Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

# Evaluación del Modelo Neuronal de Atención Visual en la Descripción Automática de Imágenes en Español

Rafael Gallardo-García  Beatriz Beltrán-Martínez  Darnes Vilariño

Language & Knowledge Engineering Lab, Benemérita Universidad Autónoma de Puebla

7 de agosto de 2020

# Índice


1. Introducción
   - ¿Qué es el *image captioning*?
   - ¿Qué es un mecanismo de *atención*?
   - Trabajo previo
   - Justificación del trabajo experimental
2. Descripción de la arquitectura
   - Detalles generales
   - Codificador
   - Decodificador
   - Mecanismos de Atención Visual
3. Métodos
   - Conjunto de datos Flickr8k
   - Implementación
   - Entrenamiento del modelo
   - Resultados
4. Conclusión
   - Conclusiones
   - Trabajo futuro



Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

BUAP


1 Introducción
- ¿Qué es el *image captioning*?
- ¿Qué es un mecanismo de *atención*?
- Trabajo previo
- Justificación del trabajo experimental

2 Descripción de la arquitectura
- Detalles generales
- Codificador
- Decodificador
- Mecanismos de Atención Visual

3 Métodos
- Conjunto de datos Flickr8k
- Implementación
- Entrenamiento del modelo
- Resultados

4 Conclusión
- Conclusiones
- Trabajo futuro



Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

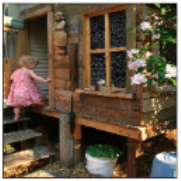
# ¿Qué es el image captioning?

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- Se puede traducir literalmente como "subtitulado de imagen".
- Consiste en la generación de descripciones textuales de una imagen o de sus escenas.
- Comúnmente se utilizan técnicas de procesamiento del lenguaje natural y de visión artificial de forma conjunta.
- Es parte de los esfuerzos para dar a las máquinas la capacidad de comprender e interpretar información de escenas visuales.

# ¿Qué es el image captioning?

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión



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 .

Figura: Ejemplo de descripción automática de una imagen. Tomado de *Flickr8k dataset*[1].

[1]HODOSH, Micah; YOUNG, Peter; HOCKENMAIER, Julia. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013, vol. 47, p. 853-899.

# ¿Cómo solucionarlo?

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

Es común que las soluciones a este problema utilicen más de una técnica de aprendizaje automático [2]:

- Reconocimiento de objetos.
- Clasificación de imágenes.
- Reconocimiento de patrones y texturas.
- Comprensión sintáctica y semántica.
- Generación de lenguaje natural.

[2] HOSSAIN, MD Zakir, et al. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR), 2019, vol. 51, no 6, p. 1-36.

# ¿Qué es un mecanismo de atención?

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- Es uno de los conceptos más importantes en el campo del aprendizaje profundo.
- Está basado en el comportamiento psicológico humano de atender ciertas partes del universo de información que se está procesando[3].
- En este trabajo se utiliza la *atención visual*, sigue el mismo principio, aplicado a información visual [4].

[3] VAN ZOMEREN, Adriaan H.; BROUWER, Wiebo H. Clinical neuropsychology of attention. Oxford University Press, 1994.

[4] ALLPORT, Alan. Visual attention. 1989.

# Trabajo previo

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

Pocos autores han trabajado con el problema de descripción de imágenes para el idioma español:

- GOMEZ-GARAY, Alejandro; RADUCANU, Bogdan; SALAS, Joaquín. Dense captioning of natural scenes in spanish. En Mexican Conference on Pattern Recognition. Springer, Cham, 2018. p. 145-154.
- MARTINEZ GUTIERREZ, Maria Fernanda. Automated Image Captioning: Exploring the Potential of Microsoft Computer Vision for English and Spanish. 2019. Tesis Doctoral. University of Geneva.

# Justificación del trabajo experimental

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- Se necesita medir el desempeño de los modelos *estado del arte* en distintos idiomas. El alcance de la aplicación de los modelos dependerá en gran parte de su desempeño en una variedad de idiomas.
- No existe una base sólida de investigación que permita comparar los resultados de estos modelos en el idioma español. En este trabajo se describe a profundidad el método de experimentación y las métricas utilizadas.
- No hay conjuntos de datos para descripción de imágenes en español, se propone y hace público el primer conjunto de datos sintético para este problema.

