con imágenes con sus correspondientes descripciones (que pueden tomarse de datasets abiertos), de esta forma se obtienen mejores resultados. La importancia de emplear imágenes con sus descripciones en el entrenamiento es que la variedad de las frases generadas y su estructura es mejor. Asimismo conceptos un poco más complejos se pueden relacionar mejor con patrones visuales gracias a este tipo de entrenamiento [7] (Ver Figura.1).

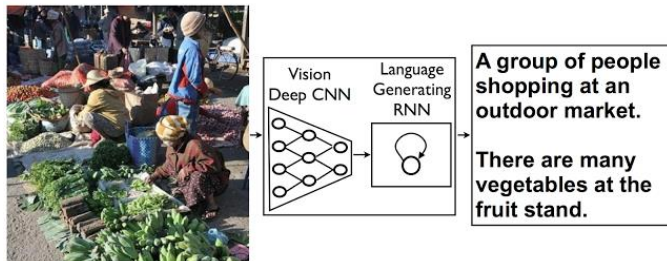


Fig. 1: Diagrama de bloques general de un generador de descripciones de imágenes [1]

Este tema es ampliamente estudiado hoy en día debido a que es de gran importancia por su gran número de aplicaciones. Por ejemplo, Puede ayudar a mejorar la búsqueda de imágenes, o a las personas con problemas de visión, ya que, teniendo la descripción general de la imagen, podrían entender que ocurre en la misma. También, con sistemas más avanzados, al proporcionarles una secuencia de imágenes se podría obtener una historia completa, y con resultados más acertados sistemas robóticos en campos como de carros autónomos, podrían hacer un mejor reconocimiento de sus alrededores y tomar mejores decisiones. Por otro lado, también podrían emplearse en educación a temprana edad [6].

Hay mucho trabajo que hacer en esta área y no es fácil, dado que aparte de reconocer los objetos se debe distinguir la relación entre los mismos, para lo que sus atributos y las actividades con las que están relacionados necesitan ser conocidas también,

pero las investigaciones en esto son muchas y los avances continuos.

## II. OBJETIVOS

### A. Objetivo General:

Implementar y analizar el funcionamiento de generadores de leyendas para imágenes simples con pocos sujetos y acciones.

### B. Objetivos específicos:

- Poner diferentes modelos pre entrenados a prueba y verificar su desempeño.
- Implementar una métrica de evaluación similar a las de los modelos oficiales, para verificar que tan bueno es cada modelo.

## III. TRABAJO RELACIONADO

Las siguientes son imágenes de programas que tienen el mismo propósito o función que los que vamos a poner a prueba. Se les dan imágenes de entrada con las que nunca entrenaron y el sistema les da descripciones relacionadas con las mismas. Es claro que la confiabilidad en los resultados puede variar.



Fig. 2: Ejemplo de imágenes con descripción generada. Grupo de investigación de Stanford.



Fig. 3: Ejemplo de imágenes con descripciones generadas con distintos grados de certeza, agrupadas por clasificación humana. Grupo de investigación de Google.

En los estudios los investigadores usualmente emplean diferentes datasets para validar la eficacia de su modelo [6]. Se sabe que entre más grandes sean los datasets que emplean para entrenar, los sistemas tienen más capacidad de dar mejores textos a mayor cantidad de imágenes. El otro punto importante es el poder de las redes que usan, ya que algunos llegan a decir que su sistema puede ser mejorado al incorporar redes profundas más fuertes [6].

Hay varias formas de generar las descripciones de imágenes. Algunos métodos dividen las frases en partes y luego cada parte se asocia a un objeto o atributo, similar a rellenar plantillas [6,7]. Otros métodos usan una densidad de probabilidad que permite generar frases más flexibles y con mejor estructura, como las máquinas de Deep Boltzmann [6].

