ISSN 1870-4069

Evaluación del modelo neuronal de atención visual en la descripción automática de imágenes en español

Rafael Gallardo-García, Beatriz Beltrán-Martínez, Darnes Vilariño Ayala

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

rafael.gallardo@alumno.buap.mx,
{bbeltran,darnes,}@cs.buap.mx

Resumen. Este artículo presenta la evaluación del desempeño del modelo de atención visual neuronal presentado por Kelvin Xu, et al. en 2016. Se entrena y prueba el modelo con una nueva versión del conjunto de datos Flickr8k con una traducción al español. Así, se realiza el primer análisis cuantitativo del desempeño de este tipo de modelos en el idioma español. El artículo presenta la puntuación de similitud de las descripciones obtenidas, dicha puntuación se calculó con el algoritmo Bilingual Evaluation Understudy (BLEU) e incluye comparaciones entre las descripciones reales y las generadas de modo que se pueda realizar una evaluación directa. Se provee el conjunto de datos traducido y el código a través de GitHub.

Palabras clave: Análisis cuantitativo, BLEU, codificador-decodificador, lenguaje natural, subtitulado.

Evaluation of the Neural Model of Visual Attention in the Automatic Description of Images in Spanish

Abstract. This paper presents a performance analysis of the neural model of visual attention presented by Kelvin Xu, et al. in 2016. The model was trained and tested with a new Spanish translated version of the Flickr8k dataset. This is the first quantitative analysis of the performance of the visual attention models when captioning images in Spanish. This paper presents an evaluation of the quality of the predicted descriptions, the score was calculated with the Bilingual Evaluation Understudy (BLEU) and includes comparisons between the ground truth captions and the generated ones in order to allow a direct analysis of the results. The Spanish translated dataset and the code for the experiments is provided through GitHub.

Keywords: Quantitative analysis, BLEU, encoder-decoder, natural language, captioning.

1. Introducción

Una de las principales motivaciones que han llevado al humano a realizar investigación en el área de la inteligencia artificial es poder replicar las capacidades humanas en las computadoras. El área de la visión artificial busca replicar aquellas capacidades humanas de comprender el entorno a través de estímulos visuales. Los resultados del trabajo es esta área han sido tan variados como impresionantes: clasificación de imágenes con desempeños impresionantes [17,24], generación de contenido realista [14], traducción de estilos con redes condicionadas [21], súper fluidez y resolución en videos [3,5], entre muchos otros avances. Sin embargo, aunque las computadoras han llegado a dominar una gran variedad de áreas en el contenido visual hacía falta replicar una de las características más remarcables del humano: comprender escenas visuales e interpretar información detallada de dicha escena con exactitud.

El problema de la descripción automática de imágenes, bien conocido en el idioma inglés como *image captioning*, es un problema interesante que debe abordarse desde la interdisciplinaridad de las técnicas de aprendizaje automático: las mejores soluciones se obtienen utilizando híbridos entre las técnicas de visión artificial y procesamiento del lenguaje natural [6,18]. Para que un sistema de inteligencia artificial replique la capacidad humana de interpretar y describir una escena es necesaria más de una técnica de aprendizaje automático: reconocimiento de objetos, clasificación de imágenes, comprensión sintáctica y semántica de las escenas, además de la capacidad de poder expresar todo lo anterior en forma de lenguaje natural a través de técnicas como los modelos del lenguaje.

a little girl in a pink dress going into a wooden cabin .
a little girl climbing the stairs to her playhouse .
a little girl climbing into a wooden playhouse .
a girl going into a wooden building .
a child in a pink dress is climbing up a set of stairs in an entry way .

Fig. 1. Descripción de una escena con lenguaje natural en inglés, tomado de Flickr8k [15].

Uno de las características más interesantes de los sistemas de visión humana es la presencia de la *atención* [7,23], en vez de comprimir toda la información contenida en una escena en una representación estática, la atención permite analizar características sobresalientes de dicha escena de forma dinámica, similar al funcionamiento de la atención humana en ciertas áreas de una escena. Este artículo reporta las evaluaciones sobre el modelo de descripción automática de imágenes basado en atención presentado por Kelvin Xu, et al. [26] pues aunque el modelo basado en atención entrenado en el idioma inglés obtiene resultados fascinantes, es de interés analizar de forma cuantizada su desempeño en el lenguaje español.