1 Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental

2 Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual

3 Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados

4 Conclusión
Conclusiones
Trabajo futuro

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

# Detalles generales

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- Propuesto por Kelvin Xu et al. en su publicación *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*[5] en 2015.
- Busca reproducir la capacidad humana de enfocar su atención en las características sobresalientes de una escena de forma dinámica.
- La arquitectura alinea una imagen de entrada con una palabra de salida.
- Es uno de los primeros intentos de utilizar mecanismos de atención en problemas distintos a la traducción automática.

[5] XU, Kelvin, et al. Show, attend and tell: Neural image caption generation with visual attention. En International conference on machine learning. 2015. p. 2048-2057.

# Arquitectura


Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

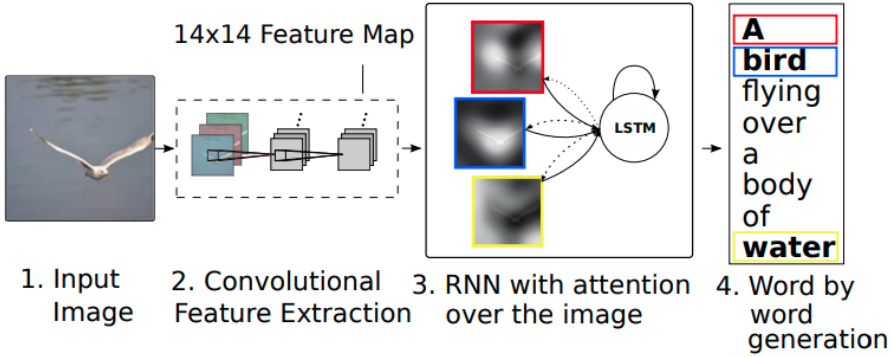




Figura: Estructura del modelo, tomado del artículo de Kelvin Xu et al.[6]

[6] XU, Kelvin, et al. Show, attend and tell: Neural image caption generation with visual attention. En International conference on machine learning. 2015. p. 2048-2057.

# Características convolucionales: el codificador

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- Se utiliza una ConvNet para extraer el conjunto de vectores de características de la imagen de entrada.
- Cada uno de estos vectores es una representación $D$-dimensional que corresponde a una parte de la imagen.
- Se extraen las características utilizando una capa convolucional anterior en vez de la completamente conectada.
  1. Permite obtener una correspondencia entre los vectores de características y las porciones de la imagen en 2 dimensiones.
  2. El decodificador puede ponderar un subconjunto de todos los vectores de características para enfocarse selectivamente en ciertas partes de la imagen.

# Generando las descripciones: el decodificador

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- Se utiliza una red Long Short-Term Memory (LSTM) para generar una palabra en cada paso de tiempo condicionado por:
  1. Vector de contexto.
  2. Estado oculto anterior.
  3. Las palabras generadas anteriormente.
- El vector de contexto es una representación dinámica de la parte relevante de la imagen de entrada en un instante de tiempo.
- La definición de la función $\phi$ determina si el mecanismo de atención ($f_{att}$) es estocástico o determinista.

# Mecanismos de atención

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

Se proponen dos enfoques para el entrenamiento del mecanismo de atención:

- "Soft" attention: Entrenable mediante métodos estándar de retropropagación del error. Basado en el mecanismo de Bahdanau et al. [7]
- "Hard" attention: Entrenable con aprendizaje por refuerzo, maximizando el límite inferior variacional. Equivalente al método REINFORCE propuesto por Williams [8].

[7] BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[8] WILLIAMS, Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 1992, vol. 8, no 3-4, p. 229-256.

1 Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental

2 Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual

3 Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados

4 Conclusión
Conclusiones
Trabajo futuro

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

# Traducción de Flickr8k

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

El primer paso para entrenar un modelo de descripción de imágenes en español es contar con un conjunto de datos que describa imágenes en este idioma.

- Consiste en un total de 8,092 imágenes, cada una con 5 descripciones para las entidades y eventos en las escenas. Tiene un total de 40,460 oraciones.
- La traducción se realizó de forma automática utilizando el sistema *Google Neural Machine Translation*[9](GNMT).
- Las traducciones pudieron heredar los sesgos del modelo GNTM, abriendo la posibilidad a errores gramáticos o sintácticos en la versión traducida.

