Universidad Complutense de Madrid
Facultad de Informática

TRABAJO DE FIN DE GRADO

Daniel Gamo Alonso

Generación de descripciones de imágenes mediante *Deep Learning*

Doble Grado en Ingeniería Informática y Matemáticase

Directores: Alberto Díaz Esteban, Gonzalo Méndez

5 de julio de 2017

TODO - Agradecimientos

Resumen

El objetivo de este trabajo es construir un sistema basado en *deep learning* con el propósito de generar descripciones textuales a partir de las imágenes que se suministren como entrada. Para ello se utilizarán distintas técnicas relacionadas con la visión artificial y el procesamiento del lenguaje natural como redes neuronales y otras metodologías basadas en modelos estadísticos para, a partir de los elementos presentes en las imágenes, predecir y generar descripciones coherentes con su contenido. Entre las posibles aplicaciones podría estar el facilitar la accesibilidad a contenido en forma de imagen a usuarios con algún tipo de discapacidad visual.

Palabras clave: *deep learning*, red neuronal, descripción de imágenes.

TODO: ENGLISH

Lista de contenidos

Capítulo 1

Introducción

A menudo se dice que una imagen vale más que mil palabras, y este dicho no podría ser más acertado. Los humanos nos apoyamos en el sentido de la vista para gran parte de las tareas que realizamos en nuestra vida cotidiana. Esta importancia motiva el contenido de nuestro trabajo, en el que pretendemos construir un modelo usando técnicas de *deep learning* (con las que se han conseguido grandes avances en los últimos años, e incluso meses, dentro de este campo) para estudiar la tarea de analizar y extraer datos de las imágenes. Sin embargo, la importancia de la visión para los humanos no se basa exclusivamente en reconocer objetos, que es la primera aproximación que se hace en este sentido, sino que también contamos con esa poderosa herramienta que es el lenguaje, y que nos permite no solo reconocer, sino describir lo que vemos. Esta idea de conexión entre la vista y el lenguaje es la que constituye el grueso de este trabajo, y nuestro objetivo va a ser estudiar y construir un modelo que relacione el contenido de una imágen con una descripción textual de la misma.

En este proceso de análisis y descripción hay que tener diversos factores en cuenta. Para empezar, hay que definir que vamos a entender por descripciones. Esta tarea es difícil debido al amplio significado que tiene la tarea de describir; no está claro cómo de larga puede ser la descripción, si debe centrarse en todos los detalles o dar una "idea general" del contenido de la imágen, etc. Todas estas razones convierten la tarea de la descripción de imágenes en algo mucho más complejo, alejado del planteamiento algo más sencillo (que no trivial) de extraer elementos de las imágenes y organizarlos en una frase bien estructurada.

1.1 Objetivos

Nuestro objetivo principal en este trabajo es construir un sistema que sea capaz de relacionar el contenido de una imagen con una descripción textual que se acerque a la que daría una persona. Para ello vamos a estudiar, desde el enfoque del *deep learning*, las diferentes técnicas y modelos existentes para el análisis de imágenes y la descripción de su contenido. Destacamos dos conceptos claves para el desarrollo de nuestro trabajo, que son las redes neuronales recurrentes (*Recurrent Neural Network*, RNN) y las redes neuronales convolucionales (*Convolutional Neural Network*, CNN), que han probado su eficacia en las tareas relacionadas con procesamiento del lenguaje natural y análisis de imágenes, respectivamente. Hablaremos en profundidad sobre ello en los capítulos 2 y 3 de esta memoria.

Para esta tarea necesitamos analizar una gran cantidad de datos. Cuando se suministra a una máquina, una imagen queda representada como una matriz de píxeles y una sentencia (descripción) como una lista de palabras (*tokens*). Cada una de estas unidades no da información por si misma; necesita del resto para conformar una unidad con sentido. Además, para inferir las reglas que permitan detectar las relaciones entre los distintos elementos (entre elementos de la imagen, entre elementos de la sentencia y entre elementos de la imagen con su correspondiente sentencia) necesitamos un gran número de imágenes y de sentencias, con lo que la capacidad de computación necesaria para llevarlo a cabo es inmensa. En nuestro caso, contamos con una tarjeta gráfica donada por NVIDIA que será clave en la realización de todos los cálculos involucrados en un tiempo aceptable.