2. Estado del arte

Las primeras propuestas para resolver el problema consistían en generar descripciones utilizando características manuales que operan sobre lenguajes formales [27], algunas otras propuestas consistían en utilizar detección de objetos y plantillas de descripciones [11], los enfoques más actuales se han respaldado en arquitecturas como las redes neuronales convolucionales (CNN) con esquemas de codificador-decodificador para codificar la imagen y de redes neuronales recurrentes (RNN) para generar texto a partir de la codificación obtenida con la CNN.

En los últimos 10 años se ha publicado una gran variedad de propuestas para solucionar el problema de descripción automática de imágenes, dichas propuestas afrontan el problema desde diversas áreas del aprendizaje automático: enfoques convolucionales [1], representación visual recurrente [4], con vecinos más cercanos [8], con redes recurrentes convolucionales a largo plazo [9], aprendizaje de instancias múltiples [10], con redes *long-short term memory* (LSTM) bidireccionales [25], aprendizaje reforzado [12,22] o con atención [26,28]. Gómez-Garay et al, reportan haber encontrado sistemas de descripción automática de imágenes en idioma inglés, japonés, chino, alemán y francés [13].

Gomez-Garay et al. presentaron un sistema [13] que presenta descripciones verbales en español sobre el entorno, dicho sistema recibía estímulos del entorno en formato de imágenes. Su sistema tardaba en promedio 0.96 segundos para generar la descripción para una imagen. El sistema generó un total de 8,693 frases y 47,061 palabras (838 únicas) para un conjunto de datos de 103 imágenes. Los autores no proveyeron un puntaje o métrica específica que evalúe las descripciones generadas.

En 2019, Martinez-Gutierrez evaluó el potencial de la API de visión computacional de Microsoft Azure en los idiomas español e inglés [20], concluyendo con un puntaje BLEU de 21.09 para el idioma español, siendo en su experimento, un puntaje mayor al obtenido por la API para el idioma inglés, sin embargo, la autora concluye que es un puntaje BLEU bajo, lo que significa que la correlación entre las descripciones generadas por la máquina y los humanos es débil.

3. Descripción del modelo de generación automática de imágenes

El modelo de descripción automática de imágenes que se evalúa en este artículo recibe el nombre de *Image Caption Generation with Attention Mechanism*, publicado con el nombre *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention* [26]. Los autores introdujeron un modelo basado en atención que aprende a describir el contenido de las imágenes automáticamente, en su artículo, ellos describen como entrenar el modelo de forma determinista utilizando las técnicas estándar de retropropagación y maximizando el límite inferior variacional con técnicas estocásticas.

En el artículo original, los autores describen distintos enfoques para generar las descripciones que intentan incorporar dos variantes de los mecanismos de atención: un mecanismo de atención "duro" y uno "suave". Además, defienden la ventaja de que incluir un mecanismo de atención permite visualizar lo que el modelo "ve" en un momento del análisis. El modelo que presentan puedan atender una parte destacada de la imagen mientras generan la descripción.

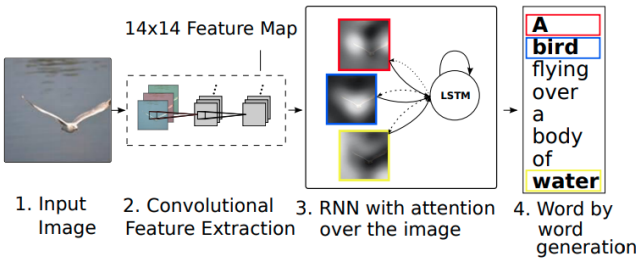

Fig. 2. Modelo de descripción automática con atención [26].

La figura 2 describe el funcionamiento del modelo propuesto en el artículo mencionado. El primer paso consiste en ingresar la imagen de entrada, posteriormente, un decodificador convolucional extrae las características de dicha imagen, el paso 3 consiste en generar las zonas de atención sobre la imagen con un decodificador basado en redes neuronales recurrentes, por último, el sistema genera una palabra basada en el vector de contexto entregado por el mecanismo de atención.

3.1. Codificador

El modelo en cuestión toma una imagen y genera una descripción y codificada como una secuencia de 1 de K palabras codificadas:

$$y = y_1, ..., y_C, y_i \in \mathbb{R}^K, \tag{1}$$

dónde K es el tamaño del vocabulario y C es la longitud de la descripción.