[9]WU, Yonghui, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.

# Ejemplo de Flickr8k en español

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

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.

Figura: Ejemplo de descripción automática de una imagen en español[10].

- La versión traducida tiene un vocabulario de 12,439 palabras, 3,521 palabras extras que la versión en inglés.

[10] Disponible en: https://github.com/gallardorafael/ShowAttendTell_Flickr8k_Spanish

# Implementación del modelo

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- La implementación se realizó utilizando el lenguaje Python 3.
- Se utilizó el framework Tensorflow en su versión 2.
- El código está disponible como libreta de *Jupyter* en el directorio del artículo[11].

[11] Disponible en: https://github.com/gallardorafael/ShowAttendTell_Flickr8k_Spanish

# Características de la implementación

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- La extracción de características se realizó utilizando *transfer learning*, haciendo uso de una capa convolucional anterior a la capa completamente conectada del modelo InceptionV3, previamente entrenada sobre ImageNet.
- El codificador es una ConvNet completamente conectada.
- El decodificador es una red neuronal recurrente (RNN), específicamente una *Gated Recurrent Unit*(GRU).
- La implementación utiliza el mecanismo de atención determinista, basado en el propuesto por Bahdanau.

# Detalles sobre el entrenamiento

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- Se entrenó utilizando una GPU NVIDIA Tesla P100 con 16GB de VRAM.
- El procedimiento de extracción de características tomó 450 iteraciones y un total de 35 minutos de procesamiento.
- El entrenamiento del mecanismo de atención tomó un tiempo total de 56 minutos. A partir de la época 18 se comienza a ver un sobre ajuste, se da por terminado el entrenamiento.

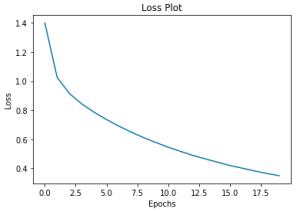



Figura: Pérdida del modelo a través de las épocas de entrenamiento.

# Resultados con puntaje BLEU $> 0{,}7$


Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión



Figura: Ejemplos de los resultados que obtuvieron un puntaje BLEU $> 0{,}7$.

# Plot del mecanismo de atención


Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

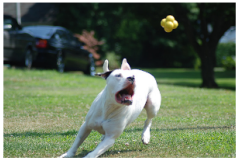




Figura: Visualización de los focos de atención.

# Resultados cuantitativos

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- La evaluación se realizó sobre 300 imágenes.
- Para calcular el puntaje BLEU, se tomaron como hipótesis las descripciones generadas y como referencias a las descripciones esperadas.

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

Figura: Mediciones del desempeño del modelo.

1 Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental

2 Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual

3 Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados

4 Conclusión
Conclusiones
Trabajo futuro

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

# Conclusiones

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- La media de 0.356 no es mala según nuestra referencia que considera una descripción mala a aquellas con BLEU $< 0{,}3$, sin embargo, está bastante alejada de una media que pueda ser considerada como buena (BLEU $> 0{,}7$).
- Un conjunto de datos no sintético, generado o supervisado por expertos del español podría eliminar los errores morfológicos y sintácticos de las predicciones.
- Las predicciones podrían mejorar si se utilizaran más ejemplos de entrenamiento.
- Es necesario construir bases para una evaluación de estos modelos en el idioma español, de modo que a futuro exista una investigación sólida en el área.

# Trabajo futuro

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

- Revisión por expertos del conjunto de datos Flickr8k para mejorar su efectividad en el español.
- Traducción de un conjunto de datos de mayor escala como Flickr30K (158,915 oraciones) o MS COCO (1,026,459 oraciones).
- Evaluación y comparación de modelos estado del arte en descripción de imágenes, entrenados sobre un conjunto de datos en español.
- La evaluación de modelos utilizando métricas más completas y variadas.

Introducción
¿Qué es el *image captioning*?
¿Qué es un mecanismo de *atención*?
Trabajo previo
Justificación del trabajo experimental
Descripción de la arquitectura
Detalles generales
Codificador
Decodificador
Mecanismos de Atención Visual
Métodos
Conjunto de datos Flickr8k
Implementación
Entrenamiento del modelo
Resultados
Conclusión

# ¡Gracias!