Anteproyecto: Descripción de Imágenes con Mecanismo de Atención

Micaela Del Longo

Facultad de Ingeniería, Universidad Nacional de Cuyo
Inteligencia Artificial II, Licenciatura en Ciencias de la Computación

Dr. Rodrigo Gonzalez
Dr. Jorge Guerra

23 de Enero de 2025

Contenidos

1. Introducción

En la era de los datos visuales, las imágenes han tomado un rol fundamental en la comunicación y el intercambio de información [1]. Un modelo de descripción automática de imágenes (Image Captioning) [2] combina procesamiento de imágenes y lenguaje natural para generar textos que describan contenido visual. Este tipo de modelos tiene aplicaciones diversas, como en tecnologías de accesibilidad para personas con discapacidades visuales, organización automática de imágenes en plataformas digitales, y mejora de la búsqueda basada en imágenes.

El presente anteproyecto propone el desarrollo de un modelo que combine Redes Neuronales Convolucionales (CNN) para la extracción de características de imágenes y Redes Neuronales Recurrentes (RNN) para la generación de texto que describa el contenido visual. Además, se incluirá un mecanismo de atención que permita al modelo enfocar regiones específicas de la imagen al generar cada palabra [3]. Inicialmente, se desarrollará una versión del modelo sin mecanismo de atención, la cual se utilizará como base para comparar los resultados obtenidos al implementar la versión con atención.

2. Descripción del Problema

El problema se enmarca en la categoría de aprendizaje supervisado [4, c. 12] y clasificación secuencial [4, c. 4]. Dado un conjunto de datos etiquetados con imágenes y sus respectivas descripciones textuales, el modelo debe aprender a asociar las características visuales de una imagen con las palabras y frases que describen su contenido.

3. Objetivos

Como se describió anteriormente, el objetivo principal de este proyecto es desarrollar un modelo basado en CNNs y RNNs que genere descripciones textuales precisas y coherentes para una imagen dada.

Para lograr este objetivo, se plantean los siguientes objetivos específicos:

- Extraer características visuales de las imágenes utilizando una CNN preentrenada (ResNet50 [5], [6]).
- Implementar una RNN (LSTM [7, p. 297]) para generar descripciones secuenciales de texto basadas en las características visuales.

- Crear un mecanismo de atención [7, c. 11.4] para permitir al modelo enfocar diferentes regiones de la imagen durante la generación de texto.
- Probar los modelos con distintas cantidades de capas y neuronas para evaluar su impacto en el desempeño.
- Evaluar los modelos utilizando métricas de evaluación de lenguaje natural como BLEU [8], METEOR [9] y/o CIDEr [10].
- Comparar los resultados obtenidos con y sin mecanismo de atención para analizar la mejora en la calidad de las descripciones generadas.

Además, se plantean objetivos adicionales: implementar beam search [11] para mejorar la generación de texto y realizar fine-tuning de ResNet [7, p. 234] para optimizar la extracción de características visuales según el dominio del conjunto de datos. Dichos objetivos se abordarán en caso de cumplir con los objetivos principales en tiempo y forma.

4. Datos de Entrada

Los datos de entrada al modelo consistirán en:

- *Imágenes*: Fotografías en formato RGB, que serán procesadas a través de una CNN para extraer mapas de características.
- *Descripciones textuales*: Secuencias de palabras en formato texto que describen el contenido visual de cada imagen.

Como posible fuente de datos, se considera el dataset Flickr8K [12] o Flickr30K [13], que contienen imágenes de *Flickr* con descripciones en inglés. También, se considera el conjunto de datos MS COCO [14], que contiene imágenes de escenas cotidianas y sus descripciones asociadas.

5. Algoritmos a Implementar

El modelo implementará una arquitectura que combina diferentes técnicas:

- *Redes Neuronales Convolucionales* (CNNs): ResNet50, previamente entrenada, para extraer representaciones de alto nivel de las imágenes.
- *Redes Neuronales Recurrentes* (RNNs): Una LSTM (Long Short-Term Memory) que tomará como entrada el vector de características de la imagen y generará las palabras de la descripción de forma secuencial.
- *Mecanismo de atención*: Una capa que calculará pesos de atención para cada región de la imagen, permitiendo que el modelo enfoque diferentes partes del mapa de características mientras genera cada palabra. Este mecanismo será implementado tras evaluar la versión inicial sin atención.

6. Descripción del Sistema

6.1. Preprocesamiento de Datos

En el caso de las *imágenes*, estás serán redimensionadas y normalizadas para cumplir con los requisitos de entrada de ResNet50. Posteriormente, se extraerán características de las capas intermedias de ResNet para obtener un mapa de características de alto nivel.

Mientras que las *descripciones textuales* serán tokenizadas, y se convertirá cada palabra en un índice utilizando un vocabulario creado a partir del conjunto de datos. Se aplicará padding para unificar la longitud de las secuencias de texto.