1.2 ¿Qué es una descripción?

Ya hemos comentado la importancia que tiene definir correctamente lo que nuestro modelo va a entender por una descripción. Tenemos un modelo generativo, que necesita descripciones en el entrenamiento con un formato más o menos similar para producir buenos resultados.

Según la RAE, describir se define como:

1. Representar o detallar el aspecto de alguien o algo por medio del lenguaje.
2. Moverse a lo largo de una línea.
3. Definir imperfectamente algo, no por sus cualidades esenciales, sino dando una idea general de sus partes o propiedades.
4. Delinear, dibujar, pintar algo, representándolo de modo que se dé perfecta idea de ello.

En el proceso de descripción de una escena, si nos atenemos a la tercera acepción, no se hace un análisis detallado de todos los objetos y acciones que se reflejan en ella, sino que se resume la información, y se tiende a describir los elementos que más llaman nuestra atención. Ya sea por su importancia o tamaño en la escena, por la impresión subjetiva que nos causan o por el contexto en el que sucede la escena y que los dota de mayor o menor relevancia. Pensemos por ejemplo en una imagen de una persona con el cielo de fondo; un humano destacaría a la persona que aparece en ella y daría menos importancia a otras cosas como el cielo que aparece detrás de la imagen (no es algo que llame la atención, siempre esta ahí), mientras que una máquina podría centrar su atención en ese cielo que aparece de fondo (por ejemplo, porque ocupa un porcentaje de la imagen más alto que la persona).

En los *datasets* que vamos a utilizar, cada imagen va acompañada de cinco frases con una longitud media de ///CALCULAR CON EL CODIGO. La anotación de imágenes se ha realizado utilizando operarios humanos a través de la plataforma Amazon Mechanical Turk. Se pidió a los trabajadores que describiesen el contenido de la imagen con una frase. Se ha probado empíricamente que

en esta colección de datos se suele describir los aspectos más relevantes de la imagen, con especial incapié en descripción de personas, sus acciones, interacciones con la gente o el entorno. [Karpathy (2016)].

1.3 Problemas

La tarea que nos proponemos presenta una seria de desafíos más allá de la construcción del modelo o de la capacidad de cálculo de la que disponemos.

Cuesta decidir qué es una buena descripción de una imagen y qué no lo es, pues dos personas distintas podrían dar dos descripciones distintas de la misma imagen, ya sea por la importancia que dan a ciertos elementos de la escena o por como describan el mismo elemento. Sin embargo, dos descripciones distintas pueden ser igualmente válidas y esto pone de relieve la importancia de tener buenas métricas para que el sistema sepa cuándo está describiendo algo bien, cuándo está describiendo algo mal, cuando lo hace mejor y cuando lo hace peor. Desarrollaremos en el capítulo 4 con más detalle qué métricas utilizamos, cuál es el razonamiento que hay detrás de ellas y cómo de fiables son.

Otro inconveniente que se nos presenta es la dificultad de obtener imágenes con buenas descripciones asociadas. Aunque hay sitios como Flickr que contienen muchas imágenes con descripciones, a menudo estas últimas no dan información fiable sobre el contenido de la imagen. Cuando subimos una fotografía, no solemos describir su contenido, sino que adjuntamos texto sobre la situación en la que se produce, las personas que nos acompañan o los sentimientos que nos evocan. Por esta razón, no es fácil obtener automáticamente un *dataset* lo bastante bueno como para entrenar al sistema, y precisamos de trabajo humano para acompañar las imágenes (o partes concretas de las imágenes) de descripciones adecuadas. En este sentido, muchos de los *datasets* que se utilizan en este campo han requerido del trabajo de muchas personas, principalmente utilizando la herramienta Amazon Mechanical Turk. Además, en relación con lo que hemos expuesto en el párrafo anterior, necesitamos más de una descripción por imagen para dar perspectiva a nuestro sistema sobre las diferentes formas de describir el mismo contenido. Como en el caso de las métricas, la información acerca de los *datasets* utilizados se desarrollará en el capítulo ??.