El codificador utiliza una red neuronal convolucional para extraer un conjunto de vectores característicos. El extractor produce L vectores, cada es una representación D-dimensional que corresponde a una parte de la imagen:

$$a = a_1, ..., a_n, a_i \in \mathbb{R}^D. \tag{2}$$

Para obtener la correspondencia entre los vectores característicos y las porciones de la imagen $2-D$, se extraen las características de una capa convolucional menor (a diferencia de trabajos anteriores que utilizan una capa completamente conectada). Esto último permite al decodificador enfocarse de forma selectiva en algunas partes de una imagen seleccionando un subconjunto de todos los vectores característicos.

3.2. Decodificador

El decodificador utiliza una red LSTM que produce una descripción, para generar la descripción, la red LSTM genera una palabra en cada paso de tiempo condicionado por el vector de contexto, el estado oculto y las palabras generadas anteriormente.

En términos generales, cada celda LSTM de la red implica proyecciones con un vector de pesos previamente aprendido. Cada celda aprende como pesar sus compuertas de entrada, mientras aprende como modular esa contribución a la memoria. Además, aprende los pesos que debe borrar de cada celda de memoria y los pesos que controlan como debería ser emitida esa memoria.

Para entender a profundidad la implementación de la red LSTM para el decodificador, se recomienda leer el artículo de Kelvin Xu et al. en la sección 3.1.2.

3.3. Mecanismos de atención

Se introducen dos técnicas de descripción automática de imágenes:

1. Un mecanismo de atención determinista "suave", entrenable mediante los métodos estándar de retropropagación.
2. Un mecanismo de atención estocástico "duro", entrenable mediante la maximización del límite inferior variacional o su equivalente mediante refuerzo.

Los detalles matemáticos y la demostración de cada uno de los mecanismos de atención se pueden encontrar en el artículo de Kelvin Xu et al [26].

4. Metodología de evaluación

Para poder evaluar el rendimiento en el idioma español del modelo descrito en la sección anterior de este artículo, es necesario entrenarlo con un conjunto de datos en pares que contenga imágenes y sus posibles descripciones en el lenguaje deseado, una vez que se cuente con ese conjunto de datos, se puede proceder con un análisis del vocabulario y la estructura del lenguaje contenido en las descripciones de las imágenes, posteriormente, será posible realizar el entrenamiento del modelo y las pruebas para medir cuantitativamente como se desempeña dicho modelo en el idioma español.

4.1. Traducción del conjunto de datos Flickr8k

Como se mencionó anteriormente, para poder evaluar el desempeño del modelo en el idioma español primero es necesario contar con un conjunto de datos que nos permita entrenar el modelo para generar descripciones en español.

No existen conjuntos de datos para descripción automática de imágenes en español que fueran de utilidad para los fines de este artículo, por lo que los autores generaron un conjunto de datos en pares para *image captioning* en

español. El conjunto de datos base recibe el nombre de Flickr8k y es resultado del trabajo de investigación [15] realizado por Hodosh et al. en 2013.

El conjunto Flickr8k contiene un total de 8,092 imágenes con 5 posibles descripciones para cada imagen. Las imágenes pueden contener animales, personas y objetos. Las posibles descripciones de cada imagen se enfocan en distintos aspectos de la escena y cuentan con distintas construcciones lingüísticas.

Proceso de traducción El conjunto de datos tiene un total de 40,460 descripciones de cada imagen, en idioma inglés. Para generar el conjunto de datos con las descripciones en español se utilizó la API de Google Cloud para traducción de texto en su versión básica. Se tradujo automáticamente el total de 40,460 descripciones y se emparejó cada quinteto de descripciones en español con su correspondiente imagen.

Por ahora, el modelo que se entrene a partir de estas descripciones podría heredar los sesgos del modelo de traducción automática de texto que contenga la API de Google. En trabajos posteriores se buscará verificar cada una de las traducciones manualmente para aumentar la fiabilidad del modelo de descripción de imágenes en español. Así mismo, se traducirán conjuntos de datos con una mayor cantidad de ejemplos de entrenamiento, como podría ser el conjunto de datos COCO [19] u 80K Flickr [16].

una niña en un vestido rosa de entrar en una cabina de madera.
una niña que sube las escaleras hasta su casa de juegos.
una pequeña muchacha que sube en una casa de juegos de madera.
una niña de entrar en un edificio de madera.
un niño en un vestido rosa está subiendo por una escalera en una puerta de entrada.