6.2. Arquitectura del Modelo

El modelo constará de las siguientes capas y componentes:

1. Extracción de características visuales:
   - Se utilizará una ResNet50 preentrenada en ImageNet.
   - El mapa de características extraído tendrá dimensiones fijas, representando las características visuales de diferentes regiones de la imagen.
2. Codificación del texto:
   - Una capa de embedding para convertir las palabras en vectores densos.
   - Una LSTM para procesar las secuencias de palabras generadas hasta el momento.
3. Mecanismo de atención (en la versión avanzada):
   - Una capa de atención que calculará pesos para cada región del mapa de características basado en la etapa actual de la LSTM.
   - Los pesos se combinarán con las características visuales para producir un vector de contexto.
4. Generación de texto:
   - La LSTM combinará el vector de contexto (características visuales ponderadas) con el estado actual para predecir la siguiente palabra.
   - Una capa densa con softmax se utilizará para predecir la probabilidad de cada palabra en el vocabulario.

6.3. Flujo de Trabajo

El flujo de trabajo del sistema se divide en dos etapas principales: entrenamiento e inferencia.

Durante el *entrenamiento*, las imágenes y las descripciones tokenizadas se pasan al modelo. Se calcula la pérdida basada en la predicción de palabras del modelo y las palabras reales de la descripción. El optimizador ajusta los pesos de las redes para minimizar esta pérdida.

En la etapa de *inferencia*, una imagen se pasa por ResNet para obtener características visuales. Luego, la LSTM y el mecanismo de atención (si está implementado) generan una descripción palabra por palabra. Si se implementa beam search, se generarán múltiples descripciones candidatas, y se seleccionará la más probable.

Bibliografía

[1] B. Bajarin, «The New Era of Visual Communication». Accedido: 19 de enero de 2025. [En línea]. Disponible en: https://www.vox.com/2015/6/16/11563610/the-new-era-of-visual-communication

[2] P.-Y. Chen y C.-J. Hsieh, «Image Captioning». Accedido: 16 de enero de 2025. [En línea]. Disponible en: https://www.sciencedirect.com/topics/computer-science/image-captioning

[3] «Image Captioning with Visual Attention». Accedido: 16 de enero de 2025. [En línea]. Disponible en: https://www.tensorflow.org/text/tutorials/image_captioning

[4] G. James, D. Witten, T. Hastie, R. Tibshirani, y J. Taylor, *An Introduction to Statistical Learning with Applications in Python*. 2023. [En línea]. Disponible en: https://www.statlearning.com/

[5] «ResNet-50 v1.5». Accedido: 20 de enero de 2025. [En línea]. Disponible en: https://huggingface.co/microsoft/resnet-50

[6] K. He, X. Zhang, S. Ren, y J. Sun, «Deep Residual Learning for Image Recognition». Accedido: 20 de enero de 2025. [En línea]. Disponible en: https://arxiv.org/abs/1512.03385

[7] F. Chollet, *Deep Learning with Python*, 2nd ed. Manning Publications, 2021.

[8] K. Papineni, S. Roukos, T. Ward, y W.-J. Zhu, «Bleu: a Method for Automatic Evaluation of Machine Translation». Association for Computational Linguistics, 1 de julio de 2002. doi: 10.3115/1073083.1073135.

[9] S. Banerjee y A. Lavie, «METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments». Association for Computational Linguistics, 1 de octubre de 2005. Accedido: 21 de enero de 2025. [En línea]. Disponible en: https://aclanthology.org/W05-0909/

[10] R. Vedantam, C. L. Zitnick, y D. Parikh, «CIDEr: Consensus-based Image Description Evaluation». 3 de junio de 2015. Accedido: 21 de enero de 2025. [En línea]. Disponible en: https://arxiv.org/abs/1411.5726

[11] K. Doshi, «Foundations of NLP Explained Visually: Beam Search, How It Works». Accedido: 18 de enero de 2025. [En línea]. Disponible en: https://towardsdatascience.com/foundations-of-nlp-explained-visually-beam-search-how-it-works-1586b9849a24

[12] C. Rashtchian, P. Young, M. Hodosh, y J. Hockenmaier, «Collecting Image Annotations Using Amazon's Mechanical Turk». Accedido: 21 de enero de 2025. [En línea]. Disponible en: https://hockenmaier.cs.illinois.edu/8k-pictures.html

[13] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, y S. Lazebnik, «Flickr30K Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models». Accedido: 21 de enero de 2025. [En línea]. Disponible en: https://bryanplummer.com/Flickr30kEntities/

[14] T.-Y. Lin *et al.*, «Microsoft COCO: Common Objects in Context». Accedido: 21 de enero de 2025. [En línea]. Disponible en: https://cocodataset.org/#home

[15] V. Patel, «IDM - The Evolution of Visual Communication: A Look into the History and Future of the Field». Accedido: 19 de enero de 2025. [En línea]. Disponible en: https://www.itm.edu/blog/idm-the-evolution-of-visual-communication-a-look-into-the-history-and-future-of-the-field/#:~:text=Visual%20communication%20has%20a%20long,through%20visual%20symbols%20and%20signs.