Aunque tengamos un conjunto de pares imagen-descripción lo suficientemente grande y bueno para nuestra tarea, todavía queda algo que dificulta nuestra tarea, y es el poder asociar elementos de la descripción con zonas concretas de la imagen y no con toda ella. En este sentido, se han publicado trabajos que cuentan con *datasets* más completos en este sentido, como *Visual Genome* [Krishna et al. (2016)], de manera que las relaciones entre lo descrito y su localización en la imagen son más fáciles de aprender por el sistema.

1.4 Estructura del documento

En el capítulo 1 hemos introducido el tema sobre el que trata esta memoria. El capítulo 2 consiste en un resumen sobre los trabajos más importantes publicados acerca de análisis de imágenes, de sentencias y de relación entre ambas. En el capítulo 3 describiremos nuestro modelo, y hablaremos sobre la teoría de

deep learning que hay detrás del mismo. Además describiremos la estructura y el funcionamiento de la redes neuronales que vamos a implementar. En el capítulo 4 explicamos la metodología que seguimos en nuestros experimentos y analizamos los datos y los resultados, comparándolos con los de trabajos existentes. Se muestran además ejemplos concretos de resultados que nuestro modelo a proporcionado. Por último, en el capítulo 5 exponemos las conclusiones del trabajo y planteamos lineas de trabajo futuras. ///////HASTA CAP 5?

Capítulo 2

Trabajo relacionado

Tanto la tarea de análisis de imágenes como la de generación de sentencias en lenguaje natural mediante *deep learning* está en desarrollo constante, y cada mes aparece un nuevo artículo o tesis sobre este tema.

La idea de utilizar redes neuronales recurrentes para construir modelos relacionados con el procesamiento del lenguaje natural está muy presente en nuestro trabajo. En Kombrink et al. (2011) se propone un modelo que utiliza RNNs con este propósito, y analiza su comportamiento en esta tarea. Existen numerosos estudios sobre la generación de frases, donde nos interesan especialmente aquellos que las construyen en base a unas etiquetas [Bahdanau et al. (2014)], más cercano a la idea de combinar análisis de imágenes y sus descripciones (los elementos que aparecen en ellas se usan para generar las frases).

Análogamente, se ha probado que las redes neuronales convolucionales dan buenos resultados cuando estamos tratando con datos en forma de imagen, como en Krizhevsky et al. (2012). Sin embargo, también se plantea el problema de conseguir buenos tiempos de entrenamiento en estas redes con tantos parámetros que ajustar.

Para trabajar con las técnicas de *deep learning*, han aparecido numerosas APIs que automatizan gran parte de los cálculos necesarios en la red. Entre ellas destacamos Theano (la que usaremos en este proyecto) y TensorFlow, que cuenta con el apoyo de Google.

En el trabajo de Barnard et al. (2003) se explora la conexión entre imágenes (o regiones de imágenes) y palabras mediante diferentes modelos, aún sin utilizar técnicas de *deep learning*, presentes en las publicaciones más actuales.

Los trabajos que combinan ambos enfoques para generar descripciones de imágenes son más recientes, Estos son los que nos interesan, y que vamos a tomar como base para construir nuestro propio modelo. El elemento común entre todos ellos es que se basan en redes neuronales convolucionales para el análisis de imágenes y redes neuronales recurrentes para la generación de frases [Karpathy and Fei-Fei (2015); Vinyals et al. (2015)]. En el trabajo de Karpathy and Fei-Fei (2015) se trabaja con una modificación de las RNNs tradicionales (introduciendo la bidireccionalidad en la red) y de las CNNs, para que ambas trabajen bien en conjunto.

También es importante destacar la importancia de contar con buenos *datsets* y metodologías definidas para la evaluación de parejas imagen-frase. En el trabajo de Hodosh et al. Hodosh et al. (2013) se recopilan anotaciones de 8000

imágenes (5 para cada una de las imágenes) que deben describir las entidades y los eventos. Como en el resto de *datasets* que vamos a estudiar en este trabajo, las anotaciones se han obtenido mediante trabajadores humanos con plataformas de *crowdsourcing*. Este *dataset* se denomina Flickr8k. Mencionamos el trabajo relacionado con otro de los *datsets* que vamos a utilizar: MSCOCO Lin et al. (2014).