Fig. 3. Descripción de una escena con lenguaje natural en español, tomado de Flickr8k traducido.

La figura 3 es un ejemplo de entrenamiento de la versión de Flickr8k traducida al español automáticamente. Como se puede observar, el contenido de la imagen sigue siendo expresado en términos generales pero se pierden algunas de las características gramaticales y morfológicas del español.

4.2. Análisis de Flickr8k en español

El conjunto de datos de Flickr8k traducido al español varía en cantidad de vocabulario a su versión en inglés. Las descripciones de las imágenes en inglés acumulan un vocabulario total de 8,918 palabras, siendo el artículo *a* el más común con un total de 62,989 apariciones, por otra parte, la versión en español tiene un vocabulario de 12,439 palabras, siendo el artículo *un* el más común con un total de 39,366 apariciones.

Como se puede observar, la traducción del conjunto de datos añadió 3,521 palabras al vocabulario. Solo se removerán signos de puntuación durante la limpieza. Remover palabras cerradas resultaría en descripciones pobres en sintaxis y morfología.

La figura 4 muestra las 50 palabras más frecuentes en todo el conjunto, las más comunes llegan casi a una frecuencia de 40,000 mientras que las menos frecuentes apenas aparecen una vez en todas las descripciones.

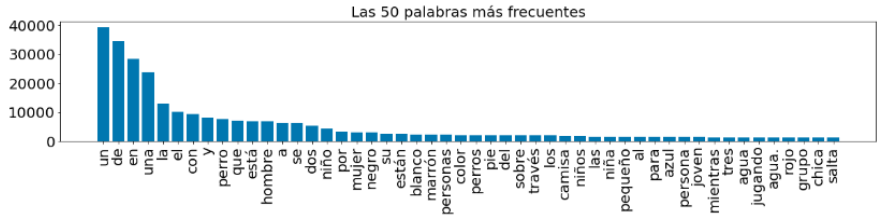

Fig. 4. 50 palabras más comunes en Flickr8k en español.

4.3. Implementación del modelo

El modelo de descripción de imágenes basado en atención se implementó en el lenguaje Python3 con soporte de la plataforma para aprendizaje automático TensorFlow [1]. El código utilizado para la implementación está basado en cuaderno de Jupyter *Image captioning with visual attention* [2], la versión modificada de dicha libreta para el conjunto de datos en español y la versión de Flickr8k en español se pueden encontrar en el siguiente repositorio de GitHub: *https://github.com/gallardorafael/ShowAttendTell_Flickr8k_Spanish*.

Estructura del conjunto de datos El conjunto de datos utilizado cuenta con 8,092 imágenes, como se mencionó anteriormente, cada imagen cuenta con 5 descripciones con el siguiente formato:

nombre_imagen,id_descripción,descripción_textual

En dónde *nombre_imagen* contiene el nombre del archivo .jpg de la imagen a la que esa descripción corresponde, *id_descripción* es el identificador para la descripción en cuestión (5 por cada imagen) y *descripción_textual* contiene el texto crudo que describe a la imagen y sobre el cuál se entrenará el modelo.

El archivo con el conjunto de datos es un CSV que contiene 8,092 líneas, cada línea es un ejemplo de entrenamiento como el descrito anteriormente, cada

[1] https://www.tensorflow.org/
[2] https://www.tensorflow.org/tutorials/text/image_captioning

línea está dividida por una ",", de modo que pueda ser importada con facilidad sin necesidad de realizar un procesamiento adicional.

Implementación del modelo Para implementar el modelo se utilizó un cuaderno de *Jupyter* de modo que la ejecución se pueda realizar con facilidad en una sesión de *Google Colaboratory* [3]. La información detallada sobre el código se encuentra en el repositorio mencionado con anterioridad, pero por motivos de claridad, a continuación se detallan los puntos importantes de la implementación:

- Se utilizó una capa convolucional menor de InceptionV3 (a su vez preentrenada sobre Imagenet [17]), con dimensiones (8,8,2048), mismas que se redimensionaron a (64,2018).
- El codificador implementado es una CNN que consiste en una capa completamente conectada.
- El decodificador implementado es una RNN, específicamente una *Gated Recurrent Unit* (GRU), que será la encargada de predecir la siguiente palabra.
- Se implementó un mecanismo de atención Bahdanau [2].
- Se utilizó el método *teacher forcing* para acelerar de forma eficiente el entrenamiento del modelo.