Hay sistemas creados hoy en día, como el OFFICIAL MICROSOFT COCO BENCHMARK que han probado generar

descripciones con una calidad igual o mejor que las creadas por personas, un 34% de las veces. Al igual que al probarse con el método de evaluación BLEU-4 (que evalúa la calidad de traducciones automáticas o frases generadas, con buena robustez), obtienen una calificación de 29,1% [7].

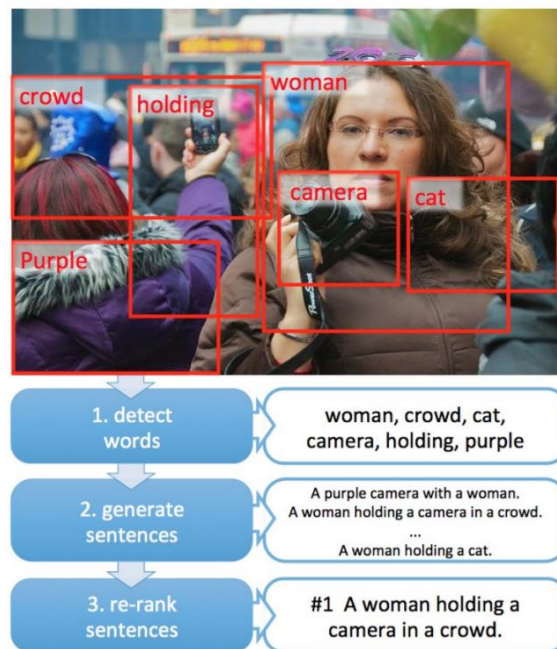


Fig. 4: Ejemplo diagrama de bloques de generador de descripciones de Microsoft.

#### IV. ACERCAMIENTO PROPUESTO

Se escogió implementar el modelo realizado por un grupo de investigación de Google denominado Show and Tell debido a que como se observa en la Figura 5, según la competencia de generadores de descripciones, es el que tiene mejores resultados [8].

	M1	M2	M3	M4	M5	date
Human <sup>[5]</sup>	0.638	0.675	4.836	3.428	0.352	2015-03-23
Google <sup>[4]</sup>	0.273	0.317	4.107	2.742	0.233	2015-05-29
MSR <sup>[11]</sup>	0.268	0.322	4.137	2.662	0.234	2015-04-08
Montreal/Toronto <sup>[10]</sup>	0.262	0.272	3.932	2.832	0.197	2015-05-14
MSR Captivator <sup>[12]</sup>	0.25	0.301	4.149	2.565	0.233	2015-05-28
Berkeley LRCN <sup>[2]</sup>	0.246	0.268	3.924	2.786	0.204	2015-04-25

Fig. 5: Ranking COCO Captioning Challenge 2015.

Este modelo sigue una arquitectura como la que se muestra en la Figura 6. Se divide en dos grandes etapas: la clasificación de los objetos por parte de una CNN (Convolutional Neural Network) y un conjunto de RNN (Recurrent Neural Network) encargadas de la formulación de la frase en lenguaje natural.

En la primera etapa capas de redes neuronales convolucionales se encargan de representar la imagen como un vector de features con tamaño fijo. Durante el entrenamiento cada palabra es asociada a uno de estos vectores de features con tamaño fijo y



de esa forma al final de la CNN el vector de features se lleva al espacio vectorial de palabras para que sirva de entrada a la etapa de RNNs.

En dicha etapa se usan redes LSTM o Long Short Term Memory networks las cuales son capaces de conectar información previa con la presente y de esa manera determinar las palabras y el orden de ellas en la descripción a generar.