Capítulo 3

Descripción de imágenes

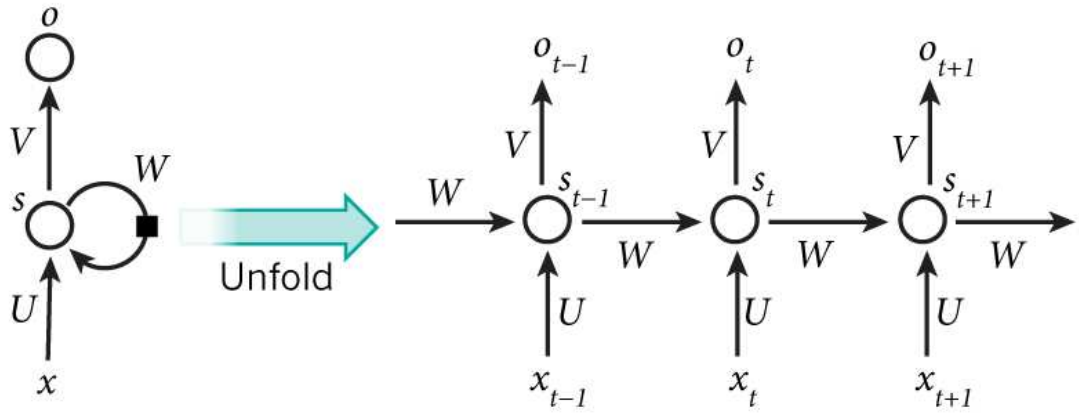
3.1 Recurrent Neural Networks - RNNs

A diferencia de las redes neuronales tradicionales, en las redes neuronales recurrentes se considera que los datos de entrada (y salida) no son independientes entre sí. Entre los principales usos de estas redes destacamos dos:

1. Clasificar sentencias de acuerdo a su probabilidad de aparecer en una situación real, dándonos una medida de su corrección sintáctica y/o gramática.

2. Generar texto nuevo (original) tras entrenar el sistema con frases de prueba.

Observamos la importancia de considerar dependencias entre las entradas y las salidas de la red: en el caso de las frases, si queremos generar una nueva palabra, tendremos que tener en cuenta la parte de la frase ya generada, pues esta influirá en el resto de la sentencia.



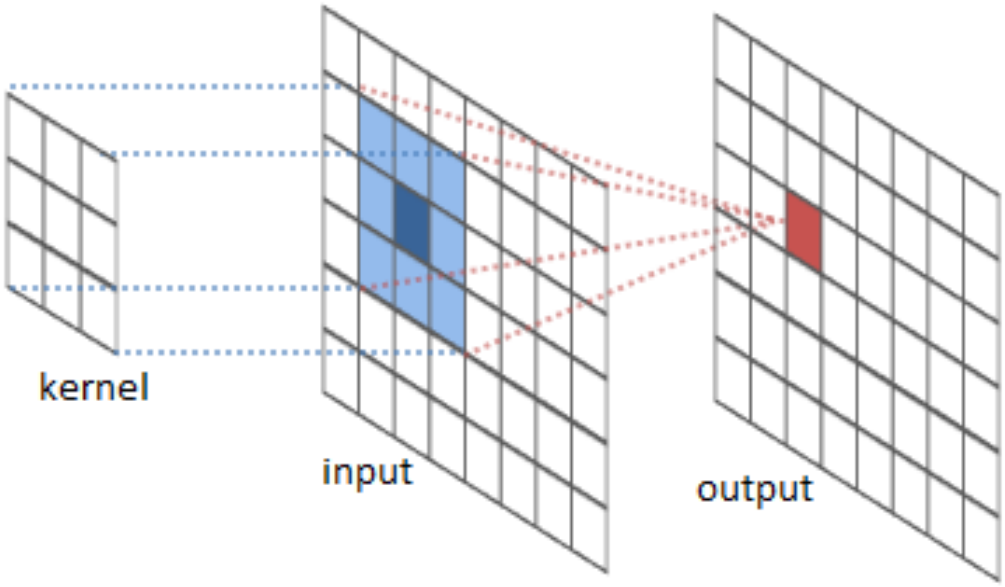
En este caso, x_t representa la entrada de la red, s_t el estado oculto y o_t la salida en el paso t. En la figura vemos que el estado s_t se calcula como función del estado anterior s_{t-1}, la entrada en el paso actual x_t. La red posee "memoria" en el sentido en que los estados anteriores condicionan el estado actual. Sin embargo, esta memoria no se mantiene durante muchas fases. Existe un tipo concreto de RNN, las conocidas como *long short-term memory* (LSTM) que favorece la persistencia de los datos de los estados anteriores durante un número de mayor de fases, lo que las hace especialmente indicadas para comprensión de lenguaje natural, análisis de textos manuscritos y reconocimiento de voz.