5. Resultados experimentales

El modelo se entrenó sobre un total de 8,000 imágenes y sus 40,000 descripciones correspondientes, de las cuales el 20 % se utilizó para validación y el otro 80 % para entrenamiento del modelo. Las 91 imágenes y sus 455 descripciones fueron utilizadas para la evaluación del modelo, esto es, calcular los puntajes de secuencia-secuencia y los análisis cualitativos de los resultados.

5.1. Proceso de entrenamiento

La primera parte del entrenamiento consistió en extraer las características de las imágenes utilizando la última capa convolucional (codificador) de InceptionV3, para extraer dichas características se utilizó una GPU NVIDIA Tesla P100 con 16 gigabytes de VRAM. Para extraer las características de las 8000 imágenes fueron necesarias 450 iteraciones y 35 minutos de procesamiento, con un promedio de 4.5 segundos por iteración.

La segunda parte del entrenamiento consistió en entrenar los mecanismos de atención con las características extraídas de las imágenes y de los textos de modo que el sistema aprendiera a describir cada imagen. Sobre la misma plataforma (NVIDIA Tesla P100), el entrenamiento con 20 épocas tomó un tiempo total de 56 minutos. La figura 5 muestra el gráfico de la función de pérdida del entrenamiento a través de las épocas. Es fácil observar que aumentar la cantidad de épocas más allá de 20 en un conjunto de datos reducido como este nos llevaría a un sobre ajuste del modelo.

[3] https://colab.research.google.com/

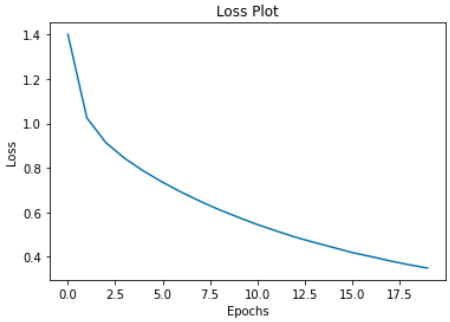

Fig. 5. Gráfico del entrenamiento a través de las épocas.

Una vez concluido el proceso de entrenamiento se puede proceder con la fase de pruebas para evaluar el desempeño del modelo entrenado.

5.2. Análisis cualitativo

El objetivo de este artículo es realizar una evaluación cuantitativa y cualitativa del desempeño del modelo en español. Se presentarán ejemplos de los peores resultados (con puntaje BLEU < 0.3) y de los mejores resultados (con puntaje BLEU > 0.7), así como un ejemplo del desarrollo del mecanismo de atención sobre una imagen tomada del set de prueba.

La figura 6 es una demostración del mecanismo de atención empleado. En la parte superior se encuentra la imagen mostrada al modelo, la parte inferior es la visualización del mecanismo de atención enfocando diversas partes de la imagen para generar una palabra a partir de sus *observaciones*. La descripción real para esta imagen es la mostrada en el primer renglón: *un perro blanco corriendo detrás de una pelota amarilla*, mientras que la generada por el modelo es: *un perro blanco está jugando con una pelota amarilla en su boca*. No es necesaria una métrica para concluir que esta descripción es acorde al contenido de la imagen, describe tanto los objetos como las acciones presentes en la escena.

La figura 7 muestra algunos ejemplos de malas descripciones generadas por el modelo, todas con puntaje BLEU menor a 0.3.

La figura 8 muestra ejemplos de descripciones con puntajes BLEU mayores a 0.7. Al igual que en el ejemplo de la figura 6, no es necesaria la métrica para comprobar que la descripciones coinciden de forma adecuada (aunque no completa) con el contenido de las escenas.

5.3. Análisis cuantitativo

Para realizar el análisis cuantitativo de los resultados se utilizará la medida BLEU. BLEU es un método de evaluación de la calidad de las traducciones realizadas por sistemas de traducción automática, en este sistema, una traducción tiene mayor calidad cuanto más similar es con respecto de otra de referencia que

Fig. 6. Visualización del mecanismo de atención del modelo.