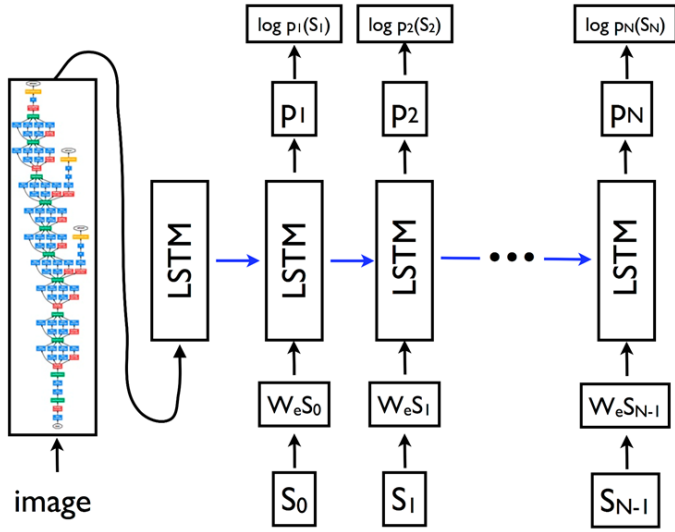


Fig. 6: Arquitectura de Show and Tell.

Las RNNs calculan la distribución de probabilidad de la siguiente palabra en la oración y al final se escoge la que presente la máxima respuesta de la función de verosimilitud.

$$\log p(S|I) = \sum_{t=0}^N \log p(S_t|I, S_0, \dots, S_{t-1})$$

El cálculo de la probabilidad depende tanto de las palabras generadas en el pasado como de la imagen y esta dependencia se representa por una función “memoria” la cual denominan como un vector de tamaño fijo de estado oculto y que se actualiza cada vez que llega un nuevo vector proveniente de la imagen es decir la CNN.

## V. EXPERIMENTOS

### A. Detalles de implementación (librerías usadas, lenguajes de programación) y datasets seleccionados:

El modelo descrito anteriormente fue implementado como un modelo de TensorFlow (librería de Machine Intelligence para Python) llamado im2txt, disponible la cuenta oficial de TensorFlow en GitHub. Este repositorio brinda todo lo necesario para lograr obtener un generador de descripciones de imágenes, aunque no dispone de modelos pre-entrenados oficiales.

La documentación de im2txt sugiere usar el modelo pre-entrenado de Inception V3 para inicializar el modelo y luego reentrenar hasta 2.5 millones más de steps con el dataset MS COCO. Sin embargo, realizar este entrenamiento requiere

mucho tiempo y capacidad computacional por lo tanto se buscaron modelos pre-entrenados no oficiales que se publicaron en una de las cadenas de Issues en el repositorio de GitHub antes mencionado [9]. Dos usuarios publicaron los checkpoints del modelo entrenado con el dataset MS COCO, uno de ellos con 1 millón de steps más que el Inception V3, es decir que en total presenta 2 millones y por lo tanto se referirá a él como ckp-2 [10]. El otro lo entrenaron con 2 millones de steps más y por lo tanto se denominará ckp-3 [11].

Una vez con estos modelos, se procedió a realizar un notebook en Jupyter donde se evalúa cada modelo como lo sugiere la documentación de im2txt. Es importante resaltar que debido a la actualización de TensorFlow los modelos pre-entrenados encontrados no eran compatibles con la última versión y por lo tanto fue necesario actualizarlos como se sugirió en el mismo Issue de GitHub.

El modelo im2txt genera tres subtítulos por cada imagen entregada, para visualización se escoge la leyenda con mayor probabilidad y se inserta una caja con el texto generado en la parte superior de la imagen. Para lograr esto se usó la librería Pillow y Textwrap para garantizar la visualización correcta de los subtítulos en caso que fuesen muy largos para ocupar una sola línea. Un ejemplo de la salida del código se observa en la Figura 5.



Fig. 7: Imagen de salida.

### B. Métricas de evaluación

Como métrica se tomó por ejemplo lo encontrado en la página de MSCOCO, y se hizo un estudio humano para evaluar qué tan satisfactorias eran las descripciones resultantes.

Se realizó una evaluación que buscaba la validez promedio de 1 a 5 de las leyendas de las imágenes. A partir de esto se

podieron comparar los dos modelos y determinar cuál era “Mejor”.

La encuesta constaba de 5 imágenes por cada dataset, los cuales eran, PASCAL (Pattern Analysis Statistical Modeling and Computational learning) [12], FLICKR 8K [13] y MSCOCO [14].

