Universidad de La Habana
Facultad de Matemática y Computación

Textos alternativos para las colecciones de imágenes del repositorio digital de la Oficina del historiador de La Habana

Autores:
Amanda Cordero Lezcano
Ana Paula González Muñoz
Carlos Antonio Bresó Sotto
Christopher Guerra Herrero
Dennis Daniel González Durán
Marian Susana Álvarez Suri

Proyecto final de la asignatura
Aprendizaje de Máquinas

Enero 2025

Resumen

Abstract

Índice general

Capítulo 1

Introducción

Capítulo 2

Estado del Arte

El campo del *image captioning* ha evolucionado significativamente en la última década, impulsado por el avance de modelos de aprendizaje profundo. En este capítulo, se presentan los principales enfoques y modelos que han marcado hitos en esta área, destacando sus arquitecturas, metodologías y contribuciones.

2.1. Modelos Basados en Redes Recurrentes

Los primeros avances en generación de descripciones de imágenes se apoyaron en arquitecturas *encoder-decoder* con redes neuronales recurrentes (RNN). Uno de los primeros modelos destacados fue **Show and Tell** [1], que propuso un enfoque generativo basado en una arquitectura recurrente profunda. Este modelo utilizó una red convolucional (CNN) para la extracción de características visuales y una red LSTM para la generación de texto. Fue entrenado para maximizar la probabilidad de generar una descripción textual dada una imagen, aprendiendo únicamente a partir de descripciones de imágenes.

Posteriormente, **Show, Attend and Tell** [2] introdujo mecanismos de atención visual, permitiendo que el modelo enfocara diferentes regiones de la imagen en cada paso de generación. Este enfoque mejoró la calidad de las descripciones y presentó una formulación matemática más avanzada para el cálculo de la atención.

2.2. Modelos Basados en Redes Convolucionales

En un intento por superar las limitaciones de las RNN, **Convolutional Image Captioning** [3] propuso una arquitectura basada en CNNs para la generación de texto. Este modelo demostró que las CNNs pueden superar a las LSTM en tareas de *captioning*, especialmente cuando se combinan con mecanismos de atención, mitigando problemas como el desvanecimiento del gradiente. Se realizó un análisis detallado que proporcionó razones convincentes para preferir los enfoques de generación de lenguaje convolucional.

2.3. Atención y Modelos Jerárquicos

Otro enfoque relevante fue el modelo **Bottom-Up and Top-Down** [4], que implementó un mecanismo de atención jerárquico basado en la segmentación de objetos dentro de la imagen. Este modelo propuso una estrategia combinada de atención Bottom-Up y Top-Down, permitiendo un análisis más profundo de la imagen a través de múltiples pasos de razonamiento. En el mecanismo Bottom-Up, basado en Faster R-CNN, se proponen regiones de la imagen, cada una con un vector de características asociado. Por otro lado, el mecanismo Top-Down determina las ponderaciones de estas características, lo que permite una atención más precisa a nivel de objetos y regiones destacadas.

Knowing When to Look [5] propuso un mecanismo de atención adaptativo mediante un centinela visual, que decide cuándo prestar atención a la imagen y cuándo confiar en el contexto textual generado previamente. Este modelo abordó una limitación clave de los enfoques tradicionales de atención, que forzaban la atención visual en cada palabra generada, incluso cuando no era necesaria. En lugar de ello, el modelo introdujo un centinela visual que, en cada paso de tiempo, determina si es necesario extraer información de la imagen y, de ser así, selecciona las regiones relevantes. Esto permite al decodificador alternar de manera inteligente entre la información visual y el contexto lingüístico, mejorando la precisión y fluidez de las descripciones generadas.

2.4. Transformers y Modelos Multimodales

Con la llegada de los Transformers, los modelos de *captioning* han adoptado arquitecturas más avanzadas. **Multimodal Transformer** [6] exploró la representación visual multi-vista para mejorar la generación de texto, am-

pliando el éxito del modelo Transformer en traducción automática a la tarea de subtítulos de imágenes. A diferencia de los enfoques tradicionales basados en codificador-decodificador, que utilizan una CNN para extraer características visuales y una RNN con mecanismos de atención para generar texto, este modelo propone un bloque de atención unificado que captura simultáneamente interacciones intra e intermodales. Esto permite un razonamiento multimodal más complejo, integrando tanto la autoatención (interacciones intramodales) como la coatención (interacciones intermodales) en una arquitectura modular y profunda.