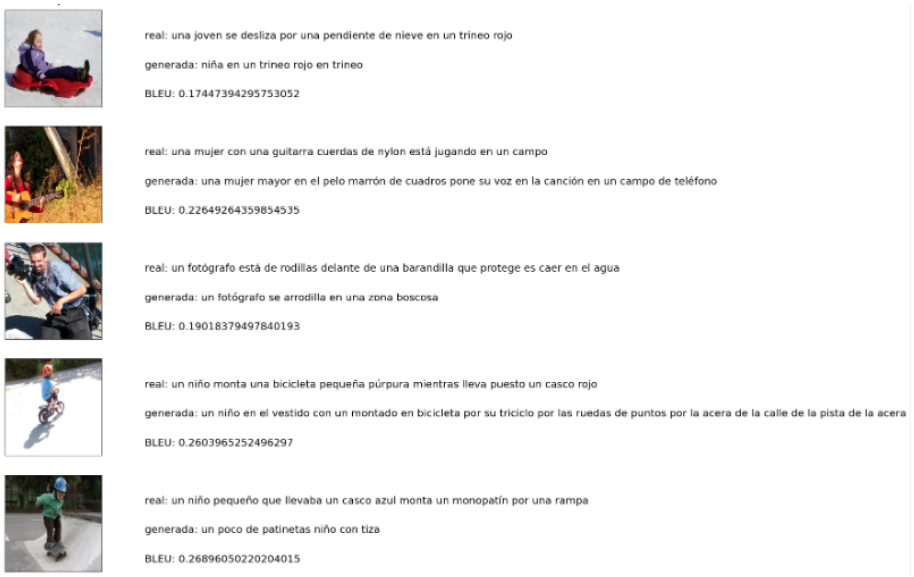

Fig. 7. Descripciones generadas con un puntaje BLEU < 0.3.

Fig. 8. Descripciones generadas con un puntaje BLEU > 0.7.

se supone correcta, BLEU permite evaluar con más de una referencia. Aunque este método surgió para evaluar traducciones, es ampliamente utilizado por las investigaciones del estado del arte mencionados en este artículo para evaluar la similitud de dos textos del mismo idioma.

Para este análisis se describieron 300 imágenes del conjunto de prueba y se calculó el puntaje BLEU para cada una de ellas: las hipótesis son las descripciones generadas y las referencias las descripciones reales obtenidas del conjunto de datos. La Tabla 1 muestra los resultados obtenidos por este método para el español.

Tabla 1. Evaluación de los puntajes BLEU obtenidos por el método.

Media	Mediana	Más alto	Más bajo	Desviación estándar
0.356	0.346	0.816	0	0.192

Una media de 0.356 no es buena, sin embargo es mayor que el reportado por Martinez-Gutierrez para la API de Microsoft Azure (obtuvo un 0.21). El puntaje

BLEU más alto obtenido por el modelo es de 0.816 y el más bajo de 0, dada la dispersión de los puntajes se decidió calcular la varianza, misma que tiene un valor de 0.192, lo que significa que los puntajes no están demasiado alejados de su media.

6. Conclusión

Aunque la media de los puntajes obtenidos es baja, existen descripciones con puntajes BLEU muy altos, lo que indica que una mayor cantidad de datos de entrenamiento podría mejorar satisfactoriamente los puntajes. Por otro lado, las descripciones mejor puntuadas describen con una exactitud alta el contenido de las imágenes tanto en objetos como en las acciones que se están llevando a cabo en la escena, por lo tanto, esas descripciones con puntajes BLEU mayores a 0.7 satisfacen los criterios cualitativos y cuantitativos de la descripción automática de imágenes.

El modelo de atención implementado en este trabajo y entrenado sobre el idioma español obtuvo resultados positivos muy satisfactorios y aunque la media de los puntajes es baja, no es menor que el límite impuesto (0.3) para ser considerada una media mala. Como se mencionó en el párrafo anterior, los resultados de este modelo podrían mejorar añadiendo más ejemplos de entrenamiento y mejorando la calidad de dichos ejemplos. Posiblemente, con una traducción al español realizada por expertos en vez de la automática se puedan obtener descripciones que respeten mejor la morfología y sintaxis del idioma, de forma que suenen coherentes y *más humanas*. Una propuesta para mejorar los resultados del modelo de atención para la descripción automática de imágenes es la utilización de un conjunto de datos con más ejemplos de entrenamiento, como 80K Flickr o MS COCO, aunque esto implique incrementar drásticamente los tiempos de entrenamiento del modelo. De igual forma, la calidad de las descripciones puede mejorar si se cambia la traducción automática de los conjuntos de datos por una traducción manual o semi-supervisada de las descripciones.