Esta fue realizada por 15 individuos, y como se tenían los modelos checkpoint 2 y 3 se realizaron 2 encuestas M3.

Por otro lado, se quiso evaluar qué modelo era más rápido, por lo que se miró, cuanto demoraba cada modelo en cargar y, cuanto se demoraba cada uno en generar la descripción de cada imagen en promedio.

### C. Resultados

A continuación, se pueden ver los resultados de las encuestas. Lo calculado para evaluar el desempeño fue el promedio. Este se halló según el dataset empleado y según el modelo. Al final se incluye asimismo una comparación del rendimiento de ambos modelos en general.





Pascal					
ckp-2 3.66	a large jetliner sitting on top of an airport runway	a group of people standing around a fire hydrant	a bike parked on a beach next to the ocean	a group of people sitting around a table eating food	a colorful bird sitting on a tree branch
	4.66	2.06	3.93	3.66	4
ckp-3 3.48	a large jetliner sitting on top of an airport runway	a group of people standing around a fire hydrant	a bicycle parked next to a body of water	a man and a woman sitting at a table eating food	two parrots sitting on a branch in a tree
	4.46	2.26	3.73	2.73	4.2

Tabla 1: Datos de la encuesta con imágenes de dataset Pascal.






MS COCO					
ckp-2 3.88	a couple of people standing next to a stop sign	a man sitting on a bench in front of a building	a table topped with plates of food and drinks	a little boy standing on a sidewalk	a large brown bear walking across a lush green field
	4.6	4.26	4.53	2.2	3.8
ckp-3 4.12	a woman standing next to a stop sign	a man sitting on a bench in front of a building	a table topped with plates of food and drinks	a little girl standing on a sidewalk holding an umbrella	a large brown bear walking across a lush green field
	4.26	4.06	4.6	4.06	3.6

Tabla 2: Datos de la encuesta con imágenes de dataset MSCOCO.






Flickr 8K					
ckp-2 2.42	a man sitting on a bench in front of a building	a black and white dog with a frisbee in its mouth	a man in a suit and tie standing next to a woman	a little boy holding a frisbee in his hand	a group of people sitting on top of a sandy beach
	1.4	3.2	2.73	2.86	1.93
ckp-3 2.34	a man riding skis down a snow covered slope	a dog with a frisbee in its mouth	a man and a woman standing next to each other	a little boy holding a red frisbee in a yard	a group of people sitting on top of a sandy beach
	2.73	2.46	2	2.73	1.8

Tabla 3: Datos de la encuesta con imágenes de dataset Flickr 8K.

	Valoración
ckp-2	2.94
ckp-3	3.19

Tabla 4: Datos de la encuesta comparando los dos modelos.

Se puede ver que como era de esperarse que el modelo entrenado con un millón más de imágenes es algo mejor que el otro, sin embargo, la diferencia no es lo suficientemente grande como para justificar el tiempo que consume en entrenamientos, etc.

El dataset con el que se probó que tuvo más éxito fue el MS COCO, lo que también es de esperarse ya que es con este que se entrenó, por lo que es posible que estemos probando con imágenes con las que fue entrenado.

	t carga modelo	t descripción
Ckt 2	9,5 seg	2,9 seg
Ckt 3	10,5 seg	2,8 seg

Tabla 5: Datos de tiempo para cada modelo.

Así entonces vemos que a pesar de que al momento de carga del modelo el checkpoint 2 es más rápido que el 3, el tiempo que consumen ambos modelos es realmente similar, por lo que ninguno de los dos representa una ventaja notable sobre el otro en este aspecto.

## VI. CONCLUSIONES Y TRABAJO FUTURO

### A. Una implementación rápida del generador de descripciones para video.

Hacer los modelos más rápidos a partir de la modificación de la red y su re entrenamiento para así lograr que sea lo suficientemente rápido para usarse con videos.