BLIP [7] amplió la capacidad de los modelos al abordar tanto la generación como la comprensión de imágenes y videos, logrando mejoras significativas en una amplia gama de tareas de visión y lenguaje. A diferencia de los modelos preentrenados existentes, que suelen especializarse en tareas de comprensión o generación, BLIP propone un marco de preentrenamiento unificado que se transfiere de manera flexible a ambas.

Además, **CLIP** [8] introdujo el aprendizaje multimodal a gran escala utilizando correspondencias entre imágenes y texto en internet. A diferencia de los sistemas de visión por computadora tradicionales, que se entrenan para predecir un conjunto fijo de categorías de objetos, CLIP aprende representaciones visuales directamente a partir de texto sin procesar, lo que le permite capturar una gama más amplia de conceptos visuales. Este enfoque se basa en una tarea de preentrenamiento simple pero efectiva: predecir qué descripción de texto corresponde a una imagen dada.

ViT [9] aplicó Transformers directamente a imágenes dividiéndolas en secuencias de parches, lo que representó un cambio significativo en el procesamiento de información visual. A diferencia de los enfoques anteriores, que combinaban Transformers con redes convolucionales (CNNs) o reemplazaban solo ciertos componentes de las CNNs, ViT demostró que una arquitectura basada únicamente en Transformers puede lograr un rendimiento excepcional en tareas de clasificación de imágenes. Este enfoque elimina la dependencia de las CNNs, procesando las imágenes como secuencias de parches lineales y aplicando mecanismos de atención pura para capturar relaciones globales entre ellos.

2.5. Resumen y Tendencias Actuales

La evolución del *image captioning* ha pasado de modelos basados en RNNs con atención visual a enfoques más sofisticados que integran Transformers y aprendizaje multimodal. Modelos recientes como BLIP y CLIP han

demostrado que la combinación de visión y lenguaje en grandes volúmenes de datos puede llevar a mejoras sustanciales en la generación y comprensión de imágenes.

Las tendencias actuales apuntan a modelos más eficientes y escalables, con capacidades mejoradas en la generación de texto y una mayor comprensión del contexto visual. El impacto de estas tecnologías se extiende más allá del *image captioning*, beneficiando tareas como búsqueda visual, generación de contenido y asistencia en accesibilidad.

Capítulo 3

Propuesta

Se propone un sistema para la generación de texto alternativo a partir de las imágenes del repositorio digital del patrimonio cultural de la Oficina del Historiador, combinando múltiples modelos de aprendizaje profundo. La metodología se basa en un enfoque de evaluación comparativa entre diferentes modelos de generación de texto a partir de imágenes, utilizando un criterio de selección basado en la similitud semántica con el contenido visual.

En primer lugar, se emplea el modelo BLIP (Bootstrapped Language-Image Pretraining) para generar una primera descripción de la imagen. Posteriormente, la imagen es procesada por el modelo ViT (Vision Transformer) en conjunto con GPT-2, obteniendo una segunda descripción independiente. Finalmente, el modelo CLIP (Contrastive Language-Image Pretraining) se utiliza como mecanismo de selección, comparando las dos descripciones generadas y eligiendo la que presente una mayor correspondencia semántica con la imagen de entrada.

Capítulo 4

Desarrollo de la propuesta

Capítulo 5

Resultados

Capítulo 6

Conclusiones

Referencias

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2015. (Citado en la página 6).

[2] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. (Citado en la página 6).

[3] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. *arXiv preprint arXiv:1805.09019*, 2018. (Citado en la página 7).

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *arXiv preprint arXiv:1707.07998*, 2018. (Citado en la página 7).

[5] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887*, 2017. (Citado en la página 7).

[6] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *arXiv preprint arXiv:1905.07841*, 2019. (Citado en la página 7).

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. (Citado en la página 8).

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin,

Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. (Citado en la página 8).

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021. (Citado en la página 8).