3.2 Convolutional Neural Networks - CNNs

Las redes neuronales convolutivas se utilizan en tareas como la clasificación y reconocmiento de imágenes.

Podemos ver que el modelo asigna la mayor probabilidad a "barco" de entre las cuatro categorías existentes. En el modelo de la figura observamos cuatro operaciones en la red:

- **Convolución.** El principal objetivo de la operación de convolución es extraer características de una imagen.La convolución preserva la relación espacial entre los pixels de la imagen usando pequeños cuadros como datos de entrada.



  Consideramos una imagen como una matriz bidimensional de píxeles (input), y otra matriz (*kernel* o filtro), normalmente de tamaño $3x3$ que "recorre" la imagen de entrada. Con los valores del kernel y la porción de imagen que cubre, se computa la convolución y esto da como resultado otra imagen (mapa de activación).

- **No linealidad.** Se aplica una función de activación no lineal operando sobre cada pixel del mapa de activación. Aunque pueden usarse funciones como la sigmoide, se ha probado que la función ReLU (Rectified Linear Unit) da mejores resultados en este tipo de redes neuronales [REF].

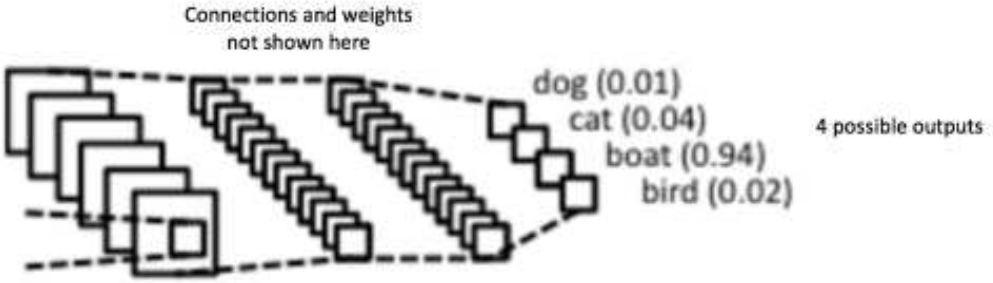
- **Pooling.** Se encarga de reducir el tamaño del mapa de activación conservando los elementos más importantes. El *Pooling* puede ser de distintos tipos: Max, Sum, Avg...

  En el caso del *Max Pooling*, se define un espacio (por ejemplo una matriz $2x2$) y para cada bloque $2x2$ se coge el mayor valor de entre los 4 existentes.

  La función del *Pooling* es reducir las imágenes y convertirlas en objetos más manejables por las siguientes capas de la red.

- **Fully Connected Layer.** Tras la convolución y el *Pooling*, obtenemos características de alto nivel de la imagen de entrada. En esta fase, y usando

dichas características como entrada, clasificamos la imagen en una serie de categorías basadas en el *dataset* de entrenamiento.



Bibliografía

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *Journal of machine learning research*, 3(Feb):1107–1135, 2003.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

Andrej Karpathy. *Connecting Images and Natural Language*. PhD thesis, Stanford University, 2016.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

Stefan Kombrink, Tomáš Mikolov, Martin Karafiát, and Lukáš Burget. Recurrent neural network based language modeling in meeting recognition. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. URL https://arxiv.org/abs/1602.07332.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.