B. Subtitulado con memoria y hasta generación de historias: Por ejemplo, usando el dataset: Visual Storytelling [15]. El cual contiene un set de imágenes relacionadas con una historia generada y evaluada por humanos.

#### Story # Descriptive Text

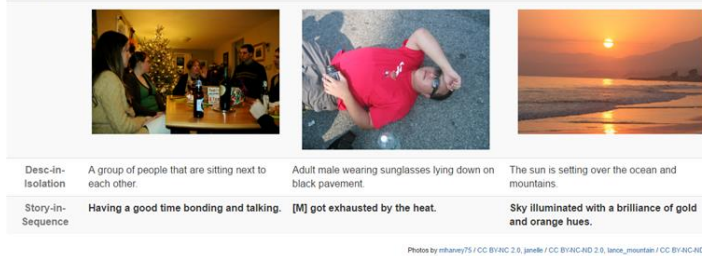


Fig. 8: Ejemplo de generador de historias de Microsoft.

#### VII. REFERENCES

- [1] D. Harris. (2014, November 18). *Google, Stanford build hybrid neural networks that can explain photos*. [Online]. Available: <https://gigaom.com/2014/11/18/google-stanford-build-hybrid-neural-networks-that-can-explain-photos/>
- [2] J. Mannes. (2016, September 22). *Google open sources image captioning model in TensorFlow*. [Online]. Available: <https://techcrunch.com/2016/09/22/google-open-sources-image-captioning-model-in-tensorflow/>
- [3] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, (2015). *Show and Tell: A Neural Image Caption Generator*. Google. [Online]. Available: <https://arxiv.org/pdf/1411.4555.pdf>
- [4] A. Karpathy, L. Fei-Fei. (2016). *Deep Visual-Semantic Alignments for Generating Image Descriptions*. Stanford University. [Online]. Available: <http://cs.stanford.edu/people/karpathy/deepimagesent/devisagen.pdf>
- [5] J. Zaino. (2014, November 18). *Google Researchers Use End-to-End Neural Network To Caption Pictures*. [Online]. Available: <http://www.dataversity.net/google-researchers-use-end-end-neural-network-caption-pictures/>
- [6] J. Mao, W. Xu, Y. Yang, J. Wang, A. Yuille. (2014). *Explain Images with Multimodal Recurrent Neural Networks*. Baidu research, University of California. [Online]. Available: <https://arxiv.org/pdf/1410.1090.pdf>
- [7] X. He, J. Gao, L. Deng. (2015). *From Captions to Visual Concepts and Back*. Microsoft Research. [Online]. Available: <https://arxiv.org/pdf/1411.4952.pdf>
- [8] Microsoft COCO 2015 Captioning challenge results. [Online]. Available: <http://mscoco.org/dataset/#captions-leaderboard>
- [9] Discusión modelos pre-entrenados. [Online]. Available: <https://github.com/tensorflow/models/issues/466>
- [10] Modelo checkpoint de 2'000.000. [Online]. Available: [https://drive.google.com/file/d/0Bw6m\\_66JSYLIRFVKQ2tGc\\_UJaWjA/view?usp=sharing](https://drive.google.com/file/d/0Bw6m_66JSYLIRFVKQ2tGc_UJaWjA/view?usp=sharing)
- [11] Modelo checkpoint de 3'000.000. [Online]. Available: [https://drive.google.com/file/d/0B\\_qCJ40uBfjEWVItOTdyNU\\_FOMzg/view](https://drive.google.com/file/d/0B_qCJ40uBfjEWVItOTdyNU_FOMzg/view)
- [12] Pascal dataset. <http://vision.cs.uiuc.edu/pascal-sentences/>
- [13] Flickr 8K dataset. [Online]. Available: <http://nlp.cs.illinois.edu/HockenmaierGroup/8k-pictures.html>
- [14] MSCOCO dataset. [Online]. Available: <http://mscoco.org/>
- [15] [Online]. Available: <http://visionandlanguage.net/VIST/>