Desafortunadamente, el modelo no puede ser comparado de forma cuantitativa con otros modelos de descripción automática de imágenes pues las condiciones de entrenamiento y experimentación no hacen viable la comparación para este idioma, por esta razón, se ha decidido hacer público el conjunto de datos utilizado en este trabajo, de modo que futuros trabajos tengan la opción de evaluar modelos en idioma español que las condiciones de experimentación hagan viable la comparación.

Otra consideración que se deberá tomar en cuenta en trabajos similares o derivados de la descripción automática de imágenes en español, es el tipo de métricas utilizadas, en un futuro, la evaluación a este tipo de modelos debería enriquecerse con métricas que miden secuencia-secuencia como *Syntatic Features for Evaluation of Machine Translation* (SMT), *Metric for MT Evaluation with Improved Correlation with Human Judgments* (METEOR), con técnicas que consideren la complejidad morfológica como *Languaje-independent Model for Machine Translation Evaluation with Reinforced Factors* (hLEPOR), con

métricas que consideren paráfrasis, lemas y sinónimos como *Translation Edit Rate-Plus* (TER-Plus) o con métricas que no necesiten referencias como *Machine Translation Evaluation without Reference Texts* (MEWR).

Referencias

1. Aneja, J., Deshpande, A., Schwing, A.G.: Convolutional image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5561–5570 (2018)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Bao, W., Lai, W.S., Ma, C., Zhang, X., Gao, Z., Yang, M.H.: Depth-aware video frame interpolation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3703–3712 (2019)
4. Chen, X., Zitnick, C.L.: Learning a recurrent visual representation for image caption generation. arXiv preprint arXiv:1411.5654 (2014)
5. Chu, M., Xie, Y., Leal-Taixé, L., Thuerey, N.: Temporally coherent gans for video super-resolution (tecogan). arXiv preprint arXiv:1811.09393 1(2), 3 (2018)
6. Cui, Y., Yang, G., Veit, A., Huang, X., Belongie, S.: Learning to evaluate image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5804–5812 (2018)
7. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annual review of neuroscience 18(1), 193–222 (1995)
8. Devlin, J., Gupta, S., Girshick, R., Mitchell, M., Zitnick, C.L.: Exploring nearest neighbor approaches for image captioning. arXiv preprint arXiv:1505.04467 (2015)
9. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
10. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1473–1482 (2015)
11. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: European conference on computer vision. pp. 15–29. Springer (2010)
12. Feng, Y., Ma, L., Liu, W., Luo, J.: Unsupervised image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4125–4134 (2019)
13. Gomez-Garay, A., Raducanu, B., Salas, J.: Dense captioning of natural scenes in spanish. In: Mexican Conference on Pattern Recognition. pp. 145–154. Springer (2018)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Advances in Neural Information Processing Systems 3 (06 2014)
15. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research 47, 853–899 (2013)

16. Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., Winnemoeller, H.: Recognizing image style. arXiv preprint arXiv:1311.3715 (2013)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
18. Kumar, A., Goel, S.: A survey of evolution of image captioning techniques. International Journal of Hybrid Intelligent Systems 14(3), 123–139 (2017)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
20. Martinez Gutierrez, M.F.: Automated Image Captioning: Exploring the Potential of Microsoft Computer Vision for English and Spanish. Ph.D. thesis, University of Geneva (2019)
21. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
22. Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L.J.: Deep reinforcement learning-based image captioning with embedding reward. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 290–298 (2017)
23. Rensink, R.A.: Change blindness: Implications for the nature of visual attention. In: Vision and attention, pp. 169–188. Springer (2001)
24. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision 115(3), 211–252 (2015)
25. Wang, C., Yang, H., Bartz, C., Meinel, C.: Image captioning with deep bidirectional lstms. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 988–997 (2016)
26. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057 (2015)
27. Yao, B.Z., Yang, X., Lin, L., Lee, M.W., Zhu, S.C.: I2t: Image parsing to text description. Proceedings of the IEEE 98(8), 1485–1508 (2010)
28. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4651–4659 (2016)